OULU 2017
A 701

ACTA UNIVERSITATIS OULUENSIS

Tiina Mattila

POST-GLACIAL COLONIZATION, DEMOGRAPHIC HISTORY, AND SELECTION IN *ARABIDOPSIS LYRATA*

GENOME-WIDE AND CANDIDATE GENE BASED APPROACH

UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF SCIENCE

ACTA UNIVERSITATIS OULUENSIS
A Scientiae Rerum Naturalium 701

TIINA MATTILA

POST-GLACIAL COLONIZATION, DEMOGRAPHIC HISTORY, AND SELECTION IN *ARABIDOPSIS LYRATA*

Genome-wide and candidate gene based approach

Academic dissertation to be presented with the assent of the Doctoral Training Committee of Health and Biosciences of the University of Oulu for public defence in the Arina auditorium (TA105), Linnanmaa, on 10 November 2017, at 12 noon

UNIVERSITY OF OULU, OULU 2017

Acta Univ. Oul. A 701, 2017

Supervised by
Professor Outi Savolainen
Docent Tanja Pyhäjärvi

Reviewed by
Doctor Maud Tenaillon
Associate Professor Thomas Bataillon

Opponent
Professor Peter Tiffin

ISBN 978-952-62-1708-6 (Paperback)
ISBN 978-952-62-1709-3 (PDF)

ISSN 0355-3191 (Printed)
ISSN 1796-220X (Online)

Cover Design
Raimo Ahonen

JUVENES PRINT
TAMPERE 2017

Mattila, Tiina, Post-glacial colonization, demographic history, and selection in *Arabidopsis lyrata*. Genome-wide and candidate gene based approach
University of Oulu Graduate School; University of Oulu, Faculty of Science
Acta Univ. Oul. A 701, 2017
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

Demographic history and natural selection are central forces shaping the genetic diversity of populations. Knowledge on these forces increases understanding of processes shaping genetic variability of populations. In this PhD thesis I investigated demographic history and selection in multiple populations of *Arabidopsis lyrata*, an outcrossing herbaceous plant species of the Brassicaceae family. Due to its wide distribution in the temperate and boreal regions, *A. lyrata* serves as a good model system to study population genetic consequences of colonization of northern latitudes. The first aim of this study was to characterize the demographic and colonization history of the species using site frequency spectra estimated from whole-genome diversity data. Another aim was to detect genetic loci targeted by recent selective sweeps at genome-wide scale as well as at candidate flowering time genes. Patterns of genome-wide selection at linked sites (linked selection) were also compared between populations of *Capsella grandiflora* and *A. lyrata* with contrasting demographic histories.

Evidence for strong effective population size decline in the past few hundred thousand years was detected in *A. lyrata* populations species-wide. This study also suggests recent Scandinavian colonization from an unknown refugium, distinct from the Central European source population. Selection analyses revealed loci targeted by positive selection in two Scandinavian lineages after the recent population split as well as selective sweeps in flowering time genes in the colonizing populations. In comparison with the studied *C. grandiflora* population, the Norwegian *A. lyrata* population had weaker purifying selection and no evidence for reduction of diversity around genes was found. This thesis offers novel information on species colonization history and its genome-wide effects, which is important for understanding the framework of local adaptation.

Keywords: colonization history, demographic history, genetic diversity, linked selection, selective sweep, site frequency spectrum

Mattila, Tiina, Mattila, Tiina, Jääkauden jälkeinen kolonisaatio, demografinen historia ja valinta idänpitkäpalon (*Arabidopsis lyrata*) populaatioissa. Koko perimän laajuinen ja kandidaattigeeniperusteinen lähestymistapa

Oulun yliopiston tutkijakoulu; Oulun yliopisto, Luonnontieteellinen tiedekunta

Acta Univ. Oul. A 701, 2017
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Populaation demografinen historia ja luonnonvalinta ovat keskeisiä populaation perinnöllisen muuntelun muokkaajia. Näiden tekijöiden tutkimus on tärkeää eliöiden sopeutumisen ymmärtämiselle. Tässä väitöskirjassa tutkin demografista historiaa ja valintaa monivuotisen ristisiittoiseen ruohovartisen Brassicaceae-heimon kasvilajin idänpitkäpalon (*Arabidopsis lyrata*) useissa eri populaatioissa. Idänpitkäpalko on erinomainen mallilaji pohjoiseen ympäristöön sopeutumisen tutkimukseen, koska sen toisistaan eristäytyneet paikalliset populaatiot ovat levittäytyneet laajalle boreaalisella ja lauhkealla ilmastovyöhykkeellä. Tutkimuksen tarkoituksena oli luonnehtia populaatioiden demografista historiaa ja kolonisaatioreittejä käyttäen koko perimän laajuisesta muunteluaineistosta estimoituja alleelifrekvenssispektrejä. Lisäksi koko perimän laajuista aineistoa sekä kukkimisaikaa ohjaavien geenien sekvenssejä käytettiin positiivisen luonnonvalinnan merkkien tunnistukseen. Genominlaajuista kytkeytynyttä valintaa vertailtiin toiseen ristisiittoiseen Brassicaceae-heimon lajin *Capsella grandifloran* populaatioon, jonka demografinen historia poikkeaa huomattavasti tutkituista idänpitkäpalon populaatioista.

Tutkimuksessa havaittiin, että kaikissa tutkituissa idänpitkäpalon populaatioissa tehollinen populaatiokoko oli pienentynyt viimeisen muutaman sadantuhannen vuoden aikana. Kolonisaatiohistorian tarkastelu osoitti, että idänpitkäpalon skandinaaviset populaatiot ovat todennäköisesti peräisin keskieurooppalaisesta refugiosta erillisestä läntisestä refugiosta. Skandinavian kolonisaation yhteydessä vaikuttaneen positiivisen luonnonvalinnan merkkejä havaittiin useissa eri genomin osissa sekä erityisesti valojaksoa mittaavissa geeneissä. Tämä kertoo erilaisiin valojaksoihin sopeutumisen tärkeydestä skandinaavisen kolonisaation yhteydessä. Verrattuna tutkittuun *C. grandifloran* populaatioon, idänpitkäpalolla puhdistavan valinnan havaittiin olevan heikompaa ja muuntelun vähenemistä geenien ympärillä ei havaittu. Tämä tutkimus tarjoaa uutta tietoa Skandinavian kolonisaatiohistoriasta ja sen genominlaajuisista vaikutuksista. Tutkimuksessa tuotettua tietoa voidaan hyödyntää paikallisen sopeutumisen ymmärtämisessä.

Asiasanat: alleelifrekvenssispektri, demografinen historia, kolonisaatiohistoria, kytkeytynyt valinta, perinnöllinen muuntelu, valinnan pyyhkäisy

To my family

Acknowledgements

First, I would like to thank my supervisor Outi Savolainen for offering me an opportunity to work with this interesting PhD project. I would also like to thank my supervisor Tanja Pyhäjärvi for being such a friendly, supporting and practical mentor. You have both been amazing role models and I admire your skills and knowledge.

I further like to acknowledge the collaborators of my thesis work. The expertise and knowledge of Esa Aalto, Helmi Kuittinen, Anne Niittyvuopio, Susanna Piltonen, Tuomas Toivainen, Jaakko Tyrmi, Benjamin Laenen, Tuomas Hämälä and Tanja Slotte have been crucial for the success of my PhD thesis. Special thanks to Jaakko and Benjamin for the practical help and support.

This work was financially supported by the Finnish Population Genetics Doctoral Programme and Emil Aaltonen foundation. CSC IT Center for Science and Uppsala Multidisciplinary Center for Advanced Computational Science have provided computational resources and support for the thesis work. Finnish Population Genetics Doctoral Programme, University of Oulu Graduate School (UniOGS) and Oulun Luonnonystäväin Yhdistys are acknowledged for travel grants enabling me to attend several international courses and conferences during my studies. The pre-examiners, Maud Tenaillon and Thomas Bataillon, are acknowledged for the careful revision of my thesis and encouraging comments.

A warm thanks to former and current members of the plant genetics group for useful discussion and advice related to my thesis work. It has been great to work with people sharing the same interest and passion to plant population genetics. Special thanks go to Ulla Kemi for travelling with me to Plech in 2011 for collecting seeds with only a few days' notice, Soile Alatalo for proficient laboratory assistance and Tuomas Hämälä for kindly offering unpublished data for the analyses. I would also like to thank my colleagues and friends Heidi Aisala-Aalto, Jesper Bechsgaard, Sonja Kujala and Lumi Viljakainen who have always been ready for scientific discussion and to give their advice.

I greatly appreciate the help and support my family and friends have offered. Special thanks to my brother Kalle for helping me with various Python scripting related issues. I am also very grateful to my parents for always being there for me and my family. Finally, I want to express my deepest gratitude to Nooa and Sami. Thank you for always supporting me and accepting me as I am.

Oulu, November 2017 Tiina Mattila

Abbreviations

θ	population mutation rate
θ_W	Watterson's theta
π	nucleotide diversity
μ	mutation rate
BM	bottleneck model
CI	confidence interval
DAF	derived allele frequency
DM	decline model
GM	growth model
HWE	Hardy–Weinberg equilibrium
LGM	last glacial maximum
MLHKA	maximum likelihood Hudson–Kreitman–Aquadé test
MRCA	most recent common ancestor
N_e	effective population size
PBS	population branch statistics
SFS	site frequency spectrum
SNM	standard neutral model
tMRCA	time since most recent common ancestor

Original publications

This thesis is based on the following publications, which are referred throughout the text by their Roman numerals:

I Mattila TM, Tyrmi J, Pyhäjärvi T, Savolainen O (2017) Genome-wide analysis of colonization history and concomitant selection in *Arabidopsis lyrata*. Mol Biol Evol 34(10): 2665–2677.

II Mattila TM, Aalto EA, Toivainen T, Niittyvuopio A, Piltonen S, Kuittinen H, Savolainen O (2016) Selection for population-specific adaptation shaped patterns of variation in the photoperiod pathway genes in *Arabidopsis lyrata* during post-glacial colonization. Mol Ecol 25(2): 581–597

III Mattila TM**,** Laenen B, Hämälä T, Savolainen O, Slotte T (2017) The genome-wide impact of selection in two outcrossing Brassicaceae species with contrasting demographic history. Manuscript.

Author contributions

Paper	Study design	Data collection	Data analysis	Manuscript preparation
I	**TM**, OS	**TM**	**TM**, JT	**TM**, TP, OS
II	AN, HK, OS	**TM**, AN, TT, SP, EA	**TM**, EA	**TM**, AN, EA, OS
III	**TM**, BL, TH, OS, TS	**TM**, TH, BL	**TM**, BL	**TM**, BL, TS

Anne Niittyvuopio (AN), Susanna Piltonen (SP), Tuomas Toivainen (TT), Esa Aalto (EA), Tanja Pyhäjärvi (TP), Helmi Kuittinen (HK), Jaakko Tyrmi (JT), Outi Savolainen (OS), Benjamin Laenen (BL), Tuomas Hämälä (TH), Tanja Slotte (TS), **Tiina Mattila** (**TM**)

Contents

1 Introduction

Heritable genetic variation existing in a population is the basis for populations' ability to evolve over time (Darwin 1859), which is crucial for species survival in a spatially and temporally heterogeneous environment. Genetic variation, ultimately produced by mutation, persists in a population for a period of time after which it becomes either lost (only the ancestral form exists in the population) or fixed (only the derived variant exists in the population). If the carrier of a new mutant has a fitness difference in comparison with the carriers of the other alleles, natural selection can either decrease (negative selection) or increase (positive selection) the frequency of the mutant allele deterministically. In finite populations, allele frequencies fluctuate through time, even in the absence of any selection, due to the random sampling of gametes in each generation, a phenomenon known as genetic drift. The strength of the drift effect is dependent on the size of the population (effective population size, N_e) (Fisher 1930, Wright 1931). Further, selection is also influenced by the effective population size, since in small populations, selection is not efficient enough to favor alleles with small fitness effect and they behave as expected under neutrality (Ohta 1973). Hence, to comprehend the molecular evolution of populations it is important to understand the neutral and selective forces acting on populations.

This thesis aimed at studying the neutral and selective processes shaping genetic variation in the perennial outcrossing herb *Arabidopsis lyrata* with a focus on demographic and selection associated with post-glacial Scandinavian colonization. The patterns of genome-wide selection were further contrasted with the data from another outcrossing herb *Capsella grandiflora*. The ultimate goal was to make inference on the neutral and selective processes acting on these species and understand these in the context of the species history.

