OULU 2015
A 661

ACTA UNIVERSITATIS OULUENSIS

Sonja Kujala

DISSECTING GENETIC VARIATION IN EUROPEAN SCOTS PINE (*PINUS SYLVESTRIS* L.) – SPECIAL EMPHASIS ON POLYGENIC ADAPTATION

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

ACTA UNIVERSITATIS OULUENSIS
A Scientiae Rerum Naturalium 661

SONJA KUJALA

DISSECTING GENETIC VARIATION IN EUROPEAN SCOTS PINE (*PINUS SYLVESTRIS* L.) – SPECIAL EMPHASIS ON POLYGENIC ADAPTATION

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 Kuusamonsali (YB210), Linnanmaa, on 11 December 2015, at 12 noon

UNIVERSITY OF OULU, OULU 2015

Acta Univ. Oul. A 661, 2015

Supervised by
Professor Outi Savolainen
Professor Katri Kärkkäinen

Reviewed by
Professor Teemu Teeri
Professor Peter Tiffin

Opponent
Assistant Professor Andrew Eckert

ISBN 978-952-62-1035-3 (Paperback)
ISBN 978-952-62-1036-0 (PDF)

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

Cover Design
Raimo Ahonen

JUVENES PRINT
TAMPERE 2015

Kujala, Sonja, Dissecting genetic variation in European Scots pine (*Pinus sylvestris* L.) – special emphasis on polygenic adaptation

University of Oulu Graduate School; University of Oulu, Faculty of Science; University of Oulu, Biocenter Oulu

Acta Univ. Oul. A 661, 2015

University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

Adaptation through polygenic selection is a prominent feature in nature. Still, the genetic backgrounds of polygenic adaptations are often unknown. The challenges of resolving adaptive processes are related to selection being distributed over several loci with often small effect sizes. Also, even a low level of population substructure can obstruct the inference. Further, demographic factors in the history of the species, such as population size changes and range expansions leave a confounding footprint in the background genomic variation. In this thesis, polygenic adaptation was studied with Scots pine (*Pinus sylvestris* L.), a widespread ecologically and economically important conifer.

In this thesis, timing of bud set – an adaptive polygenic trait – was studied at the level of the phenotype in a common garden study, and at the genomic level by examining the sequence and allele frequency variation patterns in bud set timing related loci, with a sampling across a latitudinal transect in Europe. An association study, combining these two levels, was carried out with a new Bayesian multipopulation method. The congruence of allozyme and nucleotide level diversity was estimated, the level of neutral genetic population structure surveyed, and a demographic background model for statistical inference of selective signals redefined.

Allozyme variation seemed to correlate well with the nucleotide level variation at the between species level, but within population, at the individual allozyme coding loci, allozyme heterozygosity does not describe the underlying level of nucleotide variation well. Indications of recent colonization history affecting the level of differentiation between populations were seen, and the need to control for the background effects of simultaneous range expansion and adaptation shown. Lower phenotypic and additive genetic variation in timing of bud set was found in northern compared to central European populations. Signs of heterogeneity in genetic basis of this trait were also found between these areas, which could indicate different timekeeping mechanisms due to different environmental cues in the two regions. The results in this thesis are of value to the study of adaptation, but also for breeding, conservation and prediction of responses of forest trees to future climate change.

Keywords: adaptation, allelic covariance, allozyme, association, cline, demography, F_{ST}, genetic heterogeneity, Pinus sylvestris, polygenic

Kujala, Sonja, Geneettinen vaihtelu eurooppalaisessa metsämännyssä (*Pinus sylvestris* L.) – erityistarkastelussa polygeeninen sopeutuminen

Oulun yliopiston tutkijakoulu; Oulun yliopisto, Luonnontieteellinen tiedekunta; Oulun yliopisto, Biocenter Oulu

Acta Univ. Oul. A 661, 2015
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Sopeutuminen perustuu usein polygeenisiin ominaisuuksiin. Näiden ominaisuuksien geneettiset taustat ovat silti vielä pitkälti selvittämättä. Sopeutumisominaisuuksien genetiikan selvittämistä vaikeuttaa valinnan vaikutusten jakautuminen usean, usein pienivaikutuksisen lokuksen kesken. Lisäksi vähäinenkin populaatiorakenne hankaloittaa geenien tunnistamista. Myös lajin historiassa tapahtuneet demografiset muutokset, kuten populaatiokoon vaihtelut ja kolonisaatio jättävät jälkensä genomiin. Väitöskirjassani tutkin polygeenistä sopeutumista ekologisesti ja taloudellisesti tärkeän havupuulajin, metsämännyn (*Pinus sylvestris* L.) avulla.

Väitöskirjassani tutkin metsämännyn silmunmuodostuksen ajoitusta sekä fenotyypin että sekvenssimuuntelun tasolla. Ajoitusta mitattiin eri leveysasteilta peräisin olevista eurooppalaisista populaatioista yhteiskasvatuskokeessa, ja sekvenssimuuntelua sekä alleelifrekvenssien jakautumista tutkittiin vastaavasta näytteestä. Geenikartoituskokeessa yhdistettiin nämä kaksi muuntelun tasoa hyödyntäen uutta, usean populaation tutkimiseen soveltuvaa analyysimenetelmää. Lisäksi tutkin allotsyymimuuntelun ja nukleotidimuuntelun keskinäistä tarkkuutta geneettisen diversiteetin kuvaajina, neutraalin populaatiorakenteen tasoa, sekä demografian vaikutusta metsämännyn genomissa.

Allotsyymimuuntelun todettiin kuvaavan hyvin lajien välisiä diversiteettieroja. Populaation sisällä yksittäisten allotsyymien heterosygotia ei korreloinut entsyymiä koodaavan geenin muuntelun määrän kanssa. Pohjoisten populaatioiden vähäisemmät keskinäiset erot verrattuna keski-eurooppalaisiin antoivat viitteitä siitä, että viimeisimmän jääkauden jälkeiset kolonisaatiotapahtumat voivat edelleen vaikuttaa populaatioiden erilaistumisasteeseen. Assosiaatiotutkimuksessa osoitettiin, kuinka tärkeää yhtäaikaisen sopeutumisen ja kolonisaation huomioiminen on sopeutumisominaisuuksien tutkimisessa. Fenotyyppinen muuntelu silmunmuodostuksen ajoituksessa oli vähäisempää pohjoisissa populaatioissa. Lisäksi löysimme merkkejä geneettisestä heterogeenisuudesta silmunmuodostuksen taustalla pohjoisten ja keskieurooppalaisten metsämäntyjen välillä, mikä voi johtua vuodenajan vaihtelun mittaamiseen käytettävien ympäristösignaalien eriytymisestä näiden alueiden välillä. Väitöskirjassani saadut tulokset hyödyttävät paitsi sopeutumistutkimusta, jalostus- ja luonnonsuojelututkimusta sekä ilmastonmuutoksen vaikutusten arviointia.

Asiasanat: alleelikovarianssi, allotsyymi, assosiaatio, demografia, F_{ST}, geneettinen heterogeenisuus, kliini, Pinus sylvestris, polygeeninen, sopeutuminen

To my beautiful daughters

Acknowledgements

I want to start by thanking my main supervisor and coauthor professor Outi Savolainen for taking me onboard to the plant genetics group. I am very grateful for your expert guidance in this challenging field of research. You have been most patient and encouraging, thank you for having faith in me! I also thank my other supervisor professor Katri Kärkkäinen for collaboration and coauthorship. Your expertise on forest genetics has been very valuable and motivating.

I want to thank professor Teemu Teeri and professor Peter Tiffin for reviewing this thesis. My coauthors Tanja Pyhäjärvi, Timo Knürr, Mikko Sillanpää and David Neale are highly appreciated for their collaboration. I am grateful also for Natural Resources Institute Finland for cooperation. I acknowledge financial support from EU projects Treesnips and Evoltree, Department of Genetics and Physiology, Biocenter Oulu Doctoral Programme, University of Oulu Graduate School, Finnish Graduate School in Population Genetics and the Emil Aaltonen foundation.

A warm thank you for all the current and former coworkers, especially my "brothers and sisters in pine" Tanja, Timo, Komlan, Matti, Yongfeng and Jaakko. Also Tiina, Lumi and Kukka, to mention just a few, are thanked for their companionship. I have had many fun moments with all of you over the years, not to mention the value of discussing various science and life issues. I warmly thank Soile Alatalo for all the laboratory work she has done for the studies in this thesis, and also for great, fun moments. The department of Genetics and Physiology has offered a professional and comfortable environment for learning.

Thank you family and friends! Mother and Father, your help in various aspects of life through different times has been invaluable. You have always supported me when I wanted to see how far I can reach, and also encouraged me to finish what I have started. And yet you have always made me feel that if I fall, you will catch me. Kerttu and Jaakko, thank you for your help and support, too. Jaana, thank you for being my friend, for all your help, and for helping me see things in so many different lights.

Heikki, we have practically grown up together, and gone through many things in life together. And many more to come! You have been very patient through these final stages of my thesis work. For better, for worse, My Love! Anni, Iiris and Ella, my best work on inheritance, you are my most precious treasures <3

Oulu, November 2015 Sonja Kujala

Abbreviations

ABC	approximate Bayesian computation
ADT	Assay Design Tool
BLUP	best linear unbiased predictor
bp	base pair
CRSP	Comparative Re-Sequencing in Pinaceae
DNA	deoxyribonucleotide acid
EMBR	extended bottleneck model with recombination
EST	expressed sequence tag
F_{ST}	population differentiation at genetic markers
GWAS	genome wide association study
LGM	last glacial maximum
MAF	minor allele frequency
PEMR	population expansion model with recombination
SFS	site frequency spectrum
SNMR	standard neutral model with recombination
SNP	single nucleotide polymorphism
SPC	selection, pleiotropy and compensation
Q_{ST}	population differentiation at quantitative traits
QTL	quantitative trait locus

List of original articles

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

I Pyhäjärvi T*, Kujala ST* & Savolainen O (2011) Revisiting protein heterozygosity in plants – nucleotide diversity in allozyme coding genes of conifer *Pinus sylvestris*. Tree Genetics & Genomes 7: 385–397.

II Kujala ST & Savolainen O (2012) Sequence variation patterns along a latitudinal cline in Scots pine (*Pinus sylvestris*): signs of clinal adaptation? Tree Genetics & Genomes 8: 1451–1467.

III Kujala ST*, Knürr T*, Kärkkäinen K, Neale DB, Sillanpää MJ & Savolainen O (2015) Indications of genetic heterogeneity in a locally adaptive clinal trait in *Pinus sylvestris* revealed by novel multipopulation association method. Manuscript.

IV Kujala ST & Savolainen O (2015) Shallow allele frequency clines and lack of allele frequency covariation in timing of bud set associated SNPs in Scots pine (*Pinus sylvestris*). Manuscript.

*authors contributed equally to the manuscript

Author contributions

Paper	Study design	Experiments	Data analyses	Manuscript preparation
I	OS, TP	**SK**, TP	TP, **SK**	TP, **SK**, OS
II	OS, **SK**	**SK**	**SK**	**SK**, OS
III	OS, KK, MS	KK, **SK**, DN	TK, **SK**	OS, **SK**, TK, MS
IV	OS, **SK**		**SK**	**SK**, OS

Sonja Kujala (SK), Outi Savolainen (OS), Tanja Pyhäjärvi (TP), Timo Knürr (TK), Mikko Sillanpää (MS), Katri Kärkkäinen (KK), David Neale (DN)

Table of contents

1 Introduction

The ability of organisms to adapt to different environmental conditions is one of the cornerstones of evolution. Adaptation can be seen all around us; bird species have adapted to different food sources with morphological changes in their beaks (Darwin 1845), Tibetan people have acquired features in their physiology that allow life in the high altitudes with low oxygen pressure (Lorenzo *et al.* 2014), bacteria are constantly evolving and gain resistance to new antibiotics (Toprak *et al.* 2012), only plants that have developed cold tolerance survive the northern winter (Oakley *et al*. 2014). Adaptation by definition requires that it has a genetic basis and that different genetic variants have different probabilities to survive and produce offspring in particular environmental conditions. The proportion of these favourable genetic variants will increase in the following generations if chance effects do not override this effect of natural selection. Adaptation can have a single gene basis, or be polygenic, meaning that variation in multiple genes contribute to the variation in a phenotypic trait that confers adaptation (Kawecki & Ebert 2004).

Adaptation often takes place simultaneously with colonization of new habitats (Davis & Shaw 2001). As an example, the retreating glacial ice at the end of the last glacial maximum (LGM), provided open land in Fennoscandia to be colonized. The environment was harsh, though, and any species that migrated northwards from more southern latitudes had to tolerate the cold. The seasonal changes between summer and winter were also more extreme than in the south, and the environmental cues that warn about the approaching winter were different. A species that already had genetic variants enabling some individuals to tolerate the cold and to interpret the different environmental cues had a better chance of colonizing the new space. Some new favourable mutations conferring these abilities might also have occurred while the species was spreading to north. Adaptation occurred as these variants increased in frequency in successive generations as the species spread towards new areas.

In this thesis I have studied climatic adaptation and genetic variation with Scots pine, a species that has migrated and adapted to northern latitudes in Fennoscandia after the last glacial maximum, but exists in various other environmental conditions as well. I will start by introducing the species in order to give an overall picture of the biological features of Scots pine important in the context of studying polygenic adaptation, followed by the theory and methodological considerations.

1.1 Scots pine (*Pinus sylvestris* L.)

Scots pine *(Pinus sylvestris*; genus *Pinaceae*, subgenus *Pinus*, section *Pinus*, subsection *Pinus*) is one of the several pine species occurring in Eurasia. *P. sylvestris* is the most widespread among all pines (altogether some 115 species in the world) with a wide distribution from Western Europe to the eastern parts of Siberia (Fig. 1). The species extends to the harsh northern latitudes in northern Finland, Sweden and Norway. The southernmost populations are found in Spain and Turkey. A large part of the range has a rather continuous distribution. Mostly in the southern parts of the range, the species splits into more isolated patches (Mirov 1967, Richardson & Rundel 1998). In some western part of the range, the species has gone extinct due to human activities, but has later been reintroduced.

A large number of varieties within this species have been suggested along the long history of Scots pine studies, but only three official varieties are recognized nowadays; var. *mongolica* (in Mongolia and close by regions in southern Siberia and China), var. *hamata* (in Balkans, northern Turkey and Caucasus) and var. *sylvestris*, which is the most widespread variety and covers the rest of the range (Farjon 1998). The plethora of unofficial varieties is often based on phenotypic characters, and most likely reflects the large phenotypic variation found in this species, and the obvious capacity to adapt to variety of environmental conditions (Giertych & Mátyás 2013).

1.1.1 Early demographic history of European Scots pine

The areas that Scots pine occupies nowadays have gone through dramatic climatic changes in the past as multiple glacial and interglacial periods have alternated (Petit 1999, Cheddadi *et al*. 2005), which have influenced the population size and distribution of Scots pine (and other species). Most likely recurrent extinction and recolonization events have followed these climatic oscillations throughout the history of the species. Population size fluctuations leave a footprint in the amount and/or pattern of nucleotide variation. Certain features in the genome, such as the amount of diversity, the site frequency spectrum (SFS) and extent of linkage disequilibrium can deviate from neutral expectations due to past demographic events. For example, a population expansion is known to cause an excess of low frequency variants, and a recent bottleneck a reduction of diversity and an excess of intermediate frequency variants. The effects of demography extend to the whole genome (Przeworski 2002, Huber *et al.* 2014).

Pyhäjärvi *et al.* (2007) studied the signs of past demography in the nuclear genome of Scots pine. They found that the patterns of variation fit a demographic model with an old, severe bottleneck. Similar finding is common among many other forest trees, too. In Norway spruce (Heuertz *et al*. 2006), European aspen (Ingvarsson 2008a) and Maritime pine (Lepoittevin 2009) similar old bottlenecks fit the nuclear sequence data. The inferred models surely are simplifications of the actual course of events (see Jesus *et al.* 2006). It seems, however, that at least for Scots pine, the effect of the recolonization events after the last glacial maximum are not seen in the SFS of nuclear sequence data, as similar demographic model were inferred to both northern Scots pine populations and populations closer to potential LGM refugia (Pyhäjärvi *et al.* 2007). Neutral evolution of long-lived species is slow, i.e. mutation-drift equilibrium is reached very slowly. Therefore signs of very old events are seen in the genome (Savolainen & Pyhäjärvi 2007).

1.1.2 Postglacial history and present day population structure

The last glacial maximum occurred 26500–19000 years ago (Clark *et al*. 2009). In Europe the glacial ice sheet covered Scandinavia and most of the British Isles with its southern edge running through Germany and Poland (Svendsen *et al.* 2004). Permafrost extended even further south. Post-LGM events have been studied with maternally inherited organelle DNA markers in Scots pine (Sinclair *et al.* 1999, Cheddadi *et al*. 2006, Naydenov *et al.* 2007, Pyhäjärvi *et al.* 2008) and in other trees (e.g. Ferris *et al.* 1998, Petit *et al.* 2002, 2003) along with fossil evidence (e.g. Willis & van Andel 2004). These studies have suggested that cold tolerant trees, such as Scots pine, might have had LGM refugia quite close to the edge of the ice sheet, and that more southern refugia in Spain and Turkey have not contributed to the recolonization of central and northern Europe. Northern Europe was colonized within the last 10 000 years (Huntley & Birks 1983).

