Demography and Polyploidy in *Capsella*

KATE ST.ONGE
Dissertation presented at Uppsala University to be publicly examined in Zootissalen, Evolutionary biology centre, Villavägen 9, Uppsala, Friday, February 4, 2011 at 10:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

Studies of demography and population structure give insight into important evolutionary processes such as speciation and diversification. In the present work I perform such studies in the genus Capsella, which has three species: C. grandiflora, an outcrossing diploid, C. rubella a selfing diploid, and C. bursa-pastoris a selfing tetraploid. These three species make a good model system for evolutionary studies because they encompass two major plant evolutionary processes: mating system shifts and polyploidization. To conduct my studies I have gathered a large number of samples across the distributions of each species and scored them both phenotypically and genotypically: more specifically we measured flowering time and collected DNA sequence data.

In the tetraploid C. bursa-pastoris we applied an association mapping approach which takes population structure into account to search for genetic variation associated with variation in flowering time. Flowering time is an important and highly adaptive trait which is frequently subject to natural selection. We found evidence of association between flowering time and several single nucleotide polymorphisms (SNPs) within the flowering locus C (FLC) and cryptochrome 1 (CRY1). In the case of FLC these SNPs code for nonconsensus splice site variation in one of the two copies of the gene. The SNPs could potentially have functional consequences and our results imply that non-functionalization of duplicate genes could be an important source of phenotypic variation.

Using a novel coalescent based approach, we investigated the polyploid origin of C. bursa-pastoris and find evidence supporting a recent autoploidy origin of this species. In the two diploid species, I use sequence data to investigate population structure and demographic history and to assess the effects of selfing on C. rubella. Observed patterns of population structure and genetic diversity in C. rubella can be explained by a combination of both demographic history and mating system. Observed patterns in C. grandiflora suggest that the investigated populations do not deviate strongly from the SNM, which has rarely been found in modern demographic studies.

Finally, we investigate the effect of sampling strategy on demographic inference. Extensive sampling both within and across our populations allow us to empirically test the effect of sampling strategy on demographic inference. We complement the empirical analysis with simulations and conclude that the effect of sampling strategy is in many cases weak compared with that of demographic events. Nevertheless, these effects are real and have the potential to lead to false inference and therefore sampling strategy should be carefully considered in demographic studies.

Keywords: Approximate Bayesian Computation, Mating system, Brassicaceae, flowering time, genetic diversity, Evolution

Kate St.Onge. Department of Ecology and Genetics, Evolutionary Functional Genomics, Norbyvägen 18 C, Uppsala University, SE-75236 Uppsala, Sweden.

© Kate St.Onge 2010

ISSN 1651-6214
ISBN 978-91-554-7979-4
urn:nbn:se:uu:diva-136696 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-136696)
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


*These authors contributed equally to this work.

Reprints were made with permission from the respective publishers.
1 Introduction

Searching for the signature of natural selection in genetic data has been a central goal of molecular evolutionary biology since the field was first established. Which mutations make an organism more fit? Like any scientific question a null hypothesis is required. Such a model was presented in 1968 by Kimura in the form of his neutral theory of evolution. The neutral theory assumes that the variation we observe is mostly due to neutral mutations on their way to fixation or loss under random drift as deleterious mutations quickly disappear from populations and advantageous ones are rare. From this theory the Standard Neutral Model (SNM) was created, which has been the dominant null hypothesis in molecular evolution and population genetics since its inception.

The SNM describes a population, which is not experiencing selection, is mating randomly and is not experiencing population size changes. Departures from this model can be caused by violation of any of these assumptions, which clearly leads to several alternative explanatory hypotheses, only one of which is selection. The assumptions of random mating and constant population size can be violated if there is population structure and complex demographic histories. Importantly, demographic departures from the SNM can leave signatures in the genome that are identical to those left by selection and demographic changes and selection often occur simultaneously. For instance colonization of new environments are often accompanied by population bottlenecks and expansions as well as new selective pressures. Here lies the importance of complementing studies searching for the signature of selection with information on population structure and demographic history.

Studies of demography and population structure are also interesting in their own right, as they give insight into important evolutionary processes such as speciation and diversification. Mechanisms of speciation and diversification have always been of general interest to evolutionary biologists and are of importance to the more contemporary work of conservation biologists. Demographic studies using coalescent analysis are an important part of understanding these processes. In the current work, I use coalescent analysis to study demographic history, population structure and speciation within the genus *Capsella*, laying the ground for future studies searching for the signature of selection in candidate genes. I use my empirical data in combination with simulated data to investigate what type of sampling strategy is most appropriate for such studies. Finally, I use an association mapping approach
that accounts for population structure to search for genetic variation which is associated with variation in flowering time, an important quantitative trait in plants.

1.1 *Capsella* as a model for evolutionary genomic studies

*Capsella* is a small genus including only three species according to current classification. *Capsella grandiflora* (Fauche & Chaub) Boiss, is a self-incompatible diploid (2n=16) and this ploidy and mating system are likely the ancestral state of the genus (Paetsch et al. 2006). *Capsella rubella* Rent., which has recently evolved from *C. grandiflora* (Foxe et al 2009, Guo et al 2009, Paper II), is also a diploid (2n=16) but has lost its self-incompatibility system and become predominately selfing (0.90-0.94 Paper II). Finally, *Capsella bursa-pastoris* (L.) Medik, is also a selfer, but is a tetraploid species (2n=4x=32).