1.1 Genetic diversity under neutrality

To be able to make inference on the genetic diversity studied in real populations it is useful to have a theoretical model describing the sequence evolution against which real data can be tested (Wakeley 2008). Although several population genetics models exist, one widely used paradigm in population genetics is the coalescent theory that was formulated in 1980s and 1990s by several scientists, most notably John Kingman (Kingman 1982a, Kingman 1982b), Richard Hudson (Hudson 1991) and Fumio Tajima (Tajima 1983). In the coalescent model, each sampled gene copy

has a random parent in the previous generation. If two gene copies descend from the same parent they are said to coalesce (Fig. 1). The tracking of population ancestry is done backward in time until all the linages have coalesced, i. e., have found the most recent common ancestor (MRCA). In a population of *N* diploid individuals, the coalescent probability of a pair of gene copies is 1 / *2N*, since the number of possible parents in the previous generations is *2N* (Nielsen & Slatkin 2013). For a sample of n haploid genes the expected time to most recent common ancestor (tMRCA) measured in $2N$ generation is

$$E[tMRCA] = \sum_{k=2}^{n} \frac{2}{k(k-1)} \quad (1)$$

If n = 2 the coefficient of *2N* is 1 and the expected coalescent time for two gene copies is simply *2N* generations (Hein *et al.* 2004, Nielsen & Slatkin 2013). The derivation of some essential results considering genetics of populations is more straightforward in coalescent framework in comparison with forward time models. Furthermore, only a sample from the current population is needed to characterize the population properties, making simulation under coalescent framework effective, and hence it has become a widely used tool in population genetics (Wakeley 2008).

For understanding the variation existing in a population, mutation needs to be incorporated in the model. Assuming a mutation rate per generation of μ and infinite sites model (each mutation hits a new position), the expected number of pairwise difference between two haplotypes is $4N\mu$, since the expected waiting time for two lineages to coalesce is *2N*. Hence, populations with larger *N* harbor higher diversity, as the expected coalescence time between two pairs of haplotypes is longer (Nielsen & Slatkin 2013: 40–41).

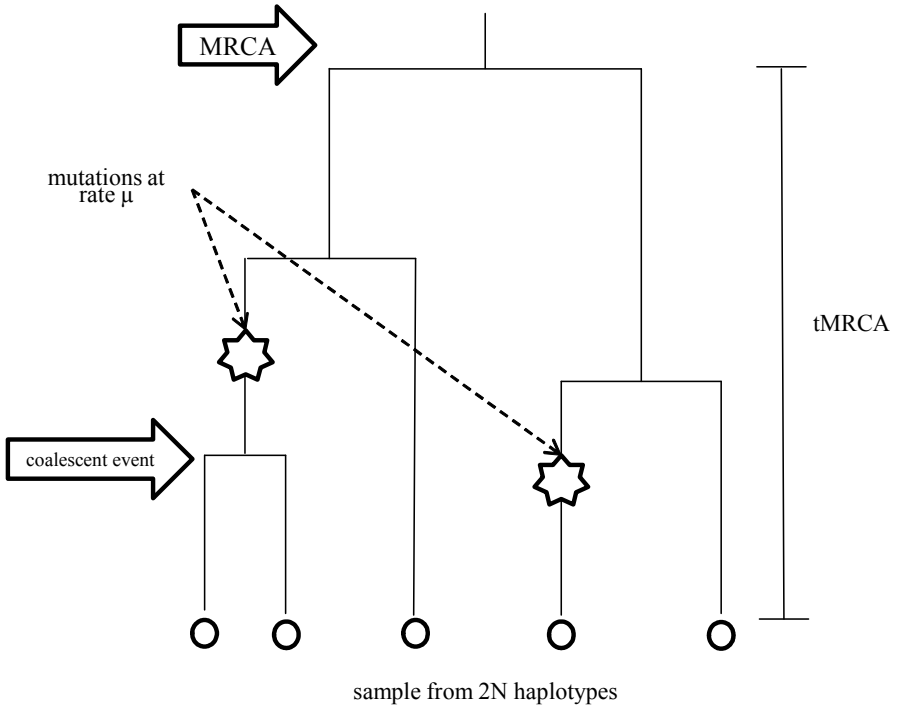

Fig. 1. Example coalescent tree of 5 haplotypes, 2*N* = haploid effective population size, MRCA = most recent common ancestor, tMRCA = time since most recent common ancestor.

The parameter $4N\mu$ (population mutation rate), often denoted by θ, is of great interest in population genetics since it informs on the amount of genetic variation in a given population. Several methods for directly estimating this from DNA sequence polymorphism data exist. Two widely used estimators are π (Nei & Li 1979, Tajima 1983) and Watterson's θ (θ_W) (Watterson 1975), which can be estimated as follows (Equations 2, 3)

$$\pi = 2\sum_{i<j}\frac{\hat{d}_{ij}}{[n(n-1)]} \quad (2)$$

where, $\hat{d}_{ij}$ is the average number of nucleotide differences between a pair of sequences i and j and n is the number of sequences sampled.

$$\theta_W = \frac{S}{\sum_{i=1}^{n-1}\frac{1}{i}} \quad (3)$$

where S is the number of segregating sites. The denominator is the harmonic number for (n – 1).

The different θ estimators are sensitive to the variants in different frequencies. Under the standard neutral model these estimators should give approximately equal results but if these assumptions are violated, e.g. due to recent change in the effective population size, this is not necessarily the case. Hence, by comparing the difference in the values of θ estimators, it is possible to test whether a given population is in mutation-drift equilibrium. One commonly used such so called neutrality test is based on D statistics (Tajima 1989), which compares π and $\underline{\theta_W}$. This statistics, known as Tajima's D, is often used as a summary statistics describing the skew in site frequency spectrum (SFS) in comparison to neutrality. It is also possible to use the full SFS, which is a vector of counts of all variants of different frequency classes (see Fig. 2A lower panel), to estimate population parameters such as changes in the effective size of a population (Gutenkunst *et al.* 2009, Excoffier *et al.* 2013, Liu & Fu 2015) as well as estimate selection (Nielsen *et al.* 2005, Keightley & Eyre-Walker 2007). Depending on whether the ancestral allele of each site can be deduced, SFS can be either folded (no ancestral information) or unfolded (ancestral information). The unfolded SFS contains more information but the ancestral inference is challenging (Baudry & Depaulis 2003, Hernandez *et al.* 2007) and hence depending on the situation both forms can be used.

The two major factors that can violate neutrality assumptions are selection and population size change, with both affecting diversity in similar ways (Tajima 1989). Hence, for selection inference, understanding of the demographic background is essential. On the other hand, investigating the neutral allele frequency distribution per se may be of interest for example from a biogeographical or a conservation point of view. Using coalescent simulations it is easy to demonstrate the effect of demographic history on genetic variation, such as θ estimates and SFS. Here the effect of four basic single population demographic models is compared with respect to the summary statistics described above (Fig. 2). The details used for the simulations can be found in Appendix 1. In general, diversity is reduced if the mean effective size over time is smaller (Fig. 2A & B). In the SFS, the frequencies of high and intermediate frequency variants are increased after population decline (DM and BM) while the opposite occurs after population expansion (Fig. 2A). π is higher than θ_W and mean Tajima's D is positive after population size decline while θ_W is higher after population expansion and mean D negative (Fig. 2B & 2C). One

notable difference in diversity under different demographic scenarios is the high variance in D after population decline (Fig. 2C).

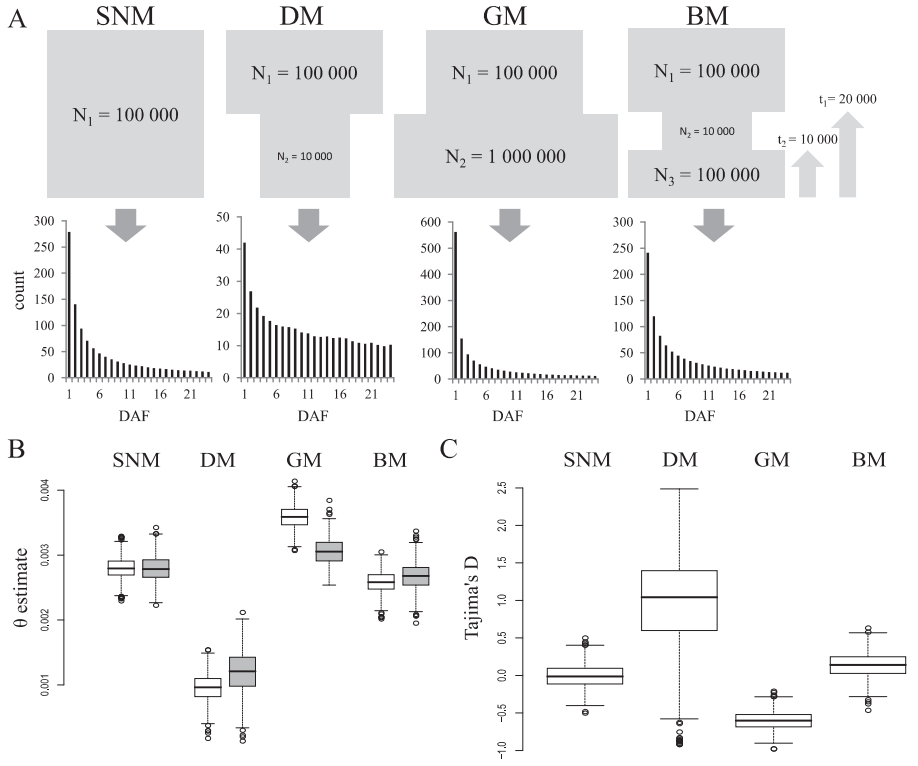

Fig. 2. Patterns of genetic diversity under four demographic scenarios SNM = standard neutral model, DM = decline model, GM = growth model, BM = bottleneck model. A) Parameters of the demographic models and distribution of derived allele frequency (DAF) counts in the simulated data averaged over 1000 independent replicates, 100 Mb region and 25 individuals with mutation and recombination, B) Watterson's θ (white) and π (grey) estimated from the simulated data, and C) Tajima's D estimated from the simulated data.

1.2 Population structure and distribution of allele frequencies

The basic population genetics models often assume random mating as does the standard coalescent model. However, in reality this assumption is often unrealistic. The individuals of a species rarely mate equally likely with all the possible mates

due to for example geographical factors or mating system (Hein *et al.* 2004) and hence allele frequencies between the sub-groups may deviate considerably from the neutral expectation. In a diploid population the deviation from random mating can be tested using the Hardy-Weinberg principle. In a random mating diploid population the expected genotype frequencies in a di-allelic locus with alleles A and a with frequencies p and q under Hardy-Weinberg equilibrium (HWE) (Hardy 1908, Weinberg 1908 according to Kimura 1983) are

$$\begin{aligned} f_{AA} &= p^2 \\ f_{Aa} &= 2pq \\ f_{aa} &= q^2 \end{aligned} \tag{4}$$

In the case of population structure, the allele frequencies may differ between sub-populations and the total population deviates from the HWE while HWE holds within sub-populations. In turn, inbreeding decreases the number of heterozygotes within population. A set of statistics, F-statistics (Wright 1951), is a common method for quantifying differentiation of sub-populations and inbreeding within sub-populations. Of these statistics the measure quantifying between population differentiation is F_{st}, which measures what proportion of the existing variation is explained by the between population variance in comparison with the total variance. Hence, the scale of F_{st} is between 0 and 1, low F_{st} indicating low levels of population genetic differentiation (Holsinger & Weir 2009).

In many cases, the structure of the populations studied is of interest but it may not be known a priori. Assuming HWE within sub-populations, it is possible to deduce the number of sub-populations from genotype frequencies of a set of genetic loci (Pritchard *et al.* 2000). With the population structure information it is possible to evaluate the genetic relationship and the degree of isolation between populations which can further be used as background information in genetic mapping studies and conservative decision making.

In the coalescent framework population structure is often modelled as a group of district random mating units (demes) which exchange migrants at arbitrary rate and merge at some point forming a single random mating population (Hudson 1991, Excoffier *et al.* 2000). The coalescence probability between genes sampled from different demes is 0 unless genes are migrants between demes. If isolation is strong enough, independent stochastic processes occur within each deme leading to differentiation of gene frequencies between demes. The amount of neutral genetic differentiation between populations is determined by the time since isolation, the

amount of genetic drift and gene flow (Hudson 1991, Nielsen & Slatkin 2013). This is because the probability of the first between deme coalescence is decreased, if the divergence time is increased. Second, in large populations the probability of coalescence is smaller and hence the tMRCA is higher, more likely exceeding the population split time. Finally, since coalescnce can only occur within a deme, the rate of gene flow is inversely correlated with population differentiation (Hein *et al.* 2004).

1.3 Effects of selection on genetic variation

To be able to make inference of selection acting on a given genetic locus it is important to understand how the different forms of selection influence the amount and distribution of genetic variation in addition to the neutral processes described above. Estimating the relative role of selection shaping the genetic diversity in natural population has been of great interest ever since the diversity data started to accumulate (Kimura 1983).

Natural selection is due to mean difference in the reproductive output (fitness) of individuals with different genotypes. In the case of directional selection alleles are fixed faster than under genetic drift alone (Nielsen & Slatkin 2013) eliminating the variation at the selected locus. This can be either due to lower (purifying selection) or higher (positive selection) fitness of the carriers of a new variant, compared to the ancestral allele. On the other hand, some forms of selection cause maintenance of variation for longer than expected under drift (balancing selection), for example due to heterozygote advantage or frequency dependent selection (Charlesworth 2006). Due to non-random association of linked loci (linkage disequilibrium), loci near the selected site are also affected by the selective event and hence selection leaves a distinguishable signal on the patterns of genetic variation. This signal decreases with increasing distance from the selected site due to recombination. The linked selection effect can be due to purifying selection when the selection at linked sites is known as background selection (Charlesworth *et al.* 1993, Charlesworth 1994) or selection for advantageous mutation (an effect known as hitchhiking or selective sweep) (Maynard Smith & Haigh 1974, Kaplan *et al.* 1989).

