OULU 2008
A 506

ACTA UNIVERSITATIS OULUENSIS

Tanja Pyhäjärvi

ROLES OF DEMOGRAPHY AND NATURAL SELECTION IN MOLECULAR EVOLUTION OF TREES, FOCUS ON *PINUS SYLVESTRIS*

FACULTY OF SCIENCE,
DEPARTMENT OF BIOLOGY,
UNIVERSITY OF OULU

ACTA UNIVERSITATIS OULUENSIS
A Scientiae Rerum Naturalium 506

TANJA PYHÄJÄRVI

ROLES OF DEMOGRAPHY AND NATURAL SELECTION IN MOLECULAR EVOLUTION OF TREES, FOCUS ON *PINUS SYLVESTRIS*

Academic dissertation to be presented, with the assent of the Faculty of Science of the University of Oulu, for public defence in Kuusamonsali (Auditorium YB210), Linnanmaa, on April 11th, 2008, at 12 noon

OULUN YLIOPISTO, OULU 2008

Acta Univ. Oul. A 506, 2008

Supervised by
Professor Outi Savolainen
Professor Martin Lascoux

Reviewed by
Research Associate Professor Thomas Bataillon
Professor Craig Primmer

ISBN 978-951-42-8767-1 (Paperback)
ISBN 978-951-42-8768-8 (PDF)
http://herkules.oulu.fi/isbn978951428768 8/
ISSN 0355-3191 (Printed)
ISSN 1796-220X (Online)
http://herkules.oulu.fi/issn03553191/

Cover design
Raimo Ahonen

OULU UNIVERSITY PRESS
OULU 2008

Pyhäjärvi, Tanja, Roles of demography and natural selection in molecular evolution of trees, focus on *Pinus sylvestris*
Faculty of Science, Department of Biology, University of Oulu, P.O.Box 3000, FI-90014 University of Oulu, Finland
Acta Univ. Oul. A 506, 2008
Oulu, Finland

Abstract

Natural selection, mutation, recombination, demographic history and chance all have a role in evolution. In natural populations, the outcome of these forces is seen as adaptations, differences between geographic varieties, and as genetic diversity in populations—both at the phenotypic and molecular levels. In this thesis I wanted to examine the roles of the evolutionary forces shaping molecular genetic diversity in trees, with emphasis on a boreal conifer, Scots pine (*Pinus sylvestris*).

Phylogeographic history and past population size changes have a dominant role in molecular diversity of *P. sylvestris*. The effect of the Last Glacial Maximum (37 000–16 000) was observed in the distribution of mitochondrial DNA variation. In contrast, nuclear DNA was not much affected by the last glacial period. Instead, more ancient demographic events that took place millions of years ago can still be observed in the variation of *P. sylvestris* nuclear DNA.

Not much evidence of positive natural selection was found in pines or trees in general. This is in contrast to strong natural selection that is observed at the phenotypic level. Positive selection is difficult to prove, especially when the genome is still affected by demographic history. Mutation–drift equilibrium may rarely be reached in tree populations.

Keywords: demographic history, molecular evolution, nucleotide diversity, Pinus sylvestris, population genetics, Scots pine, tree

To my grandparents, Aila and Aarre

Acknowledgements

First I'd like to thank Outi Savolainen for opportunity to conduct my doctoral studies under her skilful supervision. I want to thank her and my another supervisor Martin Lascoux for hours of enthusiastic discussion on population genetics, encouraging comments, respect for my opinions and spurring when I otherwise could have settled for less ambitious solutions.

I express my gratitude to reviewers Thomas Bataillon and Craig Primmer for their comments on the thesis and manuscripts. The thesis was financially supported by Research Council for Biosciences and Environment, Finnish Graduate School in Population Genetics and European Commission projects TREESNIPS and EVOLTREE. Torgny Persson from The Forestry Research Institute of Sweden, Leena Yrjänä and Jukka Lehtonen from Finnish Forest Research Institute, Aleksei Fedorkov from Russian Academy of Sciences and Eduardo Notivol from SIA have provided the seed material for the research.

The plant genetics group: Helmi Kuittinen, Johanna Leppälä, Päivi Leinonen, Anne Niittyvuopio, Annaleena Okuloff, Ulla Kemi, Esa Aalto *et al.* have been a joyful company during past five years and I'd like to thank all of them for sharing their views on plant population genetics —especially on Tuesday mornings. My co-pine-troopers Sonja Kujala and Timo Knürr are irreplaceable when one needs understanding for difficulties in pine genetics or a bit of discussion on the essence effective population size. Thanks and roses to Soile Finne and Hannele Parkkinen, they have provided me with most of the molecular data by high quality level lab work.

Pekka Pamilo's group has been a supportive peer group for me. Special thanks to Lumi and Jonna who never get tired to listen how me and my research are doing. I'd like to thank also Anna Palmé, Witold Wachowiak, Matti Salmela, Rosario Garcia-Gil, Päivi Komulainen, Katri Kärkkäinen, Merja Mikkonen and Esa Läärä for fruitful collaboration.

I want to warmly thank my parents, Kirsi and Aaro, and my sister Tiina for their support and encouragement in everything I do. My little brother Tuukka is barely as old as my PhD studies, but much more amusing and loved. My warmest thanks and apologies belong to my fiancé Jukka and ANI the dog. You have to be the most self-possessed family. I know I've been maddening during the last weeks. Love you!

Abbreviations

θ	population mutation rate
μ	mutation rate per base pair
ρ	population recombination rate
bp	base pair
LD	linkage disequilibrium
LGM	last glacial maximum
mtDNA	mitochondrial DNA
MYA	million years ago
N_e	effective population size
PCR	polymerase chain reaction
RFLP	restriction fragment length polymorphism

List of original articles

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

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

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

III Pyhäjärvi T, Kujala ST & Savolainen O What accounts for high protein heterozygosity and low nucleotide diversity in trees? (Manuscript).

IV Savolainen O & Pyhäjärvi T (2007) Genomic diversity in forest trees. Curr Opin Plant Biol 10: 162-167.

V Palmé A, Pyhäjärvi T, Wachowiak W & Savolainen O Selection on nuclear genes along a *Pinus* phylogeny. (Manuscript).

Contents

1 Introduction

Since the publication of Darwin's *On the Origin of Species* (1859), the key question in evolutionary biology has been what are the relative roles of advantageous, neutral and deleterious mutations in evolution (Bernardi 2007, Nei 2005). Even though Darwin's theory accepted the existence of neutral variation, natural selection acting on advantageous mutations (Darwinian selection) had a major role in his original idea. The view that selection is the major force in evolution was later supported by so-called neo-Darwinists in the first decades of the twentieth century (Nei 2005) and for example in the pseudo-hitchhiking model by Gillespie (2000).

A less dominant role of natural selection was already suggested by Thomas Morgan in 1930's (Morgan 1925, Morgan 1932). However, the neutralist view really came forward after 1960's when data on molecular variation started to accumulate. At that time, it was generally believed that most of the genetic polymorphism in a population is a result of some sort of balancing selection (Dobzhansky 1955, Mayr 1963). However, the data revealed so much molecular variation in populations that it could no longer be explained by balancing selection. The most important non-selectionist theory on the genetic variation was developed by Motoo Kimura (1968), who argued that most of the mutations are deleterious or neutral.

The basic idea in Kimura's neutral theory of molecular evolution (Kimura 1983) was that mutation creates new alleles, and drift either fixes or deletes the variation from the finite population just by chance. Deleterious alleles are purged from the populations, and therefore do not contribute to evolution in the long run. Neutral alleles do not have fitness effects and their allele frequencies depend only on effective population size. According to Kimura's neutral theory, the amount of polymorphism in the population depends on the equilibrium between mutation and drift. Further, the divergence between species depends only on the mutation rate. It has to be emphasized that neutral theory did not suggest the absence of positive natural selection as such; only that it is rare compared to neutral and deleterious mutations. The advantage of the neutral theory was that it provided clear theoretical predictions that could be tested with empirical data—and at first, the data seemed to agree with the theory (Nei 2005). Even though it has been recently called "a theory in retreat" (Chamary *et al.* 2006), it is still widely used as a null model in molecular evolutionary biology (Fay *et al.* 2002, Kreitman 1996).