*Capsella* is an excellent model for molecular evolutionary studies for several reasons. First, this small genus presents the opportunity to study several key phenomena in plant evolution and speciation. Important to this study are the change in mating system between *C. grandiflora* and both *C. rubella* and *C. bursa-pastoris* and the polyploidisation of *C. bursa-pastoris*. Both of these changes have occurred independently in many plant lineages, often leading to speciation. They can also have large effects on the amount and structure of genetic diversity. A second important advantage of working with *Capsella* is its phylogenetic relations within the Brassicaceae family. *Capsella* is closely related to the model plant species *Arabidopsis thaliana*, which has a two-fold advantage. Because of their close relation, many of the tools and methods developed for *Arabidopsis* work with little adjustment in *Capsella*; in fact most of the PCR primers that we have designed for the current work were designed using the *Arabidopsis thaliana* full genome sequence. Additionally, because of the popularity of *Arabidopsis* in the past decade, quite a large amount of molecular work has been done on the genus and its close relatives, making comparative genomic studies between *Arabidopsis*, *Capsella* and their relatives possible. And finally, work within the *Capsella* genus has real tangible use to agriculture because it belongs to the Brassicaceae family, the major crop family which includes mustard, cabbage and oil seed rape, and also because *C. bursa-pastoris* is a common weed of agriculture. In fact, the evolution of polyploidy and weediness in *C. bursa-pastoris* is of direct interest to agricultural research.
1.2 Sampling and identification of *Capsella* species

*Capsella grandiflora* has the most limited distribution of the three *Capsella* species. It occurs locally in northern Italy, but is mainly found in northern Greece and Albania, where it grows in disturbed habitats but also in less disturbed mountain meadows (Hurka and Neuffer, 1997; Paetsch et al. 2006). The natural distribution of *C. rubella* is thought to be Western Europe, the Balkans, Greece, North Africa and the Middle East, but it also occasionally followed settlers into the Americas and Australia where it occurs in Mediterranean climates (Hurka and Neuffer, 1997; Paetsch et al. 2006). *C. bursa-pastoris* has the widest distribution of the three *Capsella* species, it can be found worldwide and avoids only the tropics (Hurka and Neuffer, 1997). This species is believed to have originated in the Middle East, and to have spread across the world with European colonizers reaching the Americas and Australia in the 18th century (Neuffer and Hurka 1999; Ceplitis et al. 2005). However, evidence presented in this thesis suggest that *C. bursa-pastoris* may have originated in Greece, the only region where all three species currently thrive, making it the likely origin of the genus.

*C. grandiflora* represents the ancestral obligate outbreeding species of the genus (Paetsch et al. 2006), and is clearly distinguishable by its large and scented flowers. *C. bursa-pastoris* and *C. rubella*, on the other hand, both display reduced, unscented flowers, classic traits of the selfing syndrome (Hurka and Neuffer, 1997). *C. bursa-pastoris* is known for its extensive variability in many morphological traits, such as leaf and fruit shape and flowering time (Shull 1929; Neuffer and Hurka, 1989 a,b; Neuffer and Eschner, 1995, Ceplitis et al. 2005). The morphological characters of *C. bursa-pastoris* overlap broadly with those of *C. rubella*, often making it difficult or impossible to distinguish the two morphologically. Although *C. rubella* more often displays red tipped flower buds, red tinged fruits and a more concave shape to the lower tapering half of the fruit (personal communications with Valerie LeCorre and personal observations), testing for ploidy level remains the only definitive method for distinguishing the two species.

Luckily, sampling *Capsella* species is often as easy as flying to a Mediterranean country, renting a car and heading for the mountains. All three species thrive in disturbed habitats, such as roadsides, grazed fields, animal or foot paths and the periphery of agriculture fields, making them easy to find. *C. bursa-pastoris* seems to be the most weedy of the three and is commonly found even in highly populated cities. During my studies we have accumulated a large collection of *Capsella*. I have sampled *Capsella* in Greece, Albania, Montenegro, Bosnia-Herzegovina, Croatia, Italy, Switzerland, France and Spain. We have also received samples from colleagues in Spain, France, Morocco, Bulgaria, Greece, Luxembourg, Poland, Belgium, Germany, the Netherlands, Iceland, Algeria, Israel, Iran, Palestine, Japan, the United States, Russia and China. A. Ceplitis collected samples from Turkey and
Syria, T. Slote from Finland, Sweden and Greece and M. Lascoux from France, Great Britain, China and Russia, in particular Siberia. We have today a collection of over 300 populations, most of which are represented by many individuals and many have been tested for ploidy level using flow cytometry (Figure 1.1).
1.3 Mating System

Self-incompatibility (SI) systems are widespread in angiosperms. These systems arise to prevent self-fertilization in plants with hermaphroditic flowers and thereby promote out-crossing. Capsella has a sporophytic self-incompatibility (SSI) system, which is basal to the Brassicaceae family. In this system, self pollen is rejected based on its diploid genotype, and control of this mechanism lies in the S-locus. The transition from self-incompatibility to self-compatibility has occurred many times in angiosperms (Barrett, 2002), and according to our work has occurred twice among the three species of Capsella.

The long term evolutionary consequences of selfing vs. outcrossing have been of great interest to plant evolutionary biologists (Escobar et al. 2010; Goldberg et al. 2010; Igic et al. 2008; Takebayashi and Morrell, 2001; Wright and Barrett, 2010). It is traditionally thought that outcrossing is advantageous in the long term as it limits the impact of inbreeding depression (Porcher and Lande, 2005), while selfing is advantageous in the short term because selfers have an inherent transmission advantage over outcrossers (Fisher, 1941) and reproductive assurance eliminating reliance on mates, thereby increasing colonization potential (Baker, 1955). These advantages are the cause of the repeated loss of self-incompatibly in various plant lineages. Nevertheless selfers are often viewed as evolutionary dead-ends because they have limited potential for adaptation and speciation, due to inbreeding depression, the accumulation of deleterious mutations and overall low efficiency of selection as a result of small effective population size (Takebayashi and Morrell, 2001). It is also unlikely that selfers will revert to outcrossing, due to the complex genetic mechanisms that allow self-pollen recognition (Steinbachs and Holsinger, 2002). This “dead-end” hypothesis of selfing is theoretically logical, and is supported by the observation that most selfing lineages are young (Schoen et al. 1997). On the other hand, the increased rate of deleterious mutation accumulation in selfers is questionable and has so far not been supported by studies of protein evolution (Escobar et al. 2010; Wright et al, 2008), possibly because strict selfers do not exist. Nevertheless, both selfing and outcrossing lineages persist in many plant genera making the study of the evolutionary dynamic of these two opposing mating systems very interesting.