The role of linked selection affecting the patterns of genetic variation is of great interest in population genetics (Gillespie 2001, Cutter & Payseur 2013). Early empirical evidence supporting large scale contribution of selection on genome-wide variation is the observation of positive correlation between neutral diversity

and recombination rate in *Drosophila melanogaster* (Begun & Aquadro 1992) which is expected under linked selection model. Further, several lines of evidence suggest that a large proportion of new mutations have negative effects on survival (Kimura 1977, Sanjuán *et al.* 2004, Eyre-Walker & Keightley 2007) and hence a large proportion of the genome is likely affected by background selection. On the other hand, the relative role of positive selection is more variable between species (Bustamante *et al.* 2005, Gossmann *et al.* 2010).

1.3.1 Detecting loci affected by positive selection

From an ecological point of view, detecting loci targeted by positive selection is often of interest (Sabeti *et al.* 2006) offering possibility to understand the molecular basis of evolutionary change. These efforts have been greatly facilitated in the past 10 years by the development of new DNA sequencing methods, as the whole genomes can be scanned without knowledge of the underlying adaptive traits (Siol *et al.* 2010).

The basic population genetics model for positive selection (the hitchhiking model by Maynard Smith & Haigh 1974) considers the effect of fixation of a single advantageous new mutation. Due to positive selection favoring a new variant, it increases in frequency rapidly sweeping out the variation in this locus. Due to linkage disequilibrium, selection affecting a single site also has an effect on its nearby markers, with a decreasing effect as a function of distance from the selected site (Maynard Smith & Haigh 1974). In addition to the decreased variation, the sweep also increases the frequency of rare alleles (Tajima 1989, Braverman *et al.* 1995) and produces a distinguishable pattern of linkage disequilibrium around the selected site (Kim & Nielsen 2004, McVean 2007). After a selective sweep, the selective signature persists in the population for a certain time (Biswas & Akey 2006, Hohenlohe *et al.* 2010) with signatures disappearing at different rates (Przeworski 2002), setting limits to how recent selection can be detected with different method.

One important aspect of selection inference is to separate locus specific selective signals from neutral processes. For this purpose studies often use a null distribution derived from demographic simulations (e.g. Ometto *et al.* 2005) or from background genomic control set (e.g. Nielsen *et al.* 2005) as a neutral model. More recent methods estimate demography and selection simultaneously (e.g. Sheehan & Song 2016) overcoming some complications related to the two-step (demography first, selection second) approach (Li *et al.* 2012).

A classic example of a single locus showing strong evidence for a positive selective sweep is the *LCT* gene in humans, which has been detected in selection scans with various combinations of methods (Bersaglieri *et al.* 2004, Nielsen *et al.* 2005, Voight *et al.* 2006, Sabeti *et al.* 2007). Certain alleles of *LCT* gene are associated with human lactase persistence phenotype (Enattah *et al.* 2002) and the spatial and temporal distribution of these variants suggests that the variants were positively selected in cultures with extensive dairy consumption late after the Neolithic revolution (Allentoft *et al.* 2015). Another well described example is the *Agouti* gene in Nebraska Sandy Hill deer mice (*Peromyscus maniculatus*) (Linnen *et al.* 2009, Pfeifer *et al.* 2017). Variation in the *Agouti* locus is associated with light coat coloration in deer mice and it is hypothesized that selection has favored the light form of this gene in sandy environments for camouflage.

1.3.2 Adaptation from standing genetic variation and polygenic adaptation

The original sweep theory considers a single adaptive new mutation that is advantageous and eventually becomes fixed. However, it is also possible that selection starts favoring a variant already existing in the population (standing variation) and the outcome of this form of selection is known as a soft sweep first introduced by Hermisson & Pennings (2005). Such adaptation can occur if certain previously neutral (conditional neutrality) or deleterious allele (antagonistic pleiotropy) becomes adaptive due to for example environmental change or a new pathogen pressure. In addition, many traits under selection are polygenic and selection affecting a single locus may be relatively weak, causing only small allele frequency differences rather than fixation (Stephan 2016). The effect of soft sweeps and polygenic selection is much less pronounced in comparison with the classic (hard) sweep. Hence the power for detecting such loci may be reduced using the standard neutrality tests (Messer & Petrov 2013). Some statistics may however be more efficient in finding such selection, for example Garud *et al.* (2015) studied the haplotype frequencies under soft and hard sweeps and showed that the simple haplotype frequency statistics were efficient in detecting soft sweeps and also separating hard and soft sweeps. Methods detecting polygenic adaptation are also under active development (Pritchard & Di Rienzo 2010, Le Corre & Kremer 2012, Stephan 2016). Combining multiple aspects of the data, for example genetic mapping and allele frequencies (Berg & Coop 2014) can also be used in such inference.

1.3.3 Spatially varying selection

Due to environmental heterogeneity selection pressures often vary spatially which may cause local adaptation between populations of the same species. In such case, in addition to within population patterns, variation between populations can help to detect selection. Correlation of heritable phenotypic variation with environmental factors is often taken as the first evidence for adaptation to environmental gradient. For example human pigmentation correlation with UV radiation gives evidence that selection has favored lower pigmentation in high latitudes with low UV radiation levels while the higher pigmentation is advantageous in low latitudes with high UV radiation levels (Sturm & Duffy 2012). Another example comes from coniferous trees, where growth cessation in common garden experiments is strongly correlated with latitude of origin suggesting adaptation to the local photoperiodic environment during the growth season (Savolainen *et al.* 2007, Chen *et al.* 2014). At the level of genetic variation, spatially varying selection increases differentiation between populations and can produce a correlation between gene frequencies and environmental factors (Lewontin & Krakauer 1973, Coop *et al.* 2010). Variety of methods aiming at detecting such adaptation exists, differing in how they model the background demography and genetic structure (Foll & Gaggiotti 2008, Excoffier *et al.* 2009, Bonhomme *et al.* 2010, Coop *et al.* 2010).

1.4 The geological framework of Northern European populations

Adaptive as well as neutral population processes can only be fully understood in the context of the past environmental conditions. The fluctuation in earth's climate in the recent past has heavily influence the species communities and distribution ranges especially in regions undergoing glaciation cycles (Hewitt 2000) which has affected the adaptive environment of the organisms. The glaciation history of Northern Europe is especially well characterized and only approximately 19 000–22 000 years ago during the last glacial maximum (LGM) (Yokoyama *et al.* 2000) big parts of northwestern Europe were covered with ice (Hughes *et al.* 2016). Hence, all the current communities in this region have formed relatively recently, after the LGM, which has influenced the patterns of variation and structure of the current populations (Davis & Shaw 2001). In general, after range expansion the colonizing population is assumed to harbour less variation due to founder effect (Hewitt 2000).

The colonization patterns of European biota after the LGM have been divided into three main categories a) colonization of western and northern Europe from the

Balkan region b) colonization from multiple southern refugia c) colonization from south-western Europe and from the east with hybrid zones in Scandinavia (Hewitt 1999, Hewitt 2000). Evidence for several glacial refugia has been found for example in Norway spruce (Tollefsrud *et al.* 2008). Alternatively, it is also likely that colonization was a range shift of a large more or less continuous population rather than colonization from specific refugia (Davis & Shaw 2001). Fossil pollen sediment samples and high-throughput genetic information have shed light on the details of colonization and the composition of change in the species communities through time in general. For example Seppä *et al.* (2002) studied pollen records from Finnish peat sediments and characterized the post-glacial dynamics of treeline. Further, (Willerslev *et al.* 2014) investigated ancient DNA from sediment as well as megafauna gut and coprolite samples and described the vegetation history of some parts of the Arctic over the last 50 000 years. Interestingly, some studies have suggested survival within the glaciated regions, for example of several tree species and arctic-alpine pioneer plant species (Kullman 2002, Westergaard *et al.* 2011, Parducci *et al.* 2012b) although whether the evidence is conclusive remains under active discussion and research (Birks *et al.* 2005, Birks *et al.* 2012, Parducci *et al.* 2012a, Tzedakis *et al.* 2013).

1.5 Timing of developmental events in plant adaptation and its molecular control

In plants, to match the timing of developmental processes, such as flowering, germination and growth, with the favorable season of a year is of adaptive importance (Donohue 2003, Putterill *et al.* 2004, Donohue 2005, Amasino 2010, Anderson *et al.* 2011). Plants track information on the optimal growing conditions to a large degree by daylength (Garner & Allard 1920), which is a good predictor of average weather conditions. Other ambient factors such as temperature also modulate the plant responses. As the seasonal processes vary between sites (e.g. latitudinal variation) and through time (climate change) it is likely that spatial and temporal changes are needed for tracking the optimal growing season (Stinchcombe *et al.* 2004, Kubota *et al.* 2015). For example, as predicted, the populations from high latitudes require longer days to grow while for populations from lower latitudes will grow in shorter days (Savolainen *et al.* 2007). Latitudinal variation in selection is also likely in other components controlling developmental transitions, for example in whether vernalization (exposure to a period of cold) is

required for flower induction (Caicedo *et al.* 2004, Shindo *et al.* 2005, Kuittinen *et al.* 2008).

To explore the molecular basis of such adaptation it is useful to consider the molecular pathways controlling seasonal transitions. In the annual molecular biology model plant *Arabidopsis thaliana* dozens of genes controlling flowering time have been characterized (Andrés & Coupland 2012). Many of these genes have also been found to be associated with the timing of flowering and growth in other even distantly related species (Böhlenius *et al.* 2006, Wang *et al.* 2009, Avia *et al.* 2014) suggesting that the same evolutionarily very old general pathways control the plant seasonal rhythms universally even though the functional details may be species specific (Andrés & Coupland 2012, Koskela *et al.* 2012, Mouhu *et al.* 2013). In brief, the flowering time pathway is divided into distinct sub-pathways of which the vernalization pathway, the photoperiodic pathway, the light quality pathway, the autonomous pathway and the gibberellic acid pathway are central. The information from the different pathways comes together at specific integrator genes, such as *FLOWERING LOCUS T* (*FT*), which control the transformation between developmental stages e.g. from vegetative stage to flowering stage (Ausín *et al.* 2005).

A concrete example of how these genes may function in adaptive evolution is the variation in the vernalization requirement between *A. thaliana* accessions. In some accessions of *A. thaliana,* exposure to cold is required for rapid flowering while other accessions will flower rapidly in long days without vernalization (e. g. Col and *Ler*) (Levy & Dean 1998). Recessive loss-of-function mutations in the *FRIGIDA* gene, which is a positive regulator of the flowering repressor *FLOWERING LOCUS C* (*FLC*) (Michaels & Amasino 1999), have been shown to remove the requirement for vernalization. Such rapid flowering is likely adaptive under certain conditions (Johanson *et al.* 2000). In the accessions with functional *FRIGIDA,* the upregulation of *FLC* causes repression of flowering, but vernalization represses *FLC* by removing methylation and thus allowing flowering after vernalization (Michaels & Amasino 1999).

1.6 Brassicaceae as a model system for plant evolutionary studies

The Brassicaceae family is a widely used plant model system in plant evolutionary and population genetics. According to the current estimates Brassicaceae includes 3 600 mostly herbaceous plant species that diverged approximately 65 million years ago from core Brassicales (Beilstein *et al.* 2010, Koenig & Weigel 2015). The most

widely studied species in this family is clearly the thale cress (*Arabidopsis thaliana*). The self-fertilizing mating system, small genome size (~135 Mb, The Arabidopsis Information Resource 2017), short generation time as well as the simple growing requirements and the convenient transformation procedures have made it ideal for molecular studies (Meinke *et al.* 1998, Koenig & Weigel 2015). From a molecular point of view, studies of the other species in this family benefit from the extensive knowledge obtained from *A. thaliana,* but having their own special characteristics, they offer possibilities to study traits and phenomena that are not accessible with *A. thaliana* (Mitchell-Olds 2001). For example *Arabis alpina* has been used in studies of perennial flowering (e.g. Wang *et al.* 2009), *Arabidopsis halleri* is exploited understating adaptation into heavy metal soils (e.g. Turner *et al.* 2010) and studies in *Capsella bursa-pastoris* & *Arabidopsis arenosa* have shed light on the evolution of polyploidization (e.g. Arnold *et al.* 2015, Douglas *et al.* 2015). In addition, this family contains several important crop plants such as cabbages (*Brassica*) and mustard (*Sinapis*) (Franzke *et al.* 2011). In this work two out-crossing species of the Brassicaceae family, *Arabidopsis lyrata* and *Capsella grandiflora*, were used as model species to study the roles of selection and demography in colonization and genome-wide diversity patterns. The special characteristics of these two species are hence described below.

