POPULATION GENETICS AND SPECIATION IN THE PLANT GENUS *SILENE* (SECTION *ELISANTHE*)

by

ANDREA LOUISE HARPER

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Biosciences
The University of Birmingham
2009

ABSTRACT

This thesis is concerned with speciation and population genetics in the plant genus *Silene* (section *Elisanthe*). The introductory chapter is a literature review covering characteristics of the species studied, and the current literature on their evolutionary dynamics and population genetics.

The second and third chapters cover techniques used in all experiments, such as DNA extraction, sequencing and genotyping protocols, and explain the rationale behind the initial experimental design.

The fourth chapter focuses on the multi-locus analysis of autosomal gene sequences from *S. latifolia* and *S. dioica*. The relationship between the two species was investigated using various analyses such as isolation modeling and admixture analysis providing estimates of evolutionary distance and extent of historical gene flow. The maintenance of the species despite frequent hybridization at present-day hybrid zones is discussed.

The fifth chapter discusses *S. diclinis*, a rare endemic found only in Valencia, Spain. The nature of population structuring and the evolutionary history of this species were investigated using a multilocus approach incorporating individuals

from *S. diclinis* populations. The causes of the restricted distribution and low population size of this species is discussed

The concluding chapter discusses how the species evolved from a common ancestor amidst changing climatic and environmental conditions.

This thesis is dedicated to Mum and Dad for their never-ending love and support. All my love and thanks. x

ACKNOWLEDGEMENTS

Firstly, I would like to thank Dr. Dmitry Filatov and Dr. Sue Armstrong for their patience, supervision and assistance, which is greatly appreciated.

I also owe thanks to the BBSRC for supporting me financially through the last three years.

I would also like to thank several members of my lab that have helped me in various ways. Special mention to Dr. Graham Muir for his contribution to this project and the long discussions about it! Also, Katie Ridout for her advice and expertise on programming matters, Serene Hargreaves for helpful discussion and advice, and Dr. Antonina Vontintseva for her technical expertise. I would also like to thank Drs Elaine Howell, Maxim Kapralov and Chris Dixon, and Constantinos Groutides, all of which have at some time have been forced to listen to me talk shop.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 General

The genus *Silene*, the Campions, is composed of around 700 individual species including annuals, biennials and perennials, which favour habitats as varied as meadow, woodland and mountain, and species in this genus have colonised Asia, Europe, Australasia and the Americas. Several sections and species in this genus have become the focus of research. *Silene vulgaris*, the Bladder Campion, is interesting due to its well-documented heavy-metal tolerance. *Silene latifolia* is at the centre of research on the early evolution of sex chromosomes and dioecy, and is also used as a model for studying its frequent infection from the anther smut *Microbotryum violaceum*. Several members of the genus such as *S. latifolia*, *S. gallica*, *S. vulgaris* and *S. noctiflora* have also become invasive in certain regions. *S. latifolia* for instance, has become an invasive pest in North America where it was introduced around 200 years ago.

1.2 *Silene* section *Elisanthe*

Many of the interesting characteristics associated with this genus are found within the section *Elisanthe* including the species *S. latifolia*, *S. dioica*, *S. diclinis*, *S. heuffelii* and *S. marizii*. This section is fairly diverse with a variety of favoured habitats and ranges. *S. latifolia* has a broad range which covers much of Eurasia and North America. It is pollinated by the Lychnis Moth *Hadena*

bicruris and is commonly found growing in cultivated and disturbed ground (Prentice, 1988). Its close relative *Silene dioica* shares much of the same range, yet has become largely pollinated by bumblebees (*Bombus* spp.) and favours a more established woodland habitat. The two species are also easily distinguished from each other by their flower colour, *S. latifolia* being white and *S. dioica* pink. This may be complicated however, as it is known that these two species may readily hybridize in regions where the two species are found in close proximity, producing an array of intermediate flower shades (Baker, 1950).

Silene diclinis is another closely related species which grows in Europe. Unlike *S. latifolia* and *S. dioica*, however, this species is found only in a small area of Valencia, Spain, and may be close to extinction. Consequently, it has been entered on the 2008 IUCN Red List as an endangered species probably numbering less than 2000 individuals and thought to be decreasing in numbers (Montesinos & Güemes, 2006). It shares characteristics with both *S. latifolia* and *S. dioica* (see Figure 1.1. *Silene* Flowers.). Like *S. dioica* its bright pink flowers are mainly pollinated by bumblebees, but it prefers cultivated and disturbed ground like *S. latifolia*. The two remaining species in the section, *S. heuffelii* and *S. marizii* are also endemics. *S. heuffelii* occurs in the northern Balkan regions and the Carpathian Mountains at altitudes above 700–800 m. *S. marizii* is found in Portugal and Spain associated with rocky habitats (Prentice 1976).

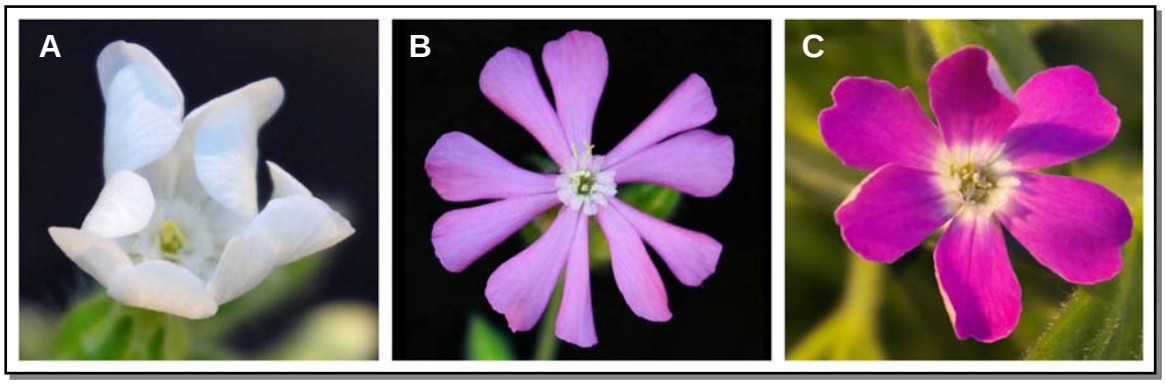

Figure 1.1. *Silene* Flowers.
A. *Silene latifolia*, B. *Silene dioica*, C. *Silene diclinis*

Despite the variation in pollinators, habitats and ranges exhibited in this section, all species are still inter-fertile, suggesting that these species have diverged from each other relatively recently. We can make an approximate estimate of the age of the section from the level of synonymous divergence between *S. latifolia* and the close relative *S. vulgaris* (which is not a member of the section *Elisanthe).* This indicates that the age of the section is no older than around 10-20 million years (Filatov & Charlesworth, 2002).

Little is known about the evolutionary history of the species in the section *Elisanthe*. The species may have diverged relatively recently and the fact that they can still hybridize to produce fertile offspring provides the possibility that the evolution of these species may be more complicated than it at first appears. It is not uncommon for species to continue to hybridize during and immediately after the speciation process, as can be seen in *Helianthus* species (Yatabe et al., 2007; Strasburg & Rieseberg, 2008). In fact, there is evidence that this may

have occurred in our own evolutionary history following the divergence of humans and Neanderthals (Wall & Hammer, 2006). The species within the section *Elisanthe* have become phenotypically and phenologically distinct however, displaying a range of morphological (particularly flower) differences, distributions, habitat preferences, pollinators and scent and flowering-time traits among others (Waelti *et al*., 2008).

In principle, gene flow could occur during and after speciation due to a persisting "porous" species boundary such as has been suggested to occur in butterflies and crickets (Kronforst, 2008; Shaw & Danley, 2003). This boundary would allow transfer of genes between species whilst protecting speciation important genes from introgression. Alternatively, reproductive barriers may have developed in these closely-related species as they adapt to different ecological niches reducing introgression on a large scale.

1.3 Gene Flow between *S. latifolia* and *S. dioica*

If gene flow and introgression have been occurring since the divergence of the species, this is likely to be most evident between *S. latifolia* and *S. dioica*. Both of these species have large effective population sizes and overlapping distributions, and are known to hybridize at natural hybrid zones, as well as the ability to be cross-fertilised with high efficiency in the greenhouse (Baker, 1950).

Many studies have focused on the nature of introgression between these two species by studying the characteristics of plants at and around hybrid zones. One such study (Minder *et al*., 2007) investigated *S. latifolia* and *S. dioica* individuals from within hybrid zones, outside of hybrid zones, and suspected hybrids. They found that there was a lack of true intermediate hybrids, and that hybrids were commonly back-crossing into parent species. Despite the lack of intermediate hybrids, they estimated that introgression was occurring at extremely high levels. Assuming that this has been occurring at various hybrid zones since the divergence of the two species, the signature of gene flow would be expected to be visible throughout the genome of both species, not only around hybrid populations but all over their shared distribution. Conversely, if this hybridization has only been occurring for a relatively short time due to more recent secondary contact it may not be possible to detect gene flow and shared alleles would be an indication of shared ancestry.

There is some evidence from outside of hybrid zones that gene flow has occurred between *S. latifolia* and *S. dioica* in the past. Ironside & Filatov (2005) investigated population structure and the relative introgression of Y and X linked gene DD44X/Y in *S. latifolia* and *S. dioica*. Population sub-division was found to be higher in the Y-linked copy of the gene rather than the X-linked copy (Table 1.1). Monte Carlo Markov Chain simulations also suggested that introgression had occurred on the X-linked copy, but not the Y, and that background selection

was the likely cause of the low Y diversity seen in *Silene* rather than selective sweeps, as strong population structure remains on the Y (Table 1.1).

Table 1.1. FST estimates of population structure for DD44X and Y.
(Ironside & Filatov, 2005).

Locus	F_{ST} between *S. latifolia* and *S. dioica.*
DD44X	0.05<0.17
DD44Y	0.82<0.91

Muir and Filatov (2008) concluded that gene flow has also occurred between the two species in respect to the chloroplast genome, as *S. latifolia* and *S. dioica* have been shown to have lower levels of diversity than expected coupled with low population structure. The marked difference between the population structure of the Y and chloroplast is illustrated by the respective neighbour-joining trees (see Figure 1.2 and Figure 1.3). Partitioning of the variation across different hierarchical levels in the chloroplast showed that most of the variation was partitioned within populations rather than between species or among populations. This may be explained if a selective sweep crossed the sequence boundary, with the advantageous alleles having dragged linked alleles (effectively the whole chloroplast genome due to extremely low recombination rate) along with it. As a result, the chloroplast sequence is homogenized across both species.

The above studies have suggested that there has been (or is) introgression of the X chromosome and chloroplast but not the Y chromosome between these

two species. It is not possible from these studies to conclude definitively that the Y chromosome is unusual in this way, but limitation of Y chromosome introgression is well characterized in model species such as rodents (Vanlerberghe *et al*., 1986, Jaarola *et al*., 1997).

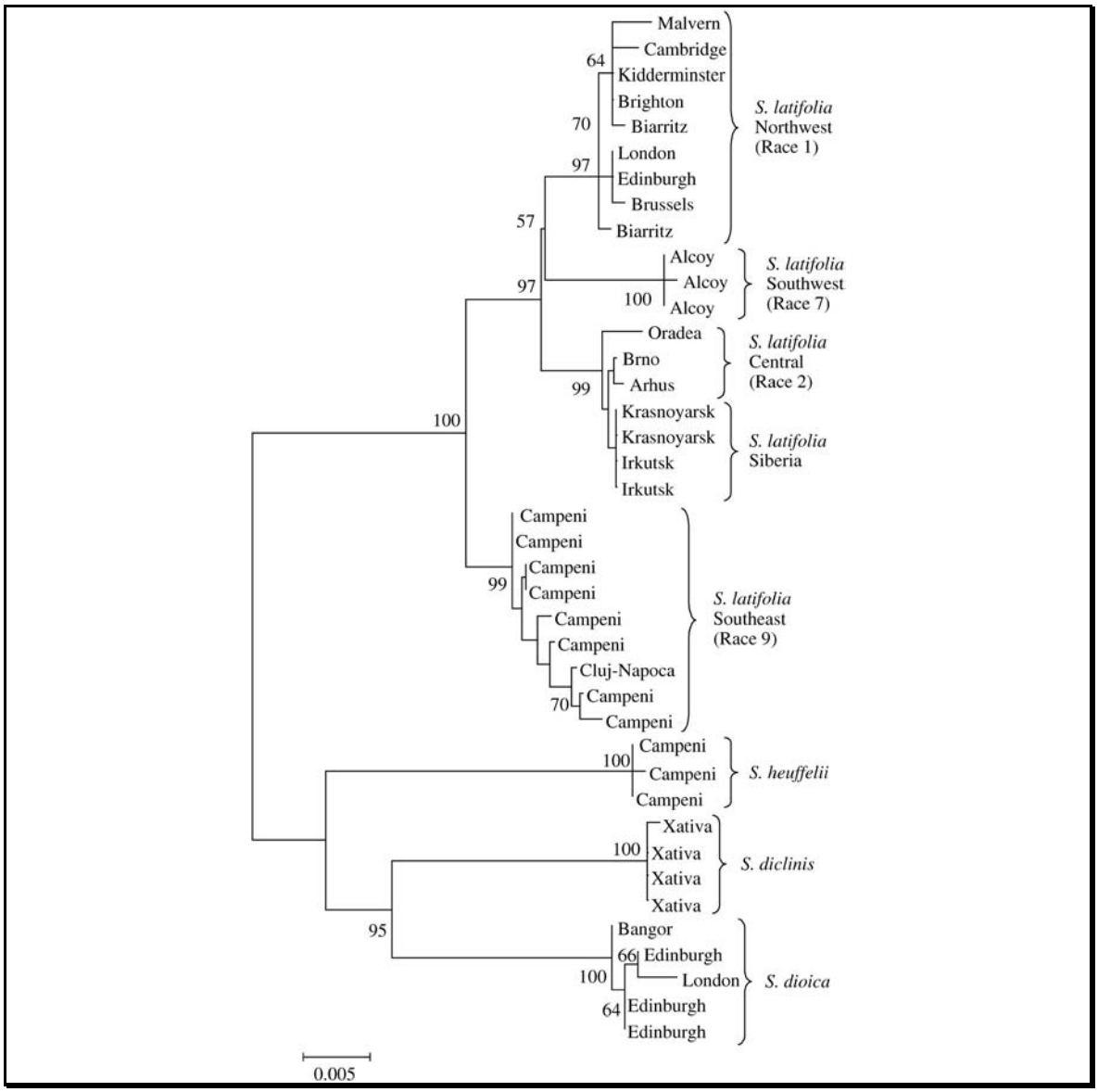

Figure 1.2 Neighbour-joining tree of *Silene* DD44Y sequences, showing bootstrap support for each node
(Ironside & Filatov, 2005).

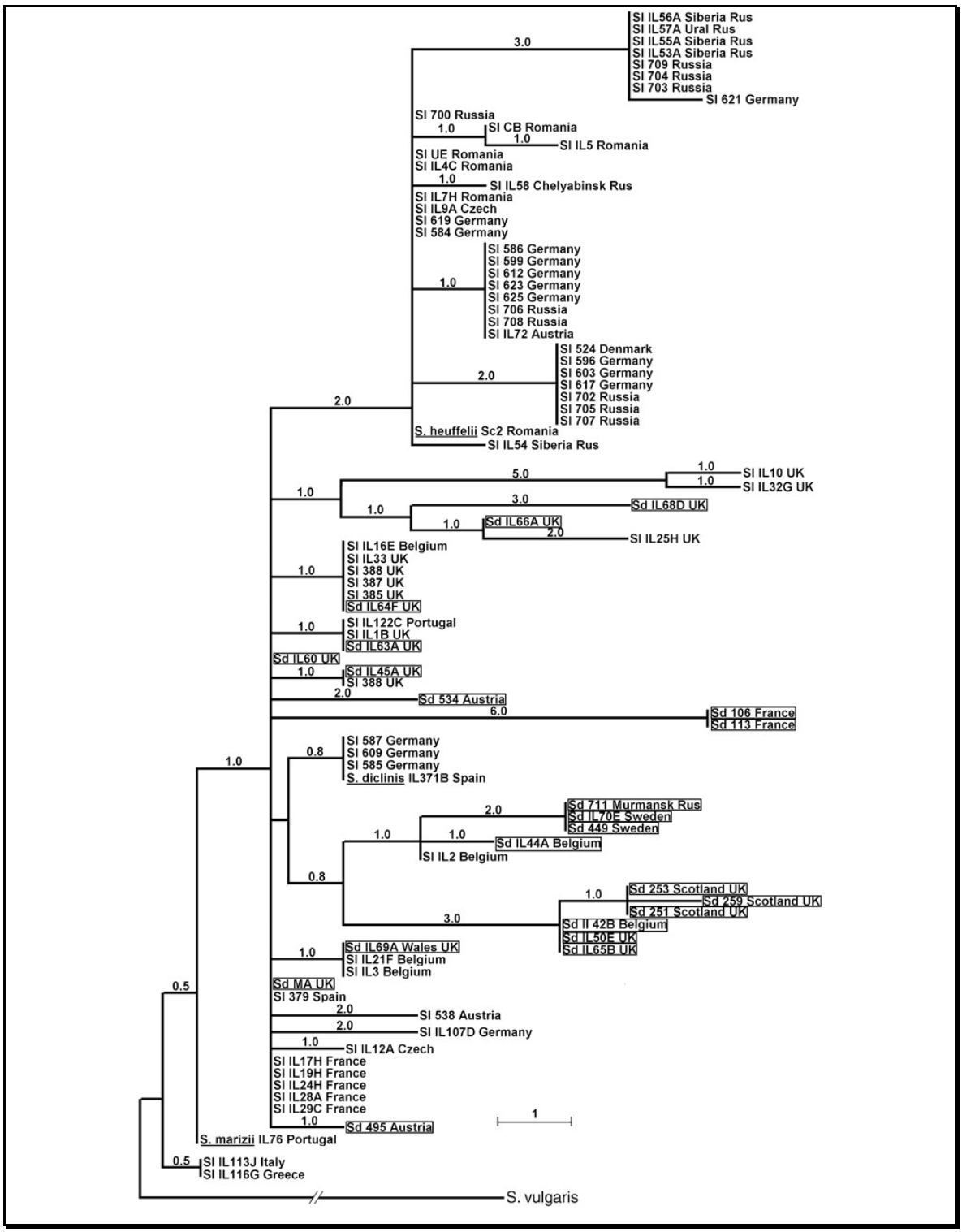

Figure 1.3 Neighbour-joining tree of dioecious *Silene* (section *Elisanthe*) chloroplast sequences (concatenated matK + trnT-trnL-trnF) rooted by an outgroup *S. vulgaris*.
(Muir & Filatov, 2008)

Studies concerned with gene flow between species may also provide indications of the level of reproductive isolation. There may be several pre-zygotic barriers between *S. latifolia* and *S. dioica*. They have different habitat preferences, reducing the occasions that the two species are likely to come into close contact. They have developed different flowering time phenologies, with *S. latifolia* pollinated mainly during the night, and *S. dioica* during the day, although both species may have flowers open during the day, particularly dawn or dusk, and in certain weather conditions (Jurgens *et al*., 1996). Perhaps most importantly, the two species have become principally pollinated by different insects. Although *S. latifolia* is primarily pollinated by the Lychnis Moth *Hadena* and *S. dioica* by the bumblebee, the presence of natural hybrids suggests that one or both of these insects will pollinate the opposite species or that other minor pollinators may be involved, as reproductive isolation of the two species is incomplete (Waelti *et al*., 2007).

Post-zygotic reproductive isolation is harder to determine. There may be some evidence within recent hybrid zone studies. Both Minder *et al*. (2007) and Karrenberg and Favre (2008) have found evidence that intermediate hybrids are rare or largely absent. This could either suggest that the hybrid zones are old and introgression has ceased or that post-zygotic barriers have evolved producing less fit hybrids. Minder *et al*. found that one of their populations was in linkage equilibrium suggesting an old zone, yet another population had genes in linkage disequilibrium (suggesting recent introgression) yet still with far fewer

intermediate hybrids than would be expected. Karrenberg and Favre saw a similar lack of intermediate hybrids, and found little evidence for large-scale introgression into pure populations. These studies suggest that the two species are not able to produce a large number of hybrids and the hybrid zones may be transient. Although greenhouse experiments have shown that F1 hybrids between the two species are highly fertile and can cross and backcross easily, this does not mean that post-zygotic barriers are not in effect. Hybrid fitness may be affected by ecological factors. For example, available resources may not be favourable for intermediate hybrids.

A genome-wide study focusing on hybrid zone effects could provide a wealth of interesting material. It is important to note however, that it would describe only the specific hybrid zone studied which may be more or less conducive to gene flow than hybrid zones found elsewhere. It would also tend to characterize the recent history of the species and the level of introgression seen in these relatively rare locations will be vastly inflated compared to the introgression seen in the species as a whole. It is possible however that these hybrid zones could act as bridges, allowing a much smaller number of introgressed genes to seep into the allopatric individuals, and that once introgressed the patterns of linkage disequilibrium will be erased by recombination.

A study based only on a single gene will also be limited in that it will not be able to compensate for demography, and results must be cautiously interpreted.

Perhaps the best way to understand the history of *S. latifolia* and *S. dioica* in relation to each other may be to adopt a multilocus approach to account for demography and compensate for the natural variation seen between genes. Samples from across as much of the natural distribution of these species as possible would also be more useful. Looking at random individuals across the distribution allows a “baseline” measurement of gene flow to be calculated i.e. the historical level of seepage of genes across the species boundary since their divergence from each other. Once again caution must be employed however as the species’ shared ancestry could lead to a false signal of introgression due to incomplete lineage sorting. For this reason, recently developed coalescent methods providing a means of distinguishing between incomplete lineage sorting and interspecific gene flow should be utilized.

1.4 Characteristics of the endemic *S. diclinis*

Silene diclinis grows in one particular area of South Eastern Spain in a region of Valencia near the town of Xàtiva. The species is sub-divided into several populations which have been monitored since 1986 and are classified on the International Union for Conservation of Nature and Natural Resources (IUCN) 2008 Red List as endangered (Montesinos & Güemes, 2006). The sizes of these populations vary from a few dozen to several hundred individual plants. The largest population is situated in Plà de la Mora, and numbers somewhere in the region of 1000 individuals. Some of the populations are included inside the

micro-reserves that were established in the area in the mid-1990s, but the majority (particularly the smaller) populations are located outside these reserves. Conservation measures currently in place are limited, and are centred on a seed bank collection from individuals in confirmed populations (of which there are five, although unconfirmed populations also exist).

There are several reasons behind the apparent decline in numbers of *Silene diclinis*. Firstly it is not a competitive species, and can only survive and multiply where the native thicket and scrub are not dominant. For this reason, it is often found on the borders of cultivated dry-lands (often carob plantations) where the scrub has been cleared and the ground may be slightly disturbed, but not intensively farmed. This niche has its associated problems however, and one of the most common factors frequently cited as the reason for the decline of *S. diclinis* is destruction of this habitat by change of farming practices in the area. It is also threatened by fire to a lesser extent, but all threats must be taken seriously when the number of remaining individuals is so low.

The level and organization of genetic diversity remaining in *S. diclinis* is unclear, but there are several hypotheses that can be drawn about endemic species such as *S. diclinis*. The first is that there will be some level of genetic differentiation between the spatially separated populations. As yet, little is known about the nature of population structure in *S. diclinis*. Allozyme studies such as those by Prentice (1984a) suggest that population structure between

populations is very low and that the majority of the genetic variation is situated within the populations which may indicate undetected fine-scale sub-population structuring.

We can also speculate that there will probably be some level of inbreeding occurring. This is an intuitive assumption due to the low numbers, subdivision of populations and the distance limited foraging behaviour of the principal pollinator, bumblebees (Osborne *et al*., 2008). The only possible escape from severe inbreeding depression would be migration of new alleles into the populations from genetically distinct populations (which may not exist) or other species. *S. latifolia* is the only other species in the section *Elisanthe* currently overlapping in distribution, and despite hybrids possible in greenhouse crosses, they are not known to occur under natural conditions. There is evidence that chromosomal translocations have evolved in *S. diclinis* producing a neo-XY sex chromosome system which may be responsible for reproductive isolation between the two species (Howell *et al*. in press). This would limit migration of alleles into *S. diclinis* from *S. latifolia*.

It is also important to note that *S. diclinis*, like *S. latifolia*, appears to have a stable sex-ratio bias both in its natural habitat and greenhouse conditions (Prentice, 1984). The bias favours females in a 60:40 ratio. Prentice hypothesized that this could be due to incidental effects of the recently evolved X-Y sex determination system with the X-transmitting pollen being more

competitive than that of the Y-transmitting pollen. This ratio will have the effect of further reducing the effective population size in *S. diclinis*.

The effect of low effective population size is that the power of selection is much reduced, and drift effects begin to shape the patterns of variation (Barrett & Kohn, 1991). Deleterious mutations are able to subsequently build up in the species without the purifying effects of background selection, and overall fitness will decline. Left unchecked this process could lead to the demise of the species. The implications for *S. diclinis* will therefore be serious. As well as possible damage to its habitat being a danger to its survival, a further contribution to its possible extinction could be genetic degeneration due to the extremely low effective population size. It therefore becomes important to establish the patterns and extent of remaining variation in the species to ensure that conservation efforts can encompass as much variation as possible. To capture the level of diversity it is necessary to sample from as many individuals from as many locations across the distribution of the species as possible. Spatially isolated populations with varying sizes may well harbour different variants to each other. It is also important to incorporate different loci into the analysis due to the expected stochastic effects of drift in small populations and to compensate for demography.

1.5 Phylogeography

Many species families that exist today have evolved from common ancestors due to environmental pressure. In some cases, this may be localized pressure such as adaptation to different ecological niches due to lack of resources, or geographical separation creating rapid species explosions such as that of the genus Scheidea on the Hawaiian Archipelago (REF). Occassionally, however, global events may be attributed to the emergence of many different species in many locations. One of these major events was the last great ice age during the Pliocene which saw ice sheets growing from the Arctic and temperature drops affecting much of the Northern Hemisphere.

The major ice sheets associated with the Pliocene cooling began to expand around 2.4 million years ago (Webb & Bartlein, 1992) and the severity of the ice ages increased around 700,000 years ago. This period was punctuated by relatively short interglacials which saw the ice sheets recede slightly and the amount of land habitable to flora increase, only to decrease again with the next cycle. During the colder periods, the large Scandinavian ice sheet covered much of the British Isles and Northern Europe with smaller ice sheets also forming across the mountain ranges of the Cantabria, Pyrenees, Alps, Transylvania and Caucasus. Between the ice sheets, most of Northern Europe was polar desert, cold-steppe and tundra (see Figure 1.4).

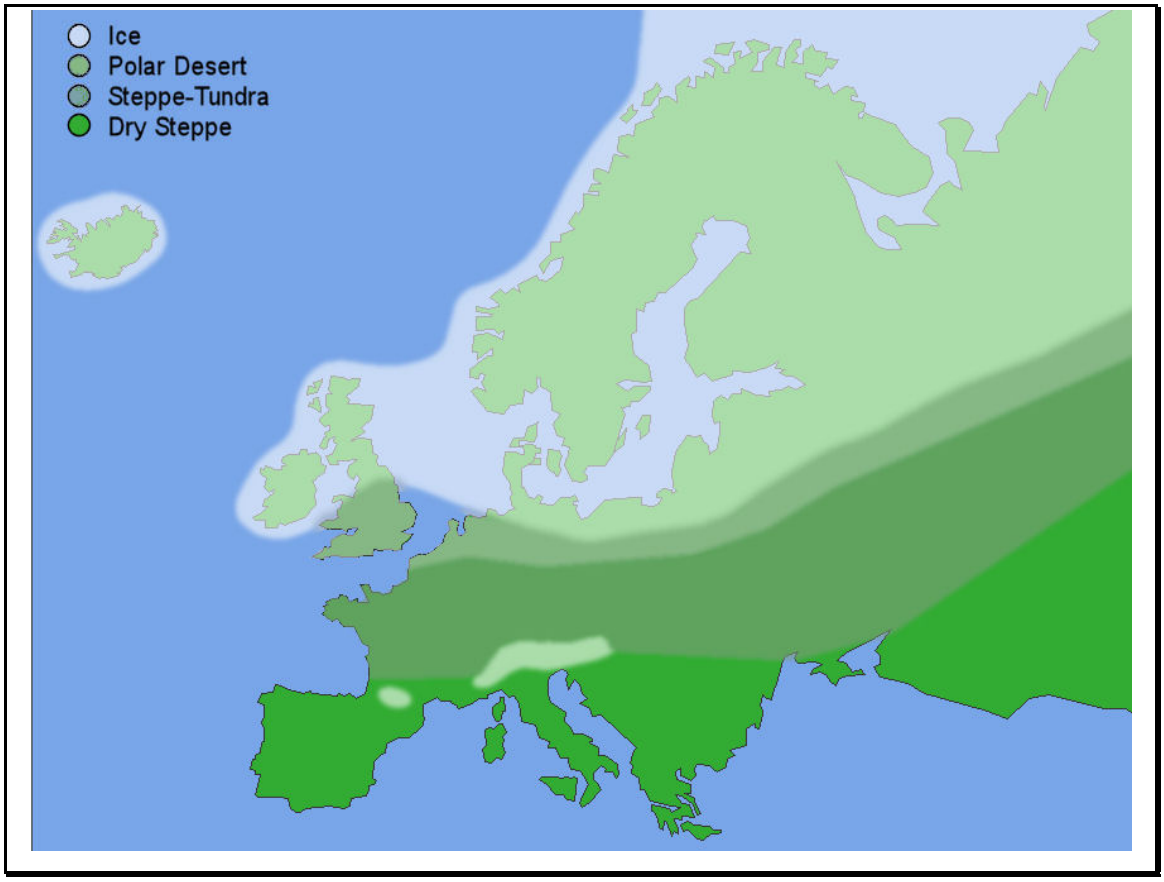

Figure 1.4. European Glacial Maximum, 150,000 years before present. Present coastlines are shown, although sea levels were approximately 100m lower than present. (adapted from van Andel & Tzedakis, 1996)

Plants now found in Europe would have been forced south of the ice and tundra into refugia such as the South of Spain, Italy, the Balkans and North Africa, allowing new species to begin evolving under isolation and with different environmental conditions. The climate began to warm around 18,000 years ago, and the present interglacial that we are now in stabilized around 8-10,000 years ago (Hewitt, 1996). This milder climate saw the ice sheets recede and the land become reclaimed by species that were previously limited to more southern

regions. The new species that had been formed by the ice age could now begin range and population expansions further north.

It is likely that the Pliocene cooling would have been responsible for the evolution of many flowering plants such as the Campions, and if not responsible, it will have had a profound impact on them forcing them into extreme bottlenecks.

As well as climatic events, humans may also have had an effect on the evolution and spread of new species. With the onset of the interglacial, Man could also expand his range, bringing agriculture to Northern Europe via several paths as shown in Figure 1.5. As well as bringing new crops and animals, they will also have created new ecological niches by clearing and disturbing the land, and possibly allowing easier dispersal of plant species from the south, allowing them to overcome geographical obstacles such as mountain ranges and rivers.

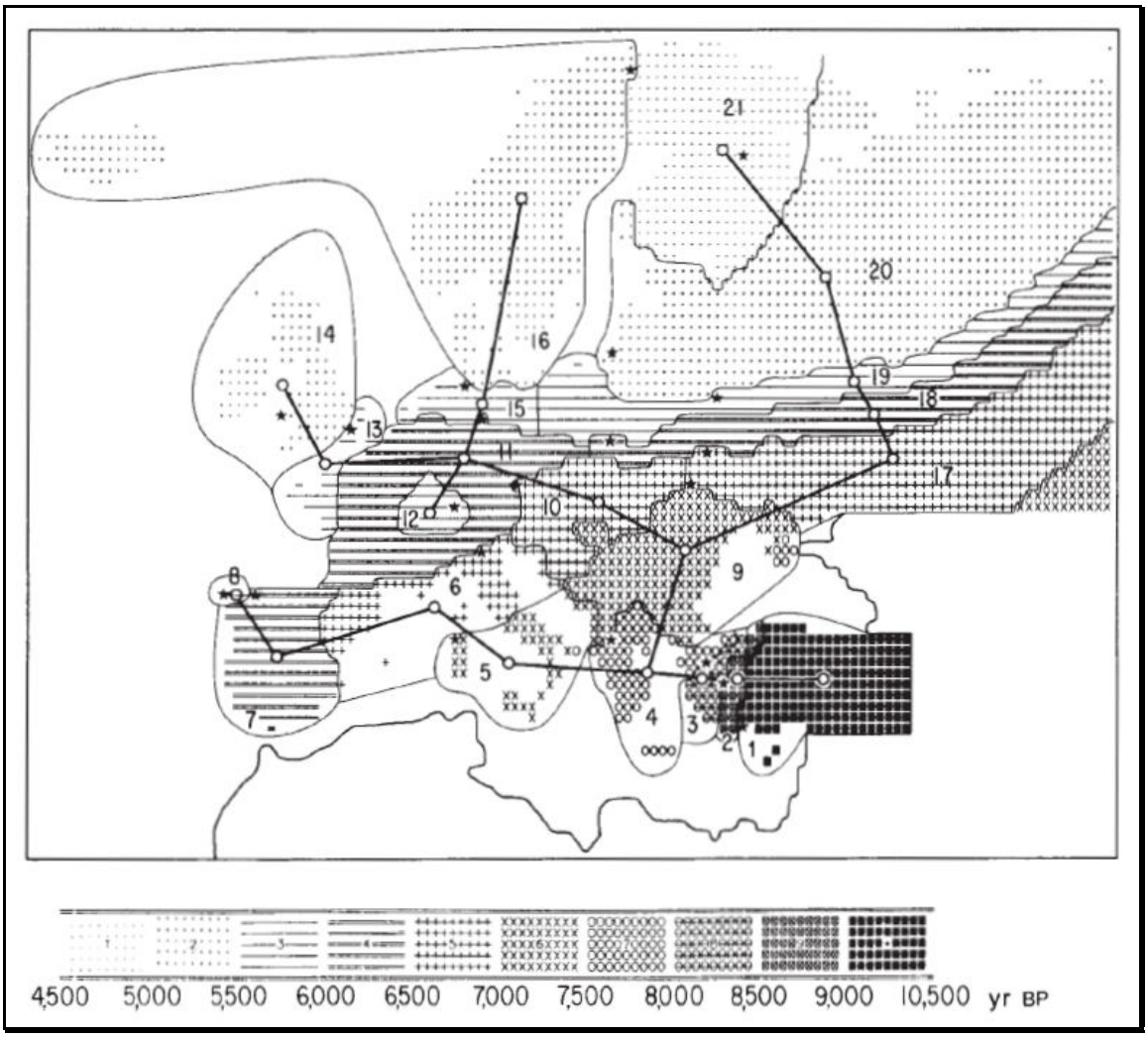

Figure 1.5. Contour map of the onset of agriculture.
(Sokal *et al.*, 1991) Contours mark 500 year intervals as identified in the key. Latest pixels identified with a star.

This thesis attempts to uncover the secrets of the evolution of the species of *Silene* section *Elisanthe*. It was decided that the best approach for studying the nature of variation, population structure and evolution of these three species was to incorporate samples covering as much of their natural distributions as possible. The samples could then be used for extraction of genomic DNA, PCR

amplification and sequencing of coding sequence, and subsequent identification of autosomal loci by segregation analysis. The methods used to do this are discussed in Chapter 2.

The neutral markers harvested from these processes could then be used to calculate the levels of genetic diversity, divergence, population differentiation, linkage disequilibrium and neutrality. To assess levels of migration between *S. latifolia* and *S. dioica* and to look for structure in the dataset, statistical modeling was also incorporated into the analysis. This analysis is discussed in Chapter 3. Many of these techniques were also used when assessing the level of variation and the nature of population structuring in the endemic *S. diclinis*, but further analyses were required to investigate the distribution of variation amongst the remaining *S. diclinis* individuals, and the possibility of a recent bottleneck in the species. This is discussed further in Chapter 4.

These experiments and analyses were designed not only to provide a snapshot of how each individual species has evolved, but to provide an overall picture of the history of the section *Elisanthe* from its beginnings as a newly dioecious group of individuals emerging from glacial refugia, to the subsequent adaptation into separate niches and the foundations of the reproductive isolation that accompanies speciation. The implications of the findings from the previous chapters in relation to the section as a whole are discussed in Chapter 5.

2. GENERAL PROTOCOLS FOR ISOLATION OF GENES AND INITIAL ANALYSIS

2.1 Introduction

Various methods were modified from existing protocols (Sambrook & Russell, 2001) and were used for all experiments in the following chapters. These methods were for extracting DNA from fresh plant material, general Polymerase Chain Reaction protocols for amplification of genomic DNA, SNP genotyping system appraisal and optimisation, sequence editing and segregation analysis of possible autosomal genes to be used for further analysis.

The decision to identify and use several autosomal genes was to enable subsequent analyses to be as effective as possible. Single loci have been found to produce error-prone estimates of historical demography and the timing of speciation events. These errors are reduced when multiple loci are used (Takahata & Satta, 2002; Edwards & Beerli, 2000). Use of numerous markers is also an excellent way to take into account demographic effects as they will be expected to affect all loci equally. Results affecting only a single gene are unlikely to have been caused by demographic effects, but may be due to natural selection. Conversely, effects that are seen across all loci are likely to be due to demographic effects such as population size changes (Emerson et al, 2001).

Autosomal loci are particularly useful for population genetic studies on models such as *Silene*. Many studies focus on uni-parentally inherited markers such as chloroplast and mitochondrial markers. In plants, these markers are most useful for looking at deep levels of evolution over many millions of years due to their lower substitution rates compared to nuclear markers (Wolfe *et al*., 1987), but synonymous divergence between *S. latifolia* and the non-dioecious relative *S. vulgaris* suggests that the *Silene* section *Elisanthe* is likely to have evolved relatively recently, around 10-20 million years ago (Filatov & Charlesworth, 2002). For this reason, the faster evolving nuclear loci are likely to provide more information as they will have accumulated more polymorphic sites.

The use of multiple loci also enables analytical tools to be used to estimate population genetic parameters such as migration and gene flow. The power of these coalescent analyses is improved with an increase in the number of genetic markers, although the time and power needed to compute the statistics involved (which also increases with the number of markers) is a major limiting factor. DNA sequences from multiple genes are therefore a useful resource as they include many polymorphic sites, particularly single nucleotide polymorphisms (SNPs) and microsatellites. A major advantage of SNPs over microsatellites in a population genetics study is that microsatellites are often so variable even between closely-related species that they are not as useful as SNPs for comparing levels of interspecific variation (Hedrick, 1999).

It is also possible to incorporate a SNP genotyping system into experiments to reduce cost, time and manpower. Such systems often rely on allele-specific competitive PCR systems such as ARMS (Amplification Refractory Mutation System, Newton *et al.*, 1989), and subsequent detection of attached fluorescent dyes. Although this provides a quick method for scoring multiple individuals, it is also a possible source of ascertainment bias. This occurs when too small a subset of the overall sample has been used (such as a single population) to identify the polymorphic sites for genotyping. Some SNPs may have been absent in this subset, but present elsewhere in the sample set. This would result in a false drop in diversity outside of the initial subset which would not occur if all samples were sequenced (Brumfield *et al*., 2003).

The following protocols were incorporated into the initial steps of the investigations into the evolutionary history of the species of the *Silene* section *Elisanthe*. They describe how several single-copy autosomal genes were identified, amplified and sequenced, and how SNPs were subsequently scored.

2.2 Methods

2.2.1 DNA Extraction

S. latifolia and *S. dioica* leaves were collected for DNA extraction from glasshouse plants derived from seed collected in the field by D. Filatov. *S. diclinis* leaf material was collected directly from plants in the field and extracted. 100mg of frozen leaf tissue was homogenized and extracted using either the Invitrogen™ Chargeswitch gDNA Plant Kit or the DNAzol Plant DNA extraction kit (also Invitrogen), following manufacturer's instructions.

2.2.2 PCR Amplification and Sequencing

Primers were designed from already available male flower bud cDNA library sequences (Atanassov, Tan and Filatov, unpublished) by myself and Dr. D.A. Filatov (see Appendix 1). Thermocycling was performed using an Eppendorf Mastercycler. PCR protocols were as follows:

PCR Reaction Mix		Thermocycling Conditions	
1µl	gDNA	2 mins 94°C	
10µl	Biotaq Red (Bioline)	30 secs 94°C	x 35 cycles
5µl	water	30 secs 53°C	
2µl	forward primer (5µM)	1 min 72°C	
2µl	reverse primer (5µM)	7 mins 72°C	

PCR products were run on a 1% agarose gel containing 10µl of 10mg/ml Ethidium bromide, and visualized by UV transillumination before gel extracting using the QIAquick Gel Extraction Kit (Qiagen) according to manufacturer's protocols.

Sequencing reactions for purified amplified fragments were performed using the Big Dye Teminator v3.1 Cycle Sequencing Kit (Applied Biosystems) and Eppendorf Mastercycler, and electrophoresis of products using an ABI PRISM 3700 DNA Analyser by the University of Birmingham Functional Genomics and Proteomics Laboratories. Sequencing PCR protocols were as follows;

Sequencing Reaction Mix		Thermocycling Conditions	
2µl	Purified PCR product	10 secs 94°C	x 30 cycles
1µl	Forward or reverse primer (5µM)	10 secs 53°C	
2µl	Big Dye Terminator Ready	4 mins 60°C	
Reaction Mix (Ready Reaction		7 mins 60°C	
Premix:Sequencing Buffer 2:1)		Hold 4°C	

To purify, 2µl of 0.05M EDTA was added to the amplified product to minimize unincorporated dye, and samples then precipitated by addition of 25µl 100% ethanol and frozen at -20°C for 30 mins. Samples were then centrifuged for 20 mins at 13,000rpm and ethanol poured off. Samples were then washed in 50µl 70% ethanol, centrifuged for 5 mins at 13,000 rpm, ethanol poured off and samples dried at 50°C before dissolving in 10µl HiDi Formamide (ABI), and submitting to the genomics laboratory for capillary electrophoresis at the ABI 3700 automated sequencer.

2.2.3 Sequence Editing and Segregation Analysis

Sequences were trimmed, edited for mis-called bases, re-coded for heterozygous sites using IUB (International Union of Biochemistry) codes and aligned by eye using ProSeq sequence editing program version 3 (Filatov,

2002). Exons and introns were assigned by either aligning to BLASTx hits (Altschul *et al*., 1990) against transcripts in the NIH GenBank sequence database or using open reading frames (ORFs) predicted using NCBI ORF finder. Sequences were then assigned functional regions in ProseqV3.

Parents and offspring were genotyped to test segregation of polymorphisms and establish autosomal or sex-linked inheritance. The crosses used for this were DF33 (*S. latifolia* ♀ IL9F x *S. latifolia* ♂ IL25H) and DF37 (*S. latifolia* ♀ Sa12 x *S. dioica* ♂ IL42). The parents and offspring were PCR amplified using the primers listed in Appendix 1 and sequenced before being checked at all polymorphic loci for sex-linked inheritance.

2.2.4 KASPar genotyping

The KBiosciences KASPar SNP Genotyping Kit (a dye-coupled competitive allele PCR based system) was tested as a method for genotyping offspring and species samples for known polymorphisms found during sequence analysis of genes amplified from *S. latifolia* and *S. dioica* individuals.

Primers were designed from the sequences of *S. latifolia* and *S. dioica* individuals using KBiosciences Primer Picker program (www.kbioscience.co.uk). There are two primers corresponding to each of the relative SNP alleles and a choice of two primers common to both sequences. Loci were chosen for testing as they had fixed allele differences between the *S. latifolia* and *S. dioica*

individuals. Primers were designed against these fixed allele differences so that clustering could easily be evaluated. The assay mix was produced by mixing the primers in the proportions shown in Table 2.1, and the final reaction mixtures for a 96-well and 384-well format are shown in Table 2.2.

Table 2.1. Assay mix for KASPar genotyping protocol.

Assay Mix	Volume (µl)
100µM Allele specific primer #1	12
100µM Allele specific primer #2	12
100µM Common primer	30
dH2O	46

Table 2.2. KASPar reaction mixtures for the PCR for 96-well and 384-well formats.

Reaction Mix	96 well typing	384 well typing
DNA	4µl	2µl
4X reaction mix	2µl	1µl
Assay mix	0.15µl	0.075µl
Taq	0.05µl	0.025µl
50mM $MgCl_2$	0.064µl	0.032µl
H2O	1.736µl	0.868µl
Total	8µl	4µl

Thermocycling conditions for KBiosciences K-Taq were as follows;

94°C 4 mins
94°C 10 secs
57°C 5 secs
72°C 10 secs
} x 20 cycles + 94°C 10 secs
57°C 20 secs
72°C 40 secs
} x 18 cycles

Plate reading was performed on a BMG LABTECH Fluostar Galaxy. Excitation and emission settings are shown in Table 2.3. Data was plotted FAM (x) against VIC (y). Data was normalised by dividing both sets of data by the reference (ROX) value of that particular well. Data was then called, if possible, according to the sample clusters.

Table 2.3. Excitation and emission wavelengths for FAM, VIC and ROX dyes used in the KASPar system.

Dye label	Excitation (nm)	Emission (nm)
FAM (allele 1 label)	495	520
VIC (allele 2 label)	538	554
ROX (reference label)	588	608

It was necessary to attempt to optimise this protocol to improve sample clustering. 24 samples of *S. latifolia* and 24 of *S. dioica* were used for optimisation. The following conditions and reagents were changed during the optimisation process;

Taq choice

Both proofreading and non-proofreading taq polymerases were tested with this protocol. They were the following;

- K-Taq (KBiosciences)
- Recombinant Taq (Helena Biosciences)
- Platinum Taq (Invitrogen)
- Taq (Promega)
- Hot-Taq (Eppendorf)

DNA

Different amounts of DNA were used in the reactions (0.5-2µl).

Annealing temperature

Gradient PCR was performed between 55°C and 68°C to establish whether annealing temperature would improve clustering.

Primers

Two loci were amplified (SlY4-ad1 and C2C4-ad2), and each set of primers was tested using both first and second choice common primers (C1 and C2) as produced by the KBiosciences Primer Picker program. This was to establish variability due to primer design (see Appendix 1).

$MgCl_2$ concentration

$MgCl_2$ was adjusted between 0.4 and 2.5mM to compensate for AT rich oligonucleotide designs as instructed in the manufacturer's protocol.

2.3 Results

2.3.1 DNA Extraction

Both the Invitrogen Chargeswitch gDNA Plant Kit and the DNAzol Plant DNA extraction kit successfully extracted DNA from *Silene* leaf and flower bud material consistently, and in high enough concentrations for PCR reactions. Both kits were subsequently used for extracting the samples.

2.3.2 Segregation Analysis

Five new autosomal genes were added to the pool of previously confirmed genes (see Table 2.4). These genes were clearly identified as autosomal by their familial inheritance patterns from the sequence data. The new genes are listed in Table 2.4 along with the corresponding NCBI BLASTx (Altschul *et al*., 1990) and ORF-finder hits which were subsequently used for assignment of functional domains. Figure 2.1 shows a typical autosomal pattern of inheritance.

Table 2.4. New autosomal genes confirmed by segregation analysis.