The presence of adaptive genetic variation at the phenotypic level is undeniable (for examples in plant species, see Linhart & Grant 1996). What has proven to be difficult is finding the molecular basis of the adaptations (Caicedo *et al.* 2007, Hamblin *et al.* 2006, Schmid *et al.* 2005), because non-selective forces like mutation, recombination and drift that also shape the molecular variation, easily blur the signal of selection. The challenge is to distinguish the effect of selection from the patterns that non-selective forces have created.

The amount of drift can vary in time and space depending on breeding system, population structure, population size changes and other demographic events. Mutation and recombination rates vary along the genome (Gaut *et al.* 2007, Lercher & Hurst 2002). Usually the demographic history of a study organism is poorly known. Certain forms of selection are hard to find and prove based on molecular data alone (Kreitman & Di Rienzo 2004, Hughes 2007). Furthermore, parts of the genome that have been assumed to be neutral, like synonymous or non-coding positions, have been argued to evolve under selection too (Bernardi 2007, Chamary *et al.* 2006, Kimchi-Sarfaty *et al.* 2007). Despite the growing amounts of molecular data, the extent to which natural selection affect nucleotide polymorphisms at the genomic level is not really known in any species and continues to be at the core of molecular evolutionary science.

1.1 Coalescent theory

Coalescent theory is essential in interpreting DNA sequence polymorphism data in population genetics. In brief, it is a model describing the connection between patterns of observed DNA polymorphism and gene genealogy. For a comprehensive introduction, see Wakeley (2008). The idea was first presented by Kingman (1982) and also independently by Tajima (1983) and Hudson (1983b). The model can be used for example to derive estimates of population parameters (e.g. scaled mutation and recombination rate). It is also useful in describing the effect of chance in evolution. This is important, because DNA polymorphism data from one locus are in principle a random sample from many possible evolutionary scenarios (Rosenberg & Nordborg 2002). Importantly, coalescent simulations together with data can be used in hypothesis testing and in likelihood methods (Rosenberg & Nordborg 2002).

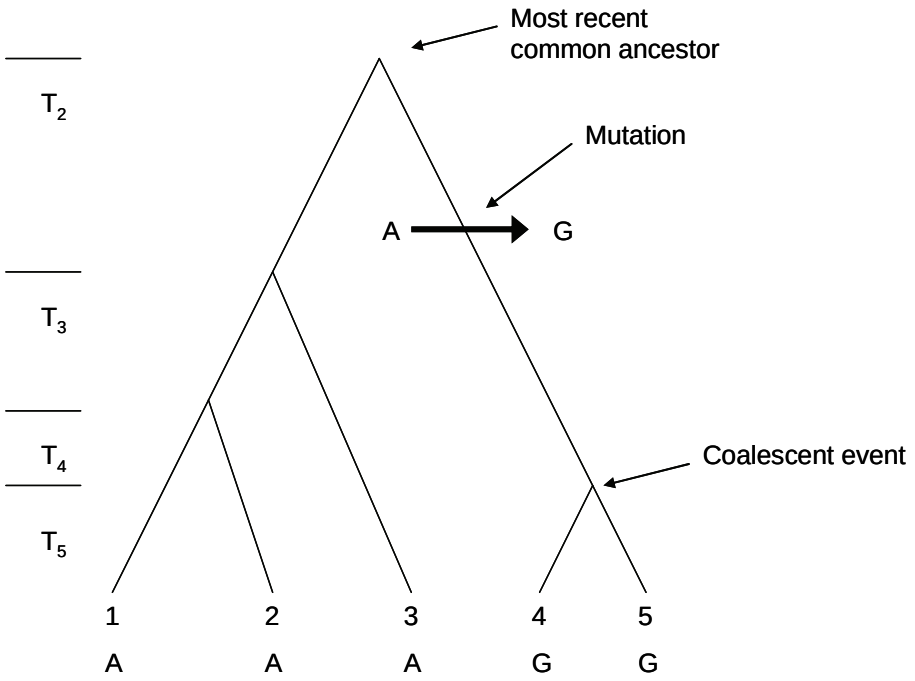

Fig. 1. An example of coalescent tree.

In coalescent model samples are followed back in time to their most recent common ancestor. At the beginning of the process (present) a genealogy of *n* samples has *n* external branches (see Figure 1 for example of tree where *n*=5). The first coalescent event joins together two of them after a time period of T_n, indicating the common ancestry of the samples in the history. After *n*-1 coalescent events, the most recent common ancestor of all samples is reached and no more coalescence events happen. Events in coalescent tree in diploids are scaled in units of $2N_e$ generations. Coalescent process is a Poisson process, where each pair coalesces independently of others with constant rate, and branching structure and coalescent times are independent random variables. N_e is called the *coalescent effective population size* and is equal to N in large Wright-Fisher population (Fisher 1930, Wright 1931). Results on branching structure of tree, expected total length of tree, expected time to most recent common ancestor and their variances are derived based on the above-mentioned Poisson processes.

To model for example how DNA sequence polymorphism emerges, gene genealogy is not enough; mutation also needs to be incorporated. Mutation at the

DNA sequence level is often considered to follow the infinite sites model (ISM) (Kimura 1969, Watterson 1975). In the ISM, each mutation hits one site only once and always creates a new allele, which is usually a reasonable assumption at the intra-specific level. Mutations are distributed stochastically on the coalescent tree according to Poisson distribution, conditional on the lengths of the branches, with rate $\theta/2$, where $\theta=4N_e\mu$ in a diploid case. Provided that only neutral polymorphisms are considered, mutation and coalescence events are independent processes.

The basic coalescent makes several simplifying assumptions about a population: no selection, no population structure, no intra-locus recombination and no population size changes over time. In addition the Wright-Fisher model assumes non-overlapping generations. Even though the assumptions under the basic coalescent model are very unlikely to be met in natural populations, it has proven to be an excellent tool for population geneticists (Kingman 2000, Nordborg 2001, Rosenberg & Nordborg 2002, Wakeley 2008). Many deviations from the basic assumptions above, such as population size changes, structure and selfing can be easily incorporated by rescaling the time by adjusting N_e (reviewed in Nordborg 2001). Recombination (e.g. Hudson 1983a) and selection (e.g. Neuhauser & Krone 1997) have also been included in the coalescent framework.

Further details about how coalescent model and methods based on it have been used in this thesis in testing for neutrality of polymorphism/divergence data and inferring past demographic history are described in Chapter 2.

1.2 Trees from evolutionary viewpoint

Trees are large, long-living woody plants. They do not form a monophyletic group, but the life form has developed independently several times during plant evolution. Trees often dominate their habitat and can even affect the microhabitat of their growing site (Petit & Hampe 2006). They have large census sizes and distributions, consisting of as much as thousands of millions of individuals (Petit & Hampe 2006). However, tree population sizes are not stable. According to pollen fossil abundances, population sizes and distribution ranges, trees have undergone drastic changes several times during climatic cycles (Cheddadi *et al.* 2005, Davis & Shaw 2001).

Most trees are outcrossing and have effective dispersal mechanisms (Petit & Hampe 2006, Savolainen *et al.* 2007). They reach maturity relatively late and stay fertile very long, which results in long generation times. Trees can live hundreds,

even thousands of years. After they reach maturity, fecundity is high and gets higher as trees grow older (Franco & Silvertown 1996). For example a mature *Pinus sylvestris* forest in southern Finland produces about 90 seeds per square meter per year (Sarvas 1962). Effective dispersal and high numbers of individuals have been argued to be main reasons for observed high genetic diversity and low differentiation among tree populations (Hamrick *et al.* 1981, Nybom 2004).

Tree evolution is slow compared to annual and herbaceous species. Speciation and extinction rates are low and many tree species are old (Petit & Hampe 2006). Also the molecular divergence between species and mutation rates per time unit in trees are low compared to other plants (Willyard *et al.* 2007). Generation time effect and low metabolic rate have been suggested to explain the difference between trees and other plants (Petit & Hampe 2006). However, trees can be phenotypically adapted to various environmental conditions and have large genetic differences among geographic regions (Howe *et al.* 2003). For example, populations of several European tree species are adapted to changes in growing season by varying timing of bud set, bud flush and cold tolerance (Savolainen *et al.* 2007). The adaptations can arise relatively quickly: in northern Europe, the locally adapted populations have emerged during less than 10 000 years after the LGM.

1.3 Overview of the genus *Pinus*

There are approximately 111 pine species (Gernandt *et al.* 2005), most of them trees (Richardson & Rundel 1998). Pines are distributed mainly in the Northern hemisphere where they can be found under various demanding environments. Typically they are light demanding and survive in nutrient poor soil better than other trees (Richardson & Rundel 1998). They dominate various forest ecosystems and are also economically important as a source of pulp, timber and other products.