The transition to selfing is expected to have major effects on genome diversity and structure. For example, effects have been seen in population diversity, population divergence, linkage disequilibrium, codon bias and efficiency of selection (Charlesworth 2003; Charlesworth and Wright 2001). Many of the effects of selfing stem from the reduction in effective population size (Ne), which is expected to be reduced by half. This decrease in Ne reduces genetic variation, increases the effect of genetic drift and can lead to the fixation of slightly deleterious mutations. Because homozygosity is in-
increased, recombination will also have much less impact in selfers than in outcrossers, as only double heterozygotes are affected by recombination (Nordborg 2000). Invasiveness and increased colonization ability are often associated with selfing and can further reduce Ne due to bottlenecks during colonization and metapopulation dynamics (Ingvarsson, 2002). In this thesis I compare and contrast genetic diversity, population structure and demographic history in *C. rubella* and *C. grandiflora*, the two diploid species with different mating systems.

### 1.4 Polyploidy

Polyploidy is extremely common among plants. It has been estimated that between 30-70% of all angiosperms have polyploidy in their history (Masterson, 1994), and that 15% of speciation events in angiosperms and 31% in ferns are accompanied by changes in ploidy level (Wood et al. 2009), making polyploidization the most common mechanism of sympatric speciation in plants (Otto and Whitton, 2000). Speciation events resulting from polyploidization have been traditionally thought to be instantaneous because backcrosses between newly formed polyploids and their progenitor result in offspring with odd numbered ploidy levels which have very low fertility. However, more recent studies have been changing this view and demonstrate that multiple origins and even gene flow across ploidy levels are more common than originally thought (Ramsey and Schemske, 1998; Chapman and Abbott, 2010 and Soltis and Soltis, 2010 and reference therein). For instance in *Capsella*, there is evidence of gene flow from *C. rubella* into *C. bursapastoris* (Slotte et al. 2008). Polyploids have been theorized to have several evolutionary advantages. For example, Otto and Whitton (2000) theorize that polyploids have a greater chance of harbouring new beneficial alleles and are therefore more likely to evolve novel gene function. Although the evolutionary advantage and flexibility of polyploidy remain to be proven empirically, it is supported by the fact that many polyploids display heterosis, have broader ecological tolerances (Stebbins 1950, Stebbins, 1980) and are often successful weeds and crops (for example wheat, potato, cotton, apple, tobacco and many Brassiceace).

Polyploids are formed in two ways 1) by merging of two genomes from different species, resulting in allopolyploids or 2) by whole genome duplication within a single species, resulting in autopolyploids. After genome doubling, autopolyploids are expected to form multi-valents at meiosis and therefore display polysomic inheritance. On the other hand, it has commonly been assumed that duplicated chromosomes in allopolyploids are different enough to avoid multivalents at meiosis, and therefore allopolyploids are expected to display disomic inheritance. These differences in inheritance patterns have been considered by many to be a diagnostic trait in distinguish-
ing between allo- and autopolyploids. However, with renewed interest in genome doubling as an evolutionary process in the genomics era, recent genomic studies had proven that polyploid genomes can be highly dynamic and undergo rapid structural and functional alterations (Doyle et al. 2008; Leitch and Leitch, 2008), for example diploidization. How rapid diploidization is in autopolyploids remains to be tested.

Capsella bursa-pastoris, which is arguably the most notorious weed in the world, is an example of a polyploid with a broader range of ecological tolerances than its diploid relatives. In addition to co-inhabiting the ranges of both C. grandiflora in the western Balkans and C. rubella in central and southern Europe, C. bursa-pastoris has spread to nearly all habitable corners of the planet save the tropics. Early genetic studies found that C. bursa-pastoris shared alleles with both C. rubella and C. grandiflora suggesting that it was an allopolyploid between these two species (Hurka et al, 1989). This theory was not supported by later studies demonstrating a very recent divergence of C. rubella from C. grandiflora, much later than the presumed time of the split of C. bursa-pastoris from its common ancestor (Foxe et al, 2009, Guo et al, 2009, Paper II). Further studies hypothesized two alternatives, namely an autopolyploid origin from the ancestor of C. grandiflora (Hurka and Neuffer, 1997) or an allopolyploid origin with C. grandiflora and an unknown related species as parents (Slotte at al, 2008). These possibilities are tested in Paper IV using coalescent methods and the first hypothesis is deemed most likely. Understanding the demographic history and timing of the polyploid speciation of C. bursa-pastoris is a vital part of elucidating how polyploidy has enabled the success of this species as we will then be able to correct for the age of the different species in pairwise comparisons.