In general, nuclear markers are considered to be less informative about the recent colonization history than maternal organelle markers that are transferred through seeds only. They are, however, useful in describing the current state of differentiation between populations that depends on both the sharing of common ancestors and the amount of effective gene flow since the beginning of spatial separation (Whitlock & McCauley 1999). Nuclear allozyme markers (Gullberg *et al.* 1985, Muona & Smidt 1985, Muona & Harju 1989, Wang *et al.* 1991, Goncharenko *et al.* 1994, Prus-Glowacki & Stephan 1994, Shigapov *et al.* 1995, Puglisi & Attolico 2000, Dvornyk 2001), microsatellites (Karhu *et al.* 1996) and

sequence data (Dvornyk *et al.* 2002, Garcia-Gil *et al.* 2003, Pyhäjärvi *et al.* 2007, Wachowiak *et al*. 2009) have indicated very little population structure among current European Scots pine populations when estimated through F_{ST} (generally below 0.02). Especially the northernmost populations are very undifferentiated from each other. Gullberg *et al.* (1985) and Dvornyk *et al.* (2001) have suggested that this reflects their shorter occupation time when compared to the central populations that have had slightly more time to diverge. The southern isolated Spanish populations have been shown to be more differentiated from the rest of the Europe (Prus-Glowacki & Stephan 1994, Dvornyk *et al.* 2002, Pyhäjärvi *et al*. 2007).

1.1.3 Scots pine as a study species

Scots pine is a long lived tree with an average age at first reproduction of approximately 25 years. As it is an economically important species, many results have been obtained from growth trials, such as the transfer trial data of Eiche (1966). Scots pine disperses through seeds, and the pollen is wind dispersed. Pollen dispersal has been estimated both by pollen capture and by genetic methods. Long-distance pollen dispersal occurs, but most of the pollen lands within few hundred meters at most (Koski 1970, Robledo-Arnunzio & Gil 2005, Robledo-Arnunzio 2011). Seeds disperse less than pollen. Migration capacity must, however, be significant to explain the rapid colonization of the north indicated in pollen data (Austerlitz *et al.* 2000, Austerlitz & Garnier-Géré 2003, Mimura & Aitken 2007, Savolainen *et al.* 2011).

The genome of Scots pine is very large, approximately 22.4 gigabases (Plant DNA C-values Database, release 5.0, December 2010), which is approximately seven times the size of the human genome. A huge genome size is a typical feature of conifers. As a comparison, the genome of an angiosperm tree *Populus trichocarpa* is approximately 485 megabases (Tuskan *et al.* 2006). The number of genes in conifers (about 50000) is however not different from an average angiosperm. Most of conifer genomes consist of repetitive retrotransposon sequences, which can cause some difficulties in molecular genetics. There is, however, no evidence of recent genome duplications (Kovach *et al.* 2010, Nystedt *et al*. 2013, Wegrzyn *et al.* 2014, reviewed in De La Torre *et al*. 2014).

In earlier diversity studies with allozyme and microsatellite markers, the genetic diversity of Scots pine, as of many other forest trees, has been considered among the highest of all plants (Hamrick & Godt 1996). Early sequencing studies

later suggested that nucleotide diversity is not higher than in other plants (approximately 0.005/bp in Pyhäjärvi *et al*. 2007, see also Savolainen & Pyhäjärvi 2007). This is, however, at least five times more than in humans (Cargill *et al*. 1999, The 1000 Genomes Project Consortium 2015). Linkage disequilibrium generally decays rapidly, within a few hundred base pairs. So far, estimates for larger distances (beyond genes) are not available. The fact that linkage disequilibrium does not extend far is both an advantage and a challenge in association mapping studies (described below). Accuracy of mapping improves as the markers recognized in the analysis must reside close to the causal polymorphism. A high density of markers is, however, required to adequately cover the areas of interest (Neale & Savolainen 2004).

1.2 Adaptive traits are often polygenic

In contrast to the generally low F_{ST} estimates, there is ample phenotypic differentiation between populations of Scots pine and in other trees as well; genetic differentiation between populations of many adaptive and economically important traits (such as growth, cold tolerance, morphology) are high (Morgenstern 1996). These traits have a complex, polygenic genetic background with moderate to high heritability (Howe *et al*. 2003, Savolainen *et al.* 2007). Selection on the traits is in the form of diversifying selection towards different optima along environmental gradients. At the local level selection is stabilizing; the individuals closest to the local optimum have the highest fitness.

1.2.1 Theory on polygenic traits

Much theory on polygenic traits exists – traditionally within the framework of quantitative genetics – that can be applied to adaptive polygenic traits. A large number of loci are expected to contribute to the trait variation (Fisher 1930, Turchin *et al*. 2012). The effect sizes vary. Exponential distribution (with few loci with big effects and a large number of small effect loci) is often expected after an episode of directional selection (Orr 2005, Alonso-Blanco & Méndez-Vigo 2014). Genetic redundancy occurs, i.e. the same phenotypic value can be gained with multiple different genotypic combinations (Wright 1935, Goldstein & Holsinger 1992). Population genetics adds another layer of theory with issues such as migration-selection balance (Haldane 1930, Wright 1931), the origin of adaptive variation (adaptation from new mutations vs. adaptation from standing genetic variation,

Hermisson & Pennings 2005, Barrett & Schluter 2007) and the trajectory of beneficial alleles in time and space (Pavlidis *et al*. 2012, de Vladar & Barton 2014).

The underlying allele frequency changes in space (along an environmental gradient) has been modelled in detail by Slatkin (1973), Barton (1999), Bridle *et al.* (2010), Polechova & Barton (2011, 2015) and Geroldinger & Burger 2015. According to these models steep sequential allele frequency clines (from near zero to near fixation) are expected to form in part of the loci, while the rest of the loci remain near fixation. Each of the alleles is favoured towards a different end of the gradient. While the phenotypic optimum might change slowly along the environmental gradient, the individual allele frequency changes can occur at a narrower spatial scale. The allele frequency clines form in the timescale of few hundred generations, assuming fairly weak selection on individual loci (Barton 1999).

Latta (1998), Le Corre & Kremer (2003, 2012) and Kremer & Le Corre (2011) emphasize a different aspect of spatial processes, concentrating more on the early stages of adaptation. This model puts more weight on the positive allelic covariation and stresses that in many conditions selection on beneficial allele combinations is more important than the frequency changes at individual loci. The allele covariation originates from the between population component of linkage disequilibrium (Ohta 1982) and leads to a lack of significant differentiation, i.e. low between population F_{ST} estimates in the trait defining loci. The amount of gene flow between populations, the strength of diversifying (between populations) and stabilizing (within population) selection, the number of loci and time from the onset of selection all influence the amount of covariation. These studies are based on the island model (see also Merilä & Crnokrak 2001, McKay & Latta 2002, Leinonen *et al*. 2013).

1.2.2 Timing of bud set as an example

Timing of growth is another adaptive polygenic trait showing high differentiation, and has been studied in many tree species, often using timing of bud flush and timing of bud set as proxies for start and end of the active growing period (see e.g. Cooke *et al.* 2012). In Scots pine, as in other trees (Clapham *et al.* 1998, Viherä-Aarnio *et al.* 2005, Ingvarsson *et al.* 2006, Mimura & Aitken 2010), timing of bud set forms a latitudinal cline, studied in detail by Mikola (1982) in Finnish populations, and by Notivol *et al.* (2007) and Oleksyn *et al.* (1992). Timing of bud set is measured in the first year seedlings, but correlates with the end of the period

of active growth in older trees (Oleksyn *et al.* 1998). Photoperiod is the critical environmental cue that the tree uses to determine the proper time to end the yearly growth (Vaartaja 1959). The actual adaptation is to the interpretation of light conditions and varies according to the natural light environment in the place of origin at the time of the year when preparation for the winter must start. The genes are therefore expected to be found mainly from the light perception and timekeeping functional networks. These networks have been studied extensively in *Arabidopsis thaliana* (e.g. Andrés & Coupland 2012, Song *et al.* 2014) and in some crop plants (Nakamichi 2014). In coniferous trees the networks have been examined by Gyllenstrand *et al.* (2007), Lagercrantz (2009), Karlgren *et al.* (2011), Avia *et al.* (2014) and Gyllenstrand *et al.* (2014).

1.3 Searching for the genetic basis of an adaptive polygenic trait

The factors described above, i.e. the past demographic events, population structure, and the polygenic architecture of adaptive traits all influence the patterns of sequence variation in the trait controlling loci. Consequently they also influence our ability to successfully find the trait related genes with population genetic or association based methods (see e.g. Savolainen *et al.* 2013, Bank *et al.* 2014, Tiffin & Ross-Ibarra 2014, Pardo-Diaz *et al.* 2015). As an example, an excess of low frequency variants caused by past demographic events can affect the success of association tests, since the power is, in part, related to the allele frequencies (Long & Langley 1999).

1.3.1 Controlling for the effects of past demographic events

Past demographic events can produce similar changes in the amounts and site frequency spectrum of nucleotide variation as positive selection (Przeworski 2002, Huber *et al.* 2014). As an example, the excess of low frequency variants can be due to past population expansion, but can also be due to recent directional selection that has rapidly increased a frequency of a specific haplotype. Since commonly used sequence based tests of neutrality, such as Tajima's *D* (Tajima 1989) and Fay & Wu's *H* (Fay & Wu 2000), use the properties of the site frequency spectrum to infer selection in a locus, false positive (or negative) results are likely under nonequilibrium demography. This problem can be alleviated by using a background model that describes the overall patterns of the genome due to demography and contrasting the findings in genes of interest against this

background; the effects of the demography are genomewide, while effects of selection are constrained around the selected locus.

This kind of model inference has been conducted for many species, with many different methods (e.g. Voight *et al*. 2005, François *et al.* 2008, Gronau *et al.* 2011, Duchen *et al.* 2012, Liu & Fu 2015). Approximate Bayesian computation (ABC; Beaumont *et al.* 2002, Bertorelle *et al*. 2010, Csillery *et al*. 2010) is one tool used for the inference with reference sequence data. In this approach the data are reduced to summary statistics (instead of using the full sequence data) to decrease the computational load. Coalescent simulations for specific models are performed, using parameter values from prior distributions, and models that produce summary statistics closest to the observed empirical reference data are favored. Once the model has been built, it can be used as a null model to test the deviation from neutrality in sequence variation of genes of interest.

As mentioned above, this type of background modelling in long lived trees has often revealed signs of old (pre-LGM) demographic events. This does not mean that younger events are negligible, but rather that the study design used for the model inference has not been well suited for detecting recent events (Pyhäjärvi *et al.* 2007, but see Städler *et al*. 2009 and Holliday *et al.* 2010b). It is important to recognize and control also for younger events because they can generate neutral genomewide population structure that mimics the structure caused by natural selection (Meirmans 2012, De Mita *et al.* 2013). This kind of situation can arise, for instance, when colonization and adaptation to conditions in the new habitat have occurred simultaneously and along the same environmental axis. This is especially important with association methods and analysis based on allele frequencies (such as F_{ST} outlier tests). Correcting for structure will reduce the number of false positives, but unfortunately also reduces the power to detect the true outliers and associations (Lotterhos & Whitlock 2015).

1.3.2 Challenges in finding the signal of polygenic selection

The polygenic nature of adaptive traits essentially means that most individual genes influencing the trait variation receive only weak selective pressure. The tests often used to test for signals of selection (such as Tajima's *D* and Fay and Wu's *H*) are designed to detect the recovery phase after classical selective ("hard") sweeps. These tests therefore have low power to detect weaker selection, soft sweeps or incomplete sweeps (Kelly 2006, Chevin & Hospital 2008, Stephan 2015). Further, the temporal allele frequency trajectories can be nonmonotonic, (i.e. a polymorphic

equilibrium frequency can be achieved quickly, but might change to another equilibrium state or become extinct or fixed), which can further complicate the nucleotide variation patterns (Pavlidis *et al*. 2012). In *Pinus taeda*, it was however shown that as a group, loci associated with adaptive polygenic traits show more extreme values of these statistics than nonassociated loci (Eckert *et al.* 2013).

Further, considering selection towards different optima along an environmental gradient, genes and alleles can be experiencing dissimilar selective pressures in different locations. A naïve view based on this aspect would be that at a broad spatial scale the overall selective scheme would resemble balancing selection maintaining variation. At the local scale alleles would be experiencing directional selection (Hedrick *et al.* 1976, Hedrick 2006, Eckert *et al.* 2009b, Moeller & Tiffin 2008). Therefore, increased diversity and an excess of intermediate frequency variants would be expected when samples from different environments are analyzed jointly, and effects resembling selective sweeps could be detected locally.

This scenario, however, assumes that allele frequencies between environments at the selected sites differ enough so that the effects of selection pushing alleles in different directions can be seen in the surrounding genomic region. In the case of clinal polygenic selection, the majority of loci experience only weak selection and do not necessarily have significantly different frequencies (Barton 1999, LeCorre & Kremer 2003). Also other population genetic factors, such as the amount of gene flow between populations, have an effect on the patterns of variation (Kelly 2006, Städler *et al.* 2009). The expectations are thus hard to define for these small effect loci, and new population genetic models for polygenic selection are needed (Pritchard & Di Rienzo 2010, Jain & Stephan 2015, Matuszewski *et al.* 2015, Remington 2015, Stephan 2015, Yeaman 2015).

1.3.3 Association mapping can reveal the underlying loci

The logic in association mapping (Risch & Merikangas 1996) is the same as in traditional QTL (quantitative trait locus) mapping; the phenotype of interest is measured, and genotypes at the genetic markers determined in the study population. Linkage disequilibrium (Nordborg & Tavare 2002) between the markers and the causative polymorphism enables sorting out the marker that is closest to the causative site. The biggest difference between association and QTL studies is the nature of the study population; while in QTL mapping the progeny of controlled crosses are used, association mapping uses population samples. This leads to a difference in accuracy between these two approaches. In association mapping

population multiple rounds of recombination have segmented the genome into small haploblocks which enables improved resolution. In addition, all variation segregating within the population sample can be used instead of just the variation in the crossing parents of the QTL mapping family (see e.g. Cardon & Bell 2001, Balding 2006).

Association studies have been widely used in human disease genetics (Robinson *et al*. 2014), but also in studies of adaptation in many species (Atwell *et al.* 2010, Ingvarsson & Street 2011). Adaptive variation along environmental gradients in forest trees has been studied with association mapping in e.g. González-Martínez *et al*. (2007, 2008), Ingvarsson *et al*. (2008b), Eckert *et al*. (2009a, 2012), Holliday *et al*. (2010a), Ma *et al*. (2010), Cumbie *et al*. (2011), Olson *et al*. (2013), Prunier *et al.* (2013) and Evans *et al*. (2014). In addition, environmental associations (instead of phenotypic associations) have been examined by Eckert *et al*. (2010a,b, 2015), Keller *et al*. (2012) and Jaramillo-Correa *et al*. (2015).

In humans and model species with sufficient genome data, association studies have been carried out as genome wide association studies (GWAS), in which the genetic markers – most often single nucleotide polymorphisms (SNPs) – from the whole genome are used in the analysis. In recent years, next generation sequencing techniques have enabled GWAS studies in many non-model species, too (Ellegren 2014). In species whose genome is still mostly uncharacterized and/or very large and complex, a candidate gene approach is more suited. In this approach markers from preselected areas of the genome (e.g. certain genes or regulatory regions) potentially related to the trait of interest are used. Additional markers from other areas can be used as reference loci. This approach is also useful when the level of linkage disequilibrium is low and, consequently, high marker density is required (Neale & Savolainen 2004).

The success of association mapping is heavily influenced by careful design of the study and the choice of the analysis method. The mixed-model method (Yu *et al.* 2006, Kang *et al.* 2008) has been widely used in cases also typical to forest trees where there is a potential continuum of relatedness in the sample. While this method seems to perform better (Zhao *et al.* 2007) than the basic association methods with correction for population structure (Devlin & Roeder 1999, Pritchard *et al.* 2000, Price *et al.* 2006) it still suffers from inflation of test statistics as do the other single-locus methods, too (Yang *et al.* 2011b). Some multi-locus methods for association have been developed, such as the multi-locus mixed models by Segura

et al. (2012) and Yang *et al*. (2011a), and Bayesian models by Li *et al.* (2011) and Kärkkäinen & Sillanpää (2012) (see also Würschum & Kraft 2015).