The genus has been divided into two groups, Pinus and Strobus, soft and hard pines and further into four sections (*Trifoliae*, *Pinus*, *Quinquefoliae* and *Parrya*) (Gernandt *et al.* 2005). Pine fossils are abundant, but there is still some uncertainty about the timing of diversification between lineages (Eckert & Hall 2006, Willyard *et al.* 2007). For example, the oldest pine fossil *P. belgica* has been estimated to be about 130 million years old (Alvin 1960, Willyard *et al.* 2007), but the recent estimates of *Pinus-Picea* split still vary between 190-72 MYA, depending on how the phylogeny is calibrated (Willyard *et al.* 2007).

Pines are monoecious, their pollen is wind dispersed and they are practically outcrossing (Richardson & Rundel 1998). They do not have a real incompatibility system, but most self-pollinated individuals die at the embryonic stage (Ledig 1998). As trees in general, also pines are genetically highly variable at allozyme coding loci (Hamrick & Godt 1990, Ledig 1998). Large portions of their genome consist of repetitive DNA and they typically have large gene families and large genomes (Ahuja *et al.* 1994, Friesen *et al.* 2001, Grotkopp *et al.* 2004, Kinlaw & Neale 1997, Valkonen *et al.* 1994). Despite effective gene flow, local adaptation and clinal phenotypic variation in pines is common (Ledig 1998).

1.4 *Pinus sylvestris*

1.4.1 Population geneticist view

P. sylvestris, Scots pine, is one of the most widely distributed pine species in the world. It can be found from Spain to northern Finland and from Scotland to eastern Siberia. The distribution is mostly continuous, though some marginal isolated populations exist, for example in Spain and Turkey. It has a huge population census size, which only in Finland consists of tens of milliards of trees. *P. sylvestris* is highly outcrossing. The proportion of progeny due to random outcrossing, t_m has been estimated to be around 0.95 (Kärkkäinen *et al.* 1996, Muona & Harju 1989). The pollen flow is also extensive. For example, Robledo-Arnuncio & Gil (2005) estimated that that 4.3% of fertilizing pollen in an isolated Spanish population came from more than 30 km distance. Presumably due to effective pollen flow, the species is not genetically highly structured (Gullberg *et al.* 1985, Karhu *et al.* 1996, Muona & Harju 1989, Szmidt & Muona 1985). Even populations that are separated by thousands of kilometres show little genetic differentiation at neutral markers (Dvornyk *et al.* 2002). As a conclusion, it can often be assumed that *P. sylvestris* has a large, panmictic population.

P. sylvestris has several useful characteristics from a population genetics viewpoint. The haploid maternal megagametophyte tissue in seeds is practical, because haplotypic phase can be acquired easily. For example recombination rate and extent of LD can be measured reliably using data from direct DNA sequencing of megagametophyte tissue. The different modes of inheritance of mitochondrial (maternal), chloroplast (paternal) (Neale & Sederoff 1989 and citations therein) and nuclear (biparental) DNA can be exploited to differentiate

between pollen and seed mediated components of gene flow or in inferring population history at several time-scales.

Despite low differentiation at neutral markers, *P. sylvestris* is clearly phenotypically adapted to different environmental conditions. Bud set timing and cold tolerance varies clinally with latitude in Europe (García-Gil *et al.* 2003, Hurme *et al.* 1997, Mikola 1982). Common garden experiments have shown that populations are locally adapted (Savolainen *et al.* 2007) and proved that strong natural selection has affected genetic variation in these traits (Knürr *et al.* manuscript). The combination of low differentiation at neutral markers, and strong natural selection at phenotypic trait makes finding the molecular basis of selection easier. Low genetic differentiation is useful because population structure can cause false positives in association analysis (Aranzana *et al.* 2005, Neale & Savolainen 2004). However, the large genome and gene families make finding the loci causing adaptive phenotypic variation a demanding task.

One advantage of *P. sylvestris* as a study species is the amount of background data available. Database search in ISI Web of Science with terms “Pinus sylvestris OR Scots pine” produced 7055 scientific articles. The extensive provenance trials as well as first studies of local adaptation and flowering phenology have been largely motivated by forestry, but nonetheless benefit also evolutionary studies on the species (Eiche 1966, Langlet 1936, Mikola 1982, Morgenstern 1996, Sarvas 1962, Shutyaev & Giertych 1998).

1.4.2 Phylogeographic history

According to pollen fossil data, Pinus abundance in Europe has fluctuated over past hundreds of thousands of years in conjunction with climate cycles (Cheddadi *et al.* 2005, Müller *et al.* 2003). During the LGM (37 000 - 16 000 years ago), Fennoscandia and western parts of northern Russia were covered by ice. The region was colonized by *P. sylvestris* as the ice sheet retreated during the past 10 000 years. Suggestions about phylogeographic history of *P. sylvestris* during and after the LGM have been made based on paleoecological and genetic evidence.

The exact colonization routes of *P. sylvestris* are not known, but based on pollen fossil, *P. sylvestris* reached the northernmost part of Scandinavia about 7800 BP (Huntley & Birks 1983, Willis *et al.* 1998). Later, about 5000 BP, the northern limit of the distribution retreated southwards (Willis *et al.* 1998). The earlier view has been that tree refugia were located only in southern Europe, but

more recent macrofossil evidence suggests that there have been forests in central and eastern Europe even the during the LGM (Cheddadi *et al.* 2006, Willis & van Andel 2004). The pollen data from Russia indicates also eastern refugia in Siberia (Kremenetski *et al.* 1998). The history of *P. sylvestris* prior to the LGM is not clear, because older pollen data are scarce. At present, the northern forest line of *P. sylvestris* in Finland is at the latitude 67°30'N (Seppänen & Norokorpi 1998), and the species is colonizing more northern areas, probably due to climate warming (Neuvonen *et al.*, manuscript).

Studies on distribution of genetic diversity and especially mitochondrial DNA types have revealed some isolated populations and putative locations of refugia. Spanish populations seem to be isolated from the mainland *P. sylvestris* populations and have been suggested to be remnants of LGM refugial population in Iberian Peninsula (Dvornyk *et al.* 2002, Sinclair *et al.* 1999, Soranzo *et al.* 2000). There is also genetic evidence of an Italian refugium (Cheddadi *et al.* 2006). In addition, a distinct, more northern origin of Scottish *P. sylvestris* population has been suggested based on genetic data (Sinclair *et al.* 1999, Sinclair *et al.* 1998).

1.5 Aims of the study

Because trees are both economically and ecologically important group of plants, it is important to understand how the natural selection has affected their evolution. This helps to understand and predict how and at what pace the trees would respond to environmental changes in the future. The aim of the study is to assess the relative roles of selection and demography on tree genomes at various time scales with special focus on the demographic history of *P. sylvestris*.

In *P. sylvestris* the effect of natural selection is clear at the phenotypic level. However, the effect of natural selection at the molecular level is difficult to identify without knowledge about the species demographic history. I wanted to examine the demographic history of *P. sylvestris* and especially study how post-glacial colonization of northern Europe has affected the molecular variation in *P. sylvestris*. I also wanted to study how differences in the histories of natural populations can be observed in their genomic variation. A goal is also to get better estimates of scaled mutation and recombination rates and how they vary in the *P. sylvestris* genome, because they are essential in understanding the molecular evolution at the genomic level.

Another, inter-species level approach is taken to assess the effect of selection in different pine lineages and generally in trees. The phylogenetic analysis is used to compare frequencies of positive and negative selection in genus of pines at time scale extending tens of millions of years from present. Meta-analysis of studies on nucleotide variation in trees is used to evaluate what specific features arise from tree-like life form compared to other plants.

2 Material and methods

Only a brief outline of the methods is presented in this chapter. Details about sampling, and molecular and statistical methods are given in the original papers (I-V).

2.1 Sampling

Generally, samples from natural populations were preferred in order to minimize possible anthropogenic effects on molecular diversity of *P. sylvestris*. Seed orchard samples were used in five cases in paper I, the rest of the samples are from natural populations. Due to the wide geographical distribution of *P. sylvestris*, the sampling of the whole distribution and especially the most eastern part of it was demanding. Therefore most of the results are based on the European part of the *P. sylvestris* distribution.