Locus Name	BLASTx	ORF-finder (frame)
C1A11	XP002269099.1	
C1E3	EAY79984.1	
C1E4	ABW91147.1	
C1H1	No significant matches	1-282 (-1)
C2C4	ABY74431.1	

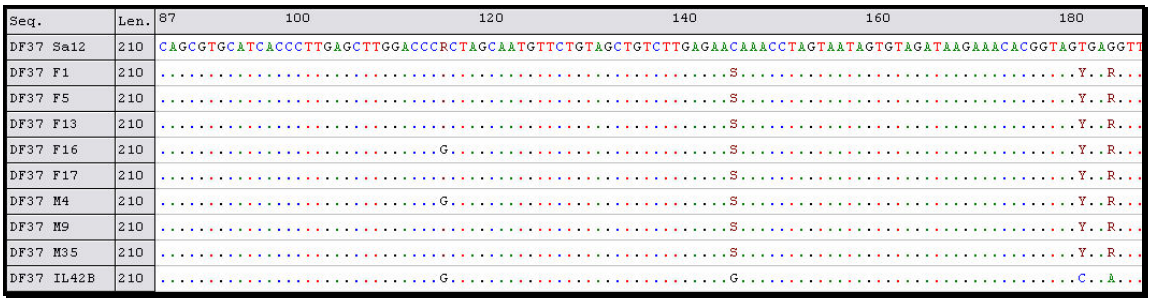

Figure 2.1. Sequence data for autosomal gene C1A11.
Female parent (Sa12), male parent (IL42B), female offspring (F) and male offspring (M). (ProseqV3, Filatov, 2002).

2.3.3 PCR Amplification and Sequencing

The loci in Table 2.5 were selected as suitable genes for polymorphism analysis following confirmation of autosomal inheritance by segregation analysis.

Table 2.5. Autosomal loci selected for further analysis.

Locus	Amplified by *	Length (bp)			BLAST Accession	Description
		S.lat.	*S.dio.*	*S.dic.*		
C37	GM	319	319	288	CAN62667.1	Hypothetical protein [*Vitis vinifera*]
C109	GM	248	248	337	XP002322538.1	Predicted protein [*Populus trichocarpa*]
C1D7	GM	208	208	N/A	No significant matches	-
C1F6	GM	210	210	N/A	No significant matches	-
C2D5	GM	575	575	N/A	XP002308937.1	ABC transporter family, cholesterol/phospholipid flippase [*Populus trichocarpa*]
C18	GM	290	290	N/A	BAE07183.1	Putative serine decarboxylase [*Beta vulgaris*]
C110	GM	193	193	217	EEF32792.1	tfiif-alpha, putative [*Ricinus communis*]

Locus	Amplified by *	Length (bp)			BLAST Accession	Description
		S.lat.	*S.dio.*	*S.dic.*		
C158	GM	432	432	N/A	EEF46877.1	DNA-directed RNA polymerase subunit, putative [*Ricinus communis*]
C34	GM	332	332	327	XP002306035.1	Predicted protein [*Populus trichocarpa*]
C79	GM	369	369	N/A	BAE07182.1	S-adenosyl-L-homocysteine hydrolase [*Beta vulgaris*]
C1A8	DF	829	829	N/A	EEF40868.1	Transporter, putative [*Ricinus communis*]
C1A11	ALH	191	191	328	XP002269099.1	Predicted: hypothetical protein [*Vitis vinifera*]
C1E3	ALH	258	258	N/A	EAY79984.1	Hypothetical protein OsI_35149 [*Oryza sativa* Indica Group]
C1E4	ALH	234	234	N/A	ABW91147.1	ACC oxidase 2 [*Ziziphus jujuba*]
C1H1	ALH	288	288	N/A	No significant matches	-
C2C4	ALH	303	303	N/A	ABY74431.1	Inositol methyl transferase [*Oryza coarctata*]
C1G11	DF	2036	2036	1098	EEF35149.1	Oligopeptidase A, putative [*Ricinus communis*]
Total		7603	7603	2595		

* GM = G. Muir, DF = D. Filatov, ALH = A. L. Harper

2.3.4 KASPar Genotyping Method

Initial runs using the KASPar system using manufacturer's instructions did not lead to reliable genotyping of the *S. latifolia* and *S. dioica* test samples. Optimisation experiments to improve clustering were as follows;

Taq

Different taq polymerarases gave slightly different results when used. Best separation of clusters was observed when K-Taq (KBiosciences) and Platinum Taq (Invitrogen) were used. Following this experiment K-Taq was used as standard.

DNA

The amount of DNA added to the 384 well reaction volume was adjusted from the suggested 2µl down to 0.5µl. 1µl was sufficient for successful reactions and achieved similar sets of results as when 2µl was added.

Primers

The four different combinations of primers do not appear to give dramatically different results. The C1 primer is marginally better at separating the clusters, but ad1 and ad2 show very similar results (Figure 2.2). None of the primer sets show good clustering, so from this point onwards, ad1c1 primer combination was used as standard.

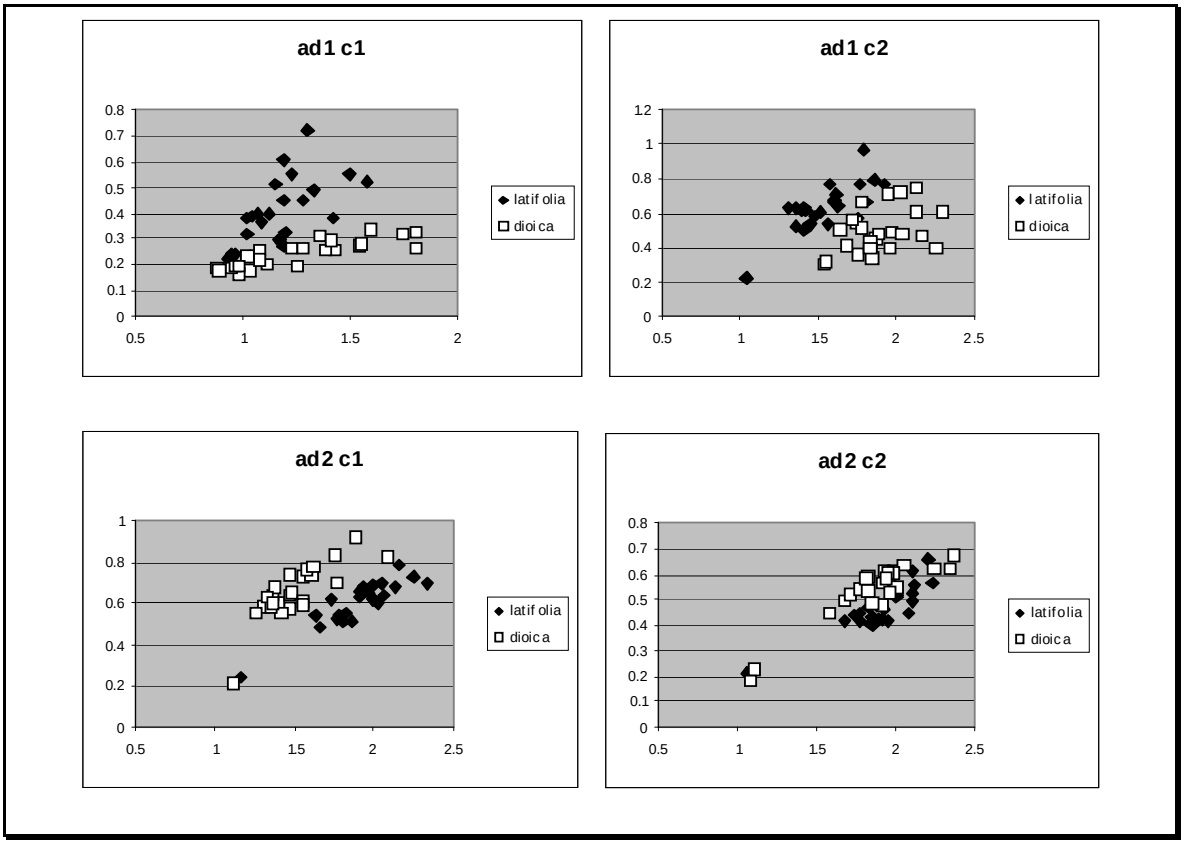

Figure 2.2. Plots of KASPar genotyping using different primer sets. Black diamonds represent *Silene latifolia* and white squares represent *Silene dioica*. Allele specific primer sets ad1 and ad2 were designed against SNPs that distinguish between the two species. First and second choice Common Primers (c1 and 2) were also tested.

Annealing temp

Annealing temperature seemed to have little effect on the efficiency of the genotyping protocol. Although fewer samples of each species were used for each temperature (due to the number of wells that could be at used at each temperature in the gradient cycler), It is clear that none of the annealing temperatures improved the efficiency of the clustering (Figure 2.3).

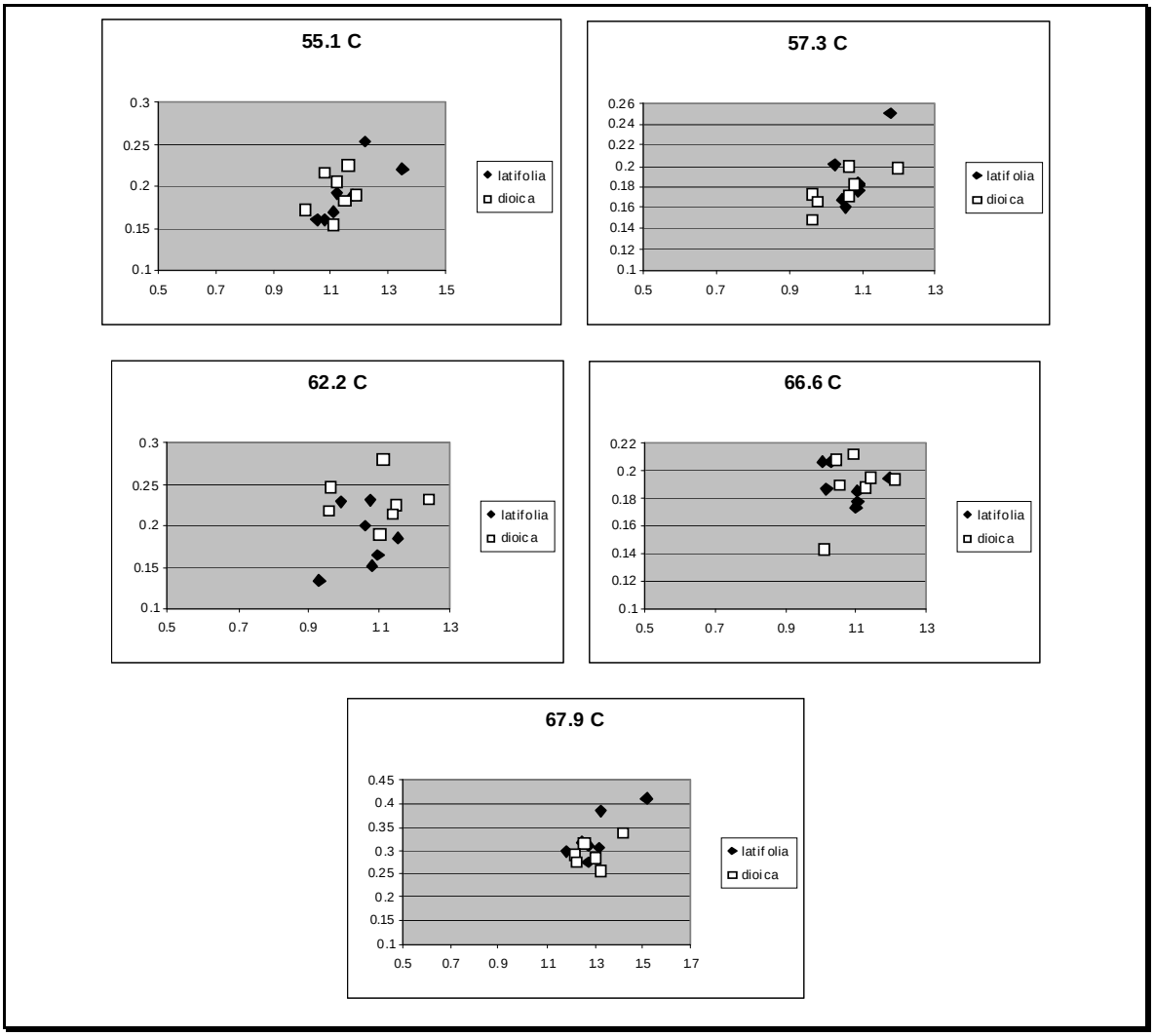

Figure 2.3. Plots of KASPar genotyping using Gradient PCR to test annealing temperature.
Black diamonds represent *Silene latifolia* and white squares represent *Silene dioica*. Samples were amplified using the ad1c1 primer set in an Eppendorf Mastercycler Gradient.

<u>$MgCl_2$</u>

Better separation of clusters was achieved when the $MgCl_2$ concentration was increased from the suggested 0.4mM to 1.2mM. Interestingly, poorer results were seen when 2.5mM $MgCl_2$ was added, despite this being the concentration suggested by the kit manufacturer in cases of high AT primer ratios (Figure 2.4).

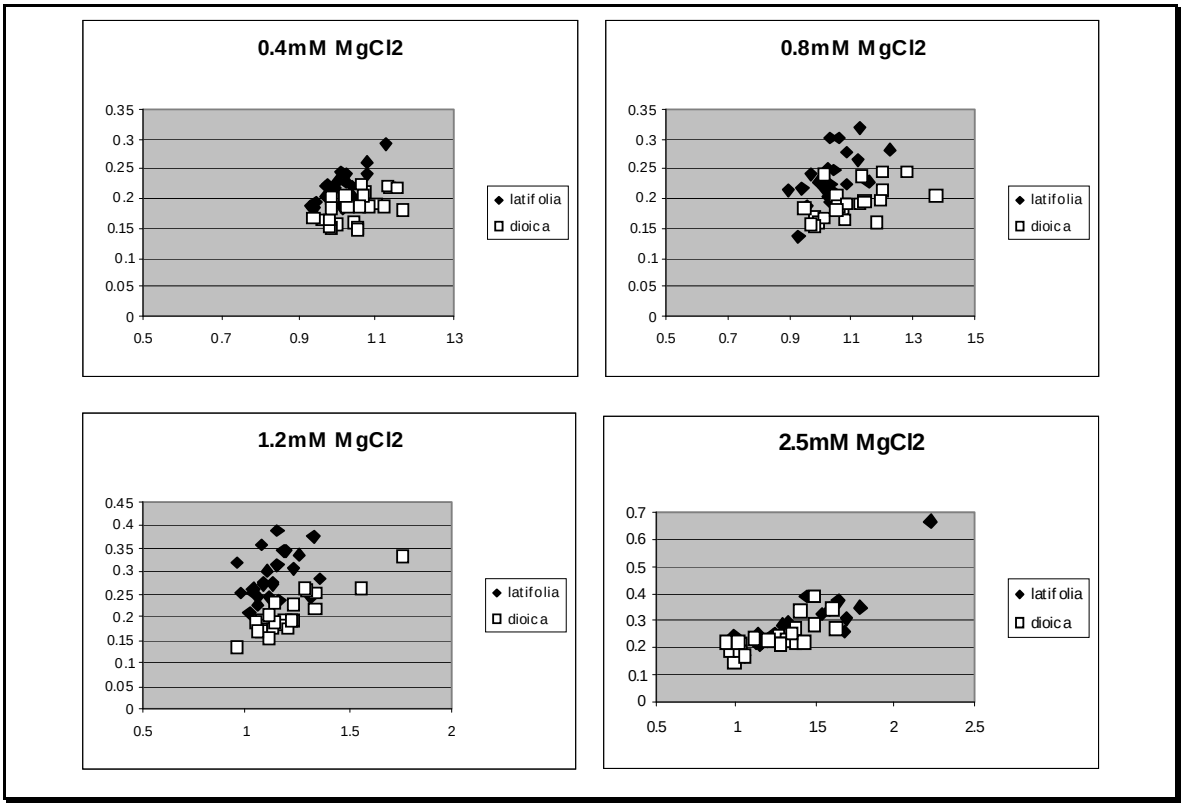

Figure 2.4. Plots of KASPar genotyping using different concentrations of $MgCl_2$.
Black diamonds represent *Silene latifolia* and white squares represent *Silene dioica*.

None of the optimisation steps taken tightened clustering or separated the clusters sufficiently for accurate allele-calling, and repeatability of experiments was poor despite automated robotic pipetting. Sequencing of all loci was therefore used to genotype the individuals used in all subsequent experiments.

2.4 Discussion

5 new autosomal genes were successfully identified using segregation analysis and PCR amplified from *Silene* leaf and flower bud material following design of primers from a *Silene latifolia* flower bud cDNA library, and these were added to previously sequenced autosomal genes provided by G. Muir and D. Filatov. All 18 loci could be amplified from *S. latifolia* and *S. dioica*, but only six of the 18 provided full length sequence in *S. diclinis* that could be aligned with the other species. This was a disappointing result considering that the divergence between *S. latifolia* and *S. diclinis* was expected to be low. The high amplification efficiency in those genes that could be amplified suggests that this was not a problem with the DNA extraction procedure, but most likely to be due to oligonucleotide primer mis-matches in *S. diclinis*. *S. diclinis* is a highly restricted endemic with very few numbers remaining (Montesinos & Güemes, 2006). In this situation, it is possible that insertions and deletions (indels) may have risen to high frequency in the species due to drift, reducing the inter-specificity of the primers that were designed in *S. latifolia*. One remedy would be to create a separate cDNA library for *S. diclinis*, but this would not necessarily provide loci that could be directly compared with *S. latifolia* and *S. dioica*. As a result of this problem, the length of total sequence in *S. diclinis* was around 30% that of *S. latifolia* and *S. dioica*.

A further disappointment was the failure of the KASPar genotyping kit to provide accurate clustering of SNP alleles. During optimization of the protocol, all protocol variables were manipulated, but in all cases no single combination produced clear reproducible clustering. It is possible that the problem was associated with the plate reader that was used, as it is different to the type used by KBiosciences for in-house genotyping. For this reason, sequencing of the individuals in this study was used instead.

Although sequencing is more expensive, numerous polymorphisms can be detected in a single sequencing run, negating the need for multiple primers per locus. Also, ascertainment bias that is commonly introduced via a genotyping study is eliminated by full sequencing of all individuals. It would have been very difficult to minimise ascertainment bias in this study if a genotyping kit were used, as there are individuals from several species, and even within species there are individuals from different races and populations. All populations and species would need to be included in the subset for the initial SNP screen to reduce this bias. Consequently, a large proportion of the individuals would need to be sequenced anyway.

Sequencing also should decrease the unknown errors in the dataset from incorrect allele clustering that may occur when genotyping SNPs using a system such as KASPar (Estimated to be around 0.3%; www.kbioscience.co.uk). The advantage of sequencing is that all of the sequence chromatograms can be

checked by eye and corrected if necessary. Overall, what was lost in speed and efficiency with the genotyping system failure was gained in the cost and accuracy of sequencing.

3. THE SEARCH FOR SEX-LINKED GENES IN *S. latifolia*

3.1 Introduction

Sex chromosomes are normally recognized as sex factors inherited via a large chromosomal segment that exhibits a low level of crossing over in the heterogametic sex (Bull, 1983). In the case of mammals, the heterogametic sex is the male, which normally carries a single X and a single Y chromosome. Females have two copies of the X chromosome. A similar system exists in birds and some reptiles, the ZZ/ZW system. In this case, the female is the heterogametic sex (ZW), and the male is homogametic ZZ (Bull, 1983). It is thought that these sex chromosomes may have evolved from a pair of autosomes (Ohno, 1967).

Other sex chromosomal systems also exist, notably the sex ratio system seen in *Drosophila*, whereby sex is determined by the ratio of X chromosomes to autosomes (Brown and Chandra, 1977), and XX/XO systems seen in some insects such as the Fire Bug (*Pyrrhocoris apterus*), which was one of the first sex chromosomal systems to be discovered (Henking, 1891).

Sex chromosomes have also been characterized in some plant species. They are less common in plants, with only a fraction of dioecious angiosperms possessing them (and only a fraction of angiosperms exhibiting dioecy), and they are often absent in many species within a genus. Some species in the genus

Silene (Campions) have evolved an XX/XY system that resembles that of mammals (Westergaard, 1958), and other models such as Sorrel (*Rumex acetosa*) have an XX/XY_1Y_2 system (Kihara and Ono, 1923). Bryophytes such as the liverwort *Morchantia polymorpha*, also possess sex chromosomes, although their lifecycle is more complex, with the dominant phase of the lifecycle producing haploid gametophytes. In this case, the sex of the individual is determined by which sex chromosome was inherited, the X or the Y. In most plant species, the male will possess the Y, but in some families such as Asteraceae and Rosaceae, it is the female (Bull, 1983).

This great variety of sex chromosome systems in many different organisms suggests that sex chromosomes have evolved separately multiple times in different families. Despite this, sex chromosomes often have notable character similarities.

In cases where extreme heteromorphy between the sex chromosomes is seen (such as in mammals), the X (or Z) chromosome is often comparable in size and gene content to the autosomes (Bull, 1983). The size and the shape of the Y or W chromosome vary, but the Y is usually smaller in size than the X in the case of animals. In plants, the Y (or in the case of multiple chromosome systems, the combined length of the two Y chromosomes) is usually larger than the X.

Gene content appears to be much lower on the Y or W chromosome in extremely heteromorphic systems. This is expected, as if the Y chromosome encoded for any genes essential to both sexes, females could not be XX. Also, YY individuals are normally inviable, or in the case of rare viable YY individuals (such as in *Mercurialis*), they are sterile (Westergaard, 1958). This suggests that most essential genes are located only on the X and not the Y. Linkage analysis in organisms such as *Drosophila* and humans has indicated vastly more genes linked to the X than the Y, and very few Y-linked gene functions. In the case of humans, they are mainly limited to such things as testis-determining factors, and genes associated with viable sperm production.

Heterochromatin is also an interesting feature associated with the sex chromosomes. Normally, chromatin is predominantly euchromatin, remaining diffuse unless in cell division. The Y chromosome, however, is mainly composed from constitutive heterochromatin, which renders it almost completely genetically inactive and late replicating. Few genes are known to be located inside constitutive heterochromatin, and this follows the argument for low gene content on the Y. In the case of the X chromosome, euchromatin will become functionally altered to become facultative heterochromatin at interphase in somatic cells. This can be seen to occur in mammalian females as Barr Bodies (Barr, 1959).

In placental mammals, this process of genetic inactivation is random in the embryonic tissues, so females are a mosaic for the active X chromosome in their

somatic cells, with active female Xs in certain places and active male in others. Interestingly our most distant mammalian relatives, the monotremes (marsupials), always inactivate the paternal X in their somatic tissues (Cattenach, 1975).

This process of X inactivation is a method of dosage compensation, which is also a feature of some organisms possessing sex chromosomes. It becomes necessary to dosage compensate when one of the sex chromosomes become degenerate, as so few genes are located on it. This means that in females, there would be roughly twice the amount of sex chromosome gene expression due to the two functional X chromosomes compared to the male's single X. In mammals, this is compensated by almost complete inactivation of one of the chromosomes in the female somatic tissues. Some genes escape inactivation however, and in the case of marsupials, different loci may show differing levels of inactivation, and some tissues may even show different levels associated with the same loci (VandeBerg *et al.*, 1983).

Other methods for dosage compensation exist, however. Compensation in *Drosophila melanogaster*, for instance, compensates for varying numbers of X chromosomes by reducing the levels of expression across all of the X chromosomes in the cell. This suggests that dosage compensation is not reliant on sex, but on the number of X chromosomes present (Brown and Chandra, 1977).

Possibly the most common feature uniting different chromosomal systems, is the reduced level of crossing over at meiosis in the heterogametic sex. In the highly evolved human sex chromosomes, the X and Y chromosomes are only homologous in small regions at the end of the chromosomes. These regions are known as the pseudoautosomal regions (PARs).

The 2.6Mb PAR at the tip of the short arm of the human Y chromosome (Yp-PAR) is necessary for adequate pairing with the X at male meiosis, and contains 13 genes. The 320kb region at the tip of the long arm (Yq-PAR) does not pair as efficiently with the X, and crossing over is less common. The X chromosome is able to cross over at female meiosis along its entire length with its partner X. In male meiosis, recombination has ceased along the majority of the length of the chromosome. This is thought to have occurred to protect the fitness of the sexes due to combinations of genes on the chromosomes associated with the sex determining loci (Bull, 1983). These linked genes may have been advantageous to one sex, and disadvantageous to the other. For one sex to keep the advantageous gene without reducing the fitness of the other sex, recombination was prevented in this region, fixing the genes in the sex that confers the advantage (Rice, 1987). Interestingly some invertebrates such as some *Diptera*, lack the ability to cross over entirely (Eloff, 1932). In humans, however, recombination still occurs, but all chiasmata normally distributed over the length of a chromosome are forced to be squeezed into the pseudoautosomal regions

at the tips of the chromosome arms at male meiosis. Consequently, the amount of crossing over in these regions is very high (Lien *et al*, 2000).

In other systems, such as that of *Drosophila miranda*, the cessation of recombination is due to a Y-autosome fusion, immediately stopping recombination due to the lack of a partner in male meiosis (Steinemann and Steinemann, 1998).

This reduced rate of recombination along much of the length of the Y chromosome, is also a key to understanding some of the genetic characteristics of the Y. Recombination is the chromosomal way to rid itself of the deleterious mutations that naturally build up over time, and to provide every opportunity to make new combinations of genes that may confer an advantage, and fix them into a population. The X chromosome is able to do this as it can recombine with the other X during female meiosis. The Y however, has stopped itself from recombining to protect its male associated combinations from the ravages of a crossover event within them. In so doing, it has allowed itself to become sensitive to the effects of Muller's Ratchet and drift, gradually reducing diversity in the Y population, but at the same time diverging from the X chromosome by genetic hitchhiking effects and selective sweeps (reviewed in Charlesworth and Charlesworth, 2000).

The Y chromosome may shelter recessive deleterious mutations due to its constant state of heterozygosity. Over time, this can lead to the genes on the Y becoming less well adapted. It is no longer possible once recombination has ceased, for the Y to regulate the build up of mutations in these genes. At some point these mutations will reach a level where the gene will no longer be functional on the Y, and will become inactivated. The presence of a functional copy on the X means that organisms may become viable only when they have an X chromosome in their karyotype. It is possible to see these old homologues of X and autosomal linked genes on the Y, now functionless pseudogenes. Once inactivated, selective forces have less of an effect, and mutations can completely scramble a gene beyond recognition. In addition to these mutations, it seems that the Y chromosome can become a dumping ground, accumulating tandem arrays of satellite DNA, and transposable elements (Steinemann and Steinemann, 1992; Skaletsky *et al*., 2003; Bachtrog, 2003).

It appears that there may be a method for preserving important genes on the Y chromosome in-tact, however. Y chromosome repeats are often organized as palindromes, separated by a spacer sequence in the centre (Skaletsky *et al*., 2003; Rozen *et al*., 2003). This allows gene conversion between the arms via a hairpin fold, preserving their sequence similarity and actually allowing them to undergo a form of recombination.

The final result of these effects is something similar to what we see in highly evolved sex chromosomes like our own. Evolutionary strata may be detected as a scrambled gene order compared to the X, as large inversions occur at different times ceasing recombination in the inverted segment. This allows synonymous mutations to build up that can be used to gauge the time since the inversion events (Lahn and Page, 1999; Lawson-Handley *et al*., 2004). The population diversity of the Y is lowered as drift acts upon its non-recombining region, and tandem arrays and transposable elements become fixed onto the Y. The Y chromosome becomes a degenerate both in size and functional gene content, and will eventually only carry functional genes involved in sex-determination, sex-linked traits and sex-specific fitness effects, and a few remnant genes that have been lucky enough to survive inactivation. Organisms carrying a Y and no X will normally be inviable, or at the very least sterile.

Studying a chromosome in such an evolved state becomes difficult, as only the genes that have escaped inactivation are useful. Apart from a few pseudogenes, the vast majority of genes that are thought to have once existed on the Y chromosome, have long since been scrambled beyond recognition, and are lost to us. It has therefore became necessary to find more useful models for studying sex chromosomes, and *Silene* (campions) has become one of the widely used models for a variety of reasons.

Silene is a diverse plant genus, with species that have monoecious, dioecious and gynodioecious sex-determination. Of the relatively few dioecious species, some such as *Silene latifolia*, have been found to also carry morphologically distinct sex chromosomes, but in the case of *S. latifolia*, the Y is actually larger than the X, probably due in part to the accumulation of repetitive DNA sequences. They have an XX/XY sex determination mechanism like our own, and divergence between the X and Y chromosomes has been detected in the few genes detected so far (Delichere *et al*., 1999; Atanassov *et al*., 2001; Moore *et al*., 2003; Filatov, 2005a; Bergero et al., 2007). It appears that recombination has ceased along most of the length of the *S. latifolia* X and Y chromosomes, but unlike our own sex chromosomes, the Y appears to be mainly largely euchromatic (Grant *et al*., 1994).

This all suggests that *Silene* sex chromosomes are at a much earlier stage of evolution than our own, and in fact they are thought to be at a tenth of the age of human sex chromosomes. They are consequently a good model to work with, as it is hoped that many more genes will be detectable on the *Silene* Y than the human Y, providing a wealth of molecular data to be studied.

Classical cytogenetic analysis of deletion mutants has suggested that the Y chromosome in *Silene* has at least three functions (Westergaard, 1958). The first of these functions is suppression of pistil development. The second is the initiation of the stamen development (deletion mutants in this region have no

stamen), followed by the third region which appears to be involved in the completion of stamen development (characterised by incomplete stamen development in mutants). Further studies have more or less confirmed these conclusions (Negrutiu *et al*., 2001; Lebel-Hardenack *et al*., 2002).

The genes currently known to be located on the Y chromosome in *Silene* have been mapped and synonymous and non-synonymous mutation data collected. This has provided insight into the location and timing of the events causing the cessation of recombination, allowing prototype evolutionary strata models to be proposed for *Silene* (Filatov, 2005b). The most useful genes have also been used to establish the levels of divergence between the X and the Y, and the possible mechanisms for the effects seen. The data also allows for the genes to be placed into phylogenetic trees to study the relationships between closely related dioecious species and their hermaphroditic cousins (Filatov and Charlesworth, 2002).

SlY1 was isolated by screening a *Silene latifolia* early male flower cDNA library using Y-derived probes made from pooled DOP-PCR products amplified from Y chromosomes micro dissected from metaphase root spreads (Delichere *et al*, 1999). Positive clones were then tested for Y-linkage by hybridization with restricted genomic DNA from male and female individuals, and their progeny. A similar cDNA corresponding to a highly homologous gene on the X chromosome was also detected (*SlX1*) and open reading frames encoding a 472 amino acid

polypeptide were identical at all but two amino-acid positions. These polypeptides were found to be part of a family of WD-repeat proteins, and shared a common origin with a similar protein seen in *Arabidopsis thaliana*. Immunolocalisation analysis also revealed that this protein is expressed in the nuclei of actively dividing cells, or in cells beginning to differentiate.

SlY4 and *SlX4* were identified as sex-linked genes in the same way (Atanassov *et al*., 2001). In this case it was discovered that *SlX4* had two allelic forms, and these genes were thought to encode Fructose-2,6-Biphosphatases. The genomic organisation of both *SlX1*/*SlY1* and *SlX4*/*SlY4* were compared, and it was found that the four genes were very similar to their respective pairs, containing the same number of exons. Whilst the introns of *SlX1* and *SLY1* were very similar in size and sequence, the introns of *SlX4* and *SlY4*, were quite different. The second intron differed in size by over 1000bp, and they shared very little sequence identity.

The silent and non-silent substitution rates were calculated for both of these genes, and were used for comparing both between the X and Y copies, but also between *S. latifolia*, and a close hermaphroditic relation *S.conica*. The results showed that the divergence between *SlX4* and *SlY4* was much greater than between *SlX1* and *SlY1* (Atanassov *et al*., 2001).

Moore *et al* (2003) isolated another sex-linked gene from *Silene latifolia*, *DD44X* and *Y*, homologous to the oligomycin sensitivity-conferring protein. The gene was isolated by conducting a differential display of mRNA transcripts from pre-meiotic male and female flower buds in an attempt to identify transcripts involved in the early stages of sex differentiation. These transcripts were then used as probes for a Southern blot of genomic DNA pooled from male and females, and restricted using one of three enzymes. Those transcripts which gave a male-specific restriction fragment length polymorphism (RFLP) were then used for a more detailed segregation analysis using male and female F1 plants from a mating between *S. latifolia* plants from two different populations. This gene is ubiquitously expressed in both sexes, and the genetic map (using deletion mutants characterised by Lebel-Hardenack *et al*., 2002) suggests linkage to the carpel-suppression locus. Physical maps produced from Fluorescent In-Situ Hybridization (FISH) analysis place this gene on the distal end of the long arm on the X, and the opposite arm on the Y. This suggests major chromosomal rearrangements have taken place on the *S. latifolia* Y chromosome.

A gene with strong homology to spermidine synthase in other species was isolated by Filatov (2005a). *SlssY* and *X* were discovered using segregation analysis of random cDNA clones. PCR primers were designed from the cDNA sequences, and used to amplify genomic DNA from male and female parents. Size polymorphisms between the parents were detected on agarose gels, and SNPs (Single Nucleotide Polymorphisms) were detected by sequencing of the

PCR products. Segregation of these polymorphisms could then be tested in the offspring from these plants. A low Ka/Ks ratio indicated that both the X and Y copies are functional at the moment, although analysis of the *SlssY* sequence shows that three amino-acid residues normally conserved are mutated. This may have reduced the efficiency of this enzyme. Silent divergence is much higher in the Y linked copy of this gene, suggesting that the Y is mutating more quickly than the X. Synonymous divergence of *SlssX/Y* genes was determined to be 4.7%, which is between the estimates for *SlX1/Y1* (3%) and *DD44X/Y*, with *SlX4/Y4* having the highest rate (16%). When these values are calculated from intronic sequences only (which have a higher rate of divergence), the values suggest evolutionary strata may be present (Filatov 2005b).

The evolutionary history of these four genes was revealed by Filatov (2005b). The four homologous genes were seen to be linked in *S. vulgaris* as well as in *S. latifolia*. A genetic map was produced for the X-linked genes by testing co-segregation of alleles. This placed the genes in an order which correlated with the calculated Ks values, suggesting at least three evolutionary strata (see figure 3.1).

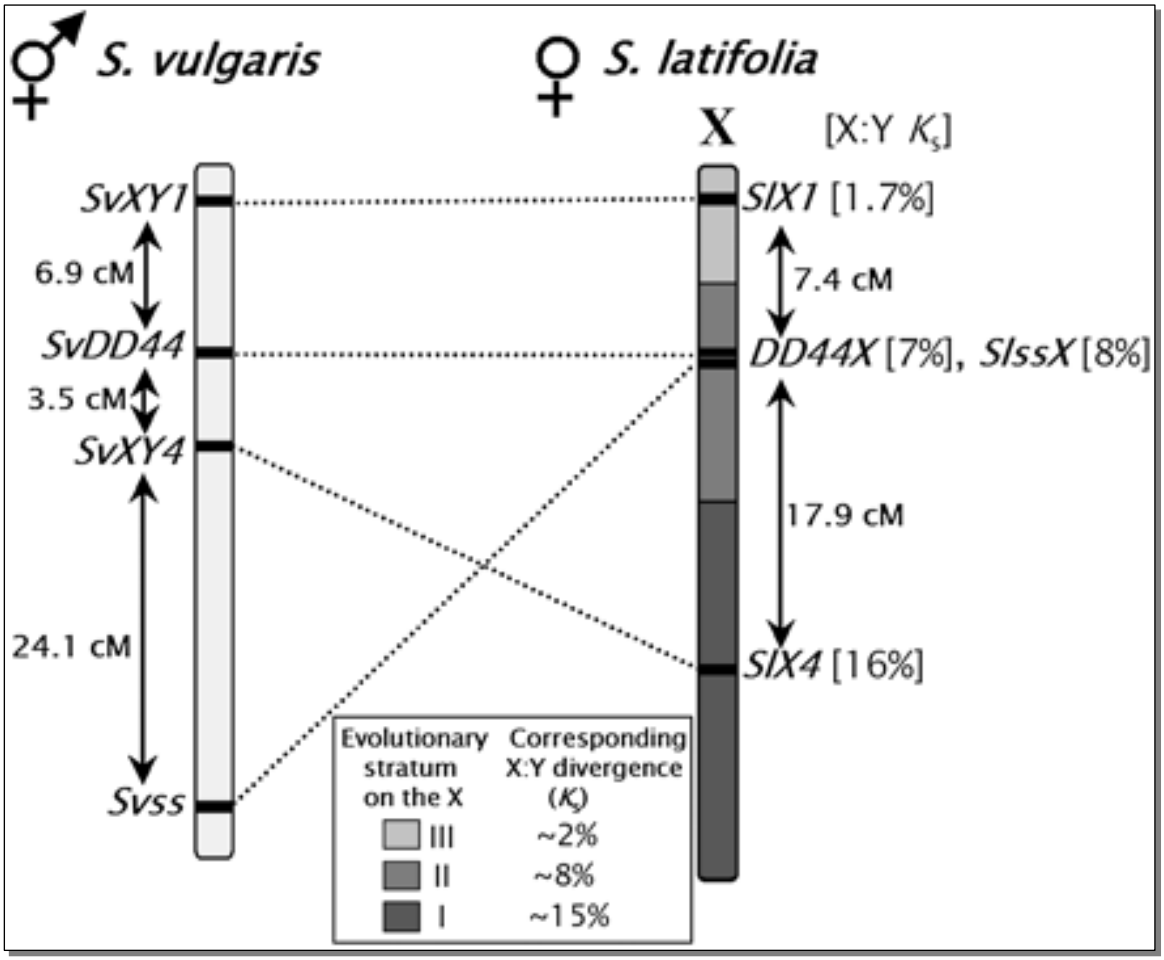

Figure 3.1. Genetic map for four X-linked genes in dioecious *S. latifolia* and their homologous genes in *S. vulgaris*. (Filatov, 2005b)

Other genes that have been detected on the Y chromosome of *Silene* include *MROS3X/Y* and *SlAP3Y/A*. MROS3X/Y (Guttmann and Charlesworth, 1998) was described as a X-linked gene with a degenerate Y-linked homologue, but as a member of a large multicopy gene family, with copies on autosomes as well as the sex chromosomes, this gene is more difficult to use for sex chromosome evolution analysis (Kejnovsky *et al*., 2001). SlAP3Y/A (Matsunaga *et al*., 2003) is a complete MADS box gene with significant similarity to the *Arabidopsis* AP*ET AL*A3. There is a copy on the Y chromosome which has been duplicated from an

autosomal copy, and that has survived degeneration, possibly due to fitness effects on the male.

Most recently, three new genes on the Y were discovered by Bergero *et al*. (2007) using analysis of Intron Size Variants and SNPs. One of these genes SlY6 is duplicated and shows best hits to the A. thaliana peptidyl-prolyl cis-trans isomerase locus *Cyp2.* The third gene identified SlX/Y7 corresponds to an *A. thaliana* unknown protein (locus AT5G48020). These new genes were mapped to the X loci, and suggested that divergence values had been saturated, confirming the age of the sex chromosomes at 10-20 million years.

From these few genes, it has been possible to infer these brief hypotheses about the history of the *Silene* sex chromosomes. We suspect that the Y is at an early stage of degeneration and that most genes have remained functional so far. It is also possible that evolutionary strata have been formed by successive rounds of recombination cessation along the X and Y, possibly caused by inversion events along the Y. Some of the most recent research regarding *Silene*, however, has thrown up some interesting findings.

Previous research has focused on the diversity of X and Y-linked homologues (Filatov *et al*., 2000, 2001; Laporte *et al*., 2005) and has suggested that diversity is much reduced on the Y chromosome, which appears to fit in with the

suggestions that either genetic hitchhiking or background selection are responsible for the degeneration of the Y.

Genetic hitchhiking (Rice, 1987) is proposed to allow fixation of certain haplotypes when they are linked to a strongly advantageous allele. This allele rapidly increases in frequency to fixation, as are any linked mutations, during a selective sweep. Conversely, the background selection model (Charlesworth *et al*., 1993, 1995) suggests that only those Y chromosomes with the fewest deleterious mutations will survive in the population. As the accumulation of mutation is stochastic, different chromosomes may prevail in different sub-populations.

To distinguish between these processes it is therefore necessary to look at polymorphism on a geographical scale. We would expect to see selective sweeps occurring across the entire species (Slatkin and Wiehe, 1998). Background selection however, can act on a finer scale producing increased population structure between sub-populations of a species.

This situation is complicated, however, by the action of interspecific gene flow. *S. latifolia* and *S. dioica* have been found to interbreed at hybrid zones under natural conditions (Baker, 1948). This migration of genes may act to further reduce Y diversity within species, whilst increasing the divergence between species. This is because Y-linked genes have been found to introgress less

readily than autosomal and other genes in rodent hybrid zones (Vanlerberghe *et al*., 1986, Jarola *et al*., 1997). If this effect is also true in *Silene*, then the lack of Y diversity may be due in some part to this effect in samples originating from hybrid zones.

Ironside and Filatov (2005) investigated some of these problems by looking at the population structure on a wider geographical scale than previously, and also looking at the relative introgression of Y and X linked gene DD44X/Y. Population sub-division was found to be higher in the Y-linked copy of the gene rather than the X-linked copy. Monte Carlo Markov Chain simulations also suggested that the Y-linked copy had introgressed less than the X-linked copy in *Silene* hybrid zones. Their results point towards background selection as the likely cause of the low Y diversity seen in *Silene* rather than selective sweeps, as strong population structure remains on the Y.

Unfortunately, demographic factors such as previous population bottlenecks further complicate these findings. Bottlenecks have the ability to mimic the effects of selection by reducing diversity in the population. These factors should, however, affect the whole population genome-wide. To eliminate demography as a factor, it is therefore necessary to check for these drops in diversity throughout the genome. If diversity is significantly lower on the Y when compared to other genomic markers, we can therefore discount demography as the cause.

Generation of new sex-linked genes in *Silene* will provide further information about the nature and evolution of the sex-chromosomes in *Silene*. Due to the lack of a *Silene* genome project, it was necessary to attempt to identify novel sex-linked genes using segregation analysis of loci amplified from families of *Silene latifolia* and *S. dioica* with primers designed from a young male flower bud cDNA library. During this process, it was expected that many of these loci would turn out to have an autosomal pattern of segregation. These autosomal genes could be used to study population and evolutionary genetics in members of the *Silene* section *Elisanthe*.

3.2 Methods

3.2.1 PCR and sequencing

PCR primers were designed from young leaf bud cDNA library clones. Clones were not included if they were predicted to be short (<200bp), pseudogenes or retrotransposons following alignment to database accessions using BLASTx. Loci were PCR amplified in four individuals that had been used as parents in crosses (see below). Initially, agarose and polyacrylamide gels were run to identify length polymorphisms between the parents, followed by sequencing to identify single nucleotide polymorphisms if no length polymorphisms could be detected. Those candidate loci that amplified and sequenced well were then amplified and sequenced in the offspring from the genetic crosses.

3.2.2 Segregation Analysis

The parents and offspring of the crosses used for segregation analysis are listed below;

Parental samples	**Crosses**
Sa12 (female *S. latifolia*)	DF37 (Sa12xIL42B), 8 females, 3
IL9F (female *S .l*atifolia)	males
IL25H (male *S. latifolia*)	DF33 (IL9FxIL25H), 3 females, 4
IL42B (male *S. dioica*)	males

Sequence chromatograms of the parent and offspring were compared to identify the segregation pattern of the polymorphisms for each candidate locus. Y-linked polymorphisms would be expected to be transmitted only through the male line, making them easy to identify when passed from father to male offspring. X-linked genes would be harder to identify as their polymorphisms would be transmitted in both the male and female offspring. They could still be identified if a polymorphism was passed from the father to all females (which would be heterozygotes) and no male offspring.

3.3 Results

No sex-linked genes were identified, but several candidates for the species comparison analysis have been identified. These were clean, single copy sequences with a reasonable amount of polymorphisms detectable in the parental sequences. In all cases, segregation analysis was used to confirm an autosomal inheritance pattern. The sequences were then amplified from the 12 unrelated *S. latifolia* and 10 *S. dioica* individuals, plus the *S. diclinis* and *S.marizii* plants. The details of these gene sequences are listed in Chapter 4.

3.4 Discussion

Unfortunately, no new sex-linked genes have been discovered to date. This is not altogether unexpected, however, as the relatively small number already characterized in *Silene* is a testament to the difficult nature of isolating these genes. Segregation analysis seems to be one of the more reliable methods for isolating them, but this is reliant on useful diagnostic SNPs inside the amplified region of the gene.

The segregation analysis could have been made much faster by use of a genotyping method such as the KBiosciences KASPar SNP Genotyping Kit. This method would allow samples to be genotyped according to their SNP allele by use of a single PCR step. Only parental samples would need to be sequenced, and then allele specific and common primers designed around diagnostic SNPs. It would then be relatively simple to perform PCR on the offspring from parental crosses in a plate format, read the wavelengths of the associated dye for each sample using a plate reader, correct against the control dye (ROX) and plot the results to cluster the alleles. Unfortunately, SNP gentyping using the KASPar kit was unsuccessful due to unsatisfactory clustering of alleles (see Chapter 2 for details).

Sequencing of the samples yielded several good single copy autosomal genes. These were added to previously amplified genes for population genetic analysis (see Chapter 4).

Discovery of new sex-linked genes has proved to be an extremely difficult and slow process in *S. latifolia*. The method employed by Bergero *et al* (2007) seems to be the most efficient, but is dependent on finding conserved introns with length polymorphisms. Until the *Silene* genome can be sequenced, discovery of new sex-linked genes will continue to be a haphazard process, and the true nature of the sex-chromosomes will be difficult to determine.

4. GENE FLOW BETWEEN *S. latifolia* AND *S. dioica*

4.1 Introduction

Hybridization between *S. latifolia* and *S. dioica* is known to occur with reasonably high frequency at hybrid zones (Baker, 1950) and so there is a possibility that this may have had some effect on the evolution of one or both species. Introgression of new genes is caused by a combination of hybridization followed by backcrossing. If intermediate (F_1) hybrids such as those between *S. latifolia* and *S. dioica* are fertile, they are likely to become backcrossed with one of the parent populations. If these backcrossed individuals are able to subsequently mate with the same parent population, genes may be able to introgress into the new background. This is how the hybrid zone can act as a "bridge" for interspecific gene flow.

The level of introgression that occurs is dependent firstly on the frequency of intermediate hybrids that are produced at the hybrid zone, and secondly by the fitness of the backcrossed individuals. A low frequency of hybrids will subsequently lower the number of backcrossed individuals, thus lowering the chances of introgression. Similarly, if hybrids are on average less fit than the parents, this may reinforce any reproductive barriers that may have evolved

since the divergence of the two species. This may be particularly relevant for species which have only recently made secondary contact with each other. However, even if the fitness of hybrids is on average lower than that of the parents, individual genotypes may be fitter than both parents in some environments, and may be able to introgress.

Reproductive barriers are often strong enough to severely restrict gene flow between even very closely related populations or species. In the case of *Mimulus guttatus,* (Lowry *et al*., 2008) coastal and inland races have become almost completely reproductively isolated due to selection against immigrants and flowering time differences. These ecological reproductive barriers have the potential to lead to formation of new reproductively isolated species. Evidence of this can be found in other plant species such as the irises *Iris brevicaulis* and *I. fulva* which have developed substantial barriers to gene flow involving both flowering time phenologies and pollinator preference (Martin *et al*., 2007; Martin *et al*., 2008).