One of the most important things to control for in association studies is the underlying population structure, whether obvious or hidden. The mixed model (Yu *et al.* 2006, Kang *et al.* 2008) allows correcting for population structure and relatedness by using a genomic relationship matrix. The correction can however lead to missing some true variants as they can be mistaken as confounding background variation. As described above, the F_{ST} estimates in Scots pine are generally low within the continuous part of the range. The neutral background population structure thus seems to be fairly weak. However, as the colonization and adaptation have occurred simultaneously, some spatial patterns may have formed also in neutral markers (Klopfstein *et al.* 2006, Excoffier *et al.* 2009, Slatkin & Excoffier 2012). Further, other traits that have adapted along the same environmental gradient as the trait in interest may show similar spatial genetic structure. Some confounding population structure may thus exist despite the low F_{ST} values (see also Salmela *et al*. 2008, Städler *et al.* 2009, Leslie *et al.* 2015).

1.4 Aims of the study

The aim of this thesis work was to study population genetic questions related to polygenic adaptation with Scots pine, a species that has adapted to various environmental conditions, is of ecologic and economic importance, shows many biological features that make it an excellent study system, and has ample allozyme diversity and growth trial data from earlier studies. A special trait of interest in this thesis was the timing of bud set, an important adaptive trait that shows fitness related geographical variation, and is of interest to the forest breeding community as well.

Resolving a genetic basis of a complex trait can ideally lead to the identification of all the loci affecting the trait variation, and to determining the effect sizes and potential interactions between these loci. Many population genetic factors, however, affect our ability to find these effects. In this thesis the aim was to study and characterize these factors, and to find part of the loci controlling the variation in timing of bud set. The following themes were addressed: What are the relationships between the three levels of genetic variation studied here; nucleotide variation, allozyme variation and adaptive trait variation? What are the consequences that the historical demographic events, time since colonization and

the dispersal capacity have had on these three levels of genetic variation? More specifically, the following research questions were asked:

1. What is the level of population structure of Scots pine within Europe inferred with nucleotide sequence data and SNP markers? (II, IV)
2. How do allozyme polymorphism data reflect the nucleotide level genetic variation in Scots pine? (I)
3. How is the genetic variation in a quantitative trait (timing of bud set) distributed across a wide latitudinal transect in Europe, and how does this compare to the distribution of molecular variation? (III, IV)
4. How have the past demographical events in the history of the species shaped the genetic variation? (II, IV)
5. What is the underlying genetic background of timing of bud set? (II, III)
6. Do we see signals of selection in sequence and SNP frequency data in loci controlling variation in timing of bud set and in allozyme coding genes? (I,II,IV)

2 Material and methods

A short description of the material and methods is given here. The detailed information can be found in the original papers I–IV.

2.1 Scots pine population samples

In studies I–IV we used partly overlapping sets of populations to describe genetic variation at different levels. In Study I we examined the within population diversity and haplotype structure of six allozyme coding genes with a sample of 35 trees from a southern Finnish natural standard forest. In Study II samples were drawn from ten natural populations of Scots pine natural range in Europe (9 populations) and northern Russia (1 population) to study sequence variation in candidate genes for timing of bud set and cold tolerance. From nine to 20 trees were representing each population. In studies I and II the sequences were obtained from haploid tissue, and one allele per tree was sequenced.

In studies III and IV also ten populations (partly overlapping with study II) from northern and central Europe were included. Four of these originated from seed orchards, which are collections of trees (clonal genotypes) from limited close-by area. The seeds of these genotypes are produced by open pollination by the nearby trees and should thus resemble a local natural stand (Muona & Harju 1989). Six populations were from natural forests. Sample sizes ranged from 18 to 30 mother trees per population. With this set we examined phenotypic variation in timing of bud set in European Scots pine, and conducted an association study to search for loci underlying the trait variation (study III). The phenotype was measured from open pollinated families with 25 seedlings in each, the genotype was determined for the mother tree only. This same set of mother trees was used in study IV to examine the allele frequency patterns of the genotyped loci in detail. The locations and details of the population samples in each study are shown in Fig. 1 and Table 1.

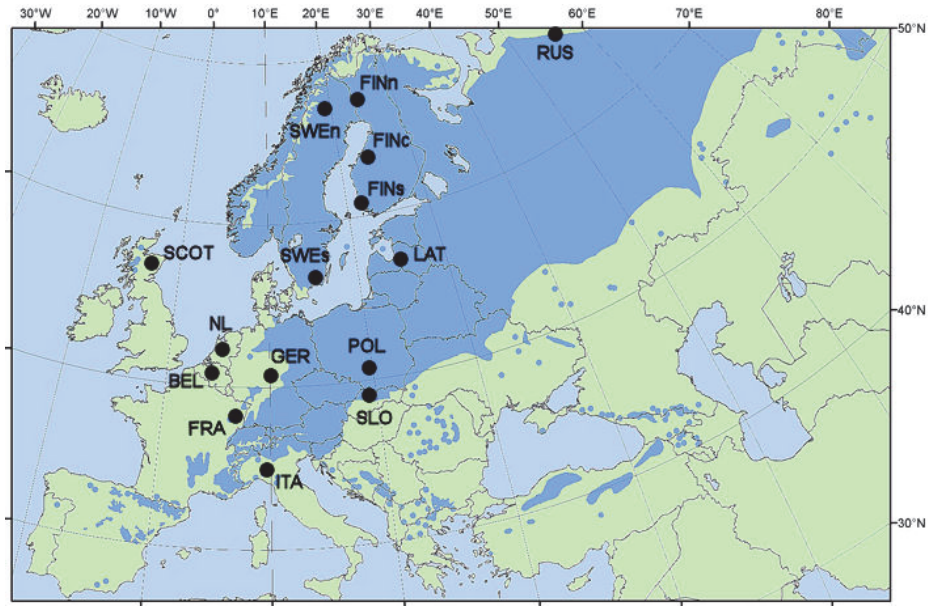

Fig. 1. Population origins in studies I–IV. The western part of Scots pine distribution in blue. Modified from: Distribution map of Scots pine (*Pinus sylvestris*) EUFORGEN 2009, www.euforgen.org.

2.2 Generating sequence, phenotype and genotype data

For sequencing in studies I and II the megagametophyte tissue of the seeds of Scots pine was used as a starting material. This tissue is haploid and therefore represents one copy of the mother tree genome. This is a great advantage for sequencing as only one haplotype is present. One contiguous fragment per individual tree was always sequenced from a single megagametophyte. PCR primers for studies I and II (when not available from previous experiments) were designed based on *Pinus* ESTs (expressed sequence tags) accessible in GenBank and EVOLTREE databases, and *P. sylvestris* sequences from previous experiments. Standard procedures for DNA isolation, PCR, and Sanger sequencing were used in both studies.

Table 1. Population samples in studies I–IV.

Population	Country	Latitude	Longitude	Number of trees in each study[1]		
				I	II	III & IV[2]
FINn	Finland	67°11’	24°03’		20	26
FINc	Finland	63°45’	24°05’			30
FINs	Finland	60°52’	21°20’	35	10	30
SWEn	Sweden	66°04’	19°06’		9	
SWEs	Sweden	56°28’	15°55’		10	30
LAT	Latvia	56°45’	25°53’		10	
RUS	Russia	66°05’	57°30’		10	
SCOT	UK	57°03’	03°16’		10	
NL[3]	Netherlands	52°06’–52°30’	05°39′–06°27’			29
BEL[3]	Belgium	49°58’–50°20’	04°55’–05°38’			19
POL	Poland	50°41’	20°05’		20	30
GER[3]	Germany	50°15’	10°30’			18
SLO[3]	Slovakia	47°44’–49°37’	16°50’–22°34’			30
FRA	France	48°51’	07°52’		10	29
ITA	Italy	44°37’	10°09’		10	
				35	119	271

[1] Number of trees refers to haploid sample size in studies I and II, and to diploid sample size in studies III and IV, [2] In study III this corresponds to the number of mother trees (half-sib families) for which both genotype and phenotype data was available, [3] Seed orchard

In study I, the identity of the allozyme coding genes was first verified in a separate set of 18 trees (Muona & Smidt 1985) with known allozyme genotypes; both enzyme electrophoresis (with standard procedures; Vallejos 1983) and DNA sequencing was conducted in this set of trees using the two halves of the same megagametophytes as the starting material for each molecular method. Following this we associated the enzyme mobility changes in starch gel with corresponding charge changing replacement mutations within each megagametophyte. This initial step was then followed by resequencing the six allozyme coding genes in a sample of 35 trees to produce the sequence data for the population genetic analysis. In study II, 11 fragments from 10 candidate genes for timing of bud set and cold tolerance were resequenced in a sample of ten populations.

The phenotypic data for study III were produced by the Finnish Natural Resources Institute (Luke) in a common garden (greenhouse) experiment in Haapastensyrjä, Southern Finland (60°37', 24°26'). The trial was run in five randomized blocks with 25 open pollinated seeds per mother tree (i.e. half-sib family) sown in the beginning of June 2003. Bud set status of the seedlings was

monitored once per week until November 2003. Best linear unbiased predictors (BLUPs) were used as family-specific means, which served also as the phenotypic response variable in the association analysis.

To obtain genotype data for studies III and IV, a SNP array of 768 single nucleotide polymorphism markers was initially designed. This represented 56 gene fragments sequenced "in house" (including e.g. the candidate gene sequences of study II and allozyme coding genes of study I) and additional 341 fragments from a conifer resequencing project CRSP (Comparative Re-Sequencing in Pinaceae, see Wegrzyn *et al.* 2008). The SNP array design was done using Illumina ADT (Assay Design Tool, Illumina 2015). The selection of SNPs was based on the ADT score, and the within fragment distance and linkage disequilibrium patterns of the single nucleotide polymorphisms.

The genotyping was done with Illumina GoldenGate assay in CNG (Centre National de Génotypage, Evry, France). On average, 19 megagametophytes were pooled for each sample to obtain the diploid genotypes for the mother trees. The genotypes were filtered to include only those that had MAF < 0.05 in the whole sample, and that had deviations from Hardy-Weinberg equilibrium expectations in at most one population. 351 SNPs of the initial 768 survived the filtering and were included in the association analysis in study III and allele frequency pattern analysis in study IV.

2.3 Sequence material for background demography modelling

For estimating demography with approximate Bayesian computation, 32 previously published gene fragments was used (Pyhäjärvi *et al.* 2007, Palmé *et al*. 2008, Wachowiak *et al*. 2009). The sequences (maximum of five haploid sequences per population) originated from six European (three northern and three central) populations. These data were pooled together to derive a common background model for European Scots pine.

2.4 Data analysis

2.4.1 Population structure

Population structure was studied by estimating F_{ST} between pairs of populations. In study II, the sequences from ten candidate genes were used. The analysis was

done on the genewise level with Arlequin 3.11 (Excoffier *et al.* 2005). In study IV, the SNP genotype data initially produced for study III were used. As the genotype data for association analysis were chosen to include only SNPs with MAF<0.05, we recovered some SNPs with lower frequencies (but conferring to Hardy-Weinberg equilibrium expectations) for this analysis. The final number of SNPs was 286. Expected heterozygosity from these same SNP markers was also calculated for each population. Estimates from the SNP data were calculated with R (R Core Team 2015) with package 'Hierfstat' (Goudet 2014).

2.4.2 Background demography modelling with ABC

We used approximate Bayesian computation (Beaumont *et al.* 2002, see also Bertorelle *et al.* 2010 and Csilléry *et al.* 2010) to infer a demographic background model for European Scots pine. We compared the fit of three simple models; standard neutral model with recombination (SNMR), population expansion model with recombination (PEMR) and an extended bottleneck model with recombination (EBMR). The parameters used were θ (population nucleotide diversity) and ρ (population recombination rate) for all three models, α (exponential growth rate) in PEMR, and T (timing of bottleneck) and B (relative bottleneck size) in EBMR. Uniform priors were used for all these parameters. 1000000 coalescent simulations were run for each model. For each simulation, parameter values were drawn from the prior distributions, and summary statistics θ_W (Watterson estimate of θ) π (nucleotide diversity) and K (number of haplotypes) were calculated based on the simulation outcome. A subset of these simulations (0.5%) that produced summary statistic outcomes closest to the observed data in 32 reference loci was kept. Posterior distributions for the model parameters were defined from this subset. Local linear regression with log transformation was used (Beaumont *et al.* 2002). The relative fit of the three models was then estimated in two ways, i.e. determining the model probability from an acceptance rate under each model, and by posterior simulations. Demography inference was done with a Python module and application package for population genetics Seqlib 1.4 (De Mita *et al.* 2007).

2.4.3 Literature review for comparison of H_e vs. θ

In study I the relationship between expected heterozygosity (H_e) in allozymes and diversity (θ) at the nucleotide sequence level between species was studied by reviewing published estimates from 13 tree species and 14 other plant species for

which both estimates were available. Wild species were preferred over domesticated species. Studies with similar sampling design for H_e and θ for each species were included. The relationship between these two estimates was quantified with linear regression analysis.

2.4.4 Sequence analyses

In study I we characterized the within population variation in six allozyme coding loci (*6pgdB*, *aco*, *gdh*, *gotC*, *mdhA* and *mdhB*) with basic population genetics summary statistics such as S (number of segregating sites), N_h (number of haplotypes) and θ_W (Watterson estimate of θ). Synonymous diversity in these genes was compared to a set of non allozyme genes from Pyhäjärvi *et al*. (2007) with a Bayesian multilocus estimation method (Pyhäjärvi *et al.* 2007). Linkage disequilibrium was characterized with r^2 and recombination (ρ) with a composite likelihood estimator of Hudson (2001) implemented in LDhat (McVean *et al.* 2002). Silent site divergence K_{sil} was estimated with one *Pinus pinaster* individual as an outgroup. Deviations from neutrality expectations were studied with K_A/K_S ratio, *HKA* test (Hudson *et al.* 1987, calculated with a multilocus *HKA* program by Hey 2010), McDonald-Kreitman test (McDonald & Kreitman 1991), Tajima's D (Tajima 1989), Fay and Wu's H (Fay & Wu 2000), the Evens-Watterson test (*EW*; Watterson 1978), a test combining the latter three (*DHEW*; Zeng *et al.* 2007) and finally with the Hudson haplotype test (*HHT*; Hudson *et al.* 1994). The significance of *HHT* tests over all allozyme coding genes as a group was estimated with a Z transformation test (Whitlock 2005). The nonequilibrium genetic background due to historical demographic factors was taken into account in *HHT* by using the bottleneck model from Pyhäjärvi *et al.* (2007) as a null model. *DHEW* test should be relatively robust to demography. Tajima's D, Fay & Wu's H, *EW* and *DHEW* tests were done using DH.jar provided by Kai Zeng. The rest of the descriptive statistics, K_A/K_S ratios and McDonald-Kreitman tests were calculated with DNAsp 4.20.2 (Rozas *et al.* 2003).

In study II the standard diversity calculations and neutrality tests – Tajimas D, H_{norm} (a standardized version of Fay & Wu's H; Zeng *et al.* 2006) and *DHEW* – were calculated for a pooled sample of seven main range populations, and for three subgroups of those (Northern Fennoscandia: FINn, SWEn; Southern Fennoscandia and Baltic region: FINs, SWEs, LAT; and Central Europe: POL, FRA, Fig. 1). A multilocus estimate of θ was calculated for each of the ten populations separately using the same method as in study I. K_A/K_S ratios, *HKA* test and McDonald-

Kreitman tests were done for the main range sample only. One *P. pinaster* individual was used as an outgroup. The significance of Tajima's D and H_{norm} was estimated by using the bottleneck model inferred with ABC (described above) as a null model. We searched for F_{ST} outlier sites with BayeScan (Foll & Gaggiotti 2008) and for allele frequency clines by regressing allele frequencies with latitude (adjusted for multiple testing by false discovery rate with QVALUE; Storey & Tibshirani 2003). Recombination (ρ) was estimated with a composite likelihood estimator as in study I, and linkage disequilibrium characterized with r^2 and $|D'|$. Recombination and LD analyses were done only within the two biggest population samples (Northern Finland and Poland, 20 haploid samples in each). Diversity statistics were calculated with a Python module and application package for population genetics "Seqlib" (De Mita *et al.* 2007), except for silent sites, for which DNAsp 5.10.00 (Librado & Rozas 2009) was used. Also divergence and K_A/K_S ratios were calculated with DNAsp.

2.4.5 Association analysis on timing of bud set

Association analysis to find loci underlying the variation in timing of bud set was done using a novel Bayesian multilocus method introduced in study III. This method is searching for associations both from the within population variation (but combining data over multiple populations) and from the between population component of variation. The method is especially effective in studies where simultaneous colonization and adaptation processes have likely caused same kind of allele frequency patterns. The within population analysis is controlling for spurious associations due to population structure with a requirement of the association being present also within populations. The between population analysis is based on permuting the phenotypic data within populations and associating the allele frequency change with the change in the population mean. This complements the first part, as some associations might be missed if the alleles have very extreme frequencies within the populations. This approach might, however, be more prone to spurious associations. The analysis was done by implementing a Gibbs sampler based on the algorithm presented in Knürr *et al.* (2013). Associations were searched for within the whole ten population data set, and within the four northern populations (northern data set) and the six central populations (central data set).