The study on post-glacial colonization routes is based on 37 populations and 714 individual trees. 3-35 trees per population were sampled. Sampling concentrated on northern European populations that were not well covered in previous studies on *P. sylvestris* post-glacial colonization history (Sinclair *et al.* 1999, Soranzo *et al.* 2000). In a study on nucleotide diversity of *P. sylvestris* (paper II), sampling was designed to equally cover different latitudes of the species' distribution in Europe. Eight populations and five individuals per population were sampled. The populations were divided into four groups: Spanish and Turkish populations represented isolated high-altitude populations. Swedish and two Finnish populations represented populations that have colonized the common location after the LGM. Polish, Austrian and French populations represented populations with putatively more stable population history than that of the northern ones.

In the study on nucleotide diversity of allozyme loci (paper III), recombination rate was of special interest. One large sample (35 individuals) from a single southern Finnish population was used to get a reliable estimate of population recombination rate. In addition, 18 northern Swedish trees with known allozyme heterozygosities were used to deduce nucleotide polymorphisms that cause the distinct electromorphs.

2.2 Molecular methods

In studies on nucleotide diversity, DNA was extracted from megagametophyte tissue. In paper I, needles and embryos were also used for DNA extraction. This should not affect the results, because they all share the same mtDNA type.

2.2.1 Mitochondrial DNA variation

DNA variation in *P. sylvestris* mitochondria is low (Soranzo 1999). Only one known variable position, *nad*1 in *P. sylvestris* mtDNA had been found before paper I (Soranzo *et al.* 2000). The primary goal in the study I was therefore to find nucleotide polymorphism in *P. sylvestris* mtDNA. Among eleven mtDNA regions known to amplify in conifers, five were amplifying and sequenced in *P. sylvestris*.

The five mitochondrial regions were sequenced to find polymorphisms in screening set of 12 individuals of *P. sylvestris*. No single point mutation was observed, but two new variable insertion-deletion positions were found in *nad*7. Previously known polymorphism in *nad*1 (Soranzo *et al.* 2000) was also segregating in the sample. Because no additional polymorphism in these regions were found in a set of 90 individuals, the rest of the 714 individuals were genotyped for insertion-deletion polymorphisms in *nad*1 and *nad*7 by PCR or PCR-RFLP, and electrophoresis.

2.2.2 Sequencing

Papers II-V were based on sequence polymorphism in *P. sylvestris* and other tree species. Sequence data are excellent for studies of molecular evolution for several reasons. Single point mutations can be identified, mutations causing amino-acid replacements can be separated from those that do not (silent mutations) and population mutation and recombination rates can be estimated at nucleotide level. In addition, a wealth of applicable theoretical framework and statistical and computational methods for nucleotide polymorphism data are available.

Primers for amplifying nuclear genes in *P. sylvestris* were designed based on other conifer sequences in the NCBI GenBank, mostly expressed sequence tags, (ESTs) of *P. taeda* and *P. pinaster*. If no information about the intron positions was available for conifers, homologous regions of *Arabidopsis thaliana* were used to predict the positions. These positions were avoided in primer design. *P.*

sylvestris genome is known to contain a large number of gene families. Therefore long, 22–25 nucleotide primers were preferred to avoid amplifying several paralogs.

The nucleotide diversity data were acquired by direct sequencing of the PCR products. Sequencing was done in both 5' and 3' directions. Sequenced stretches were assembled into contigs individual by individual. After aligning, all chromatogram positions of polymorphic sites were verified by eye. For paper V, some of the sequence data from EST databases of conifers were used in the analyses.

Only haploid DNA from megagametophyte was used in sequencing. If overlapping peaks were observed in chromatograms, it was generally considered as amplifying several loci and the data was discarded. In few cases, secondary structures were observed as anomalies in chromatogram. In these cases sequence has been included in the analysis for those parts that are not affected by secondary structures.

2.2.3 Allozymes

In paper III, nucleotide diversity in allozyme loci was studied. Six allozyme loci were amplified and sequenced: *6pgd*, *aco*, *got*, *gdh* and two *mdh*:s. The corresponding five allozyme systems: 6PGD, ACO, ASP (GOT), GDH and MDH had been used earlier in several studies on molecular variation of *P. sylvestris* (e.g. Muona & Harju 1989).

Several allozyme systems have more than a single locus. To verify which allozyme locus corresponds to each sequenced locus, a set of trees heterozygous for these five systems (Szmidt & Muona 1989) was used. These 18 trees were studied for both nucleotide and allozyme diversity. The replacement polymorphisms that fulfilled the following criteria were considered as causative replacements behind allozyme polymorphism: 1) segregation associating with allozyme polymorphism 2) an amino acid charge change consistent with electrophoretic pattern and 3) results in similar expected heterozygosities in a population sample of 35 trees from southern Finland compared to observed earlier in nearby populations (Muona *et al.* 1988). The studied loci were, specifically: *6pgdB*, *aco*, *gotC*, *gdh*, *mdhA* and *mdhB.*

2.3 Statistical methods

Most of the statistical methods applied to sequence data in the thesis have been developed within the framework of coalescence theory (see Introduction).

2.3.1 Molecular diversity

Standard measure of molecular diversity in allozymes is expected heterozygosity H_e and for mtDNA haplotypes its haploid counterpart, gene diversity $H_{(m)}$ (Nei 1987). They are defined as a probability that two randomly sampled alleles are different. To describe nucleotide diversity at sequence level, various estimates of θ (=$4N_e\mu$), are used. Two most common are Watterson's (1975) θ_W based on the number of segregating sites, and θ_π (Tajima 1983) which is based on pairwise nucleotide diversity.

In paper II, a new Bayesian method for estimating θ from multilocus sequence data is presented. The method was used to get an unbiased estimate of genome wide level of nucleotide diversity. Simple average over multiple loci would have resulted in an upward bias. This method assumes standard equilibrium population, which may bias the estimate as well. The assumption of no recombination should not affect point estimate of θ, but widens its posterior probability.

2.3.2 Genetic and geographical structure

In papers I and II, several *P. sylvestris* populations were sampled, which enabled examining of geographical distribution of genetic diversity. Overall genetic clustering of nucleotide diversity was studied with Bayesian method, implemented in BAPS software (Corander *et al.* 2003), which partitions samples into groups based on allele frequency data. Distribution of genetic variation within and among populations was described with several F_{ST} estimates. The pattern of isolation by distance in mtDNA was tested by comparing genetic differentiation and geographical distance between populations (Mantel test), because according the IBD model, there should be positive correlation between the two (Wright 1943). Phylogeographic structure of mtDNA data was inspected by comparing two F_{ST} estimates, G_{ST} and N_{ST} (Pons & Petit 1996). The two estimates differ because, G_{ST} is based only on allele frequencies, but N_{ST} also takes into account the distance between haplotypes. If N_{ST} is higher than G_{ST}, it

indicates that alleles inside a population are more closely related than alleles compared among populations. For sequence data, F_{ST} (Hudson *et al.* 1992) estimate analogous to Weir and Cockerham's θ (not the above mentionned population mutation parameter but an actual method of moment estimator of the parameter F_{ST}) (1984) was used to describe pairwise genetic differentiation over all loci. Two statistics based on the number of differences between nucleotide haplotypes, Snn (Hudson 2000) and K_{ST}* (Hudson *et al.* 1992), were used to describe differentiation at each locus.

Genetic diversity may decrease during sequential bottlenecks or in isolated populations due to smaller N_e. Therefore, lower diversity in northern European populations and also in isolated Mediterranean populations was expected. In addition to basic F-statistics based analysis of geographic distribution, the amount of nucleotide variation in different populations was compared (paper II). This was done by comparing the credibility intervals of each geographic group's posterior distributions of θ.

2.3.3 Linkage disequilibrium and recombination

Non-random association between alleles, LD, is dependent on the recombination rate between adjacent sites *c*, and N_e. It is expected to decay along the distance between observed alleles. The squared correlation of allele frequencies of two parsimony informative sites r^2, and the distance between them was plotted to describe the decay of LD. Several methods were used to estimate the so called population recombination rate ρ ($=4N_e c$), based on nucleotide diversity data (papers II and III). In paper II, the data from each individual locus was scarce, so recombination rate estimates are based on the pooled multilocus data. A nonlinear least-squares estimate of ρ, was obtained by using expected relationship between the distance and r^2

$$E(r^2) = \left[\frac{10+\rho d}{(2+\rho d)(11+\rho d)}\right]\left[1+\frac{(3+\rho d)(12+12\rho d+\rho^2 d^2)}{n(2+\rho d)(11+\rho d)}\right]$$

(Hill & Robertson 1968). The maximum-composite likelihood method of Hudson (2001) based on two-site sample configurations to estimate ρ was also applied to pooled data. Paper III was especially designed to get reliable estimates of ρ and the composite likelihood method (McVean *et al.* 2002, Hudson 2001) was applied to each locus separately.