1.5 Searching for variation affecting flowering time in candidate genes

The timing of the transition to flowering is crucial for plants, as matching flowering time and optimal seasonal conditions is essential for insuring reproductive success. Since seasonal conditions vary widely around the globe, flowering time is a highly adaptive trait in plants, often varying along latitudinal clines as seasonal and light conditions change in a predictable way with latitude. Aside from being a very important ecological trait, flowering time is also important in agriculture and forestry. As a result, flowering time is one of the best-studied plant traits, and the genetic basis of its natural variation has started to be well described in A. thaliana (Mitchell-Olds and Schmitt, 2006).
Four main pathways are responsible for initiating the transition from vegetative to reproductive growth: the photoperiod, the vernalization, the gibberellin and the autonomous pathways (Mouradov et al. 2002; Simpson and Dean 2002; Koornneef et al. 2004). These pathways allow plants to precisely assess the correct time for flowering and, in *Arabidopsis thaliana*, involve over 50 genes, with many more genes involved in peripheral pathways. This large number of genes makes for a large number of potential targets for natural selection on the flowering time trait. In fact several different flowering time genes and variation within them have been found to be associated with variation in flowering time — or phenology in general — in various species (Ehrenreich et al, 2009; Gyllenstrand et al; 2007; Kuittinen et al. 2008; Ma et al. 2010; Kruskopf-Osterberg et al. 2002; Thornsberry et al. 2001; Xue et al, 2008)

Association mapping is a tool commonly used to unravel the genetic basis of quantitative traits by searching for associations between phenotypes and genetic markers in natural populations (Zhu et al. 2008). In non-model organisms where information is generally more limited, genetic markers are often chosen using a candidate gene approach where genes involved in the trait of interest are first identified by QTL or functional studies or by studies of the trait of interest in other species. Genetic markers within these genes, usually single nucleotide polymorphisms (SNPs), are then scored in plants grown in common garden experiments, which have also been phenotyped for the trait of interest. These phenotypic and genotypic data are then searched for statistical associations. Genetic markers yielding such associations and the neighbouring areas can then be further examined for potentially causative mutations. We employ this method in paper I to search for genetic variation responsible for the clinal variation in flowering time observed in *C. bursa-pastoris*. 
2 Research Aims

In this thesis I investigate the natural genetic diversity of the three *Capsella* species and the role of duplicated genes in the important quantitative trait flowering time. Using sequence data, coalescent theory and association mapping I address the following main topics in this work:

1. The role of duplicated genes in adaptation in the tetraploid *C. bursa-pastoris*. (paper I)
2. The effects of selfing and demographic history on population structure and genetic diversity in *C.rubella* and *C.grandiflora*. (Paper II)
3. The effect of sampling strategies on population genetic inferences (Paper III)
4. The origins of the polyploid *C. bursa-pastoris* coalescent analysis (Paper IV)
3 Results and Discussion

3.1 Association Mapping of the quantitative flowering time trait in *C. bursa-pastoris*

In this study we search for associations between genetic and phenotypic variation in the timing of flowering, a quantitative trait commonly referred to simply as flowering time. We took a candidate gene approach and focused our study on four duplicated genes involved in the flowering pathway. In two of the genes, CRY1 and LD, one homeolog (homeolog A) was located in a chromosomal region previously identified as a major flowering time QTL (Ceplitis et al., unpublished). The two other genes, FRI and FLC, are major determinants of flowering time variation in *Arabidopsis thaliana*.

We designed and used locus specific primers to amplify and sequence both homeologs of our four candidate genes and also 10 background loci. These background genes were used to assess the level of population structure and kinship among our samples. Population structure can lead to spurious associations in association mapping studies and we therefore included structure and kinship in our analysis by using a mixed model previously shown to reduce the rate of false positives in *A. thaliana* (Yu et al. 2006; Zhao et al. 2007).

We conducted our association study in two samples of *Capsella bursa-pastoris* from two geographical regions, which have previously been shown to have different histories (Slotte et al. 2008). The first population includes 60 accessions from Europe, the Middle East and North Africa, referred to here as western Eurasia, and the second population consists of 42 accessions from China. The variation in flowering time for these populations was determined under standard long day conditions and was measured as days from germination to the opening of the first flower.

Our analysis of population structure identified one northern and one southern subpopulation in the western Eurasian region, but no substructuring was found among the Chinese samples. We found a total of five SNPs that were significantly associated with flowering time in our study. Three of these SNPs occur in CRY1 homeolog A and occur only in western Eurasia. These three SNPs were in complete linkage disequilibrium and are associated with late flowering, but are all synonymous. It is therefore likely that they are linked to the causative mutation, which presumably lies outside the coding region, possibly in the promoter.
The remaining two SNPs occur in the FLC A homoeolog, at positions 452 and 495. SNP452 only occurs in the northern European subpopulation and SNP495 only occurs in China. Interestingly, both SNPs disrupt splice sites, SNP452 yielding a non-consensus splice acceptor site while the SNP492 yields a non-consensus splice donor site, and both are associated with early flowering. We performed an additional assay to assess the impact of these polymorphisms on the FLC A transcript and expression level. In accordance with cDNA predictions, Chinese accessions harbouring the non-consensus splice site SNP495, expressed an FLC A transcripts lacked a 42-bp region corresponding to exon 5, and these accessions either expressed only this transcript or both this transcript and FLC B. In Northern Europe, only one accession harboured the non-consensus splice site SNP452 and expressed an FLC A transcript. This transcript lacked the 42-bp region corresponding to exon 5. The other assayed accessions expressed only FLC B. Although little is known about the functional importance of the region encoded by exon 5,

**Figure 3.1** Estimated mean of number of days to flowering, +/- SE for accessions with difference candidate SNP genotypes at FLC A. In both western Eurasia (A) and China (B), nonconsensus splice site alleles (in boldface type) are associated with earlier flowering. In western Eurasia, FLC A splice site variation is found only in the northern cluster.

these mutations may negatively affect the function of FLC. Accessions harbouring them indeed flowered early (Fig. 3.1), which is consistent with the function of FLC as a major repressor of flowering time. These results suggest that ongoing non-functionalization of duplicated genes could provide an important source of phenotypic variation.
3.2 Demographic history

3.2.1 Divergent histories and population structure in *Capsella grandiflora* and *C. rubella* (Paper II)

I had two main objectives in paper II concerning the two diploid *Capsella* species, the outcrossing *C. grandiflora* and the selfing *C. rubella*. First, I wanted to assess the effect of mating system by comparing population structure and genetic diversity in these two species while accounting for species history. Because unaccounted for population structure can cause mis-inference in demographic studies and spurious associations in association studies, it is important that these types of studies go hand-in-hand. It is also interesting in an ecological and conservational sense to understand how natural population structure is influenced by mating system. Secondly, I wanted to infer the demographic histories of these two species, essentially testing and confirming findings and predictions of previous studies (Foxe et al. 2009; Guo et al. 2009). Specifically, I was interesting in testing if *C. grandiflora* could in fact be one of the rare species that conforms to the standard neutral model with a large and stable effective population size.