1.6.1 Arabidopsis lyrata

Lyrate rockcress (*A. lyrata*) is a small herb species widely distributed across the Northern hemisphere (Hoffmann 2005, Schmickl *et al.* 2010). Large scale population structure of *A. lyrata* is very strong, with high population isolation (Wright *et al.* 2003, Muller *et al.* 2008, Ross-Ibarra *et al.* 2008, Pyhäjärvi *et al.* 2012). North American and European groups are separated into different sub-species; North American *A. lyrata* ssp. *lyrata* and Eurasian *A. lyrata* ssp. *petraea* (O'Kane Jr. & Al-Shehbaz. 1997, Shimizu *et al.* 2005, Shimizu-Inatsugi *et al.* 2009). The whole-genome clustering based on identical-by-state block length by (Novikova *et al.* 2016) suggests frequent introgression between *A. lyrata* ssp. *petraea* and *A. halleri* and *A. arenosa* group. They further found that North American *A. arenicola* clusters with *A. lyrata* ssp. *lyrata*.

In comparison to other regions, the highest overall genetic variation in the Central European *A. lyrata* populations suggests that the species originated in this region (Clauss & Mitchell-Olds 2006). Approximately 240 000 years ago the species started spreading to the North America (Pyhäjärvi *et al.* 2012) likely

through Asia (Schmickl *et al.* 2010). The rough population relationships suggest more recent spread within Europe and post-glacial colonization of northern Europe (Muller *et al.* 2008, Pyhäjärvi *et al.* 2012) or spread from a periglacial refugia (Ansell *et al.* 2010, Falahati-Anbaran *et al.* 2014).

The genome size of *A. lyrata* is approximately 207 Mb and contains approximately 33 000 annotated genes (Hu *et al.* 2011). The estimated divergence time from *A. thaliana* is ~10 million years ago (Wright *et al.* 2002, Beilstein *et al.* 2010) with average synonymous divergence of 0.163 while the non-synonymous divergence is 0.036 (median calculated over 26 000 aligned genes, data from Ensembl Plants (Kersey *et al.* 2014)). *A. lyrata* is a perennial obligate outcrosser with a fully functioning self-incompatibility system, making it ecologically very distinct from *A. thaliana* (Savolainen & Kuittinen 2011). The difference in mating system manifest itself in highly differing pattern of population structure, genetic diversity and patterns of selection in comparison with *A. thaliana* (Wright *et al.* 2003, Wright *et al.* 2002, Clauss & Mitchell-Olds 2006, Wright *et al.* 2008). In addition, the perennial life history requires alteration of growth and dormancy synchronized with the seasonal fluctuations (Savolainen & Kuittinen 2011).

Due to the large distribution range of *A. lyrata* it also inhabits a wide range of habitats likely causing local differences in selective pressures between populations. Such differences can cause the populations of a species to be locally adapted to their home environment and indeed evidence for local adaptation has been found in reciprocal transplantation experiments between populations of *A. lyrata* in different levels of neutral divergence (Leinonen *et al.* 2009, Leinonen *et al.* 2011, Vergeer & Kunin 2013). Population differentiation in several phenotypic traits also suggests adaptation to local environment. For example (Davey *et al.* 2009) studied cold related metabolic differences in Scandinavian and West European *A. lyrata* populations. Further differences in leaf morphology (Jonsell *et al.* 1995), trichome production (Kivimäki *et al.* 2007), drought tolerance (Sletvold & Ågren 2012) and gene expression patterns (Menzel *et al.* 2015, Videvall *et al.* 2015) have been found between the European *A. lyrata* populations.

Taken together these characteristics have made *A. lyrata a* widely used model system for studies of various aspects related to plant population genetics and molecular biology (Savolainen & Kuittinen 2011). In addition, the wide distribution range and clear population structure of *A. lyrata* makes the species ideal for studying ecological adaptation associated with post-glacial colonization.

1.6.2 *Capsella grandiflora*

Capsella grandiflora is a member of *Capsella* family and very close relative of self-fertilizing model species *C. rubella.* Based on flow cytometry the genome size of *Capsella* is approximately 219 Mb (Slotte *et al.* 2013) with 8 main chromosomes. *C. rubella* diverged from self-incompatible *C. grandiflora* very recently (Guo *et al.* 2009) and hence this species pair offers possibility to study the details of very recent mating system shift. As *A. lyrata, C. grandiflora* is an outcrossing but annual species with rather restricted distribution range (Hurka *et al.* 2012). Counterintuitively, its effective population size has been estimated to be large (St. Onge *et al.* 2011, Douglas *et al.* 2015) and in contrast to many other species the estimates of proportion of adaptive substitutions have been high (Gossmann *et al.* 2010, Slotte *et al.* 2010, Williamson *et al.* 2014).

1.7 Aims of the study

The aim of this study was to investigate neutral (demography induced) and selective processes associated with post-glacial colonization in *A. lyrata* using candidate gene and genome-wide population genetics approach. In addition, the genome-wide effects of linked selection were investigated in *A. lyrata* and *C. grandiflora.* The specific research questions were

I What is the post-glacial demographic and colonization history of *Arabidopsis lyrata*? (I)
II How did the colonization affect the genome-wide patterns of variation? (I)
III Can we detect loci showing evidence for selection associated with northward colonization? (I & II)
IV Do the colonizing *A. lyrata* populations show evidence for selection in photoperiodic pathway genes? (II)
V Is there a difference in the patterns of selection in populations of *A. lyrata* and *C. grandiflora*? (III)
VI Which factors are important for determining diversity in *A. lyrata* and *C. grandiflora* and what are their relative roles (III)?

2 Materials and methods

In this section I describe the dataset and the methods briefly. The detailed methods description can be found in the original papers.

2.1 Sampling and the sequence data

The plant material for the study was collected from 10 different *A. lyrata* populations covering roughly the distribution range of the species (see study I Fig. 1A and study II Fig. 2 for sampling locations). The following population abbreviations are used throughout the text: BOH, Bohemia, Czech Republic; ICE = Reykjavik, Iceland; ITH = Ithaca; USA; KAR = Karhumäki, Russia; LOM = Lom, Norway; MA = Mayodan, USA; PL = Plech, Germany; SP = Spiterstulen, Norway; STORR = Storr, UK; STU = Stubbsand, Sweden. Mark McNair, David Remington and Philippine Vergeer are acknowledged for offering the seed material from BOH, MA and STORR.

For the candidate gene study (study II) 19 well characterized flowering time genes were sequenced from 9 *A. lyrata* populations. 11 genes were from the photoperiodic pathway, 3 from the vernalization pathway, 3 from the autonomous pathway and 2 were flowering time integrator genes. We hypothesized that these genes may be under directional selection based on their function and hence chosen for investigation. The rationale behind the choice of the candidate genes was that the growth season and especially photoperiod during the growing season differ between the study populations, and as photoperiod (among other environmental cues) controls flowering time in plants, a change in this control likely has been crucial during colonization. As empirical support of this hypothesis, previous studies have found population differences in flowering time and photoperiodic response between the populations studied (Riihimäki & Savolainen 2004). As control loci we used 19 gene fragments from Pyhäjärvi *et al.* (2012) with presumably no adaptive importance based on their molecular function. The same laboratory protocol and the same set of individuals were used in both datasets.

We further re-sequenced the whole genomes of 2 to 6 individuals from 5 *A. lyrata* populations (study I) to characterize the genome-wide patterns of polymorphism of the populations using Illumina HiSeq2000 instrument with paired-end 100 bp read length. The library preparation and sequencing was done at the Institute for molecular medicine, Finland (FIMM). We developed a bioinformatics pipeline for the data analysis in study I.

For the comparative study (III) we exploited the data from the Norwegian (SP) *A. lyrata* population (data from study I and Hämälä *et al.* In prep) and a single population of *C. grandiflora* from Greece (reference mapped data obtained from Steige *et al.* 2017).

2.2 Data analysis pipeline

The Illumina sequence reads were processes using a custom pipeline developed in the study to produce high quality population genetics data for *A. lyrata* using a tool STAPLER (Tyrmi 2016) and additional custom scripts. In brief, the raw sequence reads were quality trimmed with Trimmomatic 0.32 (Bolger *et al.* 2014) and mapped against the Ensembl plant (Kersey *et al.* 2014) version 1.0.29 *A. lyrata* reference genome (Hu *et al.* 2011) with short read alignment tool bwa-mem (Li & Durbin 2009, Li 2013). Picard tools v. 1.113 (Broad Institute 2017), BamUtil v. 1.0.13 (BamUtil documentation 2015) and Genome Analysis Toolkit version 3.2.2 (Depristo *et al.* 2011) were used in further alignment processing. All repeat annotated regions, regions with fixed heterozygotes within population (indicating putative paralogy), regions with low quality data or mapping (base quality < 20 and mapping quality < 30, not properly paired reads) and *A. thaliana* organelle aligning regions were excluded from the population genetics analysis. For all the analysis 80% of data presence was required.

For the flowering time genes (study II) primers were designed using *A. thaliana* (Arabidopsis genome initiative 2000) and *A. lyrata* reference genomes (Hu *et al.* 2011) and sequenced with ABI 377 or 3730 DNA Analyzer from forward and reverse directions. The sequence data was aligned and manually curated using either SEQUENCHER 4.0.5 (Gene Codes, Ann Arbor, MI, USA) or CodonCodeAligner 3.5.7 (CodonCode Corporation, Centerville, MA, USA). Haplotype phase for each locus and individual was inferred with PHASE v.2.1 (Stephens *et al.* 2001, Stephens & Scheet 2005).

2.3 Characterizing the patterns of population genetic variation and population structure

To characterize within population diversity and skew in site frequency spectra we calculated π, θ_W and Tajima's D either with MANVA (Heidel *et al.* 2010) (study II) or with ANGSD (Korneliussen *et al.* 2014, Nielsen *et al.* 2012) (studies I & III). F_{st} estimates between pairs of populations were calculated using the method described

in Fumagalli *et al.* (2013). Single and multi-dimensional SFS and bootstrap SFS were calculated with ANGSD (studies I & III). The ancestral state of each variable site was inferred based on *A. thaliana*. To correct for ancestral misidentification we developed a method with partially similar principles described previously (Baudry & Depaulis 2003, Hernandez *et al.* 2007) (study I) or else used folded SFS whenever possible.

The methods used for the genome-wide analyses are based on allele frequency likelihoods for each population. These methods were designed to take into account the genotype uncertainty in low and medium coverage NGS data (Fumagalli *et al.* 2013, Korneliussen *et al.* 2014, Nielsen *et al.* 2012), and have been shown to be able to estimate population allele frequencies accurately even with very low sequencing depth data (Han *et al.* 2014). As our sample varied in the amount of sequence data per individual (median read depth per individual ranged from 6 to 29, study I Table S1), these method were used analyzing the whole genome re-sequencing dataset.

The basic population structure was characterized with principal component analysis and the branching topology of the populations was studied with ABBA-BABA test (Durand *et al.* 2011) using population version of the method implemented in the ANGSD package. For population admixture proportion estimates we used maximum likelihood based method implemented in the program ADMIXTURE (Alexander *et al.* 2009) and a method designed for next generation sequencing with variable depths between individuals implemented in the program NGSadmix (Skotte *et al.* 2013).

2.4 Demographic inference

The demographic histories of the populations were studied using three different methods all based on single population or multi-dimensional site frequency spectra calculated with ANGSD. Single population demographic histories (study I) were estimated with a model-flexible method that estimates θ over time using the different site frequency categories implemented in the software Stairwayplot beta v. 2.0 (Liu & Fu 2015). Confidence intervals (CI) were estimated based on 200 bootstrap replicates. We further tested different demographic scenarios (population size decline, instant growth, exponential growth and bottleneck with growth) for SP population (study III) and the *C. grandiflora* population using diffusion approximation based method using the python library ∂a∂I (Gutenkunst *et al.* 2009). For the population split time estimation and model comparison (study I) we used

coalescent composite likelihood method implemented in the coalescent simulator fastsimcoal2 (Excoffier *et al.* 2013). For each parameter the 95% confidence interval was estimated based on 100 bootstrap replicates calculated with ANGSD. The parameter search ranges were defined based on previous studies (Pyhäjärvi *et al.* 2012, Ross-Ibarra *et al.* 2008) and geological knowledge (summarized in Hughes *et al.* 2016).

2.5 Detecting selective sweeps

The signals for selection in the flowering time genes (study II) were studied using a combination of methods sensitive for selection in different timescales. For the flowering time dataset, we used the maximum likelihood Hudson–Kreitman–Aquadé (MLHKA) method (Wright & Charlesworth 2004), population branch aware differentiation statistics FLK (Bonhomme *et al.* 2010) and compared site frequency spectra of candidate and reference genes within each population. For the MLHKA analysis, a model with selection on the candidate genes was compared with a model without selection within each population using likelihood ratio test. For FLK analysis, we used the reference gene set for building a population phylogeny and this phylogeny was used to simulate candidate gene like data. The simulated thresholds were used to test the significance of FLK statistics of each variable site.

The genome-wide data was screened for locus specific selection in SP and STU populations (PL population was an out-group) using population branch statistics (PBS) (Yi *et al.* 2010) and a 0.1% empirical significance threshold. PBS was calculated as implemented in the ANGSD package.

