Tedder, Andrew R (2011) *Comparing the consequences of mating system shifts between different species of cruciferous plants in relation to phylogeography*. PhD thesis.

http://theses.gla.ac.uk/2372/

Copyright and moral rights for this thesis are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Glasgow Theses Service
http://theses.gla.ac.uk/
theses@gla.ac.uk

Comparing the consequences of mating system shifts between different species of cruciferous plants in relation to phylogeography.

Andrew Tedder

This thesis is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy.

University of Glasgow
Institute of Biodiversity, Animal Health and Comparative Medicine
College of Medical, Veterinary & Life Sciences

October 2010

Abstract

Sporophytic self-incompatibility is a genetically controlled inbreeding prevention mechanism, which is prevalent in the Brassicaceae, and has been reported in a variety of high profile species. Despite the benefits of preventing self-fertilization in terms of maintaining genetic diversity, variation in the strength of self-incompatibility (SI) has also been well documented, as has a shift from SI to inbreeding at the species and population levels. An important underlying driving force behind a switch to inbreeding could be the reproductive assurance provided by not requiring an unrelated mating partner for sexual reproduction. This could be beneficial for a species undergoing rapid colonization, because only a single individual is required to begin a sexually reproducing colony after a long-distance dispersal event (Baker's law), which is characteristic of the plight of many species after the last glacial maxima. The purpose of my thesis was to evaluate the effects of variation in mating system on post-glacial colonization, using two model species that show intraspecific variation in outcrossing rates. The first, *Arabidopsis lyrata,* represents an excellent model system to assess post-glacial colonization history because it exhibits broad geographical and ecological ranges, and has a recently completed genome sequence. In North America, *A. lyrata* has further benefits as a model system, namely it exhibits variation in the strength of SI and shift to SC at the population level, which is not observed in Europe. The second species is *Arabis alpina,* which also appears to show population level variation in mating system strength in Europe based on variation in F_{IS}. This has been putatively linked to colonization history after the last glacial maxima. Unlike in *A. lyrata* however, its mating system has not been characterized. Mating system delimitation in *A. alpina* has the potential to aid the interpretation of patterns of ecological genetic diversity, which may in part be influenced by local or regional stochastic changes to mating system variation.

My first objective was to identify if *A. alpina* had a functioning SI system based on both self-fertilization experiments, and allozyme based outcrossing rate estimations. I found strong evidence to suggest the presence of a functional barrier preventing self-fertilization in *A. alpina*. I identified multiple putative

SRK alleles (the female determinant of self-incompatibility), suggesting that the same type of sporophytic system seen in other Brassicaceae species governs SI in this species. I also demonstrated linkage of SI phenotype to some *SRK* genotypes by diallel crosses, strengthening the case for a functional SI system in this species. Further to this I demonstrate variation in mating system strength between populations, and autonomous inbreeding was seen in a single population. I note that the potential changes in SI status coincide with areas suspected to differ in post-glacial history based on allozyme diversity reported in previous work.

While the number of populations sampled was insufficient to link mating system variation to colonization history in *A. alpina*, mating system variation has been more extensively characterized in North American *A. lyrata*, allowing more fine-scale resolution of population structure and post-glacial colonization history; an underlying objective of my thesis. I used three molecular marker systems (cpDNA, nuclear micro-satellites and allozymes) to assess these phylogeographic questions, and present evidence of three putative colonization routes for the Great Lakes region. These putative routes are congruent with those described in other species, particularly amphibians and reptiles. Further to this I considered the possible location of glacial refugia, and likelihood that plant taxa may have survived during Pleistocene glaciation in close proximity to the Laurentide Ice Sheet, particularly in Illinois, Indiana, Wisconsin and Minnesota, which may also be true for some animal taxa. I examined patterns of population structure, and scenarios that may have influenced this, and present support for the previously documented theory of multiple breakdowns in SI in this geographic region.

My final objective was to assess the suitability of the three marker systems for phylogeographic reconstruction in *A. lyrata*, by comparing and contrasting the patterns of population structure, and colonization history suggested by each system. Levels of variation observed between the marker systems used varied, and I explored how these patterns complemented and contradicted each other. As expected, the nuclear micro-satellite loci represent the system with the greatest genetic diversity, but do not allow meaningful conclusions to be drawn regarding colonization history because of low levels of shared variation between populations. Conversely, the allozyme loci presented much lower levels of genetic diversity, but support population structuring conclusions based on both cpDNA data and previous studies of *A. lyrata* and other taxa in this area. The cpDNA marker (*trn*F) represents a somewhat contentious system to use for phylogeography in *A. lyrata*, as it contains a tandem array of highly variable, but complexly evolving duplications (pseudogenes). I concluded that these structural changes could be phylogenetically informative when pseudogene evolutionary relationships can be resolved This was based on variation in patterns of diversity, and the subsequent population structure change that occurred when using different methods of assessing *trn*F variation.

Table of Contents

List of Tables

List of Figures

List of Appendices

Acknowledgments

First and foremost, I would like to thank my supervisors for their help, and direction. Barbara has provided countless hours of comments on draft manuscripts and chapters, and I remain very grateful for this, as it is not a task I envy. Steve, along with Barbara, has provided in depth comments, as well as useful discussion and perhaps most importantly, continued enthusiasm for the subject matter and underlying theory. I certainly wouldn't be in a position to submit my thesis if it weren't for Barbara and Steve, so Thank you.

Secondly, I have been privileged to be part of an excellent research group, with a diverse array of both skills and personalities. Each component member has helped me solve problems (worked related and personal) over the last three and a half years, and so I extend my gratitude to you all: Marc Stift, Peter Hoebe, Clare Marsden, Ashley Le Vin, Aileen Adam, Anna Muir, Annabelle Haudry and Zha. Thanks again guys.

While a student in Glasgow, I have made it my mission to sample the many establishments that offer burgers in the West End, and in doing so have experienced the good, the bad and the ugly… It wouldn't have been possible to do this without some good friends, so thanks to Gus, Nash and Flavie. Sadly this mission has come at a price, and we are all at least 1 stone heavier than when we started (and that's being generous!)

It just remains for me to say thank you to those that mean the most. Flavie has been a constant support in and out of 'school', and although I suspect I would have got more done without her distracting me, I am eternally grateful for it. My family, well what can I say: Thank you all for all your support, who'd have thought I'd still be at university 10 years after saying I didn't want to go…

Author's Declaration

I declare that the work recorded in this thesis is entirely my own, except where otherwise stated, and has not been submitted as part of a degree elsewhere. Much of this thesis has been produced in co-authorship, and my personal contribution to individual chapters is listed below.

Chapter 2: *Submitted to Annals of Botany as:* Tedder, A., Ansell, S. W., Lao, X., Vogel, J. C ., Mable, B. K. **Sporophytic self-incompatibility genes and mating system variation in *Arabis alpina*.** AT carried out lab work (under the supervision of SWA for allozyme screening), green house pollinations, data analysis and drafted the first manuscript. The final draft was completed after detailed comments on manuscript by SWA and BKM.

Chapter 3: *Published as*: Tedder, A., Hoebe, P. N., Ansell, S. W., Mable, B. K. (2010) **Using chloroplast *trn*F pseudogenes for phylogeography in *Arabidopsis lyrata*.** AT carried out all lab and analytical work, and drafted the first manuscript. The final draft was completed after detailed comments on the manuscript by SWA and BKM.

Chapter 4: *In preparation for submission as*: Tedder, A., Vogel, J. C., Mable, B. K., Ansell, S. W. **Post-glacial expansion and population structure in North American *Arabidopsis lyrata* populations.** AT carried out all lab work (except micro-satellite PCR's, which were conducted by A. Adam), including allozyme screening, which was carried out under the supervision of SWA. AT also carried out all analysis and drafted the first manuscript. The final draft was completed after comments on the manuscript by SWA and BKM.

1 General introduction

1.1 Mating system variation

Plant mating systems, while hugely diverse, can be separated into two principle categories: asexual reproduction, creating progeny that are genetically identical to the mother; and sexual reproduction, which occurs when male and female gametes fuse, creating progeny that are genetically distinct from the mother (with the exception of those produced by self-fertilization) (Maynard Smith 1978). In species with co-sexual flowers, containing both male reproductive structures (stamens, consisting of anther and filament) and female reproductive structures (pistils, consisting of stigma, style and ovary), reproductive efficiency is increased, but so too is the risk of self-fertilization and the associated detrimental consequence, inbreeding depression (Lande and Schemske 1985).To counteract this, many co-sexual species prevent self-fertilization by being self-incompatible (SI). In homomorphic (multialleic) systems (Brewbaker 1957), SI is a genetically determined, pre-zygotic barrier, which recognizes self or self-related pollen and prevents fertilization, thus removing risk of inbreeding and maximizing the potential for outcrossing. Although the expression of SI takes a variety of forms, including variation in their time of action, floral polymorphism, gene of expression and the number of alleles involved (De Nettancourt 1977), and functional competence of SI varies between systems that employ different mechanisms (Lewis 1979), the evolution of SI is considered a significant factor in the success of early flowering plants (Whitehouse 1950). Estimates suggest that around 50-60% of angiosperms have an SI system (Brewbaker 1959; Hiscock and Kües 1999), with SI considered an ancestral state in many lineages, rather than the result of multiple independent origins from self-compatible (SC) species within a lineage (Allen and Hiscock 2008; Igic *et al* 2004; Takebayashi and Morrell 2001).

Two homomorphic SI systems are particularly well documented within the angiosperms: 1) A 'gametophytic' system (GSI), where pollen incompatibility type is dictated by the grain's own haploid genotype (reviewed by Hiscock and McInnis 2004), with single incompatibility (S) locus systems known in a variety of families, including well characterized systems in the Solanaceae and Rosaceae (reviewed by Weller *et al* 1995) and multiple loci (two or more) systems also occurring in several families, including grasses (Li *et al* 1994). 2) Single locus 'Sporophytic' SI systems (SSI), where the pollens incompatibility type is determined by the diploid plant that produces it, have been well documented in the Brassicaceae and the Asteraceae (reviewed by Hiscock and McInnis 2004).

1.2 The costs and benefits of SI

As SI is thought to evolve under conditions of extreme inbreeding depression, this will also result in an increased genetic load due to high heterozygosity (Charlesworth & Charlesworth, 1979a; Charlesworth *et al* 1990; Byers & Meagher, 1992; Charlesworth, 2006b). Thus, there is an expectation that SI loss will have detrimental fitness consequences, as a result of exposing deleterious mutations, which result from increased homozygosity and loss of genetic diversity resulting from inbreeding. Subsequently, inbreeding depression, as a mechanism of SI maintenance, has received much empirical and theoretical attention (Charlesworth & Charlesworth, 1979b; Lande & Schemske, 1985; Charlesworth & Charlesworth, 1987; Barrett, 1988; Charlesworth & Charlesworth, 1990; Barrett & Charlesworth, 1991; Uyenoyama & Waller, 1991; Charlesworth *et al* 1992a; Husband & Schemske, 1996; Charlesworth & Charlesworth, 1999; Wang *et al* 1999; Carr & Dudash, 2003), including early work by Darwin (1876).

Species that maintain a strong SI system are likely to benefit from a sheltered genetic load; however, they may subsequently suffer from a reduction in compatible mating partners compared with selfing individuals, which benefit from reproductive assurance (Baker 1955; Darwin 1876). This will be particularly apparent in small populations, where the number of different compatible partners (with different mating specificities) will likely be low. Transitions from SI to self-compatibility (SC) is frequent at the species level, and is also well documented within species, at the population level (Weller & Sakai 1999). This suggests that there are likely costs associated with having an SI system, or benefits, in certain situations, to being SC. Once a population has undergone the shift to self-compatibility, re-establishing a functional SI system is potentially very difficult, thus shifts to SC are considered unidirectional (Igic *et al* 2006).

The absence of an SI system however, is not always synonymous with complete selfing. Intermediate levels of self-fertilization can be maintained within species without functional SI systems (Lloyd 1979; reviewed in Goodwillie *et al* 2005), thus suggesting that a reversion to outcrossing could theoretically be achieved by other mechanisms (Nasrallah *et al* 2004). Despite some theory predicting that intermediate rates of outcrossing will be unstable (Lande & Schemske 1985), there is evidence to suggest that this may not always be true (reviewed in Goodwillie *et al* 2005). Intermediate rates of outcrossing (between complete outcrossing and complete inbreeding) may be facilitated by the trade-off between inbreeding, allowing reproductive assurance or increased colonization ability, and outcrossing, increasing offspring quality (Bateman 1955; Jarne & Charlesworth 1993; Charlesworth 2006b).

The consequences of inbreeding for a particular species, population or family, will be highly dependant on their own history of outcrossing. While theoretical models may predict that inbreeding depression will be exposed by selfing because of the increased effect of deleterious recessive mutations (Schemske & Lande, 1985), it should also theoretically reach a point where it is lower in selfing populations, than it is in outcrossing ones as a result of purging these deleterious recessives. However, the strength of this effect is multifaceted, and will depend on the population size and the developmental stage at which inbreeding depression acts, the magnitude and/ or duration of inbreeding, the genetic basis of inbreeding depression, the number of loci that contribute, the magnitude of effects of alleles at the contributing loci, and linkage to genes under viability selection (Charlesworth & Charlesworth, 1987; Barrett & Charlesworth, 1991; Charlesworth *et al* 1992; Husband & Schemske, 1996). These factors could allow recessive mutations, that have only a mildly deleterious effect to avoid effective purging, thus maintaining substantial inbreeding depression even in populations that are highly inbred (Charlesworth *et al* 1990, 1991).

The impact of inbreeding may be further complicated by population subdivision and biogeographical history (Vekemans *et al* 1998; Schierup *et al* 2000; Charlesworth, 2003), along with life history (Morgan, 2001) and local environmental effects (Hayes *et al* 2005). This may suggest that gene flow between populations, which differ in mating system, could affect the rate at which purging occurs. This could be particularly important when considering conservation, as it is unclear whether purging of genetic load has the ability to diminish the detrimental effects presented by reduced heterozygosity, and over what time scales it might occur (Hedrick & Kalinowski, 2000; Glémin *et al* 2001; Keller & Waller, 2002). Theoretical studies suggest that plant populations that exhibit slight inbreeding and strong population structure will be the most likely to show purging (reviewed in Keller & Waller, 2002), although there is little empirical evidence to corroborate this. Populations that demonstrate local adaptation to inbreeding could suffer from outcrossing depression (Bailey & McCauley, 2006), where the original parental gene combinations are broken up by recombination after mating. This can result in some offspring being homozygous for one parent's genes at one locus and the other parent's genes at another locus. These genome rearrangements may disrupt epistatic interactions that confer fitness in specific environments (local adaptation), as well as gene interactions that are independent of the environment (intrinsic coadaptation) (Dobzhansky 1948; Templeton 1986). This has the potential to affect the balance between selfing and outcrossing, despite the potential for inbreeding depression. The majority of documented evidence for purging comes from comparison of selfed and outcross progeny from species that do not have a genetically controlled SI system in place (Byers & Waller 1999). This likely reflects the difficulties associated with forcing SI plants to self, although this can sometimes be overcome by pollination under increased levels of CO2 (Llaurens *et al* 2009). Since SI is considered ancestral in many plant lineages (Allen & Hiscock 2008; Igic *et al* 2004; Takebayashi & Morrell 2001), species with an SI system will, in general, have experienced a long period of outcrossing. Thus, the expected magnitude of inbreeding depression will be high, which means that some purging is probably necessary to facilitate a shift to inbreeding.

To fully understand the selective forces that govern mating system regulation, it is necessary to compare the genetic consequences of inbreeding in closely related species that are predicted to have shared the same SI system in the past. As selfing is theoretically expected to reduce effective population size and effective recombination rates (Ne), it is expected to result in reduced polymorphism, increased linkage disequilibrium (non-random association of alleles at two or more loci) and hitchhiking between linked genes (Charlesworth

& Wright, 2001; Wright *et al* 2002; Charlesworth, 2003; Glémin *et al* 2006). Increased isolation between selfing populations relative to outcrossers can result directly from lack of outcrossing or indirectly from accompanying changes, such as small flowers and low pollen output (Glémin *et al* 2006). The effects of isolation on population structure means that making general predictions about the relative species-wide diversity in inbreeders and outcrossers, is not possible, despite the reduced Ne caused by homozygosity. This is because high levels of diversity can be maintained between selfing populations (Charlesworth, 2003), which in turn means that selfing species may actually show higher levels of diversity than their outcrossing relatives in certain cases.

1.3 SI in the Brassicaceae

In most genetically controlled SI systems, if proteins in the female tissues recognize surface proteins from the pollen grain as self, a signal is sent to block pollen tube growth. As the determination of self requires independent genes for the male and female components, it is essential that recognition specificity be maintained. If this lock-and-key mechanism is interrupted, it can result in SI break down.

In the SSI system present in the Brassicaceae, the genes that code for SI specificity in pollen (SCR; S-locus Cysteine Rich) and pistils (SRK; S-Related Kinase) are organized into self-recognition haplotypes, which effectively function as a single locus that can span over 100 kb (Suzuki *et al* 1999; Kusaba *et al* 2001; Shiba *et al* 2003). Low levels of recombination are essential in this region, in order to maintain the same specificity in both male and female components (Awadalla & Charlesworth, 1999). It is not known, however, how long this would be maintained for, if SI functioning were lost.

In individuals which are strongly SI, absence of recombination in the S-gene region may lead to balancing selection (maintenance of two or more alleles in equilibrium at higher frequencies than can be explained by mutation, within a population) that extends to genes in close physical proximity to those directly under selection (Charlesworth, 2006a). This has been demonstrated by isolations of genes, in close proximity to the S-locus, which show trans-specific polymorphisms (the occurrence of similar alleles in related species) (Charlesworth *et al* 2006) and high levels of linkage disequilibrium with genes flanking S-genes (Castric *et al* 2008; Hagenblad *et al* 2006; Ruggiero *et al* 2008).

In SSI systems, the haploid pollen grain carries a diploid number of SCR proteins on its surface, because diploid cells in the tapetum deposit them. Expression of SCR alleles from the sporophytic tissue is determined by dominance interactions (which can lead to one or both being expressed), as is the SRK presented to the pollen grain at the stigma surface (reviewed in Charlesworth *et al* 2000; Hatakeyama *et al* 2001). Further to this, certain haplotypes can express different dominance in the pollen and stigma: for example, one haplotype maybe dominant in the stigma, and thus expressed, but co-dominant in the pollen and so both are expressed (Bateman 1955; Thompson & Taylor 1966). This can result in offspring production between individuals that share S-haplotypes, effectively increasing the level of inbreeding in populations that maintain a strong SI system. As SRK is expressed irrespective of dominance status, it is thought that dominance is under the control of SCR (Hatakeyama *et al* 2001; Kusaba *et al* 2002; Shiba *et al* 2002).

Recognition of ‘self’ activates an SRK-mediated signalling cascade in the pistil, that is part of a ubiquitination-degradation type of cell–cell recognition system (Chapman & Goring 2010). Control of this signal transduction pathway relies on a complex interplay between promoters and inhibitors (Cabrillac *et al* 2001; Takayama & Isogai, 2003; Goring & Walker, 2004; Murase *et al* 2004; Chapman & Goring 2010) resulting in self-related pollen tube penetration being prevented. This pathway has many steps; a mutation or disruption at any of these steps could effectively cause loss of SI. With this in mind, an expectation for multiple independent SI losses, each comprising a different mechanism for loss could be justifiable. As at least some of the downstream components of the signalling pathway are located away from the S-locus, changes to SRK and SCR may not be seen for some time after the disruption to SI function.

1.4 Mechanisms of SI breakdown

Although it is clear that having a functioning SI system presents many benefits to a species, the most important of which is likely inbreeding avoidance (reviewed by Charlesworth 2003), loss of a functioning SI system is not uncommon at the species level, and has been documented in a wide range of angiosperm taxa (Weller and Sakai 1999). There have been a number of mechanisms proposed to account for this, particularly among members of the Brassicaceae, including the recently reported 213 base pair inversion of the SCR gene in *Arabidopsis thaliana* (Tsuchimatsu *et al* 2010). This inversion is associated with a high proportion of European accessions of the SC and highly inbreeding species, and its derivative haplotypes are thought to be responsible for loss of SI (Tsuchimatsu *et al* 2010). Nasrallah *et al* (Nasrallah *et al* 2002) demonstrated that introduction of functional SCR and SRK allele copies from SI *Arabidopsis lyrata* into SC *A. thaliana* can partially restore SI function in some *A. thaliana* accessions. This suggests that fixation of a non-functioning allele may be responsible for the loss of SI in this species. Mechanisms not directly associated with the S-locus may also play a role in SI loss in certain species; however, Cabrillac *et al* (2001) demonstrated that a downstream component of the pollen rejection response (ARC1) in *Brassica oleracea* might be responsible for SI loss, suggesting that regulation of signal transduction may also be important for modifying the SI/SC response.

Population-level variation in mating system is seen extensively throughout the Brassicaceae and has been reported in many genera, including *Arabidopsis* (Foxe *et al* 2010; Mable *et al* 2005a; Mable and Adam 2007), *Brassica* (Hinata *et al* 1995), *Capsella* (Paetsch *et al* 2006), *Leavenwortia* (Koelling *et al*) and *Raphanus* (Okamoto *et al* 2004). Furthermore, quantitative variation in the strength of the SI response within a species, is also well documented, and may be an important transition to SC (Good-Avila and Stephenson 2002; Levin 1996; Nielsen *et al* 2003; Vogler and Stephenson 2001). This means that the SI response of a self-incompatible plant can be ‘leaky’ to such an extent that it exhibits partial self-compatibility (PSC). PSC has been attributed to a variety of causes including: low levels of S-allele products being expressed in flowers (Mena-Ali and Stephenson 2007), the action of modifiers able to influence the efficiency of the SI reaction (Good-Avila and Stephenson 2002), or the presence of post-pollination mechanisms leading to floral abscission and fruit abortion (Vallejo-Marin and Uyenoyama 2004). PSC is typically associated with either the production of small fruits with few or small seeds (following natural or enforced self-pollination), or a weakening of the SI response, which corresponds with the flowers' age (Liu *et al* 2007; Stephenson *et al* 2003; Vogler and Stephenson 2001).

There is some evidence from populations experiencing multiple losses of SI in the genus Arabidopsis that supports the idea that PSC could be an evolutionary transition to an increased level of self-fertilization. In North American *A. lyrata*, variation in the strength of SI at both the individual and population levels is well documented (Mable *et al* 2005a), with a number of populations having shifted from predominantly outcrossing, to predominantly inbreeding (Foxe *et al* 2010; Mable and Adam 2007). Within populations, PSC individuals have been shown to exhibit similar levels of genetic variation to their SI neighbors (i.e. from the same population; Hoebe *et al* 2009). Within this species, it is possible that the loss of SI might have started with weakening of the incompatibility response in older flowers, a phenomenon known as pseudo-self-compatibility or transient SI that is well documented in *A. thaliana* (Liu *et al* 2007). This weakening may be selectively favored during post-glacial colonization (Mable *et al* 2005a). Further to this, evidence suggests multiple independent losses of the SI reaction and shift to inbreeding in *Arabidopsis* (Foxe *et al* 2010; Hoebe *et al* 2009; Nasrallah *et al* 2004; Shimizu *et al* 2004), and that the loss of variation at the S-locus may be gradual, as demonstrated in *A. thaliana* (Bechsgaard *et al* 2006). The low variability (in comparison to the genomic average) of S-locus pseudo-genes in *A. thaliana* may add support to this theory (Tang *et al* 2007) with a gradual erosion of polymorphism through genetic drift, and selection for inactivity the likely causes. Together, these independent evidence streams could tentatively suggest that natural selection, facilitated by reduced mate choice after a long distance dispersal event during colonization, may lead to the gradual fixation of multiple, independent mutations, which weaken or disable the SI system throughout the geographic range of a given species (Busch and Schoen 2008).

1.5 Glaciation and post-glacial colonization

Dispersal is a major process, which affects the ecology, genetics, and geographical distribution of species (Ridley 1930; Van der Pijl 1969; Jaquard *et al* 1984; Sauer 1988; Dingle 1996). In temperate regions, the dispersal ability of a species has directly influenced its response to glacial cycles. During the Pleistocene (2.4 x 106 yr –10 000 yr BP), at least six glacial advances (although it could be up to 20; Mann & Hamilton 1995) affected the physical and biological environments of the Northern Hemisphere (Cox and Moore 2000). In North America, the Wisconsin glaciation began 120 000 yr BP and ended approximately 8000 yr BP (Davis 1983). At its maximum (22 000 – 18 000 yr BP), the Laurentide ice sheet extended as far south as 40° N in eastern North America, and covered much of Canada, southeastern Alsaska, and the northern United States (Mann & Hamilton 1995). During the glaciation, the ice sheets, the permafrost that surrounded the ice, lower global temperatures and reduced availability of water, meant that plant and animal species were displaced. Their habitats were restructured and the subsequent distribution of taxa was radically altered (Hewitt 1999; Brunsfeld *et al* 2001). As the Wisconsin ice sheet began to retreat, which led to a shift of climatic zones, approximately 18 000 yr BP, populations of plants and animals that had survived in non-glaciated refugia expanded their ranges (Whitlock 1992). The latitudinal and altitudinal range shifts associated with glacial cycles, likely involved considerable demographic changes and provided opportunity for adaptation to occur. Range shifts of this magnitude are postulated to have had both stochastic and selective effects on the genetic variation and architecture of species (Hewitt 2004). Populations and lineages will have gone extinct, bottlenecks and founder events will have caused the loss of alleles, and mutations will have spread by both selection and population expansion (Hewitt 1996, 2004).

Somewhat paradoxically, post-glacial seed and pollen dispersal in plants is lower than would be predicted based on carbon-dated pollen density in sediment samples around eastern North America, and so appears not to explain the rapid northern movement that has clearly occurred in many species (Clark *et al* 1998).

One pertinent example of this was demonstrated by Cain *et al* (1998), which involved modelling the migration of the ant-dispersed species, Canadian wild ginger (*Asarum canadense*). They concluded that long-distance seed dispersal was necessary to explain the present-day distribution of this species in eastern North America, although the mechanism for this was unknown. Similarly, analysis of preserved pollen samples suggests that several *Acer* spp. (particularly *A. saccharum*) migrated nearly 2000 km in approximately 10 000 yr, which necessitates a migration rate of 200 m/yr (Davis 1983). Additional modelling studies have also indicated that modes of colonization, and in particular, rare long-distance dispersal events, have influenced the genetic diversity of species in glaciated regions (Ibrahim *et al* 1995). Long-distance dispersal has been demonstrated to be associated with reduced genetic diversity, and has been documented in the glaciated ranges of several European and North American species (Hewitt 1996; Soltis *et al* 1997; Bernatchez & Wilson 1998; Schmitt *et al* 2002; Michaux *et al* 2003), although in the majority of reported cases, the mechanisms of long-distance dispersal are not known. This reduction in genetic diversity has been attributed to founding effects where a reduced number of migrants are the first to arrive and then rapidly recolonize a large unoccupied area (Hewitt 2000). In many terrestrial species, regions colonized in the postglacial period have been shown to have lower genetic diversity, consistent with this model (Hewitt 1996, 2000). Such leading-edge colonization has consequences that also affect more southerly populations, by restricting population expansion once the niche space has been filled (Hewitt 1993); the greater number and smaller range of southern genomes support this (Hewitt 2004).

Fossil records show that not all temperate species followed this type of rapid expansion. Certain species will have been more dependent on certain environmental conditions or the requirement of facilitation by other species (plant-animal mutualism), which has been well documented in insects (Beattie and Culver 1979; Hanzawa *et al* 1988; Kjellsson 1985; Manzaneda and Rey 2009), birds (Horvitz and Lecorff 1993; McCay *et al* 2009; Mosandl and Kleinert 1998) and mammals (Matias *et al* 2010; Willson 1993). Some will have been more affected than others by various barriers, habitat distributions and prior colonizers (Hewitt 2000). This type of slower expansions is likely to have involved much shorter dispersal distances, and subsequently larger effective population sizes (Hewitt 2000, 2004). This would theoretically have allowed greater retention of genetic diversity (Hewitt 1996). These two extreme modes of colonization, defined as Pioneer and Phalanx, respectively (Nichols & Hewitt 1994), when combined with different habitat distributions and climatic oscillations can produce a variety of geographical genetic structures (Hewitt 2004; Soltis *et al* 2006).

Despite populations that have undergone post-glacial rapid expansion characteristically having reduced levels of genetic variation when compared with southerly conspecific populations residing in non-glaciated areas (Sage & Wolff 1986; Hayes & Harrison 1992; Meriläet *al* 1997; Soltis *et al* 1997), levels of genetic variation may be confounded by admixture among populations expanding from multiple refugia (Conroy and Cook 2000; Soltis *et al* 1997). In North America these multiple refugia are likely to have been south of the ice sheets (Soltis *et al* 1997) or from refugia further north along the Pacific Coast (Heaton *et al* 1996) or from Beringia (MacDonald & Cook 1996).

Although the phylogeography of North America has been studied for several animal species (Placyk *et al* 2007), the effects of paleoclimatic change on the distribution of plant communities are not as well understood. It is clear that populations moving from non glaciated refugia to postglacial landscapes following the retreat of the ice would have been smaller than the original source populations (Placyk *et al* 2007). This could create a bottleneck effect, resulting in depletion of genetic diversity, thus increasing the influence of genetic drift (Keller and Taylor 2008; Nei *et al* 1975). In plants with a genetically controlled self-incompatibility system, this could lead to a reduction in numbers of S-alleles present and fewer compatible mating partners in recently founded populations (Vekemans *et al* 1998). Mate choice could be further limited by low plant densities, which attract fewer insect pollinators (Levin and Kerster 1969). In these situations it is predicted that self-fertilisation will present a reproductive advantage (Baker 1967), which may outweigh the potentially negative effects of inbreeding (Baker 1955; Pannell and Barrett 1998), especially during periods of rapid postglacial expansion (Baker 1966; Busch and Schoen 2008; Pannell and Barrett 1998), even for species operating a strong SI system.

1.6 Marker system choice for phylogeographic studies

Ever since the coining of the term “phylogeography” by Avise (1986) to refer to the relationship between mitochondrial DNA (mtDNA) phylogeny and the geographic distribution of animal species, marker system choice has been a pivotal factor in the patterns elucidated. While the rate of mtDNA evolution is seemingly well suited to phylogeographic studies in animals and as such requires relatively few loci (reviewed by Avise *et al* 2000), the rate of chloroplast DNA (cpDNA) evolution in plants is much more conservative, which can lead to difficulties in delimiting evolutionary processes and patterns of colonization (Shaw *et al* 2005). While there are numerous studies to suggest that cpDNA variation is structured in some angiosperm species (see reviews by Soltis *et al* 1992, 1997), some studies appear to have problems with lack of variation at the population level, and thus they suggest that either a larger number of loci are used, or more recently the use of multiple markers systems (reviewed by Soltis *et al* 2006). This problem is emphasized by recent work on DNA barcoding, which has focused on cytochrome oxidase 1 in animal taxa, but is not appropriate for use in plant taxa due to its much slower rate of evolution (Kress *et al* 2005).

Other marker systems have also been extensively used for phylogeographic studies, including allozymes and micro-satellites; however these systems are not without their drawbacks. Allozymes for example, tend to show limited variability (Biasiolo 1992; Soltis *et al* 1992), there are problems with scoring procedures (Murphy 1993) and in some cases, loci have been shown to be under selection (Hale and Singh 1991; Krafsur 2002). Conversely, micro-satellites have a very high rate of mutation that can suggest large amounts of population subdivision, but have been shown to demonstrate homoplasy (Anmarkrud *et al* 2008; Curtu *et al* 2004; Garza and Freimer 1996; van Oppen *et al* 2000).

As marker system choice is fundamental for an accurate understanding of the evolutionary processes that have shaped the distribution of a chosen species, it stands to reason that certain criteria are met by a potential marker. For example, mtDNA has been promoted as good marker choice for animal species due to its lack of recombination, putative neutrality and small effective population size (reviewed by Hickerson *et al* 2010), however as it is maternally inherited, it only provided a picture of maternal ancestory (Hurst and Jiggins 2005). Choosing the correct marker or combination of marker systems may not only increase power to resolve phylogeographic questions, but also allow one to be more confident about predictions if multiple systems offer the same explanation.

1.7 Study species

1.7.1 Arabidopsis lyrata

Arabidopsis lyrata L. is a close relative of the model organism *A. thaliana*, from which it is estimated to have diverged, approximately five million years ago (Koch *et al* 2001; Koch *et al* 2000). Al-Shehbaz and O'kane (2002) suggested that within the species delimitation, there are three subspecies: *A. lyrata* ssp. *lyrata, A. lyrata* ssp. *petraea* and *A. lyrata* ssp. *kamchatica*. Although they also noted that distinguishing each putative species by morphological characters alone could prove difficult, especially in regions where their distributions overlap. More recent taxonomic classifications view *A. lyrata* as a "species complex" (Schmickl *et al* 2010), comprising four species (*A. petraea, A. kamchatica, A. arenicola*, and *A. lyrata*), and a number of putative subspecies. Globally, the *A. lyrata* complex has a circumpolar, arctic-alpine distribution, but adaptation to various ecological conditions is seen at the species, subspecies and population levels (Schmickl *et al* 2010). In general, members of the *A. lyrata* species complex are perennial diploid outcrossers (2n = 2x = 16; Koch *et al* 1999), but some tetraploid cytotypes have also been reported (Mable *et al* 2004; Riihimaki and Savolainen 2004; Shimizu-Inatsugi *et al* 2009).

In North America, *A. lyrata* (or *A. lyrata* ssp. *lyrata*, depending on the classification system subscribed to) has a highly disjunct distribution (Leinonen *et al* 2009), representative of its early successional life history strategy (Spence 1959). It favours exposed alvar (limestone pavements), and sand dune habitats (Mable *et al* 2003) as opposed to the colder, more alpine climates it inhabits in Europe (Spence 1959). It is distributed in Canada (throughout Ontario and west into British Columbia) and throughout eastern and central United States (Al-Shehbaz & O'kane 2002). When directly compared to European populations, A. lyrata in North America shows reduced genetic variation, detected by chloroplast DNA and nuclear ribosomal sequences (Balañá-Alcaide *et al* 2006; Clauss and Mitchell-Olds 2006; Ross-Ibarra *et al* 2008; Schmickl *et al* 2008).

The *A. lyrata* complex has been shown to be a suitable study system for the analysis of character traits such as flowering time (Riihimaki *et al* 2005; Riihimaki and Savolainen 2004) and pathogen defence (Clauss *et al* 2006), but one of the main focuses of research on North American *A. lyrata* (relative to both *A. thaliana* and European *A. lyrata*) is the molecular mechanisms controlling variation in strength of self-incompatibility (Charlesworth *et al* 2006; Charlesworth *et al* 2003; Foxe *et al* 2010; Hagenblad *et al* 2006; Hoebe *et al* 2009; Mable and Adam 2007; Mable *et al* 2005b; Mable *et al* 2003; Schierup *et al* 2006; Schierup *et al* 2001). Generally considered an obligate outcrosser due to its sporophytic SI system, populations of North America *A. lyrata* have been extensively demonstrated as exhibiting mating system variation, with highly outcrossing, highly inbreeding and mixed mating systems recorded (Mable & Adam 2007; Mable *et al* 2005; Hoebe *et al* 2009). This mating system variation is likely the result of the SI system becoming defective in certain populations, and this has been postulated to be linked to post-glacial expansion (Mable *et al* 2005). This makes North American A. lyrata an excellent model system to test hypotheses concerning the evolution and consequences of mixed mating systems.

Whole genome sequencing of *A. lyrata* has been recently completed, and the data are already available (The *A. lyrata* genome sequence assembly v1.0, http://genome.jgi-psf.org/Araly1/Araly1.info.html). This therefore enables direct comparisons with the *A. thaliana* genome, and further increases the use of A. lyrata as a model system.

1.7.2 Arabis alpina

Arabis alpina L. is a distant relative of the model plants *Arabidopsis thaliana*, *A. lyrata*, and *Capsella bursa-pastoris* (Beilstein *et al* 2006), and is being developed as a system to study the ecological genetics of alpine environments (Poncet *et al* 2010; Wang *et al* 2009). It is a perennial herb with an ecological preference for cool disturbed habitats, reproducing either sexually or asexually via stoloniferous growth (Hegi 1986). Globally, it has true artic-alpine distribution (Jalas & Suominen 1994) covering all European mountain systems, the Canary Islands, North Africa, the high mountains of East Africa (Tanzania, Uganda and Kenya) and Ethiopia, the Arabian Peninsula and mountain ranges of Central Asia in Iran and Iraq. Additionally, it is found in the northern amphi-Atlantic area including northeastern North America, Greenland, Iceland, Svalbard and northwestern Europe (Jalas and Suominen 1994). In Europe, it is distributed in the arctic areas of Iceland and Scandinavia, and alpine regions of the Pyrenees, Alps, Apennines, Balkan, Carpathian and Tatras mountain systems (Jalas & Suominen 1994). It prefers moist habitats, which are characterized by areas of open gravel and rocks, and is often found in glacier foreland, typically growing at 500–2000 m altitude (Hegi 1986).

Throughout the range of *A. alpina* the landscape has been affected by Pleistocene glacial cycles (Hewitt, 2004). For other alpine species, this has resulted in populations having a high degree of genetic structuring (Comes and Kadereit 2003; Schonswetter and Tribsch 2005; Zhang *et al* 2004), which has been extensively tested in *A. alpina* (Ansell *et al* 2008; Assefa *et al* 2007; Ehrich *et al* 2007; Koch *et al* 2006). In Europe in particular, these studies suggest that the Tatras and Apennine mountain populations most likely represent Pleistocene (c. last 2.5Mya) glacial refugia, and thus make an important contribution to the post-glacial recolonisation of the Alps (Ehrich *et al*, 2007; Ansell *et al*, 2008). The origins of *A. alpina* have been well characterized, with evidence (based on synonymous mutation rates of the t*rn*L-F cpDNA region and the ITS of nuclear encoded ribosomal DNA) suggesting that it originated in Asia Minor less than 2 million years ago (Koch *et al* 2006). Furthermore, it has been suggested that this exodus from Asia occurred along three distinct routes. The first is thought to have proceeded via migration from the Arabian Peninsula to the East African high mountains. The second group is suggested to have given rise to all European and northern populations, as well as acting as a further source for the northwest African populations. The final group, which is still principally restricted to Asia, is thought to have migrated independently southward, where it made secondary contact with the East African lineage in Ethiopia, which resulted in the high genetic diversity in this area (Koch *et al* 2006).

Generally considered an obligate outbreeder, a regional analysis of *A. alpina* populations from the Italian peninsula, considered to be one of the three principle glacial refugia in Europe (Hewitt 2000, 2004), and nearby Maritime Alps using allozymes revealed substantial local variation in within-population inbreeding (F_{IS}) (Ansell *et al*, 2008). Populations from the Apuani portion of the Apennines range were close to Hardy-Weinberg equilibrium (HWE) (F_{IS} = 0.076, 95% CI 0.014-0.131), whereas those from the narrow Maritime Alps range showed a large distortion from HWE (F_{IS} = 0.553, 95% CI 0.457-0.620), suggesting the possibility of mating system variation among European populations, which has been tentatively linked to post-glacial expansion (Ansell *et al* 2008). Local deviations in HWE may alternatively be consistent with local differences in extent of glacial survival and bi-parental inbreeding on post-glacial population recovery.

A. alpina has the potential to be a useful model system for studying the ecological genetics of alpine adaptation, based on its extensive arctic-alpine latitudinal and altitudinal range, leaf shape and flowering time variation (Wang *et al* 2009, Poncet *et al* 2010). Along with its soon to be completed genome sequence (Coupland and Weigel, unpublished), further understanding of its mating system, however, is critical in the context of interpreting patterns of ecological genetic diversity, which may in part be influenced by local or regional stochastic changes to mating system variation.

1.8 Aims of the thesis

The overall aim of my thesis is to investigate the consequences of a shift in mating system (from self-incompatible to self-compatible) on the phylogeography of cruciferous plants, namely *Arabidopsis lyrata* and *Arabis alpina*. There are three principal objectives for my thesis:

1. To establish whether *A. alpina* has a functioning self-incompatibility system or the remnants of one, and whether this follows the same mechanisms reported in other *Brassica* species.

2. To assess patterns of post-glacial expansion in relation to mating system shifts in North American populations of *A. lyrata*, based on cpDNA.

3. To compare the use different marker systems for phylogeographic study, and understand how these systems can influence the interpretation of putative colonization history.

1.9 Chapter objectives

To determine whether variation in F_{IS} values in European *A. alpina* were the result of a functioning self-incompatibility system, or remnants of one, it was necessary to perform greenhouse self-pollinations in the hope of detecting a barrier against ‘selfing’, which was carried out in several French and Italian populations. Further to this, I determined outcrossing rates, based on progeny arrays with allozyme markers, to established whether these populations were effectively outcrossing or inbreeding in the field. I supplemented these data, with amplification of known SRK alleles (the female determinant of self-incompatibility in the Brassicaceae), and a diallel cross, which effectively linked selfing phenotype to putative SRK genotypes. The results of this work are presented in **chapter 2**.

In **chapter 3** I address the use of the trnL-F chloroplast marker, which has been extensively utilized for phylogeography since its development in the early 1990’s, for phylogeographic analysis of *A. lyrata* in the Great Lakes region of eastern North America. Further to this, I question the use of the pseudogene repeat sequences contained within intronic regions of said marker, and make inferences as to their suitability for phylogeographic studies, both in populations from the genetically depauperate eastern North America, and the more genetically diverse European populations, where its use has been criticised.

As marker systems and their suitability are a key objective in my thesis, I compare and contrast the use of micro-satellite loci and allozyme loci for phylogeographic studies in **chapter 4**, again using eastern North American populations of *A. lyrata* as a study system. Also in this chapter, I look at the wider implications of mating system variation within these populations, and if this has played a role in the post-glacial colonization history of the species in this geographic region. The Chloroplast DNA marker from chapter 3 is also used to compare the patterns of population structuring elucidated by the micro-satellite and allozyme systems.

Finally, in **chapter 5**, I discuss the broader applications and benefits of using multiple marker systems in tandem for phylogeography, and highlight the broader implications of my work.

2 Sporophytic self-incompatibility genes and mating system variation in *Arabis alpina*

2.1 Abstract

Sporophytic self-incompatibility prevents inbreeding in many members of the Brassicaceae, and has been well documented in a variety of high profile species. *Arabis alpina* is currently being developed as a model system for studying the ecological genetics of alpine environments, and is the focus of numerous studies on population structure and phylogeography. However, the genetics of self-incompatibility in this species has yet to be described. Here I present strong evidence that patterns of incompatibility in this species are consistent with the action of a single-locus sporophytic self-incompatibility system. I demonstrate functional avoidance of inbreeding in three European populations of *A. alpina* based on both self-fertilisation experiments and allozyme-based outcrossing rate estimates, and also describe the presence of fifteen putative *S*-like alleles, which show high sequence identity to known *SRK* alleles (the female determinant of self-incompatibility) in *Brassica* and *Arabidopsis*, which demonstrate high levels of synonymous and nonsynonymous variation. I also identify orthologs of two other members of the *S*-receptor kinase gene family, *Aly*8 (*ARK3*) and *Aly*9 (*AtS1*). Further to this, I demonstrate co-segregation between putative *S*-alleles and compatibility phenotypes using a full-sibling cross design.

2.2 Introduction

The promotion and maintenance of outbreeding in the Brassicaceae can be attributed to a genetically controlled self-incompatibility system (SI) (Bateman 1951). The SI system is controlled by a single Mendelian locus (the *S*-locus), which is comprised of a number of tightly linked genes, coding for pollen-pistil recognition, which function as a single locus. Allelic variation at this locus provides distinct mating specificity. Recognition of shared alleles at the *S*-locus will result in rejection of 'self' pollen by the stigma, preventing individuals that share *S*-alleles from producing seeds (Bateman 1955). Incompatibility reactions therefore promote both the avoidance of inbreeding, and the maintenance of intra-population genetic variability (Charlesworth and Charlesworth 1987).

The sporophytic SI system within the Brassicaceae relies on multiple tightly linked genes that effectively function as a single locus, which encode specific proteins involved in recognition and rejection of 'self' pollen (Kusaba *et al* 2001; Shiba *et al* 2003; Suzuki *et al* 1999). These genes are highly polymorphic (Awadalla and Charlesworth 1999; Nishio and Kusaba 2000; Sims 1993), due to negative frequency-dependent selection, whereby rare alleles have a reproductive advantage in the population, which promotes the maintenance of extensive nucleotide and allelic diversity at the *S*-locus (Schierup *et al* 1998; Takahata 1990; Vekemans and Slatkin 1994). The system is also subject to complex dominance interactions among alleles (Hatakeyama *et al* 1998; Ockendon 1975; Prigoda *et al* 2005; Stevens and Kay 1989; Thompson and Taylor 1966), which affects both the male (pollen) and female (pistil) components, (Shiba 2002; Thompson and Taylor 1966; Visser *et al* 1982).

Two highly polymorphic genes are involved in the SI recognition reaction in Brassica: *SCR* (*S*-locus cysteine rich), also known as SP11 (Takayama *et al*

2000), is the male determinant of the SI response (Schopfer *et al* 1999); and *SRK* (*S*-locus receptor kinase), a plasma-membrane protein localised in the stigma, is the female determinant in the SI response. The current model of SI response suggests that *SCR* interacts with the extra-cellular domain of the *SRK* protein to form a complex which initiates a signalling cascade, inhibiting ‘self’ pollen tube growth (reviewed by Chapman and Goring 2010).

The *S* “superfamily” of genes is composed of members that share sequence similarity to genes derived from the *S*-locus (which includes *SRK* and *SCR*). In crucifers, several *S* family members have been delimited and these include the *S* locus-related genes (*SLR1*) (Lalonde *et al* 1989; Trick and Flavell 1989) identified originally in *Brassica oleracea*, with orthologs in *Arabidopsis thaliana* (*AtS1*) and *A. lyrata* (*Aly*9) (Charlesworth *et al* 2003b; Dwyer *et al* 1994; Schierup *et al* 2001), and the *Arabidopsis* Receptor Kinase gene (*ARK3*) (Dwyer *et al* 1994; Tobias *et al* 1992), identified as *Aly8* in *A. lyrata* (Charlesworth *et al* 2003b). Although more *S* family members have been characterised in *Brassica, Arabidopsis thaliana* and *A. lyrata* (Luu *et al* 2001), the majority appear not to be linked to the *S*-locus (Kai *et al* 2001). It has been demonstrated that *Aly8* is linked to the *S*-locus (Kusaba *et al* 2001), whereas *Aly9* is not (Charlesworth *et al* 2003). Both, however, have been implicated in the SI reaction.