2.4.6 Allele frequency clines and allelic covariation

In study IV we examined the prevalence of allele frequency clines and allele frequency covariation in four different SNP classes; 1) SNPs found to associate with timing of bud set in study III, 2) SNPs linked to the associated SNPs (i.e. from the same fragments), 3) SNPs from other candidate genes for timing of bud set or stress (especially cold) tolerance, and 4) reference SNPs (presumably unrelated to timing of bud set or stress). Analyses were done with the whole ten population data set, but also with only the northern data, since indication of genetic heterogeneity between northern and southern populations was found in study III. With the ten population data set, all 22 SNPs showing associations were included in the SNP class 1; in the northern subdata only the 12 SNPs associating in the northern part were included. Class 2 was adjusted for the northern data accordingly.

The slope steepness of allele frequency clines in each SNP was examined by regressing the population allele frequency on latitude of origin. The mean of absolute values of regression coefficients within each SNP class was used as a summary statistic, and the distributions of this statistic in each SNP class were compared with Kolmogorov-Smirnov test. We also checked whether the large effect alleles had steeper allele frequency clines than the small effect alleles.

Between population allelic covariation was studied with the assumption that if the allele frequency covariation is negligible, the potential for allelic covariation does not exist (Kremer & Le Corre 2011). Allele frequency covariation was calculated as sample covariance to describe the between population component of linkage disequilibrium (Ohta 1982). The Pearson correlation was also calculated to obtain a standardized measure. In cases where multiple SNPs existed per gene fragment, only one SNP per fragment was used.

3 Results and discussion

In the following I will sum up the results over the original papers I–IV. I will start by describing the results on basic population genetic issues such as the population structure, and the amounts and distribution of variation. I will then continue by discussing the results on more specific questions about the effects of demography and the adaptive sequence variation.

3.1 Population structure of European Scots pine

The level of population structure and the amount of present day gene flow influence many aspects in study design and interpretation of results in genetic studies. Scots pine has a continuous distribution covering most of northern and central Europe and extending eastwards all the way to the eastern Siberia. At range edges Scots pine is found in more isolated patches (Mirov 1967). Low F_{ST} values (generally well below 0.05) are characteristic for this species (e.g. Goncharenko *et al.* 1994, Prus-Glowacki & Stephan 1994, Karhu *et al*. 1996, Pyhäjärvi *et al*. 2007). We studied the population structure with two partly overlapping sets of populations by examining population pairwise F_{ST} estimates. In Study IV we found that the Finnish populations (northern, central and southern) and the southern Swedish population were least diverged from each other (0.0018–0.0051). Central European populations seemed more diverged from the northern populations (0.0043–0.0260). F_{ST} estimates among the central populations were also higher (0.0065–0.0246).

In study II we found, again, that Swedish and Finnish populations seem least diverged from each other. Latvian, Polish, French and Russian populations also showed some very low population pairwise F_{ST} estimates with these northern populations. Because of smaller sample sizes and fewer loci, these data were less powerful than the data in study IV and therefore had less resolution. Still, we found that the populations from Scotland, Italy and northern Russia were on average more differentiated from the other populations, especially the Italian population that consistently showed population pairwise F_{ST} values around or above 0.10 (see also Wachowiak *et al.* 2014, Scalfi *et al.* 2009). The data agree with the suggestion of Gullberg *et al.* (1985) and Dvornyk (2001) that the central populations have had more time to diverge from each other than the northern populations which have arrived to their current locations later after the last glacial maximum (LGM) and are thus further from the migration-drift equilibrium (Slatkin 1993). For large

populations it may take very long before migration-drift equilibrium is reached (see Whitlock & McCauley 1999).

The relative contributions of a common origin of recolonization and large dispersal capacity through pollen in generating the generally low F_{ST} values in Scots pine should be studied further with better modelling tools and a sampling design better suited to study this issue. According to Städler *et al.* (2009), the null hypothesis of panmixia can be tested also with site frequency spectrum based tests; the SFS is similar in local population samples, pooled population samples and scattered samples (one sample per population) only under panmixia. In this respect it is interesting that in the candidate gene sequence data of study II we observed this kind of a sampling effect, as Tajima's D values estimated from individual populations were less negative than when estimated from the pooled data. In forthcoming sequence studies of Scots pine this effect and the potential fine structure in the continuous part of Scots pine range should be examined in more detail and preferably in neutral sequence data. Low F_{ST} estimates do not guarantee a lack of fine substructure capable of biasing results on sequence variation or association studies (see Salmela *et al*. 2008, Pickrell and Pritchard 2012, Bradburd *et al*. 2015, O'Connor *et al.* 2015).

3.1.1 Population structure and diversity in comparison to allozyme level estimates

Although nucleotide level data are getting easier to obtain even for conifers with large and complex genomes, the older data sets obtained with allozyme markers are still a valuable resource. There is a wealth of diversity data measured with allozyme markers for *P. sylvestris* populations across the wide species range (Gullberg *et al.* 1985, Muona & Smidt 1985, Muona & Harju 1989, Wang *et al.* 1991, Goncharenko *et al.* 1994, Prus-Glowacki & Stephan 1994, Shigapov *et al.* 1995, Puglisi & Attolico 2000, Dvornyk 2001). From these data it has been learnt that most of the genetic variation resides within populations, and very low level of differentiation (generally $F_{ST} < 0.02$) is found between populations even over great geographical distances, as in many other tree species, too (Hamrick 1992). It is however important to understand how well allozyme level data represents the nucleotide level sequence variation.

In comparison to the allozyme markers, nucleotide level variation is a more accurate tool for measuring diversity differences and divergence. Only a fraction of all nucleotide level variation is translated into such nonsynonymous variation

that can be detected in enzyme electrophoresis (see Ramshaw *et al.* 1979). In study I we concentrated on the relationship of genetic variation at allozyme (expected heterozygosity, H_e) versus nucleotide level (nucleotide diversity, θ). At the level of diversity differences among species, a correlation between nucleotide diversity and allozyme heterozygosity was strong. Among the six allozyme coding loci that were examined in detail, a lack of correlation was observed in within population data of *P. sylvestris*. The loci we studied were mostly biallelic. Our results suggest that within population, at individual loci, allozyme heterozygosity does not predict the underlying nucleotide variation well. Rather, H_e depends on the allele frequencies of the particular site that determine the mobility of the allele, at least in biallelic loci, and does not describe the overall amount of mutation in the whole gene as θ does. In other words, H_e is affected by mutation in one or a few nucleotide sites at most, whereas θ can combine information over multiple sites. However, when combining data over larger number of allozyme loci (also multiallelic and monomorphic) and over larger taxonomical groups such as species, allozymes reflect the relative amounts of diversity fairly well and effective population size seems to be the main determinant of genetic diversity in multilocus data.

The ability of allozymes to reflect population structure (F_{ST} between populations within species) is most likely somewhere in between that of the two levels mentioned above. This is partly related to the resolution, but also to the fact that highly variable allozyme markers may have been favored. As F_{ST} is a ratio of the difference between the total and within population variation to total variation, F_{ST} values become high when within population diversity is low (Charlesworth 1998). The opposite is true for markers with high within population heterozygosity (Hedrick 1999, Muller *et al.* 2008). With nucleotide level variation in study II, somewhat higher F_{ST} values were observed in comparisons involving the three marginal populations, especially with the Italian population that had lower diversity (Fig. 2b).

Allozyme markers have often not detected a clear reduction of expected heterozygosity in marginal populations compared to populations in the continuous part of the distribution (Dvornyk 2001). With nucleotide data, in study II, the isolated Italian (Apennine) population was seen to have clearly less variation than the populations from the main range. In an earlier study the Apennine populations had lower heterozygosity in microsatellite markers (Scalfi *et al.* 2009). With allozyme markers the heterozygosity was slightly lower (Puglisi & Attolico 2000) in an Apennine population than in Alpine populations (part of the main range). Heterozygosity should be less affected by the stochastic processes at range limits

than the number of alleles, another measure of diversity (Nei 1975), some indication of which was found in a meta-analysis of Eckert *et al.* (2008). The same can hold for comparisons of the relative abilities of allozyme markers and nucleotide variation to uncover true differences in diversity.

3.1.2 Adaptive variation in timing of bud set in Scots pine

In study III we measured timing of bud set in European population samples originating from 10 different latitudes. The phenotypic cline (Fig. 2a) was clear; northern populations set bud earlier than the central ones (overall correlation R^2=0.98 between population mean of bud set timing and latitude). The cline was slightly steeper and the variation more closely associated with latitude in the north. The distributions of family means of days to bud set showed differences in the amount of phenotypic variation among populations; the central European populations had clearly larger phenotypic variance of family means in this trait. The smallest variances were found in the three Finnish populations. The population specific heritabilities had similar ranges in central and northern populations (0.35–0.63 in the northern and 0.39–0.75 in central populations). Thus the larger phenotypic variation in central populations is not caused by only larger environmental variance (V_E), but also larger genetic variance (V_A).

At the nucleotide level variation, we did not see a reduction of diversity in the north in the sequence data in study II, which agrees with earlier results on the distribution of nucleotide diversity in European Scots pine (Pyhäjärvi *et al.* 2007). The reference SNP markers of study IV also showed similar heterozygosities in all populations (Fig. 2c, low frequency <0.05 SNPs included, but note the potential ascertainment bias in the SNP data). This indicates that the reduced trait variance in the north is not a result of colonization bottlenecks, which would lead to decreased amounts of variation in random genetic markers as well. Although we cannot rule out the possibility that the greater variation in central populations is an experimental artefact due to the fact that they were grown in nonnative light conditions (see Olson *et al.* 2013), these results suggest that timing of bud set is under stronger stabilizing selection in the northern areas and perhaps under directional selection close to the northern range limit (if slightly maladapted, Garcia-Ramos & Kirkpatrick 1997, Kirkpatrick & Barton 1997, Moeller *et al.* 2011, Polechova & Barton 2015). This is in line with the results of a reanalysis based on transfer trial series data of Eiche (1966) by Savolainen *et al.* (2007) who showed that the southern genotypes transferred northward have much more decreased

survival and overall fitness compared to northern genotypes transferred southward for a similar latitudinal distance. Another explanation suggested for higher genetic variation is larger environmental heterogeneity (spatial or temporal) at the local scale (Yeaman & Jarvis 2006, Salmela 2014).

This finding of a strong cline in timing of bud set also corroborates earlier studies that find high differentiation in adaptive quantitative traits (Q_{ST}) between populations while genome differentiation (F_{ST}) is low (Latta 1998, Merilä & Crnokrak 2001, McKay & Latta 2002, Leinonen *et al*. 2013). The indication thus is that natural selection on timing of bud set is strong enough to overcome the maladaptive gene flow from surrounding areas where different phenotypic optima are found.

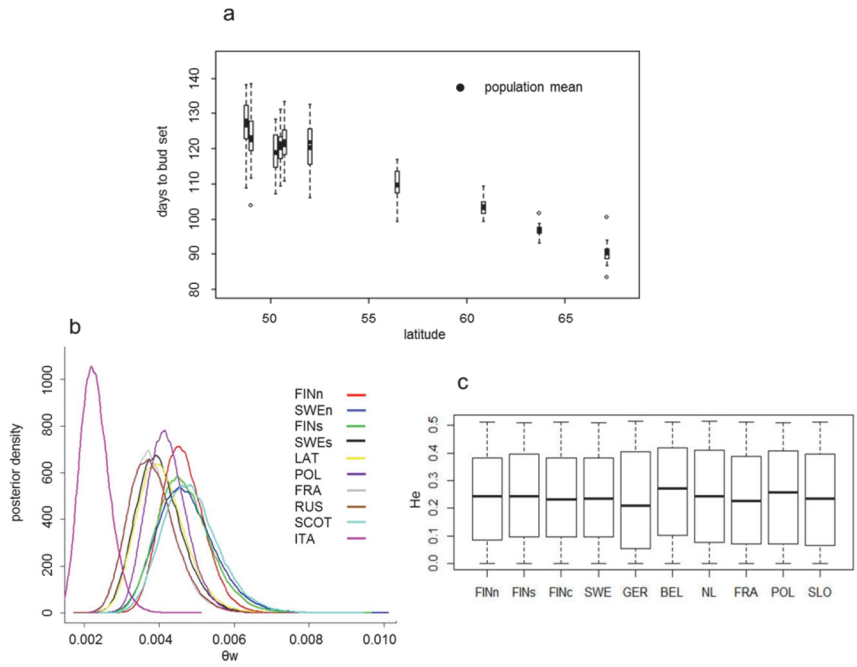

Fig. 2. The difference in variation at phenotypic and nucleotide level. a) The box and whiskers plot of population specific family means in timing of bud set (study III), b) multilocus theta estimate in sequence data (study II), c) expected heterozygosity in reference SNPs (study IV).

3.2 Impacts of demography on the genetic variation of European Scots pine populations

Demographic events such as population size fluctuations, range contractions and recolonizations are expected to affect the overall amount and pattern of genetic variation. The effects are seen throughout the genome. Controlling for these effects is important when inferring signals of selection in specific loci, since similar changes in diversity and SFS can be caused both by demographic events and selection (Przeworski 2002, Huber *et al.* 2014). The colonization history, the amount of gene flow between populations, and the resulting present day population structure also has to be taken into account.

3.2.1 Impacts of historical population size fluctuations

In study II we built a simplified background demography model for European Scots pine. The model was not intended to be an exact description of historical events, but a background model to compare the significance of neutrality tests against. Out of three simple models (constant population size, population expansion, and population bottleneck) we found most evidence in favour of the bottleneck model. This model was characterized by a size reduction to about 0.01 relative to current population size roughly about 0.10 x $4N_0$ generations ago (with duration of the bottleneck fixed to 0.006 x $4N_0$ generations), and lead to a slightly lowered level of haplotype diversity and a higher proportion of low frequency variants seen as low Tajima's *D* values when compared to the standard neutral model. The model obtained was not a perfect fit to the observed patterns, especially with respect to H_{norm} values, which had to be taken into account when interpreting the results of the tests of neutrality. The bottleneck model is, however, conservative with respect to e.g. Tajima's *D*; as the model allows for wide distribution of values, the observed test statistic must be fairly extreme to reach significance over the background.

The model obtained was in good agreement with results in previous study on the same issue by Pyhäjärvi *et al.* (2007), who compared the fit of a bottleneck model against a constant size model with a coarser grid of parameter values, and did not include an expansion model. Similar demographic scenarios were found in the northern and central European areas despite their different post-LGM colonization times. A bottleneck model with rather ancient timing for the bottleneck event has also been inferred for e.g. Norway spruce (Heuertz *et al.* 2006), European aspen (Ingvarsson *et al*. 2008a) and Maritime pine (Lepoittevin 2009). The actual

course of demographic events has most likely been more complex. As an example, the effects of recurrent range contractions (with population fragmentation) and recolonizations (a panmictic stage) were modelled by Jesus *et al.* (2006). They found that the subdivision during long interglacial periods can lead to very long gene genealogies.

The last glacial maximum could be too recent (in the viewpoint of a species with long generation times) to be seen in the SFS of nuclear genome at this point, i.e. not enough mutations have yet accumulated for the detection of the population growth after the glacial period (Wakeley & Aliacar 2001, but see Liu & Fu 2015). Some other aspects of variation might be more informative about recent events, such as distribution of F_{ST} (Lotterhos & Whitlock 2014), or age of rare variants (Mathieson & McVean 2015). Holliday *et al*. (2010b), nevertheless, found different timing of bottlenecks for different Sitka spruce populations along a postglacial colonization axis with diversity and SFS as summary statistics. The age of the bottleneck correlated with the time of colonization. However, as they noted themselves, demography inference can be strongly affected by the sampling strategy. According to Städler *et al*. (2009), only under extremely high gene flow (> 25 4*Nm*) is the species wide demography correctly inferred from local samples or from pooled population samples; compared to the scattered sampling (one sample per population) less negative Tajima's *D* values are observed. In study II we used pooled data for demography inference, and thus the model estimates here can also be affected by a sampling effect. However, with a purpose of having a background model for testing neutrality in particular genes of interest, it is best to build the background model with the sampling that resembles the structure of the sequence data to be tested for neutrality.

3.2.2 Impacts of present day gene flow and recolonization history

Under the model of stabilizing selection toward different local optima along an environmental gradient, the ability of populations to follow the optimum and locally adapt is defined by "characteristic length" $= \sigma/\sqrt{(sV_A)}$ (Slatkin 1978). This measure depends on the ratio of dispersal distance, and strength of selection and the amount of additive genetic variation. Long average dispersal distances thus make the characteristic length larger, unless the selection is very strong. On the other hand, gene flow can be viewed as an important source of genetic variation enabling adaptation (Slatkin 1987, Lenormand 2002). Previous pollen dispersal estimates (Koski 1970, Robledo-Arnuncio & Gil 2005) suggest that the average

dispersal distances in Scots pine are not extremely high despite the wind pollination. Some effective pollination has been observed over distance of 100 km and the dispersal kernel is very leptokurtic (Robledo-Arnunzio 2011, see also Kremer *et al.* 2012), but it is not known whether these long-distance pollination events lead to germination and establishment. Assuming average dispersal distance between 100 and 1000 meters, taking the estimates of V_A for timing of bud set in study III (mean = 58 days^2 for north, mean = 227 days^2 for central), and assuming fairly strong selection on the phenotype ($0.1<s<1.0$, stronger in the north than in the central areas), the characteristic length for timing of bud set becomes very small, i.e. less than a kilometre. Larger average dispersal distances or weaker selection lead to more plausible estimates. Nevertheless, the balance between gene flow and selection, i.e. the characteristic length in timing of bud set is small enough to enable the populations to adapt fairly accurately to the climatic gradient, as seen from the phenotypic data (Fig. 2a). Adaptation at small spatial scale despite low between population differentiation has been suggested e.g. in Csillery *et al.* (2014), Eckert *et al*. (2015) and Fitzpatrick *et al.* (2015) (see Richardson *et al.* 2014 and Scotti *et al*. 2015 for discussion).