2.6 Genome-wide selection inference and linked selection

The genome-wide impact of selection was contrasted in populations of *A. lyrata* and *C. grandiflora* in the study III. The genome-wide distribution of fitness effects of new mutations and proportion of mutation fixed by positive selection at 0-fold sites were estimated using maximum likelihood method implemented in DFE-alpha (Eyre-Walker & Keightley. 2009, Keightley & Eyre-Walker. 2007) assuming mutations at 4-fold degenerate sites as neutral. A two epoch demographic model was simultaneously estimated to control the effect of demographic history. For both species the 4-fold and 0-fold SFS were calculated with ANGSD. The divergence

estimates were calculated from a *A. thaliana–A. lyrata–C. rubella* whole genome alignment (Steige *et al.* 2017)

To test factors explaining neutral diversity, multiple linear regression with the diversity at 4-fold degenerate sites as responsible variable was used. Recombination rate (*A. lyrata* data obtained from Hämälä *et al.* 2017, *C. grandiflora* data obtained from Slotte *et al.* 2013), divergence at 4-fold degenerate sites, gene density, GC-content and transposable element (TE) content were used as explanatory variables. GC- and TE content was calculated from the reference genome annotations. The best model explaining diversity was chosen using stepAIC function in R (R Core Team 2016).

A decrease in diversity around functional regions is expected under linked selection model. In order to study this effect, we estimated intergenic π calculated in non-overlapping windows (study III) and for each window we defined the distance to the closest gene and plotted the diversity as a function of distance from the nearest gene. The difference between *A. lyrata* and *C. grandiflora* populations were tested in different distance bins using 95% bootstrap based confidence interval of the mean. Using the recombination maps the physical distances were transformed into genetic distances.

3 Results and discussion

In this section the main results of the original papers I–III are summarized and the key findings of these works are highlighted. I also discuss the dataset in the light of challenges given the current material and the study system in general.

3.1 Nucleotide diversity data suggest population size decline in *A. lyrata*

The within population genome-wide variation revealed large differences between the study populations reflecting differences in their demographic histories. The median estimates of θ_W at 4-fold degenerate sites ranged from 0.0024 to 0.0155 and π from 0.0027 to 0.0159 for the *A. lyrata* populations. The lowest estimates were in the North American populations (lowest in MA) and the highest estimate in the Central European populations (highest in PL) (Table 1). The three Scandinavian populations were very similar in diversity while diversity in the Icelandic population was slightly higher. Since $\theta = 4N_e\mu$ we can use these estimates to get the approximate N_e assuming constant mutation rate estimate of $7*10^{-9}$ (Ossowski *et al.* 2010). This yields long term effective population size estimates ranging approximately from 97 000 (in MA) to 590 000 (in PL) (Table 1). The scale and population ranking in Sanger based estimate for synonymous site θ estimates were concordant with the whole-genome sequencing based results (Table 1). Due to only two individuals sampled from the STORR population (Scotland), we were not able to estimate θ for this population.

Table 1. Median neutral diversity estimates in 9 *A. lyrata* populations calculated over loci (19 Sanger sequenced loci) or 100 Kb windows with at least 500 callable sites per window (Illumina based data). 95% CIs shown in parenthesis were calculated from 10 000 bootstrap replicates.

Sub-species	Region	Population	Sanger based data			Illumina based data		
			θ_W	π	N_e[1]	θ_W	π	N_e[1]
lyrata	North America	ITH	0.0026 (0–0.0088)	0.0037 (0–0.0112)	133 117 (0–399 461)	NA	NA	NA
lyrata	North America	MA	0.0019 (0–0.0048)	0.0014 (0–0.0068)	49 700 (0–244 341)	0.0024 (0.0024–0.0025)	0.0027 (0.0026–0.0028)	96 909 (91 688–100 847)
petraea	Central Europe	BOH	0.0144 (0.0066–0.0203)	0.0104 (0.0069–0.0236)	371 612 (245 039–842 869)	NA	NA	NA
petraea	Central Europe	PL	0.0232 (0.0097–0.0269)	0.019 (0.0114–0.0301)	678 488 (406 355–1 074 000)	0.0155 (0.0147–0.016)	0.0159 (0.0154–0.0166)	568 807 (55 0324–59 1320)
petraea	East Europe	KAR	0.0037 (0–0.0055)	0.0065 (0–0.0104)	230 514 (0–369 871)	NA	NA	NA
petraea	North Europe	LOM	0.0074 (0.0037–0.0132)	0.0094 (0.0027–0.0147)	334 050 (97 165–525 204)	NA	NA	NA
petraea	North Europe	ICE	0.0116 (0.0035–0.016)	0.0134 (0.0055–0.0155)	477 929 (194 884–554 391)	NA	NA	NA
petraea	North Europe	SP	0.0065 (0.0032–0.0148)	0.0079 (0.0031–0.0207)	280966 (109 572–739 459)	0.0083 (0.008–0.0086)	0.0092 (0.0089–0.0094)	327 476 (316 698–337 388)
petraea	North Europe	STU	0.0058 (0.0019–0.0152)	0.0094 (0.0013–0.0241)	334 606 (46 710–860 484)	0.008 (0.0076–0.0085)	0.0088 (0.0085–0.0092)	314 886 (303 139–328 747)

[1]Calculated from π

The genome-wide difference in θ_W and π reflected as positive Tajima's D (study I Fig. 1D, study II Fig. 5) suggests that the population size have fluctuated in the recent history of the species. The study of demographic history (study I Fig. 2) pointed to a decrease in effective population size in the past few hundred thousand years of the species as a whole. A qualitatively similar result was retained also with other demographic modelling methods (study I Table 1, study III Table 2).

The diversity estimates for the *C. grandiflora* population suggests higher diversity than any of the studied *A. lyrata* populations with $\theta_W = 0.0184$ and $\pi = 0.0173$ in 4-fold degenerate sites. Assuming the same mutation rate as for *A. lyrata* we obtain a long term N_e estimate of 620 000 (from π). In contrast to *A. lyrata, C. grandiflora* had higher diversity at intergenic sites with $\theta_W = 0.0299$ and $\pi = 0.0243$ (study III Table 1) yielding a long term N_e estimate of 870 000. The demographic inference (study III Table 2) suggested a slight increase in the effective population size of the study population of *C. grandiflora*.

The general patterns of polymorphism were in line with previous estimates for both species (*A. lyrata*: Wright *et al.* 2003, Ramos-Onsins *et al.* 2004, Ross-Ibarra *et al.* 2008, Pyhäjärvi *et al.* 2012, Vigueira *et al.* 2013; *C. grandiflora*: Slotte *et al.* 2010, Williamson *et al.* 2014, Douglas *et al.* 2015). The scale of θ estimates was comparable with estimates from other species. For example in *Drosophila melanogaster* the synonymous π estimate over more than 7 000 autosomal genes in crossover regions was 0.0141 (Campos *et al.* 2014). For three Populus species, *Populus tremula, P. tremuloides* and *P. trichocarpa,* genome-wide π estimates at 4-fold degenerate sites were 0.0108, 0.0140 and 0.0040, respectively (Wang *et al.* 2016). The estimates presented here were also in the range of genome-wide estimates obtained for 28 plant and 34 animal species (Chen *et al.* 2017). Interestingly, the highest diversities estimated in *A. lyrata* were closer to the annual outcrossing species than other perennial outcrossers (as *A. lyrata*). The results presented here indicate that the general level of diversity can vary highly even between populations of the same species with similar life histories. On the other hand, differentiation of life-histories have been found between the populations of *A. lyrata* studied, the North American populations being closer to annual life-history (Remington *et al.* 2015).

3.2 Scandinavian colonization from west European refugium

The study of population structure and admixture analysis revealed very clear genetic differentiation between the *A. lyrata* populations (study I Fig. 1B & Fig.

S1). The strongest difference was between the sub-species explaining 23% of the variance in the data (principal component analysis). The second largest proportion of variance explaining PC (explaining 16% of the variance) separated Central European population from the other populations and the other populations were separated from each other by the PC3 and PC4 (study I Fig. 1B). In the admixture analysis the model with the highest support (lowest cross-validation error) included three populations separating the sub-species and the Central European population from the group including SP, STU and STORR (study I Table S4, Fig. S1). In this model, some Central European ancestry was assigned to the Scottish individuals (study I Fig. S1). The median pairwise genome-wide F_{st} estimates were 0.33 for SP and STU; 0.36 for Scandinavian populations and STORR; 0.38 between the Northwestern European populations and PL; and 0.69 between the sub-species averaged over all pairwise comparisons (Table 2) indicating high population differentiation.

Table 2. Pairwise genome-wide F_{st} estimates at 4-fold degenerate sites. 95% CIs were calculated from 10 000 bootstrap replicates.

Population comparison	F_{st} (95% CI)
SP–STU	0.33 (0.32–0.34)
SP–STORR	0.35 (0.35–0.36)
STU–STORR	0.36 (0.35–0.38)
SP–PL	0.40 (0.40–0.41)
STORR–PL	0.34 (0.34–0.35)
STU–PL	0.41 (0.40–0.42)
PL–MA	0.59 (0.59–0.60)
SP–MA	0.71 (0.71–0.72)
STORR–MA	0.75 (0.74–0.75)
STU–MA	0.72 (0.71–0.73)

The test for different branching topologies (ABBA-BABA and fastsimcoal2 model comparison) suggested that the Central European populations separated from the Northwestern populations first and second the Scandinavian population separated from the Scottish population. The Scandinavian populations separated from each other last (study I Table 1, Table S6, S7). We estimated that the STORR–Scandinavian split occurred approximately 14 000 generations ago (95% CI 9 400–33 000) and the SP–STU split approximately 11 000 (95% CI 2 300–26 000) years ago assuming generation time 2 years (study I Fig. 5, Table 1). We also allowed gene flow between the populations but due to the current isolated populations no

current gene flow was allowed. The time when the Scandinavian populations became fully isolated was estimated to be approximately 3 000 years ago (95% CI 800–12 000).

The pattern of close relatedness with British Isles and Scandinavian population has also been found in other plant species and genetic barrier between the British Isles and Scandinavia has proposed to be weak (Eidesen *et al.* 2013, Alsos *et al.* 2015). We suggest that *A. lyrata* colonized Scandinavia recently after the last glacial maximum from western European refugium and this refugium population has long been well separated from the Central European gene pool. Based on the current distribution of the species and preference of cold environments (Hoffmann 2005) we suggested that this refugium may have been relatively far to the north, perhaps near the glacier.

3.3 Signatures of selection in the colonizing *A. lyrata* populations

Signatures of directional selection were studied in the context of Scandinavian colonization in studies I and II. In study II we investigated sequence variation in 19 flowering time genes and found strong diversity reduction on the flowering time genes in comparison with the reference gene set (study II, Fig. 3). Comparing divergence and polymorphism with models with and without selection in flowering time genes revealed that in the Northern European populations a model including selection is significantly better than the neutral model (MLHKA likelihood ratio test, study II Table S1). Further, these populations along with the other colonizing populations (MA and KAR) had significant differences between the reference and candidate gene SFS (study II, Table S3, Fig. S1) and a high number of variable sites show strong population differentiation (study II, Fig. 6).

Among the candidate genes studied, the signature of selection was strongest at the photoperiodic pathway genes in comparison with the genes from the other pathways and the integrator genes (study II, Table 2), pinpointing the importance of photoperiodic adaptation in this species. Evidence for selection on one of the genes included here, *PHYA*, was also found previously in a study where the flanking regions of this gene were studied in-depth (Toivainen *et al.* 2014). In addition, several other studies have found evidence for selection in flowering time genes for example in *A. thaliana* (Le Corre *et al.* 2002), *Populus* species (Ingvarsson *et al.* 2008, Keller *et al.* 2012, Evans *et al.* 2014) and spruce (Chen *et al.* 2014) supporting the hypothesis that this pathway has a central role in plant adaptation in general.

Using the genome-wide selection scan with PBS in the two Scandinavian populations we detected loci showing evidence for selection after the split of these lineages. Due to high variation in patterns of diversity on different parts of the genome, for which the estimated demographic models did not fully account for, an empirical significance threshold was used. Outlier loci were detected in a large region in the chromosome 1 in the STU (Swedish) lineage while the outliers in the SP (Norwegian) lineage were scattered across the genome (study I, Fig. 5). The strongest candidate genes for environmental adaptation in the detected regions were *ZAT10* which is involved in abiotic stress response (Mittler *et al.* 2006) and *U2AF^{35}A* involved in photoperiodic flowering (Wang & Brendel 2006) in *A. thaliana*.

3.4 Contrasting patterns of genome-wide selection in populations of *A. lyrata* and *C. grandiflora*

The difference in the demographic history observed in the *A. lyrata* and *C. grandiflora* populations was hypothesized to have an impact on the patterns of genome-wide selection. To quantify these effects, the genome-wide patterns of divergence and polymorphism levels at 0-fold and 4-fold degenerate sites were contrasted in populations of these species (study III). Lower π_0 / π_4 ratio (0.19 in *C. grandiflora* vs. 0.29 in *A. lyrata*) suggested stronger purifying selection in the *C. grandiflora* population. The investigation of distribution of fitness effects confirmed this with proportion of sites classified as very harmful ($N_eS > 10$) were 66% and 72% while 29% and 15% were assigned to the nearly neutral category ($N_eS < 1$) in the *A. lyrata* and the *C. grandiflora* datasets, respectively (study III Fig. 2). Further, the estimated proportion of adaptive substitutions at 0-fold sites was 1.4% (0.9%–1.9%, 95% CI) for the *A. lyrata* population while for the *C. grandiflora* population the corresponding estimate was 7.6% (7.3%–7.8%, 95% CI). Linear regression analysis revealed significant contributions of recombination, divergence at 4-fold degenerate sites, gene density, GC-content and TE% on the neutral diversity level in both species (study III, Table 4). However, the proportion of variance explained by the model was much higher for the *C. grandiflora* dataset (53%) than for the *A. lyrata* dataset (11%). This indicates much higher stochastic variation in the *A. lyrata* dataset. Further, diversity was significantly reduced around genes in protein coding regions of *C. grandiflora* as expected (study III Fig. 4), while in *A. lyrata* no evidence for reduction was found (study III Fig. 4).