Despite reproductive barriers, introgression in some species may be quite common. In the case of sunflower species *Helianthus annuus* and *H. petiolaris* , significant levels of hybridization and introgression have occurred over long periods of time since their divergence approximately 1 million years ago (Yatabe et al., 2007; Strasburg & Rieseberg, 2008). Despite this high level of

introgression, the integrity of the two species has been maintained by reproductive barriers.

Hybridization and introgression may not always be considered to be evolutionarily significant, however. For introgression to have an impact on the adaptive evolution of the parent species, it must either provide extremely rare variants or advantageous combinations of alleles, as large parental populations will be capable of producing many more variants than hybridization through mutation. It is clear, however that in some cases hybridization events are important in speciation and adaptation.

Senecio squalidus is a hybrid species derived from a hybrid zone on Mount Etna in Sicily (Abbott & Lowe, 2004; James & Abbott, 2005). It was brought to Britain for cultivation in the Oxford Botanic Garden before its escape into the wild in the 18th century. Following its escape, it came into contact with the native British species *S. vulgaris*, producing two further hybrids, and allowing a trait (ray florets) of the outcrossing *S. squalidus* to introgress into the self-pollinating *S. vulgaris*. Ray florets are a trait normally associated with outcrossers as they help to attract pollinators, so incorporation of this trait into a normally self-pollinating species does not immediately appear advantageous. However, ray florets are also associated with late germination, reducing mortality from frost (Kim et al., 2008; Abbott *et al*., 1998). In this case, hybridization is able to re-introduce

complex traits, a process which would have been unlikely to have occurred via multiple gene mutations.

As discussed in Chapter One, there appear to be varying levels of gene flow in the *Silene* genome, although the historical extent, endurance and significance of any introgression is not yet clear. Ironside and Filatov (2005) investigated population structure and the relative introgression of Y and X linked gene DD44X/Y in *S. latifolia* and *S. dioica*. Population sub-division was found to be higher in the Y-linked copy of the gene rather than the X-linked copy and Monte Carlo simulations supported gene flow after speciation on the X but not the Y.

Minder et al (2007) estimated that introgression between *S. latifolia* and *S. dioica* was occurring at a high rate at natural hybrid zones in the Swiss Alps, although few true intermediate hybrids were identified. They suggested that hybrid zones act as bridges for gene flow between the two species.

Muir and Filatov (2007) also concluded that gene flow has occurred between the two species in respect to the chloroplast genome. *S. latifolia* and *S. dioica* have been shown to have low levels of diversity and population structure on the chloroplast, suggestive of a selective sweep having crossed the sequence boundary. As a result, the chloroplast sequence is homogenized across both species, and there is a reduction in effective population size. Partitioning of the variation across different hierachical levels showed that most of the variation

was within populations. In line with predictions of population differentiation dynamics (Wright, 1931; Slatkin & Voelm, 1991; Vigouroux & Couvet, 2000), a reduction in effective population size of a subdivided population may allow faster population differentiation of a small deme (as F_{ST} is inversely related to effective population size), in relation to a large deme with a larger combined effective population size. This may explain the greater partitioning of variation within populations, as drift has more of an effect in these smaller populations.

The above studies have suggested that there has been (or is) introgression of the X chromosome and chloroplast between these two species but not the Y chromosome. It is impossible, however to conclude definitively that the Y chromosome is unusual in this way despite limited Y chromosome introgression being well characterized in other model species such as rodents (Vanlerberghe *et al*., 1986, Jaarola *et al*., 1997). To establish this, it is necessary to estimate a "general" level of gene flow in the genome. To account for demographic factors, it is preferable to use neutral autosomal markers.

In this study, eighteen such marker loci were sequenced following segregation analysis of sequences isolated from parental and offspring genomic DNA from several *Silene* crosses. Descriptive statistics, such as diversity estimates, and neutrality tests such as Tajima's D (Tajima, 1989), Fay and Wu's H (Fay & Wu, 2000) and HKA tests (Hudson *et al*., 1987) were initially generated to establish whether any loci evolved under positive selection, which could bias further

analyses. An estimate for population differentiation (F_{ST}) between *S. latifolia* and *S. dioica* was also calculated. Gene flow would be expected to push the F_{ST} values down towards those levels seen for the X chromosome and chloroplast (around 0.1-0.2) (Ironside & Filatov, 2005; Muir & Filatov, 2007).

Following these initial tests, more complex analyses were conducted. The first of the programs used to analyse the data was Structure, a Bayesian method for calculating the most likely number of population clusters in a dataset, and assigning membership of individuals to each cluster (Pritchard *et al.,* 2000). This was utilized as an indicator of the level of admixture (and shared ancestral polymorphism) in the two species. Subsequently, a Monte Carlo Markov Chain method (IM, Hey and Nielsen, 2004) designed to estimate parameters such as levels of gene flow from one population to another as well as time since species divergence was attempted to gauge the level and directionality of migration between *S. latifolia* and *S. dioica*. A stricter isolation model was also conducted using the program WH (Wakeley and Hey, 1997).

Analysis of the level of linkage disequilibrium (LD) in the total *S. latifolia* and *S. dioica* dataset was also conducted as an aid to interpreting the patterns seen between these two species. Large amounts of linkage disequilibrium would be expected to occur if there has been recent gene flow between them, and conversely, little gene flow would allow linkage disequilibrium levels to be broken

down by recombination. One other factor to be considered is that selection in some loci would also be expected to raise levels of LD.

4.2 Methods

4.2.1 Samples

Samples were provided by D. Filatov and are listed below. The samples were collected in order to cover as much of the natural distribution of the two *Silene* species as possible, particularly around natural hybrid zones such as those in the UK and Belgium. The individuals sampled are listed in Table 4.1.

4.2.2 PCR and Sequencing

Autosomal genes selected for further analysis were amplified and sequenced using the primers listed in Appendix 1. Heterozygous sites were either resolved using ProSeq for analysis, or kept as an unresolved dataset with heterozygous bases coded using the IUPAC (International Union of Pure and Applied Chemistry) notation. DNA alignments are provided as Proseq3 files on compact disc inside the back cover.

4.2.3 Sequence Analysis using DNAsp

Sequences were entered into DNAsp version 4.10.7 program (Rozas & Rozas, 1999) to calculate basic statistics from the resolved autosomal sequences. The number of segregating sites (S) for each locus was calculated and estimates of nucleotide diversity (pi) (Nei, 1987) were generated for each locus. This is the average number of nucleotide differences per site for two sequences. DNAsp

was also used to calculate the minimum number of recombination events (R_m) in the history of the sample for each locus. This is extremely conservative however, as R_m may be falsely inflated due to random reconstruction of alleles.

S. latifolia and *S. dioica* populations were separated to infer natural selection in each species for each locus by calculating Tajima's D (Tajima, 1989). This is based on the difference between two estimates of variation, the number of segregating sites and average number of pairwise differences. In a neutral equilibrium population with a constant size (ie. the null hypothesis), this figure should be zero. Balancing selection or a population decline may lower levels of both high and low frequency polymorphisms causing D to be positive. Conversely purifying selection (or population expansion) will create an excess of low frequency polymorphisms and a tendency towards negative values of D. F_{ST} was also calculated for each species and locus using DNAsp, as a measure of genetic variance (population structure) between the two species.

Table 4.1. *Silene* individuals analysed.

Identifier	*Species*	Sex	Origin
IL7f2A	*S. latifolia*	Male	Romania
IL113J	*S. latifolia*	Male	Italy
IL28f2A	*S. latifolia*	Male	France
IL3f2C	*S. latifolia*	Male	Belgium
IL116G	*S. latifolia*	Male	Greece
IL4f2M	*S. latifolia*	Male	Romania
IL19f2A	*S. latifolia*	Male	France
IL11G	*S. latifolia*	Male	Spain
IL107D	*S. latifolia*	Male	Germany
IL25H	*S. latifolia*	Male	England
IL9F	*S. latifolia*	Female	Romania
Sa12	*S. latifolia*	Female	Belgium
IL107B	*S. latifolia*	Male	Germany
IL81H	*S. latifolia*	Male	Austria
IL92	*S. latifolia*	Male	Austria
IL137C	*S. latifolia*	Male	Russia
IL139D	*S. latifolia*	Male	Russia
IL5E	*S. latifolia*	Male	Romania
IL33f2A	*S. latifolia*	Male	England
Sa777	*S. latifolia*	Male	England
IL98/7	*S. dioica*	Male	Austria
Sd106	*S. dioica*	Male	France
IL91f2A	*S. dioica*	Male	Austria
IL69f2	*S. dioica*	Male	Wales
IL42B	*S. dioica*	Male	Belgium
IL124	*S. dioica*	Male	Wales
Sd113	*S. dioica*	Male	France
IL66f2A	*S. dioica*	Male	England
Sd449	*S. dioica*	Male	Sweden
IL63f2L	*S. dioica*	Male	Wales
IL40f2D	*S. dioica*	Male	Belgium
IL60G	*S. dioica*	Male	England
IL62F	*S. dioica*	Male	England
IL70E	*S. dioica*	Male	Sweden
Sd785	*S. dioica*	Male	England
Sd780	*S. dioica*	Male	England
Sdic371B	*S. diclinis*	Male	Spain
IL74A	*S.marizii*	Male	Portugal
Sv581	*S. vulgaris*	Male	France

4.2.4 Bayesian admixture analysis

A model-based clustering method implemented in the program Structure (Pritchard *et al.* 2000) was used to assign individuals probabilistically to homogenous clusters (*K* populations) without consideration of sampling localities. Estimated posterior probabilities for the simulated model fitting the data were calculated assuming a uniform prior for *K*, where $1 \leq K \leq 5$.

An input file was created where each individual plant's identifier was followed by its haplotype for each locus recoded into a numerical format. Individuals with over 50% missing data were removed from the analysis to reduce error. To minimize the effect of the starting configuration during the Monte Carlo simulation, a burn-in of 100,000 steps was conducted, before data for the parameter estimations was collected from a further 500,000 steps. Three independent runs of the Markov chain, each of least 500,000 steps were performed to assure convergence of the chain and homogeneity among runs for each prior of *K*. The posterior probabilities of *K* were then calculated using Bayes' rule. The program was run without population identifiers and in the admixture mode which assumes that each individual has drawn some fraction of the genome from each of the populations considered. Allele frequencies were allowed to be independent.

4.2.5 IM (Isolation with Migration Model) Program

IM is used to fit genetic data to an isolation with migration coalescent model (Hey and Nielsen, 2004; based on the method originally developed by (Nielsen & Wakeley, 2001)). Resolved sequences for each locus and species that show no evidence for recombination (as recombination within the gene may disrupt the patterns of variation skewing the results), were written into a file for input into the IM program. This program is a Monte Carlo Markov Chain (MCMC) method designed to estimate the relative effects of migration and isolation on the genetic variation seen within two populations (or species). MCMC methods explore the posterior distribution landscape using a random walk which accepts steps closer to a posterior probability of 1, and rejecting a proportion of those significantly lower than 1. The chain therefore converges on the highest posterior probability for the parameter given the data and the prior probability. An Input file was generated which included the sequences for the eight loci showing no recombination.

The command line strings and their effect on the model for different runs were as follows:

<u>General MCMC run settings</u>

b=number of steps for burn in.
l=number of steps in Markov Chain.
n=number of Markov chains to run under Metropolis-coupling
k=number of Markov Chain swap attempts during Metropolis-coupling
fl=set linear heating scheme for Metropolis-coupling
g1=heating parameter

Prior probability distribution settings

q1=theta population 1 prior distribution upper bound.
q2=theta population 2 prior distribution upper bound.
m1=prior distribution upper bound for migration rate from population 1 to 2.
m2=prior distribution upper bound for migration rate from population 2 to 1.
k=prior distribution upper bound for time since population/species split

The first three runs were set up as follows;

Run 1: IM -b 1000 -l 1000000 -q1 1.0 -q2 5.0 -m1 9 -m2 14 -t10

Run 2: IM -b 100000 -l 1000000 -q1 0.5 q2 5.0 -m1 15 -m2 10 -t 20

Run 3: IM -b 100000 -l 1000000 -q1 0.5 -q2 5.0 -m1 10 -m2 15 -t 15

These runs were set up to assess the success of the program, and the robustness of the data in the model. This was done by changing parameters, and checking the effects on the distribution curves. This would have set the random walk off at different start points on the probability landscape, and the data could only be trusted if the posterior density curves were similar after each run.

Run 4 was designed to find the end point of the distribution curve for the t parameter (time since splitting of populations). The t parameter upper bound was therefore increased to a value of 50.

Run 4: IM -b 200000 -L 1000000 -q1 5.0 -q2 5.0 -m1 10 -m2 10 -t 50

Run 5 incorporated Metropolis-coupling of the MCMC chain to help improve the mixing (mapping of the distribution landscapes);

Run 5: IM -b 100000 -l 1000000 -q1 10 -m1 20 -m2 20 -t 20 –k 2 –fl -n 5 –g1 0.05

Metropolis-coupling (hereafter referred to as MCMCMC) involves setting off several Markov chains. Data is only taken from one of these chains (the cold chain). The rest of the chains are “heated” to a power β between 0 and 1 (the power by which the posterior probability is raised). As β approaches zero, this lowers the posterior probability, allowing the chain to accept more steps, and explore the landscape more fully.

Each successive chain has a greater amount of heat applied to it determined by a heating parameter h (set by the g1 parameter in the program). The heat that is applied to the i^{th} chain is $\beta = 1 / (1 + i \times h)$. After each cycle, two chains are chosen at random, and attempt to swap states and parameter values. In this way, the cold chain (from which results are recorded) can swap positions with one of the heated chains on the landscape allowing more detailed exploration of the landscape, preventing it from getting stuck in a single region of high posterior probability. Effectively it allows the chain to jump across chasms of low posterior probability to other areas of the probability landscape, thereby improving the “mixing”.

4.2.6 WH Isolation Model

The WH isolation model (Wakeley and Hey, 1997) was used to attempt to reject a null hypothesis of no gene flow between the two species. The program fits a simple speciation model (the isolation model) to multilocus datasets. The model makes several assumptions:

- The two species of interest arose from a single ancestral species t generations ago.
- The common ancestral species had a constant effective population size NA.
- The two descendent species also have constant effective population sizes N1 and N2.
- There has been no gene flow since separation from the common ancestor at time t.
- All mutations are neutral. For this reason, any genes found to be under selection using the above Multilocus Maximum Likelihood HKA method were excluded.

The program provides an output file assessing the quality of fit of the data to the simulation model with a chi-square statistic and the *wh* statistic (Wang *et al.*, 1997). Also provided is a table of parameter values with 95% confidence intervals and means and a table of observed and simulated means of variants. 10,000 simulations were run.

4.2.7 LD analysis

DNAsp was used to calculate measures of linkage disequilibrium between all informative sites both within and between loci. LD estimates may be biased upwards due to random assignment of alleles to produce haploid datasets from diploid sequences, but this should affect the entire dataset enabling LD patterns and comparisons between genes to be detected. The measure r^2 (Hill & Robertson, 1968) was chosen to provide a convenient figure between 0 and 1 for graphing. R^2 values were arranged into a triangular matrix in Microsoft® Office Excel 2007 before being converted into a linkage disequilibrium heatmap using the R (http://CRAN.R-project.org/doc/FAQ/) package LDheatmap (Shin, 2006).

4.2.8 Multilocus Maximum Likelihood HKA

A Multilocus Maximum Likelihood HKA test was used to determine if any of the autosomal loci studied are operating under selection (Wright & Charlesworth, 2004). The test conducts a maximum likelihood analysis of multilocus polymorphism and divergence data allowing selection at one or more loci for comparison to the neutral model (with no selected loci).

The first simulation to be run was the standard neutral model with no selected loci. Data was entered separately for *S. latifolia* and *S. dioica* for each locus when enough data was obtained making a sample size of 27. The simulations allowed each species at each locus in turn to be assumed to be under selection.

The output file provided the following parameter estimations;

- Value of the maximum ln likelihood.
- Maximum likelihood estimate of the divergence time parameter.
- Maximum likelihood estimate of theta for locus x.
- Maximum likelihood estimate of the selection parameter k for all loci under selection.

To test for selection, loci that produced an improved likelihood were tested for significance using the likelihood ratio test;

$$LR = 2*(lnL1-lnL2)$$

The likelihood ratio is approximately chi-squared with degrees of freedom equal to the number of selected loci.

4.3 Results

4.3.1 Sequence Analysis using DNAsp

Sequences of five genes amplified and sequenced by myself were added to 11 genes amplified by G. Muir and D. Filatov for combined analyses (see Table 4.2). DNAsp was used to measure the nucleotide diversity (pi), and to test for neutrality using Tajima's D statistic and Fay and Wu's H. See Table 4.2 for details).

Table 4.2. Statistical test results on 18 autosomal genes.

Locus	Amplified by:*	Species†	Length	Pi all	Tajima's D	Fay & Wu's H
C37	GM	*Elis.*	400	0.0137	-1.14 (NS)	0.70 (NS)
		lat	400	0.0124	-1.24 (NS)	
		dio	400	0.0079	-1.42 (NS)	-8.97 ***
C109	GM	*Elis.*	316	0.0195	-0.52 (NS)	-0.72 (NS)
		lat	316	0.0172	-0.80 (NS)	
		dio	316	0.0142	-0.32 (NS)	-0.65 (NS)
C1D7	GM	*Elis.*	211	0.0013	-1.73 (NS)	N/A
		lat	211	0	N/A	
		dio	211	0	N/A	N/A
C1F6	GM	*Elis.*	400	0.0023	-1.26 (NS)	0.62 (NS)
		lat	400	0.0028	0.47 (NS)	
		dio	400	0	N/A	
C2D5	GM	*Elis.*	576	0.0048	-1.62 (NS)	-1.55 (NS)
		lat	576	0.0034	-1.57 (NS)	
		dio	576	0.0031	-0.57 (NS)	-0.58 (NS)
C18	GM	*Elis.*	323	0.0279	-0.61 (NS)	0.02 (NS)
		lat	323	0.0213	0.82 (NS)	
		dio	323	0.0292	0.36 (NS)	3.07 (NS)
C110	GM	*Elis.*	232	0.0173	-0.39 (NS)	No Outgp.
		lat	232	0.0149	-0.69 (NS)	
		dio	232	0.0095	-1.32 (NS)	No Outgp.

Locus	Amplified by:*	Species†	Length	Pi all	Tajima's D	Fay & Wu's H
C158		*Elis.*	451	0.0032	0.14 (NS)	
	GM	*lat*	451	0.0019	-0.63 (NS)	-0.72 (NS)
		dio	451	0.002	1.02 (NS)	-1.30 (NS)
C34		*Elis.*	375	0.0035	-1.35 (NS)	
	GM	*lat*	375	0.0032	-0.51 (NS)	-0.82 (NS)
		dio	375	0.0038	-0.29 (NS)	-0.97 (NS)
C79		*Elis.*	441	0.0105	-0.56 (NS)	
	GM	*lat*	441	0.0059	0.55 (NS)	0.97 (NS)
		dio	441	0.0088	-0.64 (NS)	-0.59 (NS)
C1A8		*Elis.*	1606	0.0143	0.92 (NS)	
	DF	*lat*	1606	0.0018	-1.38 (NS)	-2.80 (NS)
		dio	1606	0.0015	-1.74 (NS)	-0.82 (NS)
C3A4		*Elis.*	317	0.033	0.54 (NS)	
	GM	*lat*	317	0.0166	-1.23 (NS)	-7.05 *
		dio	317	0.0174	0.21 (NS)	0.49 (NS)
C1A11		*Elis.*	234	0.0136	-1.12 (NS)	
	ALH	*lat*	234	0.0081	-0.68 (NS)	0.27 (NS)
		dio	234	0.0122	-0.93 (NS)	0.65 (NS)
C1E3		*Elis.*	258	0.0014	-1.96 (NS)	
	ALH	*lat*	258	0.0016	0.65 (NS)	No Outgp.
		dio	258	0.0013	-1.45 (NS)	No Outgp.
C1E4		*Elis.*	234	0.0194	-0.40 (NS)	
	ALH	*lat*	234	0.0133	-0.51 (NS)	No Outgp.
		dio	234	0.0234	0.06 (NS)	No Outgp.
C1H1		*Elis.*	288	0.001	-1.89 (NS)	
	ALH	*lat*	288	0	N/A	No Outgp.
		dio	288	0	N/A	No Outgp.
C2C4		*Elis.*	346	0.008	-0.62 (NS)	
	ALH	*lat*	346	0.0078	0.77 (NS)	0.85 (NS)
		dio	346	0.0092	1.45 (NS)	-0.21 (NS)
C1G11		*Elis.*	2780	0.0066	-1.69 (NS)	
	DF	*lat*	2780	0.0046	-1.37 (NS)	-0.03 (NS)
		dio	2780	0.0041	-1.08 (NS)	-1.93 (NS)

* GM = G. Muir, DF = D. Filatov, ALH = A.L. Harper

† *Elis.* = section *Elisanthe*, *lat.* = *S. latifolia*, *dio.* = *S. dioica*

NS = Not significant

These analyses reveal a great deal of variation between these genes. The number of segregating sites is particularly variable. All Tajima's D are non-

significant, although the majority of estimates are negative which suggests that weak purifying selection may be acting at some of these loci. The F_{ST} values for genetic differentiation are shown in Figure 4.1. The average F_{ST} value across all loci is 0.269.

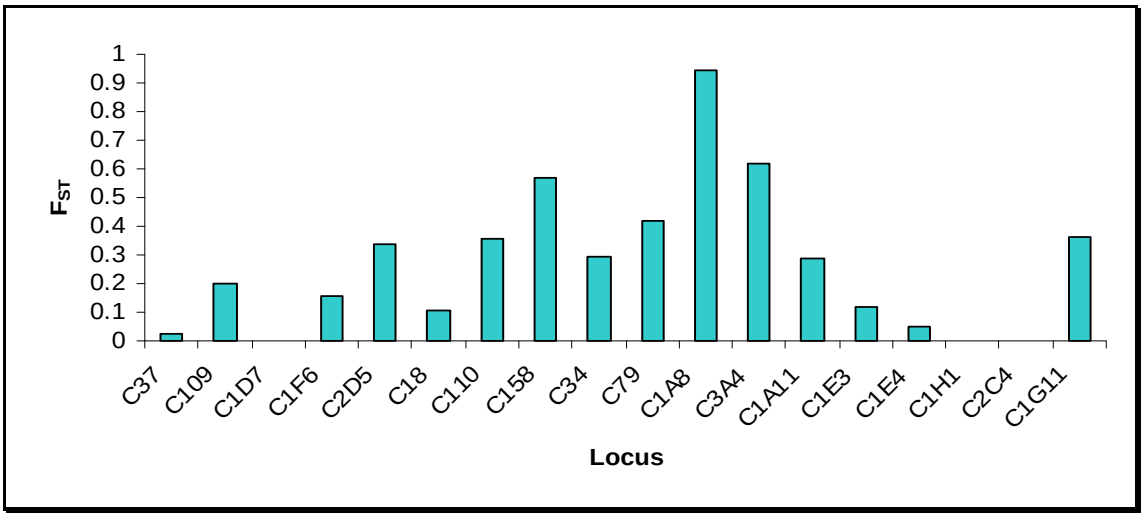

Figure 4.1. F_{ST} values for 18 autosomal loci.

4.3.2 Bayesian Admixture Analysis

Structure was used to assess the number of possible population clusters in the combined *S. latifolia* and *S. dioica* dataset (29 individuals). Results of three independent runs are shown in Figure 4.2.

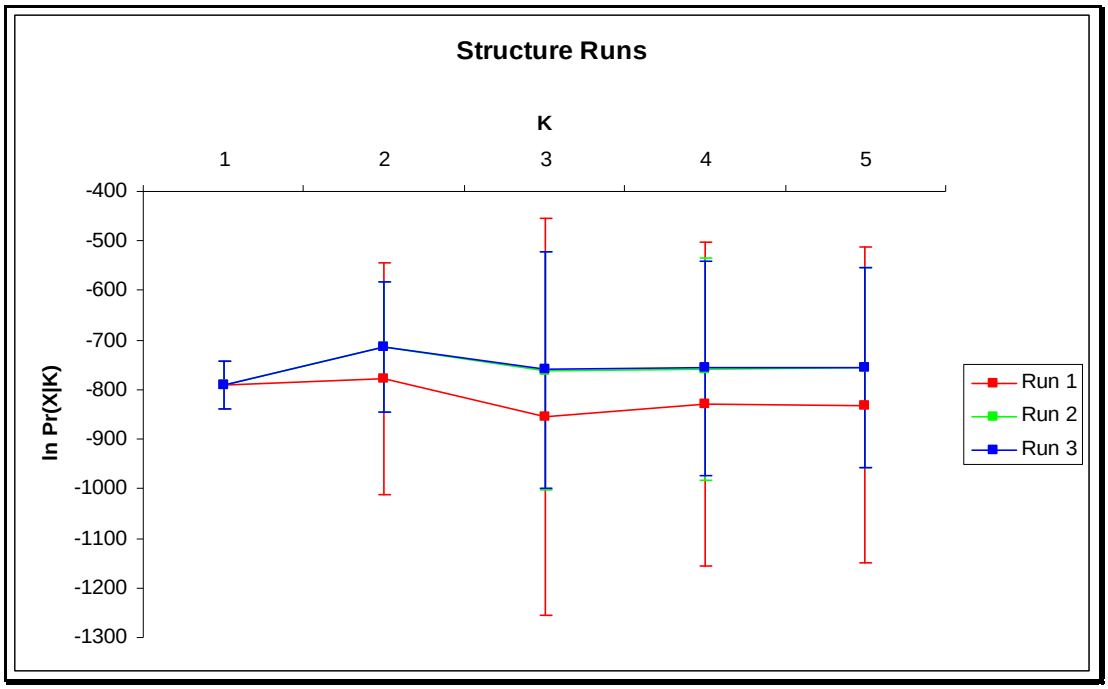

Figure 4.2. Structure analysis likelihood scores for *S. latifolia* and *S. dioica*.
K=1 to 5 for three runs of chain length 500000.

Structure runs consistently provided the highest likelihood for K=2 clusters, with likelihoods then tailing off and variances increasing. The average posterior probabilities of K from the three runs calculated using Bayes' rule, are shown in Table 4.3.

Table 4.3. Structure analysis average posterior probabilities for *S. latifolia* and *S. dioica*.

K	Average Ln Pr(X|K)	Pr(K|X)
1	-790.27	~0
2	-735.10	~1
3	-791.63	~0
4	-782.33	~0
5	-780.87	~0

The number of K clusters with the highest posterior probability was therefore K=2. Proportion membership of each individual to each of the two clusters for one of the runs is shown in Figure 4.3.

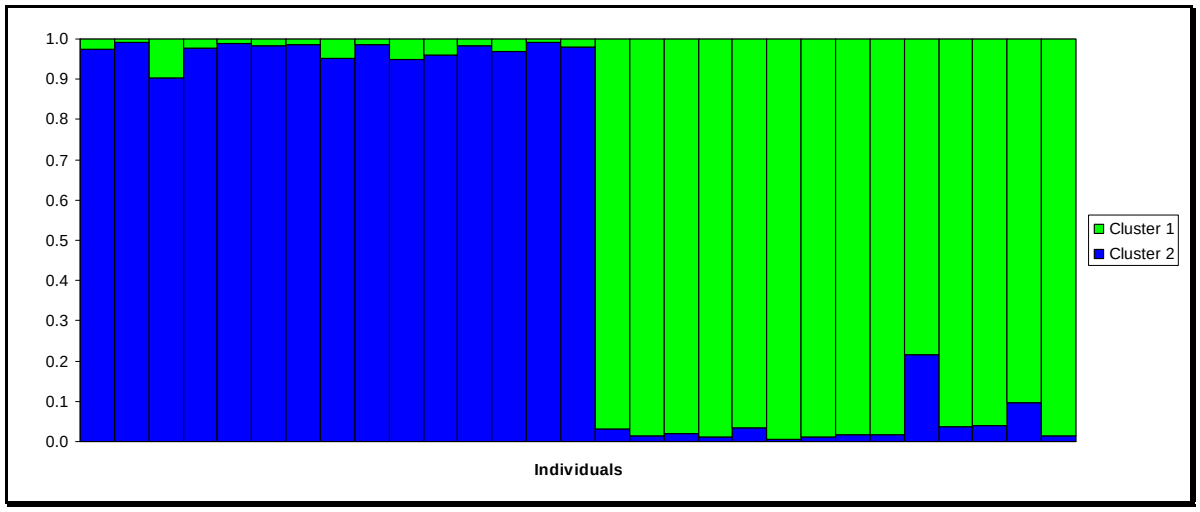

Figure 4.3. Structure analysis histogram for *S. latifolia* and *S. dioica*. Proportion membership of individuals to K=2 clusters.

The histogram shows a clear *S. latifolia*/*S. dioica* boundary with all *S. latifolia* individuals showing at least 90% membership to cluster 2, and all *S. dioica* individuals showing at least 78% membership to cluster 1.

4.3.3 IM Program

Runs 1-3 were designed to test the efficiency of the program given our data and the settings. Posterior probability curves for each parameter follow.

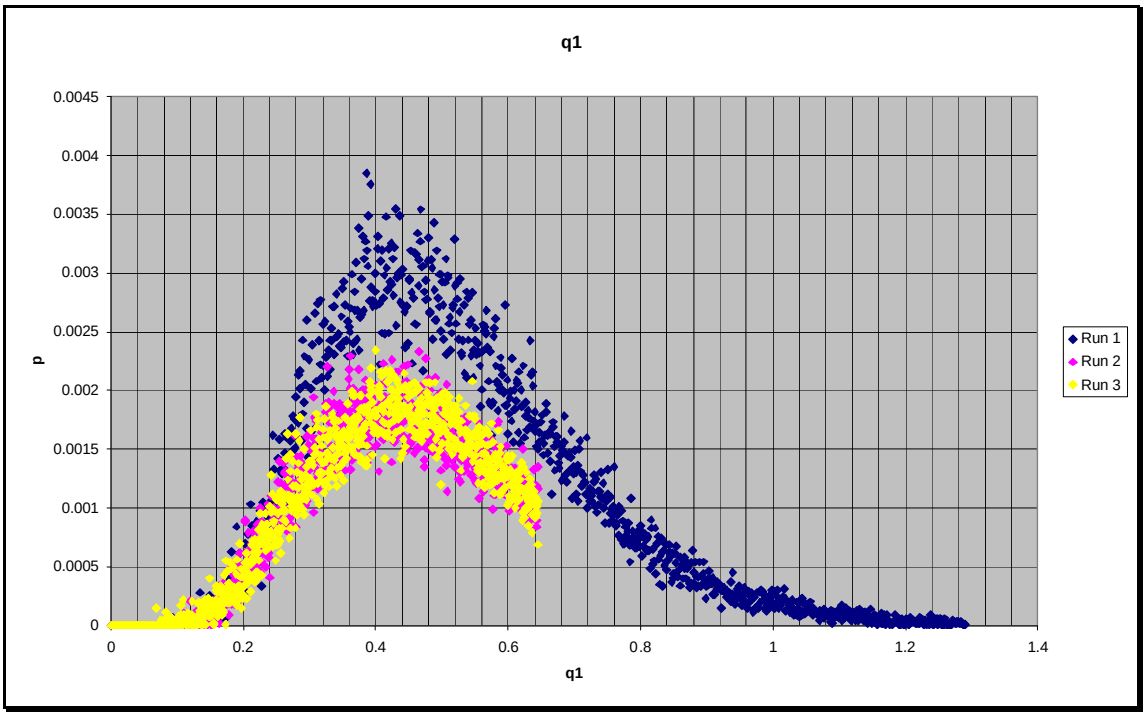

Figure 4.4. IM plot of theta for *S. latifolia,* posterior distributions for runs 1-3.
Blue=run 1, Pink=run 2, Yellow=run 3

Estimates of diversity (Theta) for *S. latifolia* for runs 1, 2 and 3 were all very similar. The mean probability graphs peak at around 0.45. The second and third run peaks do not have tails as the q1 settings (theta maximum bound for population one) were set too low. Unfortunately this means that it was not possible to get accurate upper and lower HPD90 values (90% of Highest Posterior Densities). The complete peak in run 1, however, is a good sign that there was sufficient mixing in the runs (Figure 4.4).

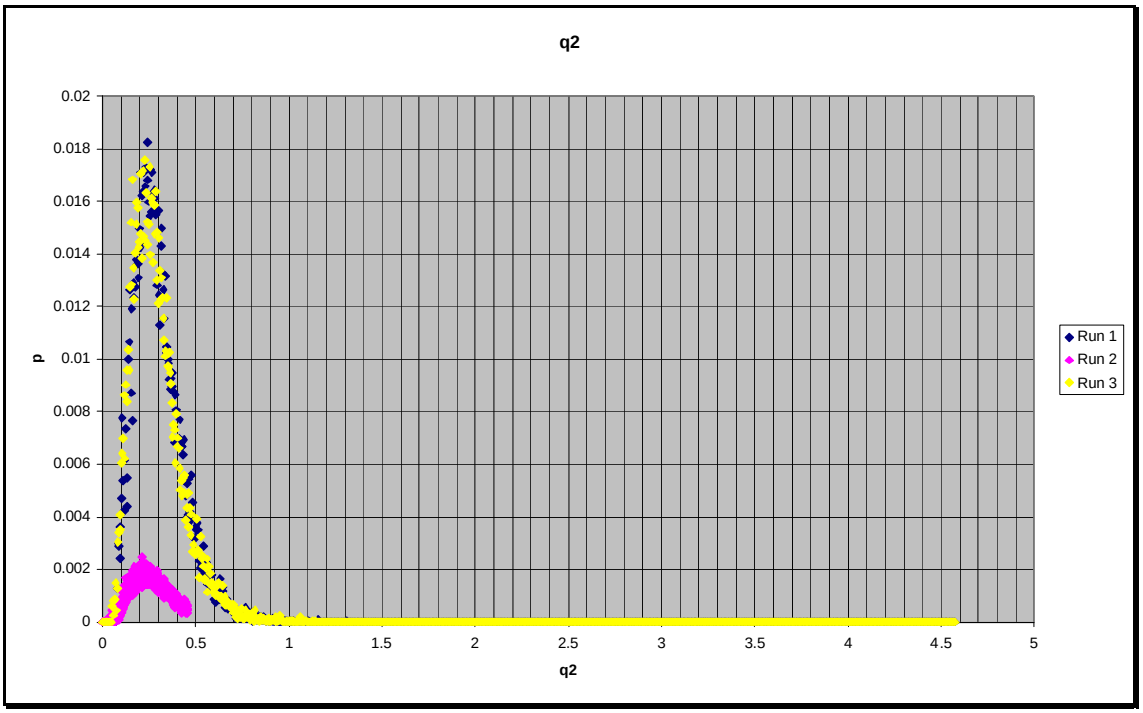

Figure 4.5. IM plot of theta for *S. dioica*, posterior distributions for runs 1-3.
Blue=run 1, Pink=run 2, Yellow=run 3

Again the posterior density peaks were very similar for all three runs for the theta estimates for *Silene dioica*. The mean of the posterior peaks for the three runs is 0.2646. Although the peaks are a good shape with low variance, it is worrying that run 2 shows such low probabilities compared to the other two peaks. The theta 2 maximum (q2) value was also too low for this run as the peak is not complete with a tail. Despite this run however, the chains appear to have mixed well again (Figure 4.5).

The posterior peaks for the migration rate from *S. dioica* to *S. latifolia* are once again very similar. The mean for these peaks is 1.39. Curiously, there appears

to be a double peak at around zero as well. Although this peak is less defined, it appears in all three runs which may indicate insufficient mixing of the chains causing the chain to get stuck on one of several peaks of high probability without exploring the landscape fully (Figure 4.6).

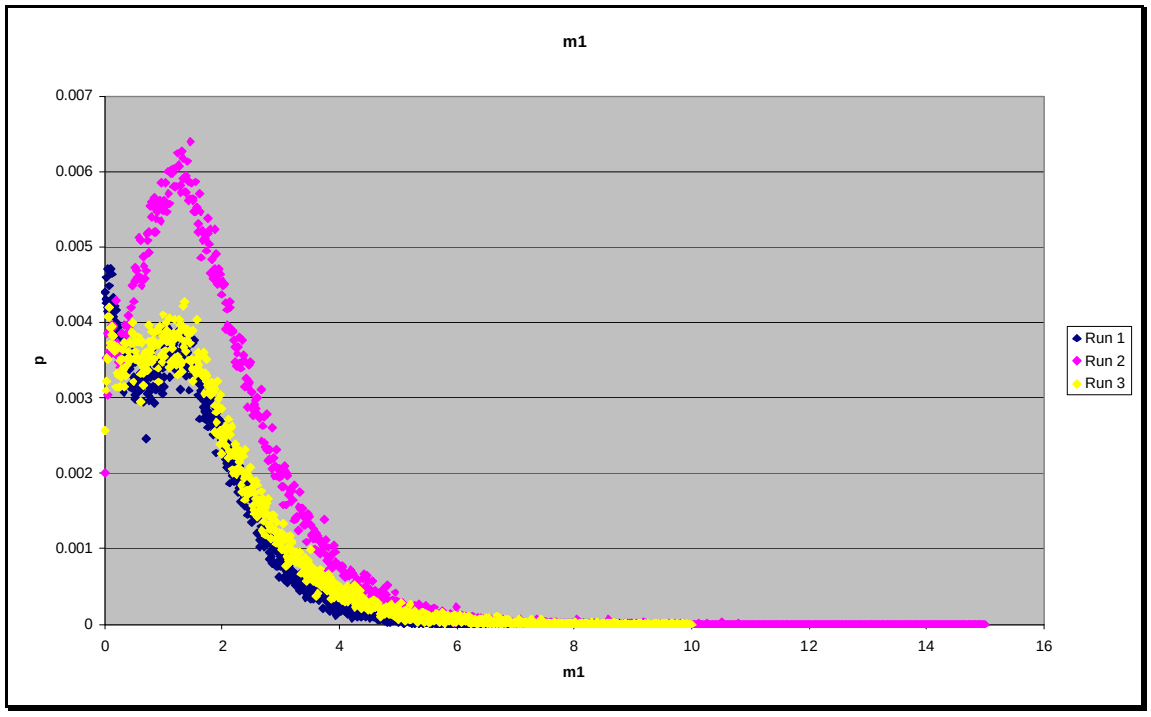

Figure 4.6. IM plot of the migration rate into *S. latifolia* from *S. dioica*, posterior distributions for runs 1-3.
Blue=run 1, Pink=run 2, Yellow=run 3

The graph for the migration rate into *S. latifolia* from *S. dioica* again shows consistent results across the three runs. It peaks very early, at its highest point the mean is at 0.866. The probabilities across all of the runs are high, and the variance is reasonably low. This is a sign of good mixing of the chains (Figure 4.7).

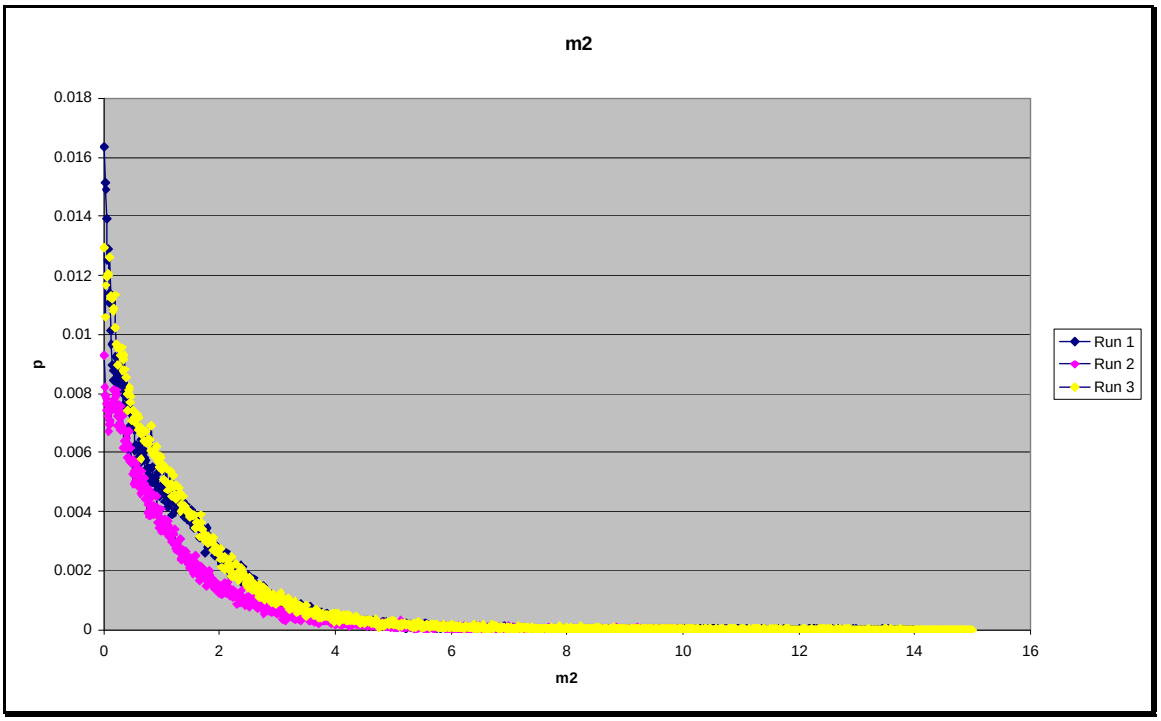

Figure 4.7. IM plot of the migration rate into *S. dioica* from *S. latifolia*, posterior distributions for runs 1-3.
Blue=run 1, Pink=run 2, Yellow=run 3

The graph for the time since splitting of the populations shows an unsuccessful attempt to estimate this parameter. The distributions show scattered points that do not form a complete peak. The distributions are also different shapes. Run 1 is slowly falling after the initial peak, but runs 2 and 3 rise after the initial peak (Figure 4.8). This could either be a sign of a complicated distribution with a long time since splitting, insufficient mixing of the chain, lack of data to fit the model, or data which does not fit an isolation with migration model.

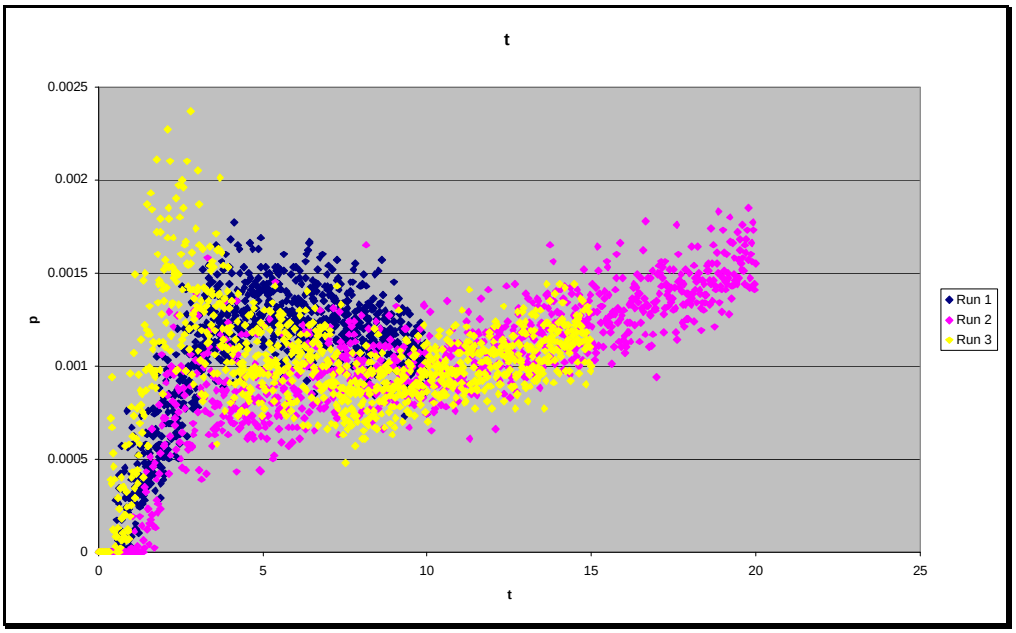

Figure 4.8. IM plot of the split time posterior distributions for runs 1-3.
Blue=run 1, Pink=run 2, Yellow=run 3

Run 4 was designed to allow larger values of t to see if convergence could be achieved at a higher value. t was therefore allowed to go up to 50 (Figure 4.9).

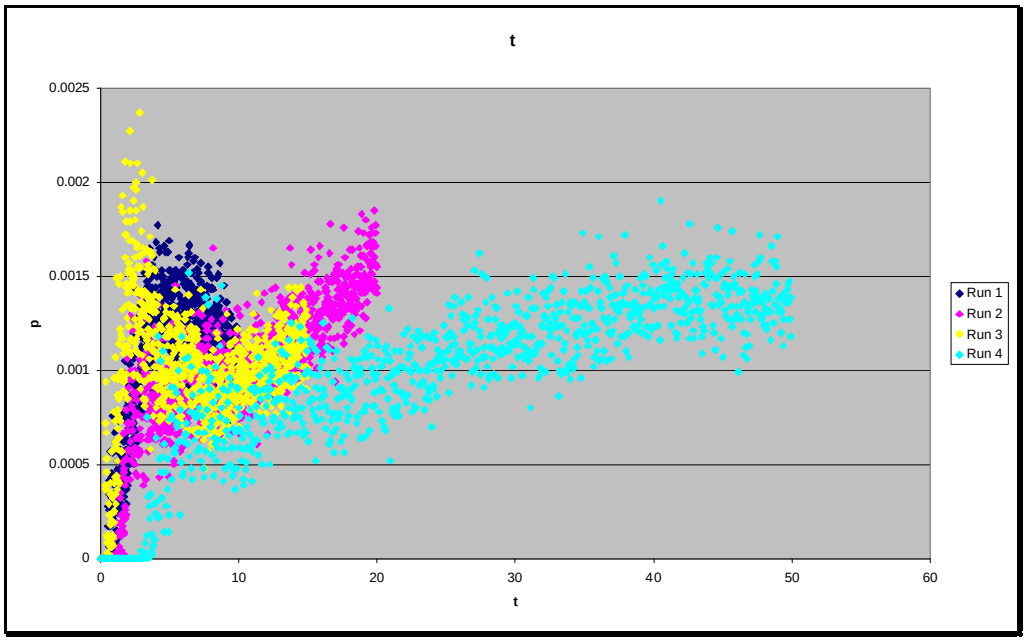

Figure 4.9. IM plot of the split time posterior distributions for runs 1-4.
Blue=run 1, Pink=run 2, Yellow=run 3, Turquoise=run 4

It is clear that the greater t max value is not sufficient for the simulation to converge on a single distribution curve for the time since splitting of the populations.

The problem of getting t to converge may have been due to insufficient mixing of the Markov chain. It was therefore decided to try a run with multiple Markov chains as part of a Metropolis-Coupled Monte Carlo Markov Chain (MCMCMC) (Figure 4.10).