In study III we showed that despite the low F_{ST} values across the main range of Scots pine there still is a need to control for population structure (Fig. 4 in study III). Apparently the colonization process has generated a shallow allele frequency cline in the SNP *CL1966Contig1_05-341* that coincides with the phenotypic cline in timing of bud set, but the allelic variation is not associated with the trait within populations in the north. Such clines is neutral alleles can form with the process of range expansion (Klopfstein *et al.* 2006, Excoffier *et al.* 2009, Slatkin & Excoffier 2012). Furthermore, in study IV, we saw that in the north the allele frequency clines were on average slightly steeper than in the central Europe, which points to demographic processes generating shallow neutral clines. In forthcoming studies, a spatially explicit model for past demography (e.g. Currat *et al.* 2004, Ray *et al.* 2010, Antoniazza *et al.* 2014) should be inferred that accounts for both the pre- and post-LGM events.

3.3 Genetic background of bud set timing in European Scots pine

Acknowledging and controlling for the demographic and other relevant population genetic factors discussed above makes the task of finding the genetic background of adaptive trait variation easier, though not easy. In Study III we examined the genetic basis of timing of bud set in an association study that combined ten

populations of Scots pine from northern and central Europe. A candidate gene based approach was used, i.e. SNPs from genes potentially involved in creating and maintaining variation in bud set timing were preferentially selected for genotyping. However, we also had a set of reference SNPs that originated from genes with functions presumably not related to clinal adaptation along a latitudinal gradient.

Associations were found among the candidate genes such as *prr1* (pseudo response regulator 1, Turner *et al*. 2005, Källman *et al.* 2014), *ftl2* (ft/tfl1 -like gene 2, Gyllenstrand *et al.* 2007, Karlgren *et al*. 2011, Avia *et al.* 2014, Chen *et al.* 2014), and *phyn* (phytochrome N, García-Gil *et al.* 2003, Franklin & Quail 2010, Pankin *et al.* 2014), but also among genes initially categorized as reference loci. A closer inspection of the putative functions of these gene fragments however revealed some potential connections to e.g. dormancy, light signalling, circadian clock and growth. Associations were found also in two allozyme coding genes of study I, *6pgdB* and *aco*. Activity of 6pgd protein seems to be associated in drought stress responses in spring wheat (Chen *et al.* 2004). Allozyme variation in *6pgd* is associated with dark respiration efficacy in perennial ryegrass, although the causal relationship was not established (Rainey *et al.* 1987). Aconitase is a target of nitric oxide (NO) and indicated in stress response reactions through inhibition by NO in *A. thaliana* (Gupta *et al.* 2012). Interestingly, SNPs in putative genes for nitric oxide synthase and calmodulin (which is involved in NO production) were also found to associate with timing of bud set in study III. In study I some closely related subgroups of haplotypes were found in allozyme coding genes unexpectedly often when compared to non-allozyme coding genes.

Some associations might arise for instance through pleiotropic effects and thus not be directly involved in this trait variation; an allele beneficial for a given trait can be selected for even when it has some harmful pleiotropic effects if complementary mutations evolve to correct for these effects (the selection, pleiotropy and compensation (SPC) model; Pavlicev & Wagner 2012). The associations that point to the compensating mutations would, nevertheless, be relevant for both understanding the evolution of the trait and for breeding (see also Rockman 2012 and Marjoram 2014 for discussion on association genetics).

3.3.1 Genetic heterogeneity underlying the timing of bud set

Interestingly, we found that different loci were associated with the trait variation in northern and central European populations. Allele frequency differences in the associating SNPs between these areas were low, and sample sizes comparable.

Since for additive genetic variation $V=2pqa^2$, the different outcomes thus arise from the SNPs having different effect sizes in different parts of the Scots pine range. One potential explanation for the genetic heterogeneity can be the different photoperiodic conditions – i.e. different information content of the light/dark cycle in northern vs. central regions of Scots pine distribution. Different timekeeping mechanisms – light-dominant and dark-dominant – have been suggested for Scots pine, the light-dominant being favoured towards north. Also different light spectral conditions in northern and central areas could lead to differences in timekeeping mechanisms (e.g. Oleksyn *et al.* 1992, Clapham *et al*. 1998, Clapham *et al.* 2002, Dueck *et al*. 2015, Strømme *et al.* 2015).

This result is also very exiting considering the continuous distribution of Scots pine within Europe and the seemingly uniform phenotypic cline in timing of bud set. This result resembles conditional neutrality (Schnee & Thompson 1984, Hall *et al.* 2010, Anderson *et al.* 2013), a situation where an allele has fitness effects in one environment but not in another environment. Maintenance of long term genetic variation based on conditional neutrality is challenged by gene flow between the different environments (Slatkin 1987). The alleles beneficial in one environment tend to fix across the range as the gene flow pushes the frequency higher also in the environment where the alleles are neutral. This fixation across the range might, however, take a rather long time in a species with long generation time. Antagonistic pleiotropy, on the other hand, can maintain polymorphism more easily in the face of gene flow (Tiffin & Ross-Ibarra 2014, Weinig *et al*. 2014, Martin & Lenormand 2015, see also Oakley *et al*. 2014). Distinguishing between these alternatives remains a challenge for the forthcoming studies. Combined studies of multiple adaptive traits (Stock *et al.* 2014, MacPherson *et al.* 2015, Oubida *et al.* 2015) and addition of more detailed landscape data (Sork *et al*. 2013, De Kort *et al*. 2014, Eckert *et al*. 2010a,b, Stucki *et al.* 2014) can shed some more light on these questions.

3.3.2 Sequence variation in timing of bud set associated genes

The association results discussed above enable us to make some comparisons on the patterns of allele frequencies and sequence variation among loci that were associated to timing of bud set versus loci that were not. Four of the ten candidate genes examined in study II were found to be associated with variation in timing of bud set in study IV. All of these associations arose in northern latitudes. All ten resequenced genes were strong candidates for timing of bud set (or cold tolerance).

The lack of association in some of these can indicate insufficient power, insufficient marker coverage or simply that the gene function is too central and/or conserved to tolerate adaptive variation (see de Montaigu *et al.* 2015). Insufficient marker coverage could well be the case in e.g. *cry1* where only one SNP was successfully genotyped in a region of 4768 base pairs. *col1* on the other hand seems very conservative at the protein level, as there were no nonsynonymous variation at all. Lack of power can affect any locus that has only small effect size, as our method of association was specifically designed to detect only the largest effect loci among the studied markers.

Among the four associated genes, only *phyn* had significant Tajima's *D* values (-2.539 in the pooled sample of seven main range populations) when tested against the background bottleneck model (discussed in section *3.2.1*). However, also *prr1* and *ftl2* had Tajima's *D* values (-2.244 and -2.039) that were amongst the lowest in study II. *lp2* had an intermediate value of *D* among the ten resequenced loci. With H_{norm} (standardized Fay & Wu's *H*) we see quite the opposite; *lp2* has a second most negative value of H_{norm} (though not significant against the bottleneck model), and *prr1*, *phyn* and *ftl2* are amongst the genes with the most positive values. *DHEW* test does not detect any of the associated genes. Haplotype diversities of these four genes were in the higher part of distribution. Among the nonassociated genes, *cry1* had the most significant values in the three neutrality tests. *dhn1* was characterized by high diversity and relatively high (less negative) Tajima's *D*. Overall, while acknowledging the small number of genes in this comparison, timing of bud set associated genes in these data were characterized by lower values of Tajima's *D*, higher values of H_{norm} and *DHEW* estimates and higher haplotype diversity in comparison to nonassociated genes. Theta estimates (per base pair) did not differ between associated and nonassociated genes (Table 2). These patterns were approximately the same also for the smaller subsets of data (Northern Fennoscandia, Southern Fennoscandia and Baltic region, Central Europe).

A closer look at the distribution of variation along the fragments (Fig. 3) indicates minor elevation of diversity ($\theta > 0.005$) around the associating SNPs in *prr1* and *lp2*. Associating SNPs in *ftl2* and *phyn* are are in regions of lower diversity ($\theta < 0.005$). *ftl2* and *phyn* were detected by the between population component of association analysis; *prr1* and *lp2* by the within population analysis. This is congruent with the notion that the between population analysis seems to capture associated SNPs with extreme population frequencies that could be missed in the within population analysis. SNPs that have clinal allelic frequency variation tend to reside near Tajima's *D* peaks in all four genes. The clinal sites were not the ones

showing the signal of association, however. The linkage disequilibrium patterns between the clinal and associated sites could be studied in more detail.

Table 2. Neutrality test statistics and diversity estimates in candidate genes (study II). The estimates are calculated from the pooled sample of seven main range populations. Ranks of associated genes in parenthesis.

gene	D[1]	Hnorm[2]	DHEW[3]	Hd[4]	θ_W[5]	length[6]	SNPs[7]
cry1	−2.626**	−1.422*	0.0035**	0.885	0.0043	4762	1
phyn	**−2.539** (2)**	**0.376 (8)**	**0.1703 (8)**	**0.969 (2)**	**0.0026 (10)**	**6683**	**4**
prr1	**−2.244 (3)**	**0.184 (7)**	**0.1024 (6)**	**0.943 (4)**	**0.0035 (8)**	**4183**	**8**
ftl2	**−2.039 (4)**	**1.042 (11)**	**0.2479 (10)**	**0.956 (3)**	**0.0045 (4)**	**2545**	**5**
col1	−1.949	−0.577	0.0237*	0.804	0.0022	3847	3
gi_f2	−1.829	−0.795	0.0175*	0.639	0.0042	1370	5
ztl	−1.383	−0.352	0.0348*	0.788	0.0075	1182	3
lp2	**−1.103 (8)**	**−1.404 (2)**	**0.0961 (5)**	**0.920 (6)**	**0.0052 (3)**	**1185**	**4**
myb	−0.578	0.848	0.2180	0.941	0.0040	3113	3
dhn1	−0.420	0.389	0.1070	0.975	0.0172	1357	2
gi_f1	−0.189	−0.446	0.3684	0.706	0.0029	402	3

[1] Tajima's D, [2] normalized value of Fay & Wu's H, [3] p-value of DHEW test, [4] Haplotype diversity, [5] Watterson estimate of theta, all sites, [6] alignment length in study II, [7] number of SNPs successfully genotyped in study III

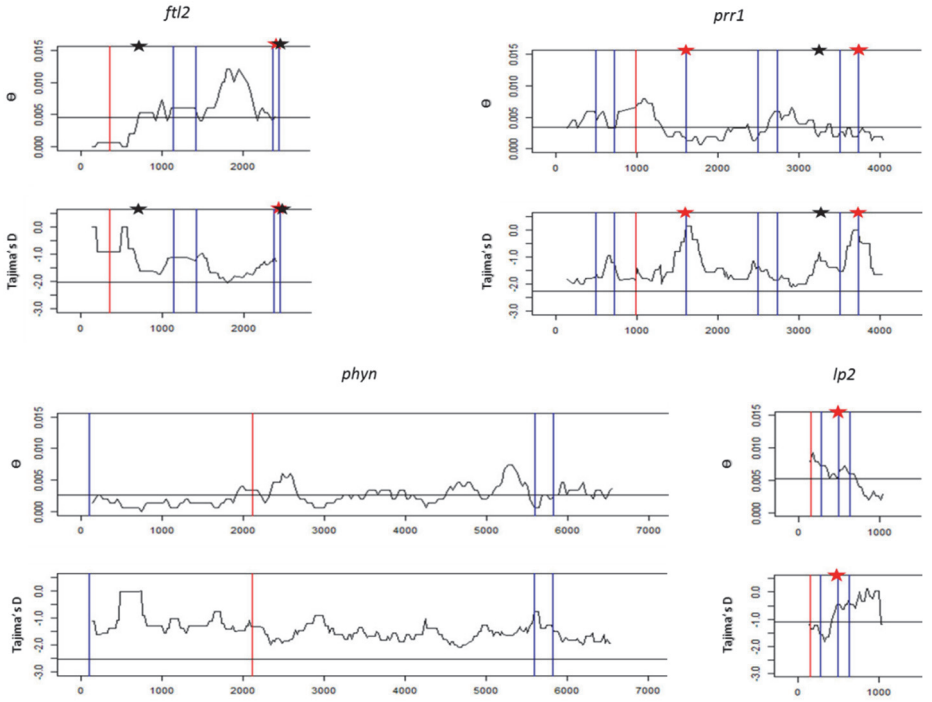

Fig. 3. Observed sequence variation in four associating genes. Sliding window plots were generated from sequence data in study II. Red vertical lines represent the associating SNPs, blue vertical lines the other SNPs in study III. Black vertical line indicates the overall value of the statistic in the fragment. Red stars indicate allele frequency clines in study III SNP data, black stars mark the locations allele frequency clines in study II sequence data.

The challenge in finding signals of selection in individual loci underlying adaptive polygenic traits (Kelly 2006, Chevin & Hospital 2008, Eckert *et al.* 2013, Stephan 2015, Yeaman 2015) is obvious in this small data, too. Neutrality tests are well powered to detect classical and complete selective sweeps. For example, Fay &Wu's *H* measures the ratio of derived high frequency variants to intermediate frequency variants, which is expected to be elevated in the surrounding areas of a site that has been already fixed due to positive directional selection. Association studies, in contrast, search for loci that are still polymorphic. Unlike Fay and Wu's *H*, Tajima's *D* is capable of detecting partial hitchhiking events that might be more likely in the scenario of selection towards different trait optima. Tests specifically designed to detect incomplete or soft sweeps (Sabeti *et al.* 2002, Voight *et al*. 2006,

Garud *et al*. 2015) might also be useful. The power of all these, however, depends on the difference between the initial and the equilibrium frequency of the favored allele and timing of selection relative to the sampling (Pavlidis *et al.* 2012). Eckert *et al*. (2013) found that loci associated with polygenic traits had more negative Tajima's *D* values as a group than nonassociated genes, as was found in our small comparison, too. As suggested by Eckert *et al.* (2013), the identification of signals of selection might be possible only through comparing groups of genes.

3.3.3 Allele frequency patterns in bud set timing associated loci

Allele frequency patterns at individual polymorphic sites were examined both in study II and study IV. None of the four associated genes (*ftl2, phyn*, *prr1*, *lp2)* had F_{ST} outlier sites in BayeScan analysis, nor high genewise F_{ST} in study II. In the SNP data (study IV) one of the associated SNPs (*CL1154Contig1_02-143)* had a particularly high F_{ST} estimate (0.1014) among the ten populations when compared to the distribution of F_{ST} estimates in the reference SNPs. It seems, however, that these high F_{ST} estimates do not necessarily reflect clinal selection, as seen from Fig. 4. When only northern populations were considered, F_{ST} of SNP *ftl2_f1-356* was on the edge of the distribution. In this case the high F_{ST} estimate agrees better with a clinal pattern. The lack of F_{ST} outliers was also evident in *Populus balsamifera* in genes related to timing of bud set (Olson *et al.* 2013, Wang *et al.* 2014).

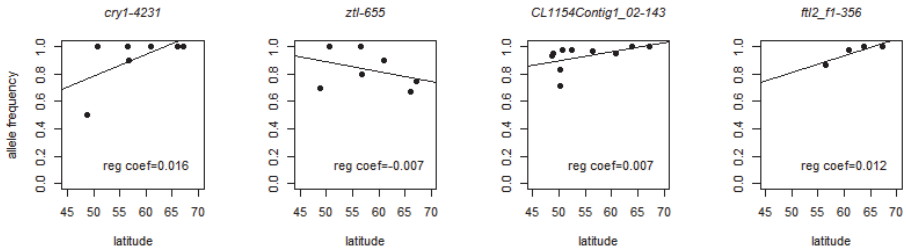

Fig. 4. Population specific allele frequencies against latitude and their respective regression lines in four SNPs that show pronounced F_{ST} estimates.

In the case of selection towards different trait optima along a clinal gradient the existence of clinal allele variation can be stronger evidence for selection than high F_{ST} alone. *prr1* and *ftl2* had significant frequency clines at some SNPs in study II. Unfortunately none of these sites were included in the genotype data in studies III and IV. The SNPs associating with timing of bud set in *prr1* and *ftl2* did not have

frequency clines. In *ftl2* one nonassociating SNP exhibited a significant cline (no multiple testing correction) and is very close to the site that showed significant cline in study II.