The results indicate a difference in the efficacy and effect of selection in the study population which is likely explained by the difference in the population demographic histories. The lower efficacy of selection is in line with the theoretical population genetics prediction suggesting that in small populations slightly deleterious mutations behave neutrally (Ohta 1973). Smaller efficacy of selection associated with smaller N_e populations has also been observed in other plant species (Strasburg *et al.* 2011, Wang *et al.* 2016). The difference on the patterns observed near genes further emphasized the demographic effect on selection at linked sites. Simulation approach on this aspect would help to shed light on this issue further.

3.5 Challenges in *A. lyrata* population genetics and genomics

In study II we found that the high genetic differentiation between the study populations (mainly the Central European, Russian, Scandinavian and North America) was so high that separating selection from neutral processes at single base pair resolution is extremely challenging. Hence, in study I we concentrated on the relatively recently diverged lineages. The high variation in diversity patterns, a characteristic of populations with recent population size decline (Fig. 1), complicates the detection of the signals of selection. Sampling more closely related populations would make it easier to account for neutral processes in *A. lyrata* selection studies. However, such populations are often also more similar in their habitat and hence the selective gradient is not so clear. For future studies aiming to detect selection in these populations would benefit using the populations from British Isles as a reference population instead of the Central European populations.

Another challenge in this study was the high complexity of the *A. lyrata* genome (studies I & III). Approximately 30% of the *A. lyrata* genome is annotated as repeats (Hu *et al.* 2011). These regions are challenging to map especially with short reads and these regions were hence masked from the analysis. Further, a peak in the strict intermediate SFS category was observed in all the populations studied. After removing sites with fixed heterozygotes (study I & III) this peak disappeared suggesting that this pattern is produced by extensive paralogous mapping. It was also found that this effect is population specific. To make the results between populations as comparable as possible the regions showing evidence for this effect in any population were excluded, limiting the dataset further. As the focus of these studies were mostly on genome-wide processes, this was not seen as a major issue here but locus specific analyses are still limited to the accessible genome. Deeper

understanding of the processes shaping genome organization in this species could help us to utilize whole-genome short read datasets more extensively.

4 Conclusions

This study aimed to investigate the neutral and selective processes in the different populations of *Arabidopsis lyrata,* an outcrossing perennial herb species with wide but patchy distribution. The contrasting patterns of diversity between the different populations even in relatively recently diverged populations emphasize the importance of local nature in population genetics studies. In addition, the comparison with a population of another Brassicaceae species, *Capsella grandiflora,* emphasized the different mixture of factors that may result in very different outcome of genome-wide variation and selection even with similar mating system and life-history traits.

The population specific demographic characteristics were found to have a strong effect on the distribution of population genetics statistics between loci (as predicted by simulations) which can have an effect on statistical power to detect signals of selection. It is not fully known whether this has an effect on the species ability to adapt or whether this is just statistical outcome but nevertheless this should be taken into account when comparing species with different demographic background.

Footprints of natural selection were investigated in this study using candidate gene and genome-wide approaches. Positive selection detected in the Scandinavian populations at photoperiodic and other loci widens the understanding of selection associated with post-glacial colonization and is intriguing from an evolutionary point of view. The results presented in this work may be a starting point for further in-depth investigations.

This study also took an approach to study biogeography using genome-wide data. Even though the relatively small sample size sets some limitations on the conclusions that can be drawn from the data and single sampling locations represent very large regional diversity in this study, it was possible to inspect other aspects of this process in great detail. The advanced DNA sequencing methods now allow unprecedented possibilities to widen the understanding of species colonization history.

List of references

Alexander DH, Novembre J & Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19(9): 1655–1664.

Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L, Malaspinas AS, Margaryan A, Higham T, Chivall D, Lynnerup N, Harvig L, Baron J, Casa PD, Dabrowski P, Duffy PR, Ebel AV, Epimakhov A, Frei K, Furmanek M, Gralak T, Gromov A, Gronkiewicz S, Grupe G, Hajdu T, Jarysz R, Khartanovich V, Khokhlov A, Kiss V, Kolár J, Kriiska A, Lasak I, Longhi C, McGlynn G, Merkevicius A, Merkyte I, Metspalu M, Mkrtchyan R, Moiseyev V, Paja L, Pálfi G, Pokutta D, Pospieszny L, Douglas Price T, Saag L, Sablin M, Shishlina N, Smrcka V, Soenov VI, Szeverényi V, Tóth G, Trifanova SV, Varul L, Vicze M, Yepiskoposyan L, Zhitenev V, Orlando L, Sicheritz-Pontén T, Brunak S, Nielsen R, Kristiansen K & Willerslev E. (2015) Population genomics of Bronze Age Eurasia. Nature 522(7555): 167–172.

Alsos IG, Ehrich D, Eidesen PB, Solstad H, Westergaard KB, Schönswetter P, Tribsch A, Birkeland S, Elven R & Brochmann C (2015) Long-distance plant dispersal to North Atlantic islands: Colonization routes and founder effect. AoB Plants 7(1): plv036.

Amasino R (2010) Seasonal and developmental timing of flowering. Plant J 61(6): 1001–1013.

Anderson JT, Willis JH & Mitchell-Olds T (2011) Evolutionary genetics of plant adaptation. Trends Genet 27(7): 258–266.

Andrés F & Coupland G (2012) The genetic basis of flowering responses to seasonal cues. Nat Rev Genet 13(9): 627–639.

Ansell SW, Stenøien HK, Grundmann M, Schneider H, Hemp A, Bauer N, Russell SJ & Vogel JC (2010) Population structure and historical biogeography of European *Arabidopsis lyrata*. Heredity 105(6): 543–553.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814): 796–815.

Arnold B, Kim ST & Bomblies K (2015) Single geographic origin of a widespread autotetraploid *Arabidopsis arenosa* lineage followed by interploidy admixture. Mol Biol Evol 32(6): 1382–1395.

Ausín I, Alonso-Blanco C & Martínez-Zapater JM (2005) Environmental regulation of flowering. Int J Dev Biol 49(5–6): 689–705.

Avia K, Kärkkäinen K, Lagercrantz U & Savolainen O (2014) Association of FLOWERING LOCUS T/TERMINAL FLOWER 1-like gene FTL2 expression with growth rhythm in Scots pine (*Pinus sylvestris*). New Phytol 204(1): 159–170.

BamUtil documentation (2015) URI: http://genome.sph.umich.edu/wiki/BamUtil. Cited 2017/10/10.

Baudry E & Depaulis F (2003) Effect of misoriented sites on neutrality tests with outgroup. Genetics 165(3): 1619–1622.

Begun DJ & Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356(6369): 519–520.

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR & Mathews S (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. PNAS 107(43): 18724–18728.

Berg JJ & Coop G (2014) A Population genetic signal of polygenic adaptation. PLoS Genet 10(8): e1004412.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE & Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74(6): 1111–1120.

Birks HH, Giesecke T, Hewitt GM, Tzedakis PC, Bakke J & Birks HJB (2012) Comment on "Glacial survival of boreal trees in Northern Scandinavia". Science 338(6108): 742.

Birks HH, Larsen E & Birks HJB (2005) Did tree-Betula, Pinus and Picea survive the last glaciation along the west coast of Norway? A review of the evidence, in light of Kullman (2002). J Biogeogr 32(8): 1461–1471.

Biswas S & Akey JM (2006) Genomic insights into positive selection. Trends Genet 22(8): 437–446.

Böhlenius H, Huang T, Charbonnel-Campaa L, Brunner AM, Jansson S, Strauss SH & Nilsson O (2006) CO/FT regulatory module controls timing of flowering and seasonal growth cessation in trees. Science 312(5776): 1040–1043.

Bolger AM, Lohse M & Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30(15): 2114–2120.

Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S & SanCristobal M (2010) Detecting selection in population trees: The Lewontin and Krakauer test extended. Genetics 186(1): 241–262.

Braverman JM, Hudson RR, Kaplan NL, Langley CH & Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140(2): 783–796.

Broad Institute (2017) Picard. A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. URI: http://broadinstitute.github.io/picard/, Cited 2017/6/6.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M & Clark AG (2005) Natural selection on protein-coding genes in the human genome. Nature 437(7062): 1153–1157.

Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J & Purugganan MD (2004) Epistatic interaction between *Arabidopsis FRI* and *FLC* flowering time genes generates a latitudinal cline in a life history trait. PNAS 101(44): 15670–15675.

Campos JL, Halligan DL, Haddrill PR & Charlesworth B (2014) The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. Mol Biol Evol 31(4): 1010–1028.

Charlesworth B (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet Res 63(3): 213–227.

Charlesworth B, Morgan MT & Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134(4): 1289–1303.

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2(4): 379–384.

Chen J, Tsuda Y, Stocks M, Källman T, Xu N, Kärkkäinen K, Huotari T, Semerikov VL, Vendramin GG & Lascoux M (2014) Clinal variation at phenology-related genes in spruce: Parallel evolution in FTL2 and Gigantea? Genetics 197(3): 1025–1038.

Chen J, Glémin S & Lascoux M (2017) Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species. Mol Biol Evol 34(6): 1417–1428.

Clauss MJ & Mitchell-Olds T (2006) Population genetic structure of *Arabidopsis lyrata* in Europe. Mol Ecol 15(10): 2753–2766.

Coop G, Witonsky D, Di Rienzo A & Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. Genetics 185(4): 1411–1423.

Cutter AD & Payseur BA (2013) Genomic signatures of selection at linked sites: Unifying the disparity among species. Nat Rev Genet 14(4): 262–274

Darwin C (1859) On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life. London, J. Murray.

Davey MP, Ian Woodward F & Paul Quick W (2009) Intraspecfic variation in cold-temperature metabolic phenotypes of *Arabidopsis lyrata* ssp. *petraea*. Metabolomics 5(1): 138–149.

Davis MB & Shaw RG (2001) Range shifts and adaptive responses to quaternary climate change. Science 292(5517): 673–679.

Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D & Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5): 491–501.

Donohue K (2003) Setting the stage: Phenotypic plasticity as habitat selection. Int J Plant Sci 164(3): S92.

Donohue K (2005) Seeds and seasons: Interpreting germination timing in the field. Seed Sci Res 15(3): 175–187.

Douglas GM, Gos G, Steige KA, Salcedo A, Holm K, Josephs EB, Arunkumar R, Ågren JA, Hazzouri KM, Wang W, Platts AE, Williamson RJ, Neuffer B, Lascoux M, Slotte T & Wright SI (2015) Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. PNAS 112(9): 2806–2811.

Durand EY, Patterson N, Reich D & Slatkin M (2011) Testing for ancient admixture between closely related populations. Mol Biol Evol 28(8): 2239–2252.

Eidesen PB, Ehrich D, Bakkestuen V, Alsos IG, Gilg O, Taberlet P & Brochmann C (2013) Genetic roadmap of the Arctic: Plant dispersal highways, traffic barriers and capitals of diversity. New Phytol 200(3): 898–910.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L & Järvelä I (2002) Identification of a variant associated with adult-type hypolactasia. Nat Genet 30(2): 233–237.

Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen JG, Tuskan GA & DiFazio SP (2014) Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. Nat Genet 46(10): 1089–1096

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC & Foll M (2013) Robust demographic inference from genomic and SNP Data. PLoS Genet 9(10): e1003905.

Excoffier L, Hofer T & Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity 103(4): 285–298.

Excoffier L, Novembre J & Schneider S (2000) SIMCOAL: A general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. J Hered 91(6): 506–509.

Eyre-Walker A & Keightley PD (2007) The distribution of fitness effects of new mutations. Nat Rev Genet 8(8): 610–618.

Eyre-Walker A & Keightley PD (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol 26(9): 2097–2108.

Falahati-Anbaran M, Lundemo S, Ansell SW & Stenoien HK (2014) Contrasting patterns of genetic structuring in natural populations of *Arabidopsis lyrata* subsp. *petraea* across different regions in Northern Europe. PLoS ONE 9(9): e107479.

Fisher RA (1930) The Genetical theory of natural selection. Oxford., Clarendon Press.

Foll M & Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. Genetics 180(2): 977–993.

Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA & Mummenhoff K (2011) Cabbage family affairs: The evolutionary history of Brassicaceae. Trends Plant Sci 16(2): 108–116.

Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A & Nielsen R (2013) Quantifying population genetic differentiation from next-generation sequencing data. Genetics 195(3): 979–992.

Garner WW & Allard HA (1920) Effect of the relative length of day and night and other factors of the environment on growth and reproduction in plants. J Agr Res 18: 553–606.