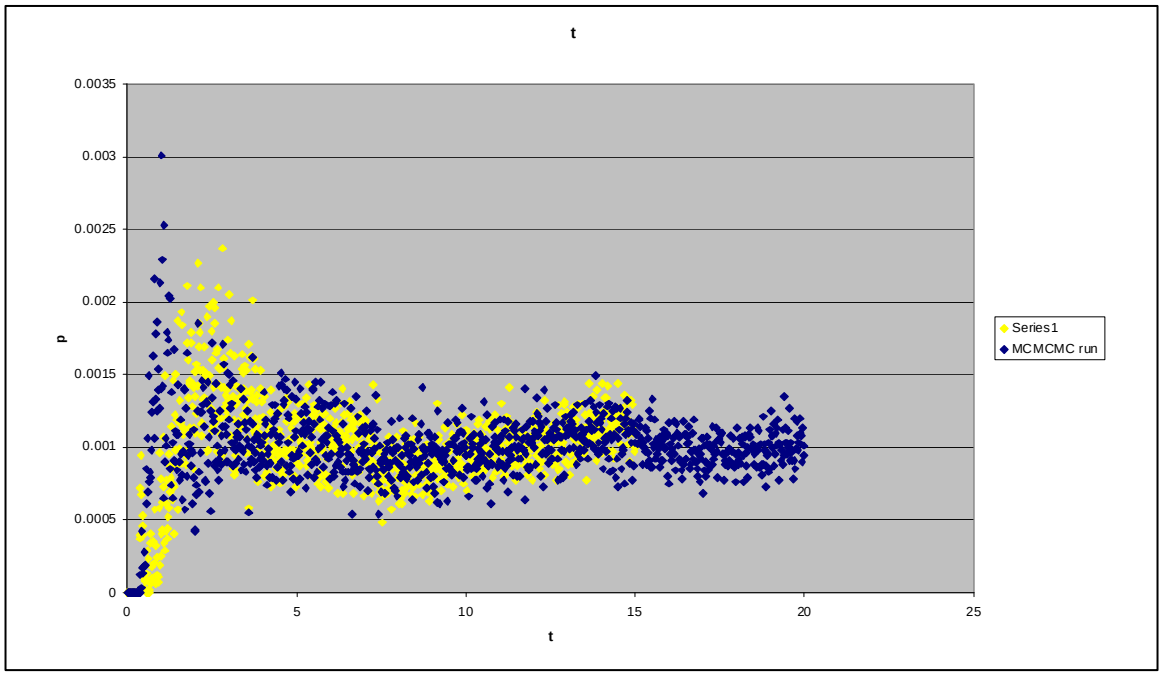

Figure 4.10. Plot of split time posterior distributions with metropolis-coupling.
Run 5 with Metropolis-coupling of Markov chains compared to uncoupled MCMC run 3. Blue=run5 (MCMCMC), Yellow=run 3 (MCMC)

MCMCMC does not appear to have improved the convergence of the Markov chain onto the posterior probability distribution for the time since splitting. This may be due to insufficient data or a poor fit to the model.

4.3.4 WH Model of Isolation

The WH Model of Isolation (Wakeley & Hey, 1997) simulates expected levels of polymorphism for each locus sampled (see appendix 3 for full results), which are then used to generate two test statistics for the fit of the data to a simple isolation model with no gene flow since time of splitting and an estimate for the time of the split from the ancestral species t generations ago. The results from this program are summarized in Table 4.4.

Table 4.4. Summarized results from the WH isolation model fitting program.
95% Confidence intervals produced by 10,000 simulations provided in brackets.

Θ *S. latifolia*	Θ *S. dioica*	Θ Ancestral	T	*P*wwh	*P*X^2
61.378 (0.013-158.160)	45.718 (0.013-98.583)	63.700 (33.650-144.296)	0.155 (0.086-0.402)	0.9310	0.8358

The probability (both X^2 and the *wwh* test statistic (Wang *et al*., 1997)) that the data fits the conservative isolation model of speciation is extremely high, and consequently this model cannot be excluded for *S. latifolia* and *S. dioica*.

4.3.5 Linkage Disequilibrium

Pairwise linkage disequilibrium was converted to a heatmap whereby red signifies LD measures close to 1, and white signifies measures close to 0.

Labels have been added to the diagonal to indicate the order of the loci in the matrix and the boundary of the within locus LD measurements (Figure 4.11).

The LD heatmap shows the predominantly low levels of linkage disequilibrium. Small pockets of LD are found in a seemingly random arrangement across the matrix, with the exception of C1A8 where there is a clear increase in LD both between sites within C1A8 and between C1A8 and the other loci.

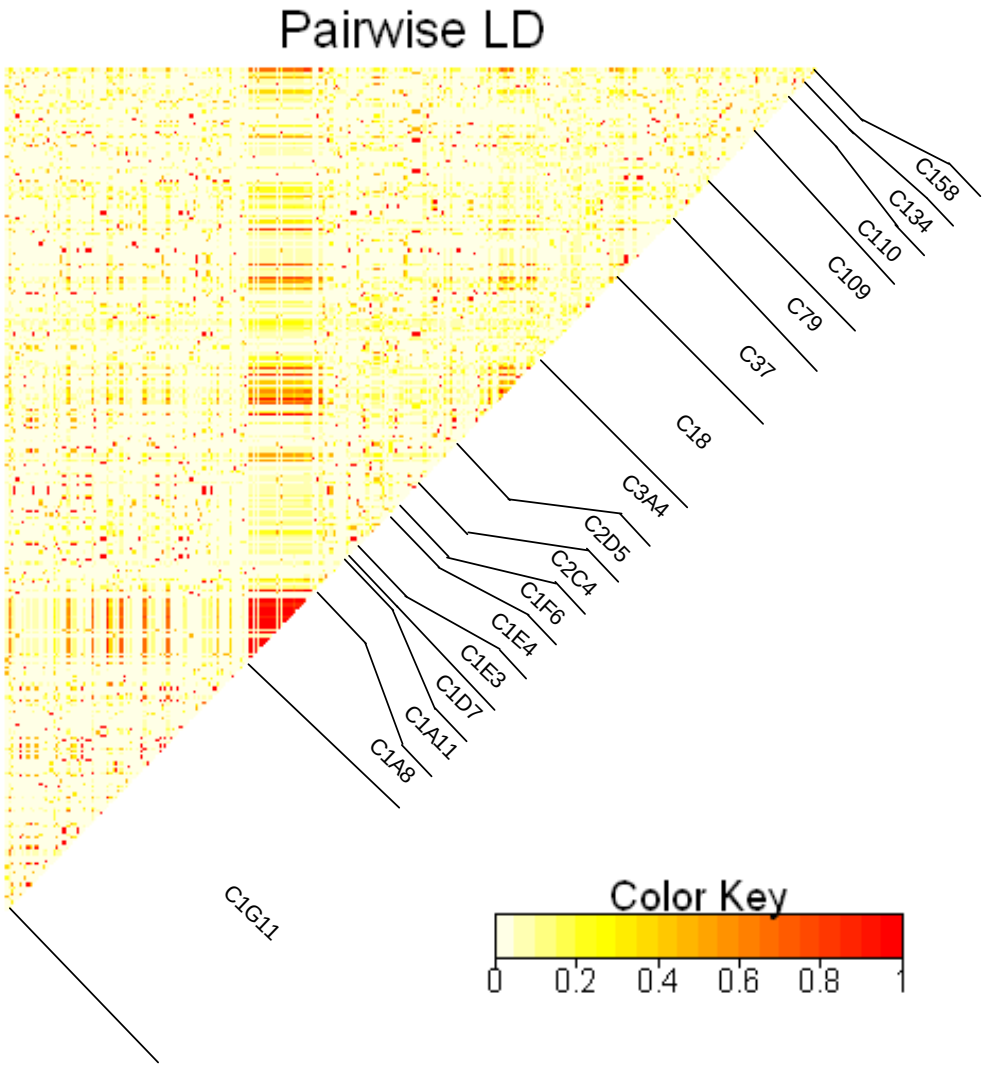

Figure 4.11. Heat map of pairwise LD measurements for 18 autosomal loci. Each coloured rectangle represents the squared correlation r^2 between a pair of polymorphic sites between *S. latifolia* and *S. dioica*. The relative locations of the SNPs and the order of the loci are indicated on the diagonal line-by-line segments. Total physical length of the genetic regions analysed = 2780bp.

4.3.6 Multilocus HKA test

The multilocus HKA test was used to identify loci under selection by allowing individual loci in each species in turn to be assumed to be under selection in the model. The model was then fitted to the data to estimate a likelihood value. The likelihood values were then compared to the neutral model with no loci selected, and a Likelihood Ratio Test used to provide a Chi-square significance estimate. Those loci that significantly improved the model are summarized in Table 4.5.

Table 4.5. Likelihood Ratio Test (LRT) for C1A8 selection variations against neutral model.
Likelihoods and Chi-Square (X^2) values. * P<0.05, ** P<0.1, *** P<0.01

Model	Likelihood	LRT X^2
Neutral – no selection	-138.924	
S. latifolia under selection at c1a8	-134.765	8.318 *** (1d.f.)
S. dioica under selection at c1a8	-134.948	7.952 *** (1d.f.)
Both species under selection at c1a8	-131.073	15.696 *** (2d.f.)

Selection at C1A8 was the only model that was a significant improvement on the neutral model with no selection. When placed under selection in the model both *S. latifolia* and *S. dioica* individually and together were significant improvements to the neutral model.

4.4 Discussion

Sequencing of the samples yielded DNA polymorphism data for 18 autosomal genes. These have been added to previously amplified genes for population genetic analysis. Several statistical tests have been applied to these genes. Firstly Tajima's D (Tajima, 1989) test was calculated using DNAsp. This test is based on the difference between two estimates of variation, the number of segregating sites and average number of pairwise differences. In a neutral equilibrium population with a constant size (i.e. the null hypothesis), Tajima's D should be around zero. Balancing selection or population decline may cause D to increase above 1, and purifying selection (or population expansion) will produce negative values of D. The results were variable across loci with both positive and negative values of D. On average the values were negative however, which would suggest that most of the genes are either under weak purifying selection or both species have undergone a recent population expansion. The D values were non-significant in all cases though, so no explicit statement can be made as to the nature of the selective forces and demography acting upon these genes using this test.

F_{ST} was also calculated for each species and locus using DNAsp, as a measure of genetic differentiation between the two species. Again, this value was variable between the loci, but most values fell within the range of zero and 0.5 which is in

line with the values seen on the X chromosome and chloroplast (Ironside & Filatov, 2005; Muir & Filatov, 2007)).

Lastly, the minimum number of recombination events for each gene was established using the 4-gamete test. Only eight loci showed an absence of recombination events since the most recent common ancestor (MRCA). These genes could then be used for the calculation of migration rate using the Isolation with Migration (IM) program (Hey and Nielsen, 2004).

Before using the IM program, a Bayesian admixture analysis was completed using the program Structure (Pritchard *et al.* 2000). This program showed that as expected, two population clusters fitted the combined *S. latifolia* and *S. dioica dataset* best. The *S. latifolia* and *S. dioica* individuals fitted into the two clusters discretely although there was a small proportion of membership to the opposite cluster for most of the individuals. As this could either be due to shared ancestral polymorphism or admixture, the IM program was used to attempt to measure the amount of migration between the species.

IM is designed to converge on the highest posterior probability for the parameters you wish to test, given the data and the prior probability that you insert. These programs are very sensitive to the type, quality and volume of data that you input, and the program parameters must be tailored to fit individual requirements. For this reason, it is usual to set up some initial runs to test the

mixing (exploration of the posterior probability landscape). Initial runs were all very similar, which is a good sign, and produced good curves for the theta and migration rate estimates. From these parameter values it is possible to estimate the population migration rate/generation (the effective rate at which genes come into a population each generation) from *S. dioica* to *S. latifolia* and vice versa using the formula Θ_x / (m_x/2). This gives us rates of 0.75 and 0.72 genes coming into *S. latifolia* and *S. dioica* respectively per generation.

Unfortunately, satisfactory parameter estimates for the time since the split of the species and ancestral theta could not be achieved despite a fourth long run with an extended tmax value, and a final run making use of the optional Metropolis-coupling algorithm. The last run used four chains of various “heats” to interchange with the “cold” recording chain randomly to explore the distribution space more fully but produced very similar results to the previous attempts. This suggests that this problem is due to a lack of data for the program to work with, and as a result extreme caution should be used when interpreting the results of the program, such as the migration rate estimates.

The WH program was used to fit an isolation model to the data as an alternative. This model assumes that there has been no gene flow since the two species split from a common ancestor at time T generation ago. The model estimates several parameters such as theta (the population mutation parameter 4Neμ estimated over all sites), which provides an estimate of the population size for

each species and the ancestral species from which they derived. *S. latifolia* appears to have a larger population size than *S. dioica*, which supports the slightly larger distribution of *S. latifolia*. The ancestral species appears to have had a larger population size than either *S. latifolia* or *S. dioica*. This is marginal for *S. latifolia*, and the confidence intervals for the parameter estimates overlap. The *S. dioica* theta value is much lower although variance is still high. The T parameter estimates produced by this analysis provide an estimate of the time since divergence in $2N_1$ (ie. *S. latifolia* population size) generations. Assuming a population size of 1 million for *S. latifolia* and a 2 year generation time, we can estimate the time since divergence at 620,000 years, with confidence intervals extending this estimate to between 300,000 and 1.6 million years.

These values are consistent with the assumption that *S. latifolia* and *S. dioica* are extremely closely related. Perhaps the most persuasive result from the WH analysis is the isolation model fit, which provides us with extremely strong evidence that these two species evolved in isolation, as the model assumes no gene flow. Although in reality, few model species would fit into such a strict model, this provides reasonable evidence that there has at least been very little recent gene flow between these two species, as the effect is undetectable using the WH model.

The linkage disequilibrium analysis confirmed that there is a globally low level of LD across this dataset, consistent with extremely limited gene flow in the recent

history of *S. latifolia* and *S. dioica*. The single exception to this is the locus C1A8, which is likely to encode a transporter protein. This locus exhibits high levels of LD between sites within the locus as well as raised LD for sites between the loci. This pattern could have been caused by introgression of this gene or by selective processes acting upon it. To investigate this, the Maximum Likelihood HKA was used to show that highly significantly better likelihoods were achieved when the C1A8 locus was placed under selection in the model for both *S. latifolia* and *S. dioica*. Considering that the previous results suggested that introgression was unlikely to have been occurring at any appreciable levels, it was important to identify whether separate selective sweeps had occurred in each species.

A single selective sweep crossing a species boundary would have a similar effect to what was seen by Muir and Filatov (2007) in the chloroplast of *S. latifolia* and *S. dioica*. We would expect to see a marked reduction in nucleotide diversity at this locus, and a low level of genetic differentiation between the species and populations. Conversely, separate selective sweeps would still reduce diversity levels, but would also increase genetic differentiation between the two species. The C1A8 F_{ST} value between *S. latifolia* and *S. dioica* is much higher than the other loci (0.946) due to the large amounts of fixed differences seen between the two species (17 and no shared polymorphic sites in ~850bp). The amount of differentiation between these two species is indicative of two separate sweeps, which supports a theory of little or no gene flow.

There is another possibility however. There are varying levels of genetic differentiation between the loci used in this study, which could indicate that there is a porous species boundary in effect, preventing genes such as C1A8 from crossing the species boundary (as perhaps they are important for retaining reproductive isolation between the species), but allowing other genes to pass through (such as those on the X chromosome and chloroplast, Ironside & Filatov, 2005; Muir & Filatov, 2007) which are not implicated in reproductively isolating the species.

It is important to note that it is unlikely that even a porous species boundary has been in place for a substantial amount of time. The estimates of the time since the divergence of *S. latifolia* and *S. dioica* generated from the WH isolation modeling suggests that the two species diverged around 300,000- 1.6 million years ago, around the time of the Pleistocene, and would have been limited to glacial refugia suited to their individual habitat preferences. Taylor & Keller (2007) found evidence that *S. latifolia* found refuge from the ice age in Southern Europe, possibly the Balkan or Iberian Peninsulas, whereas Prentice *et al*. (2008) suggested that *S. dioica* emerged from several refugia, probably in the Mediterranean, Balkans or Caucasus. The two species would then have started expanding their ranges from their glacial refugia around 10,000 years ago as the Pleistocene ended and the climate stabilized (Hewitt, (1996). *S. dioica* would most likely have spread rapidly with the expansion of deciduous forests into Northern Europe (Hewitt, 1996), while *S. latifolia* would have followed

considerably later with the spread of agriculture (Prentice, 1986b). At this point the species would have come back into secondary contact allowing hybridization where the spread of agriculture met existing deciduous forests.

The above studies suggest that although hybridization is known to occur between *S. latifolia* and *S. dioica*, it is probable that only small amounts of introgression have occurred during the history of these species. Following the separation of the two species, reproductive isolation has evolved, which will only be reinforced if hybrids are less fit than their parents. This may indeed be the case as is reflected in the lack of F1 hybrids at some hybrid zones (Minder et al., 2007). Less introgression is expected to occur if few hybrids are produced or if the two species are diverged and have only recently come into secondary contact, both of which are likely to be the case for *S. latifolia* and *S. dioica*. Any introgression that occurs is also less likely to be evolutionarily significant when parental populations are large, as many more variants will be introduced into these populations by mutation alone. In conclusion, introgression is unlikely to have had, and is unlikely to have any significant impact in the future on the evolution of *S. latifolia* or *S. dioica*.

5 THE EVOLUTION AND POPULATION GENETICS OF *S. diclinis*

5.1 Introduction

Silene diclinis is a rare endemic species of campion that is found around the town of Xàtiva in Valencia, Spain. It is thought to number less than 2000 individuals, which are spatially separated into several populations which vary in size from several hundred to a few dozen individuals (Prentice, 1976). Its habitat destruction is thought to be the major factor in its decline, but the nature of any genetic degeneration caused by the low population numbers is poorly understood.

Low population numbers in plants can be due to several factors (Harper, 1977). Firstly, available habitat sites are limited, and those that exist may be outside the natural dispersal range of the species. The habitat sites may have low carrying capacities, or there may be problems with displacement in the habitat. In the case of *S. diclinis*, all of these may have contributed to the low population numbers. *S. diclinis* is invariably associated with disturbed, well-drained ground around the carob groves and slopes of Xàtiva, and populations are thought to be endangered by disruption of these areas. There are certainly several spatially separated populations (Prentice, 1976) and *S. diclinis* is pollinated by

bumblebees which are known to have short-range foraging habits (Osbourne *et al*., 2008) that probably do not allow pollen dispersal between some or all of these populations. *S. diclinis* is also not a competitive species and may be displaced by the native thicket and scrub.

Population number is not necessarily the most informative measure of the number of individuals able to pass their genes to the next generation. Effective population size is a better measure of this, and includes many factors other than the number of individuals in a population (Wright, 1931; Wright, 1938). The effective population size is often smaller than the actual population size as it is affected by the level of inbreeding, unequal sex-ratios, population structuring and non-random mating.

It is thought that an uneven sex-ratio bias persists in *S. diclinis*, with females representing 60% of the population (Prentice (1984a). It is also known to have spatially isolated populations and allozyme studies performed by Prentice (1984b) showed that there may also be genetic structuring within populations if not between them. It is possible that there is also some level of non-random mating and inbreeding due to the population structuring. The bumblebee pollinators prefer to forage in areas that are in close proximity to the nest (Osborne *et al*., 2008), probably reducing the amount of between population migration. Considering these factors, it is likely that the effective population size

of *S. diclinis* is much smaller than the estimated few thousand remaining individuals.

The expectation for a species with low effective population size is that it will have a reduced ability for adaptation to changes in its environment, and be more susceptible to diseases and pests than species with greater population sizes (Fisher, 1930; Hamilton, 1982; Beardmore, 1983). The genetic variation within a species is controlled by four key factors: mutation, selection, migration and genetic drift, combined with the effects of recombination. In a small population, genetic drift is expected to provide the largest contribution to the structure of variation, and such populations will lose variation more quickly as a result. This can result in genetic disintegration of the species, eventually leading to extinction (Barrett & Kohn, 1991). A further consequence of small population size may be inbreeding depression, which is likely to be relevant in a dioecious species that has spatially separated populations such as *S. diclinis*.

Species which have had a historically small population size may have developed genetic systems capable of offsetting the effects of inbreeding and gained adaptations that allow them to cope with the disadvantages of their scarcity (Hopper & Moran, 1981). Of course the reverse is true for those species which have only recently become small (for instance after a recent severe bottleneck) which may be more sensitive to the hardships of a small population size.

The genetic effects of small population sizes have been investigated in natural populations and computer simulations. Lacy (1987) made several conclusions following computer simulations, namely that drift is the predominating force in reducing variation in small populations and mutation and selection only have small effects on the rate of loss of variation. Sub-divided populations lost variation from within them, but retained the overall variation across them better than a single population. Migration from a large population was able to slow down, halt and in some cases reverse the loss of variation.

The closest relatives of *S. diclinis* are *S. latifolia* and *S. dioica*, both widespread and common species, and *S. diclinis* is still able to hybridise with both to produce fertile offspring in greenhouse conditions (Baker, 1950). Only the distribution of *S. latifolia* overlaps with that of *S. diclinis* and they have similar habitat preferences, but *S. diclinis* is mainly pollinated by bumblebees (*Bombus spp*.) (Osborne *et al*., 2008), and *S. latifolia* is pollinated by the Lychnis moth *Hadena bicruris*. It has already been shown in *S. latifolia* and *S. dioica* (which is also bumblebee pollinated) that this difference in pollinators is not sufficient to prevent hybrids forming naturally at sites where the two species meet (Baker, 1950; Minder *et al*., 2007), that factors such as the flower opening times of *S. latifolia* and *S. dioica* do overlap (Hess *et al.*, 1972) and that scent compounds are somewhat similar (Waelti *et al*., 2008). It is not known whether this is the case for *S. latifolia* and *S. diclinis*, but no known natural hybridization has been recorded for the two species.

There is evidence however, that there may be some major rearrangements in the chromosomes of *S. diclinis*, which may indicate a stage of reproductive isolation. *S. diclinis* has evolved a reciprocal translocation between the ancestral Y chromosome and an autosome creating a neo-sex chromosome system (Howell *et al*., in press). Neo-sex chromosome systems have been well characterized, and several theories of how they can spread to fixation in a population have been postulated. One theory is that drift can simply allow a rearrangement to spread throughout a population eventually replacing the ancestral state (Charlesworth *et al.*, 1987). The second is that sexually antagonistic loci are advantageous when a rearrangement brings them closer to the sex-determining region (Charlesworth & Charlesworth, 1980). A third theory states that inbreeding generates associations between heterozygosities at different loci. A selective advantage can arise for a rearrangement between an autosome and a sex chromosome when there is selection in favour of heterozygotes, and in particular, Y-autosome translocations will become fixed in the population (Charlesworth & Wall, 1999). *S. diclinis* probably fits both the drift and inbreeding models due to its small numbers and subdivided populations.

The following experiments were designed to evaluate the effect that the low population size of *S. diclinis* has had on the genetic diversity and population structuring between the spatially separated populations. It was also important to assess whether migration of alleles from other species was likely to have occurred in the history of *S. diclinis* as this may have reduced the effect of

inbreeding. Lastly, it was important to postulate if the low population size of *S. diclinis* is due to a recent population bottleneck or whether *S. diclinis* has historically been scarce as this will have an impact on the likelihood of its future survival. These analyses provide information that is invaluable to the conservation strategy for this species.

5.2 Methods

5.2.1 Samples

Samples were provided by D. Filatov and are listed in Table 5.1.

Table 5.1. *S. diclinis* individuals sampled from around Xativa, Spain.

Population	Sample numbers	Sample Identifiers	GIS
1	1-7, 41-47	Sdic938-944, 1015-1021	W0.37571; N38.99797
2	8-12	Sdic945-949	W0.39562; N38.99901
3	13-16, 64-70	Sdic950-953, 1038-1045	W0.36271; N39.03642
4	17-23	Sdic954-960	W0.53122; N39.03122
5	24-28, 54-59	Sdic961-965, 1028-1033	W0.35010; N39.01398
6	29-33, 60-62	Sdic966-970, 1034-1036	W0.33954; N39.01257
7	35-40, 63	Sdic370-380, 1037	W0.53917; N38.99250
8	48-53	Sdic1022-1027	W0.35147; N39.01398

The 8 populations can be seen on a regional map of Xativa illustrating possible geographic barriers in Figure 5.1.

5.2.2 PCR and Sequencing

Autosomal genes selected for further analysis were amplified and sequenced using the primers listed in Appendix 1. Heterozygous sites were either resolved (by random assignment to one of the alleles) for analysis, or kept as an unresolved dataset with heterozygous bases coded using the IUPAC

(International Union of Pure and Applied Chemistry) notation. DNA alignments are provided as Proseq3 files on compact disc inside the back cover.

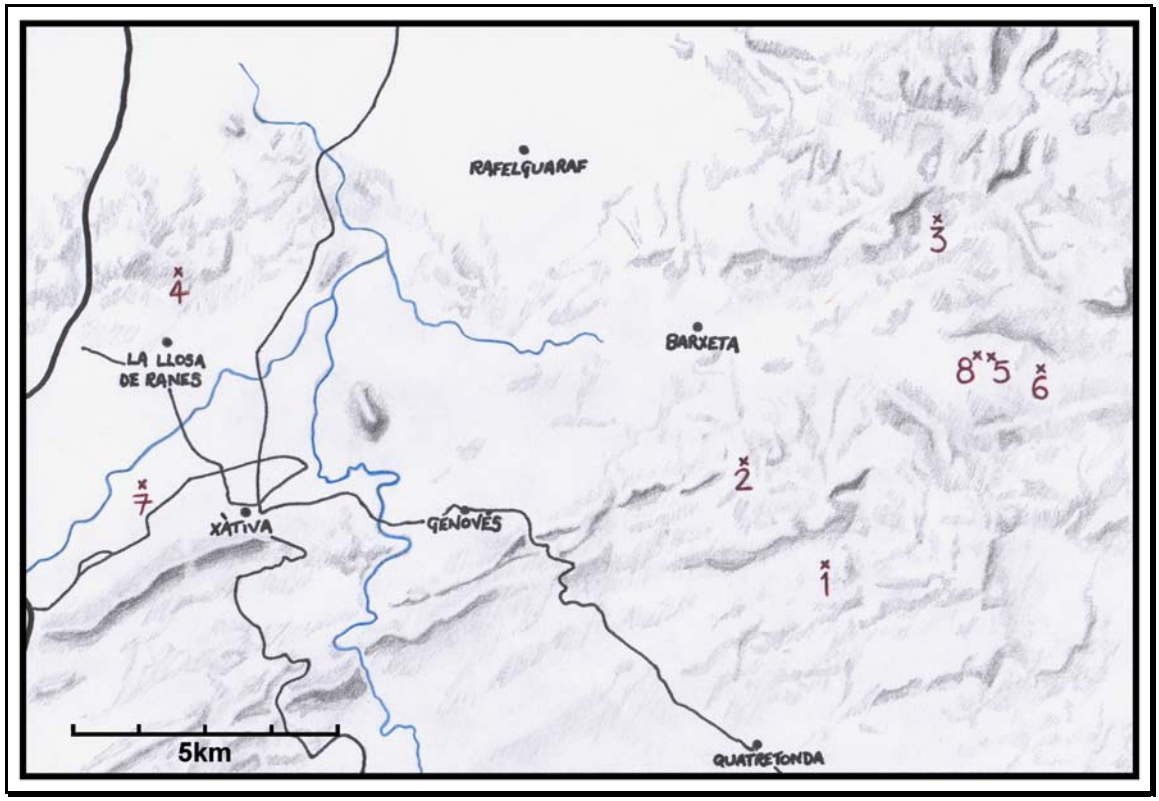

Figure 5.1. Map showing sampling locations of *Silene diclinis* populations. Populations marked in red, rivers and streams in blue, roads in black and geological features are shaded.

5.2.3 Sequence Analysis using DNAsp

Sequences were entered into DNAsp version 4.10.7 program (Rozas and Rozas, 1999) to calculate basic statistics from the resolved autosomal sequences. An *S. diclinis* dataset was created using Proseq V3, and the 8 distinct populations specified. Diversity (π), haplotype diversity and Tajima's D (Tajima, 1989) were measured for each population and locus. F_{ST} was also

calculated for population and locus using DNAsp, as a measure of genetic differentiation (population structure).

5.2.4 GenAlEX 6 Analyses

Further tests were conducted using GenAlEX6 (Peakall & Smouse, 2006). Haploid data for all loci was recoded into a numerical format and placed into a MS Office Excel document with each SNP occupying a single cell. Individual, population and locus identifiers were also inserted into the matrix as well as GIS data relating to each individual.

Geographic and genetic distance matrices could then be generated for use in further analyses such as the Mantel Test (Mantel, 1967) for isolation by distance (following the methods of (Smouse *et al.*, 1986; Smouse & Long, 1992), Analysis of Molecular Variance (AMOVA, from the methods in (Excoffier *et al.*, 1992; Huff *et al.*, 1993; Peakall *et al.*, 1995; Michalakis & Excoffier, 1996) and global spatial autocorrelation (following methods of (Smouse & Peakall, 1999; Peakall *et al.*, 2003; Double *et al.*, 2005)

The AMOVA was performed with an assumption of either 2 or 5 regions, based purely on the geographical spread of the populations and barriers such as roads and rivers (see Figure 5.1). The samples included in the 2-region (East/West) model were populations 4 and 7 in the Western region with the remainder in the Eastern region. Region 1 of the 5-region model was composed of individuals

from population 4, Region 2 individuals from population 7, Region 3 population 3, Region 4 population 8, 5 and 6, and Region 5 populations 1 and 2.

The spatial autocorrelation analysis should reduce locus to locus and allele to allele “noise” as it employs a multivariate technique to simultaneously assess the spatial signal generated by multiple loci.

Combined ProseqV3 (Filatov, 2002) sequence alignments including *S. latifolia* and *S. dioica* were also used to compare the levels of diversity (π) and divergence (Da, Fst) using DNAsp. Population structure was investigated using the Bayesian approach of Structure as described below.

5.2.5 Phylogenetic Analysis

Phylip, a maximum likelihood phylogenetic approach (Felsenstein, 2004) was also used to establish the relationship of *S. diclinis* to *S. latifolia*, *S. dioica* and outgroups such as *S. vulgaris*. Sequence alignments created from ProseqV3 for each locus were saved as Phylip file types for input into Phylip executables. To enable bootstrap analysis of the trees to be completed, each input file was read into Seqboot program to generate a multiple dataset from it by bootstrap resampling. 1000 replicates were created in this way.

The Seqboot output file was subsequently entered into the program Dnaml which estimates phylogenies from nucleotide sequences using the maximum

likelihood method. Default settings were used apart from setting the outgroup individual in each case (in most cases an *S. vulgaris* individual, but for C110 an *S. heuffelii* individual was used), and the program was set to complete the analysis using multiple datasets (1000).

The Dnaml output tree file was then entered into the program Consense which computes the majority rule extended consensus tree, once again indicating the correct individual to use as the outgroup root. Finally, the output tree from this program was drawn using the program Drawgram to plot rooted phylogenies, and the output file from Consense was used to calculate bootstrap values.

5.2.6 Bayesian admixture analysis

A model-based clustering method implemented in the program Structure (Pritchard *et al.* 2000) was used to assign individuals probabilistically to homogenous clusters (*K* populations) without consideration of sampling localities. Estimated posterior probabilities for the simulated model fitting the data were calculated assuming a uniform prior for *K*, where $1 \leq K \leq 5$.

An input file was created where each individual plant's identifier was followed by its haplotype for each locus recoded into a numerical format. To minimize the effect of the starting configuration during the Monte Carlo simulation, we conducted a burn-in of 100,000 iterations, before data for the parameter estimations were collected from a further 500,000 iterations. Three independent

runs of the Markov chain, each of least 500,000 updates were performed to assure convergence of the chain and homogeneity among runs for each prior of *K*. The posterior probabilities of *K* were then calculated using Bayes' rule. The program was run without population identifiers and either in the admixture mode which assumes that each individual has drawn some fraction of the genome from each of the populations considered (for the *S. diclinis* population runs) or in the non-admixture mode which does not assume that individuals have drawn a fraction of the genome from each of the populations (for the *S. diclinis*/*S. latifolia*/*S. dioica runs*). Allele frequencies were allowed to be independent in all runs.

5.2.7 WH Isolation Model

The WH isolation model was used to attempt to reject a null hypothesis of no gene flow between the three species (Wakeley & Hey, 1997). The program fits a simple speciation model (the isolation model) to multilocus datasets. The model makes the following assumptions;

- The two species of interest arose from a single ancestral species t generations ago.
- The common ancestral species had a constant effective population size NA.
- The two descendent species also have constant effective population sizes N1 and N2.

- There has been no gene flow since separation from the common ancestor at time t.
- All mutations are neutral. For this reason, any genes found to be under selection using the above Multilocus Maximum Likelihood HKA method were excluded.

The program provides an output file assessing the quality of fit of the data to the simulation model with a chi-square statistic and the *wh* statistic (Wang *et al*., 1997). Also provided is a table of parameter values with 95% confidence intervals and means, and a table of observed and simulated means of variants (See Appendix 3). 10,000 simulations were run.

5.2.8 Bottleneck Analysis

Bottleneck is a program for detecting recent population bottlenecks using allele frequency data to detect a relative excess in heterozygosity (Cornuet & Luikart, 1996). An input file was created for the five polymorphic genes. After a title line, each subsequent line contained a locus name, the number of alleles, the sample size and the amount of heterozygosity (Nei, 1987) as calculated in DNAsp. The program was run using the infinite alleles model (IAM) which is likely to be the best fit for the SNP data, and the stepwise mutation model (SMM) as a conservative addition. The output file provides estimated heterozygosity,

measures of heterozygote excess or deficiency along with probabilities from Sign and Wilcoxen tests (see Appendix 6).

5.3 Results

5.3.1 DNA Extraction and Sequencing

Amplification and sequence quality of the 18 loci used in the previous study was poor overall, and consequently only 6 of the 18 were successfully amplified and sequenced in enough individuals for each population to be included in this study. These were C34, C1A11, C1G11, C110, C109 and C37. The sequence lengths and number of individuals sequenced for each locus are recorded in Table 5.2.

5.3.2 Intraspecific Diversity and Neutrality analysis

Table 5.2 shows the statistics and tests conducted using DNAsp for the 8 *S. diclinis* populations. Table 5.2 shows that the *S. diclinis* individuals display the highest mean level of diversity in loci C109 and C110, both of which show the highest mean replacement site diversity, and C110 also has highest mean silent site diversity. *S. diclinis* shows a total lack of diversity for locus C37. These results are also reflected in the Haplotype diversity measure (Hd).

C110 and C109 are also unusual in that they also have the only positive mean Tajima's D value. All Tajima's D values are not significant however.

Table 5.2. Summary statistics for *S. diclinis* populations.

Locus	Samples	Length	Segregating Sites	Pi all	Pi S	Pi R	Hd	D
C34								
Total	43	221	2	0.001	0	0.00027	0.215	-0.97388
Pop1	8	232	1	0.00231	0	0.00297	0.536	1.1665
Pop2	4	221	0	0	0	0	0	NA
Pop3	9	262	0	0	0	0	0	NA
Pop4	2	327	0	0	0	0	0	NA
Pop5	8	327	2	0.00153	0	0.00027	0.464	-1.31009
Pop6	5	327	1	0.00122	0	0.00158	0.4	-0.8165
Pop7	6	327	0	0	0	0	0	NA
Pop8	2	327	2	0.00612	0.01418	0.00394	1	NA
C1A11								
Total	38	218	3	0.00219	0.00792	0.00061	0.333	-0.7301
Pop1	5	225	0	0	0	0	0	NA
Pop2	4	238	3	0.0077	0.02602	0.00294	0.833	1.08976
Pop3	9	218	2	0.00357	0.01171	0.00132	0.639	0.1959
Pop4	5	328	0	0	0	0	0	NA
Pop5	5	241	0	0	0	0	0	NA
Pop6	7	241	1	0.00119	0	0.00153	0.286	-1.00623
Pop7	3	328	0	0	0	0	0	NA
Pop8	0	NA	NA	NA	NA	NA	NA	NA
C1G11								
Total	18	425	4	0.0018	0	0	0.614	-0.72366
Pop1	5	838	1	0.00048	0	0	0.4	-0.8165
Pop2	0	NA	NA	NA	NA	NA	NA	NA
Pop3	3	425	4	0.00244	0	0	0.667	NA
Pop4	0	NA	NA	NA	NA	NA	NA	NA
Pop5	2	492	0	0	0	0	0	NA
Pop6	3	1093	5	0.00305	0	0	1	NA
Pop7	4	905	7	0.00516	0	0	1	-0.44637
Pop8	0	NA	NA	NA	NA	NA	NA	NA

Locus	Samples	Length	Segregating Sites	Pi all	Pi S	Pi R	Hd	D
C110								
Total	50	217	5	0.00823	0.02859	0.00277	0.858	1.45247
Pop1	8	217	4	0.00839	0.02795	0.00277	0.857	0.78822
Pop2	2	217	1	0.00461	0.02174	0	1	NA
Pop3	10	217	4	0.00707	0.0256	0.00208	0.867	0.32418
Pop4	5	217	3	0.00829	0.02609	0.00282	0.9	1.57274
Pop5	9	217	5	0.01126	0.04227	0.00216	0.972	1.36246
Pop6	7	217	2	0.00395	0.01242	0.0024	0.667	0.20619
Pop7	6	217	3	0.00461	0.01186	0	0.6	-1.23311
Pop8	5	217	3	0.00737	0.02087	0.00282	0.9	0.699
C109								
Total	45	337	6	0.0049	0.0039	0.0021	0.851	0.61209
Pop1	6	337	5	0.0063	0.0069	0.0013	0.8	0.36689
Pop2	4	337	3	0.0045	0.0038	0.0019	0.833	-0.75445
Pop3	9	337	3	0.003	0.003	0.0009	0.75	-0.35929
Pop4	6	337	0	0	0	0	0.8	0.76798
Pop5	5	337	5	0.0083	0.0077	0.0031	0.9	1.12397
Pop6	7	337	3	0.0034	0.0044	0	0.81	-0.30187
Pop7	5	337	2	0.003	0.0023	0.0015	0.8	0.24314
Pop8	3	337	2	0.004	0.0051	0	0.667	NA
C37								
Total	42	288	0	0	0	0	0	NA
Pop1	7	288	0	0	0	0	0	NA
Pop2	4	288	0	0	0	0	0	NA
Pop3	9	288	0	0	0	0	0	NA
Pop4	3	288	0	0	0	0	0	NA
Pop5	7	288	0	0	0	0	0	NA
Pop6	6	288	0	0	0	0	0	NA
Pop7	5	288	0	0	0	0	0	NA
Pop8	1	288	0	0	0	0	0	NA

NA = Not applicable

5.3.3 Intraspecific Genetic Differentiation

Table 5.3. *S. diclinis* population Fst values for autosomal loci.

Ave. Pop. Fst/Gene	C34	C1A11	C1G11	C110	C109	C37
1	0.234686	0.28335	0.278733	0.014757	0.0547	N/A
2	0.095233	0.22225	0.92235	0.1019	0.0333	N/A
3	0.095233	0.25	0.718817	0.0483	0.072357	N/A
4	0.095233	0.28335	0.5625	0.110857	0.119414	N/A
5	0.030614	0.28335	0.5625	0.099057	0.033614	N/A
6	0	0.28335	0.5625	0.136786	0.045971	N/A
7	0.095233	0.28335	0.525467	0.088814	0.088457	N/A
8	0.095233	0.28335	-	0.146271	0.137357	N/A
Ave. Fst/Gene	0.0927	0.272	0.517	0.093	0.073	N/A

Table 5.3 shows average Fst varied between genes from ~0.5 for C1G11 to ~0.07 for C109.

5.3.4 Intraspecific Mantel Test for Isolation by Distance

If isolation by distance has occurred in *S. diclinis*, we would expect individuals with greatest geographical distance between them to also have highest total genetic distance creating a positive correlation. No significant correlation (R^2=0.001; *P*=0.457) is evident from the Mantel Test (Figure 5.2).

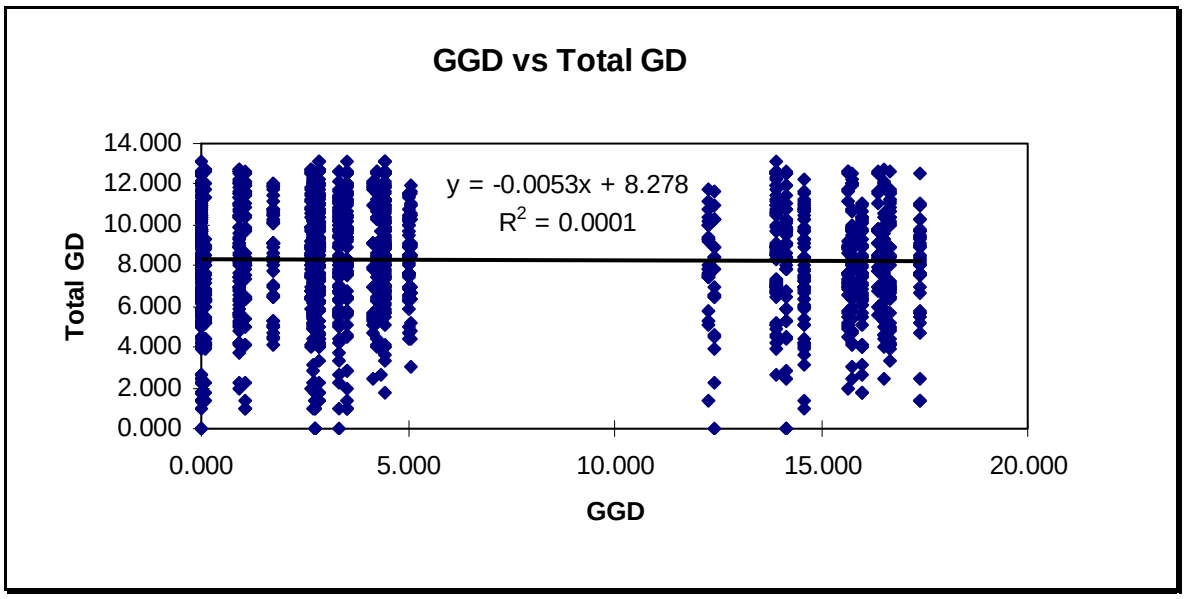

Figure 5.2 Mantel Test for association.
Total genetic (GD) and geographical (GGD) distance between *S. diclinis* individuals.

Figure 5.3 shows no evidence for significant population structure. All autocorrelation coefficient measurements (see r) fall within permutations (U) and bootstrapping intervals (bars).

5.3.5 Intraspecific Global Spatial Autocorrelation

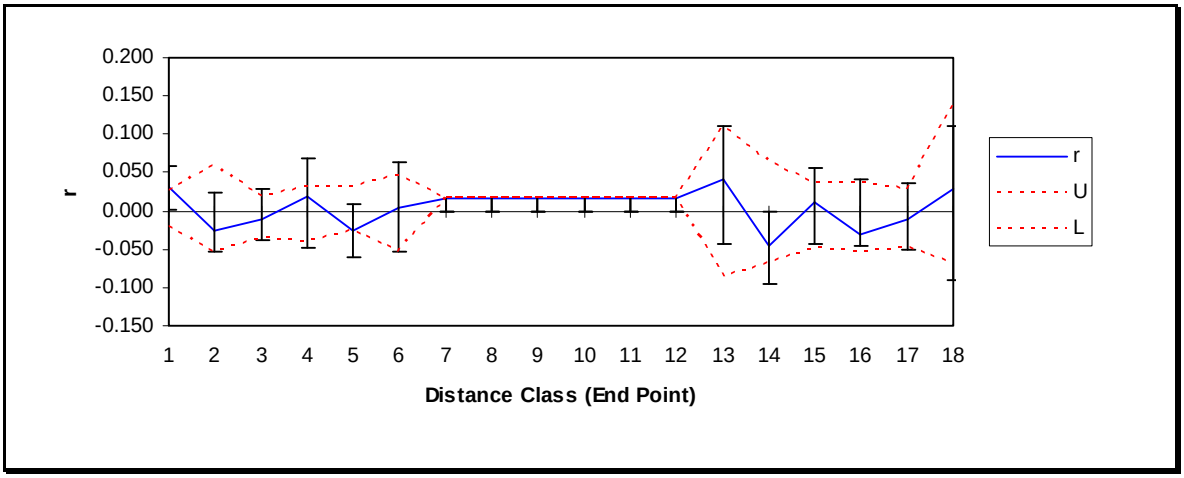

Figure 5.3. Global Spatial Autocorrelation results.

5.3.6 Intraspecific Analysis of Molecular Variance

The AMOVA tests (Figure 5.4 and Table 5.4) show that most variation is found within populations regardless of the number of regions allocated, and this result is significant in both analyses. There is a significant proportion of the variation (5%) among populations however, when a two region (east/west, see above) model is assumed. This significance disappears when a 5 region model is adopted, although the proportions of variation differ only slightly.

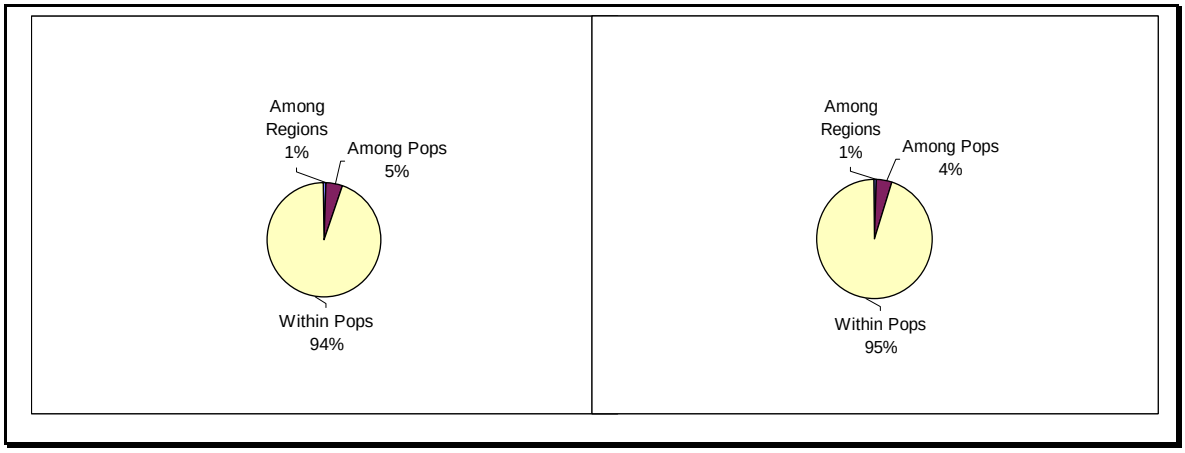

Figure 5.4. AMOVA results showing partition of molecular variation in *S. diclinis*.
Significance calculated from permutation *0.05, **0.05<0.01, ***<0.001.

Table 5.4. AMOVA Summary Statistics

Source	df	SS	MS	Est. Var.	%
Among Regions	1	5.817	5.817	0.036	1%
Among Pops	6	32.916	5.486	0.197	5%
Within Pops	54	212.742	3.940	3.940	94%
Total	61	251.476		4.173	100%

Stat	Value	P(rand >= data)
PhiRT	0.009	0.272
PhiPR	0.048	0.028
PhiPT	0.056	0.030

5.3.7 Intraspecific Bayesian Admixture Analysis

Structure analysis (Table 5.5) shows lack of population structure. Program was found to be consistent after three runs with 500000 steps following a 100000 step burn-in (Figure 5.5).