According to Barton (1999) part of the loci underlying a clinal polygenic trait are expected to show allele frequency clines from close to zero to close to fixation, while other loci contributing to the trait variation show very little differences in their frequencies among populations. In study IV we examined the evidence for this hypothesis. We found that within these data we do not see trait associated loci with steep allele frequency clines, i.e. frequency change from close to zero to close to fixation. Also there was no indication of bigger effect alleles having steeper clines. Instead we found a slight (yet statistically nonsignificant) enrichment of very shallow allele frequency clines among the trait associated loci in comparison to reference loci. Also in Norway spruce, Chen *et al*. (2012) saw a lack of enrichment in bud set candidate genes for bud set when standard regression coefficients in the candidate gene SNPs were compared against the reference genes. Chen *et al*. (2014) found a slight enrichment in Siberian spruce. Both, however, found an enrichment of positive evidence (BF, Bayes factor) for clinal variation within bud set candidate genes when using a Bayesian analysis tool Bayenv (Coop *et al*. 2010). Our data could be further tested with similar approach.

Another theory suggests that selection on beneficial allele combinations can be of more importance than the allele frequency changes at individual loci (Latta 1998, Le Corre & Kremer 2003, 2012, and Kremer & Le Corre 2011, see also Csillery *et al*. 2014). Allelic covariation between bud set timing associated SNPs was also examined in study IV. We did not find an elevated level of covariation of allele frequencies in comparison to reference SNPs. Ma *et al*. (2010) found large covariance in allelic effects in *Populus tremula*. In future studies with (closer to) genome wide data, more powerful statistical methods (such as Berg & Coop 2014) can be used to examine the contribution of allelic covariance further.

The lack of steep clines or increased covariance in these small data can be simply due to the fact that most likely only small part of the causative loci were included among the studied markers. In the model by Barton (1999) only few of the loci contributing to trait variation will show steep clines while most have only very shallow clines. Another possibility is that the time of sampling relative to the stage of selection is not optimal to see these effects; it may take longer than the time since post-LGM colonization for the allele frequency clines to settle. Allele covariance is, however, expected to have an important role especially in the early stages of selection (Kremer & Le Corre 2011). The power to detect greater

covariation among associated loci in comparison to random loci will most likely improve as the discovered proportion of loci underlying the trait increases. Post-LGM demographic modelling with spatial components included can further help to redefine the expectations on allele frequency clines and covariance.

4 Conclusions and future directions

The studies in this thesis aid in painting a more refined picture of genetic diversity in Scots pine at the levels of nucleotide, allozyme and phenotypic variation. Scots pine is in many ways an interesting species for the study of evolution. It is the most widespread of all pine species and is thus an interesting example of a gymnosperm that has managed to evolve and triumph through the history of this ancient lineage, facing many profound changes in its environment. New genetic research added to the wealth of existing older data in the form of allozyme studies and provenance trials offers a particularly rich set up.

In this thesis both neutral and selective aspects of the genetic variation in Scots pine were examined that complement the previous genetic research conducted with this species. Three neutral aspects were studied here, the influence of demographic population size fluctuations on the background sequence variation patterns, the distribution of variation at the nucleotide sequence level, i.e. population structure, and the level of correspondence of diversity estimates at the nucleotide and allozyme levels. We found that the genome of Scots pine is reflecting old demographic (pre-LGM) events and that the background variation can be modelled with a severe, ancient bottleneck. This model, here with more data and dense sampling of the parameter space, is congruent with the earlier demography model inference. An addition is the notion that population growth model does not outperform the bottleneck model. In future studies, more data and a better coverage of the genome are needed, along with new modelling tools that enable using such aspects of the genome data that are informative also about recent colonization history of a long lived species.

We found that F_{ST} estimates were generally low, as has been shown also in the earlier studies. We found indications that the marginal populations are slightly more differentiated than the populations from the main range populations. The northern populations were, however, less diverged from each other than the other main range populations. This gives support to the suggestion made already two decades ago based on allozyme data, that the central European populations have had more time to diverge after LGM than the northern populations that colonized their current location later. An implication from this is that there may exist fine population substructure also within the continuous part of the range despite the generally low F_{ST} estimates. The possibility of the fine structure should be examined with more data and more precise spatial analysis tools, and the possible violation of panmixia

should be taken into account in forthcoming population genetic and association studies.

The existing allozyme data in Scots pine has taught many things about this species, in particular that most of the neutral genetic variation is found within populations. We studied how well the allozyme diversity correlates with underlying nucleotide diversity. We found that at the between species level the data are congruent; both allozyme and nucleotide diversities reflect the effective population size. At the within population level, the allozyme heterozygosity is, however, not informative about the actual nucleotide variation in the coding loci. Comparing our new nucleotide data with older allozyme data implies that the between population variation is described fairly well with allozymes, although nucleotide data naturally adds more resolution.

Polygenic adaptation was studied with timing of bud set at the level of the phenotype in a common garden experiment, at the gene level by population genetic analysis of candidate loci, and by association methods combining these two levels. We saw that timing of bud set is well correlated with the latitude of origin. This clinal structure in this trait has been seen already in earlier studies. However, our sampling, covering both northern and central Europe, enabled us to effectively compare the variation in these two regions. Phenotypic and additive genetic variation was smaller in northern Europe, most likely as a result of stronger selection in the harsh northern environment.

Association study within this material, analyzed with a novel Bayesian multilocus method, revealed genetic heterogeneity between central and northern European populations. It is possible that the environmental cues used for the seasonal timekeeping are different in these two areas where the information content of the night length and light spectral qualities differ. Future studies can further reveal whether this heterogeneity implies conditional neutrality or antagonistic pleiotropy. We found associations within candidate genes for bud set timing, but also in genes with looser links to the phenotype. While the risk of false positives is never zero, some of the less expected associations may inform us about the complexity of the molecular networks underlying this trait, and also about the possibility of pleiotropic effects involved in adaptation processes.

The association study results enabled us to compare the population genetic properties of the associated versus nonassociated loci in across Europe. We identified some features that were often characterizing the associated genes, although at the individual loci the features did not deviate from the neutral expectations under the bottleneck demography. These trends need to be further

tested with larger number of genes. We also examined the theoretical predictions of allele frequency clines and allelic covariance among the associated SNP markers. We did not find any steep clines, nor a pronounced level of allele frequency covariation in the associated SNPs compared to the reference SNPs. We also saw that high single locus F_{ST} estimates in data sampled from multiple populations across an environmental gradient do not always reflect biologically meaningful adaptive differentiation.

The study of adaptation benefits from simultaneous analysis on multiple levels of variation. The strongest limitation in the data in this thesis was the low genome coverage that reflects the fact that genomic resources in this species have been sparse. This situation is, fortunately, constantly improving; first full genome conifer sequences have been published, which aids the development of genomic resources for Scots pine, too. Nevertheless, the results in this thesis give further guidelines for future adaptation and population genetic studies in Scots pine and other widespread forest trees. These results will also be of value for breeding and conservation purposes, and add to the discussion on the consequences of the rapid climate warming. I would like to finish with a quote:

"*Everything should be made as simple as possible, but not simpler.*"
– Albert Einstein

References

Alonso-Blanco C & Méndez-Vigo B (2014) Genetic architecture of naturally occurring quantitative traits in plants: an updated synthesis. Curr Opin Plant Biol 18: 37–43.

Anderson JT, Lee C, Rushworth CA, Colautti RI & Mitchell-Olds T (2013) Genetic trade-offs and conditional neutrality contribute to local adaptation. Mol Ecol 22: 699–708.

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

Antoniazza S, Kanitz R, Neuenschwander S, Burri R, Gaigher A, Roulin A & Goudet J (2014) Natural selection in a postglacial range expansion: the case of the colour cline in the European barn owl. Mol Ecol 23: 5508–5523.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D & Hu TT (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465: 627–631.

Austerlitz F & Garnier-Géré PH (2003) Modelling the impact of colonization on genetic diversity and differentiation of forest trees: interaction of life cycle, pollen flow and seed long-distance dispersal. Heredity 90: 282–290.

Austerlitz F, Mariette S, Machon N, Gouyon PH & Godelle B (2000) Effects of colonization processes on genetic diversity: Differences between annual plants and tree species. Genetics 154: 1309–1321.

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: 159–170.

Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7: 781–791.

Bank C, Ewing GB, Ferrer-Admettla, Foll M & Jensen JD (2014) Thinking too positive? Revisiting current methods of population genetic selection inference. Trends Genet 30: 540–546.

Barrett RDH & Schluter D (2008) Adaptation from standing genetic variation. Trends Ecol Evol 23: 38–44.

Barton NH (1999) Clines in polygenic traits. Genet Res 74: 223–236.

Beaumont MA, Zhang W & Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162: 2025–2035.

Berg JJ & Coop G (2014) A population genetic signal of polygenic adaptation. PLOS Genet 10: e1004412.

Bertorelle G, Benazzo A & Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol Ecol 19: 2609–2625.

Bradburd GS, Ralph PL & Coop GM (2015) A spatial framework for understanding population structure and admixture. bioRxiv doi: URI: http://dx.doi.org/10.1101/013474. Cited 2015/11/10.

Bridle JR, Polechová J, Kawata M & Butlin RK (2010) Why is adaptation prevented at ecological margins? New insights from individual-based simulations. Ecol Lett 13: 485–494.

Cardon LR & Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2: 91–99.

Cargill M, Altshuler D, Ireland J, Sklar, P, Ardlie K, Patil N & Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22: 231–238.

Charlesworth B (1998) Measures of divergence between populations and the effect of forces that recude variability. Mol Biol Evol 15: 538–543.

Cheddadi R, de Beaulieu JL, Jouzel J, Andrieu-Ponel V, Laurent JM, Reille M & Bar-Hen A (2005) Similarity of vegetation dynamics during interglacial periods. Proc Natl Acad Sci USA 102: 13939–13943.

Cheddadi R, Vendramin GG, Litt T, Francois L, Kageyama M, Lorentz S & Lunt D (2006) Imprints of glacial refugia in the modern genetic diversity of *Pinus sylvestris*. Global Ecol Biogeogr 15: 271–282.

Chen J, Källman T, Ma X, Gyllenstrand N, Zaina G, Morgante M & Lascoux M (2012) Disentangling the roles of history and local selection in shaping clinal variation of allele frequencies and gene expression in Norway spruce (*Picea abies*) Genetics 191: 865–881.

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: 1025–1038.

Chen K, Gong H, Chen G, Wang S & Zhang C (2004) Gradual drought under field conditions influences the glutathione metabolism, redox balance and energy supply in spring wheat. J Plant Growth Regul 23: 20–28.

Chevin LM & Hospital F (2008) Selective sweep at a quantitative trait locus in the presence of background genetic variation. Genetics 180: 1645–1660.

Clapham DH, Dormling I, Ekberg L, Eriksson G, Qamaruddin M & Vince-Prue D (1998) Latitudinal cline of requirement for far-red light for the photoperiodic control of budset and extension growth in *Picea abies* (Norway spruce). Physiol Plantarum 102: 71–78.

Clapham DH, Ekberg I, Eriksson G, Norell L & Vince-Prue D (2002) Requirement for far-red light to maintain secondary needle extension growth in northern but not southern populations of *Pinus sylvestris* (Scots pine). Physiol Plantarum 114: 207–212.

Clark PU, Dyke AS, Skakun JD, Carlson AE, Clark J. Wohlfarth B & McCabe AM (2009) The last glacial maximum. Science 325: 710–714.

Cooke JE, Eriksson ME & Junttila O (2012) The dynamic nature of bud dormancy in trees: environmental control and molecular mechanisms. Plant Cell Environ 35: 1707–1728.

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

Csilléry K, Blum MGB, Gaggiotti OE & Franȷois O (2010) Approximate Bayesian computation (ABC) in practice. Trends Ecol Evol 25: 410–418.

Csilléry K, LaLagüe H, Vendramin GG, González-Martínez SC, Facy B & Oddou-Muratorio S (2014) Detecting short spatial scale local adaptation and epistatic selection in climate-related candidate genes in European beech (*Fagus sylvatica*) populations. Mol Ecol 23: 4696–4708.

Cumbie W, Eckert A, Wegrzyn J, Whetten R, Neale D & Goldfarb B (2011) Association genetics of carbon isotope discrimination, height and foliar nitrogen in a natural population of *Pinus taeda* L. Heredity 107: 105–114.

Currat M, Ray N & Excoffier L (2004) SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. Mol Ecol Notes 4: 139–142.

Darwin C (1845) Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle round the world under the command of Capt. Fitz Roy, R.N., 2nd edition. London, John Murray.

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

Devlin B & Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.

Duchen P, Zivkovic D, Hutter S, Stephan W & Laurent S (2013) Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. Genetics 193: 291–301.

Dueck T, van Ieperen W & Taulavuori K (2015) Light perception, signaling and plant responces to spectral quality and photoperiod in natural and horticultural environments. Environ Exp Bot doi: URI: http://dx.doi.org/10.1016/j.envexpbot.2015.06.012. Cited 2015/11/10.

Dvornyk V (2001) Genetic variability and differentiation of geographically marginal Scots pine populations from Ukraine. Silvae Genet 50: 64–69.

Dvornyk V, Sirviö A, Mikkonen M & Savolainen O (2002) Low nucleotide diversity at the *pal1* locus in the widely distributed *Pinus sylvestris*. Mol Biol Evol 19: 179–188.

Eckert AJ, Bower AD, González-Martínez SC, Wegrzyn JL, Coop G & Neale DB (2010a) Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). Mol Ecol 19: 3789–3805.

Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV & Neale DB (2009a) Association genetics of coastal douglas fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. Cold hardiness related traits. Genetics 182: 1289–1302.

Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, Gonzalez-Martinez SC & Neale DB (2010b) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics 185: 969–982.

Eckert AJ, Maloney PE, Vogler DR, Jensen CE, Mix AD & Neale DB (2015) Local adaptation at fine spatial scales: an example from sugar pine (*Pinus lambertiana*, Pinaceae). Tree Genet Genom 11: 1–17.

Eckert AJ, Wegrzyn JL, Cumbie WP, Goldfarb B, Huber DA, Tolstikov V & Neale DB (2012) Association genetics of the loblolly pine (*Pinus taeda*, Pinaceae) metabolome. New Phytol 193: 890–902.

Eckert AJ, Wegrzyn JL, Liechty JD, Lee JM, Cumbie WP, Davis JM & Neale DB (2013) The evolutionary genetics of the genes underlying phenotypic associations for loblolly pine (*Pinus taeda*, Pinaceae). Genetics 195: 1353–1372.

Eckert AJ, Wegrzyn JL, Pande B, Jermstad KD, Lee JM, Liechty JD & Neale DB (2009b) Multilocus patterns of nucleotide diversity and divergence reveal positive selection at candidate genes related to cold-hardiness in coastal Douglas-fir (*Pseudotsuga menziesii* var. *menziesii*). Genetics 185: 969–982.

Eckert C, Samis K & Lougheed S (2008) Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. Mol Ecol 17: 1170–1188.

Eiche V (1966) Cold damage and plant mortality in experimental provenance plantations with Scots pine in northern Sweden. Stud For Suec 36: 1–218.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol 29: 51–63.

Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W & Chen J (2014) Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. Nat Genet 46: 1089–1096.

Excoffier L, Foll M & Petit RJ (2009) Genetic consequences of range expansions. Annu Rev Ecol Evol Syst 40: 481–501.

Excoffier L, Laval G & Schneider S (2005) Arlequin ver. 3.1: an integrated software package for population genetics data analysis. Evol Bioinform Online 1: 47–50.

Farjon A (1998) World checklist and bibliography of conifers. Kew, Royal Botanic Gardens.

Fay JC & Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.

Ferris C, King RA, Vainola R & Hewitt GM (1998) Chloroplast DNA recognizes three refugial sources of European oaks and suggests independent eastern and western immigrations to Finland. Heredity 80: 584–593.

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

Fitzpatrick S, Gerberich J, Kronenberger J, Angeloni L & Funk W (2015) Locally adapted traits maintained in the face of high gene flow. Ecol Lett 18: 37–47.

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: 977–993.

François O, Blum MGB, Jakobsson M & Rosenberg NA (2008) Demographic history of European populations of *Arabidopsis thaliana*. PLOS Genet 4: e1000075.

Franklin KA & Quail PH (2010) Phytochrome functions in *Arabidopsis* development. J Exp Bot 61: 11–24.

García-Gil MR, Mikkonen M & Savolainen O (2003) Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. Mol Ecol 12: 1195–1206.

García-Ramos G & Kirkpatrick M (1997) Genetic models of adaptation and gene flow in peripheral populations. Evolution 51: 21–28.

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: e1005004.

Geroldinger L & Bürger R (2015) Clines in quantitative traits: The role of migration patterns and selection scenarios. Theor Popul Biol 99: 43–66.

Giertych M & Mátyás C (2013) Genetics of Scots pine. Amsterdam, Elsevier.