2.2.1 Breakdown of self-incompatibility (SI)

Despite the clear benefits of inbreeding avoidance (reviewed by Charlesworth 2003), the loss of a functioning SI system is not uncommon at the species level in a wide range of angiosperms (Weller and Sakai 1999). A variety of mechanisms have been proposed to account for the loss of SI among members of the Brassicaceae, including the recently reported 213 base pair inversion of the *SCR* gene, which is associated with a high proportion of European accessions of the self-compatible (SC) species *Arabidopsis thaliana* and its derivative haplotypes are thought to be responsible for loss of SI (Tsuchimatsu *et al* 2010). Nasrallah *et al* (2002) demonstrated that introduction of functional *SCR* and *SRK* allele copies from *A. lyrata* into SC *A. thaliana* can partially restore SI function in some *A. thaliana* accessions. This suggests that fixation of a non-functioning allele may be responsible for the loss of SI in this species. However, Cabrillac *et al* (2001) demonstrated that a downstream component of the rejection response (*ARC1*) in *Brassica oleracea* might be responsible for SI loss, suggesting that regulation of signal transduction may also be important for modifying the SI/SC response.

Range expansion has been postulated to favour species (or populations) capable of self-fertilisation (Baker 1967). Glaciations have had a large impact on population structure, especially in Europe, where climate warming after the last ice-age (c.18, 000 years bp) has brought about recent rapid range expansion in many species (Hewitt 2004). Bottleneck effects in small founding populations can cause depletion of genetic diversity and increase the influence of genetic drift (Keller and Taylor 2008; Nei *et al* 1975). In plants with a genetically controlled self-incompatibility system, this could lead to a reduction in numbers of *S*-alleles present and fewer compatible mating partners in recently founded populations (Vekemans *et al* 1998). Mate choice could be further limited by low plant densities, which attract fewer insect pollinators (Levin and Kerster 1969). In these situations it is predicted that self-fertilisation will present a reproductive advantage (Baker 1967), which may outweigh the potentially negative effects of inbreeding (Baker 1955; Pannell and Barrett 1998), especially during periods of rapid postglacial expansion (Baker 1966; Busch and Schoen 2008; Pannell and Barrett 1998), even for species operating a strong SI system.

Population-level variation in mating system is seen extensively throughout the Brassicaceae and has been reported in many genera, including *Arabidopsis* (Foxe *et al* 2010; Mable *et al* 2005; Mable and Adam 2007), *Brassica* (Hinata *et al* 1995), *Capsella* (Paetsch *et al* 2006), *Leavenwortia* (Koelling *et al* 2010) and *Raphanus* (Okamoto *et al* 2004). In addition, within a species, quantitative variation in the strength of the SI response is a phenomenon that has long been recognised and may be an important transition to SC (Good-Avila and Stephenson 2002; Levin 1996; Nielsen *et al* 2003; Vogler and Stephenson 2001), which means that the SI response of a self-incompatible plant can be 'leaky' to such an extent that it exhibits partial self-compatibility (PSC).

There is some evidence from populations experiencing multiple losses of SI in the genus *Arabidopsis* that supports the idea that PSC could be an evolutionary transition to an increased level of self-fertilisation. PSC is typically characterised by the production of small fruits with few or small seeds, following natural or enforced self-pollination, or a weakening of the SI response corresponding to flower age (Liu *et al* 2007; Stephenson *et al* 2003; Vogler and Stephenson 2001). In North American *A. lyrata*, there is clear variation in the strength of SI at both the individual and population levels (Mable *et al* 2005), but only some populations have shifted to inbreeding (Foxe *et al* 2010; Mable and Adam 2007) and PSC individuals show similar levels of genetic variation to their SI neighbours (i.e. from the same population (Hoebe *et al* 2009). Within this species, it is possible that the loss of SI might have started with weakening of the incompatibility response in older flowers (Liu *et al* 2007), which may be selectively favored during post-glacial colonisation (Mable *et al* 2005). Further to this, evidence suggests multiple independent losses of the SI reaction and shift to inbreeding in *Arabidopsis* (Foxe *et al* 2010; Hoebe *et al* 2009; Nasrallah *et al* 2004; Shimizu *et al* 2004), and that the loss of variation at the *S*-locus has been gradual (Bechsgaard *et al* 2006). The low variability (in comparison to the genomic average) of *S*-locus pseudo-genes in *A. thaliana* may add support to this theory (Tang *et al* 2007). Together these independent evidence streams could tentatively suggest that natural selection, facilitated by reduced mate choice during colonisation, may lead to the gradual fixation of multiple, independent mutations, which weaken or disable the SI system throughout the geographic range of a given species (Busch and Schoen 2008).

2.2.2 Arabis alpina

Arabis alpina is a distant cousin of the model plants *Arabidopsis thaliana, A. lyrata*, and *Capsella bursa-pastoris* (Beilstein *et al* 2006), and is being developed as a system to study the ecological genetics of alpine environments (Poncet *et al* 2010; Wang *et al* 2009). *Arabis alpina* is a perennial herb with an ecological preference for cool disturbed habitats, reproducing either sexually or asexually via stoloniferous growth (Hegi 1986). In Europe, it is distributed in the arctic areas of Iceland and Scandinavia, and alpine regions of the Pyrenees, Alps, Apennines, Balkan, Carpathian and Tatras mountain systems (Jalas & Suominen 1994). All of these areas have been dramatically affected by Pleistocene glacial cycles (Hewitt, 2004), resulting in alpine plants that have genetically highly structured populations (Comes and Kadereit 2003; Schonswetter and Tribsch 2005; Zhang *et al* 2004). Traits that influence range expansion are thus expected to be important in shaping genetic structure in *A. alpina*.

At present, most molecular studies have largely focused on the population structure and phylogeography of *A. alpina* (Ansell *et al* 2008; Assefa *et al* 2007; Ehrich *et al* 2007; Koch *et al* 2006). These suggest that Tatras and Apennine mountain populations most likely represent Pleistocene (c. last 2.5Mya) glacial refugia, contributing to the post-glacial recolonisation of the Alps (Ehrich *et al*, 2007; Ansell *et al*, 2008), A regional analysis of populations from the Italian Apennine refugia and nearby Maritime Alps using allozymes revealed substantial local variation in within-population inbreeding (F_{IS}) (Ansell *et al*, 2008). Populations from the Apuani portion of the Apennines range were close to Hardy-Weinberg equilibrium (HWE) (F_{IS} = 0.076, 95% CI 0.014-0.131), whereas those from the narrow Maritime Alps range showed a large distortion from HWE (F_{IS} = 0.553, 95% CI 0.457-0.620), suggesting the possibility of mating system variation among European populations. Local deviations in HWE may alternatively be consistent with local differences in extent of glacial survival and bi-parental inbreeding on post-glacial population recovery. I note the southern Maritime Alps are a small peri-glacial refugium (Ansell *et al* 2008, Grassi *et al* 2009) in comparison to Apuani Alps, which is a component of the major Apennine refugium. At present both the characterisation of a function SI system and S-locus like genes are lacking for *Arabis alpina*. Modifications in SI status over the species may be important when interpreting patterns of ecological variation in this developing model organism.

Here I investigate whether the apparent mating system variation based on F_{IS} estimates in *A. alpina* corresponds with a functioning SI system. Specifically, I evaluated whether: (1) populations of *A. alpina* with F_{IS} estimates that suggest they are outbreeding or inbreeding are consistent with outcrossing rate estimation based on multi-locus progeny arrays; (2) those populations with low F_{IS} estimates are functionally self-incompatible, and as such are not self-fertilising based on controlled self-pollination experiments; (3) functional self-incompatibility is determined by a sporophytic SI system, based on sequence similarity to the female component of SI, *SRK*; (4) variants of *SRK*-like sequences are linked to self-incompatibility phenotypes, based on segregation within families. Since self-compatible populations of various Brassicaceae species

(Mable *et al* 2005 (Bechsgaard *et al* 2006; Kusaba *et al* 2001; Tang *et al* 2007; Guo *et al* 2009) retain *S*-like alleles with high sequence similarity to functional alleles, it is critical to demonstrate: (A) putative *S*-allele sequences should exhibit high sequence identity to known *SRK* alleles; (B) putative orthologs must exhibit high levels of synonymous and non-synonymous diversity, and (C) such sequences should exhibit co-segregation suggestive of a single Mendelian locus, as inferred from diallel crosses within families. I also present an updated genealogy of *SRK*-like sequence variants across species that vary in mating system and reiterate the importance of linking allelic to phenotypic variation.

2.3 Methods

2.3.1 Samples

Seeds of *Arabis alpina* (diploid, 2x = 16, (Koch *et al* 1999)) were collected from six populations in the Maritime Alps and Apuani Alps of the Apennine range to represent the differing levels of within-population inbreeding (F_{IS}) reported by an earlier study of allozyme variability (Ansell *et al* 2008). Three populations with significant deviations from HWE (mean F_{IS} =0.525) were sampled from the Alps in Italy (Presolana) and France (Val de Roya 1 and 2), and three populations close to HWE (mean F_{IS} =-0.023) were sampled from the Apuan-Alps of Italy (Rotondo, Colli and Porte Strazzema) (Figure 2-1; Table 2-1). Open-pollinated seeds from individual mothers were collected in the field and stored separately in paper bags.

Between 9 and 13 seeds were germinated for each mother, from between 4-25 mothers per population (3 individuals per mother for selfing phenotype determination, and 6-10 individuals for outcrossing rate estimation). A further

18 individuals, the F1 progeny of cross pollination between a single individual each from Rotondo and Colli were also germinated for the segregation analysis. Seeds were sown on Levington S2 + Sand mix (Scotts Professional, Ipswich) and were germinated in controlled climatic incubators under a regime of 16 hours light (20ºC) and 8 hours dark (16ºC), with 80% relative humidity. After six weeks, the seedlings were vernalised for a period of 12 weeks under a regime of 8 hours light and 16 hours dark, at a constant temperature of 4^{o}C. Plants used for within-family crosses were maintained within the climate-controlled incubators, and those used for selfing phenotype determinations were transferred and maintained in non-climate controlled greenhouses.

2.3.2 Outcrossing rates

The magnitude of naturally occurring outcrossing rates was determined by analysis of allozyme variation among the progeny arrays from each population. Leaf tissue was harvested at the four-week stage from between 6-10 offspring per mother and from 14-20 mothers per population. Tissue was ground in a Tris-HCl (pH 7.5) extraction buffer (Soltis 1983), and proteins were fractionated on 12.5% hydrolysed potato starch gels under standard procedures (Vogel *et al* 1999; Wendel 1989). The following enzyme systems were then resolved with the lithium borate electrode and gel buffer system 8 of Soltis *et al* (1983): acid phosphatase (ACP, E.C. 3.1.3.2), aspartate aminotransferase (AAT, E.C. 2.6.1.1), glyceraldehyde-3-phosphate dehydrogenase (G-3-PDH E.C. 1.2.1.9), leucine aminopeptidase (LAP E.C. 3.4.11.1), malic enzyme (ME, E.C. 1.1.1.40), phosphoglucose isomerase (PGI, E.C. 5.3.1.9), shikimate dehydrogenase (SKDH, E.C. 1.1.1.25), triose-phosphate isomerase (TPI, E.C. 5.3.1.1). The following enzymes were then resolved using the morpholine-citrate (pH 7.1) electrode and gel buffer system of Wendel and Weeden (1989): alcohol dehydrogenase (ADH, E.C. 1.1.1.1), isocitrate dehydrogenase (IDH, E.C. 1.1.1.42), malate dehydrogenase (MDH, E.C. 1.1.1.37), phosphoglucomutase (PGM, E.C. 5.4.2.2), 6-phosphogluconate dehydrogenase (6-PGD, E.C. 1.1.1.44) and utp-glucose pyrophosphorylase (UGPP, E.C. 2.7.9). Staining patterns were interpreted using known enzyme sub-structuring (Kephart 1990).

Multi-locus outcrossing rates (T_m) were calculated from the progeny arrays of each population, using the program MLTR v2.3 (Ritland 1981), which implements the mixed-mating model described by Ritland (2002). Using a maximum likelihood approach, this program calculates single (T_s) and multi locus (T_m) outcrossing rates and predicted allele frequencies in pollen (p) and ovules (o). The difference between T_s and T_m indicates the degree of biparental inbreeding. Standard errors were calculated by bootstrapping across progeny arrays (10,000 replicates). Values of the starting parameters were: 0.1, increasing in steps of 0.1 to a maximum of 0.9 for t; 0.1 for *rt*, *rp*, and *f*; specifying unconstrained pollen and ovule gene frequencies. Results with the lowest standard error are reported. Populations showing outcrossing rates < 0.30 were classified as inbreeding (IB); > 0.80 as outcrossing (OC), and between 0.31-0.79 as mixed mating.

2.3.3 Selfing phenotype determination

One plant per maternal seed family was tested for strength of self-incompatibility (SI), through manual self-pollination of six replicate flowers, to classify its selfing phenotype. Fruits were scored as negative (no seeds), small (short fruits with no more than 3 seeds), or positive (a full-sized fruit with more than three seeds). For each plant, selfing phenotype was classified according to the following scheme: 1) SI (self-incompatible) = individuals producing zero or one full-sized selfed fruits; 2) SC (self-compatible) = individuals producing at least five full-sized selfed fruits; and 3) PSC (partially compatible) = individuals producing two to four full-sized selfed fruits. The proportion of plants capable of setting self-seed was used to categorise each population as predominantly SI (more than 80% SI individuals), predominantly SC (fewer than 30% SI individuals), or of mixed type (between 30 and 80% SI individuals). Putatively SI individuals were tested for sterility in a sub-set of individuals used for selfing phenotype determination by performing cross-pollinations with SC individuals from Presolana. The Presolana population was chosen for this as it effectively removed the possibility that negative fruits were the result of incompatibility. Three individuals from Colli and three from Rotondo were pollinated, with three replicate pollinations per individual.

In order to distinguish “leaky” small fruits, which have been interpreted to be the result of occasional pollen tube escape from otherwise incompatible parental combinations (Mable *et al* 2003), from full sized compatible fruits with high seed abortion, mature fruit length was measured for all selfed progeny from the selfing phenotype determinations and cross progeny from the diallel cross (see below), and the number of both mature and aborted seeds was counted. A zero-inflated Poisson regression in R v2.10 (R Foundation for Statistical Computing) was used to see if discrete groups were formed between the three categories (incompatible, small, and compatible), with small fruits classified as incompatible.

2.3.4 S-allele genotyping

Leaves were collected from the individuals used to assess selfing phenotype, and were dried using silica gel (Drierite, 8 MESH, Acros Organics, New Jersey USA). DNA was extracted by two methods: (1) from 100 mg of silica-dried tissue using the FastDNA kit, following the manufacturer's instructions (QBiogene 101, MP Biomedicals); or (2) by DNeasy 96- well plate kits (Qiagen Inc), using the Genome laboratory service at the John Innes Center (Norwich, UK). Degenerate primers designed in conserved regions of the *S*-domain, which are known to amplify *SRK* in *Arabidopsis* and *Brassica* (formerly called *Aly*13 in *A. lyrata)*, using primers 13F1 + SLGR & 13-FBM + SLGR (Mable *et al* 2005; Mable *et al* 2003), were first used to amplify the putative *SRK* fragment from one individual per mother, for each mother used in the self-fertilisation experiment (Table 2-1). Amplification products of the expected size (approximately 1 kb) were cloned from six individuals per population and from different seed families (with the exception of Val de Roya 2, where four individuals were screened), using TOPO-TA cloning kits for sequencing (Invitrogen, vector PCR TOPO 4.1). In addition, DNA extracted from field-collected leaves of two individuals from Corsica, and five individuals from Italian populations that varied in F_{IS} (Table 2-1) were also amplified and cloned. For each cloned individual, between 12 and 20 colonies were screened. Plasmid DNA was extracted and cleaned using QIAprep spin miniprep kits (Qiagen Inc. Valencia, CA) and sequencing was carried out with the universal primers M13F and M13R on an ABI 3730 sequencer (by The Sequencing Service, at the University of Dundee and The Gene Pool, at the University of Edinburgh). Sequences were assembled and manually checked for base-calling errors using Sequencher v4.7 (Gene Codes, Inc.). Consensus sequences were obtained by correcting PCR mutations to the nucleotide found in the majority of clones for a given individual Each new sequence variant found in *A. alpina* was given a unique identifier. Sequences were manually aligned using SE-AL v2.0a11 (Rambaut 2002) to known *SRK* alleles (Appendix 1).

The alignment was used to design sequence-specific primers to amplify unique *A. alpina SRK*-like sequences which were present in three or more colonies using Oligo v6.86 (Molecular Biology Insights, Inc.). A single individual per maternal line for each of the six major populations used in this study, were then screened for each of the sequence-specific primers designed. Positive PCR bands were excised and sequenced to assess sequence fidelity and putative allele frequencies within populations.

2.3.5 Diallel crosses, segregation analysis and linkage

In order to test whether putative SRK genotypes were associated with SI phenotypes, a diallel crossing strategy within families was taken, using seeds raised from a cross between two SI populations, Rotondo and Colli. Cross-pollinations among full-siblings were performed by rubbing anthers on stigmas of intact flowers, after emasculation of the recipient. Each combination was pollinated reciprocally, using three replicate flowers. Compatibility was scored as fruit set 7-10 days after pollination using the criteria described above for self-pollinations. A total of 18 plants were used in the crosses, but not all plants flowered sufficiently to produce a complete diallel cross design. Pollinations were conducted over a 4-month period (April-July), because flowers are produced in small numbers at any given time; mature fruits were collected after 4-7 weeks, when mature.

Individuals were classified into incompatibility groups based on the outcome of within-family crosses only, with no prior knowledge of sequence data. Progeny were then PCR screened for the *SRK*-like sequence variants found in the parents, in order to evaluate whether self-incompatibility phenotypes corresponded with putative *S*-genotypes. In order to determine whether distribution of genotypes in the progeny was consistent with segregation at a single Mendelian locus, a χ^2 test was conducted in R v2.10 (R Foundation for Statistical Computing).

2.3.6 Phylogenetic analysis and sequence divergence estimates

Sequences were assembled using Sequencher v4.7 (Gene Codes), and then manually aligned against previously published *SRK* allele sequences (Mable *et al* 2005; Schierup *et al* 2001) from eight species (Appendix 1) using Se-al v2.0a11 (Rambaut 2002). Further adjustments were made to alignments by manual correction and by translating the data into amino-acid characters, using MacClade v4.0 (Maddison and Maddison 2000). Synonymous and non-synonymous divergences between new putative alleles and known alleles from *Brassica* and *Arabidopsis* were computed according to Nei and Gojobori (1986), using DNAsp v3.53 (Rojas and Rojas 1999).

Phylogenetic relationships among sequences were visualised using a Bayesian inference of phylogeny conducted in MrBayes (Ronquist and Huelsenbeck 2003), using a General Time Reversible model (GTR) (Tavare 1986) with gamma rate variation, chosen using ModelTest v3.7 (Posada and Crandall 1998). The MCMC algorithm used 8 heated chains, with a chain length of 2,000,000 and a burn-in of 200,000. Other *S*-domain genes were included in the phylogeny, including *Aly*8 (ARK3) from *A. lyrata*, which was used as an outgroup and *Aly*9 (SLR1 or *AtS1*) from *A. lyrata*. The phylogeny was intended for visualisation of geneological relationships only, since SRK is under strong selection pressure with the possibility of gene conversion among gene family members (Prigoda *et al* 2005), so it is not strictly appropriate to reconstruct a bifurcating phylogenetic tree.

2.4 Results

2.4.1 Outcrossing rates and selfing phenotypes

From the 13 enzyme staining systems used, I detected 16 loci, of which 11 were polymorphic and 37 alleles were discriminated. Outcrossing rates based on allozyme progeny arrays (Table 2-3) varied from the extreme inbreeding of Presolana (PRES: T_m = 0.00), where no allozyme variation was recorded, to the extreme outcrossing of population Rotondo (ROTO: T_m = 0.978 ± 0.066, T_s = 0.812 ± 0.051). Porte Strazzema (POST: T_m = 0.889 ± 0.046, T_s = 0.765 ± 0.028) and Colli (COLL: T_m = 0.779 ± 0.080, T_s = 0.652 ± 0.037) were classified as outcrossing, while Val de Roya 1 (VDR1: T_m = 0.144 ± 0.075, T_s = 0.134 ± 0.020) was classified as inbreeding. None of the populations were classified as mixed by outcrossing rate, based on a T_m criteria of <0.3 and >0.8 for inbreeding and outcrossing, respectively. In all cases, the population outcrossing rate classifications are consistent with the previous estimates of F_{IS} values from the wild populations (Table 2-3). I were unable to test the outcrossing rate of population VDR2 due to insufficient individuals that germinated; however, allozyme variation was low and the previous F_{IS} value (Ansell *et al* 2008) was relatively high (0.4), suggesting that it is probably inbreeding.

Assessment of biparental inbreeding ($S_b = T_m - T_s$) (Lu 2000) in each population revealed varying degrees of this form of inbreeding, with outcrossing populations ROTO (S_b = 0.166), POST (S_b = 0.134) and COLL (S_b = 0.127) exhibiting some evidence of biparental inbreeding and inbreeding population VDR1 (S_b = 0.010) exhibiting very low levels; however, this could be an artefact of the overall lack of variation in this population.

Hand pollinations of emasculated flowers revealed that there were clear differences between populations with respect to selfing phenotypes (Table 2-3). Although all populations tested for selfing phenotype contained at least some individuals capable of self-fertilisation, there was variation in the proportions of selfing phenotypes within populations, which corresponded well with predicted outcrossing rates and F_{IS} values. I found a single population that contained exclusively SC individuals (PRES) and functioned as an autonomous selfer. This is consistent with its low T_m and high F_{IS} estimates, which are effectively 0 and 1 respectively, due to absence of variation. VDR1 showed no individuals with an SI phenotype, but had 47% PSC individuals and 53% SC individuals; again, consistent with its low T_m and high F_{IS} estimates. For ROTO, 80% of individuals screened exhibited an SI phenotype, with 13% and 7% of individuals exhibiting PSC and SC, respectively; consistent with its high T_m and low F_{IS} estimates. A similar result was observed for the COLL population, which was also highly outcrossing (SI 73%, PSC 20% & SC 7%). Together, these results suggest that outcrossing rate is a good indicator of the proportion of self-incompatible individuals within populations. Individuals from POST and VDR2 failed to flower, so selfing phenotypes could not be assigned. Tests for fertility of apparently SI individuals, as assessed using a subset of individuals, showed no male or female sterility for any of the individuals tested, suggesting that all were actually SI.

For the Italian populations for which I had only individual DNA samples available, F_{IS} estimates (Ansell *et al* 2008) (Table 2-1) would suggest that the Madonie Mts population in Sicilly is outcrossing (F_{IS}: 0.164), as are Mt Pollino (F_{IS}: -0.046) and Mt Cava (F_{IS}: -0.005) in the Apennines. In contrast, populations in the central Alps (Largo Garda 29, F_{IS}: 0.860; and Largo Garda 30, F_{IS}: 0.654) would be predicted to be inbreeding. The population from Corsica has not been studied previously and only two individuals were sampled so no F_{IS} values are available.

Selfed fruits for the outcrossing individuals showed discrete groupings for incompatible, and compatible fruits, but not for small fruits, such as has been seen in *A. lyrata* (Mable *et al* 2003). Incompatible fruits in outcrossing populations had a mean length of 9.25mm (Range: 6.22 – 18.34mm), and fruits with 1-3 selfed seeds (classified as 'small') had a mean length of 12.86mm, with a range that overlaps that of the incompatible category (9.51 – 18.42mm). As there was a clear distinction between 'full' compatible fruits from selfing populations (mean: 20.03mm range: 15.15 – 30.30mm) and 'small' fruits (2-tailed T-test: P = <0.00), I used a cut off of less than three seeds to indicate incompatible fruits.

Compatible fruits from the outcrossing populations were significantly smaller than those from inbreeding individuals (mean: 41.51mm range: 27.68 – 49.15mm) (2-tailed T-test: P = <0.00), and aborted seed number for individuals producing compatible selfed fruits in outcrossing populations (mean: 19.93 range: 7 – 29) was significantly higher than for inbreeding populations (mean: 6.63 range: 2 – 26) (2-tailed T-test: P = <0.00), which may suggest PSC with reduced fitness. Total seed number (viable and aborted) was higher in inbreeding individuals (mean: 27.20mm) than in outcrossing individuals (mean: 21.07mm).

2.4.2 Arabis alpina S-allele genotype determination

I identified 15 sequence variants that were more similar to *SRK* than to other members of the gene family, across the 41 maternal individuals screened via cloning. Eight of these variants occurred within the six focal populations where pollination testing was available, and seven occurred in additional individuals sampled from other parts of the Italian range for which seeds stocks were not established (Table 2-1) and for which the selfing phenotype is unknown. The degenerate primers used in this study are also known to amplify other members of the *S*-gene family, including *Aly*8, *Aly*9 and *Aly*10 in *A. lyrata* (*ARK*3, *SLR*1 or *AtS*1, and *ARK*1, respectively, in *A. thaliana*). DNA sequences that showed high sequence identity to these genes were recovered during cloning, and included nine sequences with high sequence identity to *Aly*8 (ARK3) (ALP-ARK3-1 to ALP-ARK3-9), and two sequences that exhibited high sequence identity to *Aly*9 (ALP-ALY9-1 & 2). One of these *Aly*8-like sequences, ALP-ARK3-1 was present in 54% of the individuals screened. The fifteen unique putative *SRK* allele sequences were assigned sequential names based on the order in which they were detected (ALPSRK1 to ALPSRK15).

Within the six focal populations, PCR screening by allele-specific primers designed for each of the eight sequence variants showed that the putative alleles generally occurred at low frequencies within each population (Table 4-2), with a high percentage of individuals (64%) failing to produce PCR products with any of the primers tested after four PCR reactions. In addition, only six individuals were heterozygous, each containing only two of the putative *SRK* alleles and all were from outcrossing populations. Five of the heterozygous individuals occurred in the COLL population (individual 2 – ALPSRK 5&8; individuals 5 and 10 – ALPSRK 5&6; individual 7 – ALPSRK 3&6; individual 9 – ALPSRK 6&8), and one was from the ROTO population (individual 12 – ALPSRK 1&8). These heterozygous individuals also contained the *Aly*8-like sequence ALP-ARK3-1, which supports the hypothesis that this sequence is not *SRK*. It is clear from the low proportion of heterozygotes, and the high proportion of individuals that failed to generate allele-specific PCR products, that not all allelic diversity was recovered from these populations. This is further supported by the initial failure to recover amplification products for 16% of individuals (from among all populations) using the degenerate primers developed for *Brassica* and *Arabidopsis SRK* loci, suggesting that divergent alleles specific to *A. alpina* likely have been missed.

2.4.3 Diallel crosses, segregation analysis and linkage

Three seed families were raised from crosses. However, only a single seed family flowered sufficiently for the segregation of SI phenotypes and genotypes to be tested. The SI parents used in this cross (maternal: COLL, and paternal: ROTO) each had only a single putative *SRK* allele identified (ALPSRK6 in COLL; ALPSRK8 in ROTO). However, if a functioning sporophytic SI system is operating, individuals are expected to be heterozygous, so there was the potential for four genotype combinations per offspring, with unidentified alleles labelled X and Y (i.e., 6&X, 8&Y, 6&8 and X&Y). Based on compatibility scored as production of a full sized fruit, individuals were partitioned into four tentative phenotype groups. Individuals were then screened using PCR and allele-specific primers the two putative parent alleles. Of the 18 individuals used in this experiment, I assigned seven individuals to the genotype 6&X, six to 6&8, three to 8&Y and the remaining two to X&Y. The χ^2 test for distribution of genotypes showed no significant deviation from expected genotype frequencies (χ^2 = 2.0193, df = 3, p-value = 0.5684), which suggests that all four putative alleles are present at the same locus. Although a complete diallel cross could not be performed due to insufficient flowering, with some exceptions, a single genotype could be associated with each phenotype group (Figure 2-2). However, the inconsistent compatibility of individuals 3 and 12 with others of the same genotype could suggest that they are self compatible (although it is also possible that dominance in pollen and stigma is different). Individual 19 showed different patterns when it was the pollen donor or recipient, again suggesting different dominance in pollen and stigma. Individual 2 had the genotype 6&X but its compatibility phenotype suggested that it also had the ALPSRK8 allele, which failed to amplify. Selfings, represented by the diagonal (shaded black) were not performed in order to maximize the number of cross combinations possible with limited flowering.

2.4.4 Relationship between putative S-allele sequences and know SRK alleles

Figure 2-3 shows a Bayesian phylogenetic tree that summarises relationships among the putative *S*-allele sequences amplified from *A. alpina*, as well as among two other members of the *SRK* gene family (*Aly8* and *Aly9*) in *Brassica* and *Arabidopsis*. Of the fifteen identified putative *S*- alleles, all show close nucleotide similarity to *SRK* alleles amplified in *A. lyrata*, regardless of the selfing phenotype of the individual it was isolated from. The selfing phenotypes of individuals with ALPSRK4 and ALPSRK9-ALPSRK15 are unknown, as they were identified from the Italian samples that lacked seed collections. The putative *A. alpina S*-alleles fall within four clades, with the sequence variants contained within each showing high sequence identity to known *SRK*-alleles from other species. The first clade includes ALPSRK7-ALPSRK13, and shows high sequence identity to the *A. lyrata* allele ALYSRK19. The putative alleles within this clade are not isolated to a single geographic area or population. The second clade includes the putative *S*-alleles ALPSRK1, ALPSRK2, ALPSRK5, ALPSRK6 and

ALPSRK15. This clade includes known *SRK*-alleles from both *A. lyrata* (ALYSRK15, ALYSRK23 and ALYSRK42) and *A. halleri* (AHASRK12 and AHASRK21). The third clade includes only ALPSRK3, which shows high sequence identity with ALYSRK33 and ALYSRK4 from *A. lyrata* and AHASRK20 from *A. halleri*. Finally, ALPSRK4 and ALPSRK14, both identified from populations on Corsica, show high sequence identity to *A. lyrata* alleles ALYSRK9 and ALYSRK46, along with *A. halleri* allele AHASRK22. In addition to the putative *SRK*-alleles identified in *A. alpina*, I also found high sequence identity among the sequences that were more similar to other members of the S-receptor kinase gene family (Charlesworth *et al* 2003b). The putative *Aly*9 alleles (ALP-ALY9-1 and ALP-ALY9-2) cluster with *Aly*9 alleles from *Arabidopsis lyrata*. Our putative *Aly*8 (ARK3) (ALP-ARK3-1 to ALP-ARK3-8) alleles also cluster with the outgroup *Aly*8 allele from *A. lyrata*.

Table 2-5 summarises pair-wise diversity between putative *S*-alleles, and known *S*-alleles from other species (*A. lyrata*, *A. halleri*, *Brassica napus* and *B. oleracea*). It demonstrates high sequence divergence for both synonymous (Ks) (Mean: 0.3584, Range: 0.008-0.530, SE: 0.039) and non-synonymous (Ka) sites (Mean: 0.3199, Range: 0.026-0.5026, SE: 0.0166). None of the putative *S*-alleles contain stop codons, which would indicate lack of functionality.

2.5 Discussion

The observations made in this study support the hypothesis for a sporophytic SI system in *Arabis alpina,* as have been reported extensively for other species in the Brassicaceae (Bateman 1954; Bateman 1955; Llaurens *et al* 2008; Mable *et al* 2003; Schierup *et al* 2001). Both the outcrossing rates and the selfing phenotype assignments correspond to the previously calculated F_{IS} estimates from Ansell *et* *al* (2008) for the same populations. Three populations with low F_{IS} (ROTO, POST & COLL) here present clear evidence of being predominantly outcrossing (Table 2-3), rather than inbreeding, based on allozyme outcrossing rates. These data are corroborated by greenhouse self-fertilisation experiments, which demonstrate that there is a functional genetic barrier preventing individuals within the outcrossing populations from self-fertilisation (Table 2-3), suggesting a functional SI system. In contrast to this, three populations that showed high F_{IS} in Ansell *et al* (2008) (VDR1 & 2 and PRES) demonstrate an absence of an inbreeding avoidance mechanism based on both their outcrossing rate, and their selfing phenotypes (Table 2-3). At the molecular level, I amplified fifteen sequences (ALPSRK1 to ALPSRK15) that show strong sequence similarity to known *SRK* alleles sampled from species of *Arabidopsis* and *Brassica* (Mable *et al* 2003; Schierup *et al* 2001). I also identified sequence variants that show sequence similarity to other members of the *S*-receptor kinase gene family, including sequences with high similarity to *Aly* 8 (ARK3) and *Aly*9 (SLR1) (Figure 2-3). The within-family crosses demonstrate a link between the putative *SRK* alleles and the selfing phenotypes of individual plants, although limitations due to poor flowering did not allow estimate of dominance because a full diallel cross design could not be completed. Further to this, I experienced problems associated with cessation of flowering at temperatures above 20-22°C in the greenhouse, and a total lack of flowering in certain populations (POST and VRD2), which prevented selfing phenotypes being assigned.

2.5.1 Variation in strength of SI across populations

The ROTO population exhibited a large proportion of SI individuals, although despite its very high T_m, it still contained variation in mating system, with both PSC and SC individuals, albeit at low frequency (Table 2-3). Five of the putative *S*-alleles were amplified in this population (ALPSRK1, 2, 6, 7 & 8), with only a single individual identified as heterozygous. The frequency of these alleles was generally low, as would be expected from *S*-alleles, with the exception of ALPSRK 7, which was present in 25% of individuals screened. This pattern was repeated for the COLL population, which also showed variation in mating system strength, again with PSC and SC individuals. Putative alleles ALPSRK 1, 3, 5, 6 & 7 were amplified in this population, with five individuals identified as heterozygous. Allele frequency for ALPSRK 6 and ALPSRK 7 was high, with 30% and 45% of individuals tested containing each allele respectively.

The PRES population exhibited a complete lack of SI response in greenhouse selfing experiments, and produced autonomous self fruits (Table 2-3). This, coupled with a lack of allozyme variability in this and an earlier study (Ansell *et al* 2008), would suggest that this population has no functional SI system. Nevertheless, I amplified an *SRK*-like allele (ALPSRK6) from three individuals in this population, which shows high sequence identity to known *SRK*-alleles in SI populations of *A. lyrata* (ALYSRK23 and ALYSRK42) and *A. halleri* (AHASRK12). The retention of high sequence similarity to *SRK* from SI species/populations might indicate a recent (post-glacial) loss of a functioning SI system in this population, as alleles that are no longer under selection are expected to be subject to increased mutation rate (relative to *S*-alleles under selection in a functioning SI system) (Nasrallah 1997). However, SC *A. thaliana* retains recognisable *SRK* and *SCR* sequence-elements (as pseudogenes), despite the loss of SI system function having been dated between 500 thousand and 5 million years ago (Bechsgaard *et al* 2006; Nasrallah *et al* 2004; Shimizu *et al* 2004; Tang *et al* 2007; Kusaba *et al* 2001). Sequence analysis of the PRES *SRK*-like ALPSRK6 allele failed to identify premature stop-codes that would be consistent with pseudogene formation. However, as I only sequenced a fragment of the extracellular *S*-domain, it is possible that there are stop codons in another domain.

Population VDR1 demonstrated little evidence of functional inbreeding avoidance based on greenhouse selfing experiments, with exclusively SC or PSC individuals (Table 2-3). Outcrossing rate estimates from this population were less than 0.2, suggesting that inbreeding is common. Unlike PRES, there was no evidence of autonomous self-fertilisation, potentially suggesting different evolutionary histories of SI breakdown and/or shifts to selfing, or different selection pressures following the initial mating system change. This is consistent with evidence from inbreeding populations of *A. lyrata*, where there is variation in autonomous seed set within and among populations (Foxe *et al* 2010). Both of the VDR populations yielded only a single putative *S*-allele (ALPSRK8), which occurred in a clade with numerous other putative *SRK*-alleles identified from *A. alpina*, and are highly similar to the *A. lyrata SRK*-allele ALYSRK19. This allele was only amplified from a single individual from each population, so it is likely that these populations contain other, nonamplified *SRK*-like alleles.

Ansell *et al* (2008) suggested independent origin of the central Alps (which contains population PRES) and the western Alps regions (which contain the VDR populations), based on allozyme and cpDNA data. This division has also been supported by AFLP diversity structure in *A. alpina* (Ehrich *et al* 2007), and has been reported for other alpine herbaceous plants (Kropf *et al* 2003; Schonswetter *et al* 2003; Schonswetter *et al* 2004; Stehlik *et al* 2002). Our findings, which classify PRES as autonomously selfing, and the VDR1 population as capable of self-fertilisation but not autonomously selfing, tentatively agree with this phylogeographic pattern.

For the VDR1 population, it is not clear if the large proportion of individuals that exhibited partial compatibility (PSC) had a partially functional self-incompatibility system, the environment modified the self-incompatibility system, or they were SC individuals that received suboptimal pollen. Hoebe *et al* (2009) found that PSC and SI individuals of North American *A. lyrata* had similar estimates of heterozygosity by microsatellite analysis. They postulated that this was consistent with the possibility that PSC individuals are an early transition phase from an SI to SC, or that the environment could have induced the partial breakdown of the self-incompatibility system, and excluded the possibility that SC individuals received a sub-optimal pollen load. In *A. alpina*, I also found that PSC individuals were capable of producing outcrossed seeds when fertilised by unrelated individuals (data not shown), suggesting that they were not suffering decreased fertility. This could suggest that the VDR population is in a late stage of transition towards inbreeding, and both ROTO and COLL are in a very early stage of the transition.

The finding of self-compatibility and high levels of inbreeding in *A. alpina* is consistent with other studies. An AFLP study of central and northern European populations (Ehrich *et al* 2007) recovered a single multi-locus genotype among 89% of the Nordic populations (northern Germany to Greenland), suggesting that some presumably self-fertilising genotypes are highly successful or that there was a bottleneck associated with the transition to selfing in the common ancestor of these populations.

2.5.2 Diallel crosses and segregation analysis

Flowering in *Arabis alpina* is irregular and non-continuous in comparison to *A. thaliana*, reflecting its more apparent perennialised life-history (Wang *et al* 2009). Consequently, I was unable to generate complete diallelic crosses for testing the segregation of selfing phenotype and *SRK*-genotype. Even under vernalisation conditions considered appropriate for this species (Wang *et al* 2009), flowering was not sufficient to obtain a full picture of segregation of phenotypes with genotypes and dominance relationships, as has been done for other Brassica species (Charlesworth *et al* 2003a; Mable *et al* 2004; Mable *et al* 2003). The phenotypic groupings based on compatibility of crosses corresponded with the *SRK*-genotype in most cases, although not all individuals could be assigned to phenotypic groups with confidence. In particular, two genotypes were under-represented in the crosses (X&Y and 8&Y). In addition, since I only identified a single allele from each parent, individuals with poor amplification of either of these alleles might have been mistakenly assigned an X or Y allele. I suspect this may be the case for individual 2, for which PCR screening only resulted in amplification of ALPSRK 6 (subsequently leading to assignment to genotype 6&X) but its phenotype suggests that it also has ALPSRK8. Nevertheless, I was able to demonstrate a link between *SRK*-genotype and phenotype for the alleles used in this cross, which is of the utmost importance when identifying potential *S*-locus genes (Busch *et al* 2008; Charlesworth 2000). Selfing phenotype determinations would have been useful for the individuals used in the diallel cross because a few individuals showed inconsistent compatibility patterns consistent with being self-fertile. However, I decided to maximize the number of crosses possible and so selfings were not performed.

2.5.3 Multiple species SRK phylogeny

Although the use of a phylogeny for genes at the *S*-locus is not strictly appropriate, it allows visualisation of relationships between known *S*-domain alleles from well-characterised species (*A. lyrata, A. halleri* and *B. oleracea*) with putative alleles from species that have been less well characterised, including *C. grandiflora, C. rubella*, and *L. alabamica*. The relationships between putative *SRK*-alleles in *A. alpina* and known *SRK*-alleles in other species, particularly those in the genus *Arabidopsis*, appears to provide strong evidence that the sequence variants identified are orthologous to *SRK*. This phylogeny also presents evidence as to the identity of other *S*-domain genes amplified in *A. alpina* (*Aly8*-like and *Aly9*-like) and suggests that these are highly similar to those identified in other well characterised species. The *SRK* sequences for *Brassica* and *Arabidopsis* form separate clades, which likely reflects their long phylogenetic separation (Beilstein *et al* 2006). Putative *A. alpina SRK* alleles ALPSRK 4 and ALPSRK 14 show sequence identity to both genera, which is likely due to trans-specific polymorphism (i.e. sharing of alleles between species), and may reflect its phylogenetic placement intermediate to *Arabidopsis* and *Brassica* (Koch *et al* 2001).

This phylogeny also emphasizes the necessity to demonstrate linkage of putative *SRK* genotypes to mating system phenotypes, as contained within clades where linkage has been demonstrated, there are sequences which have been shown not to be linked to the *S*-phenotype. For example, unlinked sequences ALY13-2 and ALY13-7 show high similarity to a clade of alleles that have been demonstrated to be linked to the S-phenotype in both *A. lyrata* (Prigoda *et al*

2005) and *A. halleri* (Castric and Vekemans 2008). In *L. alabamica*, gene sequences were classified as *SRK* (LALSRK) due to high sequence identity to known *SRK* alleles, and linkage to self-incompatibility phenotypes (Busch *et al* 2008). LALSRK appears to show sequence identity to alleles from *A. halleri* (AHASRK14 and AHASRK25). *SRK* alleles from SC populations of *A. thaliana* also cluster with alleles from other species with functional SI systems, particularly *A. lyrata* and *A. halleri*. Alleles for the self-compatible *Capsella rubella*, which include apparent pseudogene sequences, show high sequence identity with *A. lyrata* alleles. Alleles for the self-incompatible *C. grandiflora* show high sequence identity to alleles from *A. lyrata*, *A. halleri* and self-compatible *A. thaliana*, linkage to selfing phenotype was not established however (Paetsch *et al* 2006).

2.5.4 Conclusions

Our study of Maritime Alp and Apauni Alp populations of *Arabis alpina* supports the earlier assertion, that differences in allozyme derived F_{IS} estimates are consistent with changes in mating system variation. I provide a firm case for the presence of a functioning sporophytic SI system in certain European populations, and tentative evidence for a loss of SI function in other populations. By performing transformation experiments similar to that conducted in some accessions of SC *Brassica rapa* and *A. thaliana* (Takasaki *et al* 2000; Takayama *et al* 2001), it may be possible to establish if the putative *SRK*-alleles reported in this study are necessary and sufficient to cause an SI response in the *A. alpina* populations that currently have no (or a reduced) SI response. I note that the potential changes in SI status coincide with areas suspected to differ in the extent of local glacial survival (Ansell *et al* 2008). However, it is premature to conclude the extent and context to which phylogeographic history has influenced variation in mating system of this interesting arctic-alpine species. *A. alpina* has the potential to be a useful model system for studying the ecological genetics of alpine adaptation, based on its extensive arctic-alpine latitudinal and altitudinal range, leaf shape and flowering time variation (Wang *et al* 2009, Poncet *et al* 2010). Along with its soon to be completed genome sequencing (Coupland and Weigel, unpublished), it is important that mating system variation is documented more widely across the species range. This is especially critical in the context of interpreting patterns of ecological genetic diversity, which may in part be influenced by local or regional stochastic changes to mating system variation.

Table 2-1 Sampled populations (ordered by F_{IS}), number of maternal individuals sampled, geographic coordinates, geographic region, and F_{IS} estimates (from Ansell *et al* 2008).

Sample type	Population name	Population Acronym	N	Geographic location		Region	F_{IS}
				Lat	Long		
Seeds	Presolana	PRES	18	N 45°57'59"	E 10°04'46"	central Alps	****[a]
Tissue	Lago Garda 29	LAG29	1	N 45°41' 00"	E 09°46'30"	central Alps	0.860
Tissue	Lago Garda 30	LAG30	1	N 45°48' 00"	E 10°01'00"	central Alps	0.654
Seeds	Val de Roya 1	VDR1	15	N 43°59'29"	E 07°30'17"	western Alps	0.648
Seeds	Val de Roya 2	VDR2	4	N 44°04'17"	E 07°31'20"	western Alps	0.401
Tissue	Madonie Mts 1	MAM1	1	N 37°52'38"	E 14°00'16"	Sicily	0.164
Seeds	Porte Strazzema	POST	20	N 43°59'07"	E 10°17'45"	Apuan Alps	-0.140
Seeds	Rotondo	ROTO	25	N 44°04'32"	E 10°22'50"	Apuan Alps	0.068
Tissue	Mt Pollino 5	MPO5	1	N 39°57'24"	E 16°10'56"	Apennines	-0.046
Tissue	Mt Cava 10	MCA10	1	N 42°14'48"	E 13°19'08"	Apennines	-0.005
Seeds	Colli	COLL	19	N 44°05'03"	E 10°19'09"	Apuan Alps	0.003
Tissue	Corsica	CORS	2			Corsica	N/A[b]

[a] = Due to absence of variation, F_{IS} could not be calculated for this population.

[b] = An F_{IS} estimate is not available for the population 'Corsica' because only two individuals were sampled.

Table 2-2 Primers used to amplify *S*-domain fragments for alleles present in European populations of *A. alpina*.