Table 5.5. Structure analysis average posterior probabilities for *S. diclinis*.

K	Average Ln Pr(X|K)	Pr(K|X)
1	-70.60	~1
2	-72.43	~0
3	-74.40	~0
4	-77.97	~0
5	-82.83	~0
6	-87.67	~0
7	-90.67	~0
8	-94.03	~0
9	-97.30	~0
10	-99.47	~0

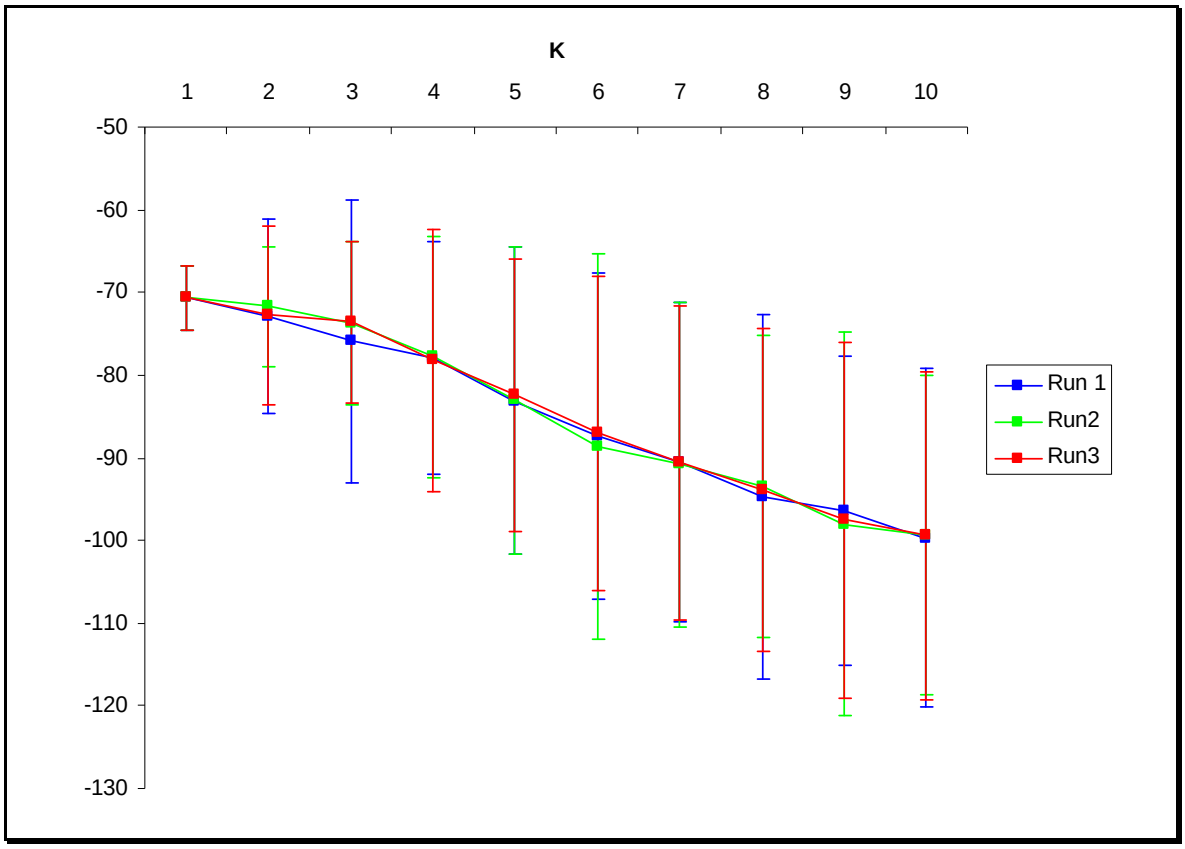

Figure 5.5. Structure analysis likelihood scores for *S. diclinis*.

5.3.8 Interspecific Diversity Analysis

Table 5.6. Nucleotide diversity (π) for *S. diclinis*, *S. latifolia* and *S. dioica*.

Gene	Species	π all	π R	π S
C110	*S. diclinis*	0.0082	0.0028	0.0286
	S. latifolia	0.0062	0.0007	0.0000
	S. dioica	0.0064	0.0000	0.0000
C109	*S. diclinis*	0.0049	0.0021	0.0039
	S. latifolia	0.0136	0.0040	0.0030
	S. dioica	0.0074	0.0000	0.0024
C37	*S. diclinis*	0.0000	0.0000	0.0000
	S. latifolia	0.0124	0.0009	0.0005
	S. dioica	0.0079	0.0011	0.0000
C34	*S. diclinis*	0.0010	0.0003	0.0000
	S. latifolia	0.0034	0.0000	0.0030
	S. dioica	0.0038	0.0009	0.0021
C1A11	*S. diclinis*	0.0022	0.0006	0.0079
	S. latifolia	0.0157	0.0000	0.0047
	S. dioica	0.0168	0.0086	0.0020
C1G11	*S. diclinis*	0.0018	0.0000	0.0000
	S. latifolia	0.0046	0.0050	0.0000
	S. dioica	0.0041	0.0045	0.0000

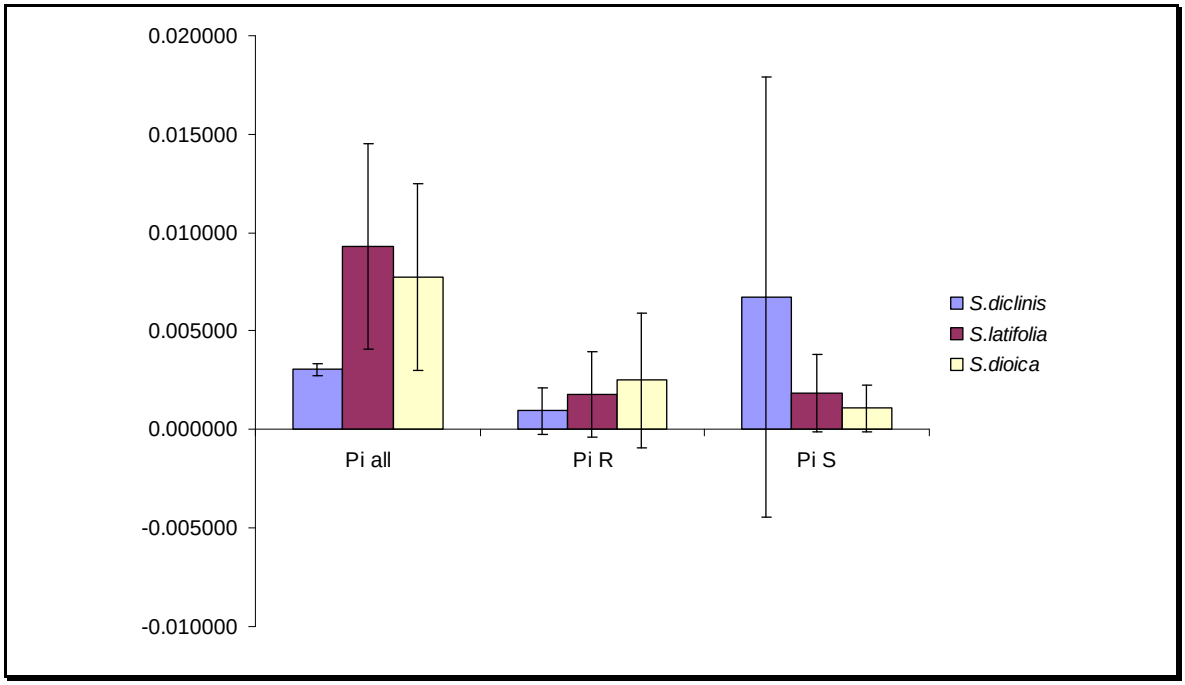

Figure 5.6. Average Nucleotide Diversity for *S. diclinis*, *S. latifolia* and *S. dioica*.
Error bars denote Standard deviation

The most striking result is the much higher silent site diversity and much lower total and replacement site diversity seen in *S. diclinis*. These differences are not significant when tested with a two tailed t-test Figure 5.6 and Table 5.7).

Table 5.7. Results of the T-test for matched pairs comparing *S. diclinis* diversity with *S. latifolia* and *S. dioica*.

	S. diclinis/*S. latifolia*	*S. diclinis*/*S. dioica*
π All	2.485 P >0.5 (NS, 5df)	1.997 P >0.5 (NS, 5df)
π Replacement Sites (R)	0.807 P >0.5 (NS, 5df)	0.934 P >0.5 (NS, 5df)
π *Silene* Sites (S)	1.011 P >0.5 (NS, 5df)	1.198 P >0.5 (NS, 5df)

5.3.9 Interspecific divergence and differentiation

As shown in Table 5.8, there are generally fewer shared differences between *S. diclinis* and either of the other two species than between *S. latifolia* and *S. dioica*, and there are consistently more fixed differences. There are also more Poly1 Mono2 (Polymorphic in species 1, monomorphic in species 2) and less Poly2 Mono1 sites reflecting the general lack of diversity in *S. diclinis*. *S. diclinis* also has slightly more shared sites with *S. dioica* than *S. latifolia*

Table 5.8. Shared, fixed and polymorphic sites between *S. diclinis*, *S. latifolia* and *S. dioica*.

	Fixed	Poly1 Mono2	Poly2 Mono1	Shared
C109				
Lat/Dic	0	14	1	4
Dio/Dic	0	10	1	4
Lat/Dio	0	8	4	10
C110				
Lat/Dic	1	9	2	3
Dio/Dic	0	7	1	4
Lat/Dio	0	5	4	7
C37				
Lat/Dic	3	16	0	0
Dio/Dic	3	9	0	0
Lat/Dio	0	14	7	2
C34				
Lat/Dic	0	4	4	0
Dio/Dic	0	2	4	0
Lat/Dio	0	4	2	0
C1G11				
Lat/Dic	3	28	10	0
Dio/Dic	4	26	10	0
Lat/Dio	0	24	20	4
C1A11				
Lat/Dic	1	3	6	1
Dio/Dic	1	11	5	2
Lat/Dio	0	2	11	2
Total				
Lat/Dic	8	74	23	8
Dio/Dic	8	65	21	10
Lat/Dio	0	57	48	25

Figure 5.7 shows the average F_{ST} values between *S. diclinis*, *S. latifolia* and *S. dioica*. Once again, the differences between these means were tested for significance using a t-test. Results showing that *S. dioica* is significantly more differentiated from *S. diclinis* than it is from *S. latifolia* can be seen in Table 5.9.

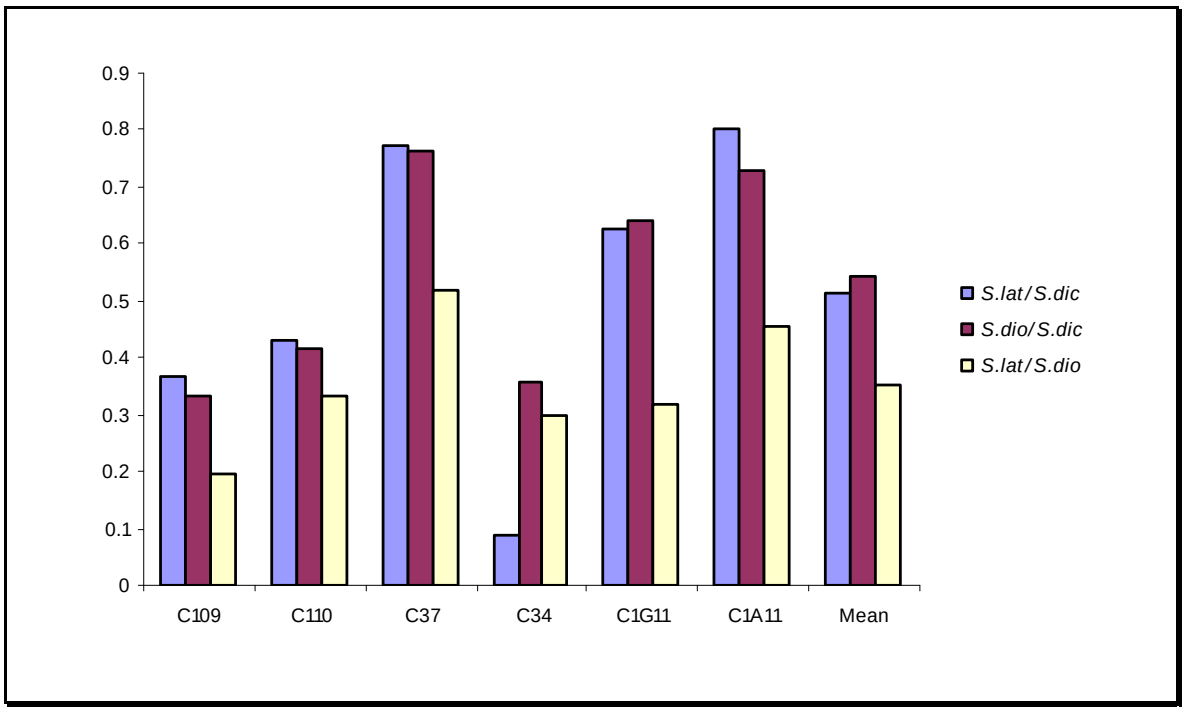

Figure 5.7. Fst between *S. diclinis*, *S. latifolia* and *S. dioica*.

Figure 5.8 shows the Dxy and Da estimates of divergence between the three species. On average, there is higher divergence of *S. diclinis* from *S. latifolia* than *S. diclinis* from *S. dioica*, although these results are not significant when tested using a T-test for matched pairs (see Table 5.9).

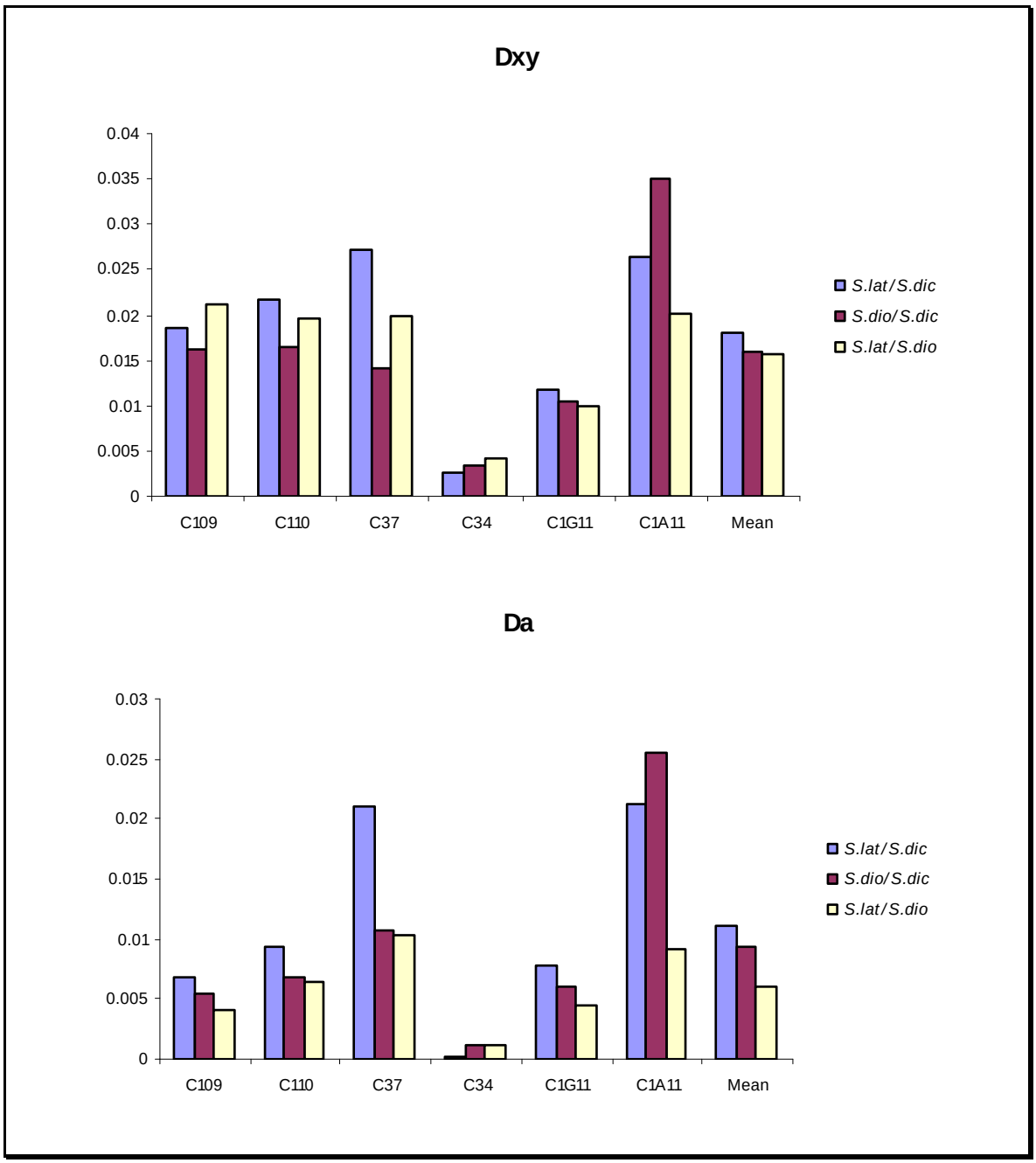

Figure 5.8. Divergence between *S. diclinis*, *S. latifolia* and *S. dioica*.

Table 5.9. Results of T-tests for divergence and differentiation between *S. diclinis*, *S. latifolia* and *S. dioica*.
S. diclinis=S. dic, S. latifolia=S. lat, S. dioica=S. dio

	S. dic / *S. lat* vs. *S. dic* / *S. dio*	*S. lat* / *S. dic* vs. *S. lat* / *S. dio*	*S. dio* / *S. dic* vs. *S. dio* / *S. lat*
D_a	0.736 *P* >0.5 (5df.)	1.403 *P* >0.5 (5df.)	0.053 *P* >0.5 (5df.)
D_{xy}	0.889 *P* >0.5 (5df.)	2.439 *P* >0.5 (5df.)	1.288 *P* >0.5 (5df.)
F_{ST}	0.518 *P* >0.5 (5df.)	1.965 *P* >0.5 (5df.)	4.306 *P >0.01* (5df.)

5.3.10 Interspecific Bayesian Admixture Analysis

The Structure analysis for the combined *S. diclinis*, *S. latifolia* and *S. dioica* dataset produced consistent results after three runs of chain length 500000 following a 100,000 burn-in (see Figure 5.9). As expected the best likelihood scores were for K=3 (Table 5.10), with variance increasing with the declining likelihoods thereafter.

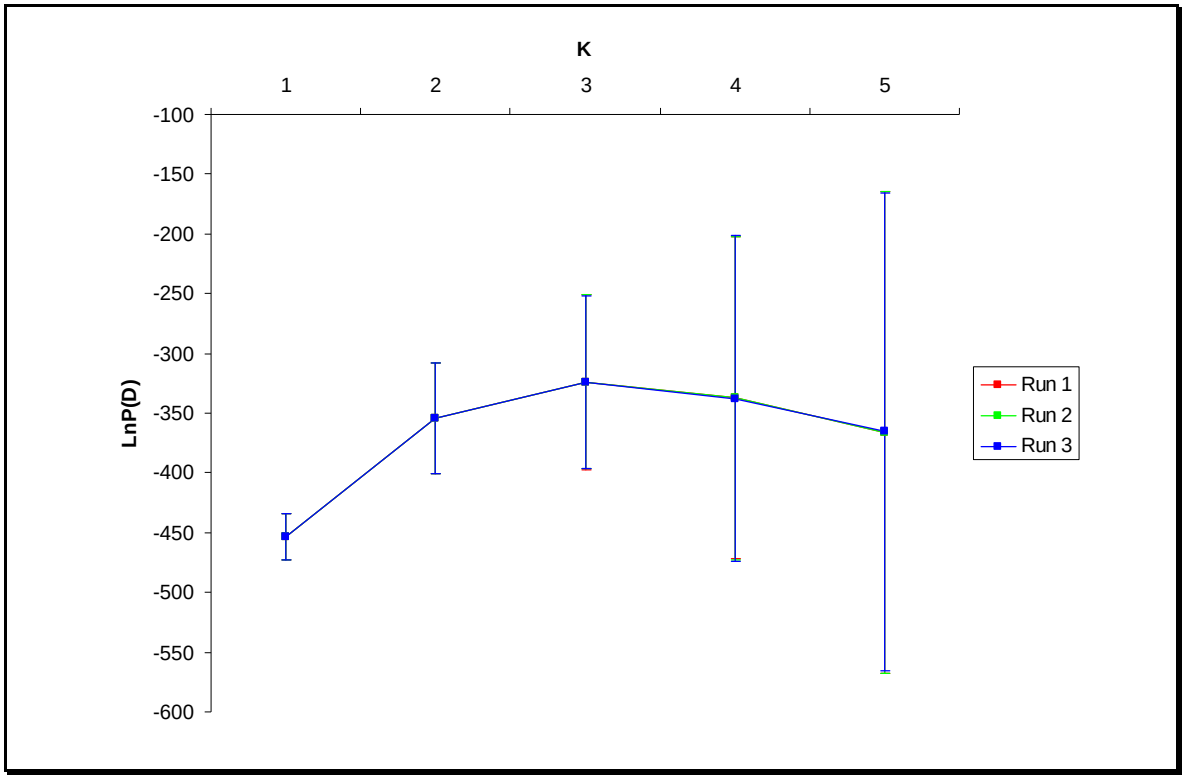

Figure 5.9. Structure analysis likelihood scores for *S. latifolia*, *S. dioica* and *S. diclinis*.

Table 5.10. Structure analysis average posterior probabilities for *S. latifolia*, *S. dioica* and *S. diclinis*.

K	Average Ln Pr(X|K)	Pr(K|X)
1	-453.70	~0
2	-354.20	~0
3	-324.00	~1
4	-337.43	~0
5	-365.93	~0

The three species fit almost exclusively into their three respective clusters, although there were some outliers (see Figure 5.10).

S. diclinis individual 69 fitted better into cluster 1 than cluster 2 with the rest of the *S. diclinis* samples. *S. latifolia* individuals IL11, IL19 and IL139 fitted better into cluster 2 than into cluster 1, and individual IL33 fitted better into cluster 3 than into cluster 1. *S. dioica* individual IL60 fitted better into cluster 1 than cluster 3, and was also one of the only *S. dioica* individuals to show any appreciable membership to cluster 2. Excluding those individuals with more than 50% missing data, only the *S. diclinis* individual and *S. latifolia* IL11 and IL19 individuals are likely outliers. The *S. latifolia* individuals are from Spain and France respectively. Cluster 1 therefore, fairly accurately represents *S. latifolia*, cluster 2 *S. diclinis,* and cluster 3 *S. dioica*.

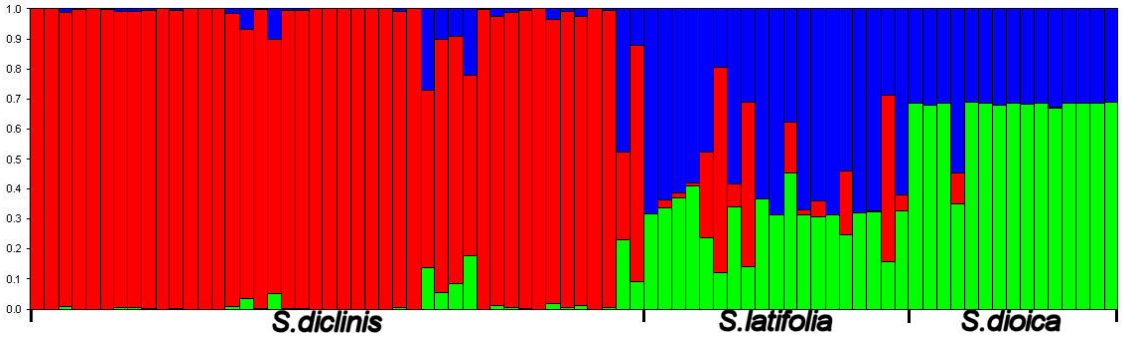

Figure 5.10. Structure analysis histogram for *S. latifolia*, *S. dioica* and *S. diclinis*.
Proportion membership of individuals to K=3 clusters.
Cluster 1= Blue, Cluster 2 = Red, Cluster 3= Green.

Table 5.11. Structure analysis proportion membership to alternative clusters.

	Proportion membership	Pairwise Average
S. diclinis in S. latifolia cluster	0.143	0.0905
S. latifolia in *S. diclinis* cluster	0.038	
S. diclinis in S. dioica cluster	0.007	0.015
S. dioica in S. diclinis cluster	0.023	
S. latifolia in S. dioica cluster	0.331	0.316
S. dioica in S. latifolia cluster	0.301	

The proportions of membership to other population clusters suggest that *S. dioica* and *S. latifolia* have the highest shared ancestry with ~30% shared cluster membership. *S. diclinis* and *S. latifolia* appear to have more shared ancestry than *S. diclinis* and *S. dioica* with ~9% and ~1% shared ancestry on average respectively (Table 5.11).

5.3.11 Interspecific Phylogenetic Analysis

The tree for locus C109 was generally not well supported by the bootstrap analysis (Figure 5.11). The only well supported branches were those grouping *S. dioica* individuals IL98 and IL42 (labelled 1) which are from Austria and

Belgium respectively, *S. marizii* and *S. latifolia* individual IL11 (2) from Portugal and Spain respectively, and *S. latifolia* IL107D and 107B which are individuals from the same line in Germany. *S. diclinis* is not in a well supported clade, but generally clusters together, whereas *S. latifolia* and *S. dioica* tend to be in mixed clusters.

C110 also produced a tree that is not well supported for the most part (Figure 5.12). The only branch with more than a 50% bootstrap score supports a clade containing exclusively *S. latifolia* individuals, namely IL28, IL113, IL19 and IL11. IL28, IL113 and IL19 are all French accessions, and IL11 is Spanish. Once again *S. diclinis* tends to cluster together, although it is also interspersed with *S. dioica* individuals.

The tree produced for locus C37 (Figure 5.13) separates the three species much more discretely, although the only well supported branch (95%) clusters all of the *S. diclinis* individuals which have no diversity at this locus. *S. latifolia* and *S. dioica* are also placed roughly into separate clades, with *S. diclinis* being placed within the mainly *S. dioica* clade.

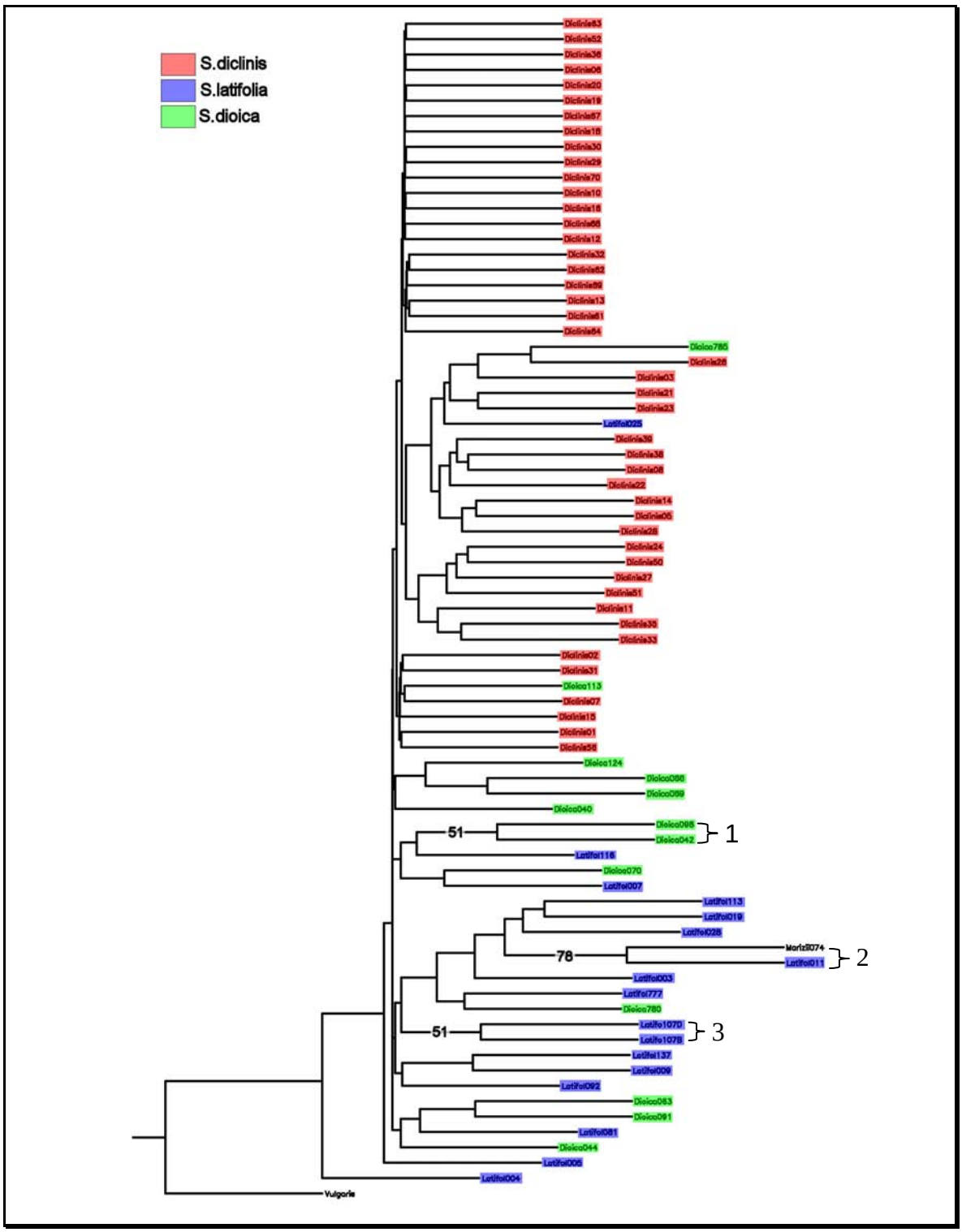

Figure 5.11. Majority Rule Extended Maximum Likelihood Tree for locus C109.

Bootstrapping values in over 50% of 1000 trees are shown on their respective branches. Well supported clades are numbered for identification.

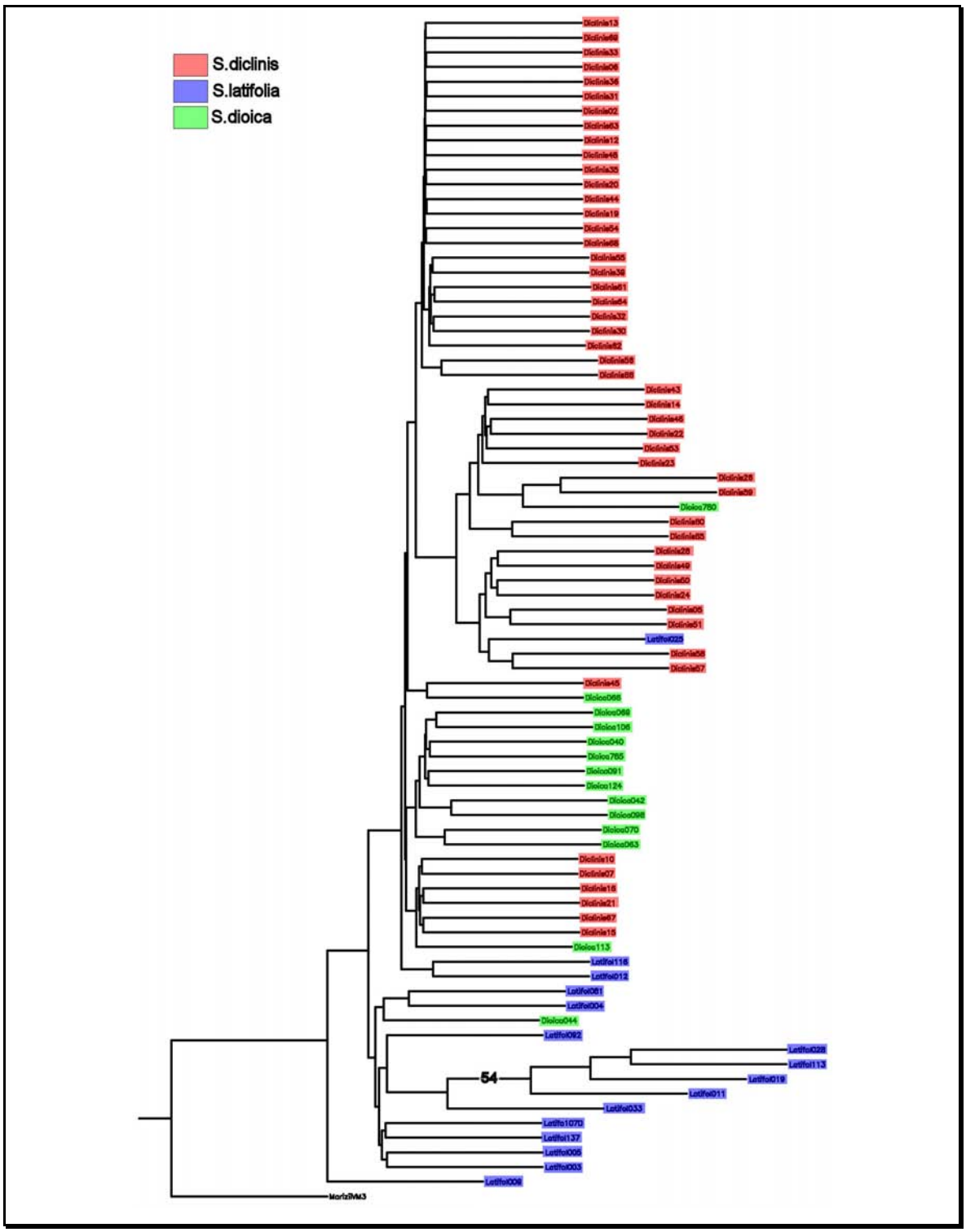

Figure 5.12. Majority Rule Extended Maximum Likelihood Tree for locus C110.

Bootstrapping values in over 50% of 1000 trees are shown on their respective branches.

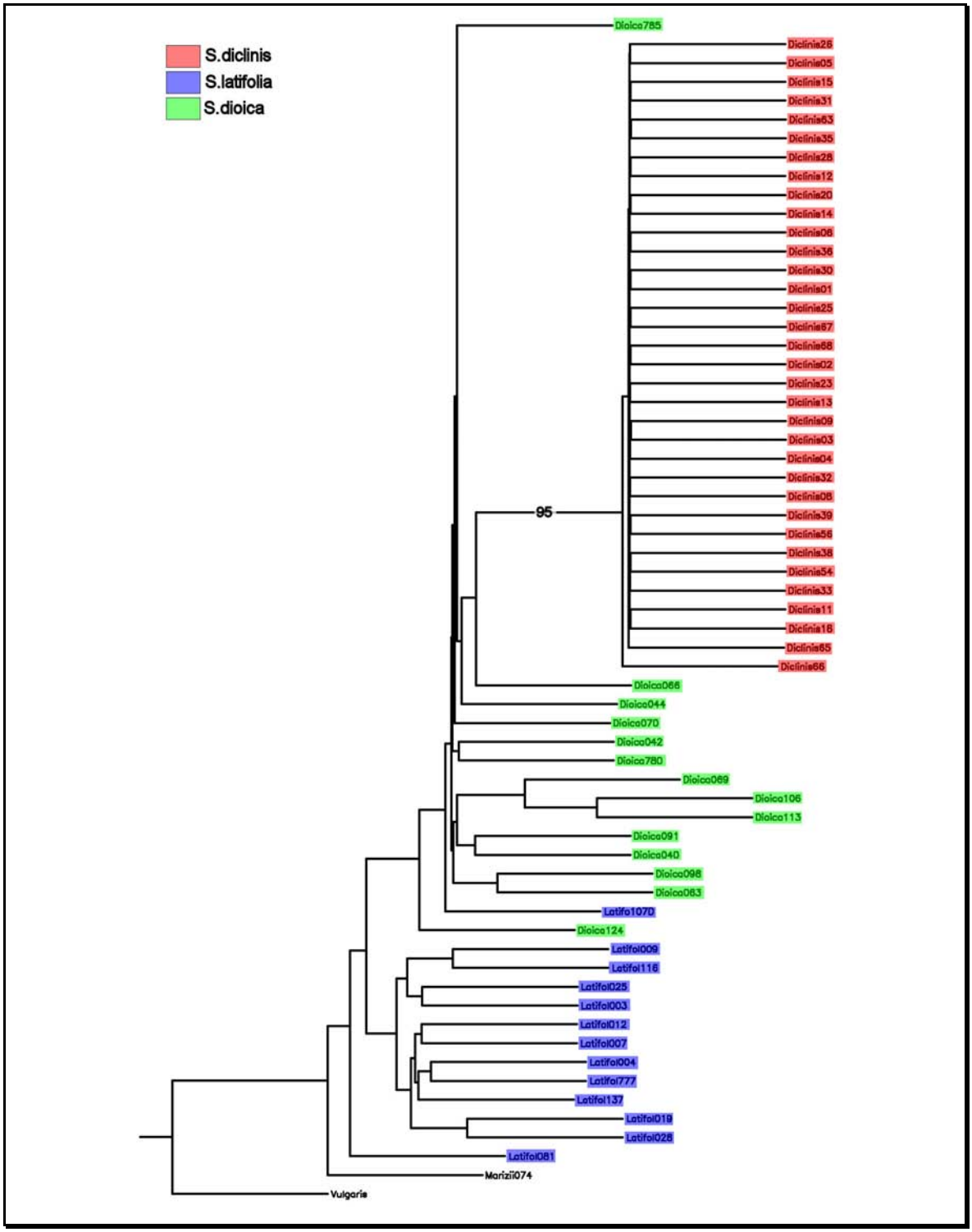

Figure 5.13. Majority Rule Extended Maximum Likelihood Tree for locus C37.

Bootstrapping values in over 50% of 1000 trees are shown on their respective branches.

The maximum likelihood tree for locus C34 is extremely mixed (Figure 5.14). The three species and the *S. marizii* sample are placed into a single mixed clade with little species clustering. There are three well-supported branches, one (labelled 1) clustering *S. diclinis* individuals 1, 2, 7 from population 1 and individual 62 from population 6. The second (2) clusters *S. dioica* individual IL42 from Belgium, and individuals IL113 and IL106 from France. The last well supported branch (3) also clusters *S. dioica* individuals IL63, IL70, IL66, IL780, IL785, IL40, IL91 and IL98. These individuals are from the UK, Sweden, Belgium and Austria.

Locus C1G11 produces a tree with several well-supported branches (Figure 5.15). The three species are well defined in this tree, with four well supported clusters. The first (labelled 1) incorporates all of the *S. diclinis* individuals. Cluster 2 incorporates all the *S. latifolia* individuals apart from the Spanish IL11 individual, and clusters 3 and 4 gather together most of the *S. dioica* individuals. One branch in particular (4) is particularly well supported (99.5%), and the *S. dioica* individuals clustered together (IL70, IL106 and IL40) are from Sweden, France and Belgium respectively.

The final tree, for locus C1A11 (Figure 5.16), has two fairly well supported branches, the first (labeled 1) clusters together most of the *S. dioica* individuals together with the *S. marizii* individual, and the other (2) clusters all but two *S. diclinis* individuals.

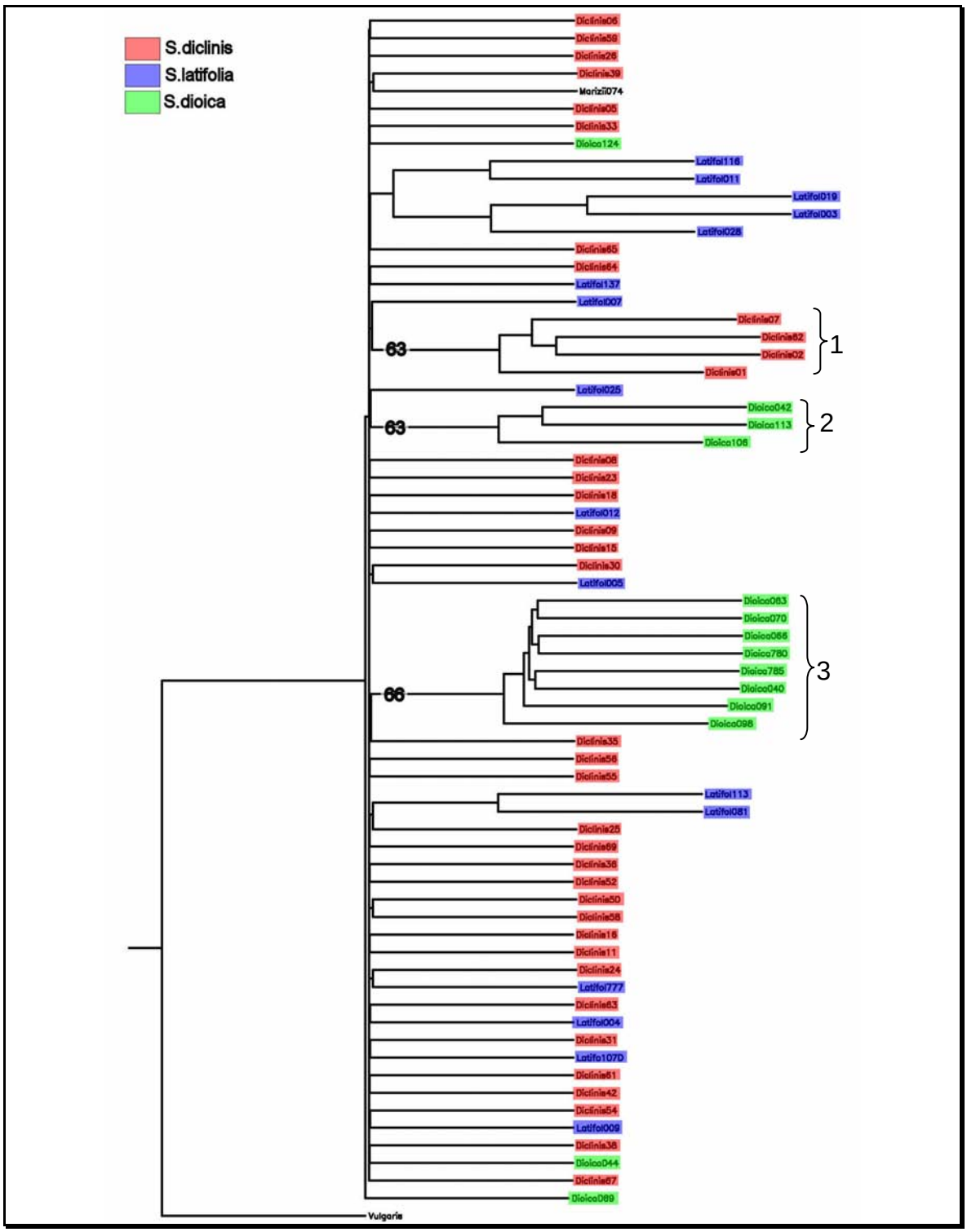

Figure 5.14. Majority Rule Extended Maximum Likelihood Tree for locus C34.

Bootstrapping values in over 50% of 1000 trees are shown on their respective branches. Well supported clades are numbered for identification.

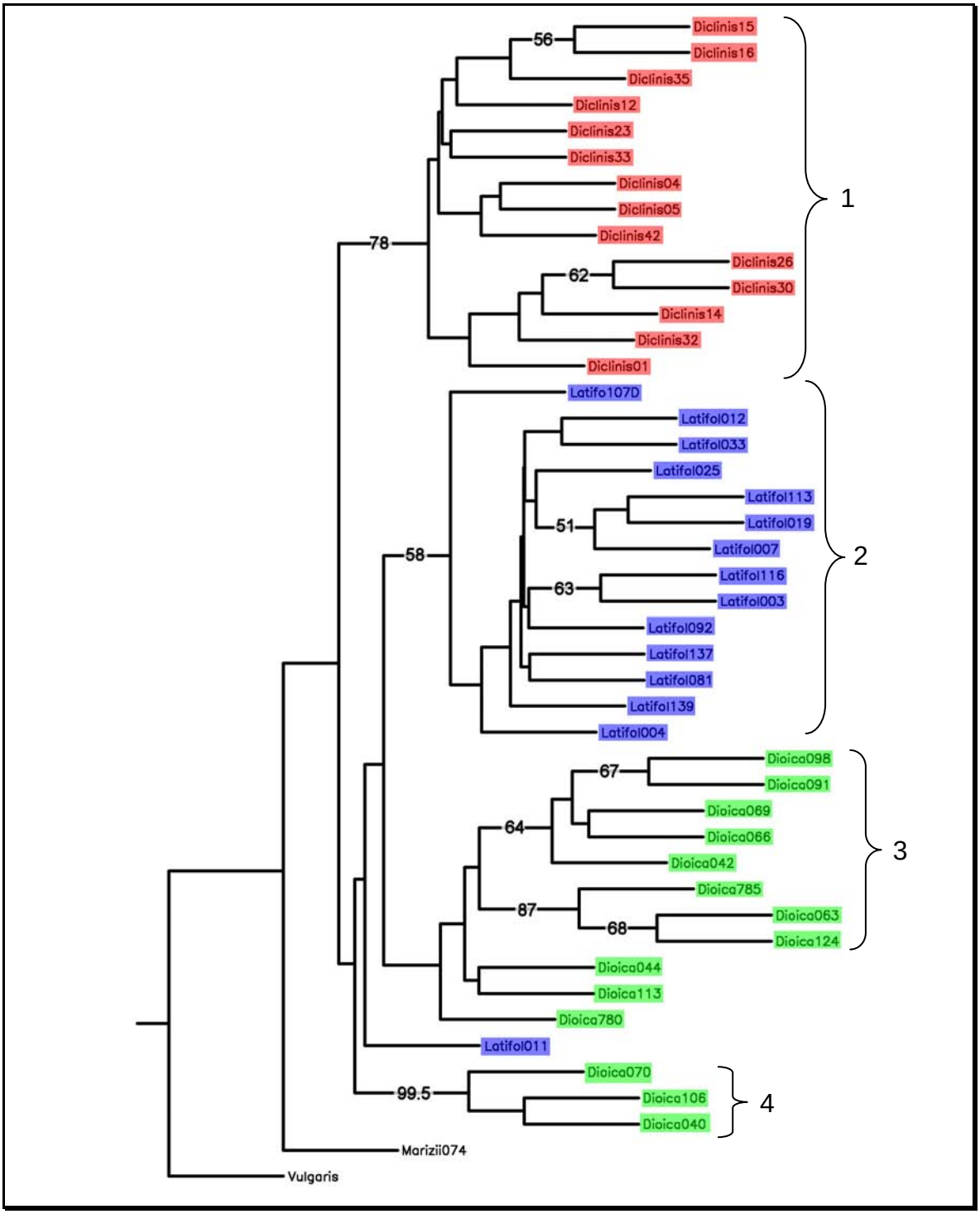

Figure 5.15. Majority Rule Extended Maximum Likelihood Tree for locus C1G11.

Bootstrapping values in over 50% of 1000 trees are shown on their respective branches. Well supported clades are numbered for identification.