Garud NR, Messer PW, Buzbas EO & Petrov DA (2015) Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. PLoS Genet 11(2): 1–32.

Gillespie JH (2001) Is the population size of a species relevant to its evolution? Evolution 55(11): 2161–2169.

Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA & Eyre-Walker A (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol Biol Evol 27(8): 1822–1832.

Guo YL, Bechsgaard JS, Slotte T, Neuffer B, Lascoux M, Weigel D & Schierup MH (2009) Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. PNAS 106(13): 5246–5251.

Gutenkunst RN, Hernandez RD, Williamson SH & Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5(10): e1000695.

Hämälä T, Mattila TM, Leinonen PH, Kuittinen H & Savolainen O (2017) Role of seed germination in adaptation and reproductive isolation in *Arabidopsis lyrata*. Mol Ecol 26(13): 3484–3496.

Han E, Sinsheimer JS & Novembre J (2014) Characterizing bias in population genetic inferences from low-coverage sequencing data. Mol Biol Evol 31(3): 723–735.

Hardy GH (1908) Mendelian proportions in a mixed population. Science, 28(706): 49–50.

Heidel AJ, Ramos-Onsins SE, Wang WK, Chiang TY & Mitchell-Olds T (2010) Population history in *Arabidopsis halleri* using multilocus analysis. Mol Ecol 19(16): 3364–3379.

Hein J, Schierup M & Wiuf C (2004) Gene genealogies, variation and evolution: A primer in coalescent theory. Oxford, Oxford University Press.

Hermisson J & Pennings PS (2005) Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics 169(4): 2335–2352.

Hernandez RD, Williamson SH & Bustamante CD (2007) Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol 24(8): 1792–1800.

Hewitt G (2000) The genetic legacy of the quaternary ice ages. Nature 405(6789): 907–913.

Hewitt GM (1999) Post-glacial re-colonization of European biota. Biol J Linn Soc 68(1–2): 87–112.

Hoffmann MH (2005) Evolution of the realized climatic niche in the genus *Arabidopsis* (Brassicaceae). Evolution 59(7): 1425–1436.

Hohenlohe PA, Phillips PC & Cresko WA (2010) Using population genomics to detect selection in natural populations: Key concepts and methodological considerations. Int J Plant Sci 171(9): 1059–1071.

Holsinger KE & Weir BS (2009) Genetics in geographically structured populations: Defining, estimating and interpreting FST. Nat Rev Genet 10(9): 639–650.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottilar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KFX, Van dP, Grigoriev IV, Nordborg M, Weigel D & Guo Y (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet 43(5): 476–481.

Hudson RR (1991) Gene genealogies and the coalescent process. In: Futuyma D & Antonovics J (eds) Oxford surveys in evolutionary biology. Vol 7. New York, Oxford University Press: 1–44.

Hughes ALC, Gyllencreutz R, Lohne OS, Mangerud J & Svendsen JI (2016) The last Eurasian ice sheets - a chronological database and time-slice reconstruction, DATED-1. Boreas 45(1): 1–45.

Hurka H, Friesen N, German DA, Franzke A & Neuffer B (2012) 'Missing link' species *Capsella orientalis* and *Capsella thracica* elucidate evolution of model plant genus *Capsella* (Brassicaceae). Mol Ecol 21(5): 1223–1238.

Ingvarsson PK, Garcia MV, Luquez V, Hall D & Jansson S (2008) Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (*Populus tremula*, Salicaceae). Genetics 178(4): 2217–2226.

Johanson U, West J, Lister C, Michaels S, Amasino R & Dean C (2000) Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. Science 290(5490): 344–347.

Jonsell B, Kustås K & Nordal I (1995) Genetic variation in *Arabis petraea*, a disjunct species in northern Europe. Ecography 18(4): 321–332.

Kaplan NL, Hudson RR & Langley CH (1989) The 'hitchhiking effect' revisited. Genetics 123(4): 887–899.

Keightley PD & Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177(4): 2251–2261.

Keller SR, Levsen N, Olson MS & Tiffin P (2012) Local adaptation in the flowering-time gene network of balsam poplar, *Populus balsamifera* L. Mol Biol Evol 29(10): 3143–3152.

Kersey PJ, Allen JE, Christensen M, Davis P, Falin LJ, Grabmueller C, Hughes DST, Humphrey J, Kerhornou A, Khobova J, Langridge N, McDowall MD, Maheswari U, Maslen G, Nuhn M, Ong CK, Paulini M, Pedro H, Toneva I, Tuli MA, Walts B, Williams G, Wilson D, Youens-Clark K, Monaco MK, Stein J, Wei X, Ware D, Bolser DM, Howe KL, Kulesha E, Lawson D & Staines DM (2014) Ensembl Genomes 2013: Scaling up access to genome-wide data. Nucleic Acids Res 42(D1): D546–D552.

Kim Y & Nielsen R (2004) Linkage disequilibrium as a signature of selective sweeps. Genetics 167(3): 1513–1524.

Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267(5608): 275–276.

Kimura M (1983) The neutral theory of molecular evolution. Cambridge, Cambridge University Press.

Kingman JF (1982a) On the genealogy of large populations. J Appl Prob 19(A): 27–43.

Kingman JFC (1982b) The coalescent. Stochastic processes and their applications 13(3): 235–248.

Kivimäki M, Kärkkäinen K, Gaudeul M, Løe G & Ågren J (2007) Gene, phenotype and function: *GLABROUS1* and resistance to herbivory in natural populations of *Arabidopsis lyrata*. Mol Ecol 16(2): 453–462.

Koenig D & Weigel D (2015) Beyond the thale: Comparative genomics and genetics of *Arabidopsis* relatives. Nat Rev Genet 16(5): 285–298.

Korneliussen TS, Albrechtsen A & Nielsen R (2014) ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinform 15(1): 356

Koskela EA, Mouhu K, Albani MC, Kurokura T, Rantanen M, Sargent DJ, Battey NH, Coupland G, Elomaa P & Hytönen T (2012) Mutation in *TERMINAL FLOWER1* reverses the photoperiodic requirement for flowering in the wild strawberry *Fragaria vesca*. Plant Physiol 159(3): 1043–1054.

Kubota A, Shim JS & Imaizumi T (2015) Natural variation in transcriptional rhythms modulates photoperiodic responses. Trends Plant Sci, 20(5): 259–261.
Kuittinen H, Niittyvuopio A, Rinne P & Savolainen O (2008) Natural variation in *Arabidopsis lyrata* vernalization requirement conferred by a *FRIGIDA* indel polymorphism. Mol Biol Evol 25(2): 319–329.
Kullman L (2002) Boreal tree taxa in the central Scandes during the Late-Glacial: Implications for Late-Quaternary forest history. J Biogeogr 29(9): 1117–1124.
Le Corre V & Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. Mol Ecol 21(7): 1548–1566.
Le Corre V, Roux F & Reboud X (2002) DNA polymorphism at the *FRIGIDA* gene in *Arabidopsis thaliana*: Extensive nonsynonymous variation is consistent with local selection for flowering time. Mol Biol Evol 19(8): 1261–1271.
Leinonen PH, Remington DL & Savolainen O (2011) Local adaptation, phenotypic differentiation, and hybrid fitness in diverged natural populations of *Arabidopsis lyrata*. Evolution 65(1): 90–107.
Leinonen PH, Sandring S, Quilot B, Clauss MJ, Mitchell-Olds T, Ågren J & Savolainen O (2009) Local adaptation in European populations of *Arabidopsis lyrata* (Brassicaceae). Am J Bot 96(6): 1129–1137.
Levy YY & Dean C (1998) The transition to flowering. Plant Cell 10(12): 1973–1989.
Lewontin RC & Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74(1): 175–195.
Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14): 1754–1760.
Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. URI: https://arxiv.org/abs/1303.3997. Cited 2016/12/23.
Li J, Li H, Jakobsson M, Li S, Sjödin P & Lascoux M (2012) Joint analysis of demography and selection in population genetics: Where do we stand and where could we go? Mol Ecol 21(1): 28–44.
Linnen CR, Kingsley EP, Jensen JD & Hoekstra HE (2009) On the origin and spread of an adaptive allele in deer mice. Science 325(5944): 1095–1098.
Liu X & Fu YX (2015) Exploring population size changes using SNP frequency spectra. Nat Genet 47(5): 555–559.
Maynard Smith J & Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23(01): 23–35.
McVean G (2007) The structure of linkage disequilibrium around a selective sweep. Genetics 175(3): 1395–1406.
Meinke DW, Cherry JM, Dean C, Rounsley SD & Koornneef M (1998) *Arabidopsis thaliana*: A model plant for genome analysis. Science 282(5389): 679–682.
Menzel M, Sletvold N, Ågren J, & Hansson B (2015) Inbreeding affects gene expression differently in two self-incompatible *Arabidopsis lyrata* populations with similar levels of inbreeding depression. Mol Biol Evol *32*(8): 2036–2047.
Messer PW & Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol 28(11): 659–669.

Michaels SD & Amasino RM (1999) *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell 11(5): 949–956.

Mitchell-Olds T (2001) *Arabidopsis thaliana* and its wild relatives: A model system for ecology and evolution. Trends Ecol Evol 16(12): 693–700.

Mittler R, Kim Y, Song L, Coutu J, Coutu A, Ciftci-Yilmaz S, Lee H, Stevenson B & Zhu JK (2006) Gain- and loss-of-function mutations in *Zat10* enhance the tolerance of plants to abiotic stress. FEBS Lett 580(28–29): 6537–6542.

Mouhu K, Kurokura T, Koskela EA, Albert VA, Elomaa P & Hytönen T (2013) The *Fragaria vesca* homolog of SUPPRESSOR OF OVEREXPRESSION OF CONSTANS1 represses flowering and promotes vegetative growth. Plant Cell 25(9): 3296–3310.

Muller MH, Leppälä J & Savolainen O (2008) Genome-wide effects of postglacial colonization in *Arabidopsis lyrata*. Heredity 100(1): 47–58.

Nei M & Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. PNAS 76(10): 5269–5273.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y & Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PLoS ONE 7(7): e37558.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG & Bustamante C (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15(11): 1566–1575.

Nielsen R & Slatkin M (2013) An introduction to population genetics: theory and applications. Sunderland Massachusetts, Sinauer Associates, Inc.

Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, Muir G, Guggisberg A, Paape T, Schmid K, Fedorenko OM, Holm S, Sall T, Schlotterer C, Marhold K, Widmer A, Sese J, Shimizu KK, Weigel D, Kramer U, Koch MA & Nordborg M (2016) Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nat Genet 48(9): 1077–1082.

Ohta T (1973) Slightly deleterious mutant substitutions in evolution. Nature 246(5428): 96–98.

O'Kane Jr. SL & Al-Shehbaz IA (1997) A synopsis of Arabidopsis (Brassicaceae). Novon 7(3): 323–327.

Ometto L, Glinka S, De Lorenzo D & Stephan W (2005) Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol Biol Evol 22(10): 2119–2130.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D & Lynch M (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science 327(5961): 92–94.

Parducci L, Edwards ME, Bennett KD, Alm T, Elverland E, Tollefsrud MM, Jørgensen T, Houmark-Nielsen M, Larsen NK, Kjær KH, Fontana SL, Alsos IG & Willerslev E (2012a) Response to comment on "Glacial survival of boreal trees in Northern Scandinavia". Science 338(6108): 742.

Parducci L, Jørgensen T, Tollefsrud MM, Elverland E, Alm T, Fontana SL, Bennett KD, Haile J, Matetovici I, Suyama Y, Edwards ME, Andersen K, Rasmussen M, Boessenkool S, Coissac E, Brochmann C, Taberlet P, Houmark-Nielsen M, Larsen NK, Orlando L, Gilbert MTP, Kjær KH, Alsos IG & Willerslev E (2012b) Glacial survival of boreal trees in northern Scandinavia. Science 335(6072): 1083–1086.

Pfeifer SP, Laurent S, Sousa VC, Linnen CR, Foll M, Excoffier L, Hoekstra HE, Jensen JD (2017) The evolutionary history of Nebraska deer mice: local adaptation in the face of strong gene flow. bioRxiv 152694; doi: https://doi.org/10.1101/152694.

Pritchard JK & Di Rienzo A (2010) Adaptation - Not by sweeps alone. Nature Rev Genet 11(10): 665–667.

Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 160(3): 1179–1189.

Putterill J, Laurie R & Macknight R (2004) It's time to flower: The genetic control of flowering time. Bioessays 26(4): 363–373.

Pyhäjärvi T, Aalto E & Savolainen O (2012) Time scales of divergence and speciation among natural populations and subspecies of *Arabidopsis lyrata* (Brassicaceae). Am J Bot 99(8): 1314–1322.

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URI: https://www.R-project.org/. Cited 2017/10/10.

Ramos-Onsins SE, Stranger BE, Mitchell-Olds T & Aguadé M (2004) Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. Genetics 166(1): 373–388.

Remington DL, Figueroa J & Rane M (2015) Timing of shoot development transitions affects degree of perenniality in *Arabidopsis lyrata* (Brassicaceae). BMC Plant Biol 15(1): 226.