Primer[a]	Sequence	Allele amplified (with SLGR)	Fragment size[b] (bp)
13-F1	ccgacggtaaccttgtcatcctc	General	~1000
13-FBM	gtgagatctccggtggtggcag	General	~950
SLGR	atctgacataaagatcttgacc	General	
ALPSRK 1	cagtgagaacagacggaacac	ALPSRK 1	938
ALPSRK 2	gagtccaatagggacattgg	ALPSRK 2	932
ALPSRK 3	cactgactaacatagggttcc	ALPSRK 3	941
ALPSRK 4	gagttttgaccgattcactcc	ALPSRK 4	639
ALPSRK 5	aaccagcctgactatgacaac	ALPSRK 5	974
ALPSRK 6	atagaacgaagacgcgccaac	ALPSRK 6	950
ALPSRK 7	gaattcgataggatggaccttg	ALPSRK 7	591
ALPSRK 8	cccggaaacttggttgggtac	ALPSRK 8	579
ALP-ARK3-1	ggaaccagttttggtacgcacc	ALP-ARK3-1	568

[a] Allele-specific primers designed for this study are indicated in bold type.

[b] Expected size when products amplified using SLGR as reverse primer.

Table 2-3 A summary of sample size, Multi-locus (T_m) and Single-locus (T_s) outcrossing rate estimates, biparental inbreeding (S_b = T_m - T_s) estimates and selfing phenotype determinations for the primary populations used in the study.

	Out-crossing rates							Selfing phenotypes			
Population	N (families/samples)	T_m	SE	T_s	SE	($S_b = T_m - T_s$)	F_{IS}[a]	Sample size	Percentage of plants		
									SI	PSC	SC
PRES	15/90	0	-	0	-	-	****	15	0	0	100
VDR1	15/117	0.144	0.075	0.134	0.082	0.010	0.648	15	-	46	54
VDR2	4/27	-	-	-	-	-	0.401	-	-	-	-
POST	15/120	0.899	0.046	0.765	0.047	0.134	0.140	-	-	-	-
ROTO	15/94	0.978	0.066	0.812	0.073	0.166	0.068	15	80	13	7
COLL	15/94	0.779	0.08	0.652	0.074	0.127	0.003	15	73	20	7

‘-‘ = Not estimated due to insufficient flowering individuals.

[a] F_{IS} estimates are based on Ansell *et al* (2008).

[b] Outcrossing rate, biparental inbreeding and F_{IS} estimation not possible due to lack of genetic variation.

Table 2-4 The number of putative *S*-alleles identified in each of the six primary populations for which selfing phenotypes were estimated and progeny raised to estimate outcrossing rates.

Population	ALPSRK 1	ALPSRK 2	ALPSRK 3	ALPSRK 4	ALPSRK 5	ALPSRK 6	ALPSRK 7	ALPSRK 8	No. Screened
PRES	0	0	0	0	0	3	0	0	20
VDR1	0	0	0	0	0	0	0	1	10
VDR2	0	0	0	0	0	0	0	1	4
POST	2	3	1	0	0	0	0	0	20
ROTO	2	1	0	0	0	3	5	1	20
COLL	3	0	1	0	4	6	9	0	20

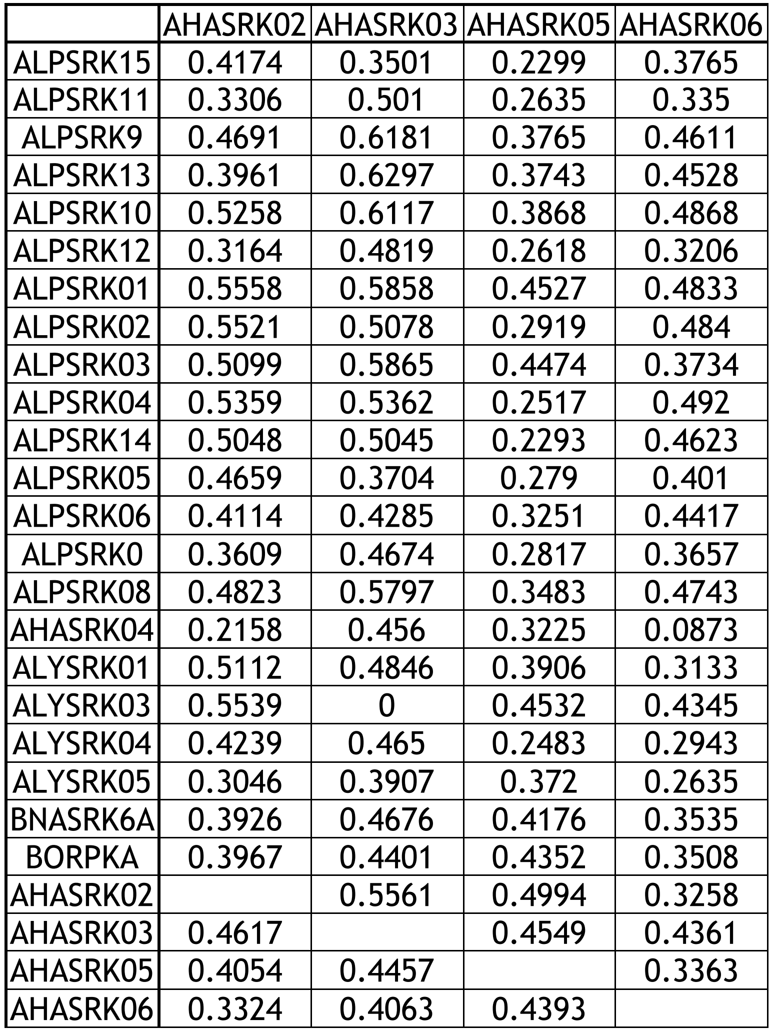
Table 2-5 Pairwise divergence values showing synonymous (Ks: above diagonal) and nonsynonymous (Ka: below diagonal) variation for the fifteen putative *SRK*-alleles described in this study. Also included are some known *SRK*-alleles from *Arabidopsis lyrata*, *Arabidopsis halleri*, *Brassica napus* and *Brassica oleracea*, for comparison of typical variation among functional *S*-alleles.

	ALPSRK15	ALPSRK11	ALPSRK9	ALPSRK13	ALPSRK10	ALPSRK12	ALPSRK01	ALPSRK02	ALPSRK03	ALPSRK04	ALPSRK14
ALPSRK15		0.2123	0.3916	0.2927	0.3603	0.2334	0.2645	0.1652	0.445	0.3145	0.2897
ALPSRK11	0.4214		0.095	0.0582	0.0852	0.008	0.3144	0.347	0.3969	0.4267	0.3992
ALPSRK9	0.369	0.1244		0.1224	0.0504	0.0853	0.4359	0.4985	0.4481	0.506	0.4752
ALPSRK13	0.3835	0.2012	0.1398		0.1215	0.0493	0.4474	0.3578	0.4122	0.5017	0.4715
ALPSRK10	0.4038	0.1523	0.0206	0.1652		0.0847	0.4319	0.462	0.4884	0.4859	0.4561
ALPSRK12	0.4495	0.0394	0.1156	0.1656	0.1401		0.3374	0.3578	0.3809	0.4237	0.3964
ALPSRK01	0.174	0.4751	0.4067	0.4119	0.4436	0.4667		0.3116	0.4449	0.53	0.4979
ALPSRK02	0.1392	0.4813	0.3813	0.4224	0.3994	0.497	0.1808		0.3969	0.404	0.3763
ALPSRK03	0.3785	0.5042	0.394	0.3705	0.4171	0.471	0.3865	0.3672		0.3412	0.3169
ALPSRK04	0.3848	0.5553	0.4302	0.4221	0.464	0.5354	0.3552	0.357	0.3767		0.0158
ALPSRK14	0.3848	0.5553	0.4302	0.4221	0.464	0.5354	0.3552	0.357	0.3767	0	
ALPSRK05	0.1434	0.4641	0.3656	0.3759	0.4005	0.4605	0.1508	0.1347	0.3105	0.3414	0.3414
ALPSRK06	0.1099	0.4234	0.3111	0.3562	0.3432	0.4243	0.1405	0.1131	0.3234	0.3293	0.3293
ALPSRK0	0.4155	0.0261	0.1434	0.1846	0.1723	0.0672	0.4673	0.4656	0.4879	0.5276	0.5276
ALPSRK08	0.3531	0.1458	0.0259	0.1492	0.0473	0.1338	0.3899	0.3692	0.4075	0.4261	0.4261
AHASRK04	0.4598	0.4383	0.3952	0.4046	0.392	0.4424	0.3977	0.3895	0.4791	0.4648	0.4648
ALYSRK01	0.3435	0.513	0.426	0.4005	0.4228	0.4846	0.3562	0.3327	0.3595	0.386	0.386
ALYSRK03	0.4431	0.388	0.3678	0.3896	0.3605	0.4196	0.437	0.4401	0.5026	0.5819	0.5819
ALYSRK04	0.4104	0.5047	0.3774	0.3408	0.4087	0.4715	0.397	0.4016	0.2232	0.2783	0.2783
ALYSRK05	0.3759	0.3892	0.3237	0.3362	0.3127	0.3771	0.3587	0.3857	0.4226	0.4204	0.4204
BNASRK6A	0.3467	0.4147	0.3624	0.344	0.397	0.4068	0.3766	0.3639	0.3587	0.3226	0.3226
BORPKA	0.357	0.5019	0.4241	0.3939	0.4625	0.4719	0.3793	0.3859	0.4011	0.3445	0.3445
AHASRK02	0.4391	0.452	0.3622	0.4014	0.3801	0.4298	0.3988	0.4092	0.4755	0.4484	0.4484
AHASRK03	0.4426	0.3876	0.3591	0.3806	0.3519	0.4103	0.4275	0.4487	0.4921	0.5701	0.5701
AHASRK05	0.3709	0.3859	0.3409	0.3741	0.3458	0.4039	0.3935	0.3667	0.3774	0.3995	0.3995
AHASRK06	0.4431	0.4314	0.3648	0.3182	0.3493	0.4188	0.4555	0.393	0.4455	0.4069	0.4069

Table 2-5 continued

	ALPSRK05	ALPSRK06	ALPSRK07	ALPSRK08	AHASRK04	ALYSRK01	ALYSRK03	ALYSRK04	ALYSRK05	BNASRK6A	BORPKA
ALPSRK15	0.1096	0.1023	0.2304	0.3902	0.3497	0.24	0.3488	0.2902	0.3953	0.477	0.4327
ALPSRK11	0.2483	0.2786	0.016	0.0856	0.2861	0.2934	0.4991	0.206	0.3307	0.3815	0.3518
ALPSRK9	0.4311	0.4911	0.1116	0.0506	0.4168	0.4015	0.6156	0.3384	0.3779	0.4433	0.4951
ALPSRK13	0.3425	0.3261	0.0861	0.1126	0.3857	0.3989	0.6271	0.2549	0.4151	0.4803	0.3923
ALPSRK10	0.3989	0.4551	0.1017	0.0331	0.4276	0.4709	0.6092	0.3481	0.3877	0.4392	0.4905
ALPSRK12	0.2812	0.301	0.024	0.0851	0.2761	0.2915	0.4801	0.1944	0.3167	0.3788	0.3537
ALPSRK01	0.1873	0.3234	0.3371	0.4643	0.502	0.4074	0.5834	0.3304	0.5773	0.6244	0.6753
ALPSRK02	0.2018	0.2421	0.366	0.4804	0.5031	0.2494	0.5058	0.368	0.5739	0.6513	0.5091
ALPSRK03	0.398	0.4684	0.363	0.4321	0.4094	0.4815	0.5841	0.271	0.4681	0.3975	0.4239
ALPSRK04	0.4002	0.3616	0.4455	0.5199	0.4533	0.3634	0.5342	0.3017	0.5047	0.44	0.471
ALPSRK14	0.3734	0.3355	0.4178	0.4886	0.4238	0.3376	0.5026	0.2786	0.4753	0.4114	0.4424
ALPSRK05		0.128	0.2776	0.4295	0.4038	0.2497	0.3691	0.3147	0.4147	0.5024	0.5179
ALPSRK06	0.1024		0.297	0.4892	0.3867	0.2486	0.4269	0.3361	0.3765	0.4232	0.3845
ALPSRK0	0.4532	0.4085		0.0841	0.3108	0.3117	0.4657	0.1854	0.3544	0.3735	0.3784
ALPSRK08	0.3577	0.2963	0.1529		0.4153	0.4286	0.5774	0.2888	0.3767	0.4367	0.4882
AHASRK04	0.4341	0.3871	0.4391	0.4042		0.3202	0.4543	0.3509	0.1911	0.3676	0.4003
ALYSRK01	0.3024	0.2729	0.4967	0.4265	0.4525		0.4828	0.3818	0.3399	0.5527	0.4761
ALYSRK03	0.4649	0.4637	0.3714	0.385	0.4278	0.4373		0.4633	0.3894	0.4658	0.4385
ALYSRK04	0.377	0.337	0.477	0.3863	0.4521	0.339	0.5313		0.3562	0.3578	0.3591
ALYSRK05	0.3867	0.3324	0.381	0.3161	0.2351	0.3856	0.3691	0.389		0.3851	0.4199
BNASRK6A	0.3392	0.298	0.3913	0.3479	0.4013	0.3436	0.4609	0.307	0.3297		0.1128
BORPKA	0.3332	0.3147	0.4725	0.4171	0.3662	0.369	0.4797	0.3137	0.3507	0.1554	
AHASRK02	0.4608	0.3628	0.4227	0.3503	0.3066	0.4949	0.4719	0.3667	0.2403	0.3357	0.302
AHASRK03	0.455	0.4538	0.371	0.3761	0.4363	0.4277	0.0051	0.5203	0.3603	0.4604	0.4694
AHASRK05	0.3472	0.3246	0.3884	0.3492	0.407	0.3529	0.4462	0.3668	0.3499	0.395	0.3808
AHASRK06	0.4086	0.3626	0.4029	0.3528	0.2158	0.369	0.4156	0.4111	0.2538	0.364	0.3735

Table 2-5 continued



	AHASRK02	AHASRK03	AHASRK05	AHASRK06
ALPSRK15	0.4174	0.3501	0.2299	0.3765
ALPSRK11	0.3306	0.501	0.2635	0.335
ALPSRK9	0.4691	0.6181	0.3765	0.4611
ALPSRK13	0.3961	0.6297	0.3743	0.4528
ALPSRK10	0.5258	0.6117	0.3868	0.4868
ALPSRK12	0.3164	0.4819	0.2618	0.3206
ALPSRK01	0.5558	0.5858	0.4527	0.4833
ALPSRK02	0.5521	0.5078	0.2919	0.484
ALPSRK03	0.5099	0.5865	0.4474	0.3734
ALPSRK04	0.5359	0.5362	0.2517	0.492
ALPSRK14	0.5048	0.5045	0.2293	0.4623
ALPSRK05	0.4659	0.3704	0.279	0.401
ALPSRK06	0.4114	0.4285	0.3251	0.4417
ALPSRK0	0.3609	0.4674	0.2817	0.3657
ALPSRK08	0.4823	0.5797	0.3483	0.4743
AHASRK04	0.2158	0.456	0.3225	0.0873
ALYSRK01	0.5112	0.4846	0.3906	0.3133
ALYSRK03	0.5539	0	0.4532	0.4345
ALYSRK04	0.4239	0.465	0.2483	0.2943
ALYSRK05	0.3046	0.3907	0.372	0.2635
BNASRK6A	0.3926	0.4676	0.4176	0.3535
BORPKA	0.3967	0.4401	0.4352	0.3508
AHASRK02		0.5561	0.4994	0.3258
AHASRK03	0.4617		0.4549	0.4361
AHASRK05	0.4054	0.4457		0.3363
AHASRK06	0.3324	0.4063	0.4393	

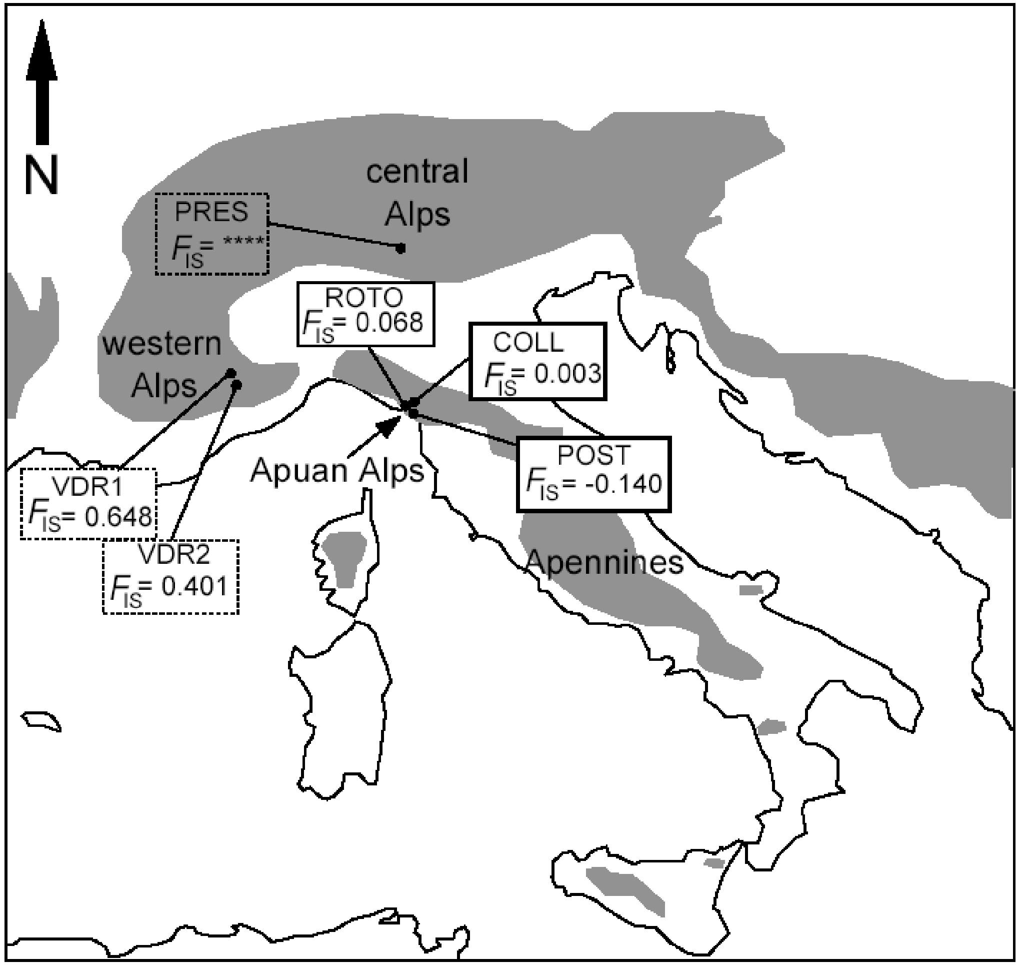
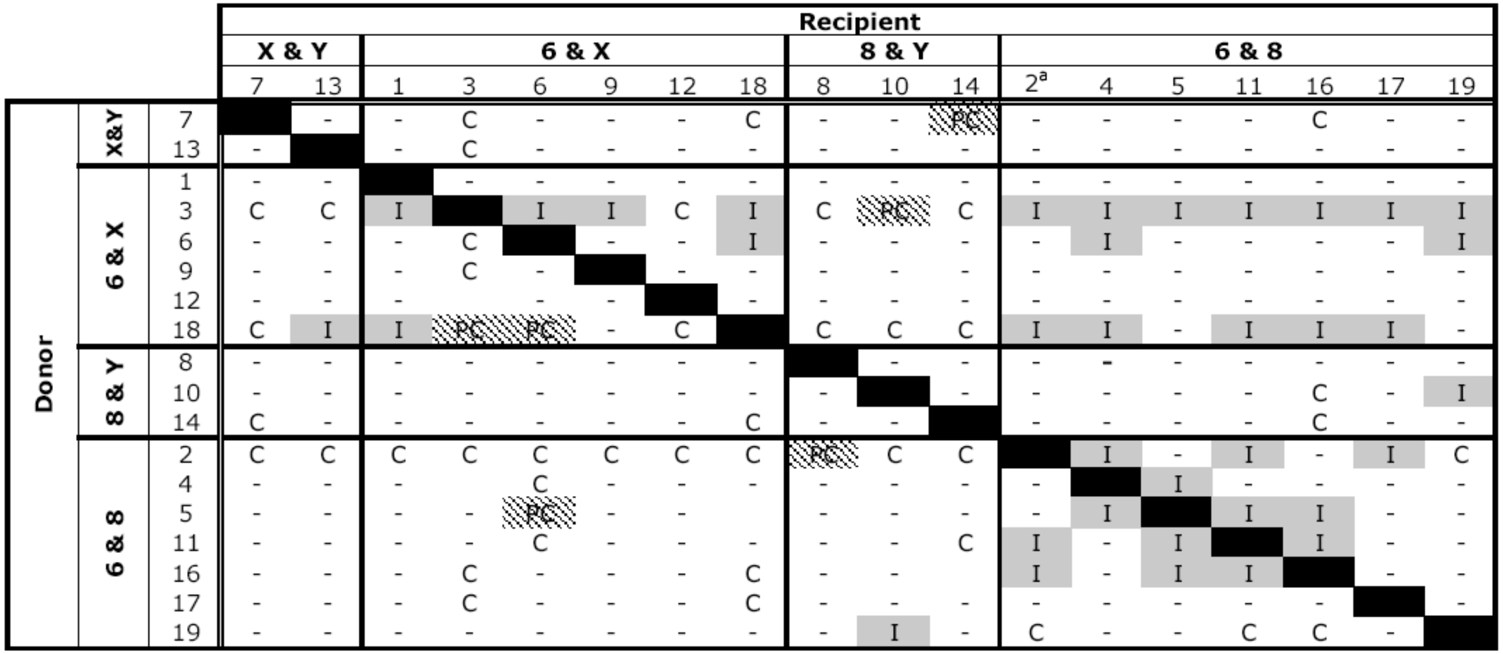


Figure 2-1 Geographic locations of the populations sampled in this study. F_{IS} estimates based on Ansell *et al* (2008) are given below each population label. PRES = Presolana; VDR = Val de Roya; ROTO = Rotondo; COLL = Colli; and POST = Porte Strazzema. Putatively highly inbred populations are in a dashed box, and putatively highly out-crossed populations are in a non-dashed box.



		Recipient																	
		X & Y		6 & X						8 & Y			6 & 8						
Donor		7	13	1	3	6	9	12	18	8	10	14	2[a]	4	5	11	16	17	19
X&Y	7	■	-	-	C	-	-	-	C	-	-	PC	-	-	-	-	C	-	-
	13	-	■	-	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6 & X	1	-	-	■	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	3	C	C	I	■	I	I	C	I	C	PC	C	I	I	I	I	I	I	I
	6	-	-	-	C	■	-	-	I	-	-	-	-	I	-	-	-	-	I
	9	-	-	-	C	-	■	-	-	-	-	-	-	-	-	-	-	-	-
	12	-	-	-		-	-	■	-	-	-	-	-	-	-	-	-	-	-
	18	C	I	I	PC	PC	-	C	■	C	C	C	I	I	-	I	I	I	-
8 & Y	8	-	-	-	-	-	-	-	-	■	-	-	-	-	-	-	-	-	-
	10	-	-	-	-	-	-	-	-	-	■	-	-	-	-	-	C	-	I
	14	C	-	-	-	-	-	-	C	-	-	■	-	-	-	-	C	-	-
6 & 8	2	C	C	C	C	C	C	C	C	PC	C	C	■	I	-	I	-	I	C
	4	-	-	-		C	-	-	-	-	-	-	-	■	I	-	-	-	-
	5	-	-	-	-	PC	-	-		-	-	-	-	I	■	I	I	-	-
	11	-	-	-	-	C	-	-	-	-	-	C	I	-	I	■	I	-	-
	16	-	-	-	C	-	-	-	C	-	-		I	-	I	I	■	-	-
	17	-	-	-	C	-	-	-	C	-	-	-	-	-	-	-	-	■	-
	19	-	-	-	-	-	-	-	-	-	I	-	C	-	-	C	C	-	■

Figure 2-2 Segregation of SI phenotypes and putative *SRK* genotypes, listing the individuals in each phenotype class and whether cross-combinations were compatible or incompatible based on three replicate pollinations. F1 individuals used in this analysis were the progeny of a maternal plant from COLL, with putative *SRK* genotype ALPSRK6/X and a paternal plant from ROTO with putative *SRK* genotype ALPSRK8/Y. Grey shading indicates an incompatible cross, with non-shaded boxes representing compatible crosses and hatched boxes partially compatible crosses. Selfings, represented by the diagonal (shaded black) were not performed in order to maximize the number of cross combinations possible with limited flowering. However, individuals 3 and 12 could be self compatible (although they could show different dominance in pollen and stigma). Individual 19 showed different patterns in when it was the pollen donor or recipient, again suggesting different dominance in pollen and stigma. Individual 2 had the genotype 6&X but its compatibility phenotype suggested that it also had the 8 allele, which failed to amplify.

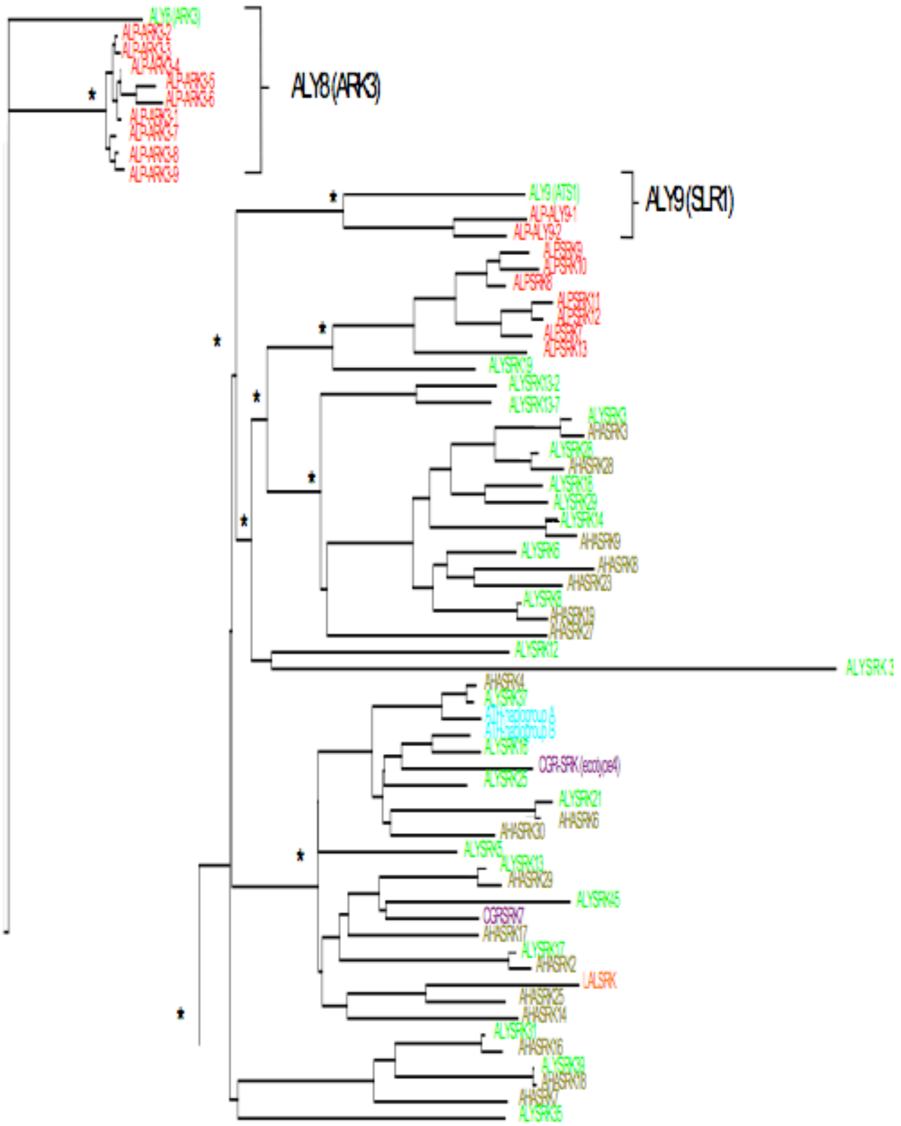
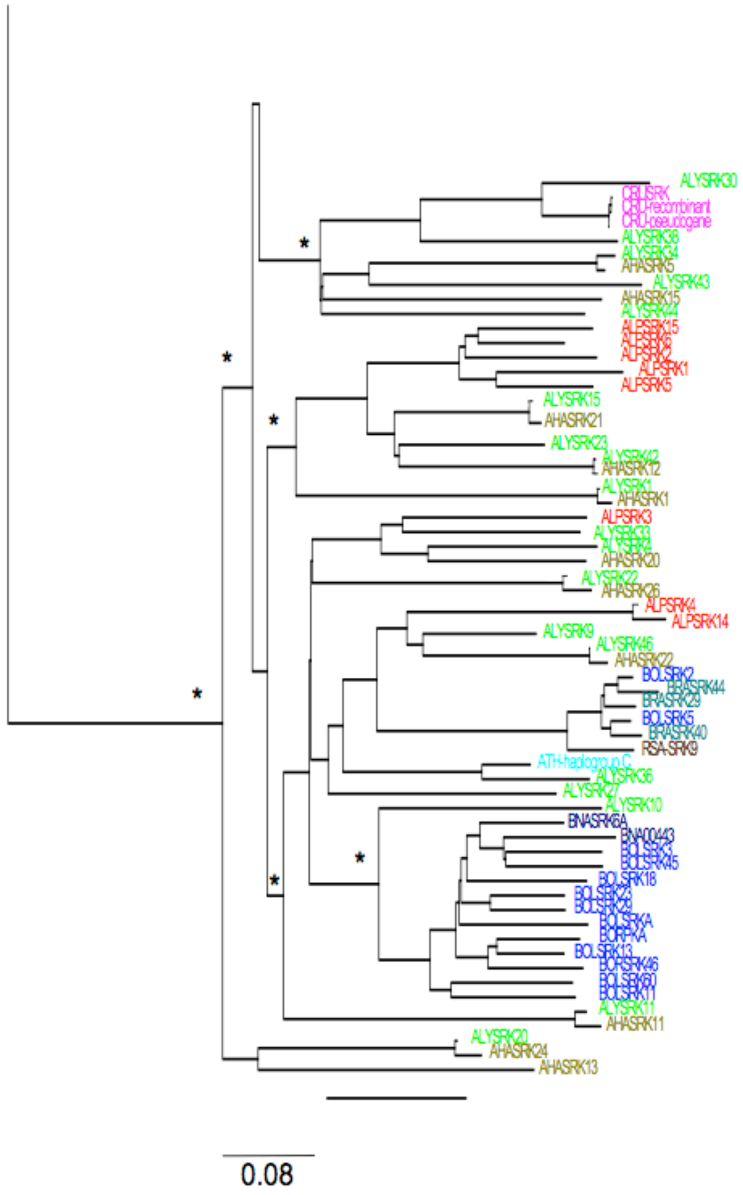


Figure 2-3 A Bayesian inference tree of Brassicaceae *S*-domain genes, constructed using the General Time Reversible model (GTR) (Tavare 1986) with gamma rate variation, using 8 heated chains, with a chain length of 2,000,000 and a burn-in of 200,000. This model was chosen using ModelTest v3.7(ref) Node support, in the form of posterior probabilities are represented by * if greater than 70%. Sequence variants for *SRK*-alleles are included for: *Arabidopsis lyrata* (Green)*,* *Arabadopsis halleri* (Gold)*,* *Arabidopsis thaliana* (Sky blue)*,* *Brassica oleracea* (Navy blue)*,* *Brassica napus* (Dark blue)*,* *Brassica rapa* (Turquoise), *Capsella rubella* (Pink)*,* *Capsella grandiflora* (Purple)*,* and *Levenworthia alabamica* (Orange). Sequence variants representing other *S*-family genes were also included, with *Aly9* (SLR1) represented by *A. lyrata*; and *Aly*8 sequence variants (which was used as outgroup, as in Prigoda *et al* 2005) from *A. lyrata*. The fifteen putative *SRK*-alleles from *Arabis alpina* (Red) are present in four clades, and cluster with both *A. lyrata* and *A. halleri* sequence variants. The putative *Aly8* and *Aly9* alleles also cluster with their respective orthologs.

Figure 2-3. *Continue*



3 Using chloroplast *trn*F pseudogenes for phylogeography in Arabidopsis lyrata.

3.1 Abstract

The chloroplast trnL-F region has been extensively utilized for evolutionary analysis in plants. In the Brassicaceae this fragment contains 1-12 tandemly repeated trnF pseudogene copies in addition to the functional trnF gene. Here I assessed the potential of these highly variable, but complexly evolving duplications, to resolve the population history of the model plant Arabidopsis lyrata. While the region 5' of the duplications had negligible sequence diversity, extensive variation in pseudogene copy number and nucleotide composition revealed otherwise cryptic population structure in eastern North America. Thus structural changes can be phylogeographically informative when pseudogene evolutionary relationships can be resolved.

3.2 Introduction

The chloroplast genome has been widely used for phylogenetic studies and its highly conserved nature allows structural changes, in the form of indels and repeat sequences, as well as base substitutions to be considered phylogenetically informative (Palmer 1991; Raubson and Jansen 2005). Low base pair substitution rates within the chloroplast genome has lead to the use of indels in population level studies, with small structural changes (<10bp) being useful for increasing phylogenetic resolution (Muse 2000), and increasing the ability to discriminate within species variation (Mitchell-Olds 2005). Larger structural changes, and complex structural rearrangements, such as inversions, translocations, loss of repeats and gene duplications are much less common, but have been reported occasionally (Hiratsuka *et al* 1989). However, there is some evidence to suggest that the value of these indel-linked, structural changes is compromised for use in evolutionary studies, due to uncertainty about the underlying mechanisms by which they arise (Ingvarsson *et al* 2003). Two main mechanisms have been suggested to account for large structural changes, such as pseudogene (non functioning duplications of functional genes) formation. 1) Intramolecular recombination between two similar regions on a single double-stranded DNA (dsDNA) molecule can result in the excision of the intermediate region, yielding a shorter dsDNA molecule and a separate circularised dsDNA molecule. It has been suggested that this will be mediated by short repeats formed by slip-strand mispairing (Ogihara *et al* 1988). 2) Intermolecular recombination between two genome copies has also been proposed because the plastid contains a large number of copies of the chloroplast genome. Recombination by this method could act to increase pseudogene variation through uneven crossing over (Pyke 1999) (Appendix 2).

Since the development of universal primers (Taberlet *et al* 1991) the non-coding trnL (UAA)-trnF(GAA) intergenic spacer region (trnL-F IGS) of the chloroplast genome has been extensively utilized for plant evolutionary analysis (Ansell *et al* 2007; Ansell *et al* 2010; Dobeš *et al* 2007; Koch *et al* 2005; Schmickl *et al* 2008) (Appendix 3). In the Brassicaceae, the trnL-F region from at least 20 genera has been documented to contain multiple copies of a duplicated trnF pseudogene (Ansell *et al* 2007; Koch *et al* 2005; Schmickl *et al* 2008). These copies are thought to be non-functional, and are made up of partial trnF gene fragments ranging between c. 50bp – c. 100bp in length. The origin of these duplications is uncertain, but Ansell *et al* (2007) identified two similar motifs that reside around the trnF gene in *Brassica* (which has no pseudogenes) and these may have functioned as sites for intermolecular recombination. There is evidence of extensive trnL-F fragment length variation in *Arabidopsis* (Koch *et al* 2005), *Boechera* (Dobeš *et al* 2007), *Cardamine* (Lihova *et al* 2004), *Rorippa* (Bleeker and Hurka 2001) and *Lepidium* (Mummenhoff *et al* 2001), and that the pseudogene duplications serve as a useful lineage marker within the Brassicaceae, as they are absent from many genera, including *Brassica*, *Draba*, and *Sinapis* (Koch *et al* 2005). Structural mutations within the Brassicacae trnL-F spacer region have caused the trnF gene copy number to be increased to 12 in certain genera (Schmickl *et al* 2008), and within the *Arabidopsis lyrata* complex (Al-Shehbaz and O'Kane 2002), the copy number varies between one and six, making this highly variable region attractive for exploring evolutionary histories.

Earlier analyses of *Arabidopsis* pseudogene variation detected complex patterns of parallel length change among trnF haplotypes (Ansell *et al* 2007; Ansell *et al* 2010; Koch *et al* 2005) and so focused on the sequence 5' to the pseudogene copy region (pre-tandem repeat region). This system was used to study the phylogeography of European populations of *Arabidopsis lyrata* (Ansell *et al* 2007; Ansell *et al* 2010), which have sometimes been considered as *A. lyrata* ssp. *petraea* (Al-Shehbaz and O'Kane 2002). However, careful analysis based on dense sampling of related taxa within both the *Arabidopsis* and *Boechera* demonstrated that it is possible to reconstruct ancestries and pseudogene copy evolution (Dobeš *et al* 2007; Koch *et al* 2005), suggesting that the pseudogenic region can increase phylogenetic resolution. This was given further weight by subsequent studies reviewing trnF evolution in cruciferous plants (Schmickl *et al* 2008) and evolution within the Brassicaceae (Koch *et al* 2007).

Within North American populations, a variation of this approach was used by Hoebe *et al* (2009), which assessed trnL-F intergenic spacer region variation among 13 populations of *A. lyrata* sampled from the Great lakes region (which are sometimes considered as *A. lyrata* ssp. *lyrata* (Al-Shehbaz and O'Kane 2002)). For this approach, the pseudogenic variation was recognized, but the relationship between copies was not addressed, and enabled three fragment length haplotypes to be distinguished among the populations sampled (Hoebe *et al* 2009). Based on these data, Hoebe *et al* (2009) concluded there had been at least two independent colonisations of glaciated areas around the Great Lakes. As part of a more extensive phylogeographic study comparing nuclear and cpDNA markers, the same TrnL-F region was sequenced from eight individuals from each of the populations used by Hoebe *et al* (2009), plus 11 additional populations sampled from more eastern and northern populations around the Great Lakes, in order to assess population structure in relation to postglacial expansion and mating system history (Foxe *et al* 2010). I found that population structure using this marker was congruent to structure elucidated from nuclear DNA markers and micro-satellites. However, while in this study considered mutations within length variants to define haplotypes, inability to infer relationships between length variants reduced the potential for resolving phylogeographic scenarios based only on cpDNA. The degree of phylogenetic resolution may be improved through more explicit incorporation of pseudogene nucleotide information, as demonstrated by wider studies of the genus (Koch *et al* 2005; Schmickl *et al* 2008).

Another limitation of the previous studies was that no samples were included from outside of the last glacial maximum. The objectives of this study were to investigate the utility of using variation in pseudogene copy number, as well as point substitutions, to infer phylogeographic patterns and relative diversity in a broader sampling of North American populations of *A. lyrata*, with European data as a reference. Specifically, I addressed the following questions: 1) do conclusions about relationships between North American and European populations change when different methods of assessing trnF variation are employed? 2) do patterns of diversity in relation to phylogeography within North American populations change when different methods of assessing trnF variation are employed?

3.3 Methods

3.3.1 Samples

Leaf material from 72 diploid plants (six per population) of *A. lyrata* ssp. *lyrata* was collected from 12 populations, from localities in eastern North America in 2007 by Yvonne Willi (Table 3-1). Field-collected tissues were desiccated prior to extraction using silica gel (Drierite, 8 MESH, Acros Organics, New Jersey USA). These samples were supplemented with sequence data from the 24 populations (eight individuals per population) reported by Foxe *et al* (2010). In addition to this, 35 herbarium samples, generously provided by the Illinois Natural History Survey (INHS), Bell herbarium (University of Minnesota) and the University of Wisconsin-Madison herbarium (Table 3-1) were also added. Herbarium tissue collection consisted of removing one leaf from each herbarium specimen that was large enough not to destroy the collection. Latitude and longitude data for each sampling location were plotted onto a map using ArcGIS v9.0 (Environmental Systems Research Institute Inc, USA).

3.3.2 Sequencing

Total genomic DNA was either extracted from 100 mg of desiccated leaf tissue using the FastDNA kit (QBiogene101, MP Biomedicals) or by DNeasy kits (Qiagen Inc) and staff of the Genome laboratory at the John Innes Center (Norwich, UK). The trnL(UAA)–trnF(GAA) IGS and trnF gene region was amplified by PCR using the 'E' and 'F' primers of Taberlet *et al* (1991). Amplifications were performed using HotStar Taq polymerase (Qiagen Inc.) under the following conditions: 94 °C for 2 min, followed by 28 cycles of 94 °C for 30 s, 50 °C for 30 s, 72 °C for 30 s, and one cycle of 72 °C for 5 min. PCR products were visualized with ethidium bromide under UV, and the bands were excised and purified with QiaQuick gel extraction kits (Qiagen Inc.). Products were sequenced on an ABI 3730 sequencer (by The Sequencing Service, University of Dundee, and The Gene Pool, University of Edinburgh). In order to compare diversity in North American populations to more extensively studied populations in Europe, I also utilised the sequence data published for 42 diploid European populations of *Arabidopsis lyrata* ssp. *petraea*; Genbank numbers DQ989814 – DQ989862 and GU456721-GU456722 (Ansell *et al* 2007) (Appendix 4).

3.3.3 TrnF Sequence alignment and pseudogene recognition

Sequences were manually aligned using Geneious 3.7 (Biomatters Ltd). Sequences containing the maximum number of tandem repeat copies detected (six for *A. lyrata* ssp. *lyrata*) were initially aligned, allowing for shorter variants to be subsequently aligned by the addition of gaps. Manual checks of the trnF pseudogene tandem repeats showed that no additional 'F' primer annealing sites were present, and all sequences contained at least 40 bases of the functional trnF gene at the 3' end of the sequence. This allowed for the identification of the complete tandem repeat region (TR), so to reliably verify the total number of pseudogene copies. Pseudogene recognition followed the protocol outlined by Ansell *et al* (2007) A slightly different approach was used by Koch *et al* (2005) and related studies.

3.3.4 TrnF sequence variation based only on the P-TR region

Pre-tandem repeat (P-TR) haplotypes (alignment positions 1-180 bp) were initially assigned according to the method described in Ansell *et al* (2007). The pre-tandem repeat region is shared throughout a number of species (Koch and Matschinger 2007), including taxa that lack the pseudogene duplications (i.e. *Brassica nigra*). New P-TR haplotypes not found in previous studies were assigned a haplotype number sequentially, following on from those previously described. A minimum spanning network (MSN), including both published European sequences and our North American sequences, was created using the minimum number of differences between P-TR haplotype pairs (Arlequin 3.11)(Excoffier *et al* 2005). A haplotype network was also constructed according to the parsimony-based algorithm developed by Templeton *et al* (1992), as implemented in the program TCS 1.13 (Clement *et al* 2000)

3.3.5 TrnL-F full sequence variation based only on mutations within length variants

The diversity of length variations in European and North American samples were compared. Our previous studies of North American populations (Foxe *et al* 2010; Hoebe *et al* 2009), did not attempted to infer evolutionary relationships between long and short fragment length haplotypes because only mutations within length variants were considered. Using this approach, full sequences (i.e. including both the P-TR and tandem repeat regions) of equal length were aligned using Sequencher 4.7 (Gene Codes, Inc.). A full sequence (FS) haplotype network was constructed according to the parsimony-based algorithm developed by Templeton *et al* (1992), as implemented in the program TCS 1.13 (Clement *et al* 2000).

3.3.6 TrnL-F full sequence variation considering evolution of copy number variants

In order to infer relationships among length variants by considering pseudogene copy number variation, as well as point substitutions and to allow comparison between European and North American populations, a new method of haplotype assignment was employed. This was achieved by: 1) assigning a number to each pseudogene variant in the tandem repeat region and scoring the presence or absence of each copy in a particular sequence; 2) dissecting each pseudogene copy present from the aligned sequence (following the alignment given in Table 3-2); 3) scoring base pair substitutions within a copy in sequential order starting at the 5’ end; 4) assigning a unique identifier to the combination of copy number variant and point substitutions; 5) combining this identifier with the haplotype assignments based on the pre-tandem repeat region to define a unique haplotype for each sequence type (Table 3-3). The probability of length changes taking place in tandem repeats increases with increasing numbers of repeats (Kelchner 2000), and consequently copy losses are more likely than gains when long TA exist, hence I consider a six copied state as more-likely ancestral.

Each full sequence haplotype was then superimposed onto the P-TR haplotype network described above. Each pseudogene copy a particular haplotype contains is then coloured based on its copy number variant (essentially, a pseudogene copy haplotype for each pseudogene copy), allowing similarities between full sequence haplotypes to be visually assessed. This allows a visual representation of haplotypes that have copy number variation, and the associated length variation this brings, which also allows inferences about parallel independent copy number evolution between continents.

3.4 Results

3.4.1 Sequence variation based on the P-TR region

Our detailed study of eastern North American A. lyrata ssp. lyrata populations identified three variable positions in the region 5' to the tandem repeat pseudogenes, allowing the discrimination of three distinct haplotypes (Table 4-3, Figure 3-2), two of which are newly reported here (P-TR11, P-TR12). P-TR 11 and 12 differ from the previously reported P-TR 1 (Ansell *et al* 2007) by two alternative single point mutations (Table 3-2). In total, 12 distinct P-TR haplotypes were recognised when our new data was combined with the existing European study of Ansell *et al* (2007; 2010) and the resulting minimum spanning network had two star-shaped clades (Figure 3-1), with all eastern North American samples restricted to clade 1. In Europe the P-TR 1 was found from a single population (Ansell *et al* 2007) at Vöslauer Hütte in Austria.

Of the three P-TR haplotypes recovered in eastern North America, P-TR 1 was the most common, detected from 96% of samples from 54/57 populations (including herbarium sample regions) (Figure 3-1; Figure 3-2a). P-TR 11 was confined to the all eight individuals sequenced in the Kitty Todd population (KTT) to the west of Lake Erie (Figure 3-2b), while P-TR 12 was present in four herbarium samples, including one individual in Houston county MN (H-HOU), one individual in Wadena county MN (H-WAA), one individual in Sheboygan county WI (H-SHE) and a final individual in Cass county MN (H-CAS). All other individuals contained P-TR haplotype 1. Thus, on the basis of this conservative analysis of trnL-F nucleotide variation, there is little genetic diversity to resolve the phylogeographic history of these populations.