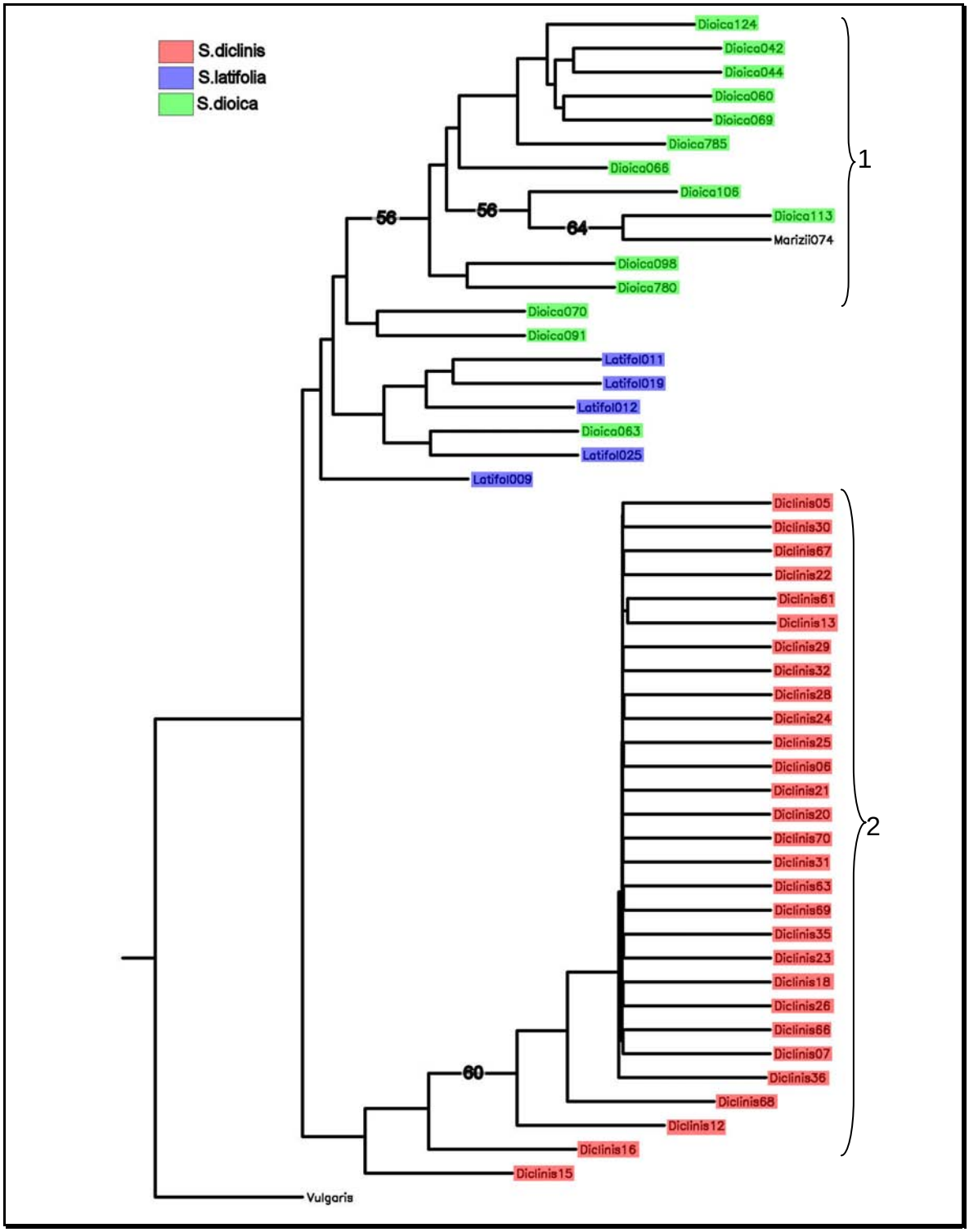

Figure 5.16. Majority Rule Extended Maximum Likelihood Tree for locus C1A11.

Bootstrapping values in over 50% of 1000 trees are shown on their respective branches. Well supported clades are numbered for identification.

5.3.12 WH isolation modeling

The probability of *S. diclinis* and *S. latifolia* and *S. dioica* evolving under an isolation model is very high (see Table 5.12) suggesting that no appreciable gene flow has occurred between *S. diclinis* and either of the other two species since their split.

Table 5.12. Summarized results from the WH isolation model fitting program.
95% Confidence intervals produced by 10,000 simulations are provided in brackets.

Species 1/ Species2	Θ Species 1	Θ Species 2	Θ Ancestral	T	*P*wwh	PX^2
S. latifolia/ *S. diclinis*	18.279 (0.529-198.306)	4.711 (2.620-11.728)	33.602 (15.691-76.972)	0.2340 (0.0029-0.525)	0.9490	0.9627
S. dioica/ *S. diclinis*	11.144 (0.390-151.923)	3.540 (0.341-13.286)	36.531 (8.620-75.700)	0.261 (0.041-0.640)	0.3834	0.5829

The theta values produced from the modeling suggest that the ancestral population size was larger than *S. diclinis*, *S. latifolia* and *S. dioica*, but reflects the much lower effective population size of *S. diclinis* than *S. latifolia* or *S. dioica*. The estimates of the time since divergence (in 2N generations) T, suggest that *S. diclinis* diverged from *S. latifolia* and *S. dioica* at approximately the same time (936,000 and 864,000 years ago respectively assuming *S. latifolia* and *S. dioica* have an effective population size of ~1million and a two year generation time)

5.3.13 Bottleneck Analysis

Table 5.13. Bottleneck analysis across the five polymorphic loci in *S. diclinis*.

Locus	Observed heterozygosity	Expected heterozygosity IAM	Expected heterozygosity SMM	Excess/Deficiency
C34	0.215	0.383	0.495	Deficiency
C1A11	0.333	0.595	0.702	Deficiency
C1G11	0.614	0.694	0.746	Deficiency
C110	0.858	0.878	0.918	Deficiency
C109	0.851	0.860	0.904	Deficiency
Wilcoxen		P=0.03125	P=0.03125	
Sign		P=0.01238	P=0.01119	

The bottleneck analysis shows that there was a heterozygosity deficiency for all loci which was significant (in both the Wilcoxen and Sign tests) for both the infinite allele and stepwise mutation models employed (see Table 5.13).

5.4 Discussion

One important feature of *Silene diclinis* is its spatially separated populations. It would be supposed that if there was limited gene flow between these populations they would begin to differentiate, and population structure should be detectable. In fact, small spatially separated populations with low effective size have been shown in computer simulations to lose variation within the populations more readily than between them (Lacy, 1987). This does not appear to be the case in *S. diclinis*. F_{ST} values show variation between genes but no isolation by distance in these populations was confirmed using a Mantel test and Global Spatial Autocorrelation, both of which showed no departures from neutrality.

An Analysis of Molecular Variance (AMOVA) was completed to assess where the variation was partitioned in the dataset. Either two or five regions were defined purely from the geographical clustering of the populations. Both of these scenarios produced similar results in the AMOVA with 94-95% of variation found within populations, and only 4-5% between them (although this was only significant when two regions were defined). This could be a false result if populations have been incorrectly assigned. It is possible that some of the populations are split, or that some (such as populations 5 and 8) are one single population. To eliminate this possibility, a Structure analysis was performed with

up to 10 population clusters tested. The best likelihoods were for a single population, and likelihoods declined and variances increased for more than one population.

From these tests, it appears that there is very little population structure in *S. diclinis*. There may well be sub-population structure present, as was suggested for the largest *S. diclinis* population by Prentice (1984b), that is not detectable in this dataset. This so-called Wahlund effect would depress the level of heterozygosity seen in the current populations but a more comprehensive study with as many different individuals as possible would need to be conducted to test this hypothesis.

S. diclinis was expected to have low levels of variation, as small populations are governed by drift which reduces levels of variation (Barrett & Kohn, 1991). The diversity in *S. diclinis* is in general lower than in *S. latifolia* and *S. dioica* and in fact C37 has no polymorphic sites at all. C110 (and to a lesser extent C109) has higher diversity than the other loci, and this is due to an increase in silent site polymorphisms. These two loci also exhibit a positive Tajima's D value, unlike the other loci, although it is important to note that all of these values are non-significant. When *S. diclinis* is compared with *S. latifolia* and *S. dioica*, the increase in silent site diversity in C110 is ten-fold higher in *S. diclinis* than the other two species.

Inbreeding is the most likely cause of reduction of variation in *S. diclinis* as selection should be weak due to the low effective population size. One of the few rescues for a population with inbreeding depression is migration of alleles from other large sources. The only candidates for this would be *S. latifolia* and *S. dioica*. *S. diclinis* seems to be equally diverged from both of them, but when Structure was run for the three species, *S. latifolia* shared around six times more proportion membership to the *S. diclinis* cluster (and vice versa) than *S. dioica* did. In fact after taking into account the individuals with over 50% missing data, two *S. latifolia* individuals are better placed within the *S. diclinis* cluster, and one *S. diclinis* individual into the *S. latifolia* cluster.

The phylogenetic analysis revealed that only some loci appear to be responsible for this effect more than others. C37, C1G11 and C1A11 all cluster well into distinct species clades, but C110, C109 and C34 show mixed clustering and poor branch support. These tests could be an indication of previous gene flow between *S. diclinis* and the other species, or it could just be due to shared ancestral polymorphism. The WH isolation model could not be rejected by the dataset, however, which suggests that recent gene flow is unlikely.

Finally, the Bottleneck analysis was used to see if there was a heterozygosity excess indicative of a recent population decline. The results showed that all loci were actually exhibiting a deficiency in heterozygosity. This is usually due to a recent population expansion or inbreeding. As the Tajima's D tests are not

consistently positive and no recent expansion is known of, it is more likely that inbreeding has reduced the levels of heterozygosity, but also suggests that the population decline was not recent. *S. diclinis* may have been historically low in numbers or alternatively, may have been through many expansions and declines in its history, thus confusing the results.

If inbreeding is the cause of this loss of heterozygosity but there has been no recent population bottleneck, could it be possible that *S. diclinis* is becoming equipped to deal with inbreeding? The reciprocal translocation of the Y and an autosome that has produced the neo-sex chromosomes in *S. diclinis* (Howell *et al*., in press) may be linked to inbreeding (Charlesworth & Wall, 1999). Inbreeding creates associations between heterozygosities at different loci and a translocation with a Y chromosome would help to retain the heterozygosity by reducing the rate of recombination around the translocation. Y-autosome translocations are particularly likely to become fixed.

This theory relies on there being selection in favour of heterozygotes (balancing selection) and there may be evidence for balancing selection in this *S. diclinis* dataset. One locus in particular stands out from the others, C110 (which is most likely the alpha subunit of the TFIIF transcription factor), exhibits very high silent site diversity (about ten times higher than *S. latifolia* and *S. dioica*) accompanied by a positive Tajima's D value. This points to possible balancing selection, although the Tajima's value is not significant.

Although difficult to prove definitively, the results from this study suggest that *S. diclinis* may have suffered low population numbers for some time, and most loci have suffered the expected loss of diversity. Structure may be present on the sub-population scale, conserving variation within populations, but creating a Wahlund effect of lowered observed heterozygosity at the population level. Most interestingly, *S. diclinis* may be coping with the inbreeding that has been forced upon it, with balancing selection acting to promote heterozygotes in certain loci, and inbreeding creating associations between heterozygosities that are subsequently fixed in the population through sex-chromosome translocations. In this respect, the greatest danger to the survival of *S. diclinis* is unlikely to be due to inbreeding depression and genetic erosion, but rather the threat of habitat loss from changes in agricultural practice.

6. CONCLUSIONS

Combined with previous research on the species of *Silene* in the section *Elisanthe* and climatic and anthropological data, we can speculate about the possible contributing factors involved in its evolutionary history. We can trace the beginnings of this back to the evolution of a sex chromosomal system in a dioecious ancestor around 10 million years ago (Filatov & Charlesworth, 2002; Bergero *et al.*, 2007). From isolation modeling results in this study, the effective population size of this ancestor appears to be large and is possibly larger than any of the current species. We can speculate that it was also widely distributed because, during the subsequent ice age, the ancestor must have inhabited several glacial refugia following the onset of global cooling during the late Pliocene.

6.1 Historical Range Expansions

During this study, *S. diclinis* was estimated to have split from the ancestor of *S. latifolia* and *S. dioica* around 900,000 years ago with *S. latifolia* and *S. dioica* splitting some 400,000 years later, consistent with the cold period during the Pleistocene and the growth of the ice sheets. In the case of *S. latifolia*, studies suggest that it emerged from refugia in Southern Europe, possibly the Balkan or Iberian Peninsulas (Taylor & Keller, 2007), whereas *S. dioica* probably spread

from several refugia around the Mediterranean, Balkans or Caucasus (Prentice *et al.*, 2008).

The present warm interglacial period began around 18,000 years ago, causing the ice sheets to recede. There were still climatic fluctuations with warmer and cooler periods causing brief range expansions and subsequent contractions until the climate stabilized approximately 8-10,000 years later (Hewitt, 1996). At this point significant range extensions could be made by plants residing in glacial refugia, although in Europe the advances would have been slowed by the barriers of mountain ranges and the Mediterranean Sea. Deciduous forests, which were probably limited to the Balkans, Calabria, and the Caucasus, rapidly spread northwards when the ice began to retreat. This allowed woodland flora with relatively slow dispersal rates such as *S. dioica* to follow the spread northwards into already established forest (Prentice *et al*., 2008). 6,000 years ago the vegetation distribution had become similar to what we see today (Hewitt, 1996).

Following the natural colonisation by flora, farming by man would also have begun its spread into Europe, reaching northern Europe around 5,000 years ago (see Figure 3.2, Sokal *et al*., 1991), bringing with it more new species that perhaps could not overcome the barriers of mountain ranges and sea without human intervention, or had become adapted to the new niches created by man. Species such as *S. latifolia* and *S. diclinis* which favour disturbed ground

probably expanded their ranges by following the spread of agriculture in this way (Vellekoop *et al.*, 1996).

In the case of *S. latifolia* and *S. dioica*, their range expansions were extensive, permitting secondary contact when *S. latifolia*, spreading with farming, came close to the forests where *S. dioica* had become established. This suggests that those populations in southern Europe would probably have come into contact several thousand years before those in the north, and introgression between them would have been possible from 6,000-8,000 years ago.

S. diclinis was probably not as successful as *S. latifolia* and *S. dioica* in its expansion following recession of the ice. Although the possible locations of its glacial refugia have not been speculated, the lack of surviving pockets of the species outside of Spain suggests that its range expansion was not extensive.

There are several possible reasons why *S. diclinis* did not expand its range as fully as its two close relations. It has a similar requirement for disturbed ground like *S. latifolia*, and is often associated with the cultivated lands around Xàtiva. This would suggest that like *S. latifolia*, it would have expanded its range along with the spread of agriculture as *S. latifolia* did. *S. latifolia* may have had several advantages over *S. diclinis* in its ability to expand its range, however. Firstly, *S. latifolia* was likely to have been expanding from Iberia or the Balkans (Taylor & Keller, 2007). Similarly, agriculture probably originated in Asia and spread to the

rest of Europe via Iberia and the Balkans (see Figure 3.2), allowing *S. latifolia* to ride the wave of farming spreading across Europe. Spain, however, was the last stop of one of the advancing paths of agriculture. If *S. diclinis* had a refuge close to its current distribution in Spain it may have spread more slowly than *S. latifolia*, and over less distance as it would also have been more likely to have found its preferred niches already inhabited by other possibly more invasive species.

S. diclinis may also have been hampered by the fact that its bumblebee pollinators prefer to forage over short distances (Osbourne *et al*., 2008), preventing pollination of any plants colonizing sites too far away from established populations. *S. latifolia* is more likely to have been able to overcome this problem thanks to its wide-ranging moth pollinators. The Pyrenees, which have been identified as a suture zone in other species (Taberlet *et al.*, 1998), may have acted as a barrier, preventing *S. diclinis* from escaping Spain after the onset of the current interglacial.

6.2 The Present Day Species

Today *S. latifolia* and *S. dioica* are reunited and thriving across Eurasia, whereas its sister species *S. diclinis* is trapped in a small area of Spain, and is limited by low numbers and a specific niche. Currently it seems that, although hybridization between *S. latifolia* and *S. dioica* is able to occur at hybrid zones,

the effects of this have not yet seeped into the genome of either species. They show little evidence for gene flow following the species divergence until their relatively recent secondary contact, fitting isolation models extremely well and exhibiting species-specific selective sweeps. During their isolation there is also evidence that they have become pre-zygotically reproductively isolated with different habitat preferences, pollinators, flower phenotype and, even more importantly, reduced hybrid fitness, as shown by the lack of intermediate hybrids at hybrid zones, which may suggest the two species have become post-zygotically reproductively isolated (Hess *et al*., 1972; Prentice, 1988; Jürgens *et al.*, 1996; Minder *et al*., 2007; Waelti *et al*., 2008).

It is unlikely that *S. diclinis* has been able to hybridize with either *S. latifolia* or *S. dioica* since it split from them. Isolation models between *S. diclinis* and the other two species cannot be rejected, and no known incidences of natural hybrids are known. Unfortunately, hybridization of *S. diclinis* with other species could have enabled it to avoid the inbreeding depression caused by its low numbers (Lacy, 1987). The lack of evidence for a recent bottleneck in this species suggests that it had been historically low in numbers, possibly since the climate stabilized 6000 years ago. The species may have expanded and contracted numerous times with the previous climate oscillations, but each contraction and expansion would have destroyed the evidence of any preceding ones. Despite its isolation and small, subdivided population, *S. diclinis* seems to have been coping with low

population size for some time and now shows possible evidence for balancing selection and spread of heterozygosity-fixing sex chromosome rearrangements.

6.3 Future Prospects

The immediate future of *S. latifolia* and *S. dioica* appears to be solid. Both have large enough population sizes and ranges to cope with all but the most catastrophic ecological disaster. Eventually it is likely that the two species will cease to hybridize in the wild as they have clearly become reproductively isolated in the time that they have been separated, and this reproductive isolation is only likely to become strengthened as more time passes, possibly becoming postzygotic.

S. diclinis has a more uncertain future. Although it appears to have adapted to cope in some part with the genetic effects of its low effective population size, the fact remains that it will struggle to adapt to any environmental change, and as such is completely dependent on its current habitat remaining stable. Its fate therefore, is reliant on conservation measures to preserve its habitat for the immediate future. Beyond this, there is little that could be done for the current natural population of *S. diclinis* should the climate change again for instance. If its current habitat were to become inhospitable, only reintroduction of specimens from seed banks to other suitable habitats would be an option to preserve the species in the wild.

APPENDICES

Appendix 1 - Primers

KASPar Genotyping primers

Primer Name	Sequence	Specificity
SlY4 ad1_C1	GCCATGGGCATCTGTTGCACAATTT	*SlY4* *S. latifolia & S. dioica*
SlY4 ad1_C2	TAAGCTGTCGTTGTCATGTGGCCAT	*SlY4* *S. latifolia & S. dioica*
SlY4 ad1_ALC	GAAGGTCGGAGTCAACGGATTCCCCA TTTGCTGTAA	*SlY4* Allele specific *S. latifolia*
SlY4 ad1_ALT	GAAGGTGACCAAGTTCATGCTAACCC CATTTCCTGT	*SlY4* Allele specific *S. dioica*
C2C4 ad2_C1	TAGCCGAAGCATACGATCCAGCAA	*C2C4* *S. latifolia & S. dioica*
C2C4 ad2_C2	TAGCTGGTCAGTAGCCGAAGCATA	*C2C4* *S. latifolia & S. dioica*
C2C4 ad2_ALT	GAAGGTGACCAAGTTCATGCTGGTCT AAGAACCCAAATGCTCCT	*C2C4* Allele specific *S. latifolia*
C2C4 ad2_ALC	GAAGGTCGGAGTCAACGGATTGGTCT AAGAACCCAAATGCTCCC	*C2C4* Allele specific *S. dioica*

Gene flow Primers

Locus and Direction	Primer Sequence
C1A11Forward	ACA GTG TTC AAT ATG TGC CAA AAT C
C1A11 Reverse	GAG CCA ATT TCA ACT TCA TAC CAG
C1E3 Forward	GGT TTT CCA TGA TAC TCG AT
C1E3 Reverse	GTG AGA GAT TGC GAA GAT G
C1E4 Forward	GCA GCA GAG ATA GAG AGG TT
C1E4 Reverse	ATG CTA TGG ACA TCC TGT TT
C1H1 Forward	TAC CGC GAA GAA GCA GTA G
C1H1 Reverse	CCC AGA CCG TTG AGT TTC
C2C4 Forward	ATC AGT CTA GTG AAT GGT AAC GGT G
C2C4 Reverse	CAT GTG CTC TCT TGA ATG GTA CTT C

Appendix 2 - IM Program Output Files

Run 1

INPUT AND STARTING INFORMATION

Command line string : -b 1000 -l 1000000 -t10 -q1 1.0 -q2 5.0 -m1 9 -m2 14
Comment at runtime :
Input filename : infile.txt
Output filename: outfile.txt
Random number seed : 1151339087
IM Model:
- each population is constant in size
- All Loci Share the Same Two Migration Rate Parameters
" - Inheritance Scalars are treated as constants, as given in input file"
- Run Duration -
" Burn period, # steps: 1000 "
" Record period, # steps: 1000000 "
- Metropolis Coupling -
None

Text from input file: Autosomes

- Population Names -
Population 1 : lat
Population 2 : dio
- Locus Information -

Locus#	Locusname	MutationRatesPerYear	samplesize1	samplesize2	Model	InheritanceScalar
0	C1d7	11	8	IS	1	
1	C1f6	11	9	IS	1	
2	C34	10	9	IS	1	
3	C79	11	10	IS	1	
4	c158	10	9	IS	1	
5	C1e3	12	10	IS	1	
6	C1h1	11	9	IS	1	
7	C1a8	11	9	IS	1	

- Maximum Parameter Values -
Max for q1 : 1.29
Max for q2 : 4.58
Max for m1 : 9.00
Max for m2 : 14.00
Max for qA : 1.29
Max for t : 10.00

RUN INFORMATION

Number of steps in chain following burnin: 1000000
Number of steps between recording: 10 Number of record steps: 90909
Number of genealogy updates per step: 1
"Time Elapsed : 0 hours, 7 minutes "

"Highest Total P(D|G, Params) (log) : -80.173 "

"Highest P(D|G, Params) (log) for each Locus "

Locus	"P(D|G, Params)"
0	-3.743
1	-6.949
2	-12.175
3	-1.087
4	-8.969
5	-5.563

6	-1.098
7	-21.12

Highest P(G|PARAMS) (log) for each Locus

Locus 0: 47.603
Locus 1: 50.624
Locus 2: 49.380
Locus 3: 49.600
Locus 4: 44.622
Locus 5: 56.069
Locus 6: 53.162
Locus 7: 43.034

Mean Time of Most Recent Common Ancestor

Locus	Mean Time	Variance
0	0.818	0.937
1	0.784	0.814
2	0.513	0.34
3	1.148	1.841
4	0.764	0.756
5	0.804	0.914
6	1.022	1.519
7	2.531	7.37

Demographic Parameter Update Rates

Param	Updates	Attempts	%
q1	5.83E+04	1.67E+05	35.01
q2	1.81E+04	1.67E+05	10.86
qA	1.64E+05	1.67E+05	98.37
t	5.41E+03	1.67E+05	3.25
m1	1.31E+05	1.67E+05	78.39
m2	1.35E+05	1.67E+05	81.2

Genealogy and Branching Update Rates

Locus#	G updates	G attempts	G%	B updates	B attempts	B%
0	8.47E+05	1.00E+06	84.75	6.80E+05	1.00E+06	67.96
1	5.75E+05	1.00E+06	57.47	4.16E+05	1.00E+06	41.57
2	3.95E+05	1.00E+06	39.54	2.55E+05	1.00E+06	25.48
3	7.15E+05	1.00E+06	71.49	5.44E+05	1.00E+06	54.41
4	4.47E+05	1.00E+06	44.72	2.87E+05	1.00E+06	28.69
5	6.07E+05	1.00E+06	60.68	4.68E+05	1.00E+06	46.83
6	8.41E+05	1.00E+06	84.14	6.66E+05	1.00E+06	66.58
7	3.69E+05	1.00E+06	36.89	2.51E+05	1.00E+06	25.07

Mutation Update Rates

Locus#	u#	outof#	u%
0	1.40E+05	1.67E+05	83.76
1	1.36E+05	1.67E+05	81.87
2	1.33E+05	1.67E+05	79.59
3	1.41E+05	1.67E+05	84.49
4	1.35E+05	1.67E+05	80.84
5	1.37E+05	1.67E+05	81.94
6	1.41E+05	1.67E+05	84.62
7	1.02E+05	1.67E+05	61.43

Autocorrelations and Effective Sample Size Estimates

step	L[P()]	q1	q2	qA	t	m1	m2
1	0.9981	1	1	0.9789	1	0.994	0.9974
10	0.946	0.9287	0.9741	0.8875	0.9913	0.9693	1
50	0.8063	0.8101	0.9342	0.5959	0.9593	0.9133	0.9659
100	0.7059	0.7452	0.8619	0.3867	0.9309	0.84	0.9076
500	0.4246	0.4048	0.5602	0.0556	0.7548	0.6696	0.695

1000	0.3059	0.23	0.4038	-0.0038	0.6832	0.5409	0.5905
5000	0.0415	0.0168	0.0492	-0.0157	0.3787	0.1927	0.1731
10000	0.0468	0.0139	-0.0348	0.003	0.1831	0.0654	0.0749
50000	-0.0183	-0.005	0.0129	0.0056	0.0342	0.0219	-0.0888
100000	-0.0225	-0.0093	0.0028	-0.0729	-0.0906	-0.0704	-0.0488
500000	-0.0006	0.0166	0.0074	0.018	-0.1478	-0.1906	-0.0855
ESS	214	537	304	3024	53	122	114

Correlations among Parameters

	q1	q2	qA	t	m1	m2	0u	1u	2u	3u	4u	5u	6u	7u
q1		-0.11	0	0.11	-0.25	0.01	0.01	-0.17	-0.27	0.14	-0.22	-0.15	0.12	-0.36
q2			0	0.06	0.01	-0.31	0	-0.12	-0.25	0.11	-0.19	-0.15	0.11	-0.28
qA				-0.07	0.01	0.01	-0.01	0	-0.01	0	0.01	0	-0.01	-0.01
t					-0.06	-0.02	0.01	-0.02	-0.05	0.02	-0.05	-0.03	0.02	-0.14
m1						-0.24	-0.02	0.03	0.09	-0.11	0.12	0.12	-0.04	0.31
m2							0.01	0.07	0.07	-0.05	0.1	-0.02	-0.03	0.24
0u								-0.07	-0.05	-0.16	-0.08	-0.09	-0.18	-0.05
1u									0.04	-0.14	0.01	-0.01	-0.15	0.07
2u										-0.17	0.07	0.06	-0.15	0.17
3u											-0.16	-0.15	-0.19	-0.14
4u												0.02	-0.14	0.16
5u													-0.16	0.09
6u														-0.12
7u														

MARGINAL HISTOGRAMS

Summaries

	q1	q2	qA	t	m1	m2	0u	1u	2u	3u	4u	5u	6u	7u
Minbin	0.0833	0.0572	0.0006	0.505	0.0045	0.007	0.0025	0.0685	0.3162	0.0002	0.1028	0.0294	0.0003	2.1086
Maxbin	1.2907	1.3111	1.2907	9.995	8.9685	13.545	7.1121	13.8038	19.2309	5.9156	15.9956	13.8038	4.8306	49.204
HiPt	0.3855	0.2403	0.3209	4.125	0.1035	0.007	0.6486	1.3552	2.9923	0.2312	1.8197	1.2823	0.2489	8.3946

HiSmth	0.388	0.2494	1.2687	5.425	0.0765	0.007	0.6607	1.2589	3.0479	0.2399
1.7865	1.2359	0.2312	8.3946							
Mean	0.4875	0.2769	0.6579	5.815	1.2735	0.903	0.5598	1.2823	2.9376	0.1854
1.6904	1.2359	0.2032	8.7096							
95Lo	0.2292	0.1075	0.0329	1.445	0.0495	0.021	0.0855	0.3532	1.028	0.0117
0.5297	0.3404	0.0126	4.2462							
95Hi	0.9562	0.6201	1.261	9.755	3.8925	4.403	2.2284	3.9446	7.9433	1.0471
5.0119	3.9446	1.1066	19.9526							
HPD90Lo	0.2228	0.0892	0.0006?	2.5350?	0.0045	0.007	0.1432	0.4487	1.2134	0.0273
0.6368	0.4246	0.0305	4.4055							
HPD90Hi	0.8013	0.4965	1.2907?	9.9750?	2.7585	2.723	2.1086	3.5318	7.1121	1.028
4.4055	3.4674	1.0864	17.5388							
Tail?	complete	complete	rising	falling	complete	complete	complete	complete	complete	
complete	complete	complete	complete	complete						

Value	q1	L	q2	L	qA	L	t	L	m1	L
m2	L	0u	L	1u	L	2u	L	3u	L	4u
L	5u	L	6u	L	7u	L				
HiPt	0.3855	0.00385	0.2403	0.01824	0.3209	0.0014	4.125	0.00177	0.1035	0.00471
0.007	0.01636	0.6486	0.00991	1.3552	0.01257	2.9923	0.01496	0.2312	0.00733	1.8197
0.01348	1.2823	0.0126	0.2489	0.00762	8.3946	0.01876				

Run 2

INPUT AND STARTING INFORMATION

Command line string : -b 100000 -l 1000000 -t 20 -q1 0.5 q2 5.0 -m1 15 -m2 10
Comment at runtime :
Input filename : infile.txt
Output filename: outfile.txt
Random number seed : 1151340960
IM Model:
- each population is constant in size
- All Loci Share the Same Two Migration Rate Parameters
- Inheritance Scalars are treated as constants, as given in input file
- Run Duration -
Burn period, # steps: 100000
Record period, # steps: 1000000
- Metropolis Coupling -
None

Text from input file: Autosomes

- Population Names -
Population 1 : lat
Population 2 : dio
- Locus Information -

Locus#	Locusname	samplesize1	samplesize2	Model	InheritanceScalar	MutationRatesPerYear
0	C1d7	11	8	IS	1.000000	
1	C1f6	11	9	IS	1.000000	
2	C34	10	9	IS	1.000000	
3	C79	11	10	IS	1.000000	
4	c158	10	9	IS	1.000000	
5	C1e3	12	10	IS	1.000000	
6	C1h1	11	9	IS	1.000000	
7	C1a8	11	9	IS	1.000000	

- Maximum Parameter Values -
Max for q1 : 0.65
Max for q2 : 0.46
Max for m1 : 15.00
Max for m2 : 10.00
Max for qA : 0.65
Max for t : 20.00

RUN INFORMATION

Number of steps in chain following burnin: 1000000
Number of steps between recording: 10 Number of record steps: 90909
Number of genealogy updates per step: 1
Time Elapsed : 0 hours, 8 minutes

Highest Total P(D|G, Params) (log) : -80.889

Highest P(D|G, Params) (log) for each Locus

Locus	P(D|G, Params)
0	-2.945
1	-6.168
2	-12.026
3	-1.084
4	-9.009
5	-5.690
6	-1.104
7	-23.349

Highest P(G|PARAMS) (log) for each Locus

Locus 0: 57.216
Locus 1: 55.761
Locus 2: 52.419
Locus 3: 58.456
Locus 4: 47.333
Locus 5: 61.666
Locus 6: 52.336
Locus 7: 48.961

Mean Time of Most Recent Common Ancestor

Locus	Mean Time	Variance
0	0.730	0.770
1	0.726	0.714
2	0.447	0.255
3	1.084	1.709
4	0.667	0.577
5	0.690	0.656
6	0.932	1.285
7	2.270	6.155

Demographic Parameter Update Rates

Param	Updates	Attempts	%
q1	9.01e+004	1.67e+005	54.07
q2	7.81e+004	1.67e+005	46.88
qA	1.66e+005	1.67e+005	99.65
t	8.31e+002	1.67e+005	0.50
m1	1.30e+005	1.67e+005	78.09
m2	1.37e+005	1.67e+005	82.10

Genealogy and Branching Update Rates

Locus#	G updates	G attempts	G%	B updates	B attempts	B%
0	8.51e+005	1.00e+006	85.08	6.81e+005	1.00e+006	68.06
1	5.79e+005	1.00e+006	57.87	4.17e+005	1.00e+006	41.67
2	3.97e+005	1.00e+006	39.65	2.54e+005	1.00e+006	25.43
3	7.17e+005	1.00e+006	71.68	5.45e+005	1.00e+006	54.50
4	4.49e+005	1.00e+006	44.92	2.88e+005	1.00e+006	28.76
5	6.10e+005	1.00e+006	61.03	4.71e+005	1.00e+006	47.05
6	8.44e+005	1.00e+006	84.44	6.67e+005	1.00e+006	66.68
7	3.71e+005	1.00e+006	37.14	2.53e+005	1.00e+006	25.26

Mutation Update Rates

Locus#	u#	outof#	u%
0	1.40e+005	1.67e+005	84.20
1	1.37e+005	1.67e+005	82.17
2	1.33e+005	1.67e+005	79.96
3	1.42e+005	1.67e+005	85.11
4	1.36e+005	1.67e+005	81.35
5	1.37e+005	1.67e+005	82.38
6	1.42e+005	1.67e+005	85.27
7	1.02e+005	1.67e+005	61.22

Autocorrelations and Effective Sample Size Estimates

step	L[P()]	q1	q2	qA	t	m1	m2
1	0.9916	0.9757	1.0000	0.9826	1.0000	1.0000	0.9897
10	0.9298	0.9285	0.9587	0.8930	0.9855	0.9642	0.9847
50	0.7967	0.7826	0.8280	0.5194	0.9608	0.9085	0.9479
100	0.6798	0.6790	0.7382	0.2706	0.9359	0.8501	0.9169
500	0.4737	0.3481	0.4607	-0.0524	0.8096	0.6450	0.7327
1000	0.3041	0.2184	0.2675	-0.0388	0.7430	0.4407	0.6633
5000	0.1181	0.0709	0.0630	-0.0417	0.5363	0.1613	0.3672
10000	0.0453	0.0515	0.0063	0.0471	0.4021	0.0606	0.1326
50000	0.0369	-0.0005	0.0567	0.0224	0.0720	0.0595	0.0005
100000	-0.0581	-0.0347	-0.0642	0.0462	-0.0011	-0.0220	-0.0774
500000	0.0064	0.0653	0.0104	0.0055	0.0623	0.0093	-0.0438

	ESS	127	235	415	4952	32	87	82			

Correlations among Parameters

	q1	q2	qA	t	m1	m2	0u	1u	2u	3u	4u	5u
		6u	7u									
q1		-	0.21	0.01	-0.02	-0.31	-0.03	-0.05	-0.20	-0.29	0.11	-0.22
		-0.20	0.10	-0.42								
q2		-	-	0.01	0.01	-0.04	-0.27	-0.06	-0.15	-0.27	0.10	-0.22
		-0.18	0.10	-0.34								
qA		-	-	-	-0.00	-0.01	-0.01	0.01	0.01	-0.01	0.02	-0.02
		-0.02	-0.00	-0.02								
t		-	-	-	-	0.03	0.02	-0.01	0.01	-0.02	-0.00	0.02
		0.01	-0.01	0.02								
m1		-	-	-	-	-	-0.18	0.01	0.08	0.12	-0.07	0.13
		0.14	-0.09	0.37								
m2		-	-	-	-	-	-	-0.02	0.09	0.07	-0.01	0.10
		0.01	-0.01	0.14								
0u		-	-	-	-	-	-	-	-0.05	-0.03	-0.17	-0.04
		-0.04	-0.14	-0.01								
1u		-	-	-	-	-	-	-	-	0.05	-0.12	0.03
		0.01	-0.14	0.11								
2u		-	-	-	-	-	-	-	-	-	-0.14	0.11
		0.07	-0.11	0.19								
3u		-	-	-	-	-	-	-	-	-	-	-0.13
		-0.15	-0.23	-0.12								
4u		-	-	-	-	-	-	-	-	-	-	-
		0.06	-0.14	0.15								
5u		-	-	-	-	-	-	-	-	-	-	-
		-	-0.13	0.16								
6u		-	-	-	-	-	-	-	-	-	-	-
		-	-	-0.11								
7u		-	-	-	-	-	-	-	-	-	-	-
		-	-	-								

MARGINAL HISTOGRAMS

Summaries

	q1	q2	qA	t	m1	m2	0u	1u	2u	3u
4u	5u	6u	7u							
Minbin	0.0823	0.0323	0.0003	1.0100	0.0075	0.0050	0.0063	0.0530	0.3597	0.0001
0.1107	0.0360	0.0002	2.1086							
Maxbin	0.6453	0.4574	0.6453	19.9900	10.7925	9.9950	10.0925	15.7036	29.9226	3.9446
25.3513	17.5388	6.3680	71.1214							
HiPt	0.4652	0.2144	0.1062	19.7900	1.4625	0.0050	0.5808	1.2823	3.1046	0.2399
1.6596	1.3804	0.2109	10.2802							
HiSmth	0.4678	0.2332	0.1850	19.8500	1.2675	0.0050	0.6486	1.3305	3.2810	0.2270
1.9953	1.4322	0.2109	9.9083							
Mean	0.4271	0.2492	0.3219	12.4700	1.5225	0.8050	0.5702	1.3062	3.2810	0.1570
1.8535	1.3552	0.1660	9.9083							
95Lo	0.2018	0.0972	0.0165	2.2900	0.1125	0.0250	0.0855	0.3467	1.1695	0.0061
0.5702	0.3532	0.0071	4.4055							
95Hi	0.6305	0.4363	0.6286	19.6900	4.5825	4.7450	2.3988	4.0926	9.0365	0.9727
5.7016	4.2462	1.1066	23.5505							
HPD90Lo	0.2599?	0.1110	0.0068?	2.9100?	0.0075	0.0050	0.1355	0.4487	1.3552	0.0163
0.6855	0.4571	0.0189	4.7424							
HPD90Hi	0.6453?	0.4121	0.6447?	19.9900?	3.2925	2.7550	2.1878	3.7325	7.9433	0.9908
4.8306	3.8019	1.0864	20.7014							
Tail?	falling	falling	rising	rising	complete	complete	complete	complete	complete	
complete	complete	complete	complete	complete						

Value	q1	L	q2	L	qA	L	t	L	m1	L
m2	L	0u	L	1u	L	2u	L	3u	L	4u
L	5u	L	6u	L	7u	L				
HiPt	0.4652	0.00233	0.2144	0.00246	0.1062	0.00139	19.7900	0.00185	1.4625	0.00639
0.0050	0.00930	0.5808	0.00956	1.2823	0.01260	3.1046	0.01505	0.2399	0.00701	1.6596
0.01318	1.3804	0.01261	0.2109	0.00677	10.2802	0.01795				

Run 3

INPUT AND STARTING INFORMATION

Command line string : -b 100000 -l 1000000 -q1 0.5 -q2 5.0 -m1 10 -m2 15 -t 15
Comment at runtime :
Input filename : infile.txt
Output filename: outfile.txt
Random number seed : 1151322501
IM Model:
- each population is constant in size
- All Loci Share the Same Two Migration Rate Parameters
- Inheritance Scalars are treated as constants, as given in input file
- Run Duration -
Burn period, # steps: 100000
Record period, # steps: 1000000
- Metropolis Coupling -
None

Text from input file: Autosomes

- Population Names -
Population 1 : lat
Population 2 : dio
- Locus Information -

Locus#	Locusname	samplesize1	samplesize2	Model	InheritanceScalar	MutationRatesPerYear
0	C1d7	11	8	IS	1.000000	
1	C1f6	11	9	IS	1.000000	
2	C34	10	9	IS	1.000000	
3	C79	11	10	IS	1.000000	
4	c158	10	9	IS	1.000000	
5	C1e3	12	10	IS	1.000000	
6	C1h1	11	9	IS	1.000000	
7	C1a8	11	9	IS	1.000000	

- Maximum Parameter Values -
Max for q1 : 0.65
Max for q2 : 4.58
Max for m1 : 10.00
Max for m2 : 15.00
Max for qA : 0.65
Max for t : 15.00

RUN INFORMATION

Number of steps in chain following burnin: 1000000
Number of steps between recording: 10 Number of record steps: 90909
Number of genealogy updates per step: 1
Time Elapsed : 0 hours, 9 minutes

Highest Total P(D|G, Params) (log) : -76.516

Highest P(D|G, Params) (log) for each Locus

Locus	P(D|G, Params)
0	-3.775
1	-6.837
2	-11.875
3	-1.099
4	-8.688
5	-6.016
6	-1.091
7	-20.116

Highest P(G|PARAMS) (log) for each Locus

Locus 0: 48.981
Locus 1: 50.719
Locus 2: 44.741
Locus 3: 55.926
Locus 4: 44.833
Locus 5: 54.738
Locus 6: 51.088
Locus 7: 44.721

Mean Time of Most Recent Common Ancestor

Locus	Mean Time	Variance
0	0.770	0.847
1	0.731	0.720
2	0.466	0.284
3	1.079	1.651
4	0.700	0.644
5	0.749	0.795
6	0.933	1.263
7	2.348	6.474

Demographic Parameter Update Rates

Param	Updates	Attempts	%
q1	9.06e+004	1.67e+005	54.34
q2	1.77e+004	1.67e+005	10.63
qA	1.64e+005	1.67e+005	98.14
t	3.68e+003	1.67e+005	2.21
m1	1.31e+005	1.67e+005	78.66
m2	1.36e+005	1.67e+005	81.62

Genealogy and Branching Update Rates

Locus#	G updates	G attempts	G%	B updates	B attempts	B%
0	8.49e+005	1.00e+006	84.87	6.80e+005	1.00e+006	67.97
1	5.73e+005	1.00e+006	57.32	4.15e+005	1.00e+006	41.51
2	3.95e+005	1.00e+006	39.55	2.54e+005	1.00e+006	25.40
3	7.16e+005	1.00e+006	71.59	5.45e+005	1.00e+006	54.48
4	4.49e+005	1.00e+006	44.93	2.89e+005	1.00e+006	28.92
5	6.09e+005	1.00e+006	60.92	4.70e+005	1.00e+006	47.00
6	8.45e+005	1.00e+006	84.47	6.69e+005	1.00e+006	66.87
7	3.70e+005	1.00e+006	37.03	2.52e+005	1.00e+006	25.15

Mutation Update Rates

Locus#	u#	outof#	u%
0	1.40e+005	1.67e+005	83.95
1	1.37e+005	1.67e+005	82.07
2	1.33e+005	1.67e+005	79.69
3	1.41e+005	1.67e+005	84.85
4	1.35e+005	1.67e+005	81.24
5	1.37e+005	1.67e+005	82.16
6	1.41e+005	1.67e+005	84.86
7	1.02e+005	1.67e+005	61.41

Autocorrelations and Effective Sample Size Estimates

step	L[P()]	q1	q2	qA	t	m1	m2
1	0.9973	0.9814	1.0000	0.9652	1.0000	0.9793	1.0000
10	0.9532	0.9231	0.9738	0.8821	0.9928	0.9610	0.9888
50	0.8199	0.7704	0.9573	0.5604	0.9641	0.8754	0.9416
100	0.7334	0.7127	0.8880	0.3770	0.9590	0.8689	0.8998
500	0.5172	0.3596	0.6070	0.0479	0.8789	0.6128	0.8181
1000	0.3969	0.2390	0.4123	0.0284	0.8181	0.5068	0.6163
5000	0.1400	0.0952	0.1262	0.0112	0.6392	0.1801	0.2173
10000	0.1036	0.0937	0.0989	-0.0084	0.5172	0.0728	0.0680
50000	-0.0286	-0.0316	0.0583	0.0151	0.1430	-0.0138	0.0009
100000	-0.0216	-0.0167	-0.0115	-0.0135	0.0003	-0.0197	-0.0534
500000	-0.0693	-0.0119	-0.0507	0.0748	-0.4073	-0.0170	-0.0123

ESS	127	159	80	3480	23	133	121	

Correlations among Parameters

	q1	q2	qA	t	m1	m2	0u	1u	2u	3u	4u	5u	6u	7u
q1		-	0.17	0.03	-0.06	-0.36	0.05	-0.06	-0.19	-0.31	0.12	-0.22	-0.18	0.13
q2			-	-0.01	0.01	-0.07	-0.28	-0.05	-0.14	-0.27	0.12	-0.21	-0.17	0.15
qA				-	-0.06	-0.05	0.04	-0.01	-0.00	-0.01	0.02	0.00	-0.01	-0.02
t					-	0.14	-0.19	-0.01	0.00	0.01	-0.01	0.02	0.01	0.02
m1						-	-0.23	0.04	0.12	0.14	-0.07	0.16	0.15	-0.08
m2							-	0.02	0.04	0.06	-0.04	0.05	0.01	-0.01
0u								-	-0.04	-0.02	-0.17	-0.03	-0.05	-0.16
1u									-	0.07	-0.12	0.04	0.01	-0.15
2u										-	-0.13	0.10	0.09	-0.15
3u											-	-0.14	-0.16	-0.21
4u												-	0.07	-0.14
5u													-	-0.16
6u														-
7u														

MARGINAL HISTOGRAMS

Summaries

	q1	q2	qA	t	m1	m2	0u	1u	2u	3u	4u	5u	6u	7u
Minbin	0.0681	0.0526	0.0003	0.3525	0.0050	0.0075	0.0043	0.0673	0.2312	0.0001	0.1107	0.0417	0.0001	2.0701
Maxbin	0.6453	1.2104	0.6453	14.9925	9.9050	13.7625	9.7275	13.0617	37.3250	5.2966	18.1970	11.9124	4.3251	51.9996
HiPt	0.4000	0.2265	0.5788	2.7975	1.3650	0.0075	0.7112	1.4060	3.3420	0.2535	1.8535	1.3062	0.2399	9.3756
HiSmth	0.3922	0.2357	0.6137	2.1075	1.3650	0.0075	0.7244	1.4859	3.2211	0.1923	1.8197	1.3804	0.2312	8.8716
Mean	0.4329	0.2677	0.3361	7.4775	1.3750	0.8925	0.5702	1.3552	3.2211	0.1629	1.7865	1.2823	0.1820	9.5499
95Lo	0.1966	0.1030	0.0171	1.2975	0.0750	0.0375	0.0794	0.3597	1.1066	0.0069	0.5297	0.3404	0.0072	4.4055
95Hi	0.6298	0.6201	0.6311	14.6775	4.5150	4.7325	2.4434	4.2462	9.0365	1.0666	5.4954	4.1687	1.1272	21.8776
HPD90Lo	0.2612?	0.0801	0.0003?	1.3425?	0.0050	0.0075	0.1330	0.4571	1.2823	0.0189	0.6607	0.4406	0.0211	4.6559
HPD90Hi	0.6453?	0.4874	0.6453?	14.9925?	3.1450	2.7975	2.2284	3.7325	7.7983	1.0471	4.8306	3.7325	1.1482	19.5884
Tail?	falling	complete	rising	rising	complete	complete	complete	complete	complete	complete	complete	complete	complete	complete

Value	q1	L	q2	L	qA	L	t	L	m1	L	m2	L	5u	L	6u	L	7u	L
HiPt	0.4000	0.00234	0.2265	0.01757	0.5788	0.00150	2.7975	0.00237	1.3650	0.00427	0.0075	0.01294	0.7112	0.00964	1.4060	0.01244	3.3420	0.01440
0.2535	0.00701	1.8535	0.01337	1.3062	0.01235	0.2399	0.00722	9.3756	0.01877									