To address these questions I collected sequence data from 16 gene fragments in 7 populations of each species, consisting of approximately 10 individuals each. This depth of within populations sampling has not been used in previous studies based on sequence data and is important when a study’s aim is to assess patterns of population structure and diversity within and between populations. Population structure and genetic diversity was assessed using the programs STRUCTURE, Arlequin and DNAsp. Demographic inference was carried out using Approximate Bayesian Computation implemented in the program Seqlib v. 1.0. Demographic models are depicted in Figure 3.3.

My analysis of population structure in *C. grandiflora* revealed three clearly delineated subpopulations, which fit well with geographic sampling localities. There is also a clear pattern of isolation-by-distance among these subpopulations. In *C. rubella* the correspondence between genetic and geographical distance is much less clear and I failed to observe a significant pattern of isolation-by-distance. Some of the *C. rubella* populations are genetically well defined while individuals in other populations are highly admixed (Fig. 3.2). Additionally, high levels of genetic diversity were observed in the outcrossing *C. grandiflora* with the relatively low F<sub>ST</sub> (0.14), while a low level of diversity and a very high F<sub>ST</sub> (0.53) is observed in the selfing *C. rubella*. These results are consistent with the mating system of each species. Outcrossers are expected to have high diversity while maintaining low fixation indices. Selfers, on the other hand, are expected to have much lower levels of diversity and high fixation indices. The seemingly erratic patterns
of population structure in *C. rubella* are likely due to multiple independent founding events and low migration rates.

**Figure 3.2** Bar plots from the clustering analysis conducted in STRUCTURE on the *C. rubella* (A) and the *C. grandiflora* (B) data sets. Each bar represents the estimated membership of each individual in the inferred genetic clusters and populations are delineated with black lines. Here, only results of the most likely number of genetic clusters (K) are presented.

In my analysis of demographic history in *C. grandiflora*, I accounted for the observed population structure by analyzing each of the three subpopulations separately. I found that 2 of the three subpopulations in this species do not deviate from the standard neutral model, while the third subpopulation shows some evidence for weak expansion (Table 3.1). This makes *C. grandiflora* nearly unique as, since likelihood and Bayesian methods of inferring

**Figure 3.3** Diagram of the different demographic models evaluated for the two Capsella species. The SNM is determined by a single parameter $\theta$, the population mutation rate, whereas the PEM has two parameters $\theta$ and $\alpha$, the exponential growth rate. The BNM is parameterized by 4 parameters, $\theta$; $T$, the time since the change of the population size; $f$, the size of the population during the bottleneck; $d$, the duration of the bottleneck. The ICM has three parameters; $\theta$; $T$, time since the change of the population size; $\theta_A$, the ancestral population mutation rate.
demographic history have been developed, most studied organisms have shown evidence of departure from the standard neutral model (Schmid et al. 2005; Heuertz et al. 2006; Pyhäjärvi et al. 2007; Ingvarsson, 2008; Ross-Ibarra et al. 2008; Platt et al. 2010), although studies in coniferous tree species from the Tibetan plateau often failed to detect departures from the standard neutral model (Li et al. 2010a; Li et al. 2010b). Analysis in *C. rubella* confirmed earlier predictions of a strong bottleneck in this species (Table 3.2). My results show an approximate 18-fold size reduction approximately 18.6 kya (5-35 kya). This is consistent with previous studies (Foxe et al. 2009; Guo et al. 2009), which found that *C. rubella* speciated from *C. grandiflora* in association with the loss of self-incompatibility and large amount of genetic diversity at or near the last glacial maximum (~20,000 ya, Clark et al. 2009).

**Table 3.1** Bayes factors of the demographic models tested against the standard neutral model in *C. grandiflora*. All models incorporate recombination and were tested separately in each genetic cluster identified by STRUCTURE.

<table>
<thead>
<tr>
<th></th>
<th>Group A</th>
<th>Group B</th>
<th>Group C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard neutral model</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Population expansion model</td>
<td>16.4</td>
<td>0.05</td>
<td>0.13</td>
</tr>
<tr>
<td>Extended bottleneck model</td>
<td>1.8</td>
<td>0.18</td>
<td>0.30</td>
</tr>
</tbody>
</table>

**Table 3.2** Bayes factors of the demographic models tested against the standard neutral model in *C. rubella*. Models do not consider recombination.

<table>
<thead>
<tr>
<th></th>
<th>C. rubella</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard neutral model</td>
<td>1*</td>
</tr>
<tr>
<td>Population expansion model</td>
<td>1*</td>
</tr>
<tr>
<td>Instant Change model</td>
<td>628</td>
</tr>
<tr>
<td>Bottleneck model</td>
<td>243</td>
</tr>
</tbody>
</table>

*Because zero samples were accepted for SNM and PEM models, an acceptance rate of 0.000002 was assumed for the purpose of calculating the Bayes Factors*
3.2.2 What sampling strategy is most appropriate for demographic studies? (Paper III)

What type of sampling scheme or strategy that is most appropriate for recovering the true history of a species has been much debated recently among population geneticists. The effect of pooling limited samples over genetically structured populations for demographic studies in particular has been a target of criticism and was recently tested by Städler et al. (2009). Building on previous theoretical works of Wakeley (1999, 2001), Wakeley and Aliacar (2001), Ray et al. (2003) and De and Durrett (2007) they used coalescent simulation and empirical data to highlight the potential problem of such a sampling scheme. They described and tested three different types of sampling strategy i) Local, where all individuals are taken from a single deme, ii) Pooled, where several individuals from several demes are pooled, and iii) Scattered, where a single individual is taken from each deme. Their simulated results under simple stepping-stone models and island models demonstrated that local and pooled sampling can lead to an excess of intermediate-frequency polymorphisms (positive Tajima’s D) relative to expectations in a panmictic population. This led to the suggestion that scattered sampling gives the best estimates of population history. Their empirical data seem to follow the trends seen in their simulated data, except that their sampling was too limited to include a scattered sample.