3.4.2 TrnL-F full sequence variation based only on mutations within length variants

PCR amplification of the IGS region from North American populations indicated that there are three main length variants, a 'long' fragment of 741 bp and two 'short' fragments of 515 bp and 498 bp. The European samples show eight distinct length variants ranging from 339 bp to 741 bp (haplotypes which include P-TR 2 in their full sequence haplotype will be four base pairs longer due to the insert in the pre-tandem repeat region), which include two of the three main lengths present in North America (the 515 bp variant and the 741 bp variant). There is, however, variation in base pair composition relating to mutations between European and North American sequences within these length variants so none are identical between continents. One of the two short length variants (498 bp, consisting of pseudogene copies one, two and six) is not present in Europe. Alignment of the full sequences (including both the P-TR and tandem repeat pseudogene region) from the North American sequences showed that there were 13 distinct full sequence haplotypes (Figure 3-3a; Figure 3-4) compared to 46 distinct full sequence (FS) haplotypes in Europe. Only two of the North American FS haplotypes are similar to those in Europe, L4 and S2, although both of these haplotypes are present in European haplotype P-TR2 samples (PET1A and PET9 respectively), and so they vary by the four base pair insert in the pre-tandem repeat region. Each FS haplotype is delimited from its counterparts by at least one point mutation, which can occur in the P-TR or the tandem repeated duplications.

The FS haplotype network (Figure 3-3a) for North America shows the clustering of nine 'long' haplotypes in a single clade, and FS haplotype L6 is identified as ancestral by the TCS analysis, based on both number of connections and relative frequency using the principles of parsimony. Two distinct 'short' clades were also recovered, each containing two FS haplotypes (515 bp fragments: S1&S4 and 495 bp fragments: S2&S3). Haplotypes S1, L1 and L2 had previously been described by Hoebe *et al* (2009) and S2, S3, L3 and L4 by Foxe *et al* (2010).

S1 is principally restricted to the north (BEI, ISR, PIC, PUK, OWB and LSP) and northeast (TC, TSS and TSSA) of the Great lakes (Figure 3-2b), with a single population located in the extreme southeast (DOP in New York state). S3 is present in a single other population, on the opposite shore of Lake Michigan from White Fish Dunes at Sleeping Bear Dunes (SBD). Full sequence haplotype S2 is found in a single population in North Carolina (NCM), which is the only population that was sampled far to the south of the glaciated region. The final 'short' full sequence haplotype, S4, is found in two populations in New York state, in the south east of the Great lakes region, Indian Ladder (INL), where all six individuals share this full sequence haplotype and the mixed Dover Plains (DOP) population, where three of the eight individuals sequenced contain S4, and the remaining five have S1.

L1 is generally restricted to the tip of lake Michigan in the south and the northern shore of lake Erie south east of the Great lakes. L2 is widespread and occurs on Lake Michigan on the southeastern shore (Saugatuck; SAU) the south west shore (Illinois Beach; ILB). On Lake superiors southern shore at Pictured Rock (PIR) and Apostle Island (API), and on Lake Hurons northern shore at Manatoulin (MAN) and southern shore at Port Cresent (PCR). Further L2 haplotypes were found in herbarium specimens from Wisconsin (H-OCO & H-TRE) and from Minnesota (H-MOR, H-COK & H-WAH) (Figure 3-2b). Indiana dunes (IND), on the southern shore of Lake Michigan is a mixed population containing both L1 and L2. There are two further mixed populations containing L1, both on Georgian Bay in southern Ontario: Tobermory Singing Sands (TSS), where three of the eight samples sequenced were L1 and the remaining five were S1; and Wasaga beach (WAS), where two of the eight individuals sequenced were L1 and the remaining six L4. L3 is present in a single population, Kitty Todd nature reserve (KTT), which is located on the remnants of a sand flat towards the western edge of lake Erie in Toledo, Ohio. L4 has a relatively large range, spreading from the extreme south east of the Great lakes in New York state and Pennsylvania (IOM, FOM and FDV) in a north westerly direction encompassing two populations on the south shore of Lake Erie (HDC and PRI) and a population on the eastern shore of Lake Michigan (LUD). L4 is also present in White Fish Dunes (WFD), which is located on the western shore of lake Michigan and also contains a short haplotype, S3 (with four individuals of each haplotype present). The remaining full sequence haplotypes (L5 – L9) are all found in herbarium samples to the west of the Great Lakes in Minnesota and Wisconsin. The most common of these is the putative ancestral sequence for the long haplotypes, L6. It is found throughout the west of the Great Lakes region in both Minnesota and Wisconsin. Full sequence haplotypes L8 and L9 are each only found in a single individual: L8 in Houston county, Minnesota (H-HOU) and L9 in Filmore County, Minnesota (H-FIL).

3.4.3 TrnL-F full sequence variation considering evolution of copy number variants

The 13 distinct FS haplotypes for North America, and 46 FS haplotypes from Europe, were mapped onto the haplotype network recovered from the conservative analysis of the P-TR nucleotide variation. This revealed parallel independent changes in pseudogene copy number in both Europe and North America, with tandem repeats that contain various combinations of particular pseudogene copies (one, three, four, or six) are present in both clades (e.g PET 12 and PET 25), and also between continents (L4 and PET 5). Only a small proportion of IGS variants have undergone changes in the P-TR region, but this can also be seen on both continents (e.g. Europe - PET 31A compared to PET 31B/31C, North America – L2 compared to L3).

By comparing pseudogene copy variation, and the mutations contained within, evolutionary links between ‘short’ and ‘long’ clade haplotypes could be established. The resultant full sequence pseudogene structure (FSPS) haplotype network contains a single clade (Figure 3-3b). Within the “long” clade, each of the full sequence haplotypes contains the same six pseudogene copies (one, two, six, seven, nine and ten), with each haplotype being separated from one another by at least a single point mutation. These mutations equate to variation within a particular pseudogene copy or, in the case of L3, L5 and L9, in the P-TR region. Each of the four 'short' full sequence haplotypes contains three pseudogene copies. S1 and S4 contain copies 1, 2 and 6 whereas S2 and S3 contain copies 1, 9 and 10. S2 and S3 are separated by a single point mutation and S1 and S4 separated by two point mutations. Based on this analysis, haplotype L6 remains the most parsimonious potential ancestral haplotype based on the number of single base connections, with the structuring among 'long' FS haplotypes remaining unchanged. Both 'short' clades from the FS haplotype network are now incorporated into the 'long' clade, but the relationship between 'short' haplotypes has changed quite significantly. S2 and S3, previously linked by a single base pair substitution, are now separate. S2 can be linked to L1 based on shared pseudogene mutations with a putative pseudogene loss event, and S3 can be linked to L4 based on the same principles, with the loss of pseudogene copies six and seven. The second short clade from the FS haplotype network remains together, with haplotype S1 now linked to 'long' haplotype L2, and haplotype S4 remains linked to S1 (separated by two base pair changes).

L1 populations, being located principally on the northern shore of Lake Erie and the southern tip of Lake Huron are in close proximity to the related 'long' haplotype L3, and although the suggested closely related 'short' haplotype S2 is some distance away, it falls well below the estimated extent of the last glacial maxima (Holman 1992)(Figure 3-2b), suggesting it is a much older population. FSPS haplotype L2's broad distribution means it is geographically located in close proximity to its suggested 'short' counterpart, S1, on the shores of both Lake Superior (PIR, BGB, API and H-COK) and Lake Huron (MAN). L4's wide, narrow distribution in a north-westerly direction (from New York/ Pennsylvania in the east to the north of Lake Michigan) culminates in its close proximity to, and mixture with S3, its suggested 'short' relation in WFD (of the eight individuals screened, four are L4 and four are S3) and SBD (non mixed S3 population). Haplotype S4 is found in two populations only, both in New York State. Indian Ladder (INL) contains all S4 individuals (six) and is in close proximity to Dover Plains (DOP), which contains two S4 individuals and four S1 individuals, supporting the evolutionary link between these two haplotypes (Figure 3-2b; Figure 3-3a).

3.5 Discussion

3.5.1 Summary of chloroplast variation within Arabidopsis lyrata

It is clear from this study that there is a relatively large amount of variation within the trnL-trnF IGS region, across *A. lyrata* as a whole, and this pattern is also seen between related species and genera (Dobeš *et al* 2007; Schmickl *et al* 2008). Much more variation was found in European samples relative to eastern North America, but this is consistent with earlier nucleotide studies on A. lyrata (Balañá-Alcaide *et al* 2006; Clauss and Mitchell-Olds 2006; Ross-Ibarra *et al* 2008), and is concordant with the suggestion that the origins of North American populations are firmly centred in Europe (Clauss and Mitchell-Olds 2006; Ross-Ibarra *et al* 2008; Wright *et al* 2003), with lower levels of genetic diversity found in North American (with respect to European *A. lyrata*) being suggested to demonstrate a possible long-term population bottleneck associated with the colonization of North America from European populations (Clauss and Mitchell-Olds 2006).

While variation appears to be maintained in both the P-TR (with three distinct variants in NA and 10 variants in Europe) and the pseudogene tandem array portions of the trnL-F fragment (with 13 distinct haplotypes in NA and 46 distinct variants in Europe) (Figure 3-4), variation within pseudogene copies globally was higher, with up to 10 distinct variants for each copy. How this variation is considered has important implications for use of the trnF region for phylogeographic studies.

3.5.2 TrnF sequence variation based only on the P-TR region

If we assign haplotypes based only on variation in the P-TR region, as has been suggested previously (Ansell *et al* 2007), then populations in Europe show a high degree of variation, and North American populations are almost completely devoid of variation. North American populations appear to be more closely related to European haplotypes in Clade 1 than Clade 2 (Figure 3-4), which is consistent with Europe being the centre of genus diversity and source for the North American populations, as has been suggested previously (Clauss and Mitchell-Olds 2006; Koch *et al* 2005; Ross-Ibarra *et al* 2008). However, in North America, despite sampling similar numbers of “populations” as in Europe, we sampled only a very small part of the range, whereas in Europe the sampling has been more comprehensive; this could partly explain the increased diversity in the latter. Nevertheless, the comparatively low amounts of variation in North America compared to Europe could be related to their different histories of post-glacial expansion. Glacial events in both regions are relatively well understood. However, post-glacial expansion in plants has not been as well studied in North America as it has in Europe. It is possible, therefore, that, differences in glacial maxima, glacial retreat times or the suspected bottleneck event associated with colonization of North America (Clauss *et al* 2002; Foxe *et al* 2010), gives rise to the disparity between the two continents. Although there is also a shift towards inbreeding in some of the North American populations sampled, our previous work has demonstrated that there is not a resulting difference in chloroplast diversity (Hoebe *et al* 2009).

Based only on the P-TR haplotype allocations, the trnL-trnF IGS region would not be a good candidate for studying phylogeography in North America. It shows very little variation, with only three pre-tandem repeat haplotypes present, only one of which occurs at high frequency. Each is comprised of a single base pair mutation from the common P-TR haplotype 1. However, from the analysis in this study, I consider that using only the P-TR region to assess phlyogeograpy does not represent all the available variation found in the Great Lakes region. This is given extra weight by the findings of Foxe *et al* (2010), which found clear phylogeographic structuring based on microsatelite and nuclear data in this species; although cpDNA data did not conflict with conclusions based on the nuclear loci, I was unable to draw strong conclusions about phylogeography based on the distribution of length variants in the cpDNA data alone.

3.5.3 TrnF full sequence variation based only on mutations within length variants

Considering variation across the tandem repeat region, as well as the P-TR region allows the assignment of more phylogeographically informative haplotypes. Both North America and Europe exhibit variation within full sequence haplotypes. European haplotypes form eight different length variations due to varying pseudogene copy numbers present in the tandem repeat region. Tandem repeat composition varied from a single pseudogene copy, to six pseudogene copies, with variation in pseudogene position being common also (e.g. PET 14 contains four pseudogene copies in positions one, two, six and seven; PET 10 also contains four pseudogene copies but in positions one, six, nine and ten). This variation in length variants in Europe represents greater diversity than is present in North America, where only three length variants are found, two of which are present in the European samples and a third length variant which is not found in Europe. When I consider all unique haplotypes, which fall into these length variants, then European samples form 46 full sequence haplotypes. Again this level of variation represents much greater diversity than present in North America, whereas only 13 full sequence haplotypes have been described. When considering FS haplotypes like this, no North American haplotypes are identical to those in Europe, and so suggesting potential origins of North American populations becomes very difficult with this method.

When I consider the full sequence haplotypes for North America but without considering shared copy variants the FS haplotype network separates sequences into three distinct clades: A “long” clade containing L1 – L9, and two distinct “short” clades containing S1 & S4 and S2 & S3, respectively. Both ‘short’ clades contain two full sequence haplotypes, with S2 and S3 separated by a single point mutation and S1 and S4 separated by two point mutations. Neither haplotype can be considered ancestral based on the number of connections, unless haplotype frequency data are added (TCS assumes a haplotype to be ancestral by calculating both the number of individuals which contain it and the number of connections it has, based on the principles of parsimony) but our focus on sampling single herbarium specimens does not allow this. Whilst this network appears justified at the sequence level based on shared pseudogene copies, there are aspects of it that do not make strict geographical sense. One example of this is that full sequence haplotypes S2 and S3, which are considered closely related in this network, are actually spatially separated (Figure 3-2b), with no evidence to support a shared colonization route history.

3.5.4 Trn-F full sequence variation considering evolution of copy number variants

When considering the FSPS haplotype assignment method, I find that globally, variation within a pseudogene copy is large, with up to 10 variants being found at any given position. Unique variants are found in both Europe and North America; however, as for all of the other analyses, diversity (in this case the number of unique within copy variants present on either continent) is much higher in Europe. The complexity of variation in copy number within Europe still makes it difficult to infer which full sequence variants are more closely related to North American haplotypes.

However, within North America, when shared pseudogene copy variation between full sequence haplotypes is considered to be evidence of shared ancestry, full sequence haplotypes previously included in either of the ‘short’ clades can be included in a single network with the ‘long’ sequences (FSPS haplotype network; fig 3-4). The loss of pseudogenes from the tandem array is considered a single evolutionary step, and is indicated with a dashed line (Figure 3-3b). Using this approach, the S2-S3 ‘short’ clade is separated, with S2 being linked to L1 and S3 being linked to L4. S1 and S4 are still linked together, and now join the single clade network at L1. This single clade full sequence haplotype network represents a better geographical fit, with closely related sequences separated by a pseudogene loss event being spatially closer than in the FS haplotype network (Figure 3-2b).

Although limited variation within populations restricts the utility of formalized phylogeographic tests for the North American samples, I can draw some conclusions about historical patterns of colonization of the Great Lakes region following the last glacial maximum. Following the retreat of the Wisconsin glaciation 10,000 bp, various scenarios of post-glacial colonization have been suggested. For reptiles and amphibians in particular, it has been suggested that at least two routes of colonization is most parsimonious with molecular and geographic data (Holman 1995; Placyk *et al* 2007; Soltis *et al* 2006), with likely routes being to the east, and the west of Lake Michigan. Hoebe *et al* (2009) agreed that two colonization routes also were likely for A. lyrata, suggesting that individuals with the S1 full sequence haplotype moved northwards along the west side of Lake Michigan, colonizing the area to the north of Lake Superior and those with full sequence haplotype L1 was suggested to have moved northwards on the west side of Lake Michigan. In this study, I have found that the divide between ‘short’ and ‘long’ haplotypes presented by Hoebe *et al* (2009) is much less clear when further haplotypes are added. Foxe *et al* (2010) found that FS haplotype distributions for 24 of the 33 populations used here (Table 3-1) were consistent with Bayesian clustering patterns based on combined nuclear DNA markers and micro-satellites, which supported the findings of Hoebe *et al* (2009) suggesting at least two routes of colonization. In that study, we only considered mutations within length variants, and by using this method we were unable to link ‘short’ and ‘long’ clade haplotypes together. However, if we consider that full sequence haplotype S1 is linked to full sequence haplotype L2 by a single pseudogene loss event, this suggests a northern post-glacial colonisation on either the east or west side of Lake Michigan, with S1 principally clustering at the north of Lake Superior, and L2 showing a more scattered distribution. (Figure 3-2b). It is possible that there may be a third colonization route not seen in previous studies, moving westwards from the New York state area towards Wisconsin. This is supported by current distributions for full sequence haplotype L4, and its probable derivatives S3 and L7. This colonization route may be supported by the fact that the L4 population at Friedensville (FDV) is below the estimated position of the Wisconsin glacial maxima (Holman 1992).

It is possible that the Friedsville population (FDV) in Pennsylvania (Figure 1b) and a number of populations in Minnesota and Wisconsin (H-EAU, H-WAB, H-WAU, H-MAR, H-SAU, H-RIC, H-TRE, H-WIN, H-GOH, H-HOU, H-FIL and H-HEN) are much older than the other populations, as they are currently present in what should have been non-glaciated areas. This might support the view that full sequence haplotype L6 is ancestral.

3.5.5 Pseudogenes in a population level study

Earlier phylogeographic studies on *Arabidopsis* based on the trnL-F IGS employed conservative approaches to analysing sequence variation, due to structurally mediated parallel changes in pseudogene copies, and elected to base their interpretations on the pre-tandem array portion of the fragment (Ansell *et al* 2010; Koch *et al* 2007; Schmickl *et al* 2010). This approach was appropriate for Europe, where there is more extensive variation in the pre-tandem array region, and correspondingly, strong evidence for multiple changes in pseudogene copy number (Ansell *et al* 2007; Schmickl *et al* 2008), In the relatively genetically depauperate eastern North American populations of *A. lyrata* there is limited nucleotide diversity in the pre-tandem array region, allowing for greater certainty when assessing relationships between sequences that differ by pseudogene content. As a consequence, by following a simple set of assumptions, I show that the nucleotide informative available within the tandem array can be used to make inferences about evolutionary links between populations that have varying sequence length, and to recover otherwise cryptic population structure.

Table 3-1 Sampled populations, geographic co-ordinates, date collected, collector and *trn*F haplotypes. Haplotype assignments are based on the P-TR assignment method and the full sequence haplotype method.

Population name	Abbrev.	Lat.	Long.	Holmgren list acronym	Number of Individuals	Date collected	Collector	Herbarium (inc. Date)	Full sequence haplotype 1 (Number)	Full sequence haplotype 2 (Number)	P-TR haplotype (Number)
Point Pelee	PTP	41.93	-82.51	-	8	2003	BM	-	L1	-	1
Iona Marsh	IOM	41.30	-73.98	-	8	2007	YW	-	L4	-	1
Pictured Rock	PIR	46.67	-86.02	-	8	2003	BM	-	L1	-	1
Presque Isle	PRI	42.17	-80.07	-	8	2007	BM. AT. PH.	-	L4	-	1
Long Point	LPT	42.58	-80.39	-	8	2007	BM. AT. PH.	-	L1	-	1
Indiana Dunes	IND	41.62	-87.21	-	8	2007	BM. AT. PH.	-	L1 (4)	L2 (4)	1
Rondeau	RON	42.26	-81.85	-	8	2007	BM. AT. PH.	-	L1	-	1
Sleeping Bear	SBD	44.94	-85.87	-	8	2007	BM. AT. PH.	-	S3	-	1
North Carolina, Mayodan	NCM	36.41	-79.97	-	8	2007	DM	-	S2	-	1
Tobermoray Alvar	TSSA	45.19	-81.59	-	8	2007	BM. AT. PH.	-	S1 (4)	L1 (4)	1
Lake Superior Park	LSP	47.57	-84.97	-	8	2003	BM	-	S1	-	1
Tobermory Cliff	TC	45.25	-81.52	-	8	2007	BM. AT. PH.	-	S1	-	1
Old Woma Bay	OWB	47.79	-84.90	-	8	2003	BM	-	S1	-	1
Pic River	PIC	48.60	-86.30	-	8	2003	BM	-	S1	-	1

Table 1-3 continued

Pukaskwa National Park	PUK	48.40	-86.19	-	8	2003	BM	-	S1	-	1
Manitoulin Island	MAN	45.67	-82.26	-	8	2003	BM	-	L2	-	1
Tobermoray SS	TSS	45.19	-81.58	-	8	2007	BM. AT. PH.	-	S1	-	1
Pinery	PIN	43.27	-81.83	-	8	2007	BM. AT. PH.	-	L1	-	1
Beaver Island	BEI	45.76	-85.51	-	8	2007	YW	-	S1	-	1
Headland Dunes	HDC	41.76	-81.29	-	8	2007	BM. AT. PH.	-	L4	-	1
Kitty Todd	KTT	41.62	-83.79	-	8	2007	BM. AT. PH.	-	L3	-	11
Port Cresent	PCR	44.00	-83.07	-	8	2007	BM. AT. PH.	-	L2	-	1
Wasaga Beach	WAS	44.52	-80.01	-	8	2003	BM	-	L4 (6)	L1 (2)	1
White fish Dunes	WFD	44.92	-87.19	-	8	2007	BM. AT. PH.	-	L4 (4)	S3 (4)	1
Indian Ladder	INL	42.66	-74.02	-	6	2007	YW	-	S4	-	1
Dover Plains	DOP	41.74	-73.58	-	6	2007	YW	-	S1 (4)	S4 (2)	1
Fort Montgomery	FOM	41.33	-73.99	-	6	2007	YW	-	L4	-	1
Illinois Beach	ILB	42.42	-87.81	-	6	2007	YW	-	L2	-	1
Apostle Island	API	46.94	-90.74	-	6	2007	YW	-	L2	-	1
Bete Grise Bay	BGB	47.39	-87.96	-	6	2007	YW	-	L1 (4)	L4 (2)	1
Isle Royal	ISR	48.00	-88.83	-	6	2007	YW	-	S1	-	1
Ludington	LUD	43.96	-86.45	-	6	2007	YW	-	L4	-	1
Saugatuck	SAU	42.68	-86.18	-	6	2007	YW	-	L2	-	1

Table 1-3 continued

Friedensville	FDV	40.55	-75.41	-	6	2007	YW	-	L4	-	1
Fillmore Co. MN	H-FIL	43.67	-92.10	MIN	3	-	-	1977/1941	L6 (2)	L9 (1)	1
Houston Co. MN	H-HOU	43.67	-92.24	MIN	2	-	-	1942/1962	L6 (1)	L8 (1)	1(1) 12(1)
Wabasha Co. MN	H-WAB	44.28	-91.77	MIN	1	-	-	1997	L6	-	1
Winona Co. MN	H-WIN	43.97	-91.77	MIN	1	-	-	1992	L7	-	1
Trempealeau Co. WI	H-TRE	44.32	-91.35	WIS	2	-	-	NR	L2 (1)	L7 (1)	1
Eau Claire, WI	H-EAU	44.82	-91.50	WIS	2	-	-	NR	L6	-	1
Marquette Co. WI	H-MAR	43.82	-89.40	WIS	1	-	-	NR	L6	-	1
Richland Co. WI	H-RIC	43.38	-90.43	WIS	1	-	-	NR	L6	-	1
Sauk Co. WI	H-SAU	43.45	-89.95	WIS	1	-	-	NR	L6	-	1
Waushara Co. WI	H-WAU	44.12	-89.29	WIS	1	-	-	NR	L6	-	1
Cass Co. MN	H-CAS	46.92	-94.28	MIN	6	-	-	1992/1997	L6 (5)	L5 (1)	1(5) 12(1)
Goodhue Co. MN	H-GOH	44.42	-92.72	MIN	2	-	-	1987/1940	L6	-	1
Crow wing Co. MN	H-CRO	46.47	-94.08	MIN	1	-	-	1936	L6	-	1
Anoka Co. MN	H-ANK	45.25	-93.25	MIN	1	-	-	1960	L6	-	1
Hennepin Co. MN	H-HEN	43.08	-92.24	MIN	1	-	-	1922	L6	-	1
Sheboygan Co. WI	H-SHE	43.73	-87.93	WIS	1	-	-	NR	L5	-	12

Table 1-3 continued

Wadena Co. MN	H-WAA	46.58	-94.97	MIN	1	-	-	1992	L5	-	12
Oconto Co. Wi	H-OCO	45.00	-88.18	WIS	3	-	-	NR	L2	-	1
Washington Co. MN	H-WAH	45.03	-92.92	MIN	1	-	-	1961	L2	-	1
Morrison CO. MN	H-MOR	46.02	-94.30	MIN	1	-	-	1990	L2	-	1
Cook Co. MN	H-COK	47.92	-90.55	MIN	1	-	-	1980	L2	-	1
Milwaukee Co. WI	H-MIL	43.00	-87.97	WIS	1	-	-	NR	L1	-	1

Table 3-2 Criteria for *trn*F pseudogene alignment used in this study. P-TR refers to all sequence before the tandem repeat region (pseudogene copies). Origin species is the species in which certain pseudogene copies are found.

Psuedogene copy number	Present in *A.lyrata*	Origin species	Size (bp)	Start position (bp)
P-TR	Y	Unknown	179	1
1	Y	*A. lyrata*	93	180
2	Y	*A. lyrata*	77	273
3	N	*A.thaliana*	32	350
4	N	*A.halleri*	100	382
5	N	*A.halleri*	98	480
6	Y	*A. lyrata*	99	579
7	Y	*A. lyrata*	67	646
8	N	*A.halleri*	67	713
9	Y	*A. lyrata*	68	781
10	Y	*A. lyrata*	125	906
*trn*F gene	Y	Unknown		1032

Table 3-3 ***A. thaliana*, European *A. lyrata* (*A. lyrata* ssp. *petrea* , PET), North American *A. lyrata* (*A. lyrata* ssp. *Lyrata*; L1-L9 & S1-S4) *trn*L-*trn*F IGS sequences and their *trn*F pseudogene composition. Numbers refer to the pseudogene variant present at each copy, for each sequence.**

Species & Haplotype	P-TR Haplotype	1	2	3	4	5	6	7	8	9	10
		Arabidopsis lyrata ssp. petraea									
PET 1A	2	1.1	2.1	-	-	-	6.1	7.1	-	9.1	10.1
PET IB	2	1.1	2.1	-	-	-	6.1	7.1	-	9.1	10.1
PET 1C	2	1.6	2.1	-	-	-	6.1	7.1	-	9.1	10.1
PET 2	2	1.1	2.1	-	-	-	6.1	7.1	-	9.1	10.2
PET 3	2	1.1	2.1	-	-	-	6.1	7.1	-	9.1	10.3
PET 4	2	1.1	2.1	-	-	-	6.1	7.1	-	9.1	10.1
PET 5	2	1.6	2.1	-	-	-	6.3	7.1	-	9.1	10.1
PET 6	2	1.1	2.1	-	-	-	6.1	7.3	-	-	-
PET 7	2	1.2	-	-	-	-	-	-	-	-	-
PET 8	2	1.3	-	-	-	-	6.1	7.1	-	9.1	10.1
PET 9	2	1.4	-	-	-	-	-	-	-	9.1	10.1
PET 10	2	1.1	-	-	-	-	6.1	-	-	9.2	10.1
PET 11	2	1.6	2.1	-	-	-	-	-	-	-	10.7
PET 12	2	1.1	2.1	-	-	-	6.2	-	-	9.1	10.3
PET 13	2	1.1	2.1	-	-	-	-	-	-	-	10.5
PET 14	2	1.6	2.1	-	-	-	6.1	7.4	-	-	-
PET 15	2	1.3	-	-	-	-	-	-	-	-	10.7
PET 16	2	1.1	2.1	-	-	-	-	-	-	-	10.7
PET 17	1	1.1	-	-	-	-	-	-	-	-	-
PET 18	3	1.1	2.3	-	-	-	6.3	-	-	9.3	10.4
PET 19A	3	1.1	2.3	-	-	-	-	-	-	-	10.6
PET 19B	6	1.1	2.3	-	-	-	-	-	-	-	10.8
PET 20	5	1.5	2.1	-	-	-	6.1	7.2	-	9.1	10.4
PET 21A	6	1.6	2.1	-	-	-	6.4	7.5	-	9.1	10.4
PET 21B	6	1.6	2.1	-	-	-	6.4	7.7	-	9.1	10.4
PET 21C	6	1.6	2.1	-	-	-	6.5	7.7	-	9.1	10.4
PET 21D	7	1.6	2.1	-	-	-	6.4	7.7	-	9.1	10.4
PET 21E	10	1.6	2.1	-	-	-	6.5	7.7	-	9.1	10.4
PET 22	6	1.1	2.1	-	-	-	6.4	7.7	-	9.1	10.4
PET 23	6	1.6	2.4	-	-	-	6.4	7.6	-	9.1	10.4
PET 24	6	1.1	-	-	-	-	6.4	7.7	-	9.1	10.4
PET 25	6	1.1	-	-	-	-	6.4	7.7	-	9.1	10.4
PET 26	6	1.1	-	-	-	-	6.4	7.6	-	9.1	10.4
PET 27	6	1.7	-	-	-	-	-	-	-	9.1	10.4
PET 28	6	1.6	2.1	-	-	-	6.4	7.8	-	-	-
PET 29	6	1.6	2.4	-	-	-	6.4	7.8	-	-	-
PET 30	6	1.1	2.1	-	-	-	6.4	7.8	-	-	-
PET 31A	6	1.6	2.1	-	-	-	-	-	-	-	10.8
PET 31B	8	1.6	2.1	-	-	-	-	-	-	-	10.8
PET 31C	4	1.8	2.1	-	-	-	-	-	-	-	10.7
PET 31D	9	1.6	2.1	-	-	-	-	-	-	-	10.8
PET 31E	6	1.6	2.1	-	-	-	-	-	-	-	10.9
PET 31F	6	1.6	2.1	-	-	-	-	-	-	-	10.8
PET 32	6	1.6	2.4	-	-	-	-	-	-	-	10.9
PET 33	6	1.9	-	-	-	-	-	-	-	-	-

Table 3-3 continued

PET 34	6	1.9	-	-	-	-	-	-	-	-	-
		Arabidopsis lyrata ssp. lyrata									
L1	1	1.1	2.1	-	-	-	6.6	7.1	-	9.1	10.10
L2	1	1.2	2.1	-	-	-	6.1	7.1	-	9.1	10.10
L3	11	1.1	2.1	-	-	-	6.6	7.1	-	9.1	10.10
L4	1	1.1	2.1	-	-	-	6.1	7.1	-	9.1	10.1
L5	12	1.6	2.1				6.1	7.1		9.1	10.10
L6	1	1.1	2.1				6.1	7.1		9.1	10.10
L7	1	1.11	2.1				6.1	7.1		9.1	10.1
L8	12	1.6	2.6				6.1	7.1		9.1	10.10
L9	1	1.1	2.1				6.1	7.1		9.4	10.10
S1	1	1.1	2.1	-	-	-	6.7	-	-	-	-
S2	1	1.4	-	-	-	-	-	-	-	9.1	10.1
S3	1	1.4	-	-	-	-	-	-	-	9.1	10.10
S4	1	1.1	2.1	-	-	-	6.1	-	-	-	-
		Arabidopsis thaliana									
A.thaliana	0	1.AT	2.AT	3.AT	-	-	6.AT	-	-	9.AT	10.AT

Table 3-4 Pre-tandem repeat (P-TR) haplotype alignment criteria. * Indicates new P-TR haplotypes described in this study.

Position of *trn*L - *trn*F alignment												
			Indel									
Haplotype	46	47	A	88	96	101	111	119	136	154	156	167
0	A	T	-	C	A	T	G	C	A	T	C	C
1	A	A	-	C	A	T	G	C	A	T	C	C
2	A	A	+	C	A	T	G	C	A	T	C	C
3	A	A	-	C	T	T	G	C	A	T	C	C
4	A	A	-	C	A	T	G	C	A	T	A	C
5	A	A	-	C	A	T	G	C	A	G	C	C
6	A	A	-	T	A	T	G	C	A	T	C	C
7	A	A	-	T	A	T	G	C	C	T	C	C
8	A	A	-	T	A	T	G	T	A	T	C	C
9	G	A	-	T	A	T	G	C	A	T	C	C
10	A	A	-	T	A	T	G	C	A	T	C	C
11*	A	A	-	T	A	T	T	C	A	T	C	T
12*	A	A	-	T	A	A	G	C	A	T	C	C

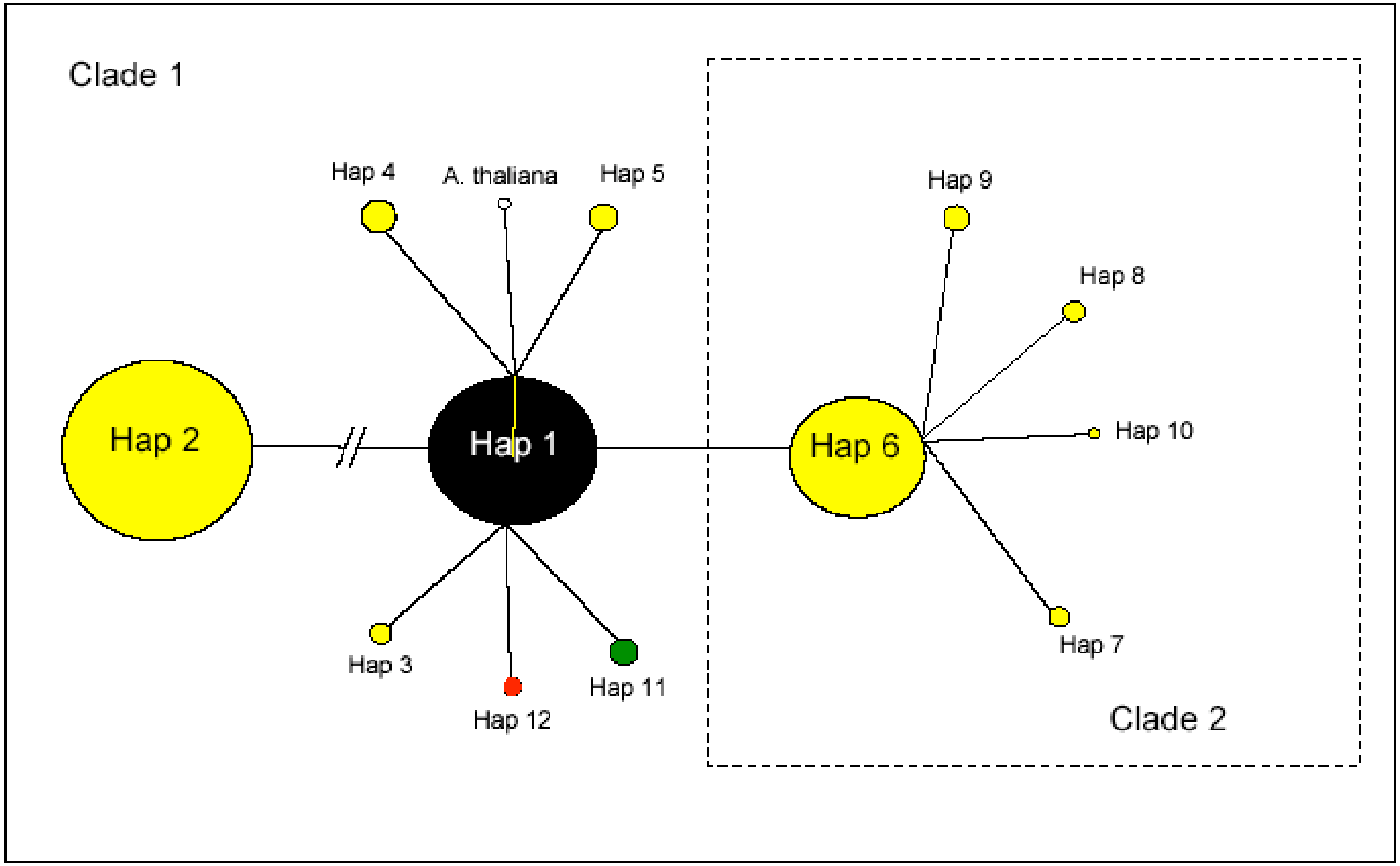


Figure 3-1 ***Arabidopsis lyrata*** **pre-tandem repeat (P-TR) haplotype network including all North American (coloured black, red and green) and European variants (coloured yellow).** ***A. thaliana*** **P-TR haplotype (open circle) included for reference. Circle size is proportional to haplotype frequency.**

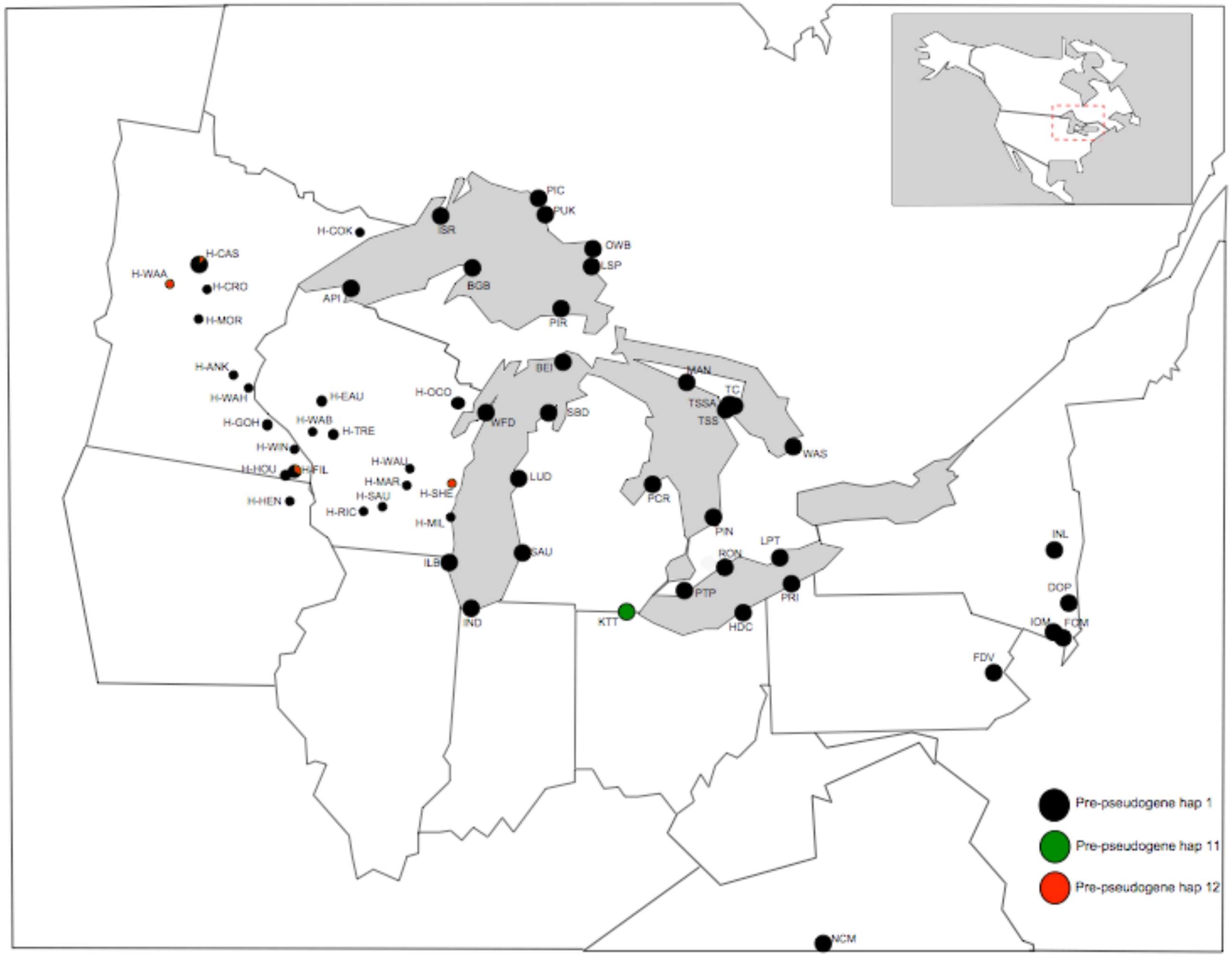


Figure 3-2a Distribution of pre-tandem repeat (P-TR) haplotypes around the Great Lakes area in North America. Full population names are given in table 1

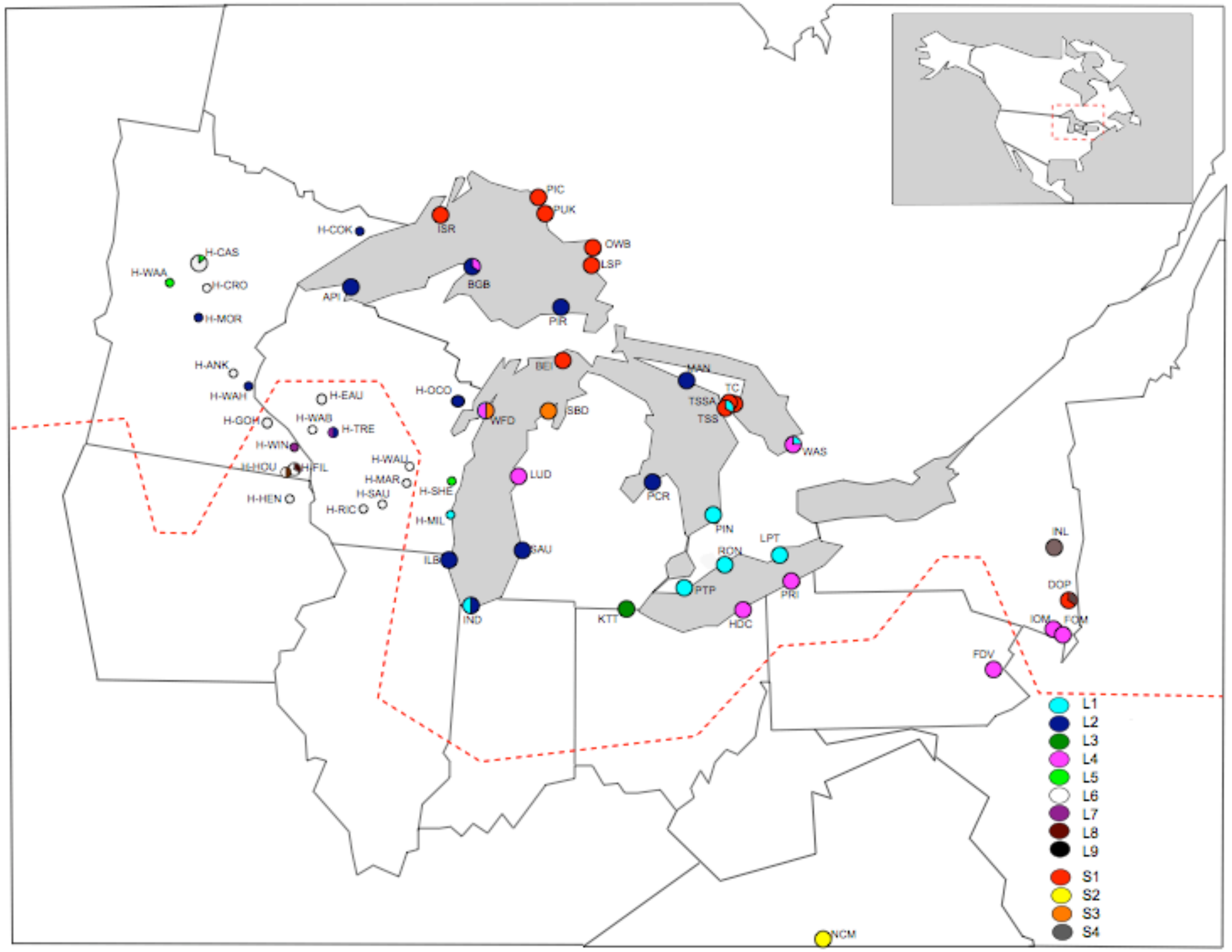


Figure 3-2b Distribution of full sequence haplotypes (L1-L9 and S1-S4) around the Great Lakes area in North America. Circles are proportional to relative frequency of haplotypes at each population. Dashed line represents the maximum southern extent of the Wisconsin ice sheets (modified from Holman, 1992).

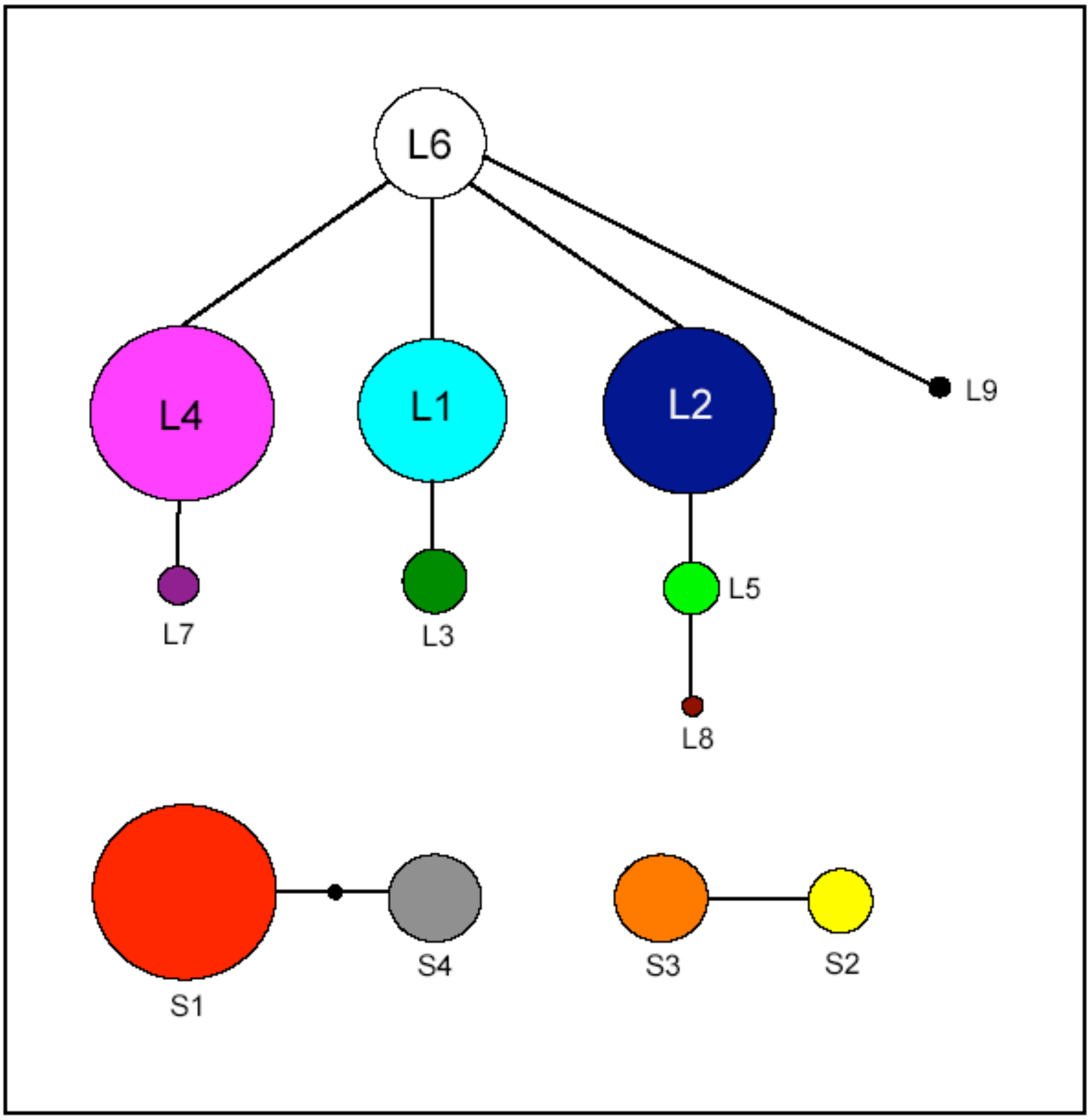


Figure 3-3a Full sequence (FS) haplotype network showing three major clades including the 'long' clade in which full sequence haplotype L6 is ancestral, and two 'short' clades (including S1 & S4 and S2 & S3 respectively) where ancestral state cannot be assigned. Circle size is proportional to frequency.