Run 4

INPUT AND STARTING INFORMATION

Command line string : -b 200000 -L 1000000 -q1 5.0 -q2 5.0 -t 50 -m1 10 -m2 10
Comment at runtime :
Input filename : infile.txt
Output filename: outfile.txt
Random number seed : 1151401168
IM Model:
- each population is constant in size
- All Loci Share the Same Two Migration Rate Parameters
- Inheritance Scalars are treated as constants, as given in input file
- Run Duration -
Burn period, # steps: 200000
Record period, # steps: 1000000
- Metropolis Coupling -
None

Text from input file: Autosomes

- Population Names -
Population 1 : lat
Population 2 : dio
- Locus Information -

Locus#	Locusname	samplesize1	samplesize2	Model	InheritanceScalar	MutationRatesPerYear
0	C1d7	11	8	IS	1.000000	
1	C1f6	11	9	IS	1.000000	
2	C34	10	9	IS	1.000000	
3	C79	11	10	IS	1.000000	
4	c158	10	9	IS	1.000000	
5	C1e3	12	10	IS	1.000000	
6	C1h1	11	9	IS	1.000000	
7	C1a8	11	9	IS	1.000000	

- Maximum Parameter Values -
Max for q1 : 6.46
Max for q2 : 4.58
Max for m1 : 10.00
Max for m2 : 10.00
Max for qA : 6.46
Max for t : 50.00

RUN INFORMATION

Number of steps in chain following burnin: 1000000
Number of steps between recording: 10 Number of record steps: 90909
Number of genealogy updates per step: 1
Time Elapsed : 0 hours, 10 minutes

Highest Total P(D|G, Params) (log) : -77.152

Highest P(D|G, Params) (log) for each Locus

Locus	P(D|G, Params)
0	-3.025
1	-6.113
2	-13.326
3	-1.042
4	-8.294
5	-5.598
6	-1.055
7	-21.222

Highest P(G|PARAMS) (log) for each Locus

Locus 0: 45.152
Locus 1: 47.128
Locus 2: 46.390
Locus 3: 48.677
Locus 4: 41.952
Locus 5: 48.780
Locus 6: 50.676
Locus 7: 44.607

Mean Time of Most Recent Common Ancestor

Locus	Mean Time	Variance
0	0.821	0.948
1	0.801	0.874
2	0.515	0.344
3	1.167	1.937
4	0.746	0.723
5	0.807	0.924
6	1.011	1.475
7	2.542	7.503

Demographic Parameter Update Rates

Param	Updates	Attempts	%
q1	2.04e+004	1.67e+005	12.25
q2	1.80e+004	1.67e+005	10.81
qA	1.67e+005	1.67e+005	100.00
t	2.00e+000	1.67e+005	0.00
m1	1.31e+005	1.67e+005	78.44
m2	1.36e+005	1.67e+005	81.61

Genealogy and Branching Update Rates

Locus#	G updates	G attempts	G%	B updates	B attempts	B%
0	8.47e+005	1.00e+006	84.70	6.79e+005	1.00e+006	67.93
1	5.75e+005	1.00e+006	57.48	4.16e+005	1.00e+006	41.62
2	3.94e+005	1.00e+006	39.43	2.54e+005	1.00e+006	25.39
3	7.15e+005	1.00e+006	71.53	5.44e+005	1.00e+006	54.44
4	4.47e+005	1.00e+006	44.67	2.87e+005	1.00e+006	28.68
5	6.07e+005	1.00e+006	60.74	4.69e+005	1.00e+006	46.92
6	8.41e+005	1.00e+006	84.07	6.66e+005	1.00e+006	66.60
7	3.70e+005	1.00e+006	36.98	2.52e+005	1.00e+006	25.18

Mutation Update Rates

Locus#	u#	outof#	u%
0	1.39e+005	1.67e+005	83.65
1	1.36e+005	1.67e+005	81.87
2	1.33e+005	1.67e+005	79.69
3	1.41e+005	1.67e+005	84.55
4	1.35e+005	1.67e+005	81.09
5	1.37e+005	1.67e+005	81.95
6	1.41e+005	1.67e+005	84.62
7	1.02e+005	1.67e+005	61.32

Autocorrelations and Effective Sample Size Estimates

step	L[P()]	q1	q2	qA	t	m1	m2
1	0.9937	0.9924	1.0000	0.9676	1.0000	0.9684	1.0000
10	0.9334	0.9400	0.9768	0.8771	0.9956	0.9427	0.9847
50	0.8076	0.8566	0.8940	0.5688	0.9704	0.9043	0.9540
100	0.6985	0.7834	0.8278	0.3005	0.9579	0.8375	0.8691
500	0.4477	0.5056	0.5478	-0.0050	0.8665	0.6961	0.7672
1000	0.3466	0.3017	0.4007	0.0095	0.8084	0.5173	0.6151
5000	0.0087	0.0340	0.0795	0.0160	0.6371	0.1458	0.2455
10000	0.0319	-0.0130	0.0311	0.0681	0.5035	0.0298	0.0170
50000	0.0036	-0.0085	0.0245	0.0083	0.1018	0.0079	-0.0381
100000	-0.0338	0.0131	-0.0364	-0.0397	-0.0255	-0.0064	0.0463
500000	-0.0106	-0.0151	0.0571	-0.0038	-0.1059	0.0727	-0.0188

1000000	0.0767	0.0227	-0.0403	0.0096	-0.3542	-0.0609	-0.0369	
ESS	499	461	244	5018	28	251	193	

Correlations among Parameters

	q1	q2	qA	t	m1	m2	0u	1u	2u	3u	4u	5u
		6u	7u									
q1		-	0.12	0.01	0.06	-0.28	-0.01	-0.00	-0.17	-0.26	0.14	-0.19
		-0.16	0.13	-0.37								
q2		-	-	-0.00	0.06	0.09	-0.32	-0.03	-0.14	-0.24	0.13	-0.19
		-0.15	0.14	-0.30								
qA		-	-	-	-0.00	-0.01	0.02	-0.01	-0.00	-0.01	0.02	0.01
		-0.01	0.01	-0.00								
t		-	-	-	-	-0.03	-0.06	0.02	-0.02	-0.07	0.02	-0.05
		-0.03	0.00	-0.11								
m1		-	-	-	-	-	-0.27	0.02	0.03	0.09	-0.08	0.12
		0.10	-0.09	0.32								
m2		-	-	-	-	-	-	-0.03	0.08	0.08	-0.02	0.07
		0.01	-0.06	0.25								
0u		-	-	-	-	-	-	-	-0.07	-0.05	-0.17	-0.05
		-0.07	-0.17	-0.02								
1u		-	-	-	-	-	-	-	-	0.04	-0.14	0.02
		0.00	-0.16	0.11								
2u		-	-	-	-	-	-	-	-	-	-0.15	0.08
		0.03	-0.14	0.16								
3u		-	-	-	-	-	-	-	-	-	-	-0.15
		-0.15	-0.19	-0.15								
4u		-	-	-	-	-	-	-	-	-	-	-
		0.02	-0.16	0.16								
5u		-	-	-	-	-	-	-	-	-	-	-
		-	-0.16	0.10								
6u		-	-	-	-	-	-	-	-	-	-	-
		-	-	-0.15								
7u		-	-	-	-	-	-	-	-	-	-	-
		-	-	-								

MARGINAL HISTOGRAMS

Summaries

	q1	q2	qA	t	m1	m2	0u	1u	2u	3u
4u	5u	6u	7u							
Minbin	0.1130	0.0526	0.0032	2.9250	0.0050	0.0050	0.0075	0.0661	0.2443	0.0001
0.1459	0.0649	0.0006	1.8880							
Maxbin	1.8498	1.2607	6.4534	49.9750	9.9950	9.9850	8.8716	22.6986	24.8886	4.7424
19.2309	16.5959	5.1050	51.9996							
HiPt	0.4487	0.2174	5.3170	40.5250	0.2050	0.0450	0.6486	1.2359	2.8314	0.2831
2.1086	1.3305	0.3664	8.3946							
HiSmth	0.4294	0.2174	2.8764	40.7250	0.1850	0.0050	0.6252	1.1912	2.8840	0.2270
1.6904	1.3552	0.3404	8.5507							
Mean	0.4875	0.2769	3.2509	31.2750	1.2050	0.8750	0.5702	1.2359	2.9376	0.1888
1.7219	1.2359	0.2109	8.7096							
95Lo	0.2292	0.1121	0.1582	5.8750	0.0650	0.0350	0.0887	0.3467	1.0093	0.0119
0.5297	0.3162	0.0119	4.0926							
95Hi	0.9524	0.6292	6.2984	49.0750	4.1050	4.3150	2.1878	3.8726	8.0910	0.9908
5.2000	3.9446	1.1695	19.5884							
HPD90Lo	0.2034	0.0892	0.0097?	7.8750?	0.0050	0.0050	0.1459	0.4406	1.1912	0.0283
0.6486	0.4169	0.0278	4.4055							
HPD90Hi	0.8103	0.5011	6.4534?	49.9750?	2.8250	2.6250	2.0701	3.4674	7.1121	0.9727
4.5709	3.5318	1.1482	17.8649							
Tail?	complete	complete	flat	rising	complete	complete	complete	complete	complete	
complete	complete	complete	complete	complete						

Value	q1	L	q2	L	qA	L	t	L	m1	L
m2	L	0u	L	1u	L	2u	L	3u	L	4u
L	5u	L	6u	L	7u	L				

HiPt	0.4487	0.01813	0.2174	0.01752	5.3170	0.00133	40.5250	0.00190	0.2050	0.00506
0.0450	0.00812	0.6486	0.01015	1.2359	0.01253	2.8314	0.01453	0.2831	0.00779	2.1086
0.01348	1.3305	0.01228	0.3664	0.00752	8.3946	0.01940				

Run 5

INPUT AND STARTING INFORMATION

Command line string : -b 100000 -l 1000000 -q1 10 -m1 20 -m2 20 -t 20 -fl -n 5 -gl 0.05 -k2 -p 8
Comment at runtime :
Input filename : infile.txt
Output filename: outfile.txt
Random number seed : 1151403243
IM Model:
- each population is constant in size
- All Loci Share the Same Two Migration Rate Parameters
- Inheritance Scalars are treated as constants, as given in input file

- Run Duration -
Burn period, # steps: 100000
Record period, # steps: 1000000
- Metropolis Coupling -
Metropolis Coupling implemented using 5 chains
Linear Increment Model term: 0.050

Text from input file: Autosomes

- Population Names -
Population 1 : lat
Population 2 : dio
- Locus Information -

Locus#	Locusname	samplesize1	samplesize2	Model	InheritanceScalar	MutationRatesPerYear
0	C1d7	11	8	IS	1.000000	
1	C1f6	11	9	IS	1.000000	
2	C34	10	9	IS	1.000000	
3	C79	11	10	IS	1.000000	
4	c158	10	9	IS	1.000000	
5	C1e3	12	10	IS	1.000000	
6	C1h1	11	9	IS	1.000000	
7	C1a8	11	9	IS	1.000000	

- Maximum Parameter Values -
Max for q1 : 12.91
Max for q2 : 9.15
Max for m1 : 20.00
Max for m2 : 20.00
Max for qA : 12.91
Max for t : 20.00

RUN INFORMATION

Number of steps in chain following burnin: 1000000
Number of steps between recording: 10 Number of record steps: 90909
Number of genealogy updates per step: 1
Time Elapsed : 1 hours, 4 minutes

Highest Total P(D|G, Params) (log) : -78.232

Highest P(D|G, Params) (log) for each Locus

Locus	P(D|G, Params)
0	-3.089
1	-6.888
2	-12.292
3	-1.027
4	-8.164
5	-5.858
6	-1.037
7	-20.870

Highest P(G|PARAMS) (log) for each Locus

Locus 0: 58.879
Locus 1: 55.729
Locus 2: 57.415
Locus 3: 61.892
Locus 4: 47.070
Locus 5: 67.695
Locus 6: 67.127
Locus 7: 58.715

Mean Time of Most Recent Common Ancestor

Locus	Mean Time	Variance
0	0.802	0.953
1	0.788	0.840
2	0.503	0.329
3	1.224	2.774
4	0.758	0.757
5	0.782	0.926
6	1.027	1.800
7	2.733	9.393

Demographic Parameter Update Rates

Param	Updates	Attempts	%
q1	1.23e+004	1.67e+005	7.36
q2	1.00e+004	1.67e+005	6.00
qA	1.64e+005	1.67e+005	98.27
t	2.14e+003	1.67e+005	1.29
m1	1.31e+005	1.67e+005	78.40
m2	1.36e+005	1.67e+005	81.82

Genealogy and Branching Update Rates

Locus#	G updates	G attempts	G%	B updates	B attempts	B%
0	8.47e+005	1.00e+006	84.71	6.79e+005	1.00e+006	67.93
1	5.75e+005	1.00e+006	57.49	4.16e+005	1.00e+006	41.63
2	3.96e+005	1.00e+006	39.56	2.56e+005	1.00e+006	25.56
3	7.15e+005	1.00e+006	71.52	5.45e+005	1.00e+006	54.53
4	4.46e+005	1.00e+006	44.61	2.86e+005	1.00e+006	28.65
5	6.08e+005	1.00e+006	60.82	4.70e+005	1.00e+006	47.03
6	8.40e+005	1.00e+006	84.02	6.66e+005	1.00e+006	66.63
7	3.71e+005	1.00e+006	37.08	2.52e+005	1.00e+006	25.23

Mutation Update Rates

Locus#	u#	outof#	u%
0	1.40e+005	1.67e+005	83.79
1	1.36e+005	1.67e+005	81.86
2	1.32e+005	1.67e+005	79.49
3	1.41e+005	1.67e+005	84.52
4	1.35e+005	1.67e+005	81.04
5	1.37e+005	1.67e+005	81.99
6	1.41e+005	1.67e+005	84.46
7	1.02e+005	1.67e+005	61.16

Autocorrelations and Effective Sample Size Estimates

step	L[P()]	q1	q2	qA	t	m1	m2
1	0.9900	0.9807	0.9381	0.9641	0.9825	0.9839	0.9899
10	0.9114	0.9115	0.8606	0.7922	0.9070	0.9557	0.9559
50	0.8234	0.8499	0.8097	0.4979	0.8470	0.8630	0.8579
100	0.8223	0.8154	0.7433	0.2745	0.8268	0.8203	0.8495
500	0.6041	0.5973	0.5620	0.0044	0.7333	0.7167	0.6640
1000	0.5030	0.4273	0.3956	-0.0157	0.6406	0.5322	0.5105
5000	0.2606	0.1419	0.1363	0.0373	0.3621	0.1354	0.1692
10000	0.1914	0.0353	0.1158	-0.0327	0.2617	0.0540	0.0902
50000	-0.0065	0.0200	-0.0476	0.0163	0.1395	-0.1029	-0.0020
100000	-0.0106	-0.0217	-0.0612	0.0406	0.0045	0.0307	0.0435

500000	-0.0347	-0.0752	-0.0117	0.0185	0.0524	-0.0994	-0.0487
ESS	77	189	120	5073	35	153	122

Beta values for chain swapping

chain#	Beta
0	1.00000
1	0.95238
2	0.90909
3	0.86957
4	0.83333

Mean # of swap attempts between pairs of chains: 220000

Overall Chain swapping - % above diagonal, counts below

	chain0	|chain1	|chain2	|chain3	|chain4
chain0	-	12.705	0.665	0.015	0.000
chain1	27913	-	21.107	1.865	0.079
chain2	1468	46448	-	26.537	3.031
chain3	33	4117	58471	-	28.945
chain4	0	174	6685	63504	-

Correlations among Parameters

	q1	q2	qA	t	m1	m2	0u	1u	2u	3u	4u	5u
	6u	7u										
q1	-	0.17	0.01	-0.02	-0.37	0.03	-0.07	-0.21	-0.32	0.20	-0.23	
	-0.24	0.16	-0.43									
q2	-	-	-0.01	-0.09	-0.04	-0.28	-0.05	-0.15	-0.29	0.16	-0.21	
	-0.21	0.13	-0.36									
qA	-	-	-	0.01	0.01	-0.02	-0.00	0.01	-0.01	0.01	0.02	
	-0.00	-0.00	-0.02									
t	-	-	-	-	-0.05	0.02	0.01	0.01	0.03	0.01	0.00	
	0.03	-0.07	0.17									
m1	-	-	-	-	-	-0.17	0.06	0.15	0.25	-0.12	0.21	
	0.23	-0.09	0.40									
m2	-	-	-	-	-	-	0.05	0.08	0.06	-0.06	0.07	
	0.03	-0.02	0.18									
0u	-	-	-	-	-	-	-	-0.03	0.01	-0.18	-0.01	
	-0.02	-0.17	0.06									
1u	-	-	-	-	-	-	-	-	0.10	-0.15	0.08	
	0.05	-0.18	0.22									
2u	-	-	-	-	-	-	-	-	-	-0.18	0.15	
	0.16	-0.17	0.32									
3u	-	-	-	-	-	-	-	-	-	-	-0.18	
	-0.18	-0.17	-0.17									
4u	-	-	-	-	-	-	-	-	-	-	-	
	0.13	-0.17	0.23									
5u	-	-	-	-	-	-	-	-	-	-	-	
	-	-0.17	0.27									
6u	-	-	-	-	-	-	-	-	-	-	-	
	-	-	-0.18									
7u	-	-	-	-	-	-	-	-	-	-	-	
	-	-	-									

MARGINAL HISTOGRAMS

Summaries

	q1	q2	qA	t	m1	m2	0u	1u	2u	3u
4u	5u	6u	7u							
Minbin	0.0968	0.0503	0.0065	0.3900	0.0100	0.0100	0.0061	0.0711	0.2188	0.0002
0.1028	0.0474	0.0001	1.1912							
Maxbin	1.8660	1.4690	12.9068	19.9900	17.6100	15.3900	8.5507	13.8038	41.6869	4.5709
29.3765	16.5959	5.2966	60.2560							
HiPt	0.3939	0.2608	2.3050	1.0100	0.0300	0.0100	0.6855	1.3804	2.7290	0.2831
1.6904	1.1272	0.3105	8.2414							
HiSmth	0.4197	0.2425	4.8231	0.9700	0.0300	0.0100	0.6368	1.3062	2.9376	0.2884
1.6904	1.3552	0.3048	8.3946							

Mean	0.4842	0.2791	6.3210	10.1300	1.2900	0.7700	0.5916	1.2823	2.9923	0.1786
1.6904	1.2823	0.2109	8.5507							
95Lo	0.1872	0.0961	0.3551	0.9300	0.0500	0.0300	0.0920	0.3597	1.0666	0.0075
0.5297	0.3281	0.0107	3.4041							
95Hi	1.0008	0.6269	12.5710	19.5100	4.8500	5.0300	2.3550	4.0179	8.7096	1.0471
5.2000	4.3251	1.1695	22.6986							
HPD90Lo	0.1743	0.0778	0.0710?	0.6700?	0.0100	0.0100	0.1459	0.4487	1.2134	0.0211
0.6368	0.4246	0.0288	3.8019							
HPD90Hi	0.8587	0.5171	12.7647?	19.9700?	2.9700	2.8500	2.1478	3.5318	7.3790	1.0666
4.4875	3.7325	1.2134	18.8799							
Tail?	complete	complete	flat	rising	complete	complete	complete	complete	complete	
complete	complete	complete	complete	complete						

Value	q1	L	q2	L	qA	L	t	L	m1	L
m2	L	0u	L	1u	L	2u	L	3u	L	4u
L	5u	L	6u	L	7u	L				
HiPt	0.3939	0.03038	0.2608	0.03440	2.3050	0.00142	1.0100	0.00301	0.0300	

Appendix 3 - WH Program Output Files

S. latifolia vs *S. dioica* (19 loci)

Data file : infileFINAL.txt This output file : outfileFINAL.wh

******** INPUT ********

Data file header : Autosomals 17 Recom

MESSAGE ADDED AT RUNTIME:

Species, Loci, Sample Sizes and Polymorphism Counts

Loci		c37	c109	c1d7	c1f6	c2d5	c18	c110	c158	c34	c79	c3a4	c1a11	c1e3	c1e4	c1h1	c2c4	c1g11
species0 =	Lat	14	17	14	14	11	11	16	11	13	14	13	6	16	8	14	11	15
species1 =	Dio	14	13	13	12	12	11	14	12	13	13	13	15	12	8	12	8	14
Sx1		16.0	8.0	0.0	4.0	6.0	2.0	5.0	3.0	5.0	4.0	13.0	2.0	1.0	3.0	0.0	6.0	49.0
Sx2		6.0	4.0	0.0	0.0	4.0	9.0	4.0	2.0	5.0	10.0	6.0	8.0	2.0	8.0	0.0	1.0	35.0
Ss		2.0	10.0	0.0	0.0	3.0	12.0	7.0	0.0	0.0	2.0	10.0	2.0	0.0	6.0	0.0	5.0	8.0
Sf		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0

Population recombination values (4Nc) given and used in simulations

Locus	species0: given	used	species1: given	used	Ancestor
c37	0.0000	0.0000	0.0000	0.0000	0.0000
c109	0.0000	0.0000	0.0000	0.0000	0.0000
c1d7	0.0000	0.0000	0.0000	0.0000	0.0000
c1f6	0.0000	0.0000	0.0000	0.0000	0.0000
c2d5	0.0000	0.0000	0.0000	0.0000	0.0000
c18	0.0000	0.0000	0.0000	0.0000	0.0000
c110	0.0000	0.0000	0.0000	0.0000	0.0000
c158	0.0000	0.0000	0.0000	0.0000	0.0000
c34	0.0000	0.0000	0.0000	0.0000	0.0000
c79	0.0000	0.0000	0.0000	0.0000	0.0000
c3a4	0.0000	0.0000	0.0000	0.0000	0.0000
c1a11	0.0000	0.0000	0.0000	0.0000	0.0000
c1e3	0.0000	0.0000	0.0000	0.0000	0.0000
c1e4	0.0000	0.0000	0.0000	0.0000	0.0000
c1h1	0.0000	0.0000	0.0000	0.0000	0.0000
c2c4	0.0000	0.0000	0.0000	0.0000	0.0000
c1g11	0.0000	0.0000	0.0000	0.0000	0.0000

Total mutation rate per year not provided
Generation time not provided
******** MODEL FITTING RESULTS ********

Basic Parameter Estimates

Theta1 est = 61.377937;
Theta2 est = 45.717841;
ThetaA est = 63.699685;
tau est = 9.536043;
f[0] = 0.077916;
f[1] = 0.069729;

f[2] = 0.000000;
f[3] = 0.013261;
f[4] = 0.045009;
f[5] = 0.080635;
f[6] = 0.050753;
f[7] = 0.017311;
f[8] = 0.033226;
f[9] = 0.052469;
f[10] = 0.099678;
f[11] = 0.044706;
f[12] = 0.009712;
f[13] = 0.066429;
f[14] = 0.000000;
f[15] = 0.044054;
f[16] = 0.295112;

Locus Specific Parameter Values

	c37	c109	c1d7	c1f6	c2d5	c18	c110	c158	c34	c79	c3a4
	c1a11	c1e3	c1e4	c1h1	c2c4	c1g11					
Theta1:	4.782321		4.279833		0.000000		0.813938		2.762558		
4.949195	3.115093		1.062522		2.039334		3.220460		6.118003		
2.743980	0.596113		4.077255		0.000000		2.703950		18.113381		
Theta2:	3.562150		3.187867		0.000000		0.606268		2.057713		
3.686447	2.320302		0.791428		1.519014		2.398785		4.557043		
2.043875	0.444020		3.036976		0.000000		2.014059		13.491895		
ThetaA:	4.963223		4.441726		0.000000		0.844727		2.867058		
5.136408	3.232928		1.102714		2.116476		3.342281		6.349429		
2.847776	0.618662		4.231486		0.000000		2.806233		18.798557		
tau :	0.743010		0.664940		0.000000		0.126458		0.429208		
0.768937	0.483980		0.165080		0.316843		0.500350		0.950530		
0.426321	0.092616		0.633467		0.000000		0.420102		2.814203		

Expected Polymorphism Levels

Locus	Sx1	Sx2	Ss	Sf
c37 Obs.	16.000000	6.000000	2.000000	0.000000
c37 Exp.	10.153151	8.370974	5.413887	0.061988
c109 Obs.	8.000000	4.000000	10.000000	0.000000
c109 Exp.	9.884378	7.148726	4.915553	0.051343
c1d7 Obs.	0.000000	0.000000	0.000000	0.000000
c1d7 Exp.	0.000000	0.000000	0.000000	0.000000
c1f6 Obs.	4.000000	0.000000	0.000000	0.000000
c1f6 Exp.	1.750101	1.338620	0.899368	0.011911
c2d5 Obs.	6.000000	4.000000	3.000000	0.000000
c2d5 Exp.	5.355232	4.661915	2.933957	0.048897
c18 Obs.	2.000000	9.000000	12.000000	0.000000
c18 Exp.	9.674366	8.055794	5.175931	0.093910
c110 Obs.	5.000000	4.000000	7.000000	0.000000
c110 Exp.	6.984184	5.387998	3.591140	0.036677
c158 Obs.	3.000000	2.000000	0.000000	0.000000
c158 Exp.	2.059705	1.793044	1.128445	0.018807
c34 Obs.	5.000000	5.000000	0.000000	0.000000
c34 Exp.	4.221790	3.491004	2.257607	0.029599
c79 Obs.	4.000000	10.000000	2.000000	0.000000
c79 Exp.	6.877771	5.472830	3.605218	0.044180
c3a4 Obs.	13.000000	6.000000	10.000000	1.000000

c3a4 Exp.	12.665369	10.473013	6.772822	0.088796
c1a11 Obs.	2.000000	8.000000	2.000000	0.000000
c1a11 Exp.	3.849835	5.489793	2.580836	0.079536
c1e3 Obs.	1.000000	2.000000	0.000000	0.000000
c1e3 Exp.	1.352985	0.968321	0.670739	0.007955
c1e4 Obs.	3.000000	8.000000	6.000000	0.000000
c1e4 Exp.	7.130081	6.023823	3.711346	0.134749
c1h1 Obs.	0.000000	0.000000	0.000000	0.000000
c1h1 Exp.	0.000000	0.000000	0.000000	0.000000
c2c4 Obs.	6.000000	1.000000	5.000000	0.000000
c2c4 Exp.	5.475337	3.818165	2.637995	0.068503
c1g11 Obs.	49.000000	35.000000	8.000000	0.000000
c1g11 Exp.	39.565714	31.505979	20.705157	0.223150

Other Calculations

T est (tau/Theta1)= 0.1554

******** SIMULATION RESULTS ********

Simulations attempted : 2000
simulations that yielded estimates : 536 (0.2680)
with zero fixed differences : 1464 (0.7320)
with zero shared differences : 0 (0.0000)

ChiSquare test statistic value : 85.0299
successful simulations with higher values : 499 (0.9310)

wh test statistic value : 13.0000
successful simulations with higher values : 448 (0.8358)

Estimated parameter values and statistics

Parameter	Estimate	2.4%	- 97.6%	Mean	Variance
Theta1	61.378	0.013	158.160	52.249	1164.00434
Theta2	45.718	0.013	98.583	38.427	544.04481
ThetaA	63.700	33.650	144.296	73.893	794.59820
tau	9.536	0.003	17.713	9.396	17.56024
T (tau/theta1)	0.155	0.086	0.402	0.210	0.00653
c37 frac	0.078	0.038	0.135	0.079	0.00063
c109 frac	0.070	0.034	0.126	0.069	0.00052
c1d7 frac	0.000	0.000	0.000	0.000	0.00001
c1f6 frac	0.013	0.003	0.034	0.014	0.00006
c2d5 frac	0.045	0.017	0.084	0.045	0.00033
c18 frac	0.081	0.029	0.138	0.076	0.00067
c110 frac	0.051	0.023	0.093	0.052	0.00034
c158 frac	0.017	0.003	0.039	0.017	0.00008
c34 frac	0.033	0.013	0.069	0.034	0.00021
c79 frac	0.052	0.023	0.098	0.053	0.00037
c3a4 frac	0.100	0.053	0.174	0.101	0.00097
c1a11 frac	0.045	0.000	0.086	0.045	0.00033
c1e3 frac	0.010	0.000	0.024	0.010	0.00004
c1e4 frac	0.066	0.000	0.123	0.065	0.00060
c1h1 frac	0.000	0.000	0.000	0.000	0.00001
c2c4 frac	0.044	0.000	0.081	0.043	0.00029
c1g11 frac	0.295	0.000	0.433	0.298	0.00401

Compare actual and simulated polymorphism levels

Locus	Sx1	Sx2	Ss	Sf
c37 Obs.	16.000000	6.000000	2.000000	0.000000
c37 Sim.	10.505597	8.220149	5.208955	0.236940
c109 Obs.	8.000000	4.000000	10.000000	0.000000
c109 Sim.	9.757463	7.322761	4.294776	0.160448
c1d7 Obs.	0.000000	0.000000	0.000000	0.000000
c1d7 Sim.	0.000000	0.000000	0.000000	0.000000
c1f6 Obs.	4.000000	0.000000	0.000000	0.000000
c1f6 Sim.	1.785448	1.389925	0.955224	0.046642
c2d5 Obs.	6.000000	4.000000	3.000000	0.000000
c2d5 Sim.	5.419776	4.789179	2.587687	0.130597
c18 Obs.	2.000000	9.000000	12.000000	0.000000
c18 Sim.	9.236940	7.688433	4.367537	0.393657
c110 Obs.	5.000000	4.000000	7.000000	0.000000
c110 Sim.	6.848881	5.376866	3.839552	0.231343
c158 Obs.	3.000000	2.000000	0.000000	0.000000
c158 Sim.	2.031716	1.845149	1.067164	0.057836
c34 Obs.	5.000000	5.000000	0.000000	0.000000
c34 Sim.	4.315299	3.531716	2.160448	0.134328
c79 Obs.	4.000000	10.000000	2.000000	0.000000
c79 Sim.	7.126866	5.563433	3.233209	0.180970
c3a4 Obs.	13.000000	6.000000	10.000000	1.000000
c3a4 Sim.	13.335821	10.371269	6.326493	0.285448
c1a11 Obs.	2.000000	8.000000	2.000000	0.000000
c1a11 Sim.	3.776119	5.626866	2.507463	0.298507
c1e3 Obs.	1.000000	2.000000	0.000000	0.000000
c1e3 Sim.	1.317164	0.957090	0.757463	0.024254
c1e4 Obs.	3.000000	8.000000	6.000000	0.000000
c1e4 Sim.	7.238806	5.951493	3.169776	0.451493
c1h1 Obs.	0.000000	0.000000	0.000000	0.000000
c1h1 Sim.	0.000000	0.000000	0.000000	0.000000
c2c4 Obs.	6.000000	1.000000	5.000000	0.000000
c2c4 Sim.	5.485075	3.735075	2.279851	0.197761
c1g11 Obs.	49.000000	35.000000	8.000000	0.000000
c1g11 Sim.	40.057836	32.796642	19.861940	0.843284

S. latifolia vs *S. diclinis* (6 loci)

Data file : infileLATDIC6.txt This output file : outfileLATDIC6.wh

******** INPUT ********

Data file header : LAT DIC 6

MESSAGE ADDED AT RUNTIME:

Species, Loci, Sample Sizes and Polymorphism Counts

Loci		c109	c110	c37	c34	c1g11	c1a11
species0 =	lat	17	16	13	16	15	5
species1 =	dic	45	50	34	37	14	29
Sx1		14.0	9.0	16.0	4.0	28.0	3.0
Sx2		1.0	2.0	0.0	4.0	10.0	6.0
Ss		4.0	3.0	0.0	0.0	0.0	1.0
Sf		0.0	1.0	3.0	0.0	3.0	1.0

Population recombination values (4Nc) given and used in simulations

Locus	species0: given	used	species1: given	used	Ancestor
c109	0.0000	0.0000	0.0000	0.0000	0.0000
c110	0.0000	0.0000	0.0000	0.0000	0.0000
c37	0.0000	0.0000	0.0000	0.0000	0.0000
c34	0.0000	0.0000	0.0000	0.0000	0.0000
c1g11	0.0000	0.0000	0.0000	0.0000	0.0000
c1a11	0.0000	0.0000	0.0000	0.0000	0.0000

Total mutation rate per year not provided
Generation time not provided
******** MODEL FITTING RESULTS ********

Basic Parameter Estimates

Theta1 est = 18.278997;
Theta2 est = 4.711396;
ThetaA est = 33.602325;
tau est = 4.284069;
f[0] = 0.157716;
f[1] = 0.125235;
f[2] = 0.166963;
f[3] = 0.067604;
f[4] = 0.364909;
f[5] = 0.117572;

Locus Specific Parameter Values

	c109	c110	c37	c34	c1g11	c1a11
Theta1:	2.882882	2.289176	3.051924	1.235740	6.670172	2.149104
Theta2:	0.743060	0.590033	0.786631	0.318511	1.719231	0.553930
ThetaA:	5.299608	4.208198	5.610358	2.271665	12.261793	3.950703
tau :	0.675664	0.536517	0.715283	0.289622	1.563295	0.503688

Expected Polymorphism Levels

Locus	Sx1	Sx2	Ss	Sf
c109 Obs.	14.000000	1.000000	4.000000	0.000000

c109 Exp.	12.490093	3.998230	1.372552	1.139125
c110 Obs.	9.000000	2.000000	3.000000	1.000000
c110 Exp.	9.751496	3.248356	1.086694	0.913454
c37 Obs.	16.000000	0.000000	0.000000	3.000000
c37 Exp.	12.277177	4.049146	1.381123	1.292555
c34 Obs.	4.000000	4.000000	0.000000	0.000000
c34 Exp.	5.272233	1.651924	0.578426	0.497416
c1g11 Obs.	28.000000	10.000000	0.000000	3.000000
c1g11 Exp.	28.216636	7.082530	2.832412	2.868421
c1a11 Obs.	3.000000	6.000000	1.000000	1.000000
c1a11 Exp.	5.992364	2.969814	0.748793	1.289028

Other Calculations

T est (tau/Theta1)= 0.2344

******** SIMULATION RESULTS ********

Simulations attempted : 2000
simulations that yielded estimates : 1714 (0.8570)
with zero fixed differences : 236 (0.1180)
with zero shared differences : 2 (0.0010)

ChiSquare test statistic value : 34.8264
successful simulations with higher values : 1696 (0.9895)

wh test statistic value : 7.0000
successful simulations with higher values : 1570 (0.9160)

Estimated parameter values and statistics

Parameter	Estimate	1.5%	- 98.5%	Mean	Variance
Theta1	18.279	0.002	223.573	470.203	163689148.30455
Theta2	4.711	0.001	11.948	3.499	10.45428
ThetaA	33.602	12.080	84.619	38.258	287.51219
tau	4.284	0.001	7.380	2.343	3.90623
T (tau/theta1)	0.234	0.012	0.595	0.214	0.01795
c109 frac	0.158	0.000	0.316	0.161	0.00364
c110 frac	0.125	0.000	0.262	0.129	0.00270
c37 frac	0.167	0.000	0.345	0.171	0.00414
c34 frac	0.068	0.000	0.164	0.071	0.00123
c1g11 frac	0.365	0.000	0.569	0.349	0.00947
c1a11 frac	0.118	0.000	0.276	0.119	0.00331

Compare actual and simulated polymorphism levels

Locus	Sx1	Sx2	Ss	Sf
c109 Obs.	14.000000	1.000000	4.000000	0.000000
c109 Sim.	12.484831	3.994749	1.250292	1.191949
c110 Obs.	9.000000	2.000000	3.000000	1.000000
c110 Sim.	2.046091	2.713536	10.225788	0.000000
c37 Obs.	16.000000	0.000000	0.000000	3.000000
c37 Sim.	12.362894	4.035006	1.274212	1.413652
c34 Obs.	4.000000	4.000000	0.000000	0.000000
c34 Sim.	5.294049	1.554259	0.541424	0.462077

c1g11 Obs.	28.000000	10.000000	0.000000	3.000000
c1g11 Sim.	28.106184	6.792882	2.343641	2.830222
c1a11 Obs.	3.000000	6.000000	1.000000	1.000000
c1a11 Sim.	5.845391	2.976663	0.665694	1.408401

S. dioica vs *S. diclinis* (6 loci)

Data file : infileDIODIC6.txt This output file : outfileDIODIC6.wh

******** INPUT ********

Data file header : DIO DIC 6

MESSAGE ADDED AT RUNTIME:

Species, Loci, Sample Sizes and Polymorphism Counts

Loci		c109	c110	c37	c34	c1g11	c1a11
species0 =	dio	13	14	14	14	14	14
species1 =	dic	45	50	34	37	14	29
Sx1		10.0	7.0	9.0	2.0	26.0	11.0
Sx2		1.0	1.0	0.0	4.0	10.0	5.0
Ss		4.0	4.0	0.0	0.0	0.0	2.0
Sf		0.0	0.0	3.0	0.0	4.0	1.0

Population recombination values (4Nc) given and used in simulations

Locus	species0: given	used	species1: given	used	Ancestor
c109	0.0000	0.0000	0.0000	0.0000	0.0000
c110	0.0000	0.0000	0.0000	0.0000	0.0000
c37	0.0000	0.0000	0.0000	0.0000	0.0000
c34	0.0000	0.0000	0.0000	0.0000	0.0000
c1g11	0.0000	0.0000	0.0000	0.0000	0.0000
c1a11	0.0000	0.0000	0.0000	0.0000	0.0000

Total mutation rate per year not provided
Generation time not provided
******** MODEL FITTING RESULTS ********

Basic Parameter Estimates

Theta1 est = 11.144097;
Theta2 est = 3.540411;
ThetaA est = 36.531170;
tau est = 2.903239;
f[0] = 0.142519;
f[1] = 0.112459;
f[2] = 0.113988;
f[3] = 0.056823;
f[4] = 0.392699;
f[5] = 0.181512;

Locus Specific Parameter Values

	c109	c110	c37	c34	c1g11	c1a11
Theta1:	1.588245	1.253254	1.270292	0.633246	4.376276	2.022785
Theta2:	0.504576	0.398151	0.403564	0.201178	1.390316	0.642626
ThetaA:	5.206383	4.108257	4.164112	2.075827	14.345753	6.630838
tau :	0.413767	0.326495	0.330934	0.164972	1.140099	0.526972

Expected Polymorphism Levels

Locus	Sx1	Sx2	Ss	Sf
c109 Obs.	10.000000	1.000000	4.000000	0.000000

c109 Exp.	9.051876	3.348434	1.475216	1.124474
c110 Obs.	7.000000	1.000000	4.000000	0.000000
c110 Exp.	7.273401	2.675608	1.182244	0.868746
c37 Obs.	9.000000	0.000000	0.000000	3.000000
c37 Exp.	7.394648	2.537541	1.175958	0.891853
c34 Obs.	2.000000	4.000000	0.000000	0.000000
c34 Exp.	3.683458	1.284347	0.589025	0.443171
c1g11 Obs.	26.000000	10.000000	0.000000	4.000000
c1g11 Exp.	25.802631	7.231499	3.723904	3.241966
c1a11 Obs.	11.000000	5.000000	2.000000	1.000000
c1a11 Exp.	11.793986	3.922570	1.853654	1.429790

Other Calculations

T est (tau/Theta1)= 0.2605

******** SIMULATION RESULTS ********

Simulations attempted : 2000
simulations that yielded estimates : 1470 (0.7350)
with zero fixed differences : 213 (0.1065)
with zero shared differences : 263 (0.1315)

ChiSquare test statistic value : 37.8777
successful simulations with higher values : 1243 (0.8456)

wh test statistic value : 8.0000
successful simulations with higher values : 962 (0.6544)

Estimated parameter values and statistics

Parameter	Estimate	1.8% - 98.2%		Mean	Variance
Theta1	11.144	0.002	231.014	305.559	100151034.45884
Theta2	3.540	0.001	16.419	4.297	21.69667
ThetaA	36.531	5.931	78.145	32.619	346.10254
tau	2.903	0.001	11.111	3.445	8.23150
T (tau/theta1)	0.261	0.017	0.766	0.266	0.03028
c109 frac	0.143	0.000	0.308	0.144	0.00416
c110 frac	0.112	0.000	0.268	0.116	0.00309
c37 frac	0.114	0.000	0.267	0.118	0.00320
c34 frac	0.057	0.000	0.152	0.062	0.00121
c1g11 frac	0.393	0.000	0.638	0.375	0.01286
c1a11 frac	0.182	0.000	0.369	0.185	0.00586

Compare actual and simulated polymorphism levels

Locus	Sx1	Sx2	Ss	Sf
c109 Obs.	10.000000	1.000000	4.000000	0.000000
c109 Sim.	8.640136	3.412245	1.734694	1.164626
c110 Obs.	7.000000	1.000000	4.000000	0.000000
c110 Sim.	6.845578	2.899320	1.378912	0.931293
c37 Obs.	9.000000	0.000000	0.000000	3.000000
c37 Sim.	7.429932	2.457143	1.302721	0.919048
c34 Obs.	2.000000	4.000000	0.000000	0.000000
c34 Sim.	3.598639	1.347619	0.729932	0.436054

c1g11 Obs.	26.000000	10.000000	0.000000	4.000000
c1g11 Sim.	24.580272	7.826531	3.987075	3.256463
c1a11 Obs.	11.000000	5.000000	2.000000	1.000000
c1a11 Sim.	11.303401	4.208163	2.258503	1.604082

Appendix 4- MLHKA Test Output Files

Neutral Outfile

ML T theta(c37latifolia) k(c37latifolia) theta(c37dioica) k(c37dioica) theta(c109latifolia) k(c109latifolia) theta(c109dioica) k(c109dioica) theta(c1f6latifolia) k(c1f6latifolia) theta(c2d5latifolia) k(c2d5latifolia) theta(c2d5dioica) k(c2d5dioica) theta(c18latifolia) k(c18latifolia) theta(c18dioica) k(c18dioica) theta(c158latifolia) k(c158latifolia) theta(c158dioica) k(c158dioica) theta(c134latifolia) k(c134latifolia) theta(c134dioica) k(c134dioica) theta(c34latifolia) k(c34latifolia) theta(c34dioica) k(c34dioica) theta(c79latifolia) k(c79latifolia) theta(c79dioica) k(c79dioica) theta(c1a8latifolia) k(c1a8latifolia) theta(c1a8dioica) k(c1a8dioica) theta(c3a4latifolia) k(c3a4latifolia) theta(c3a4dioica) k(c3a4dioica) theta(c1a11latifolia) k(c1a11latifolia) theta(c1a11dioica) k(c1a11dioica) theta(c2c4latifolia) k(c2c4latifolia) theta(c2c4dioica) k(c2c4dioica) theta(c1g11latifolia) k(c1g11latifolia) theta(c1g11dioica) k(c1g11dioica)
-138.924 1.16615 0.0124509 1 0.0112975 1 0.0142255 1 0.0134895 1 0.00224859 1 0.00429045 1 0.00327515 1 0.0167207 1 0.0206236 1 0.00229754 1 0.00183967 1 0.00311933 1 0.00384892 1 0.0038165 1 0.00375348 1 0.00544057 1 0.00869299 1 0.00415961 1 0.00424766 1 0.025507 1 0.0196147 1 0.00723477 1 0.0106371 1 0.00589001 1 0.00549905 1 0.0053748 1 0.00480681 1

C1A8 (Both species selected) Outfile

Testing for departure from neutrality at c1a8latifolia, c1a8dioica,
ML T theta(c37latifolia) k(c37latifolia) theta(c37dioica) k(c37dioica) theta(c109latifolia) k(c109latifolia) theta(c109dioica) k(c109dioica) theta(c1f6latifolia) k(c1f6latifolia) theta(c2d5latifolia) k(c2d5latifolia) theta(c2d5dioica) k(c2d5dioica) theta(c18latifolia) k(c18latifolia) theta(c18dioica) k(c18dioica) theta(c158latifolia) k(c158latifolia) theta(c158dioica) k(c158dioica) theta(c134latifolia) k(c134latifolia) theta(c134dioica) k(c134dioica) theta(c34latifolia) k(c34latifolia) theta(c34dioica) k(c34dioica) theta(c79latifolia) k(c79latifolia) theta(c79dioica) k(c79dioica) theta(c1a8latifolia) k(c1a8latifolia) theta(c1a8dioica) k(c1a8dioica) theta(c3a4latifolia) k(c3a4latifolia) theta(c3a4dioica) k(c3a4dioica) theta(c1a11latifolia) k(c1a11latifolia) theta(c1a11dioica) k(c1a11dioica) theta(c2c4latifolia) k(c2c4latifolia) theta(c2c4dioica) k(c2c4dioica) theta(c1g11latifolia) k(c1g11latifolia) theta(c1g11dioica) k(c1g11dioica)
-131.073 1.09433 0.0124439 1 0.0126873 1 0.0155504 1 0.0143 1 0.0027538 1 0.00466686 1 0.004 1 0.0176366 1 0.0222386 1 0.0023 1 0.0015 1 0.0028862 1 0.00338696 1 0.0043 1 0.0043 1 0.00573567 1 0.0104 1 0.0136702 0.135972 0.0131209 0.0667168 0.0228 1 0.0208865 1 0.00632039 1 0.0126763 1 0.0053015 1 0.00609857 1 0.0050006 1 0.00552984 1

C1A8 (Latifolia selected) Outfile

Testing for departure from neutrality at c1a8latifolia,
ML T theta(c37latifolia) k(c37latifolia) theta(c37dioica) k(c37dioica) theta(c109latifolia) k(c109latifolia) theta(c109dioica) k(c109dioica) theta(c1f6latifolia) k(c1f6latifolia) theta(c2d5latifolia) k(c2d5latifolia) theta(c2d5dioica) k(c2d5dioica) theta(c18latifolia) k(c18latifolia) theta(c18dioica) k(c18dioica) theta(c158latifolia) k(c158latifolia) theta(c158dioica) k(c158dioica) theta(c134latifolia) k(c134latifolia) theta(c134dioica) k(c134dioica) theta(c34latifolia) k(c34latifolia) theta(c34dioica) k(c34dioica) theta(c79latifolia) k(c79latifolia) theta(c79dioica) k(c79dioica) theta(c1a8latifolia) k(c1a8latifolia) theta(c1a8dioica) k(c1a8dioica) theta(c3a4latifolia) k(c3a4latifolia) theta(c3a4dioica) k(c3a4dioica) theta(c1a11latifolia) k(c1a11latifolia) theta(c1a11dioica) k(c1a11dioica) theta(c2c4latifolia) k(c2c4latifolia) theta(c2c4dioica) k(c2c4dioica) theta(c1g11latifolia) k(c1g11latifolia) theta(c1g11dioica) k(c1g11dioica)
-134.765 1.08717 0.0131728 1 0.0101768 1 0.0138413 1 0.0144277 1 0.00255533 1 0.00435593 1 0.004 1 0.0190919 1 0.0190785 1 0.00238773 1 0.00180639 1 0.00306209 1 0.00347544 1 0.0043 1 0.00334965 1 0.0051 1 0.00968951 1 0.0141017 0.118773 0.00439011 1 0.0249699 1 0.0234503 1 0.00604435 1 0.0109297 1 0.00640775 1 0.00569046 1 0.0066 1 0.0054 1

C1A8 (Dioica selected) Outfile

Testing for departure from neutrality at c1a8dioica,
ML T theta(c37latifolia) k(c37latifolia) theta(c37dioica) k(c37dioica) theta(c109latifolia) k(c109latifolia) theta(c109dioica) k(c109dioica) theta(c1f6latifolia) k(c1f6latifolia) theta(c2d5latifolia) k(c2d5latifolia) theta(c2d5dioica) k(c2d5dioica) theta(c18latifolia) k(c18latifolia) theta(c18dioica) k(c18dioica) theta(c158latifolia) k(c158latifolia) theta(c158dioica) k(c158dioica) theta(c134latifolia) k(c134latifolia) theta(c134dioica) k(c134dioica) theta(c34latifolia) k(c34latifolia) theta(c34dioica) k(c34dioica) theta(c79latifolia) k(c79latifolia) theta(c79dioica) k(c79dioica) theta(c1a8latifolia) k(c1a8latifolia) theta(c1a8dioica) k(c1a8dioica) theta(c3a4latifolia) k(c3a4latifolia) theta(c3a4dioica) k(c3a4dioica) theta(c1a11latifolia) k(c1a11latifolia) theta(c1a11dioica) k(c1a11dioica) theta(c2c4latifolia) k(c2c4latifolia) theta(c2c4dioica) k(c2c4dioica) theta(c1g11latifolia) k(c1g11latifolia) theta(c1g11dioica) k(c1g11dioica)

-134.949 0.993884 0.0120485 1 0.010854 1 0.0153623 1 0.0149045 1 0.0029643 1 0.0054 1 0.004 1 0.0187713 1 0.0216608 1 0.0023 1 0.0017031 1 0.0037 1 0.00349147 1 0.00383013 1 0.00385006 1 0.00545309 1 0.00936448 1 0.00437403 1 0.0171678 0.100163 0.0237168 1 0.0213657 1 0.00801146 1 0.0101015 1 0.00552173 1 0.00616758 1 0.00666244 1 0.00491011 1

Appendix 5 - Structure Output Files

S. latifolia and *S. dioica* dataset

K=1 Run 1

Command line arguments: bin\structure.exe -m C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
1 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined
population in each of the 1 clusters

Given Pop	Inferred Clusters 1	Number of Individuals
1:	1.000	15
2:	1.000	14

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1
1	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -61.5534

Estimated Ln Prob of Data = -790.2
Mean value of ln likelihood = -765.9
Variance of ln likelihood = 48.8

Mean value of lambda = 1.9001
Allele frequencies uncorrelated

Run 2

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
1 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined

population in each of the 1 clusters

Given Inferred Clusters Number of
Pop 1 Individuals

1: 1.000 15
2: 1.000 14

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

1
1 -

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -61.5482

Estimated Ln Prob of Data = -790.2
Mean value of ln likelihood = -765.8
Variance of ln likelihood = 48.7

Mean value of lambda = 1.9007
Allele frequencies uncorrelated

Run 3

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
1 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined
population in each of the 1 clusters

Given Inferred Clusters Number of
Pop 1 Individuals

1: 1.000 15
2: 1.000 14

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

1
1 -

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -61.5575

Estimated Ln Prob of Data = -790.4
Mean value of ln likelihood = -765.8
Variance of ln likelihood = 49.0

Mean value of lambda = 1.8994
Allele frequencies uncorrelated

K=2 Run 1

Command line arguments: bin\structure.exe -m C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
2 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined population in each of the 2 clusters

Given Pop	Inferred Clusters 1	2	Number of Individuals
1:	0.028	0.972	15
2:	0.960	0.040	14

Allele-freq. divergence among pops (Net nucleotide distance), computed using point estimates of P.