Figure 3.4 Boxplots of Tajima’s D ($D_T$) for different sampling strategies in C. grandiflora and C. rubella. The boxes range the upper and lower quartiles, the bold line indicates the median value and whiskers indicate 1.5 times the quartile range.
In my study I empirically test the broad applicability of their results using my *C. grandiflora* and *C. rubella* data and the stability of Städler et al. (2009)’s results using simulations under more complex population models. To do this I first sequenced 1-5 individuals in 5 and 6 new populations of *C. grandiflora* and *C. rubella*, respectively, and combined this data with data from paper II. From this new dataset I subsampled 4 replicates of each sampling scheme, local, pooled and scattered, for each species. Each replicate subsample contained 12 individuals. Local replicates consisted of 12 individuals originating from one population, pooled replicates consisted of three individuals from four populations, and scattered replicates consisted of 12 individuals, each from a different population. This results in a total of 24 subsampled datasets, 12 for each species. I then calculated diversity statistics and Tajima’s D and performed ABC analysis for all 24 subsamples. Second, we performed variations of Städler et al. (2009)’s simulations, altering the number loci, the number of demes and implementing two new models, one which simulates population growth within demes and one including hierarchical population structure.

The general conclusion of my study is that although sampling strategy can have an effect on some population genetic analysis, it does not have as strong an effect as proposed by some previous studies (Figure 3.4). The effect of sampling strategy is clearly over powered by the strong demographic signature in *C. rubella*, as evidenced by our ABC analysis inferring the instant size change model for all 12 subsamples over the 3 sampling strategies. For *C. grandiflora*, the demographic signal is much weaker, with probabilities over the models being similar for each subsample, and none conferring a

Figure 3.5 Tajima’s D as a function of the $4N_0m$ in an equilibrium stepping-stone model (without expansion). Simulations were run with Sampling V. 0.5. We assumed 20 demes and 16 loci, the number of loci used in the present study. See text for further details.
Bayes factor larger than 3. In such cases the simplest model is generally accepted, which in our case is the standard neutral model. So, even in species where demographic signatures are weak, sampling strategy does not necessarily alter the conclusions. One source of discrepancy between these results and the results of Städler et al. (2009) may be the number of loci used. My empirical data has 16 loci while Städler et al. (2009)’s simulations were conducted using 1000 loci. It may therefore be that the weak effects of sampling strategy are only detectable with a large number of loci (Figure 3.5).

A further concern is the simplicity of the expansion model used by Städler et al (2009), which model an instantaneous shift from an ancestral population to a subdivided population that likely exacerbates the differences between the different sampling schemes more than a progressive shift towards an n-island model would do (Figure 3.6). We made simple modifications to this model by adding first within deme expansion and second a hierarchical population structure and found rather striking alterations to the conclusions (Figure 3.7). By adding within deme expansion to the global expansion model, negative Tajima’s D values are obtained not only for scattered sampling strategies, but also for pooled and local strategies, under low migration and there is little difference on D at higher levels of migration (Fig 3.7A). Under the hierarchical model with global expansion, pooled and scattered sampling result in positive Tajima’s D while the local sampling strategy leads to values close to zero, for low migration rates. Under high migration, again, all three sampling strategies give similar results (Fig. 3.7B).

**Figure 3.6** Adapted from Städler et al. (2009) figures 1 and 2. Averages of Tajima’s D as a function of migration rates between demes in (A) an equilibrium stepping-stone model with I=100 islands and (B) a model of species wide expansion (β=10, τ=.2). Simulations were carried out with 100 demes without recombination. Every plotted point is based on 1000 independently generated data sets, and vertical lines indicate standard errors. Printed with permission from the Genetics Society of America.
Figure 3.7 Averages of Tajima’s D as a function of migration rates between demes in a model including both species-wide expansion ($\beta=10$, $\tau=2$) and growth within demes (A) or hierarchical structure (B). As in Fig. 2 in Städler et al. (2009) every plotted point is based on 1000 independently generated data sets, corresponding to 1000 unlinked loci; error bars represent standard errors. In A, the exponential growth rate within each deme was $\alpha=25$. We simulated a Wright island model with 100 demes without recombination. In B, the ancestral population splits into 100 demes at time $\tau=2$ and then the 100 demes into two groups of 50 demes each at time $\tau=1$. At each stage we assume a Wright island model with no recombination.

Thus, at high or intermediate levels of migration the estimation of Tajima’s D is not largely affected by the sampling strategy. The range of populations to which this could apply should be rather large as our intermediate and high migrations rates correspond to an $F_{ST}$ between 0 and 0.33 at equilibrium and $F_{ST}$, or $G_{ST}$, in this range is very common among outcrossers (Hamrick & Godt 1996; Nybom 2004; Nybom & Bartish 2000). Together, my empirical and simulated results suggest that Städler’s results are less broadly applicable than originally implied.
3.3 Autopolyploid speciation of \textit{Capsella bursa-pastoris} (Paper IV)

This study was a collaborative work with colleagues at the University of Toronto. Together we compiled a dataset with extensive sampling in all three \textit{Capsella} species. These individuals were sequenced at 14 nuclear loci. Using this dataset we developed a novel coalescent-based approach to determine if \textit{C. bursa-pastoris} has an allopolyploid or autopolyploid origin. These two models are distinguished from each other by the branch lengths between the two \textit{C. bursa-pastoris} genomes and the \textit{C. grandiflora} genome (Fig. 3.8). Under the autopolyploid model the branch lengths are expected to be equal between \textit{C. grandiflora} and both \textit{C. bursa-pastoris} genomes, while under the allopolyploid model the branch lengths leading to one of the \textit{C. bursa-pastoris} genomes is expected to be longer than the branch between \textit{C. grandiflora} and the other \textit{C. bursa-pastoris} genome. For our study we have designated the autopolyploid model as our null hypothesis and assigned \textit{C. bursa-pastoris} homeologs most distant to \textit{C. grandiflora} to the B genome and the others to the A genome, effectively biasing our analysis toward rejecting our null hypothesis.