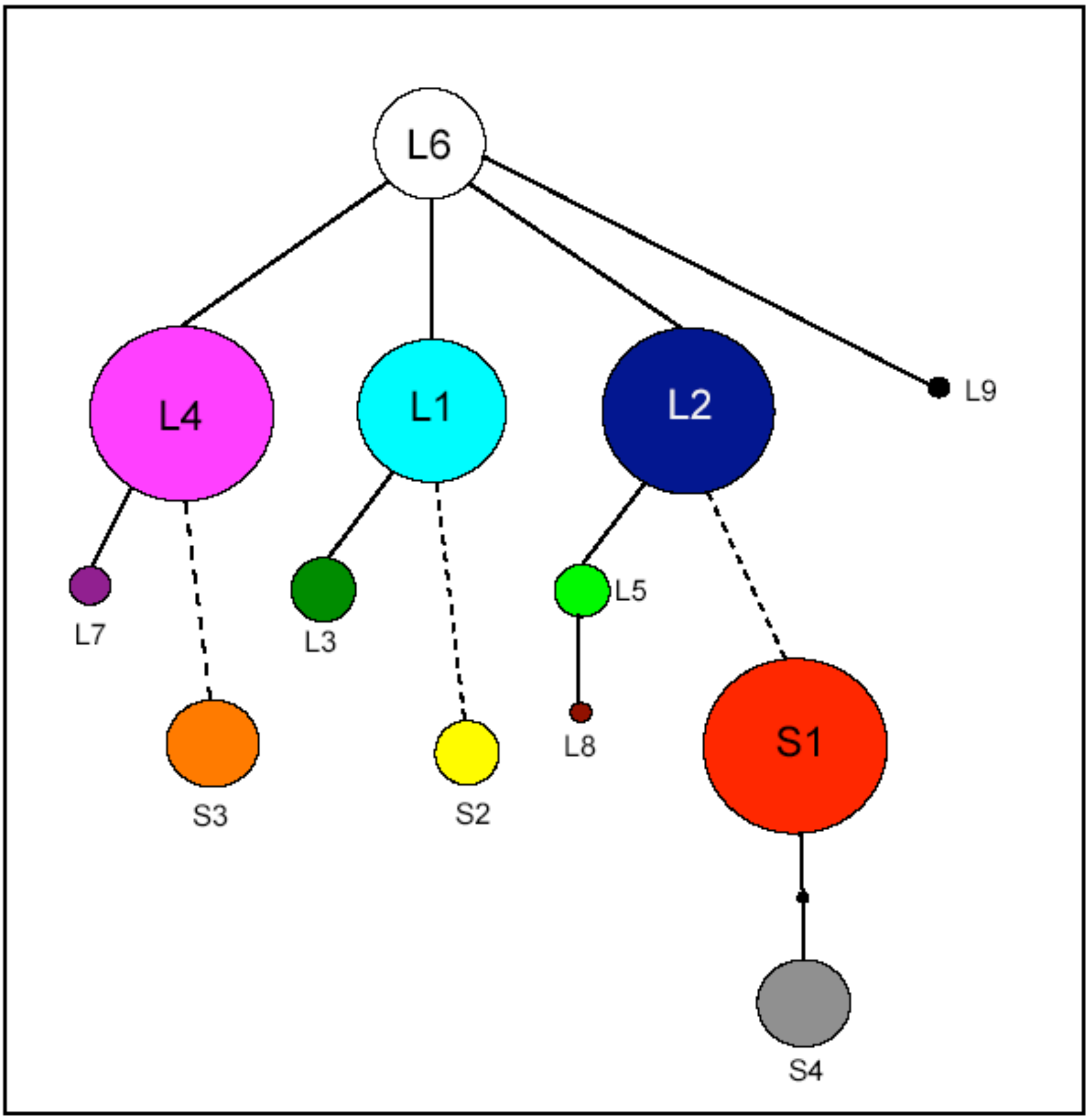


Figure 3-3b Full sequence pseudogene structure (FSPS) haplotype network showing a single major clade in which haplotype L6 is assumed ancestral. Circle size is proportional to frequency. Dashed lines represent putative pseudogene loss events.

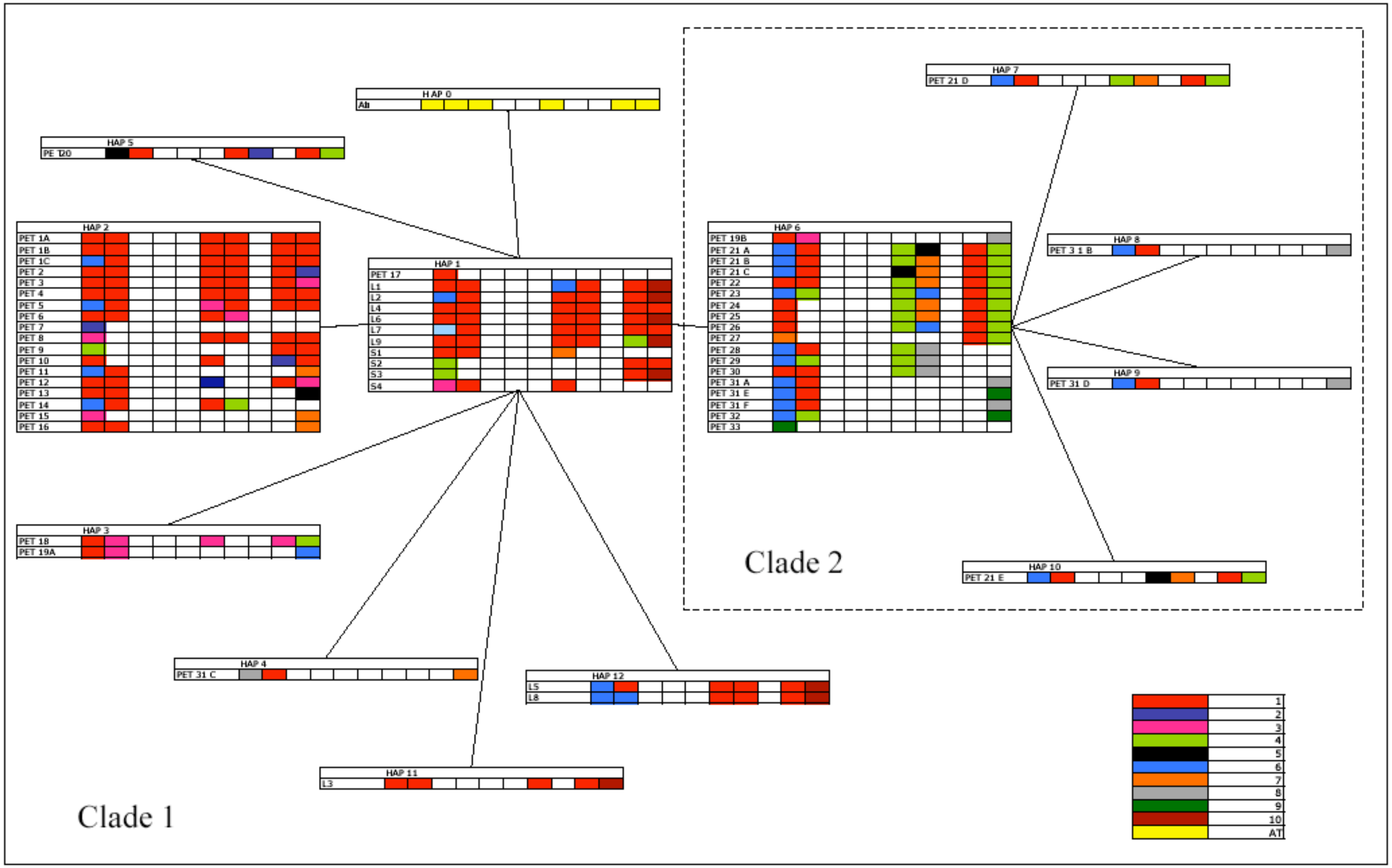


Figure 3-4 Minimum spanning network of *Arabidopsis trn*L-*trn*F pre-pseudogene haplotypes. Pre-pseudogene haplotypes 1-5 and 11-12 are Clade 1, haplotypes 6-10 are clade 2. Tandem array structure is added to each Pre-pseudogene haplotype, with each copy present in a given sequence shaded. Pseudogene copy variation is represented by colour.

4 Post-glacial expansion and population structure in North American *Arabidopsis lyrata* populations

4.1 Abstract

The perennial, outcrossing *Arabidopsis lyrata* has lost functional self-incompatibility multiple times in populations around the Great Lakes in eastern North America. It has been postulated that this is linked to different post-glacial histories, which suggests multiple routes of colonization after the last glacial maxima. I used chloroplast DNA (cpDNA), micro-satellite and allozyme markers to test if populations around the Great Lakes have different post-glacial histories, and if these are concordant with colonization routes identified in both plant and animal taxa. Chloroplast data suggested three potential lineages, which fit well with colonization routes suggested for amphibians and reptiles, however AMOVA analysis demonstrated that grouping populations based on these lineages represented the micro-satellite and allozyme data poorly. Pairwise F_{ST} values suggest significant differentiation of populations for both the micro-satellite and allozyme data sets, with Bayesian clustering analysis revealing large variation in the resultant estimates of the number of distinct groupings (K). This suggests that caution should be used when interpreting data from markers with differing levels of diversity, as it can affect levels the of population structure you find. I suggest that *A. lyrata* has colonized the Great Lakes by a minimum of two routes, one moving north along the eastern side of Lake Michigan, and the second moving northwest through New York and Ontario and around the northern side of Lake Superior.

4.2 Introduction

Dramatic climate change during the Pleistocene, 2.4 million years before present (ybp) to 12,000 ybp, contributed to a marked change in species' distributions globally (Hewitt 1996; Webb and Bartlein 1992). Temperate species in the northern hemisphere were particularly affected, as continued episodes of glacial advance and retreat resulted in latitudinal and altitudinal range shifts (Green *et al* 1996; Hewitt 1996) These range shifts were hypothesized based on chorological analysis (Dennis *et al* 1998; Hausdorf *et al* 2004), traceable through fossil records, for example pollen (Willis *et al* 1995; Willis *et al* 2004), charcoal fragments (Sumegi *et al* 2001), the elytra of beetles (Coope *et al* 1994), and remains of small mammals (Pazonyi *et al* 2004), and have in the main been confirmed by molecular studies (Hewitt 2000; Hewitt, 1999; Hewitt, 2001; Hewitt, 2004; Taberlet *et al* 1998). In North America, the most recent of these range adjustments, 18,000 ybp (Hewitt 1996), set the context for contemporary evolutionary processes, influencing the distribution of genetic diversity, the amount of genetic variation available for local adaptation and the formation of secondary contact zones between divergent lineages. Elucidating patterns of range adjustments associated with glacial cycles is thus important to our understanding of the distribution of biological diversity and the evolutionary potential of populations in temperate regions.

The biogeographical history of several North American species, particularly amphibians and reptiles, has been well examined in relation to Pleistocene climate change (Austin *et al* 2002; Carstens *et al* 2004; Church *et al* 2003; Demastes *et al* 2007; Hoffman and Blouin 2004; Nielson *et al* 2001; Placyk *et al* 2007; Steele and Storfer 2006; Wagner *et al* 2005; Zamudio and Savage 2003). These studies have led to the identification of several putative glacial refugia and postglacial colonization pathways, which are shared by a diverse range of taxonomic groups (Figure 4-1). The first, A, represents movement in a northeast direction, originating in central North America, and progressing east along the northern shore of Lake Superior. Putative route B represents a pathway spanning from refugia in non-glaciated Illinois or Indiana and moving north on both shores (east and west) of Lake Michigan along with an eastern branch radiating along the northern shore of Lake Erie. The final putative route suggested for the Great Lakes region, C, moves in a northwesterly direction along the northern shore of Lake Huron, with a branch moving west along the northern shore of Lake Erie. Its likely point of origin is refugia on the eastern coast. There is some evidence that supports two of these colonization routes for plant taxa (B & C), For example in *Arabidopsis lyrata* (Hoebe *et al* 2009; Tedder *et al* 2010) both routes have been postulated, however limited sampling prevented conclusive evidence about the location of refugia; and *Trillium grandiflorum* (Griffin & Barrett 2004), which supports route B. However in general the postglacial histories of plant taxa in the Great Lakes region, and the impact of southern refugia on patterns of genetic diversity in this region have not been well reconstructed. This is partially the result of the low number of studies that have been carried out on plant taxa in this region, and partially the result of inconclusive studies, which have suffered from insufficient sampling failing to provide conclusive evidence. Soltis *et al* (2006) provided comparative evidence (for multiple taxa) of glacial refugia, and patterns of population expansion in non-glaciated eastern North America. They suggest that while studies using mtDNA can rely on relatively few markers, cpDNA regions in angiosperms, which have been heavily utilized in studies of phylogeography, appear to have problems with lack of variation at the population level, and thus they suggest that multiple markers are required.

Taxa with northern refugia in close proximity to glaciated areas are assumed to have been the earliest species to re-invade previously glaciated terrain, acting as pioneer species (Nichols and Hewitt 1994). They may exhibit markedly different phylogeographic patterns from taxa present in more southern refugia, which become secondary or phalanx colonizers (Nichols and Hewitt 1994). The leading edge model of postglacial range expansion would suggest that primary colonizers expanded their ranges following glacial retreat via long distance dispersal events and with exponential population growth (Hewitt 1996; Nichols and Hewitt 1994), which could result in vast areas of genetic homogeneity in the northern most areas of species ranges due to successive founder effects (Hewitt 1996). This pattern differs from what is expected of phalanx colonizers, which would likely have been restricted to slower, logistic population growth as a result of habitat distribution and competition with prior colonizers, thus they will likely maintain larger effective population size and will have retained greater genetic diversity during population expansion (Hewitt 2004). It is also probable that the phylogeographic structure of pioneer colonizers may differ from that of phalanx colonizers due to differences in postglacial landscape features (i.e. geographic barriers), which influence colonization at various points in history (Nichols and Hewitt 1994; Hewitt 2004).

Long distance dispersal events likely play a key role for pioneer colonizers; Clark *et al* (1998) demonstrated that dispersal ability in temperate regions is up regulated in response to glaciations in various tree species, likely due to reduced niche space competition. Somewhat paradoxically, post-glacial seed and pollen dispersal in plants is lower than would be predicted based on carbon dated pollen density in sediment samples around eastern North America, and so appears not to explain the rapid northern movement that has clearly occurred in many species (Reid’s Paradox; Clark *et al* 1998). Modelling studies have suggested that modes of dispersal, and in particular rare long-distance dispersal events, influence the genetic diversity of glaciated regions (Ibrahim *et al* 1996). It is clear that populations moving from non glaciated refugia to de-glaciated landscapes following the retreat of the ice would have been smaller than the original source populations (Placyk *et al* 2007). This could create a bottleneck effect, resulting in decreased genetic diversity and increase the influence of genetic drift (Keller and Taylor 2008; Nei *et al* 1975).

Shifts in mating system have also been associated with colonization to new habitats opening up with glacial retreat (Stebbins 1957). Range expansion has been postulated to favour species (or populations) capable of self-fertilisation because of the benefit reproductive assurance offers (Baker 1967). In plants with a genetically controlled self-incompatibility (SI) system, this could lead to a reduction in numbers of *S*-alleles (responsible for the incompatibility response) present, and fewer compatible mating partners in recently founded populations (Vekemans *et al* 1998). Mate choice could be further limited by low plant densities, which attract fewer insect pollinators (Levin and Kerster 1969). In these situations it is predicted that self-fertilisation will present a reproductive advantage (Baker 1967), which may outweigh the potentially negative effects of inbreeding (Baker 1955; Pannell and Barrett 1998), especially during periods of rapid postglacial expansion (Baker 1955; Busch and Schoen 2008; Pannell and Barrett 1998), even for species operating a strong SI system.

Arabidopsis lyrata is a perennial plant distributed widely throughout the Northern hemisphere where it has a highly disjunct distribution. It has become a model for the study of mating system evolution, due in part to its close relatedness with the model system *A. thaliana*, and a good understanding of its sporophytic SI system (Charlesworth 2000; Charlesworth 2003; Charlesworth *et al* 2003; Kusaba *et al* 2001; Mable *et al* 2003; Nasrallah *et al* 2002; Schierup *et al* 2001). Unlike *A. thaliana*, which is a highly selfing annual (Abbott and Gomes 1989) thought to have lost its self incompatibility system independently multiple times through varying channels (Charlesworth 2005; Shimizu *et al* 2008), *A. lyrata,* is considered predominantly outcrossing. This is demonstrated by the high levels of inbreeding depression that occur when enforced selfings are implemented in European populations (Karkkainen *et al* 1999; Schierup *et* *al* 1998). In contrast to European populations of *A. lyrata*, it has been demonstrated that some North-American populations have experienced a breakdown of self-incompatibility and that this might have occurred in different genetic backgrounds (Foxe *et al* 2010; Hoebe *et al* 2009; Mable and Adam 2007). Predominantly selfing populations showed significantly lower outcrossing rates, genetic diversity and observed heterozygosity than outcrossing populations (based on nuclear micro-satellites) (Mable and Adam 2007; Mable *et al* 2005). A shift to self-compatibility within a population does not mean that the population is obligately inbreeding however. This likely represents a further evolutionary step.

Hoebe *et al* (2009) identified breakdown of SI in two distinct chloroplast (*trn*F) lineages, and remarked that the general phylogeographic pattern was consistent with colonization along a western and an eastern route, but stressed the need for increased sampling to further elucidate likely routes, and historical processes of post-glacial expansion. The *trn*F chloroplast lineages for the Great Lakes region were further delimited by Tedder *et al* (2010) (Tedder *et al* 2010) (Figure 4-2), and form the basis of a hypothesised multiple colonization route system (Figure 4-1) which is postulated to follow both branches of route B along with both branches of route C. To evaluate this, and the extent of geographic population structure in this region, I assessed micro-satellite and allozyme variation in relation to known chloroplast lineages, and compare the usefulness of each marker system when carrying out phylogeographic and population genetic studies. Differences between the resolution power of micro-satellite and allozyme loci are expected because of differences in their mutation rate and process. In particular, the higher level of variation at micro-satellite loci suggests that they should be more sensitive than allozymes to changes in population breeding size, structure and rate of dispersal. Micro-satellites are much more polymorphic than allozymes in natural populations, with more alleles, and greater heterozygosity. Mutation rates at micro-satellite loci have been estimated to be in the order of 10^{-5} to 10^{-2} (Amos *et al* 1996), which is two to four orders of magnitude higher than what is known for allozymes. Further to this, I assess population structure patterns delimited in previous studies of *A. lyrata*, including lake front of origin (Mable *et al* 2005), chloroplast lineage (Tedder *et al* 2010), mating system (Hoebe *et al* 2009) and the population structuring suggested by Foxe *et al* (2010) that featured six putative groups based on combined micro-satellite and nuclear DNA markers.

4.3 Materials and methods

4.3.1 Samples

Seeds of *Arabadopsis lyrata* were collected from 18 populations in the Great lakes region of the eastern North America. Five populations – Point Pelee (PTP), Old Woman Bay (OWB), Pictured Rock (PIR), Lake Superior Park (LSP) and Pukaskwa National Park (PUK) – were collected in 2003. The remaining 13 populations – Long Point (LPT), Rondeau (RON), Kitty Todd (KTT), Tobermoray Alvar (TSSA), Tobermoray Cliff (TC), Headland Dunes (HDC), Pinery (PIN), Presque Isle (PRI), Tobemoray Singing Sands (TSS), Sleeping Bear Dunes (SBD), Beaver Island (BEI), Port Cresent (PCR) and Indiana Dunes (IND) – were collected in 2007 (Figure 4-1; Table 4-1).

I aimed to germinate a single seed per maternal line for each population (the total number of maternal lines varied between populations), however there was a large degree of variation in germination success between populations, and so sample size varied within, and between each marker system. Between 9 and 43 individuals were raised for each population for micro-satellite analysis, with a further 6-29 raised for allozyme analysis (Table 4-1). Seeds were sown on Levington S2 + Sand mix (Scotts Professional, Ipswich) and were germinated in controlled climatic incubators under a regime of 16 hours light (20ºC) and 8 hours dark (16ºC), with 80% relative humidity.

4.3.2 Allozyme genotyping

Leaf tissue was harvested at the four-week stage from all individuals that germinated. Tissue was ground in a Tris-HCl (pH 7.5) extraction buffer (Soltis 1983), and proteins were fractionated on 12.5% hydrolysed potato starch gels under standard procedures (Vogel *et al* 1999; Wendel 1989). The following enzyme systems were then resolved with the lithium borate electrode and gel buffer system 8 of Soltis *et al* (1983): acid phosphatase (ACP, E.C. 3.1.3.2), aspartate aminotransferase (AAT, E.C. 2.6.1.1), glyceraldehyde-3-phosphate dehydrogenase (G-3-PDH E.C. 1.2.1.9), leucine aminopeptidase (LAP E.C. 3.4.11.1), malic enzyme (ME, E.C. 1.1.1.40), phosphoglucose isomerase (PGI, E.C. 5.3.1.9), shikimate dehydrogenase (SKDH, E.C. 1.1.1.25), triose-phosphate isomerase (TPI, E.C. 5.3.1.1). The following enzymes were then resolved using the morpholine-citrate (pH 7.1) electrode and gel buffer system of Wendel and Weeden (1989): alcohol dehydrogenase (ADH, E.C. 1.1.1.1), isocitrate dehydrogenase (IDH, E.C. 1.1.1.42), malate dehydrogenase (MDH, E.C. 1.1.1.37), phosphoglucomutase (PGM, E.C. 5.4.2.2), 6-phosphogluconate dehydrogenase (6-PGD, E.C. 1.1.1.44) and utp-glucose pyrophosphorylase (UGPP, E.C. 2.7.9). Staining patterns were interpreted using known enzyme sub-structuring (Kephart 1990).

4.3.3 Micro-satellite genotyping

Leaf tissue was collected from mature plants, and total genomic DNA was either extracted from 100 mg of desiccated leaf tissue using the FastDNA kit (QBiogene101, MP Biomedicals) or by DNeasy kits (Qiagen Inc) and staff of the Genome laboratory at the John Innes Center (Norwich, UK).

Eight micro-satellite loci, previously used by Mable & Adam (2007) were screened for variation: ADH-1, AthZFPG, ATTS0392, F20D22, ICE12, (Clauss *et al* 2002), LYR104, LYR133, LYR417 (obtained from V. Castric and X. Vekemans, personal communication). Products were amplified by multiplex polymerase chain reaction (PCR), using the default reagent concentrations recommended by the kit instruction manual (QIAGEN Inc). Thermocycling was performed on PTC-200 (MJ research) machines using the following programme: initial denaturation at 95°C for 15 min followed by 34 cycles of 94° for 30 s, 55°C for 90 s (ramp to 72°C at 0.7°C/s) and a final 72°C extension for 10 min. Multiplex products (1:160 dilutions) were genotyped using an ABI 3730 sequencer (by The Sequencing Service, University of Dundee). Genotypes were analysed using GENEMAPPER 4.0 (Applied Biosystems) and corrected manually.

4.3.4 Chloroplast **trn***F sequences*

We utilised full sequence chloroplast *trn*F haplotypes for eight individuals per population from previously published data (Tedder *et al* 2010) (Appendix 5). The *trn*L(UAA)–*trn*F(GAA) IGS and *trn*F gene region was amplified by PCR using the 'E' and 'F' primers of Taberlet *et al* (1991), producing three fragments with varying lengths (741bp, 5145p and 498bp) depending on pseudogene composition. Total aligned fragment length was 1076bp. Six haplotypes were represented within the samples (L1, L2, L3, L4, S1 and S3), which were split into three lineages, A (L2 & S1; coloured blue), B (L1 & L3; coloured red) and C (L4 & S3; coloured green) based on shared variation.

4.3.5 Mating system definition

Outcrossing rate estimates for the populations used in this study were taken from Foxe *et al* (2010), who used MLTR (Ritland & Jain 1981) to estimate outcrossing based on eight micro-satellite loci. Populations were classified as inbreeding if their outcrossing rate was less than 0.50 (0.01 – 0.49), and outcrossing if it was above 0.50 (0.50 – 1). Population specific outcrossing rates are given in Table 4-1.

4.3.6 Population genetic analyses

As the number of individuals screened for each marker system varied for each population, due to variation in germination success, and the specific individuals used were not the same, as whole individuals (at 4-6 weeks) were harvested for the allozyme analysis, combining the data sets was not appropriate. For this reason each of the following analyses were conducted separately for micro-satellites and allozyme data.

The program GENEPOP v4.0 (Raymond and Rousset 1995) was used to compute population allele frequencies at individual loci, as well as observed and expected heterozygosity (H_O and H_E). The average number of alleles per locus (N_a) and the number of private alleles (N_p) were calculated from the allele frequency data. The program CREATE (Coombs *et al* 2008) was used to streamline input file creation for subsequent analyses. HP-RARE (Kalinowski 2005) was used for the calculation of allelic richness (RN_a) and private allelic richness (RN_p) to estimate genetic variation using rarefaction, which uses diminution of sample size to correct for variation between populations (Kalinowski 2004). Since both the micro-satellite (which have imperfect repeats) and allozyme loci (which follow the infinite allele model) are inappropriate for models of stepwise mutation (Muller *et al* 2008; Slatkin 1995), I performed AMOVA based on pairwise F_{ST} (Slatkin 1995), as implemented in ARLEQUIN v3.11 (Excoffier *et al* 2005) to evaluate variation with respect to four different scenarios of population groupings: 1) geographic location (lakefront of origin), 2) population mating system classification, 3) cpDNA haplotype lineages, and 4) population structure suggested by Foxe *et al* (2010). They suggested six groups (K = 6) using INSTRUCT on combined nuclear DNA markers and micro-satellite data for 24 populations. As I used a subset of these populations, I used five groups (Appendix 6).

We tested for genetic isolation by distance (Slatkin 1987), by using a regression analysis of genetic similarity (M) and geographic (km) distance implemented in the computer program IBD v1.52 (Bohonak 2002). Geographic distances were generated from mean latitude and longitude values for each population (Table 4-1), using the Geographic Distance Matrix Generator v1.2.2 (http://biodiversityinformatics.amnh.org/open_source/gdmg/index.php). The WGS84 reference system was used for this conversion.

We also used ARLEQUIN v3.11 to test deviation from Hardy-Weinberg equilibrium (HWE). To test for deviation from mutation-drift equilibrium, which is expected if populations have experienced a recent demographic decline, the program BOTTLENECK v1.2 (Cornuet & Luikart, 1996) was used. The test relies on the assumption of neutral selection and mutation-drift equilibrium, and is based on the principle that when populations are near mutation-drift equilibrium, their expected heterozygosity (HEQ) is equal to the measured Hardy–Weinberg equilibrium heterozygosity (HWE) in non-bottlenecked populations. Alternatively, if a population has experienced a recent population decline, mutation-drift equilibrium is temporarily disrupted and HWE will be significantly greater than calculated from the number of alleles sampled (Luikart & Cornuet, 1998). To test for significance, I used the Wilcoxon signed-rank test under the infinite allele model (IAM), and the two-phase models. For each mutational model, 5000 replicates were performed

The population history of *A. lyrata* based on coalescent theory (Ciofi *et al* 1999) was investigated for the micro-satellite data set using the 2MOD program. This method uses the comparison of the relative likelihoods for a model of migration-drift equilibrium (gene flow model) versus ancestral population fragmentation into independent units diverging purely by drift (drift model) and estimates the Bayes factor, the ratio between the likelihoods of the gene flow model and the drift model. A Markov chain Monte Carlo simulation (100,000 iterations) was computed with the first 10% of the output discarded in order to avoid bias to the starting conditions.

Finally, I used the Bayesian clustering algorithm implemented in STRUCTURE v2.1 (Pritchard *et al* 2000) to analyze population and individual clustering. This method uses a multilocus genotype to probabilistically assign individuals to one or more (where populations are admixed) clusters. I ran 10 simulations per prior *K* (*K*=1 to *K*=18); a burn-in period of 1,000,000; number of MCMC replicates after burn-in = 500,000; and used the admixture model with the default settings. I plotted *K* vs. the likelihoods obtained over all simulations and inferred *K* as recommended by the STRUCTURE manual. I used STRUCTURE HARVESTER v0.46 to calculate the first order derivative of the distribution of –Ln probability of the data (–Ln*P*(d) (L'(*K*))), the second order derivative(L''(*K*)), and the second order derivative divided by the standard deviation (dK (L''(*K*)/stdev)) to detect the optimal number of clusters based on recommendations by Evanno *et al* (Evanno *et al* 2005). CLUMPP v1.1 was used to combine simulation outputs for each *K*.

4.4 Results

4.4.1 Genetic diversity and HWE

Measures of allelic diversity are given in Table 4-2. The rarefacted allelic richness (*RN_a*) was significantly lower for allozyme loci than for micro-satellite loci (two tailed *t*-test, *P* = 0.00). In all populations, for both marker systems, observed heterozygosity (*H_O*) was lower than the expected heterozygosity (*H_E*) when averaged across all loci (Table 4-2) with the exception of BEI, which showed a *H_O* of 0.41 and a *H_E* of 0.39 for the micro-satellite data. When considering loci individually, several populations showed higher *H_O* than *H_E* for each marker system (Tables 3a & b). When averaged within mating systems, *H_O* was 0.07 for the inbreeding populations and 0.27 for the outcrossing populations for the micro-satellites. For the allozymes, *H_O* was 0.03 for the inbreeding populations and 0.08 for the outcrossers.

Table 3a-4 shows the polymorphism results for the eight micro-satellite loci across all 18 populations of *A. lyrata*. The number of alleles per loci ranged from 1-6, and gene diversities (*H_O*) per locus per population ranged from 0.042 to 0.710. Departure from HWE was found in 25 of 99 cases where the loci were polymorphic, most commonly associated with the AthZFPG locus. Mean *H_O* for inbreeding populations was 0.07, and for outcrossing populations was 0.27. All populations except BEI, IND and PRI exhibited significant heterozygosity deficit, although only KTT, LPT, LSP, PIN, OWB, PCR, PTP, PUK, RON, SBD, TC, TSSA and TSS were significant after Bonferroni correction.

Table 3b-4 shows the results for the 12 allozyme loci. The number of alleles per locus ranged from 1-4, with gene diversities (*H_O*) ranging between 0.045 – 0.604. Departure from HWE was found in 17 of the 63 cases where loci were polymorphic. Four loci (TPI-1, ADH, AAT-1 and 6-PGD-2) were monomorphic in every population. Mean *H_O* for inbreeding populations was 0.03, and for outcrossing populations was 0.08. Ten populations deviated significantly from HWE (BEI, IND, KTT, LSP, PIN, PRI, PTP, RON, TSSA and TSS) although this was reduced to seven after bonferroni correction (BEI, KTT, PRI, PTP, RON, TSSA and TSS). HWE could not be calculated for LPT as it was monomorphic at each locus.

4.4.2 Population differentiation

Population differentiation was tested using pairwise F_{ST}. Table 4a-4 shows that all population comparisons for the micro-satellite data were significant. For the allozyme data (Table 4b-4), only a single comparison was not significant, which was between PUK and RON.

The AMOVA-based comparisons for groupings considered significant in earlier work, showed that variation was principally among populations within groups and within populations. This meant that in general, each of the structuring scenarios used in this study, did not represent either data set particularly well. For microsatellites, when population geographic location (groupings shown in Appendix 6), or more specifically the lakefront on which each population was located, was used as a grouping variable, 4% of the variation was contained among groups, 47% among populations within groups and 49% within populations. For the allozyme data set, 0% of the variation was contained among groups, with among population variation and within group variation making up 60% and 40%, respectively. Using population mating system as a grouping scenario (based on T_m from Hoebe *et al* 2010), 0% of the variation could be explained among groups, and 50% of variation was explained by both among populations within groups and within populations for the micro-satellite data. The allozyme data followed a similar pattern, with 60% of the variation explained among populations within groups and 40% within populations. When chloroplast lineage as defined by Tedder *et al* (2010) (Figure 4-2) was used to set the groupings, 13% of the variation was explained among groups, 40% among populations within groups and 47% within the populations for the micro-satellite loci. For the allozyme loci, 13% of variation was explained among groups, with 47% among populations within groups and 40% within populations for this scenario. The final scenario tested was groupings based on the findings of Foxe *et al* (2010), who delimited K=6 (using INSTRUCT v1.0 [Gao *et al* 2007]) as the most appropriate K based on combined nuclear and micro-satellite data for 24 populations of *A. lyrata* in the Great lakes region. For this scenario, the micro-satellite data showed that 16% of the variation could be explained between groups, with 36% and 48% explained among populations among populations within groups, and within populations, respectively. The allozyme data showed that 14% of the variation could be explained between groups, 45% among populations within groups and 41% within populations.

Mantel tests for association between genetic distance (based on F_{ST} /1-F_{ST}) and geographic distance showed no significant pattern of isolation by distance for either the micro-satellite data (r^2 = 0.011, P = 0.215) or the allozyme data (r^2 = 0.021, P = 0.111) (Figure 4-3). The same was true when considering the method described by Rousset (Rousset 1997) using log transformed geographical distances (data not shown).

4.4.3 Mutation-drift equilibrium and migration-drift equilibrium

For the micro-satellite data, six populations deviated significantly from mutation-drift equilibrium under the IAM with the Wilcoxon test (OWB, TSS, SBD, PUK, BEI, & PCR) all of which are outcrossing. All populations showed significant deviations for the allozyme data set (Table 4-5).

The MCMC procedure implemented in the 2MOD program was used to test the relative likelihood of gene flow versus drift models. The analysis revealed that populations of *A. lyrata* significantly adhered to a gene flow model (*P* (gene flow) = 0.997, Bayes factor = 200), indicating that the history of this species has been characterized by a balance of genetic drift and migration rather than by the action of genetic drift only.

4.4.4 Population structure

Following the Evanno *et al* (2005) method of assigning *K*, the most parsimonious and meaningful number of clusters for the micro-satellite data was *K* = 14 (Figure 4-4a). Eleven of the suggested clusters were single populations (BEI, HDC, KTT, LPT, LSP, OWB, PIR, PRI, PUK, SBD & TC), which occurred over a wide geographic range (Figure 4-1). The final three clusters were formed by PTP & RON, PCR & PIN, and TSSA & TSS. Each of these clusters was formed between populations that are geographically close (66 miles, 129 miles and 1 mile, respectively; Table 4-4). The IND population showed a high degree of admixture, which prevented it being assigned to any other cluster.

For the allozyme data, *K* = 3 was assigned as most meaningful (Figure 4-4b). The first cluster included IND, KTT, RON, PCR, SBD, BEI, PIR, PUK, TSSA & TC, the second cluster included PTP, LPT, PIN, TSS, LSP & OWB, with the final cluster containing HDC & PRI.

4.5 Discussion

4.5.1 Genetic variation

The analysis of micro-satellite and allozyme markers for 18 populations of *A. lyrata* showed considerable variation in relative levels of diversity between marker systems, characteristic of their different mutation rates. While the micro-satellite loci were all polymorphic (although not for every population), three of the allozyme loci were monomorphic, and in general (across all 12 loci) represented lower over-all diversity. Despite the difference in absolute diversity values, the patterns of diversity for rarefacted private alleles (RN_p), allelic richness (RN_a), and observed and expected heterozygosity (H_O and H_E respectively) were similar for the two marker systems. Population BEI formed the only exception to this, with a H_O greater than its H_E for the micro-satellite markers, but this was non-significant.

When compared directly to previous studies using micro-satellite markers, which used populations included in this study (nine populations by Hoebe *et al* [2009], and 24 populations (the 18 used in our study, plus six further populations by Foxe *et al* [2010]), I find the same patterns for H_O and H_E, with several exceptions. While our results agree with those of Hoebe *et al* (2009) in principle (H_O is lower than H_E), absolute values vary slightly in both directions, as do patterns of RN_p, and RN_a. When compared to those from Foxe *et al* (2010), there are two major differences. For both HDC and LSP they find H_O is greater than H_E, which is not the case for our data (Table 4-6). Further to this, KTT has no observed heterozygosity in the Foxe *et al* (2010). It is possible that these differences are the result of different sample sizes. While the sample sizes of this study and that of Hoebe *et al* (2009) are similar, those used by Foxe *et al* (2009) a much smaller, and this may account for the different patterns of diversity reported.

The chloroplast marker (*trn*F) used in this study has well documented pseudogene duplications (Ansell *et al* 2010; Koch *et al* 2007; Tedder *et al* 2010), which are responsible for much of the variation represented in the six haplotypes present in this study. Earlier studies have suggested caution when using the pseudogenes sections (tandem array) of the *trn*F gene due to structurally mediated parallel changes in pseudogene copies, and elected to base their interpretations on only the pre-tandem array portion of the fragment (Ansell *et al* 2010; Koch and Matschinger 2007; Schmickl *et al* 2010). As North American populations of *A. lyrata* are relatively genetically depauperate, and there is limited nucleotide diversity in the pre-tandem array region alone, I followed the recommendation of Tedder *et al* (2010), which suggested that including the tandem array, allowed for greater certainty when assessing relationships between sequences that differ by pseudogene content, thus allowing us to make inferences about evolutionary links between populations that have varying sequence length, and to recover otherwise cryptic population structure.

4.5.2 Population differentiation

Based on pairwise F_{ST} values, each population was significantly differentiated from each other for both marker systems, with the exception of RON-PUK in the allozyme system. As these populations are geographically isolated from each other (763 km), it is unlikely that there has been recent gene flow between them, or that there are unsampled intermediate populations between them that are admixed by gene flow. This suggests that this may be an anomaly related to sample size for the PUK population, which was very low (n = 6; Table 4-1). These findings are congruent with previous work for the micro-satellite data set, and the allozyme data set with the exception of the RON-PUK link (Hoebe *et al* 2009), which found the same general pattern of differentiation using micro-satellite loci, except that RON and PTP were not significantly different. It is possible that variation in sample sizes and marker system between this study, and that of Hoebe *et al* (2009) are responsible for this difference.

None of the four putative population history scenarios, as tested by AMOVA analysis, explained a substantial proportion of the variation based on

F_{st}. It is perhaps unsurprising that this was the case for the first scenario, lake front of origin, as the non-uniform glacial retreat would likely have opened up different areas of terrain at different times, thus allowing differential colonization of parts of the same lake shore. However, Mable *et al* (2007) suggested that lake shore of origin was a significant grouping variable. The number of populations sampled was much smaller however (n = 12), as was the effective number of micro-satellite loci used (n = 7), which may account for this. Our AMOVA findings for lake front of origin are complemented by the cpDNA lineages, which demonstrate multiple haplotypes on a single lake front (Figure 4-2). As in other studies using this species (Mable *et al* 2005; Mable & Adam 2007; Hoebe *et al* 2009), there appears to be no pattern of isolation by distance using either marker system. This is perhaps not surprising, considering that multiple colonization events may lead to multiple highly differentiated populations being located in a small geographic area. This is indicated by the F_{ST} values, which demonstrate significant differentiation between populations. Isolation by distance has been observed in other plant species with similar post-glacial colonization histories in the Great Lakes region, including the small herb, White Trillium (*Trillium grandiflora*) (Griffin *et al* 2004), where allozyme markers were also used. This may suggest that either differing patterns of drift and gene flow are working between the two systems, or modes of dispersal were different.

Mating system variation as a basis for population grouping was equally poor at explaining between group variation, with none of the variation from either marker system explained by this classification. The underlying assumption for using mating system as a grouping parameter (based on T_m), was that all populations with a T_m below 0.5 (classified as inbreeding) had arisen from a single loss of SI event and so would comprise a single group. However, it has been demonstrated that this assumption might not be valid for these populations (Hoebe *et al* 2009; Foxe *et al* 2010), which demonstrated that loss of SI and subsequent shift to inbreeding in multiple chloroplast lineages were independent, and our findings support this conclusion. Multiple transitions from SI to SC are likely to be favoured in post-glacial colonization events to allow reproductive assurance (Baker 1966).

While the cpDNA *trn*F gene has been extensively used as a marker for phylogeographic study (Ansell *et al* 2007, 2008, 2010; Koch *et al* 2007; Koch & Matschinger 2007), variation at this loci in *A. lyrata* is low. When the three pre-defined chloroplast lineages (Figure 4-2; Appendix 5), each comprising two linked haplotypes, which have previously been suggested as phylogeographically informative (Tedder *et al* 2010) are used as a grouping variables, only 13% of variation is explained between groups for both marker systems. Although this grouping is clearly not responsible for the majority of variation observed with either marker system, I feel that this may be biased by the slow rate of chloroplast evolution, and its uni-parental mode of inheritance, which is not efficient at recovering population level structure (Soltis *et al* 2006).

Perhaps the most surprising result, is for the fourth scenario, groupings based on those suggested by the INSTRUCT results of Foxe *et al* (2010). Using their combined micro-satellite and nuclear gene sequence data, they proposed *K*=6 as the most parsimonious number of clusters for their 24 sampled populations. Our modified groupings based on this result (modification was only needed as I had fewer populations; Appendix 6), showed 16% and 14% of variation explained between groups for micro-satellites and allozymes, respectively, which is relatively low. While Foxe *et al* (2010) had a larger number of populations; they only used eight individuals per population. It is possible that these eight individuals represent less of the population wide diversity than is present in our larger sample size (Table 4-1). Or that the nuclear marker data, with which the micro-satellite data was combined, shows different patterns diversity than seen in our study.

4.5.3 Mutation-drift equilibrium and migration-drift equilibrium

Given the well documented, genetically depauperate nature of North American *A. lyrata* when compared with European populations (Balañá-Alcaide *et al* 2006; Clauss and Mitchell-Olds 2006; Ross-Ibarra *et al* 2008), it is perhaps unsurprising that the test for bottlenecks provided clear evidence to their existence (Table 4-5) for both marker systems. While the allozyme system suggested that each population had been through a recent bottleneck, the micro-satellite data set provided a less general pattern. All six populations classified as inbreeding by the T_{m} estimates showed no significant evidence for a bottleneck. This could be due to insufficient power, linked to their mating system. Evidence for recent bottlenecks were found in OWB, TSS, SBD, PUK, BEI and PCR. It is possible that these populations represent early pioneers, at the end of a stepping-stone succession process whereby each subsequent population has lower genetic diversity as a direct result of its series of small bottlenecks, which have had no subsequent migration of new genetic material (Hewitt 1996). Of the remaining populations, which did not show evidence of a significant recent bottleneck (HDC, PIN, PIR, PRI, LSP and IND) it appears likely that populations on the southern tips of both Lake Michigan (IND), and Lake Erie (HDC and PRI) represent either an earlier ’stop’ on the path of colonization, and as such the population will have greater genetic diversity, or a more recent wave of colonization, with new genetic material coming into an existing population. PIR and LSP on Lake Superior may represent a zone of introgression between hypothetical colonization routes A and C (Figure 4-1), which is supported by levels of admixture seen in these populations in the allozyme structure results (Table 4b-4). Potential patterns of colonization are supported by the 2MOD results that suggest current population structure is the result of structured colonization (migration and drift), as opposed to drift alone.

4.5.4 Population structure and potential colonization routes

There is a marked difference between the level of population structure elucidated by STRUCTURE for the micro-satellite data set and allozyme data set, which directly reflects the level of genetic variation present in each system. For the micro-satellite analyses, K = 14 is supported by our F_{ST} estimates: This indicates a significant level of differentiation between populations, but is quite different from the structuring proposed by earlier studies of the species (Hoebe *et al* 2009; Foxe *et al* 2010) that assigned K = 6 as most informative. Despite the clear difference in K, there are similarities between the clusters formed in this study, and the earlier work. PTP and RON are consistently linked in all three studies (Hoebe *et al* 2009; Foxe *et al* 2010; Tedder *et al* 2010), which reflects their close geographic proximity, as are TSS and TSSA. PCR and PIN also cluster according to Foxe *et al* (2010), however they were linked to PTP and RON, which conflicts with our findings. PCR was not sampled by Hoebe *et al* (2009), but PIN did not cluster with PTP and RON, which supports our findings. While both previous studies have used the same micro-satellite loci as used here, the number of populations, and sample size within populations, has varied. The pattern of Hoebe *et al* (2009) is potentially explained by the small number of populations (n = 13) that it samples, and because of this its findings complement ours, as it recovered similar clustering patterns (PTP & RON, TSSA & TSS, and PIN and LSP forming distinct single population clusters), with the exception of OWB and LSP, which cluster for Hoebe *et al* (2009), but not for the results presented here. The findings of the Foxe *et al* (2010) study, however, are less congruent, and could be attributed to the nuclear gene sequence data they used in conjunction with the micro-satellites, the smaller sample sizes per population and/or the larger number of populations that they considered (n = 24). While nuclear sequences may be less able to find geographic structure if the time since population isolation is relatively short, they do allow you to estimate confidence limits on parameters such as times of divergence, patterns of population expansion, effects of natural selection, or levels of gene flow, which makes them a powerful tool for phylogeography (Brito and Edwards 2008), perhaps combining the slower rate of mutation (relative to micro-satellites) with micro-satellite markers allows you to fine tune the population structure resolution you achieve. Thus allowing identification of both colonization history patterns, and very local patterns of gene flow, based on distribution of rare alleles. Recent work by Fogelqvist *et al* (2010) has suggested that population sample size can have a direct effect on *K* assignment, with 5-10 samples per population likely enough to ascertain meaningful structure. The *K* = 3 assignment based on the allozyme marker system, suggests some groupings that contradict those delimited by the micro-satellites. PTP & RON, PIN & PCR and TSS & TSSA do not cluster together as they did for the micro-satellites in this, and previous studies (Hoebe *et al* 2009; Mable & Adam 2007; Foxe *et al* 2010), or the chloroplast lineages (Tedder *et al* 2010). Mutation rate may explain some of this, with the micro-satellite data representing a more recent snapshot of diversity due to its faster mutation rate than allozyme markers.