	1	2
1	-	29.8357
2	29.8357	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -84.9578
cluster 2 : -77.4125

Estimated Ln Prob of Data = -777.0
Mean value of ln likelihood = -659.8
Variance of ln likelihood = 234.4
Mean value of alpha = 0.0684

Mean value of lambda = 0.5260
Allele frequencies uncorrelated

Run 2

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
2 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined population in each of the 2 clusters

Given Inferred Clusters Number of
Pop 1 2 Individuals

1: 0.979 0.021 15
2: 0.029 0.971 14

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

1 2
1 - 29.6131
2 29.6131 -

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -76.2245
cluster 2 : -84.4761

Estimated Ln Prob of Data = -714.1
Mean value of ln likelihood = -648.4
Variance of ln likelihood = 131.5
Mean value of alpha = 0.0544

Mean value of lambda = 0.6662
Allele frequencies uncorrelated

Run 3

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
2 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined
population in each of the 2 clusters

Given Inferred Clusters Number of
Pop 1 2 Individuals

1: 0.021 0.979 15
2: 0.971 0.029 14

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

1 2
1 - 29.5984
2 29.5984 -

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -84.4560
cluster 2 : -76.2217

Estimated Ln Prob of Data = -714.2
Mean value of ln likelihood = -648.3
Variance of ln likelihood = 131.7
Mean value of alpha = 0.0536

Mean value of lambda = 0.6663
Allele frequencies uncorrelated

K=3 Run 1

Command line arguments: bin\structure.exe -m C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
3 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined
population in each of the 3 clusters

Given Pop	Inferred Clusters 1	2	3	Number of Individuals
1:	0.460	0.038	0.502	15
2:	0.072	0.857	0.071	14

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2	3
1	-	25.9233	0.0787
2	25.9233	-	26.9192
3	0.0787	26.9192	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -56.2009
cluster 2 : -92.1768
cluster 3 : -59.1715

Estimated Ln Prob of Data = -854.3
Mean value of ln likelihood = -654.3
Variance of ln likelihood = 400.0
Mean value of alpha = 0.1277

Mean value of lambda = 0.4127
Allele frequencies uncorrelated

Run 2

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
3 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined

population in each of the 3 clusters

Given Inferred Clusters Number of
Pop 1 2 3 Individuals

1: 0.275 0.474 0.251 15
2: 0.480 0.048 0.472 14

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

1 2 3
1 - 5.4713 0.0193
2 5.4713 - 5.4085
3 0.0193 5.4085 -

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -60.4532
cluster 2 : -55.2197
cluster 3 : -59.0918

Estimated Ln Prob of Data = -760.8
Mean value of ln likelihood = -640.6
Variance of ln likelihood = 240.4
Mean value of alpha = 0.0622

Mean value of lambda = 0.6346
Allele frequencies uncorrelated

Run 3

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
3 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined
population in each of the 3 clusters

Given Inferred Clusters Number of
Pop 1 2 3 Individuals

1: 0.161 0.479 0.361 15
2: 0.625 0.046 0.329 14

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

1 2 3
1 - 10.2771 2.9274
2 10.2771 - 2.3156
3 2.9274 2.3156 -

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -65.4781
cluster 2 : -55.5275
cluster 3 : -57.0407

Estimated Ln Prob of Data = -759.8
Mean value of ln likelihood = -640.3
Variance of ln likelihood = 239.1
Mean value of alpha = 0.0618

Mean value of lambda = 0.6330
Allele frequencies uncorrelated

K=4 Run 1

Command line arguments: bin\structure.exe -m C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
4 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined population in each of the 4 clusters

Given Pop	Inferred Clusters 1	2	3	4	Number of Individuals
1:	0.287	0.243	0.133	0.338	15
2:	0.367	0.061	0.510	0.062	14

Allele-freq. divergence among pops (Net nucleotide distance), computed using point estimates of P.

	1	2	3	4
1	-	4.3540	1.5046	3.6469
2	4.3540	-	7.9296	0.4679
3	1.5046	7.9296	-	8.3412
4	3.6469	0.4679	8.3412	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -58.2932
cluster 2 : -44.5032
cluster 3 : -61.5317
cluster 4 : -48.9575

Estimated Ln Prob of Data = -829.7
Mean value of ln likelihood = -666.2
Variance of ln likelihood = 327.1
Mean value of alpha = 0.1187

Mean value of lambda = 0.4473
Allele frequencies uncorrelated

Run 2

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
4 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined
population in each of the 4 clusters

Given Pop	Inferred Clusters 1	2	3	4	Number of Individuals
1:	0.190	0.223	0.301	0.286	15
2:	0.453	0.215	0.041	0.291	14

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2	3	4
1	-	1.5165	4.5312	0.8117
2	1.5165	-	0.8530	0.3532
3	4.5312	0.8530	-	1.7294
4	0.8117	0.3532	1.7294	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -55.8331
cluster 2 : -47.2656
cluster 3 : -46.4979
cluster 4 : -52.0692

Estimated Ln Prob of Data = -760.2
Mean value of ln likelihood = -648.1
Variance of ln likelihood = 224.2
Mean value of alpha = 0.0530

Mean value of lambda = 0.7274
Allele frequencies uncorrelated

Run 3

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
4 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined
population in each of the 4 clusters

Given Pop	Inferred Clusters 1	2	3	4	Number of Individuals
1:	0.377	0.117	0.285	0.222	15
2:	0.041	0.601	0.230	0.128	14

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2	3	4

1	-	8.7911	1.0000	0.7412
2	8.7911	-	3.8637	5.9003
3	1.0000	3.8637	-	0.5061
4	0.7412	5.9003	0.5061	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -49.5283
cluster 2 : -62.4503
cluster 3 : -49.8747
cluster 4 : -45.1870

Estimated Ln Prob of Data = -757.1
Mean value of ln likelihood = -648.3
Variance of ln likelihood = 217.7
Mean value of alpha = 0.0534

Mean value of lambda = 0.7289
Allele frequencies uncorrelated

K=5 Run 1

Command line arguments: bin\structure.exe -m C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\alh571\Andrea's Documents\Work\Analysis\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
5 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined population in each of the 5 clusters

Given Pop	Inferred Clusters 1	2	3	4	5	Number of Individuals
1:	0.390	0.195	0.130	0.226	0.060	15
2:	0.052	0.048	0.206	0.049	0.645	14

Allele-freq. divergence among pops (Net nucleotide distance), computed using point estimates of P.

	1	2	3	4	5
1	-	1.7699	3.1377	1.2387	14.8554
2	1.7699	-	0.9694	0.0476	13.4994
3	3.1377	0.9694	-	1.0727	7.3723
4	1.2387	0.0476	1.0727	-	13.4446
5	14.8554	13.4994	7.3723	13.4446	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -52.0936
cluster 2 : -43.0476
cluster 3 : -44.6372
cluster 4 : -44.0426
cluster 5 : -71.6907

Estimated Ln Prob of Data = -831.5
Mean value of ln likelihood = -672.0
Variance of ln likelihood = 319.0
Mean value of alpha = 0.0950

Mean value of lambda = 0.4740
Allele frequencies uncorrelated

Run 2

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
5 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined
population in each of the 5 clusters

Given Pop	Inferred Clusters 1	2	3	4	5	Number of Individuals
1:	0.181	0.082	0.187	0.246	0.304	15
2:	0.276	0.619	0.034	0.035	0.036	14

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2	3	4	5
1	-	3.0315	1.5682	1.4897	1.6061
2	3.0315	-	8.7569	8.7681	8.9392
3	1.5682	8.7569	-	0.1135	0.4345
4	1.4897	8.7681	0.1135	-	0.1048
5	1.6061	8.9392	0.4345	0.1048	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -47.7855
cluster 2 : -62.5694
cluster 3 : -42.9325
cluster 4 : -44.4225
cluster 5 : -46.3220

Estimated Ln Prob of Data = -755.6
Mean value of ln likelihood = -655.3
Variance of ln likelihood = 200.6
Mean value of alpha = 0.0466

Mean value of lambda = 0.8081
Allele frequencies uncorrelated

Run 3

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\LatDioAdmix3\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Autosomals\Structure\LatDioAdmix3\project_data

Run parameters:
29 individuals
18 loci
5 populations assumed
100000 Burn-in period
500000 Reps

Proportion of membership of each pre-defined
population in each of the 5 clusters

Given Pop	Inferred Clusters 1	2	3	4	5	Number of Individuals
1:	0.151	0.332	0.281	0.201	0.035	15
2:	0.034	0.038	0.068	0.034	0.826	14

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2	3	4	5
1	-	1.0204	0.5887	0.0745	16.2259
2	1.0204	-	0.0948	0.5446	16.2893
3	0.5887	0.0948	-	0.2552	14.8394
4	0.0745	0.5446	0.2552	-	16.0581
5	16.2259	16.2893	14.8394	16.0581	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -42.2648
cluster 2 : -47.4609
cluster 3 : -45.8699
cluster 4 : -43.2642
cluster 5 : -77.6847

Estimated Ln Prob of Data = -755.5
Mean value of ln likelihood = -655.0
Variance of ln likelihood = 201.1
Mean value of alpha = 0.0461

Mean value of lambda = 0.8056
Allele frequencies uncorrelated

3 Species dataset

K=1 Run 1

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
78 individuals
6 loci
1 populations assumed
100000 Burn-in period
500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
population in each of the 1 clusters

Given Pop	Inferred Clusters 1	Number of Individuals
1:	1.000	44
2:	1.000	19
3:	1.000	15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1
1	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -109.0172

Estimated Ln Prob of Data = -453.7
Mean value of ln likelihood = -443.9
Variance of ln likelihood = 19.5

Mean value of lambda = 0.6352
Allele frequencies uncorrelated

Run 2

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
78 individuals
6 loci
1 populations assumed
100000 Burn-in period
500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
 population in each of the 1 clusters

Given Inferred Clusters Number of
 Pop 1 Individuals

 1: 1.000 44
 2: 1.000 19
 3: 1.000 15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

  1
1 -

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -109.0132

Estimated Ln Prob of Data = -453.7
Mean value of ln likelihood = -443.9
Variance of ln likelihood = 19.6

Mean value of lambda = 0.6350
Allele frequencies uncorrelated

Run 3

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
  78 individuals
  6 loci
  1 populations assumed
  100000 Burn-in period
  500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
 population in each of the 1 clusters

Given Inferred Clusters Number of
 Pop 1 Individuals

 1: 1.000 44
 2: 1.000 19
 3: 1.000 15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

  1
1 -

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -108.9980

Estimated Ln Prob of Data = -453.7
Mean value of ln likelihood = -443.9
Variance of ln likelihood = 19.5

Mean value of lambda = 0.6352
Allele frequencies uncorrelated

K=2 Run 1

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
78 individuals
6 loci
2 populations assumed
100000 Burn-in period
500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
population in each of the 2 clusters

Given Pop	Inferred Clusters 1	2	Number of Individuals
1:	0.984	0.016	44
2:	0.404	0.596	19
3:	0.024	0.976	15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2
1	-	78.6088
2	78.6088	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -187.6945
cluster 2 : -89.3396

Estimated Ln Prob of Data = -354.2
Mean value of ln likelihood = -330.9
Variance of ln likelihood = 46.6

Mean value of lambda = 0.2934
Allele frequencies uncorrelated

Run 2

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
78 individuals
6 loci
2 populations assumed
100000 Burn-in period
500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
population in each of the 2 clusters

Given Pop	Inferred Clusters 1	2	Number of Individuals
1:	0.016	0.984	44
2:	0.597	0.403	19
3:	0.976	0.024	15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2
1	-	78.6418
2	78.6418	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -89.3857
cluster 2 : -187.8083

Estimated Ln Prob of Data = -354.3
Mean value of ln likelihood = -330.9
Variance of ln likelihood = 46.7

Mean value of lambda = 0.2929
Allele frequencies uncorrelated

Run 3

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
78 individuals
6 loci
2 populations assumed
100000 Burn-in period
500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
population in each of the 2 clusters

Given Pop	Inferred Clusters 1	2	Number of Individuals
1:	0.016	0.984	44
2:	0.597	0.403	19
3:	0.976	0.024	15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2
1	-	78.6254
2	78.6254	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -89.3739
cluster 2 : -187.7895

Estimated Ln Prob of Data = -354.1
Mean value of ln likelihood = -330.9
Variance of ln likelihood = 46.4

Mean value of lambda = 0.2928
Allele frequencies uncorrelated

K=3 Run 1

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
78 individuals
6 loci
3 populations assumed
100000 Burn-in period
500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
population in each of the 3 clusters

Given Pop	Inferred Clusters 1	2	3	Number of Individuals
1:	0.940	0.037	0.023	44
2:	0.143	0.547	0.311	19
3:	0.007	0.343	0.650	15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2	3
1	-	63.1714	81.2546
2	63.1714	-	5.3573
3	81.2546	5.3573	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -205.9619
cluster 2 : -83.2573
cluster 3 : -86.1940

Estimated Ln Prob of Data = -324.2
Mean value of ln likelihood = -287.6
Variance of ln likelihood = 73.1

Mean value of lambda = 0.2157
Allele frequencies uncorrelated

Run 2

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
78 individuals
6 loci
3 populations assumed
100000 Burn-in period
500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
 population in each of the 3 clusters

Given Inferred Clusters Number of
 Pop 1 2 3 Individuals

 1: 0.022 0.939 0.039 44
 2: 0.293 0.142 0.565 19
 3: 0.673 0.007 0.320 15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

  1 2 3
1 - 82.9458 7.0966
2 82.9458 - 62.2188
3 7.0966 62.2188 -

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -87.2458
cluster 2 : -205.8952
cluster 3 : -83.9010

Estimated Ln Prob of Data = -324.0
Mean value of ln likelihood = -287.7
Variance of ln likelihood = 72.8

Mean value of lambda = 0.2162
Allele frequencies uncorrelated

Run 3

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
  78 individuals
  6 loci
  3 populations assumed
  100000 Burn-in period
  500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
 population in each of the 3 clusters

Given Inferred Clusters Number of
 Pop 1 2 3 Individuals

 1: 0.024 0.037 0.939 44
 2: 0.327 0.531 0.142 19
 3: 0.629 0.364 0.007 15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2	3
1	-	3.9973	79.6271
2	3.9973	-	63.9657
3	79.6271	63.9657	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -85.2866
cluster 2 : -82.7275
cluster 3 : -205.8963

Estimated Ln Prob of Data = -323.8
Mean value of ln likelihood = -287.7
Variance of ln likelihood = 72.3

Mean value of lambda = 0.2168
Allele frequencies uncorrelated

K=4 Run 1

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
78 individuals
6 loci
4 populations assumed
100000 Burn-in period
500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
population in each of the 4 clusters

Given Pop	Inferred Clusters 1	2	3	4	Number of Individuals
1:	0.251	0.224	0.289	0.235	44
2:	0.264	0.261	0.220	0.255	19
3:	0.223	0.256	0.264	0.258	15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2	3	4
1	-	0.1849	0.2787	0.1563
2	0.1849	-	0.5720	0.0082
3	0.2787	0.5720	-	0.4718
4	0.1563	0.0082	0.4718	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -89.8240
cluster 2 : -85.6651
cluster 3 : -94.3053
cluster 4 : -86.2940

Estimated Ln Prob of Data = -336.9
Mean value of ln likelihood = -269.7
Variance of ln likelihood = 134.5

Mean value of lambda = 0.1860
Allele frequencies uncorrelated

Run2

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
78 individuals
6 loci
4 populations assumed
100000 Burn-in period
500000 Reps

NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
population in each of the 4 clusters

Given Pop	Inferred Clusters 1	2	3	4	Number of Individuals
1:	0.270	0.234	0.207	0.290	44
2:	0.254	0.263	0.234	0.249	19
3:	0.221	0.242	0.327	0.210	15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2	3	4
1	-	0.2640	1.2187	0.0501
2	0.2640	-	0.5194	0.4175
3	1.2187	0.5194	-	1.5886
4	0.0501	0.4175	1.5886	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -92.3348
cluster 2 : -86.5214
cluster 3 : -83.8329
cluster 4 : -93.6895

Estimated Ln Prob of Data = -337.5
Mean value of ln likelihood = -270.0
Variance of ln likelihood = 135.0

Mean value of lambda = 0.1877
Allele frequencies uncorrelated

Run 3

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
78 individuals
6 loci
4 populations assumed
100000 Burn-in period
500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
population in each of the 4 clusters

Given Pop	Inferred Clusters 1	2	3	4	Number of Individuals
1:	0.226	0.242	0.283	0.250	44
2:	0.265	0.253	0.227	0.255	19
3:	0.246	0.253	0.260	0.241	15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2	3	4
1	-	0.0480	0.4919	0.1469
2	0.0480	-	0.2357	0.0395
3	0.4919	0.2357	-	0.1261
4	0.1469	0.0395	0.1261	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -85.0667
cluster 2 : -87.2757
cluster 3 : -93.1506
cluster 4 : -89.3168

Estimated Ln Prob of Data = -337.9
Mean value of ln likelihood = -269.8
Variance of ln likelihood = 136.3

Mean value of lambda = 0.1865
Allele frequencies uncorrelated

K=5 Run 1

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
78 individuals
6 loci
5 populations assumed
100000 Burn-in period
500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
population in each of the 5 clusters

Given Pop	Inferred Clusters 1	2	3	4	5	Number of Individuals
1:	0.198	0.204	0.218	0.196	0.184	44
2:	0.203	0.192	0.199	0.208	0.197	19
3:	0.205	0.206	0.181	0.192	0.215	15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2	3	4	5
1	-	0.0146	0.0996	0.0142	0.0584
2	0.0146	-	0.0806	0.0303	0.0652
3	0.0996	0.0806	-	0.0600	0.2647
4	0.0142	0.0303	0.0600	-	0.1072
5	0.0584	0.0652	0.2647	0.1072	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -77.0083
cluster 2 : -76.4642
cluster 3 : -77.8337
cluster 4 : -77.0706
cluster 5 : -74.3783

Estimated Ln Prob of Data = -366.0
Mean value of ln likelihood = -265.3
Variance of ln likelihood = 201.5

Mean value of lambda = 0.2056
Allele frequencies uncorrelated

Run 2

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
78 individuals
6 loci
5 populations assumed

100000 Burn-in period
500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
population in each of the 5 clusters

Given Pop	Inferred Clusters 1	2	3	4	5	Number of Individuals
1:	0.197	0.199	0.187	0.210	0.207	44
2:	0.196	0.204	0.201	0.194	0.205	19
3:	0.209	0.195	0.210	0.199	0.187	15

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

	1	2	3	4	5
1	-	0.0255	0.0197	0.0282	0.0664
2	0.0255	-	0.0469	0.0187	0.0206
3	0.0197	0.0469	-	0.0821	0.1228
4	0.0282	0.0187	0.0821	-	0.0175
5	0.0664	0.0206	0.1228	0.0175	-

Average distances (expected heterozygosity) between individuals in same cluster:
cluster 1 : -76.6049
cluster 2 : -76.1959
cluster 3 : -75.0847
cluster 4 : -77.2561
cluster 5 : -77.6561

Estimated Ln Prob of Data = -366.4
Mean value of ln likelihood = -265.5
Variance of ln likelihood = 201.8

Mean value of lambda = 0.2067
Allele frequencies uncorrelated

Run 3

Command line arguments: bin\structure.exe -m C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\mainparams -e C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\noadmixindfinal\extraparams
Input File: C:\Documents and Settings\Andrea Harper\My Documents\Work\Latest\Analyses\Analysis\Diclinis\Structure\Spp 78\project_data

Run parameters:
78 individuals
6 loci
5 populations assumed
100000 Burn-in period
500000 Reps
NO ADMIXTURE model assumed

Proportion of membership of each pre-defined
population in each of the 5 clusters

Given Pop	Inferred Clusters 1	2	3	4	5	Number of Individuals

```
 1:    0.203  0.195  0.220  0.191  0.190       44
 2:    0.198  0.200  0.196  0.206  0.199       19
 3:    0.198  0.209  0.186  0.197  0.210       15
--------------------------------------------------

Allele-freq. divergence among pops (Net nucleotide distance),
computed using point estimates of P.

     1      2      3      4      5
1     -   0.0152 0.1151 0.0141 0.0237
2  0.0152    -   0.1926 0.0152 0.0109
3  0.1151 0.1926    -   0.1720 0.2069
4  0.0141 0.0152 0.1720    -   0.0118
5  0.0237 0.0109 0.2069 0.0118    -

Average distances (expected heterozygosity) between individuals in same cluster:
cluster  1  : -76.1470
cluster  2  : -76.1774
cluster  3  : -79.5879
cluster  4  : -75.4949
cluster  5  : -75.5056

--------------------------------------------------
Estimated Ln Prob of Data   = -365.4
Mean value of ln likelihood = -265.4
Variance of ln likelihood   = 200.0

Mean value of lambda        = 0.2053
Allele frequencies uncorrelated
```

Appendix 6 – Bottleneck Output File

File: C:\Program Files\Bottleneck\3.txt
Data type: Heterozygozity
Title: Haplotype 5 loci Sdilcinis
Estimation based on 1000 replications.

Date: 25/03/2009 Time: 15:03:08.
Population : Haplotype 5 loci Sdilcinis

	observed		under the T.P.M.			
locus	n	ko He	Heq	S.D.	DH/sd	Prob
Loc1	43	3 0.215	0.430	0.160	-1.347	0.1450
Loc2	38	5 0.333	0.657	0.099	-3.275	0.0140
Loc3	18	5 0.614	0.724	0.076	-1.445	0.1090
Loc4	50	15 0.858	0.902	0.025	-1.790	0.0470
Loc5	45	13 0.851	0.886	0.029	-1.201	0.1000

SIGN TEST
Assumptions: all loci fit T.P.M., mutation-drift equilibrium.
Expected number of loci with heterozygosity excess: 2.93
5 loci with heterozygosity deficiency and 0 loci with heterozygosity excess.
Probability: 0.01225

STANDARDIZED DIFFERENCES TEST
Caution: only 5 polymorphic loci (minimum 20).

Assumptions: all loci fit T.P.M., mutation-drift equilibrium.
T2: -4.051 Probability: 0.00003

WILCOXON TEST
Assumptions: all loci fit T.P.M., mutation-drift equilibrium.
Probability (one tail for H deficiency): 0.01563
Probability (one tail for H excess): 1.00000
Probability (two tails for H excess or deficiency): 0.03125

REFERENCES

Abbott, R. J., F. C. Bretagnolle and C Thebaud (1998). “Evolution of a polymorphism for outcrossing rate in *Senecio vulgaris*: Influence of germination behavior.” Evolution **52**: 1593-1601.

Abbott, R. J. and A. J. Lowe (2004). “Origins, establishment and evolution of new polyploidy species: *Senecio cambrensis* and *S. eboracensis* in the British Isles.” Biological Journal of the Linnean Society **82**: 467-474.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." Journal of Molecular Biology **215**(3): 403-410.

Atanassov, I., Delichere, C., Filatov, D. A., Charlesworth, D., Negrutiu, I., Moneger, F. (2001) “Analysis and evolution of two functional Y-linked loci in a plant sex chromosome system” Mol Biol Evol **18** (12); 2162-8.

Bachtrog, D. (2003) “Accumulation of Spock and Worf, two novel non-LTR retrotransposons, on the neo-Y chromosome of *Drosophila miranda*” Mol. Biol. Evol. **20**; 173-181.

Baker, H. G. (1950). "The inheritance of certain characters in crosses between *Melandrium dioicum* and *M. album*." Genetica **25**: 126-156.

Barr, M. L. (1959) “Sex-chromatin and phenotype in man” Science **130**; 679-685.

Barrett, S. C. H. and J. R. Kohn. (1991). Genetics and Evolutionary Consequences of Small Population Size in Plants: Implications for Conservation. Genetics and Conservation of Rare Plants. D. A. Falk and K. E. Holsinger, Oxford University Press.

Beardmore, J. A. (1983). Extinction, survival, and genetic variation. Genetics and Conservation: A Reference for Managing Wild Animal and Plant Populations. C. M. Schonewald-Cox, S. M. Chambers, B. MacBryde and W. L. Thomas. Menlo Park, California, Benjamin/Cummings**:** 125-151.

Bergero, R., A. Forrest, E. Kamau and D. Charlesworth (2007). "Evolutionary strata on the X chromosome of the dioecious plant *Silene latifolia*: evidence from new sex-linked genes." Genetics **175**(4): 1945-1954.

Brown, S. W., and Chandra H. S. (1977) “Chromosome imprinting and their differential regulation of homologous chromosomes” Cell Biology **1**; 109-189.

Brumfield, R. T., P. Beerli, D. A. Nickerson and S. V. Edwards (2003). "The utility of single nucleotide polymorphisms in inferences of population history." Trends in Ecology and Evolution **18**(5): 249-256.

Bull, J. (1983) Evolution of sex determining mechanisms, The Benjamin/Cummings Publishing Company, Inc.

Cattenach, B.M. (1975) “Control of chromosome inactivation” Ann. Rev. Gen **9**; 1-18.

Charlesworth, B. (2003) “The organization and evolution of the human Y chromosome” Genome Biol. **4** (9); 226-228.

Charlesworth, B. and Charlesworth, D. (2000) “The degeneration of Y chromosomes” Phil. Trans. R. Soc. Lond. Ser. B **355**; 1563-1572.

Charlesworth, B., J. A. Coyne and N. H. Barton (1987). "The relative rates of evolution of sex chromosomes and autosomes." The American Naturalist **130**: 113-146.

Charlesworth, B. and J. D. Wall (1999). "Inbreeding, heterozygote advantage and the evolution of neo-X and neo-Y sex chromsomes." Proceedings of the Royal Society of London. Series B **266**: 51-56.

Charlesworth, D. and B. Charlesworth (1980). "Sex differences in fitness and selection for centric fusions between the sex chromosomes." Genetical Research **35**: 205-214.

Cornuet, J. M. and G. Luikart (1996). "Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data." Genetics **144**(4): 2001-2014.

Delichere, C., Veuskens, J., Hernould, M., Barbacar, N., Mouras, A., Negrutiu, I., Moneger, F. (1999) “SlY1, the first active gene cloned from a plant Y chromosome, encodes a WD-repeat protein” Embo J **18** (15); 4169-79.

Double, M. C., R. Peakall, N. R. Beck and A. Cockburn (2005). "Dispersal, philopatry and infidelity: dissecting local genetic structure in superb fairy-wrens (*Malurus cyaneus*)." Evolution **59**: 625-635.

Drouin, G., H. Daoud and J. Xia (2008). "Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants." Molecular Phylogenetics and Evolution **49**(3): 827-831.

Edwards, S. V. and P. Beerli (2000). "Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies." Evolution **54**(6): 1839-1854.

Eloff. G. (1932) “A theoretical and experimental study on the changes in the crossing-over value, their causes and meaning” Genetica **14**; 1-116.

Emerson, B. C., E. Paradis and C. Thebaud (2001). "Revealing the demographic histories of species using DNA sequences." Trends in Ecology and Evolution **16**: 707-716.

Excoffier, L., P. E. Smouse and J. M. Quattro (1992). "Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction sites." Genetics131(479-491).

Fay, J. C. and C. I. Wu (2000). "Hitchhiking under positive Darwinian selection." Genetics **155**: 1405-1413.

Felsenstein, J. (2004). PHYLIP (Phylogeny Inference Package). V3.6. Department of Genome Sciences, University of Washington, Seattle, USA.

Filatov, D. A. (2002). "ProSeq: A software for preparation and evolutionary analysis of DNA sequence data sets." Molecular Ecology Notes **2**: 621-324.

Filatov, D. A. (2005a) “Substitution rates in a new *Silene latifolia* sex-linked gene, SlssX/Y” Molecular Biology and Evolution **22**; 402-408.

Filatov, D. A. (2005b) “Evolutionary history of *Silene latifolia* sex chromosomes revealed by genetic mapping of four genes” Genetics **170** (2); 975-9.

Filatov, D. A. and D. Charlesworth (2002). "Substitution rates in the X- and Y-linked genes of the plants *Silene latifolia* and *S. dioica*." Molecular Biology and Evolution **19**: 898-907.

Filatov, D. A., Laporte, V., Vitte C., and Charlesworth, D. (2001) “DNA diversity in sex-linked and autosomal genes of the plant species *Silene latifolia* and *Silene dioica*” Moleculat Biology and Evolution **18;** 1442–1454.

Filatov, D. A., Moneger, F., Negrutiu, I., and Charlesworth, D., (2000) “Low variability in a Y-linked plant gene and its implications for Y-chromosome evolution” Nature **404;** 388–390.

Fisher, R. A. (1930). The Genetical Theory of Natural Selection. Oxford, Oxford University Press.

Grant, S., Hunkirchen, B., Saedler, H. (1994) “Developmental differences between male and female flowers in the dioecious plant *Silene latifolia*” Plant Journal **6**; 1775-1787.

Guttmann, D. S. and Charlesworth, D. (1998) “An X-linked gene with a degenerate Y-linked homologue in a dioecious plant” Nature **393**; 263-266.

Hamilton, W. D. (1982). Pathogens as causes of genetic diversity in their host populations. Population Biology of Infectious Diseases. R. M. Anderson and R. M. May. Berlin, Springer**:** 269-303.

Harper, J. L. (1977). Population Biology of Plants. London, Academic Press.

Hedrick, P. W. (1999). "Perspective: highly variable loci and their interpretation in evolution and conservation." Evolution **53**: 313-318.

Henking, H., (1891) “Untersuchtungen uber die ersten Entwichlungsvorgange in die Eiern der Insecten. II. Uber spermatogenese und deren Beziehung Zur Entwicklung bei *Pyrrochoris apterus*” Zeit. Wiss. Zool. **51**; 685-786.

Hess, H. E., E. Landolt and R. Hirzel (1972). Flora das Schweiz. Basel, Birkhäuser verlag.

Hewitt, G. M. (1996). "Some genetic consequences of ice ages, and their role in divergence and speciation." Biological Journal of the Linnean Society **58**: 247-276.

Hey, J. and R. Nielsen (2004). "Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*." Genetics **167**(2): 747-760.

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

Hopper, S. D. and G. F. Moran (1981). "Bird pollination and the mating system of *Eucalyptus stoatei*." Australian Journal of Botany **29**: 625-638.

Howell, E. C., S. J. Armstrong and D. A. Filatov (in press). "Evolution of neo-sex chromosomes in *Silene diclinis*."

Hudson, R. R., M. Kreitman and M. Aguade (1987). "A test of neutral molecular evolution based on nucleotide data." Genetics **116**: 153-159.

Huff, D. R., R. Peakall and P. E. Smouse (1993). "RAPD variation within and among natural populations of outcrossing buffalograss *Buchloe dactyloides* (Nutt) Engelm." Theoretical and Applied Genetics **86**: 927-934.

Ironside, J. E. and D. A. Filatov (2005). "Extreme population structure and high interspecific divergence of the *Silene* Y chromosome." Genetics **171**(2): 705-713.

Jaarola, M., H. Tegelstrom and K. Fredga (1997). "A contact zone with noncoincident clines for sex-specific markers in the field vole (*Microtus agrestis*)." Evolution **51**: 241-249.

James, J. K. and Abbott R. J. (2005). "Recent, allopatric, homoploid hybrid speciation: The origin of *Senecio squalidus* (Asteraceae) in the British Isles from a hybrid zone on Mount Etna, Sicily." Evolution **59**: 2533-2547.

Jürgens, A., T. Witt and G. Gottsberger (1996). "Reproduction and pollination in Central European populations of *Silene* and *Saponaria* species." Botanica Acta: Berichte der Deutschen Botanischen Gesellschaft **109**: 316-324.

Karrenberg, S. and A. Favre (2008). "Genetic and ecological differentiation in the hybridizing campions *Silene dioica* and *S. latifolia*." Evolution **62**(4): 763-773.

Kejnovsky, E, Vrána, J., Matsunaga, S., , P., , J., Dolezel, J., Vyskot, B. (2001) "Localization of Male-Specifically Expressed *MROS* Genes of *Silene latifolia* by PCR on Flow-Sorted Sex Chromosomes and Autosomes" Genetics **158**; 1269-1277.

Kihara, H. and Ono, T. (1923) "Cytological studies on *Rumex L. I.* Chromosomes of *Rumex acetosa L.*" Bot. Mag. Tokyo **37**; 84-90.

Kim, M., M.-L. Cui, P. Cubas, A. Gillies, K. Lee, M. A. Chapman, R. J. Abbott and E. Coen (2008). "Regulatory genes control a key morphological and ecological trait transferred between species." Science **322** (5904): 1116-1119.

Kronforst, M. R. (2008). "Gene flow persists millions of years after speciation in Heliconius butterflies." BMC Evolutionary Biology **8**: 98.

Lacy, R. C. (1987). "Loss of genetic diversity from managed populations: Interacting effects of drift, mutation, immigration, selection and population subdivision." Conservation Biology **1**(2): 143-158.

Lahn, B. T. and D. C. Page (1999). "Four evolutionary strata on the human X chromosome." Science **286**(5441): 964-967.

Laporte, V., Filatov, D. A., Kamau, E., and Charlesworth, D. (2005) “Indirect evidence from DNA sequence diversity for genetic degeneration of the Y-chromosome in dioecious species of the plant *Silene*: S1Y4/SIX4 and DD44-X/DD44-Y gene pairs” J. Evol. Biol. **18;** 337–347.

Lawson-Handley, Ceplitis, H., Ellegren, H. (2004) “Evolutionary strata on the chicken Z chromosome: implications for sex chromosome evolution“ Genetics **167** (1); 367-76.

Lebel-Hardenack, S., Hauser, E., Law, T. F., Schmid, J., Grant, S. R. (2002) “Mapping of sex determination loci on the white campion (*Silene latifolia*) Y chromosome using amplified fragment length polymorphism” Genetics **160** (2); 717-25.

Lien, S., Szyda, J., Schechinger, B., Rappold, G., Arnheim, M. (2000) “Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping” Am. J. Hum. Genet. **66**; 557-566.

Lowry, D. B., R. C. Rockwood and J. H. Willis (2008). "Ecological reproductive isolation of coast and inland races of *Mimulus guttatus*." Evolution **62**(9): 2196-2214.

Mantel, N. (1967). "The detection of disease clustering and a generalized regression approach." Cancer Research **27**: 209-220.

Martin, N. H., A. C. Bouck and M. L. Arnold (2007). "The genetic architecture of reproductive isolation in Louisiana irises: flowering phenology." Genetics **175**(4): 1803-1812.

Martin, N. H., Y. Sapir and M. L. Arnold (2008). "The genetic architecture of reproductive isolation in Louisiana irises: pollination syndromes and pollinator preferences." Evolution **62**(4): 740-752.

Matsunaga, S., Isono, E., Kejnovsky, E., Vyskot, B., Dolezel, J., Kawano, S., Charlesworth, D. (2003) “Duplicative transfer of a MADS box gene to a plant Y chromosome” Molecular Biological Evolution **20** (7); 1062-1069.

Michalakis, Y. and L. Excoffier (1996). "A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci." Genetics **142**: 1061-1064.

Minder, A. M., C. Rothenbuehler and A. Widmer (2007). "Genetic structure of hybrid zones between *Silene latifolia* and *Silene dioica* (Caryophyllaceae): evidence for introgressive hybridization." Molecular Ecology **16**(12): 2504-2516.

Montesinos, D. and J. Güemes (2006). "*Silene diclinis*." IUCN 2008 Red List of Threatened Species.

Moore, R. C., Kozyreva, O. Lebel-Hardenack, S., Siroky, J., Hobza, R., Vyskot, B., Grant, S. R. **(**2003) “Genetic and Functional Analysis of *DD44*, a Sex-Linked Gene From the Dioecious Plant *Silene latifolia*, Provides Clues to Early Events in Sex Chromosome Evolution” Genetics **163**; 321-334.

Muir, G. and D. A. Filatov (2007). "A selective sweep in the chloroplast DNA of dioecious *Silene* (section *Elisanthe*)." Genetics **177**(2): 1239-1247.

Negrutiu, I., Vyskot, B., Barbacar, N., Georgiev, S., Moneger, F. (2001) "Dioecious plants. A key to the early events of sex chromosome evolution" Plant Physiology **127** (4); 1418-24.

Nei, M. (1987). Molecular Evolutionary Genetics. New York, Colombia University Press.

Newton, C. R., A. Graham, L. E. Heptinstall, S. J. Powell, C. Summers, N. Kalsheker, J. C. Smith and A. F. Markham (1989). "Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS)." Nucleic Acids Research **17**(7): 2503-2516.

Nielsen, R. and J. Wakeley (2001). "Distinguishing migration from isolation: a Markov chain Monte Carlo approach." Genetics **158**(2): 885-896.

Ohno, S. (1967) Sex chromosomes and sex-linked genes. Springer-Verlag, Berlin.

Osborne, J. L., A. P. Martin, N. L. Carreck, J. L. Swain, M. E. Knight, D. Goulson, R. J. Hale and R. A. Sanderson (2008). "Bumblebee flight distances in relation to foraging landscape." The Journal of Animal Ecology **77**(2): 406-415.

Peakall, R., P. E. Smouse and D. R. Huff (1995). "Evolutionary implications of allozyme and RAPD Variation in diploid populations of dioecious buffalograss *Buchloe dactyloides*." Molecular Ecology **4**: 135-147.

Peakall, R., M. Ruibal and D. B. Lindenmayer (2003). "Spatial autocorrelation offers new insights into gene flow in the Australian bush rat, *Rattus Fuscipes*." Evolution **57**: 1182-1195.

Peakall, R. and P. E. Smouse (2006). "GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research." Molecular Ecology Notes **6**: 288-295.

Prentice, H. (1976). "A study in endemism: *Silene diclinis*." Biological Conservation **10**: 15-30.

Prentice, H. C. (1984). "Enzyme polymorphism, morphometric variation and population structure in a restricted endemic, *Silene diclinis* (Caryophyllaceae)." Biological Journal of the Linnean Society **22**: 125-143.

Prentice, H. C. (1984). "The sex ratio in a dioecious endemic plant, *Silene diclinis*." Genetica **64**: 129-133.

Prentice, H. C. (1986). "Climate and clinal variation in seed morphology of the white campion, *Silene latifolia* (caryophyllaceae)." Biological Journal of the Linnean Society **27**: 179-189.

Prentice, H. C. (1988). "*Silene* section *Elisanthe* in the Iberian Peninsula." Monografías del Instituto Pirenaico de Ecología **4**: 321-324.

Prentice, H. C., J. U. Malm and L. Hathaway (2008). "Chloroplast DNA variation in the European herb *Silene dioica* (red campion): postglacial migration and interspecific introgression." Plant Systematics and Evolution **272**(23-37).

Pritchard, J. K., M. Stephens and P. Donnelly (2000). "Inference of population structure using multilocus genotype data." Genetics **155**(2): 945-959.

Rice W. R., (1987) “Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome” Genetics **116** (1); 161-7.

Rozas, J. and R. Rozas (1999). "DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis." Bioinformatics **15**(2): 174-175.

Rozen, S., Skaletsky, H., Marzalek, J. D., Minx, P. J., Cordum, H. S., Waterston, R. H., Wilson, R. K., Page, D. C. (2003) “Abundant gene conversion between arms of palindromes in human and ape Y chromosomes” Nature **423**; 873-876.

Sambrook, J. and D. W. Russell (2001) Molecular Cloning: a laboratory manual New York, Cold Spring Harbour Press.

Shaw, K. L. and P. D. Danley (2003). "Behavioural genomics and the study of speciation at a porous species boundary." Zoology (Jene, Germany) **106**(4): 261-273.

Shin, J. (2006). LDheatmap: Graphical display of pairwise linkage disequilibria between SNPs. V0.2-1. Department of Statistics & Actuarial Science, Simon Fraser University, Burnaby, Canada.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H.S., Hillier, L., Brown, L. G. (2003) “The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes” Nature **423**; 825-837.

Slatkin, M. and L. Voelm (1991). "FST in a hierarchical island model." Genetics **127**(3): 627-629.

Smouse, P. E., J. C. Long and R. R. Sokal (1986). "Multiple regression and correlation extensions of the Mantel test of matrix correspondence." Systematic Zoology **35**: 627-632.

Smouse, P. E. and J. C. Long (1992). "Matrix correlation analysis in anthropology and genetics." Yearbook of Physical Anthropology **35**: 187-213.

Smouse, P. E. and R. Peakall (1999). "Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure." Heredity **82**: 561-573.

Sokal, R. R., N. L. Oden and C. Wilson (1991). "Genetic evidence for the spread of agriculture in Europe by demic diffusion." Nature **351**(143-145).

Strasburg, J. L. and L. H. Rieseberg (2008). "Molecular demographic history of the annual sunflowers *Helianthus annuus* and *H. petiolaris* - Large effective population sizes and rates of long-term gene flow." Evolution **62**(8): 1936-1950.

Steinemann, M. and Steinemann, S. (1992) “Degenerating Y chromosome of *Drosophila miranda*: a trap for retrotransposons” Proceedings of the National. Academy of Science USA **89**; 7591-7595.

Steinemann, M. and Steinemann, S. (1998) “Enigma of Y chromosome degeneration: neo-Y and neo-X chromosomes of *Drosophila miranda* a model for sex chromosome evolution” Genetica **102/103**; 409-420.

Taberlet, P., L. Fumagalli, A.-G. Wust-Saucy and J.-C. Cosson (1998). "Comparative phylogeography and postglacial colonization routes in Europe " Molecular Ecology **7**: 453-464.

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

Takahata, N. and Y. Satta (2002). Pre-speciation coalescence and the effective size of ancestral populations. Modern Developments in Theoretical Population Genetics. M. Slatkin and M. Veuille, Oxford University Press**:** 52-71.

Taylor, D. R. and S. R. Keller (2007). "Historical range expansion determines the phylogenetic diversity introduced during contemporary species invasion." Evolution **61**(2): 334-345.

van Andel, T. H. and P. C. Tzedakis (1996). "Palaeolithic landscapes of Europe and environs: 150,000-25,000 years ago: and overview." Quaternary Science Reviews **15**: 481-500.

Vanlerberghe, F., B. Dod, P. Boursot, M. Bellis and F. Bonhomme (1986). "Absence of Y-chromsome introgression across the hybrid zone between *Mus domesticus* and *Mus musculus musculus*." Genetical Research **48**(3): 191-197.

Vellekoop, P., J. B. Buntjer, J. W. Maas and J. vanBrederode (1996). "Can the spread of agriculture in Europe be followed by the spread of the weed *Silene latifolia.* A RAPD study." Theoretical and Applied Genetics **92**: 1085-1090.

Vigouroux, Y. and D. Couvet (2000). "The hierarchical island model revisited." Genetics, Selection, Evolution **32**(4): 395-402.

Waelti, M. O., J. K. Muhlemann, A. Widmer and F. P. Schiestl (2008). "Floral odour and reproductive isolation in two species of *Silene*." Journal of Evolutionary Biology **21**(1): 111-121.

Wakeley, J. and J. Hey (1997). "Estimating ancestral population parameters." Genetics **145**(3): 847-855.

Wall, J. D. and M. F. Hammer (2006). "Archaic admixture in the human genome." Current Opinion in Genetics and Development **16**(6): 606-610.

Wang, R.-L., J. Wakeley and J. Hey (1997). "Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives." Genetics **147**: 1091-1106.

Webb, T. and P. J. Bartlein (1992). "Global changes during the last 3 million years: climatic controls and biotic responses." Annual Review of Ecology and Systematics **23**: 141-173.

Westergaard, M., (1958) “The mechanism of sex determination in dioecious flowering plants” Advances in Genetics **9**; 217-281.

Wolfe, K. H., W. H. Li and P. M. Sharp (1987). "Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs." Proceedings of the National Academy of Sciences of the United States of America **84**: 9054-9058.

Wright, S. (1931). "Evolution in Medelian Populations." Genetics **16**(2): 97-159.

Wright, S. (1938). "Size of population and breeding structure in relation to evolution." Science **87**: 430-431.

Wright, S. I. and B. Charlesworth (2004). "The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model." Genetics **168**(2): 1071-1076.

Yatabe, Y., N. C. Kane, C. Scotti-Saintagne and L. H. Rieseberg (2007). "Rampant gene exchange across a strong reproductive barrier between the annual sunflowers *Helianthus annuus* and *H. petiolaris.*" Genetics **175**: 1883-1893.