Our analyses are based on summary statistic first introduced by Wakeley and Hey (Wakeley and Hey, 1997), which are the number of shared polymorphisms, the number of unique polymorphisms and the number of fixed differences between each of the four genomes, where the A and B genomes of \textit{C. bursa-pastoris} are treated separately. Since it has previously been determined that \textit{C. rubella} is recently derived from \textit{C. grandiflora} we chose not to include it in our analysis.
Our first analysis testing the null hypothesis was done by first estimating the parameters of an isolation-with-migration model using MIMAR (Becquet and Przeworski, 2007) and secondly by performing coalescent simulations using the parameters estimates from MIMAR. MIMAR analysis was conducted on *C. grandiflora* and the *C. bursa-pastoris* genomes in a pairwise fashion. The model implemented by MIMAR assumes a single ancestral population of size $N_a$ splitting into two descendant populations at time $t$, each having distinct sizes. We estimate a divergence time between the two *C. bursa-pastoris* genomes that is intermediate between the divergence times between *C. grandiflora* and *C. bursa-pastoris* A and *C. grandiflora* and *C. bursa-pastoris* B. Under the allopolyploid model the divergence times between *C. bursa-pastoris* A and B are expected to be equal to that between *C. grandiflora* and *C. bursa-pastoris* B. An intermediate estimate suggests that the partitioning of the homeologs is artificial and the time of divergence between *C. grandiflora* and both *C. bursa-pastoris* copies reflects an autopolyplloid event.

Simulated data under both models further supports the autopolyplloid model. Results from the simulated data under both models demonstrate that only fixed and shared sites are useful in distinguishing the two models. We calculated two further statistics based on these sites, namely the difference in both the number of fixed and the number of shared sites between *C. grandiflora* and *C. bursa-pastoris* A, on the one hand, and *C. grandiflora* and *C. bursa-pastoris* B on the other. Both these statistics in our empirical data departed significantly from the simulated values under allopolyploidy while they fitted well with the simulated values under autopolyplloid (Figure 3.9). Thus, we can not reject the null hypothesis.
Figure 3.9: Density distribution of the simulated values of the summary statistics under (A) autopolyploidy and (B) allopolyploidy. The left column gives the distribution of the mean of \textit{fix\_diff} over the fourteen genes, where \textit{fix\_diff} is the difference between the number of fixed sites of each of the homoeologs when it is compared to \textit{C. grandiflora}. The right column gives the same for \textit{shared\_diff}, the difference between the number of shared polymorphic sites of each of the homoeologs to when it is compared to \textit{C. grandiflora}. The thick vertical line is the observed value. P values are given in the upper right corner of each plot. See text for details.

The second analysis applied an Approximate Bayesian Computation (ABC) method implemented in Seqlib v 1.0. In this analysis we estimated the parameters of a two-split model including all three genomes. The two-split model represents three populations, populations 1, 2 and 3, each with independent sizes, which coalesces in a given order: populations 1 and 2 coalesce first and their common ancestor coalesce with population 3. The model estimates the sizes of each of the three populations, the timing of both coales-
cent events and the population sizes after each event. The model was tested with two configurations; first with the A and B genomes of *C. bursa-pastoris* designated as populations 1 and 2 and *C. grandiflora* designated as populations 3, representing the autopolyploid model, and second with *C. grandiflora* and *C. bursa-pastoris* A designated as populations 1 and 2 and *C. bursa-pastoris* B designated as population 3, representing the allopolyploid model. The program failed to converge under this second configuration, demonstrating the poor fit of the data to the allopolyploid model. But reasonable parameter estimates with clear modes were produced under the first configuration, showing that data fitted well to the autopolyploid model. Interestingly, the timing of the two coalescent events under the autopolyploid model are extremely similar, showing that the two *C. bursa-pastoris* genomes diverged from each other at nearly the same time that they diverged from *C. grandiflora*, supporting a scenario where *C. bursa-pastoris* speciated from *C. grandiflora* in association with polyploidization.
4 Conclusions

The work presented in this thesis is an important addition to the establishment of the *Capsella* genus as a model for molecular evolutionary studies and we have set up a broad collection of *Capsella* populations throughout all three species distributions for future studies.

Our analysis of the two diploid species, *C. grandiflora* and *C. rubella*, show that they differ strongly not only in their mating system but also in their demographic history and population structure. The current levels of diversity and population genetic structure in these species therefore reflects a mixture of past and present factors, demonstrating the importance of studying the effect of mating system in the context of demographic history.

We have further investigated the effect of sampling scheme on these types of studies and conclude that, in many cases, the effects are minimal compared with demographic effects. However, these effects do exist and will in some situations have an effect on demographic inference; therefore we do not suggest neglecting the sampling issues in future studies. Only a good coverage of the natural range, using both within- and between-population samples and a large number of loci are likely to lead to reliable inferences of species population structure and demographic history, and thereby to correct conclusions on adaptation and the evolutionary history of the species.