As the micro-satellite population structure suggests a large degree of population level differentiation, it offers little evidence as to the colonization history of the area when interpreted alone, as reported by Mable & Adam (2007). However when interpreted along with the allozyme population structure, and chloroplast lineages, it may offer insight. Tedder *et al* (2010) postulated three potential colonization routes (A, B & C: Figure 4-1) for the Great Lakes region, which correlated with those proposed for other species (Austin *et al* 2002; Carstens *et al* 2004; Church *et al* 2003; Demastes *et al* 2007; Hoffman and Blouin 2004; Nielson *et al* 2001; Placyk *et al* 2007; Steele and Storfer 2006; Wagner *et al* 2005; Zamudio and Savage 2003). As fewer populations were used in this study, the chloroplast data only support two principle colonization routes: B) spanning from putative refugia in the south of

Indiana and Illinois (or further south in the Appalachians), moving north towards the tip of Lake Michigan, where it splits and continues north on both sides of the lake; and C) moving north from the mid-eastern region of North America, approaching the lakes from the east side, and moving west through New York, Pennsylvania and Ontario. Colonization route A is not supported by the populations used in this study, and this likely relates to sampling strategy, however the chloroplast data provided in Tedder *et al* (2010) does support this route. They show that populations to the west of Lake Michigan are dominated by a single chloroplast haplotype, L6, with high frequency in both non-glaciated areas and de-glaciated ones. They postulate that this haplotype may have been ancestral, and follows the same colonization route as various amphibian colonizers (Masta *et al* 2003; Hoffman & Blouin 2004). The allozyme data fits these putative colonization routes (B & C) well, with an east-west divide evident in the population structure (yellow and blue; figure 4-4b), characteristic of two colonization routes. This pattern is obscured slightly by a zone of introgression, where the two routes meet, particularly on Lake Huron (TSS, TSSA and TC), and this may go some way to explaining the separation of the PTP and RON populations.

4.5.5 Micro-satellites vs. allozymes markers

As expected based on mutational rate variation, there is a marked difference between the absolute levels of genetic diversity seen in both marker systems, and the subsequent structuring patterns they suggest. Observed genetic diversity within populations was always higher for micro-satellite markers (than for allozyme markers) as demonstrated by basic diversity indices (Table 4-2) and loci specific diversity (Table 3a-4 & 3b). This should theoretically allow for changes in population breeding size, structure and rate of dispersal to be more precisely delimited using micro-satellite markers than by using allozymes. Despite a difference in the magnitude of variation displayed by each marker system, the general patterns of diversity were, on the whole, the same for both systems. Pairwise F_{ST} estimates suggest that there is significant differentiation between all populations for both marker systems, with the exception of PUK-RON for the allozyme system that is likely explained by low sample size, and patterns of H_o and H_e are also consistent for both systems. The differing levels of marker system diversity are well illustrated by the test for population bottlenecks. The allozyme marker system, because of the low levels of diversity, suggests that all populations have been through a significant bottleneck (Table 4-5). While this may indeed be the case, it is in stark contrast to the micro-satellite data, which suggests that only six outcrossing populations have experienced recent population bottlenecks (Table 4-5). The difference in the ability of each marker system to distinguish population structure is evident in the STRUCTURE results; the high private allele numbers at the population level coupled with drift reduce the amount of common diversity, and thus populations are split into a large number of clusters for the micro-satellite data (K = 14). This may mask the underlying population structure, and make inferences about colonization history difficult. The allozyme data however, with its lower relative diversity, and more commonly shared alleles, appears to favour many fewer population clusters than the micro-satellite data suggests (K = 3), and may give greater ability to define colonization history. Despite the greater number of clusters proposed by the micro-satellites in this study relative to previous ones, there are similarities in the populations that form clusters between the studies. However, the number of, and populations that form clusters for the allozyme data do not support clustering patterns found by the micro-satellite data in this study, or in previous studies. The allozyme data does fit more closely with patterns of diversity elucidated by chloroplast haplotypes however, and also supports patterns of colonization suggest for other species within the Great Lakes region, which the micro-satellites in this study do not. The level of discord between the two systems in relation to population structure may suggest that the loci used in one or both of the systems are not fully appropriate to answer questions about phylogeography at this scale, in North American *A. lyrata*. Mable *et al* (2007) also reported high levels of diversity and population segregation using micro-satellite markers on a subset of the populations used in this study. Their suggestion was that different markers, with lower levels of diversity maybe required to fully realize colonization history, and routes of post-glacial expansion. The results of our study suggest that while the micro-satellite data gives a snapshot of a populations current status, it is not allowing meaningful inferences regarding colonization, as has been seen for other marker systems (Tedder *et al* 2010; Foxe *et al* 2010), and other species (Austin *et al* 2002; Carstens *et al* 2004; Church *et al* 2003; Demastes *et al* 2007; Hoffman and Blouin 2004; Nielson *et al* 2001; Placyk *et al* 2007; Steele and Storfer 2006; Wagner *et al* 2005; Zamudio and Savage 2003; Griffin and Barrett 2004). The allozyme data however appears concordant with the general patterns suggested for other marker systems, and for other species in the same geographic area.

4.5.6 Conclusions

While increasing the number of *A. lyrata* populations sampled in North America compared to previous work (Mable *et al* 2005; Mable and Adam 2007; Hoebe *et al* 2009), and the number of samples per population compared to recent work by Foxe *et al* (2010) this study has not fully resolved the colonization route history of the current populations. Rather than increase our ability to resolve population structure, adding a further marker system (allozymes) to those used in previous works (Mable *et al* 2005; Mable and Adam 2007; Hoebe *et al* 2009) has added a further layer of complexity regarding appropriate levels of diversity for answering phylogeographic questions. Perhaps contrary to general predictions about the suitability of micro-satellites for population level studies, here I find that high numbers of private alleles and low shared variation make linking populations together inappropriate. This is turn makes predictions about colonization history based on these patterns difficult. While allozymes are not without problems, the patterns elucidated in this study do support earlier work, which suggests at least two independent colonization routes in the Great Lakes (Mable and Adams 2007; Hoebe *et al* 2009; Tedder *et al* 2010). Sampling of *A. lyrata* around the great lakes region has provided some insight into colonization history, however with the exception of relatively few herbarium samples used in Tedder *et al* (2010) and a single population (eight individuals) sampled in North Carolina in Foxe *et al* (2010), no sampling has been done in areas that were not covered by ice during the last glacial maxima. I can of course speculate as to the likely location of refugia based on other species (Holman 1995; Soltis *et al* 2006; Lee *et al* 2008), but further sampling particularly in regions south of the Great Lakes in the mid-east and the Appalachians, where putative glacial refugia have been suggested is essential. This should allow the identification of potential glacial refugia relevant to this species, and further clarify colonization history as has been done more extensively in other taxa (Holman 1995; Placyk *et al* 2007; Soltis *et al* 2006).

Table 4-1 Sampled populations, population-level outcrossing rates (T_m), geographic co-ordinates, collector (initials) and collection date and sample sizes used for each marker system. Populations are ordered by increasing outcrossing rates.

					Sample Size				
Population name	Code	T_m‡	Decimal degrees N	Decimal degrees W	cpDNA†	Micro-satellite	Allozyme	Date collected	Collector
Point Pelee	PTP	0.09	41.93	-82.51	8	24	12	2003	BM
Long Point	LPT	0.13	42.58	-80.39	8	24	12	2007	BM. AT. PH.
Rondeau	RON	0.28	42.26	-81.85	8	26	29	2007	BM. AT. PH.
Kitty Todd	KTT	0.31	41.62	-83.79	8	24	24	2007	BM. AT. PH.
Tobermory Alvar	TSSA	0.41	45.19	-81.59	8	21	19	2007	BM. AT. PH.
Tobermory Cliff	TC	0.48	45.25	-81.52	8	43	12	2007	BM. AT. PH.
Old Woman Bay	OWB	0.64	47.79	-84.9	8	14	25	2003	BM
Headland Dunes	HDC	0.65	41.76	-81.29	8	24	24	2007	BM. AT. PH.
Pinery	PIN	0.84	43.27	-81.83	8	27	15	2007	BM. AT. PH.
Pictured Rock	PIR	0.88	46.67	-86.02	8	27	23	2007	YW
Presque Isle	PRI	0.89	42.17	-80.07	8	22	18	2007	YW
Tobermory Singing Sands	TSS	0.91	45.19	-81.58	8	27	15	2007	BM. AT. PH.
Lake Superior Park	LSP	0.94	47.57	-84.97	8	27	15	2003	BM
Sleeping Bear Dunes	SBD	0.94	44.94	-85.87	8	21	21	2007	BM. AT. PH.

Table 4-1. ***Continued***

Pukaskwa National Park	PUK	0.96	48.4	-86.19	8	20	6	2003	BM
Beaver Island	BEI	0.98	45.76	-85.51	8	25	22	2007	YW
Port Cresent	PCR	0.98	44	-83.07	8	14	15	2007	BM. AT. PH.
Indiana Dunes	IND	0.99	41.62	-87.21	8	9	15	2007	BM. AT. PH.

‡ Data taken from Foxe *et al* 2010, † Data taken from Tedder *et al* 2010.

Table 4-2 Sampled populations [ordered by increasing T_m (see Table 4-1)], total number of private alleles (N_p), total number of alleles (N_a), rarefacted [private (RN_p)] allelic richness (RN_a) to correct for sample size, and observed (H_O) and expected (H_E) heterozygosity for 9 micro-satellite loci and 12 allozyme loci.

	Micro-satelite loci							Allozyme loci						
Population	**N**	N_p	N_a	RN_p†	RN_a ‡	H_O	H_E	**N**	N_p	N_a	RN_p†	RN_a ‡	H_O	H_E
PTP	24	0	9	0	2.3	0.02	0.06	12	0	15	0	1.23	0.01	0.10
LPT	24	0	10	0.05	2.44	0.01	0.11	12	0	12	0	1.00	0.00	0.00
RON	26	1	10	0.12	2.44	0.04	0.11	29	0	14	0	1.16	0.03	0.06
KTT	24	0	11	0	1.25	0.02	0.05	24	1	15	0.03	1.12	0.01	0.05
TSSA	21	1	17	0.12	1.9	0.13	0.19	19	0	20	0.01	1.42	0.07	0.13
TC	43	0	21	0	2.27	0.17	0.33	12	0	17	0	1.28	0.06	0.08
OWB	14	1	14	0.12	2.98	0.11	0.27	25	1	16	0.02	1.23	0.06	0.06
HDC	24	0	11	0	2.64	0.11	0.14	24	0	13	0	1.08	0.03	0.04
PIN	27	1	18	0.04	1.94	0.19	0.27	15	0	15	0	1.22	0.06	0.08
PIR	27	3	20	0.17	2.06	0.24	0.27	23	0	16	0	1.32	0.12	0.13
PRI	22	0	12	0	2.6	0.17	0.2	18	1	16	0.03	1.28	0.06	0.12
TSS	27	0	23	0	2.66	0.38	0.43	15	0	20	0.05	1.49	0.08	0.13
LSP	27	2	18	0.23	1.94	0.13	0.2	15	0	19	0.02	1.55	0.16	0.19
SBD	21	3	31	0.33	3.41	0.5	0.58	21	2	18	0.07	1.38	0.11	0.13
PUK	20	0	17	0	3.26	0.31	0.38	6	0	13	0	1.08	0.03	0.03
BEI	25	3	23	0.17	2.42	0.41	0.39	22	0	17	0	1.34	0.05	0.11
PCR	14	1	20	0.11	2.42	0.33	0.4	15	0	18	0	1.38	0.08	0.09
IND	9	0	20	0.2	3.75	0.39	0.46	15	0	17	0	1.33	0.09	0.12

Rarefacted (private) † allelic ‡ richness: average number of private allele(s) per locus corrected for sample size using the method by Kalinowski (2004).

Table 4-3a Population diversity and HWE test based on eight micro-satellite markers.

	Micro-satellite loci							
	ADH	ATHZ	ATTS	F20D22	ICE12	LYR104	LYR133	LYR417
PTP								
N_a	1	1	1	1	2	1	2	1
H_O	-	-	-	-	0.042	-	0.083	-
H_E	-	-	-	-	0.042	-	0.383	-
P	-	-	-	-	NS	-	**	-
LPT								
N_a	1	2	1	2	1	1	2	1
H_O	-	0.042	-	0.000	-	-	0.042	-
H_E	-	0.503	-	0.082	-	-	0.191	-
P	-	**	-	*	-	-	**	-
RON								
N_a	1	2	2	1	2	1	1	1
H_O	-	0.000	0.000	-	0.308	-	-	-
H_E	-	0.075	0.265	-	0.434	-	-	-
P	-	*	**	-	NS	-	-	-
KTT								
N_a	1	2	3	1	1	1	1	1
H_O	-	0.083	0.042	-	-	-	-	-
H_E	-	0.284	0.121	-	-	-	-	-
P	-	**	*	-	-	-	-	-
TSSA								
N_a	2	3	2	2	1	2	2	3
H_O	0.263	0.263	0.056	0.105	-	0.211	0.053	0.158
H_E	0.371	0.514	0.056	0.102	-	0.341	0.149	0.152
P	NS	**	NS	NS	-	NS	NS	NS
TC								
N_a	3	4	1	5	2	2	2	2
H_O	0.098	0.186	-	0.233	0.000	0.357	0.286	0.310
H_E	0.256	0.638	-	0.560	0.092	0.380	0.506	0.313
P	**	**	-	**	**	NS	**	NS
OWB								
N_a	3	4	1	2	2	1	1	1
H_O	0.357	0.000	-	0.214	0.214	-	-	-
H_E	0.519	0.710	-	0.304	0.495	-	-	-
P	NS	**	-	NS	NS	-	-	-
HDC								
N_a	2	3	1	2	2	2	1	1
H_O	0.042	0.667	-	0.042	0.042	0.042	-	-
H_E	0.042	0.584	-	0.042	0.311	0.042	-	-
P	NS	*	-	NS	**	NS	-	-
PIN								
N_a	1	3	2	3	3	2	2	2
H_O	-	0.462	0.115	0.115	0.115	0.037	0.407	0.296
H_E	-	0.497	0.111	0.410	0.298	0.037	0.409	0.425
P	-	NS	NS	**	**	NS	NS	NS
PIR								
N_a	2	2	4	1	2	3	3	3
H_O	0.389	0.481	0.115	-	0.222	0.259	0.222	0.370
H_E	0.437	0.509	0.183	-	0.257	0.237	0.400	0.314
P	NS	NS	NS	-	NS	NS	NS	NS

Table 3a-4. *Continued*

PRI								
N_a	1	3	2	2	1	1	2	1
H_O	-	0.500	0.381	0.045	-	-	0.273	-
H_E	-	0.521	0.511	0.045	-	-	0.359	-
P	-	NS	NS	NS	-	-	NS	-
TSS								
N_a	3	6	2	4	2	2	2	2
H_O	0.481	0.655	0.207	0.345	0.036	0.379	0.414	0.400
H_E	0.492	0.774	0.189	0.540	0.363	0.354	0.479	0.467
P	NS	NS	NS	*	**	NS	NS	NS
LSP								
N_a	1	3	3	6	2	2	1	1
H_O	-	0.211	0.346	0.231	0.077	0.259	-	-
H_E	-	0.626	0.438	0.391	0.075	0.283	-	-
P	-	**	NS	**	NS	NS	-	-
SBD								
N_a	5	6	4	4	3	3	2	4
H_O	0.571	0.476	0.333	0.619	0.429	0.524	0.524	0.579
H_E	0.633	0.710	0.571	0.599	0.512	0.626	0.470	0.602
P	NS	*	*	NS	NS	NS	NS	NS
PUK								
N_a	2	3	2	4	2	2	2	1
H_O	0.200	0.300	0.150	0.350	0.400	0.450	0.300	-
H_E	0.262	0.528	0.142	0.449	0.467	0.481	0.328	-
P	NS	NS	NS	NS	NS	NS	NS	-
BEI								
N_a	3	7	1	4	2	3	2	1
H_O	0.417	0.750	-	0.583	0.375	0.542	0.625	-
H_E	0.401	0.745	-	0.577	0.439	0.414	0.510	-
P	NS	NS	-	NS	NS	NS	NS	-
PCR								
N_a	1	3	3	5	3	2	2	2
H_O	-	0.500	0.643	0.500	0.214	0.143	0.500	0.154
H_E	-	0.632	0.624	0.619	0.415	0.138	0.516	0.271
P	-	NS	NS	NS	NS	NS	NS	NS
IND								
N_a	2	4	2	4	3	3	2	1
H_O	0.222	0.444	0.111	0.444	0.222	0.778	0.333	-
H_E	0.209	0.601	0.111	0.732	0.386	0.647	0.294	-
P	NS	NS	NS	NS	NS	NS	NS	-

N_a, number of alleles at each loci; H_E unbiased estimates of expected heterozygosity; H_O observed heterozygosity; P, probability for the HWE test.

* P < .05; ** P < .01; NS, not significant; -, monomorphic loci, non estimated.

Table 4-3b Population diversity and HWE test based on 12 allozyme loci.

	Allozyme loci											
	PGI-2	**IDH**	**UPGG-2**	**TPI-1**	**TPI-2**	**ADH**	**AAT-1**	**AAT-2**	**PGM-2**	**SKD**	**PGM-3**	**6-PGD-2**
PTP												
N_a	2	1	2	1	1	1	2	1	1	1	1	1
H_O	0.000	-	0.000	-	-	-	0.083	-	-	-	-	-
H_E	0.507	-	0.159	-	-	-	0.519	-	-	-	-	-
P	**	-	*	-	-	-	**	-	-	-	-	-
LPT												
N_a	1	1	1	1	1	1	1	1	1	1	1	1
H_O	-	-	-	-	-	-	-	-	-	-	-	-
H_E	-	-	-	-	-	-	-	-	-	-	-	-
P	-	-	-	-	-	-	-	-	-	-	-	-
RON												
N_a	2	1	1	1	1	1	2	1	1	1	1	1
H_O	0.241	-	-	-	-	-	0.172	-	-	-	-	-
H_E	0.421	-	-	-	-	-	0.312	-	-	-	-	-
P	*	-	-	-	-	-	*	-	-	-	-	-
KTT												
N_a	1	2	1	1	2	1	2	1	1	1	1	1
H_O	-	0.417	-	-	0.417	-	0.083	-	-	-	-	-
H_E	-	0.417	-	-	0.417	-	0.479	-	-	-	-	-
P	-	NS	-	-	NS	-	**	-	-	-	-	-
TSSA												
N_a	3	2	2	1	1	1	2	1	2	2	2	1
H_O	0.105	0.052	0.055	-	-	-	0.166	-	0.000	0.333	0.211	-
H_E	0.104	0.052	0.055	-	-	-	0.385	-	0.203	0.489	0.273	-
P	NS	NS	NS	-	-	-	*	-	**	NS		-
TC												
N_a	3	1	1	1	1	1	2	1	1	2	2	1
H_O	0.333	-	-	-	-	-	0.083	-	-	0.083	0.250	-
H_E	0.54	-	-	-	-	-	0.083	-	-	0.083	0.228	-
P	NS	-	-	-	-	-	NS	-	-	NS	NS	-

Table 4-3b. *Continued*

OWB												
N_a	4	1	2	1	1	1	1	1	1	1	1	1
H_O	0.600	-	0.120	-	-	-	-	-	-	-	-	-
H_E	0.585	-	0.184	-	-	-	-	-	-	-	-	-
P	NS	-	NS	-	-	-	-	-	-	-	-	-
HDC												
N_a	1	1	2	1	1	1	1	1	1	1	1	1
H_O	-	-	0.333	-	-	-	-	-	-	-	-	-
H_E	-	-	0.421	-	-	-	-	-	-	-	-	-
P	-	-	NS	-	-	-	-	-	-	-	-	-
PIN												
N_a	1	1	1	1	1	1	2	1	1	2	2	1
H_O	-	-	-	-	-	-	0.133	-	-	0.133	0.400	-
H_E	-	-	-	-	-	-	0.128	-	-	0.331	0.514	-
P	-	-	-	-	-	-	NS	-	-	NS	NS	-
PIR												
N_a	2	1	1	1	1	1	2	1	2	2	1	1
H_O	0.304	-	-	-	-	-	0.478	-	0.217	0.455	-	-
H_E	0.321	-	-	-	-	-	0.450	-	0.264	0.496	-	-
P	NS	-	-	-	-	-	NS	-	NS	NS	-	-
PRI												
N_a	1	1	1	1	1	1	3	1	2	2	1	1
H_O	-	-	-	-	-	-	0.333	-	0.222	0.117	-	-
H_E	-	-	-	-	-	-	0.509	-	0.457	0.470	-	-
P	-	-	-	-	-	-	NS	-	NS	**	-	-
TSS												
N_a	1	2	2	1	1	1	3	1	2	2	2	1
H_O	-	0.067	0.067	-	-	-	0.467	-	0.133	0.133	0.267	-
H_E	-	0.067	0.186	-	-	-	0.503	-	0.342	0.497	0.405	-
P	-	NS	NS	-	-	-	NS	-	*	**	NS	-

Table 4-3b. *Continued*

LSP												
N_a	3	2	2	1	1	1	1	1	3	2	1	1
H_O	0.733	0.000	0.200	-	-	-	-	-	0.333	0.600	-	-
H_E	0.687	0.239	0.186	-	-	-	-	-	0.604	0.508	-	-
P	*	**	NS	-	-	-	-	-	*	NS	-	-
SBD												
N_a	3	2	1	1	1	1	2	1	1	2	2	1
H_O	0.238	0.095	-	-	-	-	0.286	-	-	0.143	0.524	-
H_E	0.221	0.092	-	-	-	-	0.316	-	-	0.396	0.494	-
P	NS	NS	-	-	-	-	NS	-	-	**	NS	-
PUK												
N_a	2	1	1	1	1	1	1	1	1	1	1	1
H_O	0.333	-	-	-	-	-	-	-	-	-	-	-
H_E	0.303	-	-	-	-	-	-	-	-	-	-	-
P	NS	-	-	-	-	-	-	-	-	-	-	-
BEI												
N_a	1	2	1	1	1	1	2	1	2	2	2	1
H_O	-	0.045	-	-	-	-	0.272	-	0.000	0.090	0.182	-
H_E	-	0.045	-	-	-	-	0.241	-	0.473	0.304	0.304	-
P	-	NS	-	-	-	-	NS	-	**	**	NS	-
PCR												
N_a	2	1	2	1	1	1	3	1	2	2	1	1
H_O	0.2	-	0.133	-	-	-	0.333	-	0.067	0.267	-	-
H_E	0.186	-	0.129	-	-	-	0.297	-	0.186	0.331	-	-
P	NS	-	NS	-	-	-	NS	-	NS	NS	-	-
IND												
Na	3	1	1	1	1	1	2	1	1	2	2	1
H_O	0.067	-	-	-	-	-	0.467	-	-	0.200	0.400	-
H_E	0.191	-	-	-	-	-	0.517	-	-	0.370	0.331	-
P	*	-	-	-	-	-	NS	-	-	NS	NS	-

N_a, number of alleles at each loci; H_E unbiased estimates of expected heterozygosity; H_O observed heterozygosity; P, probability for the HWE test. * $P < .05$; ** $P < .01$; NS, not significant; -, monomorphic loci, non estimated.

Table 4-4a Comparison between geographic and genetic distances for eight micro-satellite loci. Pairwise geographic distances (km) above the diagonal and pairwise F_{ST} values below the diagonal.

	BEI	HDC	IND	KTT	LPT	LSP	PIN	OWB	PCR	PIR	PRI	PTP	PUK	RON	SBD	TC	TSSA	TSS
BEI	-	523	350	443	506	237	368	262	239	118	564	453	335	450	72	296	293	292
HDC	0.56*	-	497	208	118	709	174	729	289	639	110	103	833	72	512	389	383	383
IND	0.29*	0.63*	-	304	559	578	444	604	369	430	592	395	648	442	280	541	534	534
KTT	0.39*	0.74*	0.33*	-	300	669	244	692	271	560	315	111	778	176	406	443	436	435
LPT	0.40*	0.58*	0.31*	0.54*	-	662	140	679	269	615	61	190	790	125	513	311	306	306
LSP	0.44*	0.48*	0.45*	0.63*	0.46*	-	538	25	424	153	722	658	130	640	301	369	371	371
PIN	0.35*	0.54*	0.17*	0.42*	0.09*	0.42*	-	557	129	479	196	160	663	112	373	222	215	215
OWB	0.39*	0.62*	0.31*	0.43*	0.40*	0.43*	0.33*	-	445	177	739	680	118	660	326	382	385	384
PCR	0.57*	0.60*	0.61*	0.72*	0.66*	0.57*	0.54*	0.62*	-	352	325	236	546	218	246	185	177	177
PIR	0.28*	0.51*	0.27*	0.37*	0.35*	0.31*	0.30*	0.39*	0.62*	-	674	571	224	566	162	370	369	368
PRI	0.18*	0.42*	0.23*	0.40*	0.32*	0.26*	0.25*	0.25*	0.46*	0.25*	-	205	850	150	567	371	366	366
PTP	0.14*	0.52*	0.32*	0.40*	0.41*	0.32*	0.35*	0.36*	0.51*	0.21*	0.15*	-	777	66	432	379	372	372
PUK	0.62*	0.78*	0.55*	0.75*	0.56*	0.70*	0.53*	0.68*	0.83*	0.55*	0.50*	0.60*	-	763	386	499	501	500
RON	0.55*	0.78*	0.55*	0.67*	0.48*	0.62*	0.47*	0.44*	0.80*	0.47*	0.47*	0.54*	0.82*	-	440	334	327	327
SBD	0.63*	0.71*	0.66*	0.73*	0.57*	0.71*	0.45*	0.67*	0.73*	0.62*	0.51*	0.62*	0.80*	0.78*	-	343	338	338
TC	0.54*	0.73*	0.55*	0.68*	0.46*	0.50*	0.45*	0.48*	0.76*	0.43*	0.44*	0.49*	0.76*	0.21*	0.73*	-	8	8
TSSA	0.36*	0.39*	0.32*	0.45*	0.39*	0.30*	0.36*	0.23*	0.52*	0.35*	0.22*	0.34*	0.53*	0.49*	0.61*	0.51*	-	1
TSS	0.40*	0.77*	0.44*	0.44*	0.59*	0.58*	0.49*	0.48*	0.68*	0.41*	0.35*	0.22*	0.81*	0.75*	0.80*	0.71*	0.50*	-

*** Indicates significant differences (*P* < 0.05) based on 1023 permutations.**

Table 4-4b Comparison between geographic and genetic distances for 12 allozyme loci. Pairwise geographic distances (km) above the diagonal and pairwise F_{ST} values below the diagonal. Only a single comparison is not significant, between RON and PTP (underlined).

	BEI	HDC	IND	KTT	LPT	LSP	PIN	OWB	PCR	PIR	PRI	PTP	PUK	RON	SBD	TC	TSSA	TSS
BEI	-	523	350	443	506	237	368	262	239	118	564	453	335	450	72	296	293	292
HDC	0.74*	-	497	208	118	709	174	729	289	639	110	103	833	72	512	389	383	383
IND	0.31*	0.60*	-	304	559	578	444	604	369	430	592	395	648	442	280	541	534	534
KTT	0.48*	0.83*	0.58*	-	300	669	244	692	271	560	315	111	778	176	406	443	436	435
LPT	0.55*	0.92*	0.63*	0.89*	-	662	140	679	269	615	61	190	790	125	513	311	306	306
LSP	0.26*	0.77*	0.48*	0.53*	0.58*	-	538	25	424	153	722	658	130	640	301	369	371	371
PIN	0.42*	0.83*	0.57*	0.66*	0.61*	0.36*	-	557	129	479	196	160	663	112	373	222	215	215
OWB	0.49*	0.88*	0.67*	0.77*	0.80*	0.35*	0.71*	-	445	177	739	680	118	660	326	382	385	384
PCR	0.47*	0.70*	0.27*	0.59*	0.71*	0.52*	0.52*	0.74*	-	352	325	236	546	218	246	185	177	177
PIR	0.35*	0.63*	0.23*	0.28*	0.67*	0.42*	0.46*	0.69*	0.12*	-	674	571	224	566	162	370	369	368
PRI	0.53*	0.35*	0.31*	0.63*	0.69*	0.57*	0.57*	0.73*	0.31*	0.33*	-	205	850	150	567	371	366	366
PTP	0.65*	0.81*	0.62*	0.85*	0.71*	0.57*	0.68*	0.78*	0.69*	0.68*	0.67*	-	777	66	432	379	372	372
PUK	0.59*	0.88*	0.48*	0.85*	0.95*	0.51*	0.77*	0.78*	0.59*	0.54*	0.64*	0.71*	-	763	386	499	501	500
RON	0.57*	0.79*	0.38*	0.79*	0.74*	0.60*	0.74*	0.75*	0.54*	0.52*	0.61*	0.69*	**0.09**	-	440	334	327	327
SBD	0.11*	0.67*	0.13*	0.42*	0.60*	0.35*	0.50*	0.55*	0.32*	0.20*	0.43*	0.65*	0.46*	0.42*	-	343	338	338
TC	0.54*	0.85*	0.59*	0.54*	0.88*	0.33*	0.66*	0.73*	0.56*	0.35*	0.65*	0.79*	0.70*	0.69*	0.46*	-	8	8
TSSA	0.08*	0.72*	0.29*	0.34*	0.65*	0.28*	0.47*	0.53*	0.40*	0.25*	0.49*	0.68*	0.59*	0.57*	0.07*	0.44*	-	1
TSS	0.20*	0.74*	0.42*	0.54*	0.43*	0.21*	0.18*	0.49*	0.48*	0.42*	0.50*	0.55*	0.61*	0.64*	0.34*	0.55*	0.25*	-

* Indicates significant differences ($P < 0.05$) based on 1023 permutations.

Table 4-5 Wilcoxon sign rank test assessing the presence of population level bottlenecks by comparing H_E *and* H_O under the IAM mutation process (Cornuet and Luikart, 1996) for 12 allozyme loci and eight micro-satellite loci.

Marker	PTP	LPT	RON	KTT	TSSA	TC	OWB	HDC	PIN	PIR	PRI	TSS	LSP	SBD	PUK	BEI	PCR	IND
Allozyme	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**	**
Micro-satellite	NS	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	**	NS	**	*	**	*	NS

* *P* < .05; ** *P* < .01; NS, not significant.

Table 4-6 Comparative observed (H_O) and expected (H_E) heterozygosity estimates for the micro-satellite loci used in this study, and previously published work.

Population	Micro-satellites			Hoebe *et al* (2009)			Foxe *et al* (2010)		
	N	H_O	H_E	N	H_O	H_E	N	H_O	H_E
PTP	24	0.02	0.06	23	0.03	0.08	8	0.07	0.08
LPT	24	0.01	0.11	23	0.02	0.1	8	0.07	0.13
RON	26	0.04	0.11	21	0.04	0.06	8	0.03	0.03
KTT	24	0.02	0.05	-	-	-	8	0	0.04
TSSA	21	0.13	0.19	12	0.11	0.15	8	0.10	0.18
TC	43	0.17	0.33	-	-	-	8	0.18	0.28
OWB	14	0.11	0.27	18	0.18	0.25	8	0.35	0.30
HDC	24	0.11	0.14	-	-	-	8	0.07	0.14
PIN	27	0.19	0.27	25	0.23	0.26	8	0.18	0.26
PIR	27	0.24	0.27	-	-	-	8	0.26	0.26
PRI	22	0.17	0.20	-	-	-	8	0.14	0.15
TSS	27	0.38	0.43	11	0.24	0.37	8	0.36	0.45
LSP	27	0.13	0.2	16	0.08	0.11	8	0.14	0.11
SBD	21	0.50	0.58	-	-	-	8	0.42	0.54
PUK	20	0.31	0.38	-	-	-	8	0.34	0.34
BEI	25	0.41	0.39	-	-	-	8	0.32	0.39
PCR	14	0.33	0.40	-	-	-	8	0.21	0.30
IND	9	0.39	0.46	-	-	-	8	0.43	0.47

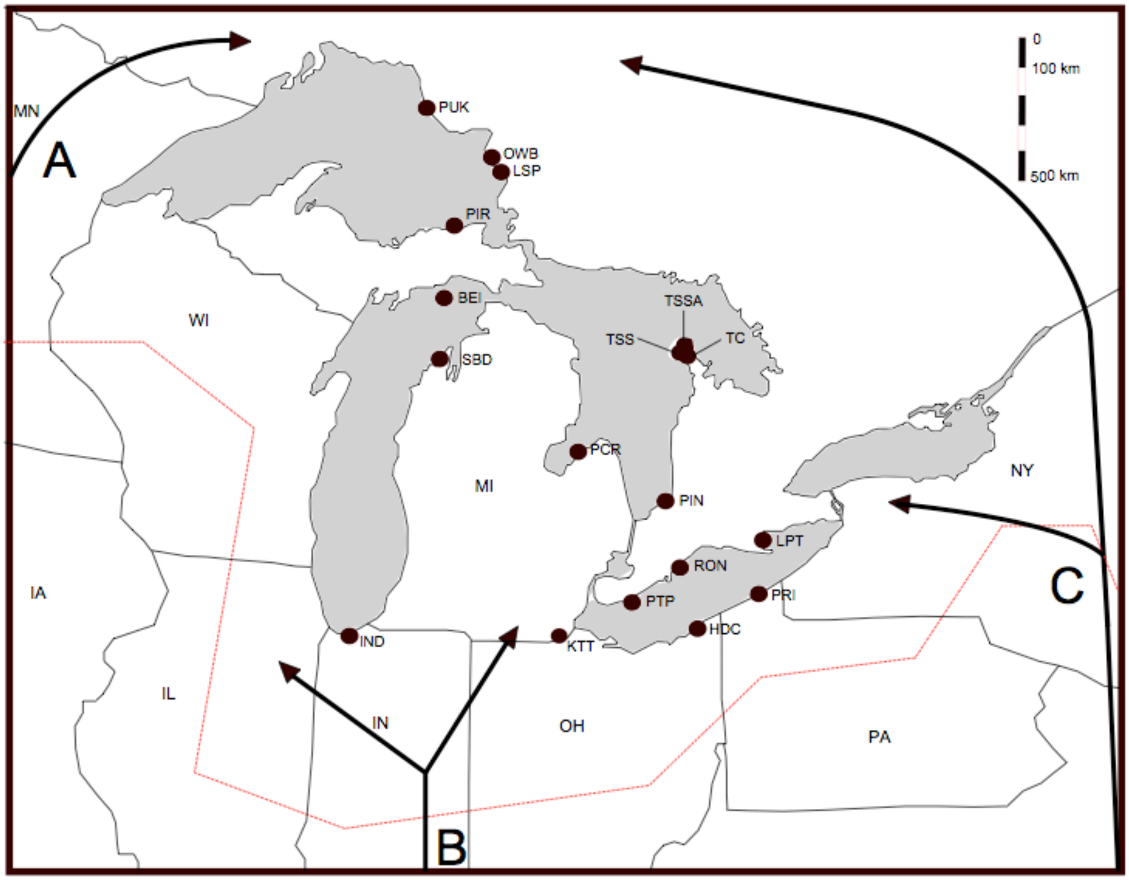


Figure 4-1 Sampling locations for 18 populations of *Arabidopsis lyrata* around the Great lakes region in eastern North America. Population names are represented by three characters, with full population details given in Table 1. The arrows denote putative colonization routes synthesized from the literature for amphibian, reptiles and and plants: A **= Great Plains: *Bufo woodhousii* (Masta *et al* 2003), *Rana pipiens* (Hoffman & Blouin 2004);** B **= Midwest: *Arabidopsis lyrata* (Hoebe *et al* 2009; Tedder *et al* 2010), *Trillium grandiflorum* (Griffin & Barrett 2004), *R. pipiens* (Hoffman & Blouin 2004), *Thamnophis sirtalis* (Placyk *et al* 2007);** C **= Coastal Plains: *A. lyrata* (Tedder *et al* 2010), *Ambystoma maculatum* (Zamudio & Savage 2003), *Ambystoma tigrinum tigrinum* (Church *et al* 2003) *Rana catesbeiana* (Austin *et al* 2004). The dashed line represents the maximum southern extent of the Wisconsin ice sheets (modified from Holman, 1992).**

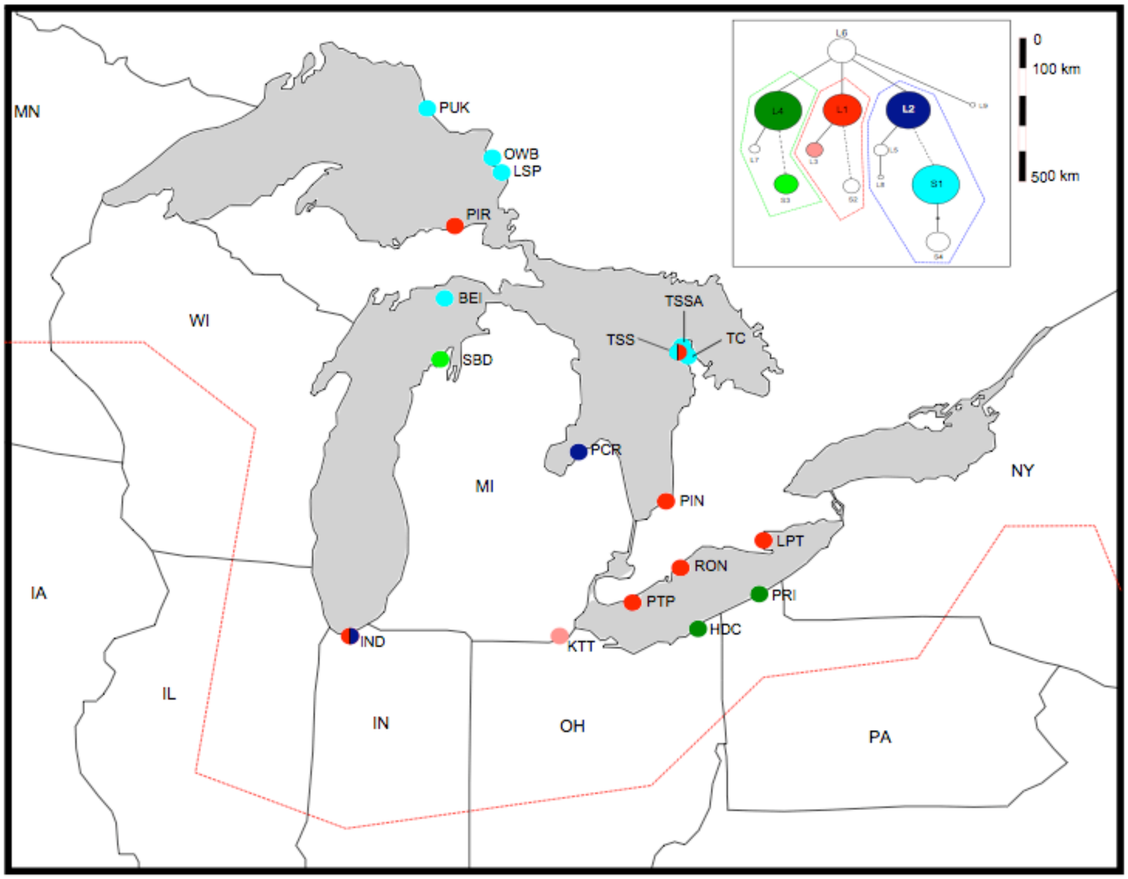


Figure 4-2 Distribution of *trn*F chloroplast haplotypes (Tedder *et al* 2010) around the Great Lakes area in North America (full haplotype details given in Table 1). Inset: Haplotype colours are representative of the chloroplast lineage from which they are derived. Non-coloured haplotypes are not represented in the sampled populations.

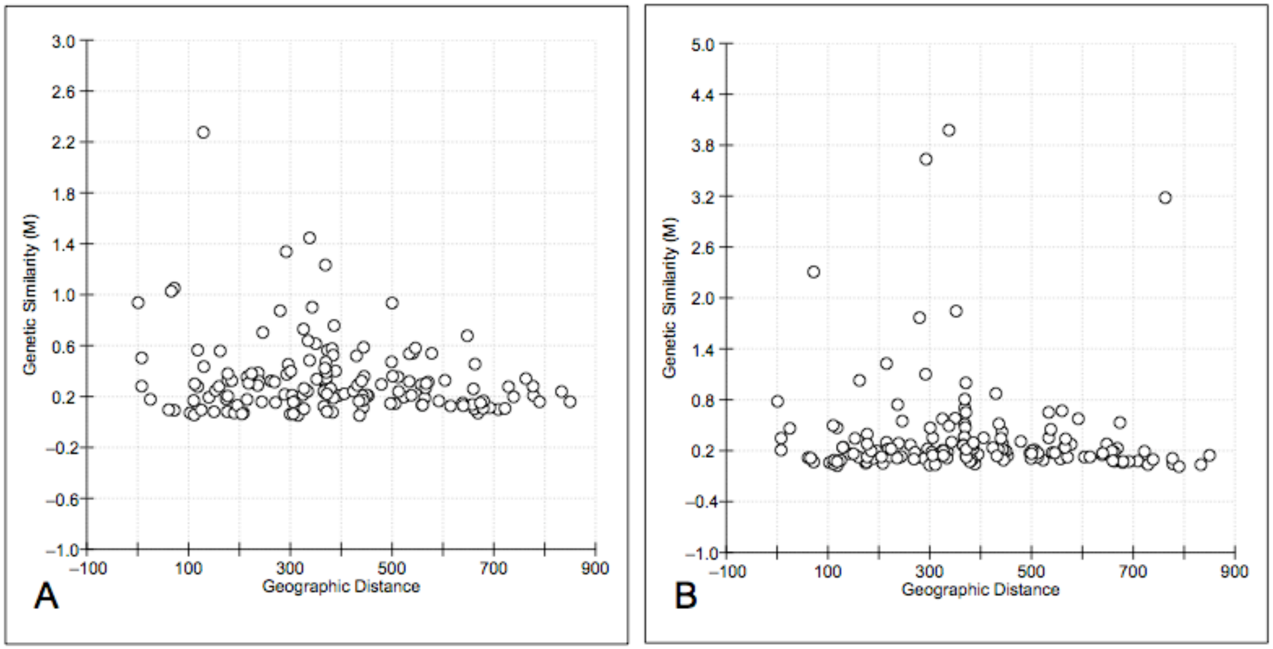


Figure 4-3 Scatter plots of pair-wise genetic similarity (M) and geographic distance (km) between 18 populations of *A. lyrata* for (a) **micro-satellites;** (b) **allozymes.**

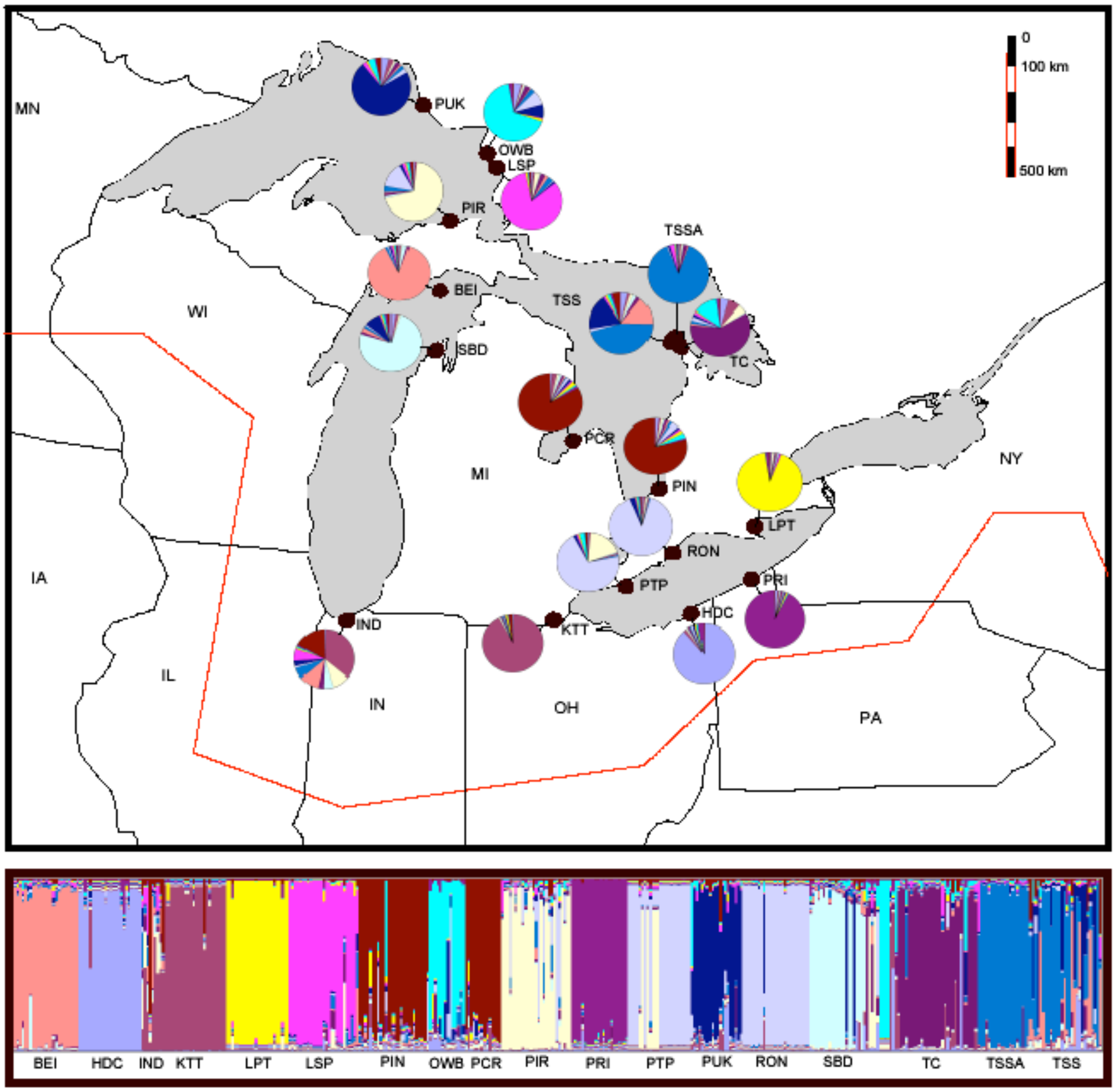


Figure 4-4a Posterior probability of Bayesian clustering analysis (STRUCTURE) using eight micro-satellite loci, based on a prior of 14 clusters (*K* = 14). Bar plots show individual posterior probabilities, and pie charts on the map show mean posterior probabilities for each population.