2.3.4 Testing of demographic models

Coalescent simulations were used to find a demographic history compatible with the observed nucleotide diversity in *P. sylvestris*. The simulated dataset had the same sample size and number of loci as the observed one. Simulations were conducted with Hudson's ms program (2002) assuming same mutation and recombination rate per site in all loci. The basic idea behind the approach is to simulate data that have the same average nucleotide diversity (θ_π) as in the 16 loci in the observed data and use other summary statistics, such as Tajima's D, and Fay and Wu's H to reject a model. For each summary statistics, the observed value was compared to a distribution from simulated dataset through a p-value. Standard neutral model and a grid of 16 bottlenecks with different timings and severities were simulated.

2.3.5 Neutrality tests

The deviation from neutral assumptions was examined based on intraspecific nucleotide diversity data in papers II and III. Many of the tests are designed to detect deviations from the standard neutral equilibrium model that assumes constant population size and absence of natural selection. In many cases, it is difficult to judge whether the deviation from neutrality is caused by natural selection or by some other violation of the neutral model.

Allele frequency spectrum based tests like Tajima's D (Tajima 1989) and Fay and Wu's H (2000) use difference between θ estimates as a test statistic. Under the standard neutral model, they should yield similar estimates. Differences between them are considered as a sign of deviation from neutrality. Tajima's D statistics is the difference between θ_π and θ_W. Fay and Wu's H is the difference between θ_π and θ_H. The latter is based on the number of derived alleles.

Hudson-Kreitman-Aguadé (HKA) –test (Hudson *et al.* 1987) and McDonald-Kreitman (MK) –test (McDonald & Kreitman 1991) compare polymorphism segregating inside species to divergence between species. Jody Hey's multilocus HKA-test detects deviations from neutrality both in the amount of diversity and divergence based on neutral coalescent simulations. In neutral situations, both divergence and diversity should be functions of mutation rate. For example balancing selection could lead to an excess diversity compared to divergence. In MK test, segregating sites are divided into four classes, replacements and silent polymorphisms, and polymorphic (inside a species) and fixed differences

(between species). If the ratio of two replacement classes is not what is expected based on silent changes, it is considered as deviation from neutrality.

Haplotype-based tests of Innan (2005) and Hudson (1994) were applied to data in paper III, where groups of closely related haplotypes were observed at several loci. Innan's haplotype configuration test (HCT) uses coalescent simulations to evaluate how unlikely the observed haplotype frequency vector is. Hudson Haplotype Test (HHT) was used to test specific haplotype groups by estimating the probability of finding a haplotype subset of size *i* that contains *j* or fewer segregating sites. The advantage of latter is that it takes into account the distance between haplotypes. In addition to standard neutral model, different non-equilibrium population histories can also be used to create a null distribution.

There are some tests that are not as sensitive to non-selective deviations from neutrality as the above-mentioned tests. A method developed by Beaumont and Nichols (1996) searches for F_{ST} outliers to detect loci that are highly differentiated and could therefore underlie the adaptive traits between populations. The test also takes into account the interdependency between heterozygosity and F_{ST}. DHEW, and other compound test statistics introduced in Zeng *et al.* (2007), combine Tajima's D, Fay and Wu's H and Ewens-Watterson (Watterson 1978) test statistics so that the effect of demographic events is diminished and the test explicitly detects deviations from neutrality caused by positive selection.

2.3.6 Phylogenetic methods

In paper V, data from several loci and conifer species were used to construct a species phylogeny and to study the effect of natural selection at the genus level. Neighbour-joining, maximum-likelihood and Bayesian methods were used to construct a species tree from concatenated dataset. Supertree analysis was used to construct a species tree from individual gene trees.

Effects of selection were examined with several tests. Neutral substitution and relative rate ratio (RRR) tests (Creevey & McInerney 2002) are based on the ratios of different classes of replacement and silent variable sites. They are similar to MK test, applied to nucleotide sequence data from several species. They can detect selection in different branches in phylogenetic trees.

Codon-based models (Goldman & Yang 1994) and maximum-likelihood method were used to examine what model of evolution best explains the data and to study the possible effects of positive natural selection. The method is focusing on the ratio of non-synonymous and synonymous divergence, d_N/d_S. Branch

models can detect differences in d_N/d_S ratios among lineages and site models can detect differences in d_N/d_S among codons. Since d_N/d_S should be around one in neutral situation, values significantly higher than one are considered to be result of positive selection.

Since a considerable number of lineages were tested in RRR, neutral, substitution test and in codon-based methods, a false discovery rate (FDR) analysis (Benjamini & Hochberg 1995) was applied to the results to correct for multiple testing. The method gives for each p-value a corresponding q-value which is the minimum false discovery rate when that particular test is considered significant. FDR can also be used to estimate the proportion of true nulls among all cases (Storey & Tibshirani 2003). With this approach, the proportion of loci and lineages under selection can be estimated, even if the signal from individual case is not strong enough to result in a statistically significant result.

3 Results and discussion

3.1 Geographical distribution of molecular diversity in *Pinus sylvestris*

Overall, there was a considerable difference in the geographic distribution and amount of molecular variation between mitochondrial and nuclear DNA. This was not surprising, because these genomes have different modes of inheritance. Nuclear DNA is biparentally inherited and is therefore dispersed by both pollen and seed. Mitochondrial DNA is maternally inherited and only dispersed via seeds. It also has half the N_e of nuclear DNA. Patterns of post-glacial colonization are clear from mtDNA, because the alleles do not spread as fast as in markers that have pollen mediated gene flow. The drawback of using mitochondrial markers is that the whole mitochondrial genome is effectively only one locus, since there is no recombination. Therefore drift can have a strong impact on the distribution of alleles and the results based only on mitochondrial data should be taken cautiously (Pakendorf & Stoneking 2005).

In paper I, altogether four mitochondrial haplotypes were found. Two new polymorphic indels were found from intron 1 of *nad*7. Together with previously detected polymorphisms in *nad*1 (Soranzo *et al.* 2000), they constitute four mitochondrial haplotypes *a*, *b*, *c* and *d* (see Figure 1 in paper I). Turkish and Spanish populations had haplotypes that were not observed in any other studied population, which was not surprising due to their geographical isolation from main distribution. Haplotype *b* was only found in Spain and haplotype *d* in Turkey. The isolation of Spanish populations was observed already in earlier studies (Cheddadi *et al.* 2006, Sinclair *et al.* 1999, Soranzo *et al.* 2000), but the haplotype *d* in Turkey had not been observed previously. In addition, there was a new haplotype *c* that was found only in northern and Central Europe and in western Russia.

In contrast to differences in the genetic diversity among populations in mtDNA, the level of nucleotide diversity at nuclear loci was similar in all four geographic groups, about 0.005/bp (see Figure 2 in paper 2). It was at the same level as observed in earlier study on *pal1* and a small dataset of 10 loci (Dvornyk *et al.* 2002). Due to sequential bottlenecks during northward post-glacial colonization, northern groups could have less diversity than central. However, this was not the case. Either the bottlenecks have not been severe (Austerlitz *et al.*

2000) or abundant gene flow has equalized the distribution of genetic diversity quickly. An option is that admixture has increased the variation in the north and counterbalanced the effect of colonization.

The overall G_{ST} estimate in mitochondrial DNA was 0.655, suggesting strong differentiation between populations. In addition, there was significant pattern of isolation by distance (p-value: 0.0012), which means that the genetic differentiation in mitochondria increases as the geographical distance between populations increases. There seemed to be a phylogeographic pattern even though the G_{ST}-N_{ST} comparison did not have power to indicate it. In contrast, based on data from nuclear genes, all European *P. sylvestris* populations belong to a single genetic cluster and have pairwise average F_{ST} values between 0 and 0.14. There was some indication that Spanish population is slightly differentiated from others as was expected based on distribution of mtDNA haplotypes and earlier study on nucleotide diversity in *P. sylvestris* (Dvornyk *et al.* 2002). On the other hand, Turkish population did not show any differentiation at nuclear genes, even though based on mtDNA and species distribution it is isolated from the main distribution.