Goldstein DB & Holsinger KE (1992) Maintenance of polygenic variation in spatially structured populations: roles for local mating and genetic redundancy. Evolution 46: 412–429.

Goncharenko GG, Silin AE & Padutov VE (1994) Allozyme variation in natural populations of Eurasian pines III. Population structure, diversity, differentiation and gene flow in central and isolated populations of *Pinus sylvestris* L. in Eastern Europe and Siberia. Silvae Genet 43: 119–132.

González-Martínez S, Huber D, Ersoz E, Davis J & Neale D (2008) Association genetics in *Pinus taeda* L. II. Carbon isotope discrimination. Heredity 101: 19–26.

González-Martínez SC, Wheeler NC, Ersoz E, Nelson CD & Neale DB (2007) Association genetics in *Pinus taeda* L. I. Wood property traits. Genetics 175: 399–409.

Goudet J (2014) hierfstat: Estimation and tests of hierarchical F-statistics. R package version 0.04-14. URI: http://CRAN.R-project.org/package=hierfstat. Cited 2015/11/10.

Gronau I, Hubisz MJ, Gulko B, Danko CG & Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. Nat Genet 43: 1031–1034.

Gullberg U, Yazdani R, Rudin D & Ryman N (1985) Allozyme variation in Scots pine (*Pinus sylvestris* L) in Sweden. Silvae Genet 34: 193–201.

Gupta KJ, Shah JK, Brotman Y, Jahnke K, Willmitzer L, Kaiser WM, Bauwe H & Igamberdiev AU (2012) Inhibition of aconitase by nitric oxide leads to induction of the alternative oxidase and to a shift of metabolism towards biosynthesis of amino acids. J Exp Bot 63: 1773–1784.

Gyllenstrand N, Clapham D, Kallman T & Lagercrantz U (2007) A Norway spruce *FLOWERING LOCUS T* homolog is implicated in control of growth rhythm in conifers. Plant Physiol 144: 248–257.

Gyllenstrand N, Karlgren A, Clapham D, Holm K, Hall A, Gould PD & Lagercrantz U (2014) No time for spruce: rapid dampening of circadian rhythms in *Picea abies* (L. Karst). Plant Cell Physiol 55: 535–550.

Haldane JBS (1930) A mathematical theory of natural and artificial selection. (Part VI, Isolation.). Math Proc Cambridge 26: 220–230.

Hall M, Lowry D & Willis J (2010) Is local adaptation in *Mimulus guttatus* caused by trade-offs at individual loci? Mol Ecol 19: 2739–2753.

Hamrick JL & Godt MJW (1996) Effects of life history traits on genetic diversity in plant species. Philos Trans R Soc Lond B Biol Sci 351: 1291–1298.

Hamrick JL, Godt MJW & Sherman-Broyles SL (1992) Factors influencing levels of genetic diversity in woody plant species. New Forest 6: 95–124.

Hedrick PW (2006) Genetic polymorphism in heterogeneous environments: The age of genomics. Annu Rev Evol Ecol Syst 37: 67–93.

Hedrick PW (1999) Perspective: highly variable loci and their interpretation in evolution and conservation. Evolution 53: 313–318.

Hedrick PW, Ginevan ME & Ewing EP (1976) Genetic polymorphism in heterogeneous environments. Annu Rev Ecol Syst 7: 1–32.

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

Heuertz M, De Paoli E, Källman T, Larsson H, Jurman I, Morgante M & Gyllenstrand N (2006) Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce *Picea abies* (L.) Karst. Genetics 174: 2095–2105.

Hey J (2010) HKA – a program for conducting the HKa test of natural selection. URI: https://bio.cst.temple.edu/~hey/software/software.htm#HKA. Cited 2015/11/10.

Holliday JA, Ritland K & Aitken SN (2010a) Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (*Picea sitchensis*). New Phytol 188: 501–514.

Holliday JA, Yuen M, Ritland K & Aitken S (2010b) Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. Mol Ecol 19: 3857–3864.

Howe GT, Aitken SN, Neale DB, Jermstad KD, Wheeler NC & Chen THH (2003) From genotype to phenotype: unraveling the complexities of cold adaptation in forest trees. Can J Bot 81: 1247–1266.

Huber CD, Nordborg M, Hermisson J & Hellmann I (2014) Keeping it local: evidence for positive selection in Swedish *Arabidopsis thaliana*. Mol Biol Evol 31: 3026–3039.

Hudson RR (2001) Two-locus sampling distributions and their application. Genetics 159: 1805–1817.

Hudson RR, Bailey K, Skarecky D, Kwiatowski J & Ayala FJ (1994) Evidence for positive selection in the superoxide-dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics 136: 1329–1340.

Hudson RR, Kreitman M & Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153–159.

Huntley B & Birks HJB (1983) An atlas of past and present pollen maps for Europe 0–13 000 years ago. Cambridge, Cambridge University Press.

Illumina (2015) Assay Design Tool (ADT) URI: https://support.illumina.com/array/array_software/assay_design_tool.html. Cited 2015/11/10.

Ingvarsson PK (2008a) Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. Genetics 180: 329–340.

Ingvarsson PK, García MV, Hall D, Luquez V & Jansson S (2006) Clinal variation in *phyB2*, a candidate gene for day-length-induced growth cessation and bud set, across a latitudinal gradient in European aspen (*Populus tremula*). Genetics 172: 1845–1853.

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

Ingvarsson PK & Street NR (2011) Association genetics of complex traits in plants. New Phytol 189: 909–922.

Jain K & Stephan W (2015) Response of polygenic traits under stabilizing selection and mutation when loci have unequal effects. G3 Genes Genom Genet 5: 1065–1074.

Jaramillo-Correa JP, Rodriguez-Quilon I, Grivet D, Lepoittevin C, Sebastiani F, Heuertz M & Gonzalez-Martinez SC (2015) Molecular proxies for climate maladaptation in a long-lived tree (*Pinus pinaster* Aiton, Pinaceae). Genetics 199: 793–807.

Jesus FF, Wilkins JF, Solferini VN & Wakeley J (2006) Expected coalescence times and segregating sites in a model of glacial cycles. Genet Mol Res 5: 466–474.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ & Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics 178: 1709–1723.

Karhu A, Hurme P, Karjalainen M, Karvonen P, Kärkkäinen K, Neale D & Savolainen O (1996) Do molecular markers reflect patterns of differentiation in adaptive traits of conifers? Theor Appl Genet 93: 215–221.

Karlgren A, Gyllenstrand N, Källman T, Sundström JF, Moore D, Lascoux M & Lagercrantz U (2011) Evolution of the PEBP gene family in plants: Functional diversification in seed plant evolution. Plant Physiol 156: 1967–1977.

Kawecki TJ & Ebert D (2004) Conceptual issues in local adaptation. Ecol Lett 7: 1225–1241.

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: 3143–3152.

Kelly JK (2006) Geographical variation in selection, from phenotypes to molecules. Am Nat 167: 481–495.

Kirkpatrick M & Barton NH (1997) Evolution of a species' range. Am Nat 150: 1–23.

Klopfstein S, Currat M & Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. Mol Biol Evol 23: 482–490.

Knürr T, Läärä E & Sillanpää MJ (2013) Impact of prior specifications in a shrinkage-inducing Bayesian model for quantitative trait mapping and genomic prediction. Genet Sel Evol 45: 24.

De Kort, H, Vandepitte K, Bruun HH, Closset-Kopp D, Honnay O & Mergeay J (2014) Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species *Alnus glutinosa*. Mol Ecol 23: 4709–4721.

Koski V (1970) A study of pollen dispersal as a mechanism of gene flow in conifers. Helsinki, Valtion painatuskeskus.

Kovach A, Wegrzyn JL, Parra G, Holt C, Bruening GE, Loopstra CA & Neale DB (2010) The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics 11: 420.

Kremer A & Le Corre V (2011) Decoupling of differentiation between traits and their underlying genes in response to divergent selection. Heredity 108: 375–385.

Kremer A, Ronce O, Robledo-Arnuncio JJ, Guillaume F, Bohrer G, Nathan R, Bridle JR, Gomulkiewicz R, Klein EK & Ritland K (2012) Long-distance gene flow and adaptation of forest trees to rapid climate change. Ecol Lett 15: 378–392.

Källman T, De Mita S, Larsson H, Gyllenstrand N, Heuertz M, Parducci L & Lascoux M (2014) Patterns of nucleotide diversity at photoperiod related genes in Norway spruce [*Picea abies* (L.) Karst.]. PLOS One 9: e95306.

Kärkkäinen HP & Sillanpää MJ (2012) Robustness of Bayesian multilocus association models to cryptic relatedness. Ann Hum Genet 76: 510–523.

Lagercrantz U (2009) At the end of the day: a common molecular mechanism for photoperiod responses in plants? J Exp Bot 60: 2501–2515.

Latta RG (1998) Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. Am Nat 151: 283–292.
Le Corre V & Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. Mol Ecol 21: 1548–1566.
Le Corre V & Kremer A (2003) Genetic variability at neutral markers, quantitative trait loci and trait in a subdivided population under selection. Genetics 164: 1205–1219.
Leinonen T, McCairns RJS, O'Hara RB & Merilä J (2013) Q_{ST}-F_{ST} comparisons: evolutionary and ecological insights from genomic heterogeneity. Nat Rev Genet 14: 179–190.
Lenormand T (2002) Gene flow and the limits to natural selection. Trends Ecol Evol 17: 183–189.
Lepoittevin C (2009) Association genetics in maritime pine (*Pinus pinaster* Ait.) for growth and wood quality traits. Dissertation, University of Bordeaux.
Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T & Lawson DJ (2015) The fine-scale genetic structure of the British population. Nature 519: 309–314.
Li J, Das K, Fu G, Li R & Wu R (2011) The Bayesian lasso for genome-wide association studies. Bioinformatics 27: 516–523.
Librado P & Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451–1452.
Liu X & Fu Y (2015) Exploring population size changes using SNP frequency spectra. Nat Genet 47: 555–559.
Long AD & Langley CH (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res 9: 720–731.
Lorenzo FR, Huff C, Myllymäki M, Olenchock B, Swierczek S, Tashi T & McClain DA (2014) A genetic mechanism for Tibetan high-altitude adaptation. Nat Genet 46: 951–956.
Lotterhos KE & Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Mol Ecol 24: 1031–1046.
Lottenhos KE & Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. Mol Ecol 23: 2178–2192.
Ma XF, Hall D, Onge KRS, Jansson S & Ingvarsson PK (2010) Genetic differentiation, clinal variation and phenotypic associations with growth cessation across the *Populus tremula* photoperiodic pathway. Genetics 186: 1033–1044.
MacPherson A, Hohenlohe PA & Nuismer SL (2015) Trait dimensionality explains widespread variation in local adaptation. Proc R Soc Lond B Biol Sci 282: 20141570.
Marjoram P, Zubair A & Nuzhdin SV (2014) Post-GWAS: where next? More samples, more SNPs or more biology? Heredity 112: 79–88.
Martin G & Lenormand T (2015) The fitness effect of mutations across environments: Fisher's geometrical model with multiple optima. Evolution 69: 1433–1447.
Mathieson I & McVean G (2015) Demography and the age of rare variants. PLOS Genet 10: e1004528.
Matuszewski S, Hermisson J & Kopp M (2015) Catch me if you can: Adaptation from standing genetic variation to a moving phenotypic optimum. Genetics 200: 1255–1274.

McDonald JH & Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351: 652–654.
McKay JK & Latta RG (2002) Adaptive population divergence: markers, QTL and traits. Trends Ecol Evol 17: 285–291.
McVean G, Awadalla P & Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160: 1231–1241.
Meirmans PG (2012) The trouble with isolation by distance. Mol Ecol 21: 2839–2846.
Merilä J & Crnokrak P (2001) Comparison of genetic differentiation at marker loci and quantitative traits. J Evol Biol 14: 892–903.
Mikola J (1982) Bud-set phenology as an indicator of climatic adaptation of Scots pine in Finland. Silva Fenn 16: 178–184.
Mimura M & Aitken SN (2010) Local adaptation at the range peripheries of Sitka spruce. J Evol Biol 23: 249–258.
Mimura M & Aitken SN (2007) Adaptive gradients and isolation-by-distance with postglacial migration in *Picea sitchensis*. Heredity 99: 224–232.
Mirov N (1967) The genus *Pinus*. New York, Ronald Press.
De Mita S, Ronfort J, McKhann HI, Poncet C, El Malki R & Bataillon T (2007) Investigation of the demographic and selective forces shaping the nucleotide diversity of genes involved in Nod factor signaling in *Medicago truncatula*. Genetics 177: 2123–2133.
De Mita S, Thuillet A, Gay L, Ahmadi N, Manel S, Ronfort J & Vigouroux Y (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. Mol Ecol 22: 1383–1399.
Moeller DA, Geber MA & Tiffin P (2011) Population genetics and the evolution of geographic range limits in an annual plant. Am Nat 178: 44–61.
Moeller DA & Tiffin P (2008) Geographic variation in adaptation at the molecular level: a case study of plant immunity genes. Evolution 62: 3069–3081.
de Montaigu A, Giakountis A, Rubin M, Toth R, Cremer F, Sokolova V & Coupland G (2015) Natural diversity in daily rhythms of gene expression contributes to phenotypic variation. Proc Natl Acad Sci USA 112: 905–910.
Morgenstern EK (1996) Geographic Variation in Forest Trees, Genetic Basis and Application of Knowledge in Silviculture. Vancouver, UBC Press.
Muller M, Leppälä J & Savolainen O (2008) Genome-wide effects of postglacial colonization in *Arabidopsis lyrata*. Heredity 100: 47–58.
Muona O & Harju A (1989) Effective population sizes, genetic variability, and mating system in natural stands and seed orchards of *Pinus sylvestris*. Silvae Genet 38: 221–228.
Muona O & Szmidt AE (1985) A multilocus study of natural populations of *Pinus sylvestris*. Lect Notes Biomath 60: 226–240.
Nakamichi N (2014) Adaptation to the local environment by modifications of the photoperiod response in crops. Plant Cell Physiol 56: 594–604.

Naydenov K, Senneville S, Beaulieu J, Tremblay F & Bousquet J (2007) Glacial vicariance in Eurasia: mitochondrial DNA evidence from Scots pine for a complex heritage involving genetically distinct refugia at mid-northern latitudes and in Asia Minor. BMC Evol Biol 7: 233.

Neale DB & Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9: 325–330.

Nei M, Maruyama T & Chakraborty R (1975) The bottleneck effect and genetic variability in populations. Evolution 29: 1–10.

Nordborg M & Tavaré S (2002) Linkage disequilibrium: what history has to tell us. Trends Genet 18: 83–90.

Notivol E, García-Gil MR, Alía R & Savolainen O (2007) Genetic variation of growth rhythm traits in the limits of a latitudinal cline in Scots pine. Can J For Res 37: 540–551.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y, Scofield DG, Vezzi F, Delhomme N, Giacomello S & Alexeyenko A (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497: 579–584.

Oakley CG, Ågren J, Atchison RA & Schemske DW (2014) QTL mapping of freezing tolerance: links to fitness and adaptive trade-offs. Mol Ecol 23: 4304–4315.

O'Connor TD, Fu W, NHLBI GO Exome Sequencing project, ESP Population Genetics and Statistical Analysis Working Group, Turner E, Mychaleckyj JC & Akey JM (2015) Rare variation facilitates inferences of fine-scale population structure in humans. Mol Biol Evol 32: 653–660.

Ohta T (1992) Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc Natl Acad Sci USA 79: 1940–1944.

Oleksyn J, Tjoelker MG & Reich PB (1998) Adaptation to changing environment in Scots pine populations across a latitudinal gradient. Silva Fenn 32: 129–140.

Oleksyn J, Tjoelker MG & Reich PB (1992) Growth and Biomass Partitioning of Populations of European *Pinus sylvestris* L. Under Simulated 50° and 60° N Daylengths: Evidence for Photoperiodic Ecotypes. New Phytol 120: 561–574.

Olson MS, Levsen N, Soolanayakanahally RY, Guy RD, Schroeder WR, Keller SR & Tiffin P (2013) The adaptive potential of *Populus balsamifera* L. to phenology requirements in a warmer global climate. Mol Ecol 22: 1214–1230.

Orr HA (2005) The genetic theory of adaptation: a brief history. Nat Rev Genet 6: 119–127.

Oubida RW, Gantulga D, Zhang M, Zhou L, Bawa R & Holliday JA (2015) Partitioning of multivariate phenotypes using regression trees reveals complex patterns of adaptation to climate across the range of black cottonwood (*Populus trichocarpa*) Front Plant Sci 6: 181.

Palmé AE, Wright M & Savolainen O (2008) Patterns of divergence among conifer ESTs and polymorphism in *Pinus sylvestris* identify putative selective sweeps. Mol Biol Evol 25: 2567–2577.

Pankin A, Campoli C, Dong X, Kilian B, Sharma R, Himmelbach A & von Korff M (2014) Mapping-by-sequencing identifies *HvPHYTOCHROME C* as a candidate gene for the early maturity 5 locus modulating the circadian clock and photoperiodic flowering in barley. Genetics 198: 383–396.