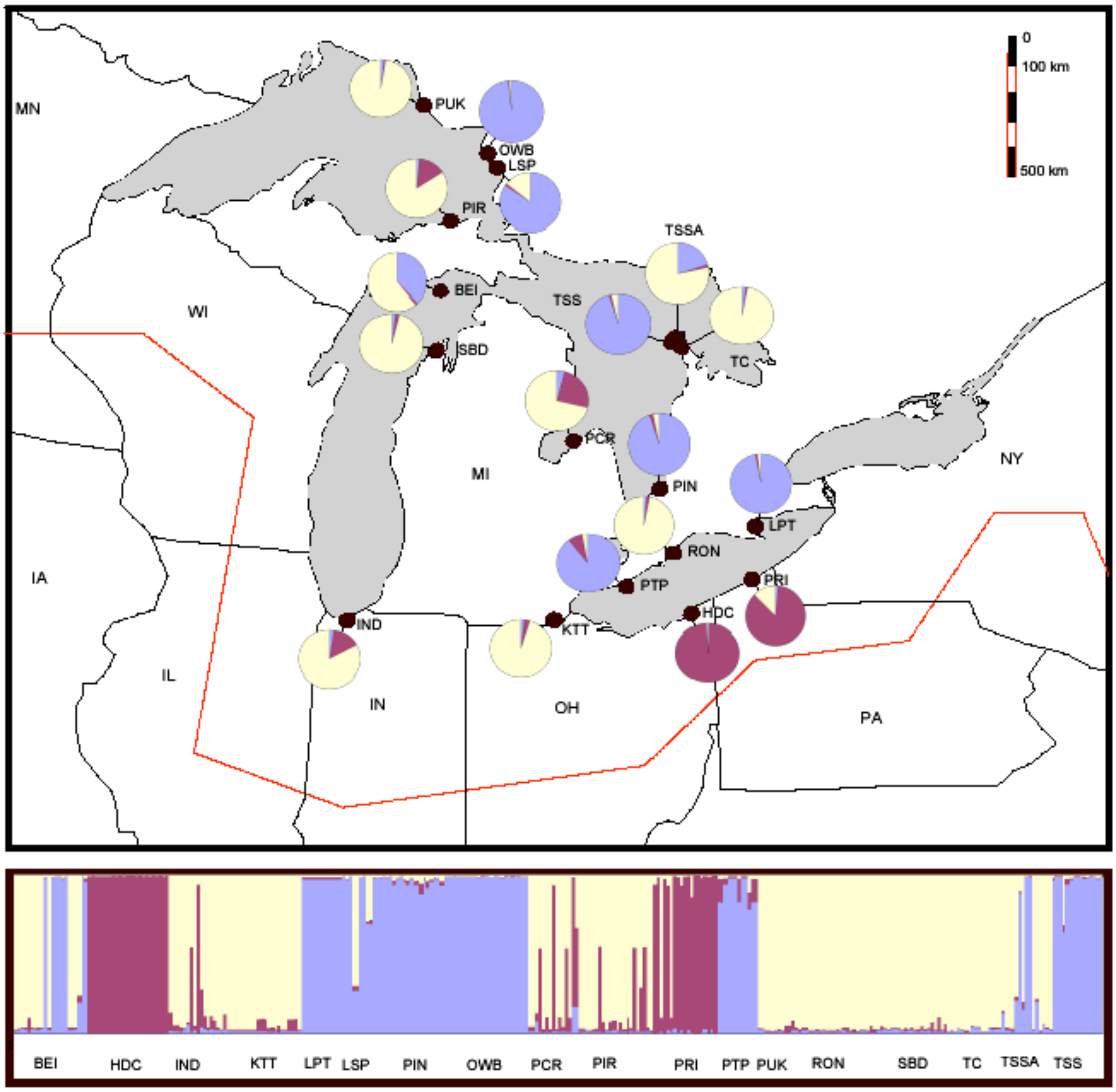


Figure 4-4b Posterior probability of Bayesian clustering analysis (STRUCTURE) using 12 allozyme loci, based on a prior of three clusters (*K* = 3). Bar plots show individual posterior probabilities, and pie charts on the map show mean posterior probabilities for each population.

5 General discussion

5.1 The colonization history of *Arabidopsis lyrata* in eastern North America

Previous work on *A. lyrata* in the Great Lakes region of eastern North America has suggested various levels of population structure, based on analysis of micro-satellite and cpDNA markers (Mable *et al* 2003, 2005; Mable & Adam 2007; Hoebe *et al* 2009; Foxe *et al* 2010). While each of these studies added to the understanding of population structure, and putative post-glacial colonization history of the area, the number of populations sampled, and the geographic distributions of these populations had restricted the conclusions that could be drawn from the data. By extending the geographic sampling area, and the number of populations sampled, along with including samples from herbarium specimens from non-glaciated regions of eastern North America, my study has allowed further insights into the post-glacial history of *A. lyrata* in the Great Lakes region.

Of the three marker systems used in this study, the chloroplast DNA (cpDNA) marker (*trn*F) was screened for in the largest number of populations, and included herbarium samples from locations which were below the area covered by the Laurentide ice sheet at the last glacial maximum (in Wisconsin and Minnesota). This area, covering much of Wisconsin, and the eastern corner of Minnesota has not previously been sampled. Although it may potentially represent a glacial refugium near the boundary of the Laurentide ice sheet (Soltis *et al* 2006), which may have relevance for the re-colonization of de-glaciated areas, after the ice began to retreat c.18 000 ybp. The predominant cpDNA haplotype in this region (L6; Fig. 2b-3) is likely ancestral to all other cpDNA haplotypes identified in this study. This region may therefore represent a post-glacial colonization route, allowing species to expand their range north along the west shorelines of both Lake Michigan and Lake Superior, as has been suggested for several amphibian species (Masta *et al* 2003; Hoffman & Blouin 2004). Unfortunately, I was unable to screen these samples for the micro-satellite and allozyme marker systems due to insufficient tissue volume and quality, and so this putative colonization route, and the potential status of these populations as ancestral, could not be tested further. Two further putative routes of post-glacial colonization were also identified using the cpDNA marker (Fig 1-3), the first (B) allowing colonization north into Michigan, from either Illinois, Indiana or Ohio, and the second (C) enabling a northward movement from the mid-eastern region of North America, approaching the lakes from the east side, and then west through New York, Pennsylvania and Ontario. Both of these putative colonization routes are supported by the allozyme data presented in this thesis (Fig. 4b-4), as well as by previous studies on *A. lyrata* (Hoebe *et al* 2009) and other plant and animal taxa that recolonized the Great Lakes area after the last glacial maxima (Griffin & Barrett 2004; Hoffman & Blouin 2004; Placyk *et al* 2007). The micro-satellite data generated in this thesis, when used alone, adds no support to any of the three putative colonization routes suggested by the cpDNA and allozyme marker systems, because the high level of variation within and between populations means that inference about relatedness between populations, and thus potential colonization routes is very difficult. Earlier studies have used micro-satellite loci to infer colonization history (Hoebe *et al* 2009; Foxe *et al* 2010), when used in conjunction with other marker systems. Hoebe *et al* (2009) found that on fewer populations (13 in total), patterns of micro-satellite diversity did complement cpDNA haplotype patterns, although they also found micro-satellite diversity was high. Foxe *et al* (2010) used a similar number of populations (24 in total), but combined micro-satellite data with less variable nuclear sequence data, also suggested two putative routes of colonization.

5.2 Marker systems in phylogeography

Phylogeography was originally introduced as a set of methods to detect recently isolated groups of individuals or populations within a given taxa, by superimposing a mitochondrial DNA (mtDNA) gene tree over the geographical area in which they were sampled (Avise *et al* 1987). Phylogeography has subsequently expanded to include estimates of levels and patterns of gene flow, population growth, strength of selection, and times of divergence. Despite the early reliance on mtDNA, it is widely acknowledged that sole reliance on a single marker system could lead to biases associated with the reliance on any single gene tree, and so studies involving multiple loci are preferred.

Unfortunately, in plant taxa, mtDNA represents a less ideal marker system than in animal taxa, with very low levels of within species sequence divergence reported (Palmer & Herbon 1989). For this reason, the chloroplast genome has been widely used for phylogenetic studies. Its highly conserved nature allows not only base pair substitutions, but also structural changes in the form of indels and repeat sequences, to be considered phylogenetically informative (Palmer 1991; Raubson & Jansen 2005). Using larger structural changes (gene duplications and loss of repeat sequences) increased my ability to discriminate population level variation in North American *A. lyrata*; something that was not possible when considering only base pair substitutions in the pre-tandem array portion of the *trn*F sequence (chapter 3). Despite the usefulness of using changes in the tandem array portion of the *trn*F sequence in North America, questions still remain as to the suitability of using such large structural changes for phylogeographic inference in Europe. This is because in Europe, the pre-tandem array portion of the gene sequence is more diverse, which may suggest its use is not required, and there is strong evidence suggesting multiple independent changes in pseudogene copy number (Ansell *et al* 2007; Schmickl *et al* 2008) that make inferring evolutionary relationships between pseudogene copies difficult.

Although allozyme markers have been extensively used in a variety of study types, including phylogeographic studies, there are questions surrounding the suitability of their use for phylogeography. These relate to uncertainty about if the loci are actually neutral (a central requirement for phylogeography), or if they are under selection, which has been demonstrated for loci used in studies of *Drosophila* (Hale and Singh 1991; Krafsur 2002; Kreitman 1996) and various plant and animal species (Karl & Avise 1992; Stuber *et al* 1980). Further difficulties involve scoring banding patterns, which can be imprecise and objective, and the potential for null alleles (Langley *et al* 1981), which could mean that variation is essentially underestimated using this type of marker system. In addition to these problems, observed differences in banding patterns may not be fully explained by variation in the coding sequence (a key assumption in their use). This is because allozymes are protein products, and can be affected by post-translational modification (for example, phosphorylation, methylation or acetylation) that can cause changes in physiochemical properties and structure (Snabel 1995). These post-transitional changes can appear as distinct bands on the starch gel, which may appear to suggest variation when none is present. Despite this array of potential pitfalls, the allozyme loci used in this study provided a picture of population structure similar to that elucidated by the cpDNA markers, which has also been demonstrated in European *A. lyrata* (Ansell *et al* 2010) and *Arabis alpina* (Ansell *et al* 2008). It is for this reason that I think that allozyme markers can still be useful for phylogeographic studies, when used appropriately, and in conjunction with other marker systems. Unfortunately, more sophisticated types of analyses, which have been recently developed, cannot be carried out on allozyme data and so its continued use looks set to decrease.

With the trend towards the abandonment of allozymes for more variable, and easier to use marker systems seen in the recent literature, micro-satellite loci have become very common in studies of phylogeography, and now rival mtDNA markers in popularity. This class of molecular markers has numerous potential advantages, the most obvious of which is that many independent and highly variable loci can be analyzed. While this is generally viewed as an improvement over mtDNA (or cpDNA) surveys that yield a single, uni-parentally inherited gene tree, which is taken to represent lineage history, there are drawbacks with the use of micro-satellite data in phylogeography (Brito and Edwards 2008). These include many of the problems that resulted in the large-scale abandonment of allozyme systems. In particular, micro-satellite systems have faced scrutiny due to the potential for size homoplasy (where different copies of a locus are identical by state, but not identical by descent due to convergent mutation), which can be a problem when using models of stepwise mutation (reviewed in Estoup & Angers 1998; Ellegren 2000a; Schlötterer 2000).

The micro-satellite markers used in this study proved to be relatively diverse, although this seemed to be amplified by relatively little shared variation, which effectively presents each population as a single unit. As such, this represented a ‘snap-shot’ of population structure that does not allow meaningful inference regarding post-glacial colonization routes in North America. This is because each population is essentially differentiated from each other populations, and because of the way micro-satellites evolve, there is no way of assessing relatedness. While the same micro-satellite loci have been used for population level analysis previously in *A. lyrata* (Mable *et al* 2003, 2005; Mable & Adam 2007; Hoebe *et al* 2009) it is my opinion that they are not particularly useful for inference about colonization history in *A. lyrata* when used independently of other marker systems. However, they may prove useful when combined with other, less variable markers, as demonstrated by Foxe *et al* (2010) with nuclear gene markers.

The field of phylogeography is currently moving away from 'allele frequency' based systems (like micro-satellites and allozymes) and towards the use of nuclear gene sequences (Lee and Edwards 2008). This is because the latter allows the estimation of confidence limits on parameters such as times of divergence, patterns of population expansion, the effects of natural selection, and levels of gene flow (Brito and Edwards 2008). Estimation of confidence limits for these parameters increases the users ability to make meaningful inferences about the life history of a species, which is a key goal of phylogeographic studies. This does not mean that micro-satellite markers are obsolete, far from it, they will continue to be useful for the assessment of parentage, patterns of heterozygosity, and localized patterns of gene flow based on the distribution of rare alleles (Zink 2010).

5.3 Broader implications

5.3.1 Mating system variation in Arabis alpina

Arabis alpina has the potential to be a useful model system for studying the ecological genetics of alpine adaptation, based on a variety of factors including an extensive arctic-alpine latitudinal and altitudinal range, variation in leaf shape and flowering time variation (Wang *et al* 2009, Poncet *et al* 2010). The genome sequencing project for *A. alpina* is also almost complete (Coupland and Weigel, unpublished). Its global colonization history is well characterized (Koch *et al* 2006), with further in-depth analysis in Europe (Ansell *et al* 2008) revealing putative evidence of mating system variation in the Alps. This variation may have implications for the conclusions drawn about local post-glacial colonization history, particularly in relation to patterns of genetic diversity, which is likely has the potential to be influenced by mating system.

The identification of a functional self-incompatibility system and the characterization of genes involved in SI in *A. alpina* add weight to the theory that SI is the ancestral state in the Brassicaceae (Paetsch *et al* 2006; Guo *et al* 2009). Further to this variation in the strength of SI in *A. alpina* has real implications for the further analysis of the evolution of loss of SI in the Brassicaceae. While population-level variation in mating system is seen extensively throughout the Brassicaceae and has been reported in many genera, including *Arabidopsis* (Foxe *et al* 2010; Mable *et al* 2005; Mable and Adam 2007), *Brassica* (Hinata *et al* 1995), *Capsella* (Paetsch *et al* 2006), *Leavenwortia* (Koelling *et al* 2010) and *Raphanus* (Okamoto *et al* 2004), SI loss in the more distantly related genus *Arabis* has not been studied, and may provide further evidence about mechanisms of loss, or factors which influence loss by known mechanisms in sporophytic systems.

5.3.2 *The importance of linking putative* SRK *genotypes to mating system phenotypes*

SRK has received much attention in the Brassicaceae due to its function in the SI response. Correspondingly, its sequence structure is well understood, and high sequence similarity is observed between species as a result of the *S*-locus being under intense selection pressure. This is the result of balancing selection, by which alleles are maintained in the population at frequencies above those expected by mutation alone (Charelsworth 2006). However, identification of sequences that demonstrate high sequence similarity to known *SRK* alleles is not enough evidence to confirm that you have isolated an *SRK* allele. High sequence similarity can also be seen with other alleles including the commonly amplified *Aly*8 and *Aly*9 (Charlesworth *et al* 2003). These genes may (Suzuki *et al* 1997, 1999; Kusaba *et al* 2000) or may not (Kai *et al* 2001) be linked to the *S*-locus, but have sufficient sequence similarity to amplify when using relatively general primer sets (demonstrated by the number of non-*SRK* alleles that cluster with know *SRK* alleles from various species; Fig. 3-2). To confirm that an identified sequence is *SRK* (although this also applies for *SCR*, the male determinant of self-incompatibility), it should also exhibit co-segregation suggestive of a single Mendelian locus, as inferred from diallel crosses within families. This has been well demonstrated in systems such as *A. lyrata* (Prigoda *et al* 2005; Charlesworth *et al* 2003) and *A. halleri* (Castric and Vekemans 2008), but less well demonstrated in *C. grandiflora* (Paetsch *et al* 2006).

5.3.3 Phylogeography in eastern North America and beyond

In previously glaciated regions of eastern North America, in particular the Great Lakes region, phylogeography has been extensively studied for amphibian taxa (Austin *et al* 2002; Carstens *et al* 2004; Church *et al* 2003; Demastes *et al* 2007; Hoffman and Blouin 2004; Nielson *et al* 2001; Steele and Storfer 2006; Wagner *et al* 2005; Zamudio and Savage 2003) with studies of other species, particularly plant species, being far less numerous. Despite the apparent disparity in the volume of studies between different taxa, and the variation in methods and speed of colonization, along with barriers to colonization that could present problems to different species, there does appear to be congruence in the phylogeographic patterns observed between *A. lyrata* used in this study, and other plant and animal taxa. It may be important to note however, that these broad, general phylogeographic patterns may be the result of different underlying causal factors (for example species specific barriers, competition with other colonizers, and speed of colonization) acting at different times, as comparative analysis of European species has shown highly individualistic responses of different species to climate change (Taberlet *et al* 1998).

Despite the relatively small geographical area studied for *A. lyrata* in North America, it is possible to make some inferences about potential refugial areas. Most notably, I find it likely that plant taxa may have survived during Pleistocene glaciation in close proximity to the Laurentide Ice Sheet, particularly in Illinois, Indiana, Wisconsin and Minnesota based on current cpDNA haplotype diversity in this area (Fig 2b-3). Which has also been postulated for and for some other plant and animal taxa (reviewed by Soltis *et al* 2006). Proposed refugial areas should be considered with caution however, as further sampling of *A. lyrata* is required before accurate delimitation of its refugia can be achieved, and perhaps additional examination including a more complete comparative examination of phylogeography in this area is also required.

In broader terms, the phylogeographic patterns of *A. lyrata* generally fit with those proposed based on species studied in a larger geographic area (eastern North America as a whole: reviewed by Soltis *et al* 2006) with particular reference to colonization history. Although it is difficult to transfer my relatively local patterns of population structure into something more meaningful on a larger scale (particularly larger scale patterns of genetic diversity) because of the relatively small sampling area, it may be possible to make predictions about the possible locations of glacial refugia based on those delimited for other species. These predictions would require extensive further sampling to test however. Despite this restricted sampling however, this study does represent the largest population level study of *A. lyrata* in North America, and is potentially very useful for comparison of *A. lyrata* phylogeography and diversity, globally. At this scale, there are marked differences between levels of genetic diversity in European *A. lyrata* and North American *A. lyrata* based on cpDNA evidence (Ansell *et al* 2007; Schmickl *et al* 2010) and also allozyme variation (Ansell *et al* 2010). This is further supported by evidence from nuclear gene marker studies (Ross-Ibarra *et al* 2008). These differences not only support the generally held belief that North American *A. lyrata* is genetically depauperate in comparison to Europe, but also that all variation in North American *A. lyrata* can be explained by colonization from Europe, with populations in Austria a likely origin due based on shared cpDNA haplotypes (taken from Ansell *et al* 2007) and allozyme variation (taken from Ansell *et al* 2010).

5.4 Future work

5.4.1 Mating system variation, loss of SI and dominance of **SRK** *alleles in* **Arabis alpina**

Six European populations, chosen on the basis of variation in inbreeding coefficient (*F_{IS}*) that suggested variation in mating system (Ansell *et al* 2008), were used to identify evidence of a functional SI system in *A. alpina*. Along with a clear barrier to self-fertilization in three populations, evidence of *SRK* alleles were found in all populations regardless of their current mating strategy (outcrossing or inbreeding) suggesting that SI may have been recently lost. Variation in mating system at the population level may have directly influenced processes relating to colonization, for example loss of SI can result in reproductive assurance, which increases the ability of individuals to colonize (Baker 1955) and so can be very important when studying the phylogeography of a given species. Global phylogeography for *A. alpina* has been well examined (Ehrich *et al* 2007), with particular reference to the three colonization routes out of Asia-minor that resulted in the present day range of the species (Koch *et al* 2006). On a finer scale, phylogeography in the Italian peninsula and the Alps has been well reconstructed (Ansell *et al* 2008). All of these studies however, may further benefit from greater understanding of how variation in mating system has affected the current range of this species, genetic diversity, and the processes that drove the colonization from Asia minor. For example, loss of SI aids the success of long-distance dispersal events, but in turn reduces genetic diversity. Conversely, maintenance of a functional SI system mediates shorter dispersal events (because of the requirement of a second, non-related partner for sexual reproduction), but maintain high levels of genetic diversity.

Although I have demonstrated firm evidence of the existence of an SI system in *A. alpina*, using only six populations was not enough to infer much about mechanisms of SI loss in the three populations that were functionally self-compatible. However, there were differences in the SI status of these populations, (two populations showed a reduced or absent barrier to self-fertilization but were not autonomous, the third population was autonomous self-fertilizing) which coincide with areas suspected to differ by colonization history (Ansell *et al* 2008). While it is currently premature to conclude the extent and context to which phylogeographic history has influenced variation in mating system (or visa versa), it would be possible to test hypotheses of multiple occurrence of SI breakdown in this species. This could be done by testing the strength of SI in relation to population level variation on a larger scale, as has been done with *A. lyrata* in North America (Mable & Adam 2007; Hoebe *et al* 2009; Foxe *et al* 2010). Specifically I would test further populations in areas which are suspected to differ in colonization history as delimited by Ansell *et al* (2008), particularly in the central Alps (where the autonomous population was located) and in the western Alps (where the two non-autonomous, but SC populations were located).

In other Brassicaceae species that employ the same type of sporophytic SI system, the number of identified *SRK* alleles is relatively high (>40 each for *B. campestris, B. oleracea, B. rapa, A. lyrata,* and *A. halleri*) as would be expected for a self-recognition system of this type. Based on these numbers, it is perhaps reasonable to expect similar numbers of alleles in *A. alpina*, although the number of alleles maintained in an SSI system, will depend on a number of factors, including the dominance relations, the frequencies of different dominance classes in populations and the population structure (Schierup *et al* 1998; Vekemans *et al* 1998). While the goal of this study was not to delimit evidence of dominance between alleles, or classify alleles into dominance classes, this would allow further understanding of the self-incompatibility reaction in this species. This could be achieved by performing multiple pairwise crosses between heterozygotes that share single alleles. Dominant alleles are expected to cause an SI reaction in any pair where they are shared, while less dominant alleles may allow production of seed set when in combination with and non-shared alleles. This could be scored by either seed production (Mable *et al* 2003) or pollen tube activity (Hatakeyama *et al* 1998).

5.4.2 Identification of A. lyrata *glacial refugia*

While the geographic area sampled in this study is relatively large, with an appropriate sample site density, I have only relatively few populations from non-glaciated areas. This makes accurate predictions about glacial refugia, and colonization routes difficult. I can of course speculate as to the likely location of refugia based on other species (Holman 1995; Soltis *et al* 2006; Lee *et al* 2008), but further sampling particularly in regions south of the Great Lakes in the mid-east and the Appalachians, where putative glacial refugia have been suggested is essential. This should allow the identification of potential glacial refugia relevant to this species, and further clarify colonization history as has been done more extensively in other taxa (Holman 1995; Placyk *et al* 2007; Soltis *et al* 2006).

Appendices

Appendix 1– Genbank accession numbers for all samples used in this study.

Species	Allele name	Genbank number	Reference
A. lyrata	ALy8 (ARK3)	XM_002867806	Bakker *et al*
A. halleri	AHASRK01	EU075124	Castric & Vekemens 2007
A. halleri	AHASRK02	EU075125	Castric & Vekemens 2007
A. halleri	AHASRK03	EU075126	Castric & Vekemens 2007
A. halleri	AHASRK05	EU075127	Castric & Vekemens 2007
A. halleri	AHASRK06	EU075128	Castric & Vekemens 2007
A. halleri	AHASRK07	EU075129	Castric & Vekemens 2007
A. halleri	AHASRK08	EU075130	Castric & Vekemens 2007
A. halleri	AHASRK09	EU075131	Castric & Vekemens 2007
A. halleri	AHASRK11	EU075132	Castric & Vekemens 2007
A. halleri	AHASRK12	EU075133	Castric & Vekemens 2007
A. halleri	AHASRK13	EU075134	Castric & Vekemens 2007
A. halleri	AHASRK14	EU075135	Castric & Vekemens 2007
A. halleri	AHASRK15	EU075136	Castric & Vekemens 2007
A. halleri	AHASRK16	EU075137	Castric & Vekemens 2007
A. halleri	AHASRK17	EU075138	Castric & Vekemens 2007
A. halleri	AHASRK18	EU075139	Castric & Vekemens 2007
A. halleri	AHASRK19	EU075140	Castric & Vekemens 2007
A. halleri	AHASRK20	EU075141	Castric & Vekemens 2007
A. halleri	AHASRK21	EU075142	Castric & Vekemens 2007
A. halleri	AHASRK22	EU075143	Castric & Vekemens 2007
A. halleri	AHASRK23	EU878008	Castric *et al* 2008
A. halleri	AHASRK24	EU878009	Castric *et al* 2008
A. halleri	AHASRK25	EU878010	Castric *et al* 2008
A. halleri	AHASRK26	EU878011	Castric *et al* 2008
A. halleri	AHASRK27	EU878012	Castric *et al* 2008
A. halleri	AHASRK28	EU878013	Castric *et al* 2008
A. halleri	AHASRK29	EU878014	Castric *et al* 2008
A. halleri	AHASRK30	EU878015	Castric *et al* 2008
A. lyrata	ALYSRK01	AF328996	Schierup *et al* 2001
A. lyrata	ALYSRK02	AY186763	Charelworth *et al* 2003
A. lyrata	ALYSRK03	AF328992	Schierup *et al* 2001
A. lyrata	ALYSRK04	AF328994	Schierup *et al* 2001
A. lyrata	ALYSRK05	AF328990	Schierup *et al* 2001
A. lyrata	ALYSRK06	DQ520281	Bechsgaard *et al* 2006
A. lyrata	ALYSRK08	DQ520282	Bechsgaard *et al* 2006
A. lyrata	ALYSRK09	AF328999	Schierup *et al* 2001
A. lyrata	ALYSRK11	AY186767	Charelworth *et al* 2003
A. lyrata	ALYSRK12	AY186768	Charelworth *et al* 2003
A. lyrata	ALYSRK13	AF328993	Schierup *et al* 2001

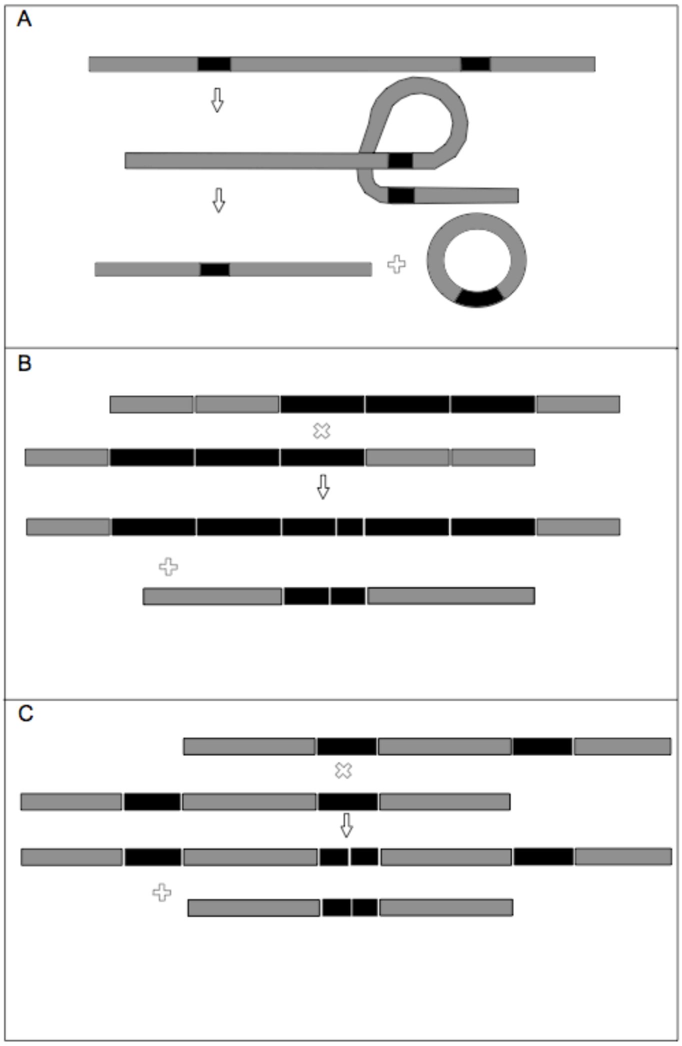
Appendix 1. ***Continued***

A. lyrata	ALYSRK13	AF328993	Schierup *et al* 2001
A. lyrata	ALYSRK13	AY186769	Charelworth *et al* 2003
A. lyrata	ALYSRK14	AY186770	Charelworth *et al* 2003
A. lyrata	ALYSRK15	AY186771	Charelworth *et al* 2003
A. lyrata	ALYSRK16	AF328991	Schierup *et al* 2001
A. lyrata	ALYSRK16	DQ520283	Bechsgaard *et al* 2006
A. lyrata	ALYSRK17	AY186772	Charelworth *et al* 2003
A. lyrata	ALYSRK18	DQ520284	Bechsgaard *et al* 2006
A. lyrata	ALYSRK19	AF328998	Schierup *et al* 2001
A. lyrata	ALYSRK20	AF328995	Schierup *et al* 2001
A. lyrata	ALYSRK21	EU878016	Castric *et al* 2008
A. lyrata	ALYSRK22	AF329000	Schierup *et al* 2001
A. lyrata	ALYSRK23	AF328997	Schierup *et al* 2001
A. lyrata	ALYSRK25	DQ520286	Bechsgaard *et al* 2006
A. lyrata	ALYSRK26	AY186774	Charelworth *et al* 2003
A. lyrata	ALYSRK27	EU878017	Castric *et al* 2008
A. lyrata	ALYSRK28	AY186775	Charelworth *et al* 2003
A. lyrata	ALYSRK29	AY186776	Charelworth *et al* 2003
A. lyrata	ALYSRK30	EU878018	Castric *et al* 2008
A. lyrata	ALYSRK31	DQ520287	Bechsgaard *et al* 2006
A. lyrata	ALYSRK33	EU878019	Castric *et al* 2008
A. lyrata	ALYSRK34	EU878020	Castric *et al* 2008
A. lyrata	ALYSRK35	EU878021	Castric *et al* 2008
A. lyrata	ALYSRK36	DQ520288	Bechsgaard *et al* 2006
A. lyrata	ALYSRK37	DQ520289	Bechsgaard *et al* 2006
A. lyrata	ALYSRK38	EU165341	Schierup *et al* 2008
A. lyrata	ALYSRK39	EU878022	Castric *et al* 2008
A. lyrata	ALYSRK42	EU878023	Castric *et al* 2008
A. lyrata	ALYSRK43	EU878024	Castric *et al* 2008
A. lyrata	ALYSRK44	EU878025	Castric *et al* 2008
A. lyrata	ALYSRK45	EU878026	Castric *et al* 2008
B. oleracea	BOLSRK05	Y18259	Cabrillac *et al* 1999
B. oleracea	BOLSRK02	AB024416	Kusaba *et al* 2000
B. oleracea	BOLSRK13	AB024420	Kusaba *et al* 2000
B. oleracea	BOLSRK60	AB032474	Suzuki *et al* 2000
B. oleracea	BOLSRK18	AB032473	Suzuki *et al* 2000
B. oleracea	BOLSRK23	AB013720	Kusaba & Nishio 1999
B. oleracea	BOLSRK03	X79432	Delorme *et al* 1995
B. oleracea	BOLSRK29	Z30211	Kumar & Trick 1994
B. oleracea	BOLSRK46	AB054747	Sato *et al* 2002
B. oleracea	BOLSRK45	AB054719	Sato *et al* 2002

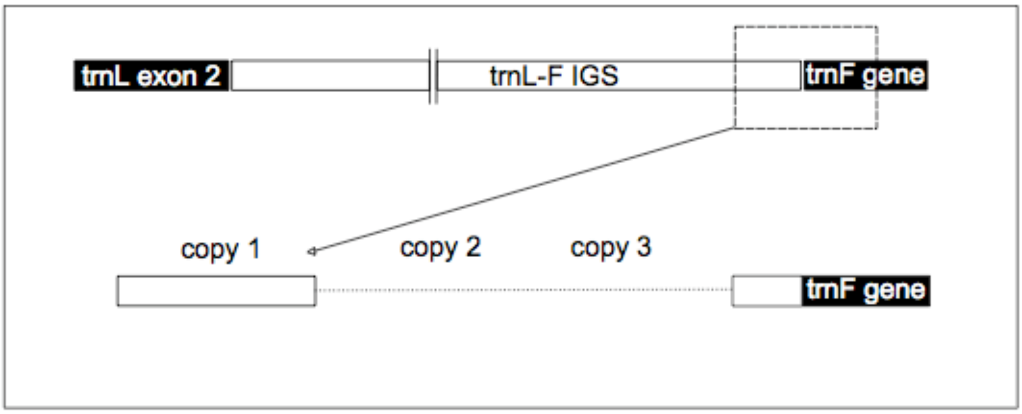
Appendix 1. *Continued*

B. oleracea	BOLSRK11	AB054709	Sato *et al* 2002
B. rapa	BRASRK40	AB201306	Sato *et al* 2006
B. rapa	BRASRK44	AB201307	Sato *et al* 2006
B. rapa	BRASRK29	AB008191	Hatakeyama *et al* 1998
R. sativus	RSASRK9	AB114851	Okamoto *et al* 2004
L. alabamica	LALSRK	EU394519	Busch *et al* 2008
C. grandiflora	CGRSRK	FJ649956	Guo *et al* 2009
C. grandiflora	CGRSRK7	EF530735	Nasrallah *et al* 2007
C. rubella	CRUSRK26	FJ649926	Guo *et al* 2009
C. rubella	CRU-recombinant	FJ649942	Guo *et al* 2009
C. rubella	CRU-pseudogene	FJ649937	Guo *et al* 2009

Appendix 2– (Adapted from Dobeš *et al* 2007). A graphical representation illustrating the main types of mutation mechanisms, potentially responsible for DNA length variation, as discussed in the text. Grey and black lines represent double stranded DNA. Direct (oriented in the same direction) DNA repeats are given in solid black. The 'x' symbol marks crossing over points. A) Intramolecular recombination between two similar regions on a single double-stranded DNA (dsDNA) molecule can result in the excision of the intermediate region, yielding a shorter dsDNA molecule and a separate circularised dsDNA molecule. B) Unequal intermolecular recombination: Recombination by crossing over between repeats of different sequential position within tandem repeat, occurring: B) within a series of direct repeats separated by intervening DNA fragments, or C) the increase in size of one DNA fragment at the expense of the other.



Appendix 3 - (Adapted from DobeŠ *et al* 2007). A graphical representation of the *trn*L-F IGS and the *trn*F (UAA) gene, detailing the region sequenced in this study.



Appendix 4 – Genbank accession numbers for all European samples used in this study.

Sequence Name	Accession number	Author
PET 1A	DQ989814	Ansell *et al* 2007
PET IB	DQ989815	Ansell *et al* 2007
PET 1C	DQ989816	Ansell *et al* 2007
PET 2	DQ989817	Ansell *et al* 2007
PET 3	DQ989818	Ansell *et al* 2007
PET 4	DQ989819	Ansell *et al* 2007
PET 5	DQ989820	Ansell *et al* 2007
PET 6	DQ989821	Ansell *et al* 2007
PET 1	DQ989822	Ansell *et al* 2007
PET 8	DQ989823	Ansell *et al* 2007
PET 9	DQ989824	Ansell *et al* 2007
PET 10	DQ989825	Ansell *et al* 2007
PET 11	DQ989826	Ansell *et al* 2007
PET 12	DQ989827	Ansell *et al* 2007
PET 13	DQ989828	Ansell *et al* 2007
PET 14	DQ989829	Ansell *et al* 2007
PET 15	DQ989830	Ansell *et al* 2007
PET 16	DQ989831	Ansell *et al* 2007
PET 17	DQ989832	Ansell *et al* 2007
PET 18	DQ989833	Ansell *et al* 2007
PET 19A	DQ989834	Ansell *et al* 2007
PET 19B	DQ989835	Ansell *et al* 2007
PET 20	DQ989836	Ansell *et al* 2007
PET 21A	DQ989837	Ansell *et al* 2007
PET 21B	DQ989838	Ansell *et al* 2007
PET 21C	DQ989839	Ansell *et al* 2007
PET 21D	DQ989840	Ansell *et al* 2007
PET 21E	DQ989841	Ansell *et al* 2007
PET 22	DQ989842	Ansell *et al* 2007
PET 23	DQ989843	Ansell *et al* 2007
PET 24	DQ989844	Ansell *et al* 2007
PET 25	DQ989845	Ansell *et al* 2007
PET 26	DQ989846	Ansell *et al* 2007
PET 27	DQ989847	Ansell *et al* 2007
PET 28	DQ989848	Ansell *et al* 2007
PET 29	DQ989849	Ansell *et al* 2007
PET 30	DQ989850	Ansell *et al* 2007
PET 31A	DQ989851	Ansell *et al* 2007
PET 31B	DQ989852	Ansell *et al* 2007
PET 31C	DQ989853	Ansell *et al* 2007
PET 31D	DQ989854	Ansell *et al* 2007
PET 31E	DQ989855	Ansell *et al* 2007
PET 31F	DQ989856	Ansell *et al* 2007
PET 32	DQ989857	Ansell *et al* 2007
PET 33	DQ989858	Ansell *et al* 2007
PET 34	DQ989859	Ansell *et al* 2007

Appendix 5 – Chloroplast *trn*F haplotypes assigned by Tedder *et al* (2010).

Population	Number of individuals	Haplotype 1 (Number)	Haplotype 2 (Number)	Lineage grouping
PTP	8	L1	-	B; Red
LPT	8	L1	-	B; Red
RON	8	L1	-	B; Red
KTT	8	L3	-	B; Red
TSSA	8	S1(4)	L1(4)	A; Blue / B; Red
TC	8	S1	-	A; Blue
OWB	8	S1	-	A; Blue
HDC	8	L4	-	C; Green
PIN	8	L1	-	B; Red
PIR	8	L1	-	B; Red
PRI	8	L4	-	C; Green
TSS	8	S1	-	A; Blue
LSP	8	S1	-	A; Blue
SBD	8	S3	-	C; Green
PUK	8	S1	-	A; Blue
BEI	8	S1	-	A; Blue
PCR	8	L2	-	A; Blue
IND	8	L1(4)	L2(4)	B; Red / A; Blue

Appendix 6 – Population groupings used for AMOVA analysis. Scenarios refer to: (1) Lakefront of origin, mating system based on T_m (2), cpDNA lineage (3) and previously defined clusters (Foxe *et al* 2010) (4).

Scenario	Group 1	Group 2	Group 3	Group 4	Group 5
1	PUK OWB LSP PIR	BEI SBD IND	PCR PIN TSS TSSA TC	PTP RON LPT KTT HDC PRI	
2	PTP LPT RON KTT TSSA TC	OWB HDC PIN PIR PRI TSS LSP SBD PUK BEI PCR IND			
3*	HDC PRI SBD	PUK OWB LSP BEI TSSA TC	LPT RON PTP PIR PIN KTT		
4‡	KTT PIR	LPT HDC PRI	PTP RON PCR PIN	BEI LSP OWB PUK SBD	TC TSSA TSS

* Populations IND and TSS not included due to mixed chloroplast lineages.

‡ IND not included due to admixture, making assigning a group inappropriate.

List of References

Abbott RJ, Gomes MF (1989). Population Genetic-Structure and Outcrossing Rate of *Arabidopsis thaliana* (L) Heynh. Heredity 62: 411-418.

Al-Shehbaz IA, O'Kane SL (2002). Taxonomy and phylogeny of Arabidopsis (Brassicaceae). In: The Arabidopsis Book. (Somerville CR, Meyerowitz EM Eds) American Society of Plant Biologist: Rockville, Maryland. 1-22.

Allen AM, Hiscock SJ (2008). Evolution and phylogeny of self-incompatibility systems in Angiosperms. In V E Franklin-Tong (Ed) Self-Incompatibility in Flowering Plants, (pp 73-101), Springer-Verlag Berlin

Anmarkrud JA, Kleven O, Bachmann L, Lifjeld JT (2008). Microsatellite evolution: Mutations, sequence variation, and homoplasy in the hypervariable avian microsatellite locus HrU10. Bmc Evolutionary Biology 8:

Ansell SW, Grundmann M, Russell SJ, Schneider H, Vogel JC (2008). Genetic discontinuity, breeding-system change and population history of *Arabis alpina* in the Italian Peninsula and adjacent Alps. Molecular Ecology 17: 2245-2257.

Ansell SW, Schneider H, Pedersen N, Grundmann M, Russell SJ, Vogel JC (2007). Recombination diversifies chloroplast trnF pseudogenes in *Arabidopsis lyrata*. Journal of Evolutionary Biology 20: 2400-2411.

Ansell SW, Stenoien HK, Grundmann M, Schneider H, Hemp A, Bauer N (2010). Population structure and historical biogeography of European *Arabidopsis lyrata*. Heredity 105: 543-553

Assefa A, Ehrich D, Taberlet P, Nemomissa S, Brochmann C (2007). Pleistocene colonization of afro-alpine 'sky islands' by the arctic-alpine *Arabis alpina*. Heredity 99: 133-142.

Austin JD, Lougheed SC, Neidrauer L, Chek AA, Boag PT (2002). Cryptic lineages in a small frog: the post-glacial history of the spring peeper, *Pseudacris crucifer* (Anura : Hylidae). Molecular Phylogenetics and Evolution 25: 316-329.

Avise JC (2000). Phylogeography: The History and Formation of Species, Harvard University Press, Cambridge.

Avise JC (1986). Mitochondrial-DNA and the evolutionary genetics of higher animals. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 312: 325-342.

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

Bailey MF, McCauley DE (2006). The effects of inbreeding, outbreeding and long-distance gene flow on survivorship in North American populations of *Silene vulgaris*. Journal of Ecology 94: 98-109.

Baker HG (1955). Self-compatibility and establishment after long distance dispersal. Evolution 9: 347-349.

Baker HG (1966). Evolution functioning and breakdown of heteromorphic incompatibility systems .I. Plumbaginaceae. Evolution 20: 349

Baker HG (1967). Support for Baker's law as a rule. Evolution 21: 853-856.

Balañá-Alcaide D, Ramos-Onsins SE, Boone Q, Aguade M (2006). Highly structured nucleotide variation within and among *Arabidopsis lyrata* populations at the FAH1 and DFR gene regions. Molecular Ecology 15: 2059-2068.

Barrett SCH (1988). The evolution, maintenance, and loss of self-incompatibility systems. In Plant Reproductive Ecology: Patterns and Strategies, pp 98-124 Oxford: Oxford University Press.

Barrett SCH, Charlesworth D (1991). Effects of a change in the level of inbreeding on the genetic load. Nature 352: 522-524.

Bateman AJ (1951). A General Theory of Self-Incompatibility. Heredity 5: 157.

Bateman AJ (1954). Self-incompatibility systems in Angiosperms .2. Iberis-Amara. Heredity 8: 305-332.

Bateman AJ (1955). Self-incompatibility in angiosperms. 3. Cruciferae. Heredity 9: 53-68.

Beattie AJ, Culver DC (1979). Neighborhood size in *Viola*. Evolution 33: 1226-1229.

Bechsgaard JS, Castric V, Vekemans X, Schierup MH, Schierup MH (2006). The transition to self-compatibility in *Arabidopsis thaliana* and evolution within *S*-haplotypes over 10 Myr. Molecular Biology and Evolution 23: 1741-1750.

Beilstein MA, Al-Shehbaz IA, Kellogg EA (2006). Brassicaceae phylogeny and trichome evolution. American Journal of Bottany 93: 607-619.

Bernatchez L (2001). The evolutionary history of brown trout (*Salmo trutta* L.) inferred from phylogeographic, nested clade, and mismatch analyses of mitochondrial DNA variation. Evolution 55: 351-379.

Bernatchez L, Wilson CC (1998). Comparative phylogeography of nearctic and palearctic fishes. Molecular Ecology 7: 431-452.

Biasiolo A (1992). Lack of Allozyme Variability among Varroa Mite Populations. Experimental & Applied Acarology 16: 287-294.

Bleeker W, Hurka H (2001). Introgressive hybridization in *Rorippa* (Brassicaceae): gene flow and its consequences in natural and anthropogenic habitats. Molecular Ecology 10: 2013-2022.

Bohonak AJ (2002). IBD (isolation by distance): A program for analyses of isolation by distance. Journal of Heredity 93: 153-154.

Brewbaker JL (1957). Pollen cytology and self-incompatibility systems in plants. Heredity 48: 271-277.

Brewbaker JL (1959). Biology of the angiosperm pollen grain. Indian Journal of Genetics and Plant Breeding 19: 121-124.

Brito P, Edwards SV (2009). Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica 135: 439-455.

Brunsfeld CW, Sullivan J, Soltis PS (2001). Compatative phylogeography of northwestern North America: A synthesis. Blackwell Science, London, UK.

Busch JW, Schoen DJ (2008). The evolution of self-incompatibility when mates are limiting. Trends in Plant Science 13: 128-136.

Busch JW, Sharma J, Schoen DJ (2008). Molecular characterization of Lal2, an SRK-like gene linked to the S-locus in the wild mustard *Leavenworthia alabamica*. Genetics 178: 2055-2067.

Byers DL, Meagher TR (1992). Mate availability in small populations of plant species with homomorphic sporophytic self-incompatibility. Heredity 68: 353-359.