Despite the overall uniformity of genetic variation of nuclear DNA, there were some differences between geographic groups. Tajima's D in silent sites had positive values in southern populations and negative ones in central and northern European group. The other (although not statistically significant) difference was in the amount of LD between central and northern group. The northern group had only one fifth of the ρ compared to southern group. This could be caused by bottleneck, admixture or differences in outcrossing rate. Selfing can increase the amount of LD, because the recombinants coalesce faster back together in populations that have more selfing than outcrossing population (Nordborg 2000). Kärkkäinen *et al.* (1996) found that the outcrossing rate was higher in southern (0.99) than in northern Finland (0.93). Coalescence simulations could be used to study whether such subtle differences could cause the difference in the amount of LD. As Nordborg (2000) pointed out, differences in outcrossing rate can easily be obscured by demographic events. Demographic hypotheses for increased LD in the north group are discussed in the following section.

3.2 Phylogeographic and demographic history of *Pinus sylvestris*

Inference about possible post-LGM colonization scenarios were mainly based on the distribution of mtDNA variation among European populations. Since haplotypes found in Turkey and Spain were not found elsewhere in Europe, it is

likely that these populations did not contribute considerably to the northward colonization. Spanish populations seem to be the most differentiated from others according to both mitochondrial and nuclear data. This implies that the Pyrenees has been a major barrier to gene flow.

The existence of haplotype *c* in central, eastern and northern Europe and its absence in Mediterranean and most eastern part of distribution reinforced the view that the origin of populations that inhabit northern Europe at present is not in Mediterranean region. The origin of haplotype *c* could be in central Europe where according to fossil evidence have been forests even during the coldest period of last glacial, or in eastern part of Eurasia, which was partly ice free during the LGM.

Results of paper I together with Soranzo *et al.* (2000), Cheddadi *et al.* (2006) and Naydenov *et al.* (2007) create a picture of post-glacial migration patterns of *P. sylvestris* that differs from many other European trees (Petit *et al.* 2003). In most temperate European forest trees, the refugia have been in Mediterranean area and these populations have colonized the rest of the Europe after the LGM. In *P. sylvestris*, there have been refugia in Mediterranean area, but these populations have not contributed much to colonization of other regions of Europe. Rather, there have been refugia somewhere else, putatively closer to the edge of the ice sheet, maybe in central Europe or somewhere west from Ural Mountains. Our results resemble those found in two other boreal trees, *Picea abies* and *Betula* (Palme *et al.* 2003, Sperisen *et al.* 1998, Vendramin *et al.* 2000). It seems that unlike the more temperate species studied in Petit *et al.* (2003), cold tolerant trees could have survived outside Mediterranean region during the LGM.

It has been suggested that there is a general suture zone in northern Sweden, where two post glacial colonization fronts, one from south and one from east, have met (Taberlet *et al.* 1998). Sinclair *et al.* (1999) have found evidence of this suture zone also in *P. sylvestris*. There is indeed a difference also in the frequency of haplotype *c* between northern Finland and Sweden, but it is not striking. Based on haplotype frequencies, it is possible that both Finland and Scandinavia have been colonized from south and the differences in haplotype frequencies are due to drift that has driven allele frequencies apart in the two regions.

In silver birch (*Betula pendula*) admixture in Finnish populations was suggested to explain the observed dimorphism in nucleotide diversity (Järvinen *et al.* 2003). Higher LD in the north group of *P. sylvestris* (paper II) compared to the central groups could result from admixture in the suture zone. However, the estimates of LD are uncertain as discussed in the next section and it has to be born

in mind that neutral coalescent process can also easily lead to dimorphism due to deep internal branches. All in all, there are several independent pieces evidence of admixture of *P. sylvestris* in Scandinavia, but the issue needs further investigation.

Nuclear data did not reveal large differences between the northern and central groups, indicating that events after the LGM have not much affected the monitored summary statistics, Tajima's D and Fay and Wu's H. The southern populations had positive values of Tajima's D in contrast to negative in the north and central group, but the data were too scarce to examine in detail what demographic history could be responsible. The northern and central groups were probably similar, because from the population genetic timescales, the last post glacial events happened during the past 500 generations or so. In coalescent time scale, colonization and other post-glacial events took place in very recent history. Most of the coalescent events of the samples and thus majority of the tree predate the LGM and therefore, no signal of colonization event can be observed (collecting and scattering phase, (Wakeley & Aliacar 2001)).

A detailed analysis of demographic history using coalescent simulations was applied to the northern and central groups. Based on nuclear data, the standard neutral model was rejected in both groups. This led us to evaluate how likely the data is produced by a few non-standard demographic scenarios. To keep the amount of parameters low, only simple bottleneck models with varying timing and severity were examined. Among 16 bottleneck scenarios simulated, only one was compatible with the data. It was a bottleneck 0.1 X $4N_0$ generations ago that reduced the population size to 1% of the present size for 0.006 X $4N_0$ generations, when N_0 is N_e at present.

Using the divergence between *Picea abies* and *P. sylvestris* and the nucleotide diversity in *P. sylvestris*, a mutation rate of 12.1 X 10^{-9} per generation and N_0 of 238 000 individuals were estimated. Based on these estimates, the bottleneck would have happened 2 MYA. Even though the timing should be considered preliminary due to uncertainties in the underlying assumptions, it shows that really ancient demographic events predating the LGM can still have an effect on nucleotide diversity in *P. sylvestris*.

The problem in the approach we chose to estimate the demographic history in paper II is the choice of scenarios. Only a limited number of scenarios can be tested by coalescent simulations and if the true or closely related scenarios are not included, the interpretations can be misleading. The parameter space for demographic histories is infinitely large and cannot be examined thoroughly with the coalescent approach. Therefore, independent data, like pollen fossils should

be used to make hypotheses and interpretations that are biologically realistic. For example, repeated cycles of bottleneck modelling glacial cycles as in Jesus *et al.* (2006) may be a realistic hypothesis for European forest species that have long generation times and have existed in Europe for hundreds of thousands of years.

Ideally, when the correct demographic scenario is known, it can be used as a null model in further analysis e.g. to detect the loci that have been affected by selection. We did this in paper III where we used the results on demographic history of *P. sylvestris* from paper II as null model in Hudson's haplotype test. However, if the null model is not correct, it can result in spurious interpretations about the amount selection.

3.3 Recombination, mutation and selection in the genome of *Pinus sylvestris*

Two nucleotide diversity datasets used in this study (paper II and III) were different in several aspects. In paper II a set of 16 loci was sequenced using 40 samples from eight different European populations. The studied fragments were on average 850 bp long partial stretches of genes, consisting mostly of coding sequence. Paper II included both candidate genes for phenological traits and so-called reference loci that *a priori* are not related to adaptively important traits. The dataset in paper III consisted of nearly full genes sequenced from a population sample of 35 southern Finnish trees and included several large introns. All genes in paper III code for allozymes and the goal of paper was to solve why trees are most variable plant species at allozyme level and at the same time harbour moderate levels of nucleotide diversity. Allozymes are usually enzymes with important housekeeping functions and are involved e.g. in sugar and nitrogen metabolism. Due to several differences it is difficult to conclude, whether the differences in their patterns of nucleotide diversity are due to mutation, recombination, selection, chance or some combination of these evolutionary forces.

According to our results, mutation rate in *P. sylvestris* varies from loci to loci. There is correlation between synonymous nucleotide diversity and synonymous divergence between *P. pinaster* and *P. sylvestris* as expected under neutrality when there is variation in mutation rate among loci (Figure 2a in paper III). However, the two sets of loci do not differ from other genes in respect to diversity or divergence. At synonymous sites, the nucleotide diversity at allozyme coding genes was slightly, but not significantly higher than in other loci. Also the

divergence from *P. pinaster* at synonymous sites was similar in the two sets of loci (Figure 2 in paper III).