Our study of speciation in *C. bursa-pastoris* confirms the usefulness of coalescent-based approaches when studying the mode of origin of polyploids and suggest that *C. bursa-pastoris* is an autopolyplloid. While these results shed much light on the evolutionary origin of *C. bursa-pastoris* and establish the molecular phylogeny of the *Capsella* genus, little is still known about the extensive phenotypic changes that have occurred in both *C. bursa-pastoris* and *C. rubella*. Understanding the genomic context and underlying evolutionary forces that have promoted these changes will be of considerable interest in future studies.

Finally, our association mapping study in *C. bursa-pastoris* identified two genes harbouring mutations associated with the quantitative trait flowering time. In CRY1 these mutations are synonymous and are unlikely to be the causative mutations. In FLC, however, these mutations confer nonconsensus splice site alleles, which effect exon-5. Although it remains for functional studies to prove that these mutations cause early flowering, this study demonstrates that association studies can identify strong candidates of functional variation.
Att hitta tecken på naturlig selektion i genetiska data har varit ett centralt mål för molekylär evolutionsbiologi sedan ämnet etablerades och modellen som kallas ”the standard neutral model” (SNM) har blivit den dominerande nollhypotesen inom ämnet. Eftersom populationsstuktur, demografisk historia och selektion alla kan leda till avvikelser från SNM, är det viktigt att studier som söker efter tecken på selektion på molekylär nivå även tar hänsyn till populationers struktur och demografiska historia. Dessa studier är dessutom intressanta i sig, eftersom de ger insikt i viktiga evolutionära processer som artbildning och diversifiering. I detta arbete utför jag sådana studier i släktet Capsella.


Blomningstid är en mycket viktig kvantitativ egenskap som påverkas av naturlig selektion. Vi har genomfört en associationsstudie för blomningstid i C. bursa-pastoris, där vi söker efter samband mellan genotypisk och fenotypisk variation samtidigt som vi tar hänsyn till den neutrala populationsstrukturen. Vi hittade fem punktmutationer som kunde knytas till blomningstid i två olika gener: FLC (flowering locus C) och CRY1 (cryptochrome 1). Två av dessa mutationer finns i FLC och resulterar i förändringar i genens struktur som potentiellt kan ha funktionella konsekvenser. I CRY1 resulterar ingen av mutationerna i proteinförändringar, vilket tyder på att de snarare är kopplade till den mutation på vilken selektionen verkar, än att de själva är föremål för direkt selektion.

separata genom och dessa behandlas här separat. Alla de olika analyserna pekar på att *C. bursa-pastoris* är en autopolyploid som relativt nyligen uppstod från *C. grandiflora*.

I de två diploida arterna, *C. rubella* och *C. grandiflora*, använder jag sekvensdata för att undersöka den genetiska diversiteten, populationsstruktur och den demografiska historien, samt för att bedöma vilken effekt självbefruktning har på dessa egenskaper i *C. rubella*. I överensstämmelse med tidigare studier, tyder dessa data på att *C. rubella* nyligen har genomgått en demografisk flaskhals som sannolikt representerar flaskhalsen vid bildandet av *C. rubella* från *C. grandiflora*. Den genetiska mångfalden är lägre hos *C. rubella* än hos *C. grandiflora* och en större del av den genetiska diversiteten orsakas av differentiering mellan populationer men vi hittar ingen korrelation mellan populationsdifferentiering och geografiskt avstånd. Detta mönster är troligen resultatet av en kombination av arternas demografiska historia och parningssystem. I *C. grandiflora* är populationsstrukturen tydligt geografiskt strukturerad och vi hittar en korrelation mellan geografiskt avstånd och genetisk differentiering. Populationerna i denna art avviker inte starkt från SNM, vilket verkar vara ovanligt bland de flesta andra arter som studerats.

Slutligen undersökte jag hur resultatet av demografiska analyser kan påverkas av provtagningsstrategin, dvs hur prover väljs ut från en arts utbredningsområde. De potentiella fel som provtagningsstrategin kan orsaka i populationsgenetiska analyser har nyligen diskuterats av flera författare. Omfattande provtagningsstrategi, både inom och mellan populationer av *C. rubella* och *C. grandiflora*, gjorde det möjligt för oss att empiriskt testa effekten av provtagningsstrategin på demografiska analyser. Vi kompletterade de empiriska analyserna med simuleringar och drar slutsatsen att effekten av provtagningsstrategin i många fall är svag jämfört med effekten av demografisk historia, och att den därför i många fall inte kommer att påverka slutsatserna från dessa analyser. Trots detta kan provtagningsstrategin i vissa fall leda till felaktiga slutsatser och bör därför noga övervägas i demografiska studier.
6 Acknowledgements

Many people have helped me along the way in this work and I very much appreciate all the support.

First, I would like to start by thanking my supervisors. Anna, for giving me this opportunity and helping me succeed. I would also like to thank Martin who has also helped along every step of the way. And I thank Stephen Wright for hosing my research visit and many follow-ups to Toronto.

Second, I would like to thank everyone at Funkgen for many interesting conversations and random help in the lab. I would like to especially thank Rose-Marie and Kerstin for making everything go smoothly, and Sofia and Anna K, for lots of fun boardgame nights. Tanja, for helping me so much in the beginning. And several students who did work within my project.

Finally, I would like to thank all my friends and family for much love and support. Especially Sandra, Sara and Emma for much needed girl talk and nights out. Adriana and Gustaf for always being just next-door when I needed them. Paul for laughs and gossip. Mom and Dad, Meg and Greg, for always supporting me in my choices even when they stressed me out and took me far way. And my John of course, for always being so so understanding.

The work presented in this thesis has been supported by: Erik Philip-Sörensens Stiftelse, Nilsson-Ehlefonden,, Carl Tryggers stiftelse, FORMAS, Foundations of the Royal Swedish Academy of Sciences, and several foundations at Uppsala University
References


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 725

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)