Byers DL, Waller DM (1999). Do plant populations purge their genetic load? Effects of population size and mating history on inbreeding depression. Annual Review of Ecology and Systematics 30: 479-513.

Cabrillac D, Cock JM, Dumas C, Gaude T (2001). The S-locus receptor kinase is inhibited by thioredoxins and activated by pollen coat proteins. Nature 410: 220-223.

Cain ML, Milligan BG, Strand AE (2000). Long-distance seed dispersal in plant populations. American Journal of Botany 87: 1217-1227.

Carr DE, Dudash MR (2003). Recent approaches into the genetic basis of inbreeding depression in plants. Philosophical Transactions of the Royal Society of London, Series B 358: 1071-2889.

Carstens BC, Stevenson AL, Degenhardt JD, Sullivan J (2004). Testing nested phylogenetic and phylogeographic hypotheses in the Plethodon vandykei species group. Systematic Biology 53: 781-792.

Castric V, Bechsgaard J, Schierup MH, Vekemans X (2008). Repeated Adaptive Introgression at a Gene under Multiallelic Balancing Selection. Plos Genetics 4(8):

Castric V, Vekemans X (2008). Evolution under strong balancing selection: how many codons determine specificity at the female self-incompatibility gene SRK in Brassicaceae? BMC Evolutionary Biology 7:

Chapman LA, Goring DR (2010). Pollen-pistil interactions regulating successful fertilization in the Brassicaceae. Journal of Experimental Botany 61: 1987-1999.

Charlesworth B, Charlesworth D (1999). The genetic basis of inbreeding depression. Genetical Research 74: 329-340.

Charlesworth B, Morgan MT, Charlesworth D (1991). Multilocus models of inbreeding depression with synergistic and partial self-fertilization. Genetical Research 57: 177-194.

Charlesworth D (2003). Effects of inbreeding on the genetic diversity of populations. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 358: 1051-1070.

Charlesworth D (2006). Evolution of plant breeding systems. Current Biology 16: 726-735.

Charlesworth D (2006). Balancing selection and its effects on sequences in nearby genome regions. Plos Genetics 2: 379-384.

Charlesworth D, Awadalla P, Mable BK, Schierup MH (2000). Population-level studies of multiallelic self-incompatibility loci, with particular reference to Brassicaceae. Annals of Botany 85: 227-239.

Charlesworth D, Bartolome C, Schierup MH, Mable BK (2003). Haplotype structure of the stigmatic self-incompatibility gene in natural populations of *Arabidopsis lyrata*. Molecular Biology and Evolution 20: 1741-1753.

Charlesworth D, Charlesworth B (1979). Evolution and breakdown of S-allele systems. Heredity 43: 41-55.

Charlesworth D, Charlesworth B (1979). Evolutionary genetics of sexual systems in flowering plants. Proceedings of the Royal Society of London, Series B 205: 513-530.

Charlesworth D, Charlesworth B (1987). Inbreeding depression and its evolutionary consequenses. Annual Review of Ecology and Systematics 18: 237-268.

Charlesworth D, Charlesworth B (1990). Inbreeding depression with heterozygote advantage and its effect on selection for modifiers changing the outcrossing rate. Evolution 44: 870-888.

Charlesworth D, Kamau E, Hagenblad J, Tang CL (2006). Trans-specificity at loci near the self-incompatibility loci in *Arabidopsis*. Genetics 172(4): 2699-2704.

Charlesworth D, Mable BK, Schierup MH, Bartolome C, Awadalla P (2003). Diversity and linkage of genes in the self-incompatibility gene family in *Arabidopsis lyrata*. Genetics 164: 1519-1535.

Charlesworth D, Morgan MT, Charlesworth B (1992). The effect of linkage and population size on inbreeding depression due to mutational load. Genetical Research 59: 49-61.

Charlesworth D, Vekemans X, Castric V, Glemin S (2005). Plant self-incompatibility systems: a molecular evolutionary perspective. New Phytologist 168: 61-69.

Charlesworth D, Wright SI (2001). Breedings system and genome evolution. Current opinion in Genetics & Development 11: 685-690.

Church SA, Kraus JM, Mitchell JC, Church DR, Taylor DR (2003). Evidence for multiple pleistocene refugia in the postglacial expansion of the eastern tiger salamander, *Ambystoma tigrinum tigrinum*. Evolution 57: 372-383.

Ciofi C, Beaumont MA, Swingland IR, Bruford MW (1999). Genetic divergence and units for conservation in the Komodo dragon *Varanus komodoensis*. Proceedings of the Royal Society of London Series B-Biological Sciences 266: 2269-2274.

Clark JS, Fastie C, Hurtt G, Jackson ST, Johnson C, King GA (1998). Reid's paradox of rapid plant migration - Dispersal theory and interpretation of paleoecological records. Bioscience 48: 13-24.

Clauss MJ, Cobban H, Mitchell-Olds T (2002). Cross-species microsatellite markers for elucidating population genetic structure in *Arabidopsis* and *Arabis* (Brassicaeae). Molecular Ecology 11: 591-601.

Clauss MJ, Dietel S, Schubert G, Mitchell-Olds T (2006). Glucosinolate and trichome defenses in a natural *Arabidopsis lyrata* population. Journal of Chemical Ecology 32: 2351-2373.

Clauss MJ, Mitchell-Olds T (2006). Population genetic structure of *Arabidopsis lyrata* in Europe. Molecular Ecology 15: 2753-2766.

Clement M, Posada D, Crandall KA (2000). TCS: a computer program to estimate gene genealogies. Molecular Ecology 9: 1657-1659.

Comes HP, Kadereit JW (2003). Spatial and temporal patterns in the evolution of the flora of the European Alpine System. Taxon 52: 451-462.

Conroy CJ, Cook JA (2000). Phylogeography of a post-glacial colonizer: *Microtus longicaudus* (Rodentia: Muridae). Molecular Ecology 9: 165-175.

Conroy CJ, Cook JA (2000). Molecular systematics of a holarctic rodent (*Microtus* : Muridae). Journal of Mammalogy 81: 344-359.

Coombs JA, Letcher BH, Nislow KH (2008). CREATE: a software to create input files from diploid genotypic data for 52 genetic software programs. Molecular Ecology Resources 8: 578-580.

Cornuet JM, Luikart G (1996). Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics 144: 2001-2014.

Cox CM, P (2000). Biogeography: an ecological and evolutionary approach. Blackwell Science, London, UK.

Curtu AL, Finkeldey R, Gailing O (2004). Comparative sequencing of a microsatellite locus reveals size homoplasy within and between European oak species (*Quercus* spp.). Plant Molecular Biology Reporter 22: 339-346.

Darwin CR (1876). The effects of cross and self-fertilization in the vegetable kingdom. London: Murray.

Davis MB (1983). Quaternary History of Deciduous Forests of Eastern North-America and Europe. Annals of the Missouri Botanical Garden 70: 550-563.

De Nettancourt D (1977). Incompatibility in angiosperms. Springer, Berlin Heidleberg New York.

Demastes JW, Eastman JM, East JS (2007). Phylogeography of the blue-spotted salamander, *Ambystoma laterale* (Caudata : Ambystomatidae). American Midland Naturalist 157: 149-161.

Dingle H (1996). Migration: the biology of life on the move. Oxford University press, New York, USA.

Dobeš C, Kiefer C, Kiefer M, Koch MA (2007). Plastidic *trnF*(UUC) pseudogenes in North American genus *Boechera* (Brassicaceae): Mechanistic aspects of evolution. Plant Biology 9: 502-515.

Dobzhansky T (1948). Genetics of natural populations. XVIII. Experiments on chromosomes of *Drosophila pseudoobscura* from different geographical regions. Genetics 33: 568-602.

Dwyer KG, Kandasamy MK, Mahosky DI, Acciai J, Kudish BI, Miller JE (1994). A Superfamily of S-Locus-Related Sequences in *Arabidopsis* - Diverse Structures and Expression Patterns. Plant Cell 6: 1829-1843.

Ehrich D, Gaudeul M, Assefa A, Koch MA, Mummenhoff K, Nemomissa S (2007). Genetic consequences of Pleistocene range shifts: contrast between the Arctic, the Alps and the East African mountains. Molecular Ecology 16: 2542-2559.

Ellegren H (2000). Microsatellite mutations in the germline: implications for evolutionary inference. Trends in Genetics 16: 551-558.

Estoup A, Angers B (1998). Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations, pp. 55–86 in Advances in Molecular Ecology, edited by G. R. Carvalho. IOS Press, Amsterdam.

Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14: 2611-2620.

Excoffier L, Laval G, Schneider S (2005). Arlequin (version 3.0): An integrated software package for population genetics data analysis. Evolutionary Bioinformatics 1: 47-50.

Foxe JP, Stift M, Tedder A, Haudry A, Wright SI, Mable BK (2010). Reconstructing origins of loss of self-incompatibility and selfing in North American *Arabidopsis lyrata*: a population genetic context. Evolution 64: 3495-3510.

Garza JC, Freimer NB (1996). Homoplasy for size at microsatellite loci in humans and chimpanzees. Genome Research 6: 211-217.

Glemin S, Bataillon T, Ronfort J, Mignot A, Olivieri I (2001). Inbreeding depression in small populations of self-incompatible plants. Genetics 159: 1217-3019.

Glemin S, Bazin E, Charlesworth D (2006). Impact of mating systems on patterns of sequence polymorphism in flowering plants. Proceedings of the Royal Society B-Biological Sciences 273: 3011-3019.

Good-Avila SV, Stephenson AG (2002). The inheritance of modifiers conferring self-fertility in the partially self-incompatible perennial, *Campanula rapunculoides* L. (Campanulaceae). Evolution 56: 263-272.

Goodwillie C, Kalisz S, Eckert CG (2005). The evolutionary enigma of mixed mating systems in plants: ocurence, theoretical expectations, and empirical evidence. Annual Review of Ecology and Systematics 36: 47-79.

Goring DR, Walker JC (2004). Self-rejection - A new kinase connection. Science 303: 1474-1475.

Green DM, Sharbel TF, Kearsley J, Kaiser H (1996). Postglacial range fluctuation, genetic subdivision and speciation in the western North American spotted frog complex, *Rana pretiosa*. Evolution 50: 374-390.

Guo YL, Bechsgaard JS, Slotte T, Neuffer B, Lascoux M, Weigel D (2009). Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. Proceedings of the National Academy of Sciences of the United States of America 106: 5246-5251.

Hagenblad J, Bechsgaard J, Charlesworth D (2006). Linkage disequilibrium between incompatibility locus region genes in the plant *Arabidopsis lyrata*. Genetics 173: 1057-1073.

Hale LR, Singh RS (1991). A comprehensive study of genetic variation in natural-populations *of Drosophila-Melanogaster* .4. Mitochondrial-DNAvVariation and the role of history Vs selection in the genetic structure of geographic populations. Genetics 129: 103-117.

Hanzawa FM, Beattie AJ, Culver DC (1988). Directed dispersal - Demographic-analysis of an Ant-Seed mutualism. American Naturalist 131: 1-13.

Hatakeyama K, Takasaki T, Suzuki G, Nishio T, Watanabe K, Isogai A (2001). The *S* receptor kinase gene determines dominance relationships in stigma expression of self-incompatibility in *Brassica*. The Plant Journal 26: 69-76.

Hatakeyama K, Watanabe M, Takasaki T, Ojima K, Hinata K (1998). Dominance relationships between *S*-alleles in self-incompatible *Brassica campestris* L. Heredity 80: 241-247.

Hayes CN, Winsor JA, Stephenson AG (2005). Environmental variation influences the magnitude of inbreeding depression in *Cucurbita pepo* ssp *texana* (Cucurbitaceae). Journal of Evolutionary Biology 18: 147-155.

Hayes JP, Harrison RG (1992). Variation in mitochondrial-DNA and the biogeographic history of Woodrats (*Neotoma*) of the eastern United-States. Systematic Biology 41: 331-344.

Heaton TH, Talbot SL, Shields GF (1996). An Ice Age refugium for large mammals in the Alexander Archipelago, southeastern Alaska. Quaternary Research 46: 186-192.

Hedrick PW, Kalinowski ST (2000). Inbreeding depression in conservation biology. Annual Review of Ecology and Systematics 31: 139-162.

Hegi G (1986). Illustrierte Flora von Mittel-Europa. Pteridophyta und Spermatophyta. Band 4 Angiosperma Dicotyledons 2, Teil 1. 230-231.

Hewitt G (1993). Postglacial distributions and species substructure: lessons from pollen, insects and hybrid zones. In Evolutionary patterns and processes Linnean Society symposium series 14 (ed D R Lees & D Edwards) pp 97-123 London: Acedemic.

Hewitt G (2000). The genetic legacy of the Quaternary ice ages. Nature 405: 907-913.

Hewitt GM (1996). Some genetic consequences of ice ages, and their role in divergence and speciation. Biological Journal of the Linnean Society 58: 247-276.

Hewitt GM (1999). Post-glacial re-colonization of European biota. Biological Journal of the Linnean Society 68: 87-112.

Hewitt GM (2001). Speciation, hybrid zones and phylogeography - or seeing genes in space and time. Molecular Ecology 10: 537-549.

Hewitt GM (2004). Genetic consequences of climatic oscillations in the Quaternary. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 359: 183-195.

Hickerson MJ, Carstens BC, Cavender-Bares J, Crandall KA, Graham CH, Johnson JB (2010). Phylogeography's past, present, and future: 10 years after Avise, 2000. Molecular Phylogenetics and Evolution 54: 291-301.

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

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M(1989). The complete sequence of the Rice (*Oryza sativa*) chloroplast genome - Intermolecular recombination between distinct transfer-RNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Molecular & General Genetics 217: 185-194.

Hiscock SJ, Kües U (1999). Cellular and molecular mechanisms of sexual incompatibility in plants and fungi. International Review of Cytology 193: 165-295.

Hiscock SJ, McInnis SM (2004). Pollen recognition and rejection during the sporophytic self-incompatibility response: *Brassica* and beyond. Trends in Plant Science 9: 64-64.

Hoebe PN, Stift M, Tedder A, Mable B (2009). Multiple losses of self-incompatibility in North-American *Arabidopsis lyrata*?: Phylogeographic context and population genetic consequences. Molecular Ecology 18: 4924-4939.

Hoffman EA, Blouin MS (2004). Evolutionary history of the northern leopard frog: Reconstruction of phylogeny, phylogeography, and historical changes in population demography from mitochondrial DNA. Evolution 58: 145-159.

Holman JA (1992). Late quaternary herpetofauna of the central Great lakes Region, USA - Zoogeographical and Paleoecological Implications. Quaternary Science Reviews 11: 345-351.

Holman JA (1995). Pleistocene amphibians and reptiles in North America. Oxford Monographs on Geology and Geophysics 32: 1-243.

Horvitz CC, Lecorff J (1993). Spatial scale and dispersion Pattern of ant-dispersed and bird-dispersed herbs in two tropical lowland rainforests. Vegetatio 108: 351-362.

Hurst GDD, Jiggins FM (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proceedings of the Royal Society B-Biological Sciences 272: 1525-1534.

Husband BC, Schemske DW (1996). Evolution of the magnitude and timing of inbreeding depression in plants. Evolution 50: 54-70.

Ibrahim KM, Nichols RA, Hewitt GM (1996). Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. Heredity 77: 282-291.

Igic B, Bohs L, Kohn JR (2004). Historical inferences from the self-incompatibility locus. New Phytologist 161: 97-105.

Igic B, Bohs L, Kohn JR (2006). Ancient polymorphism reveals unidirectional breeding system shifts. Proceedings of the National Academy of Sciences of the USA 103: 1359-1363.

Ingvarsson PK, Ribstein S, Taylor DR (2003). Molecular evolution of insertions and deletions in the chloroplast genome of *Silene*. Mol Biol Evol 20: 1737-1740.

Jalas J, Suominen J (1994). Atlas florae Europaeae. University of Helsinki Press, Helsinki Volume 10.

Jaquard PG, Heim J (1984). Genetic differentiation and dispersal in plants. Springer-Verlag, Berlin, Germany.

Jarne P, Charlesworth D (1993). The evolution of the selfing rate in functionally hermaphrodite plants and animals. Annual Review of Ecology and Systematics 24: 441-466.

Kai N, Suzuki G, Watanabe M, Isogai A, Hinata K (2001). Sequence comparisons among dispersed members of the *Brassica* S multigene family in an S-9 genome. Molecular Genetics and Genomics 265: 526-534.

Kalinowski ST (2004). Counting alleles with rarefaction: Private alleles and hierarchical sampling designs. Conservation Genetics 5: 539-543.

Kalinowski ST (2005). HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness. Molecular Ecology Notes 5: 187-189.

Karkkainen K, Kuittinen H, van Treuren R, Vogl C, Oikarinen S, Savolainen O (1999). Genetic basis of inbreeding depression in *Arabis petraea*. Evolution 53: 1354-1365.

Karl SA, Avise JC (1992). Balancing selection at allozyme loci in Oysters - Implications from nuclear RFLPs. Science 256: 100-102.

Kelchner SA (2000). The evolution of non-coding chloroplast DNA and its application in plant systematics. Annals of the Missouri Botanical Garden 87: 482-498.

Keller LF, Waller DM (2002). Inbreeding effects in wild populations. Trends in Ecology and Evolution 17: 230-241.

Keller SR, Taylor DR (2008). History, chance and adaptation during biological invasion: separating stochastic phenotypic evolution from response to selection. Ecology Letters 11: 852-866.

Kephart SR (1990). Starch-gel electrophoresis of plant isozymes - a comparative analysis of techniques. American Journal of Botany 77: 693-712.

Kjellsson G (1985). Seed fate in a population of *Carex pilulifera* L .1. Seed dispersal and Ant-Seed mutualism. Oecologia 67: 416-423.

Koch M, Bishop J, Mitchell-Olds T (1999). Molecular systematics and evolution of *Arabidopsis* and *Arabis*. Plant Biology 1: 529-537.

Koch M, Haubold B, Mitchell-Olds T (2001). Molecular systematics of the Brassicaceae: evidence from coding plastidic matK and nuclear Chs sequences. American Journal of Botany 88: 534-544.

Koch M, Matschinger M (2007). Evolution and genetic differentiation among relatives of *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences of the United States of America 104: 6272-6277.

Koch MA, Dobeš C, Kiefer C, Schmickl R, Klimes L, Lysak MA (2007). Supernetwork identifies multiple events of plastid *trnF*(GAA) pseudogene evolution in the Brassicaceae. Molecular Biology and Evolution 24: 63-73.

Koch MA, Dobeš C, Matschinger M, Bleeker W, Vogel J, Kiefer M (2005). Evolution of the *trnF*(GAA) gene in *Arabidopsis* relatives and the Brassicaceae family: Monophyletic origin and subsequent diversification of a plastidic pseudogene. Molecular Biology and Evolution 22: 1032-1043.

Koch MA, Haubold B, Mitchell-Olds T (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). Molecular Biology and Evolution 17: 1483-1498.

Koch MA, Kiefer C, Ehrich D, Vogel J, Brochmann C, Mummenhoff K (2006). Three times out of Asia Minor: the phylogeography of *Arabis alpina* L. (Brassicaceae). Molecular Ecology 15: 825-839.

Koelling VA, Hamrick JL, Mauricio R (2011). Genetic diversity and structure in two species of *Leavenworthia* with self-incompatible and self-compatible populations. Heredity 106: 310-318.

Krafsur ES (2002). Population structure of the tsetse fly *Glossina pallidipes* estimated by allozyme, microsatellite and mitochondrial gene diversities. Insect Molecular Biology 11: 37-45.

Kreitman M (1996). The neutral theory is dead. Long live the neutral theory. Bioessays 18: 678-683.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005). Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences of the United States of America 102: 8369-8374.

Kropf M, Kadereit JW, Comes HP (2003). Differential cycles of range contraction and expansion in European high mountain plants during the Late Quaternary: insights from *Pritzelago alpina* (L.) O. Kuntze (Brassicaceae). Molecular Ecology 12: 931-949.

Kusaba M, Dwyer K, Hendershot J, Vrebalov J, Nasrallah JB, Nasrallah ME (2001). Self-incompatibility in the genus *Arabidopsis*: Characterization of the *S* locus in the outcrossing *A. lyrata* and its autogamous relative *A. thaliana*. Plant Cell 13: 627-643.

Kusaba M, Tung CW, Nasrallah ME, Nasrallah JB (2002). Monoallelic expression and dominance in anthers of self-incompatible *Arabidopsis lyrata*. Plant Physiology 128: 17-20.

Lalonde BA, Nasrallah ME, Dwyer KG, Chen CH, Barlow B, Nasrallah JB (1989). A Highly Conserved *Brassica* Gene with Homology to the S-Locus-Specific Glycoprotein Structural Gene. Plant Cell 1: 249-258.

Lande R, Schemske DW (1985). The evolution of self-fertilization and inbreeding depression in plants. Evolution 39: 24-40.

Langley CH, Voelker RA, Brown AJL, Ohnishi S, Dickson B, Montgomery E (1981). Null allele frequencies at allozyme Loci in natural populations *of Drosophila melanogaster*. Genetics 99: 151-156.

Lee JY, Edwards SV (2008). Divergence across Australia's Carpentarian barrier: Statistical phylogeography of the Red-Backed Fairy Wren (*Malurus melanocephalus*). Evolution 62: 3117-3134.

Levin DA (1996). The evolutionary significance of pseudo-self-fertility. American Naturalist 148: 321-332.

Levin DA, Kerster H (1969). Density-dependent gene dispersal in *Liatris*. American Naturalist 103: 61-69.

Lewis D (1979). Genetic versatility of incompatibility in plants. New Zeland journal of Botany 17: 637-644.

Li XM, Nield J, Hayman D, Langridge P (1994). Cloning a putative self-incompatibility gene from the pollen of the grass *Phalaris coerulescens*. Plant Cell 6: 1923-1932.

Lihova J, Fuertes Aguilar J, Marhold K, Nieto Feliner G (2004). Origin of the disjunct tetraploid *Cardamine amporitana* (Brassicaceae) assessed with nuclear and chloroplast DNA sequence data. American Journal of Botany 91: 1231-1242.

Liu P, Sherman-Broyles S, Nasrallah ME, Nasrallah JB (2007). A cryptic modifier causing transient self-incompatibility in *Arabidopsis thaliana*. Current Biology 17: 734-740.

Llaurens V, Castric V, Austerlitz F, Vekemans X (2008). High paternal diversity in the self-incompatible herb *Arabidopsis halleri* despite clonal reproduction and spatially restricted pollen dispersal. Molecular Ecology 17: 1577-1588.

Lloyd DG (1979). Some reproductive factors affecting the selection of self-fertilization in plants. American Naturalist 113: 67-79.

Lu YQ (2000). Effects of density on mixed mating systems and reproduction in natural populations of *Impatiens capensis*. International Journal of Plant Sciences 161: 671-681.

Luu DT, Hugues S, Passelegue E, Heizmann P (2001). Evidence for orthologous S-locus-related I genes in several genera of Brassicaceae. Molecular and General Genetics 264: 735-745.

Mable B, Robertson A, Dart S, Di Berardo C, Witham L (2005). Breakdown of self-incompatibility in the perennial *Arabidopsis lyrata* (Brassicaceae) and its genetic consequences. Evolution: 1437-1448.

Mable BK, Adam A (2007). Patterns of genetic diversity in outcrossing and selfing populations of *Arabidopsis lyrata*. Molecular Ecology 16: 3565-3580.

Mable BK, Beland J, Di Berardo C (2004). Inheritance and dominance of self-incompatibility alleles in polyploid *Arabidopsis lyrata*. Heredity 93: 476-486.

Mable BK, Robertson AV, Dart S, Berardo CD, Witham L, Fenster C (2005). Breakdown of self-incompatibility in the perennial *Arabidopsis lyrata* (brassicaceae) and its genetic consequences. Evolution 59: 1437-1448.

Mable BK, Schierup MH, Charlesworth D (2003). Estimating the number, frequency, and dominance of S-alleles in a natural population of *Arabidopsis lyrata* (Brassicaceae) with sporophytic control of self-incompatibility. Heredity 90: 422-431.

MacDonald SO, Cook JA (1996). The land mammal fauna of Southeast Alaska. Canadian Field-Naturalist 110: 571-598.

Maddison DR, Maddison WP (2000). MacClade 4: Analysis of phylogeny and charcter evolution. Version 4.0. Sinauer Associates: Sunderland, MA.

Mann DH, Hamilton TD (1995). Late Pleistocene and Holocene Paleoenvironments of the North Pacific Coast. Quaternary Science Reviews 14: 449-471.

Manzaneda AJ, Rey PJ (2009). Assessing ecological specialization of an ant-seed dispersal mutualism through a wide geographic range. Ecology 90: 3009-3022.

Matias L, Zamora R, Mendoza I, Hodar JA (2010). Seed dispersal patterns by large frugivorous Mammals in a degraded mosaic landscape. Restoration Ecology 18: 619-627.

Maynard Smith J (1978). The evolution of sex. Cambridge University Press, Cambridge, UK.

McCay TS, McCay DH, Czajka JL (2009). Deposition of exotic bird-dispersed seeds into three habitats of a fragmented landscape in the northeastern United States. Plant Ecology 203: 59-67.

Mena-Ali JI, Stephenson AG (2007). Segregation analyses of partial self-incompatibility in self and cross progeny of *Solanum carolinense* reveal a leaky S-allele. Genetics 177: 501-510.

Merila J, Bjorklund M, Baker AJ (1997). Historical demography and present day population structure of the Greenfinch, *Carduelis chlori*s - An analysis of mtDNA control-region sequences. Evolution 51: 946-956.

Michaux JR, Magnanou E, Paradis E, Nieberding C, Libois R (2003). Mitochondrial phylogeography of the Woodmouse (*Apodemus sylvaticus*) in the Western Palearctic region. Molecular Ecology 12: 685-697.

Mitchell-Olds TA-S, I. Koch, M. Sharbel, T. (2005). Crucifer evolution in the post-genomic era. In: Plant Diversity and Evolution: Genotypic and Phenotypic Variation in Higher Plants (Henry, RJ, ed): 119–136.

Morgan MT (2001). Consequences of life history for inbreeding depression and mating system evolution in plants. Proceedings of the Royal Society of London Series B-Biological Sciences 268: 1817-1824.

Mosandl R, Kleinert A (1998). Development of oaks (*Quercus petraea* (Matt.) Liebl.) emerged from bird-dispersed seeds under old-growth pine (*Pinus silvestris* L.) stands. Forest Ecology and Management 106: 35-44.

Muller MH, Leppala J, Savolainen O (2008). Genome-wide effects of postglacial colonization in *Arabidopsis lyrata*. Heredity 100: 47-58.

Mummenhoff K, Bruggemann H, Bowman JL (2001). Chloroplast DNA phylogeny and biogeography of Lepidium (Brassicaceae). American Journal of Botany 88: 2051-2063.

Murase K, Shiba H, Iwano M, Che FS, Watanabe M, Isogai A *et al* (2004). A membrane-anchored protein kinase involved in *Brassica* self-incompatibility signaling. Science 303: 1516-1519.

Murphy RW (1993). The phylogenetic analysis of allozyme data - Invalidity of coding alleles by presence absence and recommended procedures. Biochemical Systematics and Ecology 21: 25-38.

Muse SV (2000). Examining rates and patterns of nucleotide substitutions in plants. Plant Molecular Biology 42: 481-490.

Nasrallah JB (1997). Evolution of the *Brassica* self-incompatibility locus: A look into S-locus gene polymorphisms. Proceedings of the National Academy of Sciences of the United States of America 94: 9516-9519.

Nasrallah ME, Liu P, Nasrallah JB (2002). Generation of self-incompatible *Arabidopsis thaliana* by transfer of two S locus genes from *A. lyrata*. Science 297: 247-249.

Nasrallah ME, Liu P, Sherman-Broyles S, Boggs NA, Nasrallah JB (2004). Natural variation in expression of self-incompatibility in *Arabidopsis thaliana*: Implications for the evolution of selfing. Proceedings of the National Academy of Sciences of the United States of America 101: 16070-16074.

Nei M, Gojobori T (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3: 418-426.

Nei M, Maruyama T, Chakraborty R (1975). Bottleneck effect and genetic-variability in populations. Evolution 29: 1-10.

Nichols RA, Hewitt GM (1994). The genetic consequences of long-distance dispersal during colonization. Heredity 72: 312-317.

Nielsen LR, Siegismund HR, Philipp M (2003). Partial self-incompatibility in the polyploid endemic species *Scalesia affinis* (Asteraceae) from the Galapagos: remnants of a self-incompatibility system? Botanical Journal of the Linnean Society 142: 93-101.

Nielson M, Lohman K, Sullivan J (2001). Phylogeography of the tailed frog (*Ascaphus truei*): Implications for the biogeography of the Pacific Northwest. Evolution 55: 147-160.

Nishio T, Kusaba M (2000). Sequence diversity of *SLG* and *SRK* in *Brassica oleracea* L. Annals of Botany 85: 141-146.

Ockendon DJ (1975). Dominance relationships between S-alleles in the stigmas of Brussels sprouts (*Brassica oleracea*). Euphytica 24: 165-172.

Ogihara Y, Terachi T, Sasakuma T (1988). Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in Wheat species. Proceedings of the National Academy of Sciences of the United States of America 85: 8573-8577.

Okamoto S, Sato Y, Sakamoto K, Nishio T (2004). Distribution of similar self-incompatibility (S) haplotypes in different genera, *Raphanus* and *Brassica*. Sexual Plant Reproduction 17: 33-39.

Paetsch M, Mayland-Quellhorst S, Neuffer B (2006). Evolution of the self-incompatibility system in the Brassicaceae: identification of S-locus receptor kinase (*SRK*) in self-incompatible *Capsella grandiflora*. Heredity 97: 283-290.

Palmer JD (1991). Plastid chromosomes: structure and evolution. In: Cell Culture and Somatic Cell Genetics in Plants, Vol 7 The Molecular Biology of Plastids (L Bogorad & I K Vasil, eds), pp 5-53 Acedemic Press, San Deigo, CA.

Pannell JR, Barrett SCH (1998). Baker's law revisited: Reproductive assurance in a metapopulation. Evolution 52: 657-668.

Placyk JS, Burghardt GM, Small RL, King RB, Casper GS, Robinson JW (2007). Post-glacial recolonization of the Great Lakes region by the common gartersnake (*Thamnophis sirtalis*) inferred from mtDNA sequences. Molecular Phylogenetics and Evolution 43: 452-467.

Poncet BN, Herrmann D, Gugerli F, Taberlet P, Holderegger R, Gielly G (2010). Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. Molecular Ecology 19: 2896-2907.

Posada D, Crandall KA (1998). MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817-818.

Prigoda NL, Nassuth A, Mable BK (2005). Phenotypic and genotypic expression of self-incompatibility haplotypes in *Arabidopsis lyrata* suggests unique origin of alleles in different dominance classes. Molecular Biology and Evolution 22: 1609-1620.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

Pyke KA (1999). Plastid Division and Development. Plant Cell 11: 549-556.

Rambaut A (2002). Se-Al: sequence alignment editor, version 2 alpha11. Distributed over the World wide web (http://evolve.zps.ox.ac.uk).

Raubson LA, Jansen RK (2005). Chloroplast genomes of plants. In: Plant diversity and evolution: Genotypic and phenotypic variation in higher plants (Henry, RJ, ed) pp 45-68 CABI Publishing, Cambridge, MA.

Raymond M, Rousset F (1995). Genepop (Version-1.2) - Population-genetics software for exact tests and ecumenicism. Journal of Heredity 86: 248-249.

Ridley HN (1930). The dispersal of plants throughout the world. Reeve, Asford, Kent, UK.

Riihimaki M, Podolsky R, Kuittinen H, Koelewijn H, Savolainen O (2005). Studying genetics of adaptive variation in model organisms: flowering time variation in *Arabidopsis lyrata*. Genetica 123: 63-74.

Riihimaki M, Savolainen O (2004). Environmental and genetic effects on flowering differences between northern and southern populations of *Arabidopsis lyrata* (Brassicaceae). American Journal of Botany 91: 1036-1045.

Ritland K (1981). A model for estimation of outcrossing rate and gene frequencies using n independent loci. Heredity 43: 35-52.

Ritland K (2002). Extensions of models for the estimation of mating systems using n independent loci. Heredity 88: 221-228.

Rojas J, Rojas R (1999). DNAsp version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioninformatics 15: 174-175.

Ronquist F, Huelsenbeck JP (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574.

Ross-Ibarra J, Wright SI, Foxe JP, Kawabe A, DeRose-Wilson L, Gos G (2008). Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. Plos One 3: e2411

Rousset F (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145: 1219-1228.

Ruggiero MV, Jacquemin B, Castric V, Vekemans X (2008). Hitch-hiking to a locus under balancing selection: high sequence diversity and low population subdivision at the S-locus genomic region in *Arabidopsis halleri*. Genetical Research 90: 37-46.

Sage RD, Wolff JO (1986). Pleistocene glaciations, fluctuating ranges, and low genetic-variability in a large mammal (*Ovis dalli*). Evolution 40: 1092-1095.

Sauer JD (1988). Plant migration: the dynamics of geographic patterning in seed plant species. University of California Press, Berkeley, California, USA.

Schierup MH, Bechsgaard JS, Nielsen LH, Christiansen FB (2006). Selection at work in self-incompatible *Arabidopsis lyrata*: Mating patterns in a natural population. Genetics 172: 477-484.

Schierup MH, Mable BK, Awadalla P, Charlesworth D (2001). Identification and characterization of a polymorphic receptor kinase gene linked to the self-incompatibility locus of *Arabidopsis lyrata*. Genetics 158: 387-399.

Schierup MH, Vekemans X, Charlesworth D (2000). The effect of subdivision on variation at multi-allelic loci under balancing selection. Genetical Research 76: 51-62.

Schierup MH, Vekemans X, Christiansen FB (1998). Allelic genealogies in sporophytic self-incompatibility systems in plants. Genetics 150: 1187-1198.

Schlotterer C (2000). Evolutionary dynamics of microsatellite DNA. Chromosoma 109: 365-371.

Schmickl R, Jorgensen M, Brysting A, Koch M (2008). Phylogeographic implications for the North American boreal-arctic *Arabidopsis lyrata* complex. Plant Ecology & Diversity 1: 245-254.

Schmickl R, Jorgensen M, Brysting A, Koch M (2010). The evolutionary history of the *Arabidopsis lyrata* complex: A hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. BMC Evolutionary Biology 10: 98.

Schmickl R, Kiefer C, Dobeš C, Koch M (2008). Evolution of *trn*F(GAA) pseudogenes in cruciferous plants. Plant Systematics and Evolution 282: 229-240.

Schmitt T, Giessl A, Seitz A (2002). Postglacial colonisation of western central Europe by *Polyommatus coridon* (Poda 1761) (Lepidoptera : Lycaenidae): evidence from population genetics. Heredity 88: 26-34.

Schonswetter P, Paun O, Tribsch A, Niklfeld H (2003). Out of the Alps: Colonization of northern Europe by east Alpine populations of the Glacier Buttercup *Ranunculus glacialis* L. (Ranunculaceae). Molecular Ecology 12: 3373-3381.

Schonswetter P, Tribsch A (2005). Vicariance and dispersal in the alpine perennial *Bupleurum stellatum* L. (Apiaceae). Taxon 54: 725-732.

Schonswetter P, Tribsch A, Niklfeld H (2004). Amplified fragment length polymorphism (AFLP) suggests old and recent immigration into the Alps by the arctic-alpine annual *Comastoma tenellum* (Gentianaceae). Journal of Biogeography 31: 1673-1681.

Schopfer CR, Nasrallah ME, Nasrallah JB (1999). The male determinant of self-incompatibility in *Brassica*. Science 286: 1697-1700.

Shaw J, Lickey EB, Beck JT, Farmer SB, Liu WS, Miller J (2005). The tortoise and the hare II: Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. American Journal of Botany 92: 142-166.

Shiba H (2002). The dominance of alleles controlling self-incompatibility in *Brassica* pollen is regulated at the RNA Level. The plant cell online 14: 491-504.

Shiba H, Kenmochi M, Sugihara M, Iwano M, Kawasaki S, Suzuki G (2003). Genomic organization of the S-locus region of *Brassica*. Bioscience Biotechnology and Biochemistry 67: 622-626.

Shimizu KK, Cork JM, Caicedo AL, Mays CA, Moore RC, Olsen KM *et al* (2004). Darwinian selection on a selfing locus. Science 306: 2081-2084.

Shimizu KK, Shimizu-Inatsugi R, Tsuchimatsu T, Purugganan MD (2008). Independent origins of self-compatibility in *Arabidopsis thaliana*. Molecular Ecology 17: 704-714.

Shimizu-Inatsugi R, Lihova J, Iwanaga H, Kudoh H, Marhold K, Savolainen O *et al* (2009). The allopolyploid *Arabidopsis kamchatica* originated from multiple individuals of *Arabidopsis lyrata* and *Arabidopsis halleri*. Molecular Ecology 18: 4024-4048.

Sims TL (1993). Genetic-regulation of self-incompatibility. Critical Reviews in Plant Sciences 12: 129-167.

Slatkin M (1995). A measure of population subdivision based on microsatellite allele frequencies. Genetics 139: 457-462.

Snabel V (1995). On some problems of electrophoretic methods used in isoenzyme analysis. Helminthologia 32: 29-31.

Soltis DE, Gitzendanner MA, Strenge DD, Soltis PS (1997). Chloroplast DNA intraspecific phylogeography of plants from the Pacific Northwest of North America. Plant Systematics and Evolution 206: 353-373.

Soltis DE, Morris AB, McLachlan JS, Manos PS, Soltis PS (2006). Comparative phylogeography of unglaciated eastern North America. Molecular Ecology 15: 4261-4293.

Soltis DH, C. Darrow, D. Gastony, G. (1983). Starch gel electrophoresis of ferns: a compilation of grinding buffers, gel and electrode buffers and staining schedules. American Fern Journal 73: 9-27.

Soltis PS, Soltis DE, Tucker TL, Lang FA (1992). Allozyme variability is absent in the narrow endemic *Bensoniella oregona* (Saxifragaceae). Conservation Biology 6: 131-134.

Stebbins GL (1957). Self-fertilization and population variability in the higher plants. American Naturalist 91: 337-354.

Steele CA, Storfer A (2006). Coalescent-based hypothesis testing supports multiple Pleistocene refugia in the Pacific Northwest for the Pacific giant salamander (*Dicamptodon tenebrosus*). Molecular Ecology 15: 2477-2487.

Stehlik I, Schneller JJ, Bachmann K (2002). Immigration and *in situ* glacial survival of the low-alpine *Erinus alpinus* (Scrophulariaceae). Biological Journal of the Linnean Society 77: 87-103.

Stephenson AG, Travers SE, Mena-Ali JI, Winsor JA (2003). Pollen performance before and during the autotrophic-heterotrophic transition of pollen tube growth. Philosophical Transactions of the Royal Society B-Biological Sciences 358: 1009-1017.

Stevens JP, Kay QON (1989). The number, dominance relationships and frequencies of self-incompatibility alleles in a natural population of *Sinapis arvensis* L in South-Wales. Heredity 62: 199-205.

Stuber CW, Moll RH, Goodman MM, Schaffer HE, Weir BS (1980). Allozyme frequency changes associated with selection for increased grain yield in Maize (*Zea mays*-L). Genetics 95: 225-236.

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

Taberlet P, Gielly L, Pautou G, Bouvet J (1991). Universal primers for amplification of 3 noncoding regions of chloroplast DNA. Plant Molecular Biology 17: 1105-1109.

Takahata N (1990). A simple genealogical structure of strongly balanced allelic lines and transspecies evolution of polymorphism. Proceedings of the National Academy of Sciences of the United States of America 87: 2419-2423.

Takasaki T, Hatakeyama K, Suzuki G, Watanabe M, Isogai A, Hinata K (2000). The S receptor kinase determines self-incompatibility in *Brassica stigma*. Nature 403: 913-916.

Takayama S, Isogai A (2003). Molecular mechanism of self-recognition in *Brassica* self-incompatibility. Journal of Experimental Botany 54: 149-156.

Takayama S, Shiba H, Iwano M, Shimosato H, Che FS, Kai N *et al* (2000). The pollen determinant of self-incompatibility in *Brassica campestris*. Proceedings of the National Academy of Sciences of the United States of America 97: 1920-1925.

Takayama S, Shimosato H, Shiba H, Funato M, Che FS, Watanabe M *et al* (2001). Direct ligand-receptor complex interaction controls *Brassica* self-incompatibility. Nature 413: 534-538.

Takebayashi N, Morrell PL (2001). Is self-fertilization an evolutionary dead end? Revisiting an old hypothesis with genetic theories and a macroevolutionary approach. American Journal of Botany 88: 1143-1150.

Tang CL, Toomajian C, Sherman-Broyles S, Plagnol V, Guo YL, Hu TT *et al* (2007). The evolution of selfing in *Arabidopsis thaliana*. Science 317: 1070-1072.

Tavare S (1986). (1986) Some probabilistic and statistical problems on the analysis of DNA sequences. In: Lectures in mathematics in the life sciences 17: 57–86.

Templeton AR (1986). Coadaptation and outbreeding depression. In: ME Soulé Ed, Conservation Biology: The Science of Scarcity and Diversity, Sunderland, Massachusetts pp 105-116.

Templeton AR, Crandall KA, Sing CF (1992). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA-sequence data .3. Cladogram estimation. Genetics 132: 619-633.

Thompson KF, Taylor JP (1966). Non-linear dominance relationships between S Alleles. Heredity 21: 345-355.

Tobias CM, Howlett B, Nasrallah JB (1992). An *Arabidopsis thaliana* gene with sequence similarity to the S-Locus receptor kinase of *Brassica oleracea* - Sequence and expression. Plant Physiology 99: 284-290.

Trick M, Flavell RB (1989). A homozygous S-genotype of *Brassica oleracea* expresses 2 S-like genes. Molecular & General Genetics 218: 112-117.

Tsuchimatsu T, Suwabe K, Shimizu-Inatsugi R, Isokawa S, Pavlidis P, Stadler T *et al* (2010). Evolution of self-compatibility in *Arabidopsis* by a mutation in the male specificity gene. Nature 464(7293): 1342-1343.

Uynoyama MK, Waller DM (1991). Coevolution of self-fertilization and inbreeding depression. III. Homozygous leathal mutations at multiple loci. Theoretical Population Biology 40: 173-210.

Vallejo-Marin M, Uyenoyama MK (2004). On the evolutionary costs of self-incompatibility: Incomplete reproductive compensation due to pollen limitation. Evolution 58: 1924-1935.

Van de Pijl L (1969). Principles of dispersal in higher plants. Springer-Verlag, Berlin, Germany

van Oppen MJH, Rico C, Turner GF, Hewitt GM (2000). Extensive homoplasy, nonstepwise mutations, and shared ancestral polymorphism at a complex microsatellite locus in Lake Malawi cichlids. Molecular Biology and Evolution 17: 489-498.

Vekemans X, Schierup MH, Christiansen FB (1998). Mate availability and fecundity selection in multi-allelic self-incompatibility systems in plants. Evolution 52: 19-29.

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

Visser DL, Hal JG, Verhoeven WH (1982). Classification of S-alleles by their activity in S-heterozygotes of brussels sprouts (*Brassica oleracea* var. Gemmifera (DC.) Schultz). Euphytica 31: 603-611.

Vogel JC, Rumsey FJ, Schneller JJ, Barrett JA, Gibby M (1999). Where are the glacial refugia in Europe? Evidence from pteridophytes. Biological Journal of the Linnean Society 66: 23-37.

Vogler DW, Stephenson AG (2001). The potential for mixed mating in a self-incompatible plant. International Journal of Plant Sciences 162: 801-805.

Wagner RS, Miller MP, Crisafulli CM, Haig SM (2005). Geographic variation, genetic structure, and conservation unit designation in the Larch Mountain salamander (*Plethodon larselli*). Canadian Journal of Zoology-Revue Canadienne De Zoologie 83: 396-406.

Wang JL, Hill WG, Charlesworth D, Charlesworth B (1999). Dynamics of inbreeding depression due to deleterious mutations in small populations: mutation parameters and inbreeding rate. Genetical Research 74: 165-178.

Wang RH, Farrona S, Vincent C, Joecker A, Schoof H, Turck F *et al* (2009). PEP1 regulates perennial flowering in *Arabis alpina*. Nature 459: 423-U138.

Webb T, Bartlein PJ (1992). Global Changes during the Last 3 Million Years - Climatic Controls and Biotic Responses. Annual Review of Ecology and Systematics 23: 141-173.

Weller SG, Sakai AK (1999). Using phylogenetic approaches for the analysis of plant breeding system evolution. Annual Review of Ecology and Systematics 30: 167-199.

Weller SG, Wagner WL, Sakai AK (1995). A Phylogenetic Analysis of *Schiedea* and *Alsinidendron* (Caryophyllaceae, Alsinoideae) - Implications for the evolution of breeding systems. Systematic Botany 20: 315-337.

Wendel JW, N. (1989). Visualisation and interpretation of plant isozymes. In Isozymes in plant biology (eds. Soltis, DE. & Soltis, PS.) 5-45.

Whitehouse HLK (1950). Multiple allemorph incompatibility of pollen and style in the evolution of angiosperms. Annals of Botany 14: 198-216.

Whitlock C (1992). Vegetational and climatic history of the Pacific-Northwest during the last 20,000 years - Implications for understanding present day biodiversity. Northwest Environmental Journal 8: 5-28.

Willson MF (1993). Mammals as seed-dispersal mutualists in North America. Oikos 67: 159-176.

Wright SI, Lauga B, Charlesworth D (2002). Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. Molecular Biology and Evolution 19: 1407-1420.

Wright SI, Lauga B, Charlesworth D (2003). Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. Molecular Ecology 12: 1247-1263.

Zamudio KR, Savage WK (2003). Historical isolation, range expansion, and secondary contact of two highly divergent mitochondrial lineages in spotted salamanders (*Ambystoma maculatum*). Evolution 57: 1631-1652.

Zhang LB, Comes HP, Kadereit JW (2004). The temporal course of quaternary diversification in the European high mountain endemic *Primula* sect. Auricula (Primulaceae). International Journal of Plant Sciences 165: 191-207.

Zink RM (2010). Drawbacks with the use of microsatellites in phylogeography: the song sparrow *Melospiza melodia* as a case study. Journal of Avian Biology 41: 1-7.