Under neutrality, when mutation rates vary along the genome, a positive correlation between nucleotide diversity and allozyme heterozygosity is expected. In our limited sample, there was negative non-significant correlation between nucleotide diversity at allozyme coding loci and allozyme heterozygosity. Similar observation was made by Skibinski & Ward (2004) who studied the relationship between allozyme heterozygosity and DNA sequence substitution rate in a human-mouse comparison. They concluded that allozyme heterozygosity was more related to the amount of purifying selection than to the mutation rate. Purifying selection was clearly acting on allozyme coding genes of *P. sylvestris* too. Average K_A/K_S ratio was well below one in all loci and loci seemed to be very conservative at the amino acid level. There was also positive, but non-significant correlation between K_A/K_S and allozyme heterozygosity, suggesting that selection on non-synonymous sites might be affecting the heterozygosity in these genes.

In paper II, ρ was estimated to be 0.0064 in northern group using combined data from nine loci. In paper III, all ρ estimates were below that. The complete lack of recombination that we found in six of eight regions of allozyme coding genes was surprising. Even the longest continuous fragment, *gdh*, had ρ zero for the whole gene. This is contradictory to rapid decay of LD observed in paper II as well as in other conifers, *P. taeda* (Brown *et al.* 2004, González-Martínez *et al.* 2006), *Picea abies* (Heuertz *et al.* 2006) and *Pseudotsuga menziesii* (Krutovsky & Neale 2005). In yeast, the essential housekeeping genes are clustered in regions of low recombination (Pal & Hurst 2003) and there is similar indication in *Escherichia coli* also (Boyd *et al.* 1994). It is possible that allozyme coding genes of *P. sylvestris* could reside in genomic regions where the recombination rate is reduced.

Actually, based on genetic maps' sizes and physical genome sizes, there should be regions of lowered recombination in the *P. sylvestris* genome. The genetic map of *P. sylvestris* is about 1500 cM (Komulainen *et al.* 2003) and genome size has been estimated to be 28 pg (≈ 27 Gb) (Valkonen *et al.* 1994). As a comparison, *A. thaliana* map size is approximately 500 cM (http://www.arabidopsis.org/) and genome size 0.16 pg (≈ 0.157 Gb) (Bennett *et al.* 2003). The difference in centimorgans is three-fold, but in base pairs, 170-fold. Because cM is a unit of recombination this suggests that *P. sylvestris* has much less physical recombination (*c*) per generation than *A. thaliana.*

The recombination rate in allozyme coding genes could be partly an artefact of the methods. The recombination rate estimates (ρ = 0.03 in Central Europe) in paper II were based on Hudson's maximum composite likelihood method (2001) applied on pooled data from several loci, because populations did not have much information per locus. The pooled data that included loci that had some recombination might cause bias in the overall recombination. The allozyme data had larger sample sizes and longer stretches of sequence resulting in more reliable recombination estimates that could be obtained for each gene separately.

Gene conversion could result in discrepancies in recombination rate estimates (Andolfatto & Nordborg 1998). Since the gene conversion events involve on average only a few hundreds of base pairs (Frisse *et al.* 2001, Plagnol *et al.* 2006), it could have more strongly affected dataset in paper II, that consists of shorter fragments. In longer genes, gene conversion would not affect the results as much, because the sites that are more distant from each other would not indicate recombination. However, not much is known about gene conversion in conifers.

Positive selection and selective sweeps may also increase the level of LD (McVean 2007) as in *A. thaliana fri* region (Toomajian *et al.* 2006). This is unlikely the case in the allozyme genes studied, since we did not find strong evidence of positive directional selection in these regions. In addition, the nucleotide diversity was slightly higher than that observed in other genes, which is opposite to what is expected after complete selective sweeps (Maynard Smith & Haigh 1974).

There is one more possible explanation for the difference observed between the two datasets. The HHT test indicated non-neutral haplotype clustering in several allozyme coding genes in *P. sylvestris*, observed as star shaped gene phylogenies in neighbour-joining trees. There were clusters of haplotypes that had less variation inside them than expected under neutrality. Hudson *et al.* (1994) suggested that this kind of pattern could be a result of balancing selection that has raised the frequency of the new haplotype into intermediate level so recently that new mutations have not yet accumulated into the gene. This "partial selective sweep" could also result in the observed pattern in *P. sylvestris* allozyme coding genes. The method was originally developed for nucleotide diversity data in a allozyme coding gene *Sod* in *Drosophila melanogaster* (Hudson *et al.* 1994). Several aspects in the pattern of nucleotide variation in *Sod* gene are similar to what was observed in allozyme coding genes of *P. sylvestris*: non-significant Tajima's D, diversity and divergence levels compatible with neutrality, low recombination and a group with too little variation compared to neutrality.

Balancing selection acting in allozyme genes could also explain why their heterozygosity is high at the electrophoretic level while the nucleotide diversity in other genes is not especially high.

Even though the balancing selection explanation is tempting, some caution needs to be exercised regarding the results of HHT test. The data in paper III did not fit the neutral model, but they did not fit very well with the best-fitting bottleneck model from paper II either. It is possible the demographic scenario based on 16 loci in paper II leads to a false interpretation of the demographic history, and there is a demographic scenario that fits to both datasets and explains the data without selection.

It would be surprising if most allozymes would be affected by balancing selection. On the other hand, the selected set of loci is not random, but has been used in allozyme studies for the very reason of being polymorphic, which could have given a bias for loci where balancing selection has kept the level of polymorphism high. There are not many examples of balancing selection among nucleotide diversity studies in trees. It does not prove that it is rare though, because balancing selection can be hard to infer especially in outcrossing species with low recombination rate and complex demographic history affecting the genomewide diversity (Charlesworth 2006, Garcia & Ingvarsson 2007, Nordborg & Innan 2003). There is some evidence that allozymes could be especially prone to selection caused by varying temperature (Gillespie 1991, Somero 1978). Because *P. sylvestris* is found from various temperatures, its allozymes could be prone to balancing selection caused by spatial variation in selection. However, in *P. sylvestris* allozyme diversity data, there is no indication of clinal patterns.

Whatever the reason, the datasets differed in the LD. It seems that some or several evolutionary forces have affected differently these two sets of loci. The dataset in paper II is similar to earlier results on nucleotide diversity in *P. sylvestris* and other conifers. Therefore, it seems that for some reason allozyme coding genes are different from other genes. It would be interesting to look at the nucleotide variation at allozyme coding loci in other trees to examine whether the observed pattern is shared among them.

In summary, it is difficult to prove the action of selection on any single locus in *P. sylvestris*. Tests based on detecting deviation from neutrality as a signal of natural selection are not very useful, because a genome-wide deviation from mutation drift equilibrium was observed. In addition, unlike it is implicitly assumed, it is likely that recombination and mutation rate vary along the genome

in *P. sylvestris*. The variation in recombination rate has to be taken into account in the analysis and interpretation of molecular diversity in plants (Gaut *et al.* 2007).

3.4 Effects of demography and selection in genomic diversity of trees

The main result in papers IV and V was the scarcity of strong cases of positive selection based on molecular diversity in trees. Sign of positive selection was detected only in one locus, *AbaR*, when the divergence in coding regions of 22 loci in a phylogeny of 10 Pinus species was analyzed (paper V). In a review on genomic diversity of forest trees (paper IV), where nucleotide diversity has been analyzed, signs of positive selection were found in 15% of all loci. The set of loci studied is not a random, but many studies considered were especially designed to find genes underlying adaptation. Therefore, the percentage of loci where natural selection can be detected based on molecular diversity might be even lower if unbiased set of genes was studied.

In contrast to the lack of positive selection, signs of negative selection were found at several loci. An FDR analysis on results about the frequency of selection in Pinus (paper V) implied effect of negative selection on more than half of the branches among all studied loci. Also the d_N/d_S ratio that is lower than one in all genes studied in paper V, confirms the expectation that purifying selection is common, as has been observed in many other organisms (Hughes *et al.* 2003, Skibinski & Ward 2004, Roth & Liberles 2006).

Studies on natural selection in demographic framework in plants and especially crop species have started to emerge recently (De Mita *et al.* 2007, Hamblin *et al.* 2006, Haudry *et al.* 2007, Wright *et al.* 2005). In two tree species, *Picea abies* and *P. sylvestris*, where demographic history has been examined, a bottleneck in the history of a species explains most of the deviations from neutrality observed at individual loci (Heuertz *et al.* 2006).

Inferring the correct demographic history is not simple in forest trees. Unlike the crop species studied, forest trees have not been affected much by domestication. The essential traits and adaptations are results of natural selection—not artificial breeding conducted by humans. Therefore it is not always known at what time in the history changes in the population size took place and when selection was acting on the trait of interest. In addition, there is no wild ancestor that could be used for comparative purposes as in e.g. wheat and maize. In identifying loci putatively affected by selection, methods that are not too

sensitive to demographic events, like FDIST (Beaumont & Nichols 1996) or lnRV (Schlötterer 2002) detecting outlier loci could be useful in trees. The drawback of the outlier methods is that they depend on the assumption of a relatively low frequency of selection in the genome.