Riihimäki M & Savolainen O (2004) Environmental and genetic effects on flowering differences between northern and southern populations of *Arabidopsis lyrata* (Brassicaceae). Am J Bot 91(7): 1036–1045.

Ross-Ibarra J, Wright SI, Foxe JP, Kawabe A, DeRose-Wilson L, Gos G, Charlesworth D & Gaut BS (2008) Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. PLoS ONE 3(6): e2411

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D & Lander ES (2006) Positive natural selection in the human lineage. Science 312(5780): 1614–1620.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, You QS, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, De Bakker, P I W, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'Er I, Price A, Purcell S, Richter DJ, Saxena R, Sham PC, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Johnson TA, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Matsuda I, Fukushima Y, MacEr DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Yakub I, Birren BW, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R & Stewart J (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164): 913–918.

Sanjuán R, Moya A & Elena SF (2004) The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. PNAS 101(22): 8396–8401.

Savolainen O, Pyhäjärvi T & Knürr T (2007) Gene flow and local adaptation in trees. Annu Rev Ecol Evol Syst 38:595–619.

Savolainen O & Kuittinen H (2011) *Arabidopsis lyrata* Genetics. In: Schmidt R & Bancroft I (eds) Genetics and Genomics of the Brassicaceae. New York, Springer: 347–372.

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(1): 98.

Schneider A, Charlesworth B, Eyre-Walker A & Keightley PD (2011) A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189(4): 1427–1437.

Seppä H, Nyman M, Korhola A & Weckstrm J (2002) Changes of treelines and alpine vegetation in relation to post-glacial climate dynamics in northern Fennoscandia based on pollen and chironomid records. J Quat Sci 17(4): 287–301.

Sheehan S & Song YS (2016) Deep learning for population genetic inference. PLoS Comput Biol 12(3): e1004845.

Shimizu KK, Fujii S, Marhold K, Watanabe K & Kudoh H (2005) *Arabidopsis kamchatica* (Fisch. ex DC.) K. Shimizu & Kudoh and *A. kamchatica* subsp. *kawasakiana* (Makino) K. Shimizu & Kudoh, new combinations. Acta Phytotax Geobot 56(2): 163–172.

Shimizu-Inatsugi R, LihovÁ J, Iwanaga H, Kudoh H, Marhold K, Savolainen O, Watanabe K, Yakubov VV & Shimizu KK (2009) The allopolyploid *Arabidopsis kamchatica* originated from multiple individuals of *Arabidopsis lyrata* and *Arabidopsis halleri*. Mol Ecol 18(19): 4024–4048.

Shindo C, Aranzana MJ, Lister C, Baxter C, Nicholls C, Nordborg M & Dean C (2005) Role of *FRIGIDA* and *FLOWERING LOCUS C* in determining variation in flowering time of Arabidopsis. Plant Physiol 138(2): 1163–1173.

Siol M, Wright SI & Barrett SCH (2010) The population genomics of plant adaptation. New Phytol 188(2): 313–332.

Skotte L, Korneliussen TS & Albrechtsen A (2013) Estimating individual admixture proportions from next generation sequencing data. Genetics 195(3): 693–702.

Sletvold N & Ågren J (2012) Variation in tolerance to drought among Scandinavian populations of *Arabidopsis lyrata*. Evol Ecol 26(3): 559–577.

Slotte T, Foxe JP, Hazzouri KM & Wright SI (2010) Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. Mol Biol Evol 27(8): 1813–1821.

Slotte T, Hazzouri KM, gren JA, Koenig D, Maumus F, Guo YL, Steige K, Platts AE, Escobar JS, Newman LK, Wang W, Mandkov T, Vello E, Smith LM, Henz SR, Steffen J, Takuno S, Brandvain Y, Coop G, Andolfatto P, Hu TT, Blanchette M, Clark RM, Quesneville H, Nordborg M, Gaut BS, Lysak MA, Jenkins J, Grimwood J, Chapman J, Prochnik S, Shu S, Rokhsar D, Schmutz J, Weigel D & Wright SI (2013) The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat Genet 45(7): 831–835.

St. Onge KR, Källman T, Slotte T, Lascoux M & Palmé AE (2011) Contrasting demographic history and population structure in *Capsella rubella* and *Capsella grandiflora*, two closely related species with different mating systems. Mol Ecol 20(16): 3306–3320.

Steige KA, Laenen B, Reimegrd J, Scofield DG & Slotte T (2017) Genomic analysis reveals major determinants of cis-regulatory variation in *Capsella grandiflora*. PNAS 114(5): 1087–1092.

Stephan W (2016) Signatures of positive selection: From selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. Mol Ecol 25(1): 79–88.

Stephens M & Scheet P (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76(3): 449–462.

Stephens M, Smith NJ & Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68(4): 978–989.

Stinchcombe JR, Weinig C, Ungerer M, Olsen KM, Mays C, Halldorsdottir SS, Purugganan MD & Schmitt J (2004) A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene *FRIGIDA*. PNAS 101(13): 4712–4717.

Strasburg JL, Kane NC, Raduski AR, Bonin A, Michelmore R & Rieseberg LH (2011) Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Mol Biol Evol 28(5): 1569–1580.

Sturm RA & Duffy DL (2012) Human pigmentation genes under environmental selection. Genome Biol 13(9): 248

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105(2): 437–460.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3): 585–595.

The Arabidopsis Information Resource (2017). URI: www.arabidopsis.org. Cited 2017/5/12.

Toivainen T, Pyhäjärvi T, Niittyvuopio A & Savolainen O (2014) A recent local sweep at the PHYA locus in the Northern European Spiterstulen population of *Arabidopsis lyrata*. Mol Ecol 23(5): 1040–1052.

Tollefsrud MM, Kissling R, Gugerli F, Johnsen, Skrppa T, Cheddadi R, Van Der Knaap, W O, Latalowa M, Terhrne-Berson R, Litt T, Geburek T, Brochmann C & Sperisen C (2008) Genetic consequences of glacial survival and postglacial colonization in Norway spruce: Combined analysis of mitochondrial DNA and fossil pollen. Mol Ecol 17(18): 4134–4150.

Tyrmi J (2016) STAPLER – Simple and Swift Bioinformatics Pipeline Maker. URI: https://github.com/tyrmi/STAPLER. Cited 2017/10/10.

Tzedakis PC, Emerson BC & Hewitt GM. (2013) Cryptic or mystic? Glacial tree refugia in northern Europe. Trends Ecol Evol 28(12): 696–704.

Vergeer P & Kunin WE (2013) Adaptation at range margins: Common garden trials and the performance of *Arabidopsis lyrata* across its northwestern European range. New Phytol 197(3): 989–1001.

Videvall E, Sletvold N, Hagenblad J, Ågren J, & Hansson B (2015) Strong maternal effects on gene expression in *Arabidopsis lyrata* hybrids. Mol Biol Evol, *33*(4), 984–994.

Vigueira CC, Rauh B, Mitchell-Olds T & Lawton-Rauh AL (2013) Signatures of demography and recombination at coding genes in naturally-distributed populations of *Arabidopsis lyrata* subsp. *petraea*. PLoS ONE 8(3): e58916.

Voight BF, Kudaravalli S, Wen X & Pritchard JK (2006) A Map of recent positive selection in the human genome. PLoS Biol 4(3): e72.

Wakeley J (2008) Coalescent theory: An introduction. Greenwood Village, CO, Roberts & Company Publishers.

Wang BB & Brendel V (2006) Molecular characterization and phylogeny of U2AF^{35} homologs in plants. Plant Physiol 140(2): 624–636.

Wang J, Street NR, Scofield DG & Ingvarsson PK (2016) Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. Genetics 202(3): 1185–1200.

Wang R, Farrona S, Vincent C, Joecker A, Schoof H, Turck F, Alonso-Blanco C, Coupland G & Albani MC (2009) *PEP1* regulates perennial flowering in *Arabis alpina*. Nature 459(7245): 423–427.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7(2): 256–276.

Weinberg W (1908) Über Den Nachweis Der Vererbung Beim Menschen. Jh Ver vaterl Naturk Wurttemb 64: 369–382.

Westergaard KB, Alsos IG, Popp M, Engelskj T, Flatberg KI & Brochmann C (2011) Glacial survival may matter after all: Nunatak signatures in the rare European populations of two west-arctic species. Mol Ecol 20(2): 376–393.

Willerslev E, Davison J, Moora M, Zobel M, Coissac E, Edwards ME, Lorenzen ED, Vestergrd M, Gussarova G, Haile J, Craine J, Gielly L, Boessenkool S, Epp LS, Pearman PB, Cheddadi R, Murray D, Brthen KA, Yoccoz N, Binney H, Cruaud C, Wincker P, Goslar T, Alsos IG, Bellemain E, Brysting AK, Elven R, Snsteb JH, Murton J, Sher A, Rasmussen M, Rnn R, Mourier T, Cooper A, Austin J, Mller P, Froese D, Zazula G, Pompanon F, Rioux D, Niderkorn V, Tikhonov A, Savvinov G, Roberts RG, Macphee RDE, Gilbert MTP, Kjr KH, Orlando L, Brochmann C & Taberlet P (2014) Fifty thousand years of Arctic vegetation and megafaunal diet. Nature 506(7486): 47–51.

Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M & Wright SI (2014) Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. PLoS Genet 10(9): e1004622.

Wright S (1951) The genetical structure of populations. Ann Eugen 15(4): 323–354.

Wright SI & Charlesworth B (2004) The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. Genetics 168(2): 1071–1076.

Wright SI, Lauga B & Charlesworth D (2003) Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. Mol Ecol 12(5): 1247–1263.

Wright S (1931) Evolution in Mendelian populations. Genetics 16(2): 97–159.

Wright SI, Lauga B & Charlesworth D (2002) Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. Mol Biol Evol 19(9): 1407–1420.

Wright S, Ness R, Foxe J & Barrett S (2008) Genomic consequences of outcrossing and selfing in plants. Int J Plant Sci 169(1): 105–118.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu N, Jiang H, Vinckenbosch N, Korneliussen TS, Zheng H, Liu T, He W, Li K, Luo R, Nie X, Wu H, Zhao M, Cao H, Zou J, Shan Y, Li S, Yang Q, Asan, Ni P, Tian G, Xu J, Liu X, Jiang T, Wu R, Zhou G, Tang M, Qin J, Wang T, Feng S, Li G, Huasang, Luosang J, Wang W, Chen F, Wang Y, Zheng X, Li Z, Bianba Z, Yang G, Wang X, Tang S, Gao G, Chen Y, Luo Z, Gusang L, Cao Z, Zhang Q, Ouyang W, Ren X, Liang H, Zheng H, Huang Y, Li J, Bolund L, Kristiansen K, Li Y, Zhang Y, Zhang X, Li R, Li S, Yang H, Nielsen R, Wang J & Wang (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329(5987): 75–78.

Yokoyama Y, Lambeck K, De Deckker P, Johnston P & Fifield LK (2000) Timing of the Last Glacial Maximum from observed sea-level minima. Nature 406(6797): 713–716.

Appendix

Command lines for demographic simulations (Fig. 2)

```
# requires ms, msstats and mssfs
# constant model
ms 25 1000 -t 280 -r 400 100000 | msstats > SNM_stats.txt
ms 25 1000 -t 280 -r 400 100000 | mssfs > SNM_sfs.txt

# decline model
ms 25 1000 -t 28 -r 40 100000 -eN 0.5 10 | msstats > DM_stats.txt
ms 25 1000 -t 28 -r 40 100000 -eN 0.5 10 | mssfs > DM_sfs.txt

# growth model
ms 25 1000 -t 2800 -r 4000 100000 -eN 0.005 0.1 | msstats >
GM_stats.txt
ms 25 1000 -t 2800 -r 4000 100000 -eN 0.005 0.1  | mssfs > GM_sfs.txt

# bottleneck model
ms 25 1000 -t 280 -r 400 100000 -eN 0.025 0.5 -eN 0.05 1 | msstats >
BM_stats.txt
ms 25 1000 -t 280 -r 400 100000 -eN 0.025 0.5 -eN 0.05 1| mssfs >
BM_sfs.txt
```

Original publications

I Mattila TM, Tyrmi J, Pyhäjärvi T, Savolainen O (2017) Genome-wide analysis of colonization history and concomitant selection in *Arabidopsis lyrata*. Mol Biol Evol 34(10): 2665–2677.

II Mattila TM, Aalto EA, Toivainen T, Niittyvuopio A, Piltonen S, Kuittinen H, Savolainen O (2016) Selection for population-specific adaptation shaped patterns of variation in the photoperiod pathway genes in *Arabidopsis lyrata* during post-glacial colonization. Mol Ecol 25(2): 581–597.

III Mattila TM, Laenen B, Hämälä T, Savolainen O, Slotte T (2017) The genome-wide impact of selection in two outcrossing Brassicaceae species with contrasting demographic history. Manuscript.

Original publications are not included in the electronic version of the dissertation.

Book orders:
Granum: Virtual book store
http://granum.uta.fi/granum/

UNIVERSITY OF OULU P.O. Box 8000 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

UNIVERSITY of OULU
OULUN YLIOPISTO

ISBN 978-952-62-1708-6 (Paperback)
ISBN 978-952-62-1709-3 (PDF)
ISSN 0355-3191 (Print)
ISSN 1796-220X (Online)