Pardo-Diaz C, Salazar C & Jiggins CD (2015) Towards the identification of the loci of adaptive evolution. Methods Ecol Evol 6: 445–464.

Pavlicev M & Wagner GP (2012) A model of developmental evolution: selection, pleiotropy and compensation. Trends Ecol Evol 27: 316–322.

Pavlidis P, Metzler D & Stephan W (2012) Selective sweeps in multilocus models of quantitative traits. Genetics 192: 225–239.

Petit JR, Jouzel J, Raynaud D, Barkov NI, Barnola JM & Stievenard M (1999) Climate and atmospheric history of the past 420000 years from the Vostok ice core, Antarctica. Nature 399: 429–436.

Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, Cheddadi R & Vendramin GG (2003) Glacial refugia: Hotspots but not melting pots of genetic diversity. Science 300: 1563–1565.

Petit RJ, Brewer S, Bordacs S, Burg K, Cheddadi R & Kremer A (2002) Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. Forest Ecol Manag 156: 49–74.

Pickrell JK & Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLOS Genet 8: e1002967.

Polechova J & Barton NH (2015) Limits to adaptation along environmental gradients. Proc Natl Acad Sci USA 112: 6401–6406.

Polechova J & Barton NH (2011) Genetic drift widens the expected cline but narrows the expected cline width. Genetics 189: 227–235.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA & Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.

Pritchard JK & Di Rienzo A (2010) Adaptation–not by sweeps alone. Nat Rev Genet 11: 665–667.

Pritchard JK, Stephens M, Rosenberg NA & Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67: 170–181.

Prunier J, Pelgas B, Gagnon F, Desponts M, Isabel N, Beaulieu J & Bousquet J (2013) The genomic architecture and association genetics of adaptive characters using a candidate SNP approach in boreal black spruce. BMC Genomics 14: 368.

Prus-Glowacki W & Stephan W (1994) Genetic variation of *Pinus sylvestris* from Spain in relation to other European populations. Silvae Genet 43: 7–14.

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

Puglisi S & Attolico M (2000) Allozyme variation in natural populations of the Italian range of *Pinus sylvestris* L. For Genet 7: 221–232.

Pyhäjärvi T, García-Gil MR, Knürr T, Mikkonen M, W. W & Savolainen O (2007) Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. Genetics 177: 1713–1724.

Pyhäjärvi T, Salmela MJ & Savolainen O (2008) Colonization routes of *Pinus sylvestris* inferred from distribution of mitochondrial DNA variation. Tree Genet Genom 4: 247–254.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URI: http://www.R-project.org/. Cited 2015/11/10.

Rainey D, Mitton J & Monson R (1987) Associations between enzyme genotypes and dark respiration in perennial ryegrass, *Lolium perenne* L. Oecologia 74: 335–338.

Ramshaw JA, Coyne JA & Lewontin RC (1979) The sensitivity of gel electrophoresis as a detector of genetic variation. Genetics 93: 1019–1037.

Ray N, Currat M, Foll M & Excoffier L (2010) SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. Bioinformatics 26: 2993–2994.

Remington DL (2015) Alleles versus mutations: Understanding the evolution of genetic architecture requires a molecular perspective on allelic origins. Evolution doi:10.1111/evo.12775.

Richardson JL, Urban MC, Bolnick DI & Skelly DK (2014) Microgeographic adaptation and the spatial scale of evolution. Trends Ecol Evol 29: 165–176.

Richardson DM & Rundel PW (1998) Ecology and biogeography of *Pinus*: an introduction. In: Richardson DM (ed) Ecology and biogeography of *Pinus*. Cambridge, Cambridge University Press: 3–46.

Risch N & Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516–1517.

Robinson MR, Wray NR & Visscher PM (2014) Explaining additional genetic variation in complex traits. Trends Genet 30: 124–132.

Robledo-Arnuncio JJ (2011) Wind pollination over mesoscale distances: an investigation with Scots pine. New Phytol 190: 222–233.

Robledo-Arnuncio JJ & Gil L (2005) Patterns of pollen dispersal in a small population of *Pinus sylvestris* L. revealed by total-exclusion paternity analysis. Heredity 94: 13–22.

Rockman MV (2012) The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution 66: 1–17.

Rozas J, Sánchez-DelBarrio JC, Messeguer X & Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496–2497.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ & McDonald GJ (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.

Salmela E, Lappalainen T, Fransson I, Andersen PM, Dahlman-Wright K, Fiebig A & Lahermo P (2008) Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. PLOS One 3: e3519.

Salmela MJ (2014) Rethinking local adaptation: Mind the environment! For Ecol Manag 312: 271–281.
Savolainen O, Kujala ST, Sokol C, Pyhäjärvi T, Avia K, Knürr T, Kärkkäinen K & Hicks S (2011) Adaptive potential of northernmost tree populations to climate change, with emphasis on Scots pine (*Pinus sylvestris* L.). J Hered 102: 526–536.
Savolainen O, Lascoux M & Merilä J (2013) Ecological genomics of local adaptation. Nat Rev Genet 14: 807–820.
Savolainen O & Pyhäjärvi T (2007) Genomic diversity in forest trees. Curr Opin Plant Biol 10: 162–167.
Savolainen O, Pyhäjärvi T & Knürr T (2007) Gene flow and local adaptation in trees. Annu Rev Evol Ecol Syst 38: 595–619.
Scalfi M, Piotti A, Rossi M & Piovani P (2009) Genetic variability of Italian southern Scots pine (*Pinus sylvestris* L.) populations: the rear edge of the range. Eur J For Res 128: 377–386.
Schnee FB & Thompson Jr JN (1984) Conditional neutrality of polygene effects. Evolution 38: 42–46.
Scotti I, González-Martínez SC, Budde KB & Lalagüe H (2015) Fifty years of genetic studies: what to make of the large amounts of variation found within populations? Ann For Sci doi:10.1007/s13595-015-0471-z.
Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q & Nordborg M (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44: 825–830.
Shigapov ZK, Bakhtiyarova R & Yanbaev YA (1995) Genetic variation and differentiation in natural populations of the common pine *Pinus sylvestris* L. Genetika 31: 1386–1393.
Sinclair WT, Morman JD & Ennos RA (1999) The postglacial history of Scots pine (*Pinus sylvestris* L.) in western Europe: evidence from mitochondrial DNA variation. Mol Ecol 8: 83–88.
Slatkin M (1993) Isolation by Distance in Equilibrium and Nonequilibrium Populations. Evolution 47: 264–279.
Slatkin M (1987) Gene flow and the geographic structure of natural populations. Science 236: 787–792.
Slatkin M (1978) Spatial patterns in the distributions of polygenic characters. J Theor Biol 70: 213–228.
Slatkin M (1973) Gene flow and selection in a cline. Genetics 75: 733–756.
Slatkin M & Excoffier L (2012) Serial founder effects during range expansion: a spatial analog of genetic drift. Genetics 191: 171–181.
Song YH, Estrada DA, Johnson RS, Kim SK, Lee SY, MacCoss MJ & Imaizumi T (2014) Distinct roles of FKF1, Gigantea, and Zeitlupe proteins in the regulation of Constans stability in *Arabidopsis* photoperiodic flowering. Proc Natl Acad Sci USA 111: 17672–17677.
Sork VL, Aitken SN, Dyer RJ, Eckert AJ, Legendre P & Neale DB (2013) Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. Tree Genet Genom 9: 901–911.

Städler T, Haubold B, Merino C, Stephan W & Pfaffelhuber P (2009) The Impact of Sampling Schemes on the Site Frequency Spectrum in Non-equilibrium Subdivided Populations. Genetics 182: 205–216.

Stephan W (2015) Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. Mol Ecol doi:10.1111/mec.13288.

Stock AJ, Campitelli BE & Stinchcombe JR (2014) Quantitative genetic variance and multivariate clines in the Ivyleaf morning glory, *Ipomoea hederacea*. Philos Trans R Soc Lond B Biol Sci 369: 20130259.

Storey JD & Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100: 9440–9445.

Strømme CB, Julkunen-Tiitto R, Krishna U, lavola A, Olsen JE & Nybakken L (2015) UV-B and temperature enhancement affect spring and autumn phenology in *Populus tremula*. Plant Cell Environ 38: 867–877.

Stucki S, Orozco-terWengel P, Bruford MW, Colli L, Masembe C, Negrini R, Taberlet P, Joost S & the NEXTGEN Consortium (2014) High performance computation of landscape genomic models integrating local indices of spatial association. arXiv: 1405.7658v2.

Svendsen JI, Alexanderson H, Astakhov VI, Demidov I, Dowdeswell JA, Funder S & Stein R (2004) Late quarternary ice sheet history of northern Eurasia. Quart Sci Rev 23: 1229–1271.

The 1000 Genomes project Consortium (2015) A global reference for human genetic variation. Nature 526: 68–74.

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

Tiffin P & Ross-Ibarra J (2014) Advances and limits of using population genetics to understand local adaptation. Trends Ecol Evol 29: 673–680.

Toprak E, Veres A, Michel J, Chait R, Hartl DL & Kishony R (2012) Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. Nat Genet 44: 101–105.

De La Torre AR, Birol I, Bousquet J, Ingvarsson PK, Jansson S, Jones SJ, Keeling CI, MacKay J, Nilsson O, Ritland K, Street N, Yanchuk A, Zerbe P & Bohlmann J (2014) Insights into conifer giga-genomes. Plant Physiol 166: 1724–1732.

Turchin MC, Chiang CWK, Palmer CD, Sankararaman S, Reich D, Genetic Investigation of ANthropometric Traits (GIANT) Consortium & Hirschhorn JN (2012) Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat Genet 44: 1015–1019.

Turner A, Beales J, Faure S, Dunford RP & Laurie DA (2005) The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. Science 310: 1031–1034.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I & Rokhsar D (2006) The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313: 1596–1604.

Vaartaja O (1959) Evidence of photoperiodic ecotypes in trees. Ecol Monogr 29: 91–111.
Vallejos CE (1983) Enzyme activity staining. In: Tanksley SD & Orton TJ (eds) Isozymes in plant genetics and breeding, Part A; Enzyme activity staining. Amsterdam, Elsevier: 469–516.
Viherä-Aarnio A, Häkkinen R, Partanen J, Luomajoki A & Koski V (2005) Effects of seed origin and sowing time on timing of height growth cessation of *Betula pendula* seedlings. Tree Physiol 25: 101–108.
de Vladar HP & Barton N (2014) Stability and response of polygenic traits to stabilizing selection and mutation. Genetics 197: 749–767.
Voight BF, Adams AM, Frisse LA, Qian YD, Hudson RR & Di Rienzo A (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci USA 102: 18508–18513.
Voight BF, Kudaravalli S, Wen X & Pritchard JK (2006) A map of recent positive selection in the human genome. PLOS Biol 4: 446–458.
Wachowiak W, Balk PA & Savolainen O (2009) Search for nucleotide diversity patterns of local adaptation in dehydrins and other cold-related candidate genes in Scots pine (*Pinus sylvestris* L.). Tree Genet Genom 5: 117–132.
Wachowiak W, Wójkiewicz B, Cavers S & Lewandowski A (2014) High genetic similarity between Polish and North European Scots pine (*Pinus sylvestris* L.) populations at nuclear gene loci. Tree Genet Genom 10: 1015–1025.
Wakeley J & Aliacar N (2001) Gene genealogies in a metapopulation. Genetics 159: 893–905.
Wang L, Tiffin P & Olson MS (2014) Timing of success: expression phenotype and local adaptation related to latitude in the boreal forest tree, *Populus balsamifera*. Tree Genet Genom 10: 911–922.
Wang X, Szmidt AE & Lindgren D (1991) Allozyme differentiation among populations of *Pinus sylvestris* (L.) from Sweden and China. Hereditas 114: 219–226.
Watterson GA (1978) The homozygosity test of neutrality. Genetics 88: 405–417.
Wegrzyn JL, Lee JM, Tearse BR & Neale DB (2008) TreeGenes: A forest tree genome database. Int J Plant Genomics 2008: 412875.
Wegrzyn JL, Liechty JD, Stevens KA, Wu LS, Loopstra CA & Neale DB (2014) Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. Genetics 196: 891–909.
Weinig C, Ewers BE & Welch SM (2014) Ecological genomics and process modeling of local adaptation to climate. Curr Opin Plant Biol 18: 66–72.
Whitlock MC (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol 18: 1368–1373.
Whitlock MC & McCauley DE (1999) Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$. Heredity 82: 117–125.
Willis KJ & van Andel TH (2004) Trees or no trees? The environments of central and eastern Europe during the Last Glaciation. Quat Sci Rev 23: 2369–2387.
Wright S (1935) The analysis of variance and the correlations between relatives with respect to deviations from an optimum. J Genet 30: 243–256.

Wright S (1931) Evolution in Mendelian populations. Genetics 16: 97–159.
Würschum T & Kraft T (2015) Evaluation of multi-locus models for genome-wide association studies: a case study in sugar beet. Heredity 114: 281–290.
Yang J, Lee SH, Goddard ME & Visscher PM (2011a) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88: 76–82.
Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, Smith AV, Ingelsson E, O'Connell JR & Mangino M (2011b) Genomic inflation factors under polygenic inheritance. Eur J Hum Genet 19: 807–812.
Yeaman S (2015) Local adaptation by alleles of small effect. Am Nat doi:10.1086/682405.
Yeaman S & Jarvis A (2006) Regional heterogeneity and gene flow maintain variance in a quantitative trait within populations of lodgepole pine. Proc R Soc Lond B Biol Sci 273: 1587–1593.
Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF & Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203–208.
Zeng K, Fu YX, Shi S & Wu CI (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174: 1431–1439.
Zeng K, Shi S & Wu CI (2007) Compound tests for the detection of hitchhiking under positive selection. Mol Biol Evol 24: 1898–1908.
Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C & Marjoram P (2007) An *Arabidopsis* example of association mapping in structured samples. PLOS Genet 3: e4.

Original articles

I Pyhäjärvi T*, Kujala ST* & Savolainen O (2011) Revisiting protein heterozygosity in plants – nucleotide diversity in allozyme coding genes of conifer *Pinus sylvestris*. Tree Genetics & Genomes 7: 385–397.

II Kujala ST & Savolainen O (2012) Sequence variation patterns along a latitudinal cline in Scots pine (*Pinus sylvestris*): signs of clinal adaptation? Tree Genetics & Genomes 8: 1451–1467.

III Kujala ST*, Knürr T*, Kärkkäinen K, Neale DB, Sillanpää MJ & Savolainen O (2015) Indications of genetic heterogeneity in a locally adaptive clinal trait in *Pinus sylvestris* revealed by novel multipopulation association method. Manuscript.

IV Kujala ST & Savolainen O (2015) Shallow allele frequency clines and lack of allele frequency covariation in timing of bud set associated SNPs in Scots pine (*Pinus sylvestris*). Manuscript.

*authors contributed equally to the manuscript

646. Suurkuukka, Heli (2014) Spatial and temporal variability of freshwater biodiversity in natural and modified forested landscapes

647. Cherevatova, Maria (2014) Electrical conductivity structure of the lithosphere in western Fennoscandia from three-dimensional magnetotelluric data

648. Etula, Henna (2015) Paikkatietoon perustuva reitinoptimointi metsäninventoinnin työkaluna Suomessa : menetelmän kehittäminen ja sen hyödyllisyyden arviointi

649. Romar, Henrik (2015) Biomass gasification and catalytic conversion of synthesis gas : characterisation of cobalt catalysts for Fischer-Tropsch synthesis

650. Shao, Xiuyan (2015) Understanding information systems (IS) security investments in organizations

651. Heponiemi, Anne (2015) Catalytic wet air oxidation of industrial wastewaters : oxidation of bisphenol A over cerium supported metal catalysts

652. Tolkkinen, Mikko (2015) Biodiversity and ecosystem functioning in boreal streams : the effects of anthropogenic disturbances and naturally stressful environments

653. Zoratti, Laura (2015) Effect of environmental, developmental and genetic factors on flavonoid and carotenoid profile of Vaccinium berries

654. Hekkala, Anne-Maarit (2015) Restoration of the naturalness of boreal forests

655. Li, Ying (2015) Users' information systems (IS) security behavior in different contexts

656. Grönroos, Mira (2015) Metacommunity structuring in stream systems : disentangling the roles of regional and local processes

657. Lappalainen, Katja (2015) Modification of native and waste starch by depolymerization and cationization : utilization of modified starch in binding of heavy metal ions from an aqueous solution

658. Kangas, Veli-Matti (2015) Genetic and phenotypic variation of the moose (Alces alces)

659. Prokkola, Hanna (2015) Biodegradation studies of recycled vegetable oils, surface-active agents, and condensing wastewaters

660. Halkola, Eija (2015) Participation in infrastructuring the future school. A nexus analytic inquiry

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 978-952-62-1035-3 (Paperback)
ISBN 978-952-62-1036-0 (PDF)
ISSN 0355-3191 (Print)
ISSN 1796-220X (Online)