Lack of confirmed cases of positive selection does not inherently suggest that the positive selection has not been acting on the studied genes. As pointed out by Hughes (2007), most methods designed for detecting positive selection are based on unrealistic models of selection where several amino-acid changes in a protein are expected to be favoured. Hughes claims that more likely model of adaptive evolution includes single amino-acid changes, loss-of-function mutations and changes in gene expression. Testing the genetic basis of a trait with experiments and finally testing the trait's effects on fitness to prove that the trait has adaptive importance has also been emphasized (MacCallum & Hill 2006, Hughes 2007). In the hunt for local adaptation in trees, methods utilizing phenotype data, such as association analysis focusing on variation in candidate genes for certain traits are more useful than whole genome scans. The low genetic structure in neutral markers facilitates the detection of loci behind the adaptations.

The methods for detecting positive selection in molecular data have been dominated by Kimura's neutral theory. Neutrality is used as a null hypothesis and selection is often implied as a deviation from neutral expectation. The well founded statistical methods maintain this approach even though for example in Drosophila species, selection has been claimed to affect almost 50% of the amino acid substitutions (Smith & Eyre-Walker 2002). Local selective sweeps may be common also in plants due to their sessile lifeform and strong selection for local conditions, as pointed out by Kane & Rieseberg (2007) who found evidence of selection in 17 loci in *Helianthus annuus*.

Gillespie (2000) has proposed that positive selection in the form of recurrent hitchhiking (genetic draft) could be the dominant force in the genomes of some organisms. He proposes genetic draft as an explanation to the similarity of enzyme variation across species even though they should have very different effective population sizes, as already pointed out by Lewontin (1974). Effective population size estimates based on nucleotide diversity for trees are at the level of hundreds of thousands, when the true current census sizes can be thousands of millions of individuals. Genetic draft offers an explanation for observed levels of diversity in trees: efficient selection would diminish the variation that a high population mutation rate would create. Heuertz *et al.* (2006) argued that recurrent selective sweeps are not likely scenario for a species where LD decays fast and

high frequency of derived alleles are found, like in *Picea abies*. However, our results on low rates of recombination in allozyme coding genes (paper III) and lack of recombination in two japanese conifers, *Cryptomeria japonica* and *Chamaecyparis obtusa* (Kado *et al.* 2008) imply that regions of high LD can be found in tree genomes, which would facilitate the action of genetic draft.

3.5 The role of generation time and effective population size in genomic diversity of trees

In coalescent process, the time is scaled in N_e generations. In the analysis of the demographic history of *P. sylvestris* (paper II), generation time estimate of 20 years was used, which led to a conclusion that events millions of years ago could still have an impact on its genomic diversity. In reality, the average generation time in *P. sylvestris* is probably even longer due to longevity of the species and the timescale could reach even more ancient period in time.

In addition to generation time, the time to the most recent common ancestor of a sample depends also on N_e. The larger the N_e, the further back in history the sample is carrying information from. Mitochondrial and chloroplast DNA (in monoecious species) have half the N_e of nuclear markers and therefore are affected by more recent events than nuclear DNA. In *P. pinaster* and *Quercus suber* even the chloroplast genomes have been claimed to reflect the geographic events that took place 15 million years ago (Magri *et al.* 2007). In many species, this is close to timescales where interpretation of data requires taking into account also speciation events. Considering the pace of changes in the environment, the generation times and genetic diversity in trees, the situation where they would reach equilibrium is very unlikely. Changes in environment happen faster relative to coalescence events in species with long generation time, like trees compared to annual species, for example *A. thaliana*.

The idea of a single value of N_e is confusing. A long term effective population size which takes into account the past population size changes is affecting the observed amount of diversity, since the polymorphisms are accumulating over time. Varying population size affects also the efficacy of selection. In paper IV, it is concluded that as a contrast to long term effects of demographic events, selection can have very rapid effects on the large tree populations. This can happen in at least two ways in a population that has experienced recent expansion: 1) the probability of fixation a new beneficial allele is increased in growing population (Otto & Whitlock 1997) and 2) the time to fixation is lower for larger

populations, assuming the same original frequency of an allele (Kimura & Ohta 1969). One can imagine a scenario where a beneficial allele is segregating in a smaller population and behaves as if it were neutral. When population size grows, the same allele is no longer neutral because $N_e s$ is now larger. Thus relative roles of chance (drift) and deterministic processes (selection) vary over time according to population size. Neutral variation is dominated by periods of low population size during which drift is stronger and purges variation. Short periods of high population size may not produce much new neutral variation, but if important adaptations emerge, they could be quickly raised to high frequency. Thus periods of high population size could be important for genes underlying adaptation.

4 Concluding remarks

P. sylvestris has experienced population size changes in the past. The species' genetic diversity is not at mutation-drift equilibrium. Considering the pace of environmental changes, it is unlikely that a species with long generation time and large population size, like *P. sylvestris* would ever reach equilibrium. The non-equilibrium situation has to be borne in mind when studying molecular evolution, because many statistical methods in the field assume the standard neutral equilibrium model.

P. sylvestris populations in Mediterranean refugia have not contributed as much northward colonization as previously assumed. Exact dynamics of post-glacial colonization history were not resolved, but some evidence of admixture in northern Finnish populations was found. More mitochondrial polymorphisms would give a better resolution on the issue. Especially the role of putative eastern refugia is unclear, even though its effect could be substantial in the genetic composition of northern European populations.

Surprisingly low recombination was found in allozyme coding genes of *P. sylvestris*. This implies that the decay of LD in trees is not necessarily rapid in all genomic areas. This can have important implications in the efficiency of association analyses. Better picture of recombination rate in trees requires more genomic areas, longer stretches of nucleotide diversity data and larger sample sizes.

Signs of positive selection are not common in molecular diversity of trees. This could be either due to the actual low rate of positive selection, but also because the methods detect selection only in limited cases and the sample sizes are small. Selection e.g. on gene expression or single nucleotide sites is ignored. Nucleotide diversity data from non-genic genomic (as e.g. in Ometto *et al.* 2005) regions could help to evaluate the effect of selection as well as amount and variation of mutation and recombination rate in tree genomes.

References

Ahuja MR, Devey ME, Groover AT, Jermstad KD & Neale DB (1994) Mapped DNA probes from loblolly pine can be used for restriction fragment length polymorphism mapping in other conifers. Theor Appl Genet 88: 279-282.

Alvin KL (1960) Further conifers of the *Pinaceae* from the Wealden formation of Belgium. Mem Inst R Sci Nat Belg 146: 1-39.

Andolfatto P & Nordborg M (1998) The effect of gene conversion on intralocus associations. Genetics 148: 1397-1399.

Aranzana MJ, Kim S, Zhao KY, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang CL, Toomajian C, Traw B, Zheng HG, Bergelson J, Dean C, Marjoram P & Nordborg M (2005) Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. PLoS Genet 1(5): 531-539.

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(3): 1309-1321.

Beaumont MA & Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Philos Trans R Soc Lond, B 263(1377): 1619-1626.

Benjamini Y & Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B 57(1): 289-300.

Bennett MD, Leitch IJ, Price HJ & Johnston JS (2003) Comparisons with *Caenorhabditis* (~100 Mb) and *Drosophila* (~175 Mb) using flow cytometry show genome size in *Arabidopsis* to be ~157 Mb and thus ~25% larger than the *Arabidopsis* Genome Initiative Estimate of ~125 Mb. Ann Bot 91(5): 547-557.

Bernardi G (2007) The neoselectionist theory of genome evolution. Proc Natl Acad Sci U.S.A. 104(20): 8385-8390.

Boyd EF, Nelson K, Wang FS, Whittam TS & Selander RK (1994) Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) in natural populations of *Escherichia coli* and *Salmonella enterica*. Proc Natl Acad Sci U.S.A. 91(4): 1280-1284.

Brown GR, Gill GP, Kuntz RJ, Langley CH & Neale DB (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci U.S.A. 101(42): 15255-15260.

Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, Polato NR, Olsen KM, Nielsen R, McCouch SR, Bustamante CD & Purugganan MD (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet 3(9): 1745-1756.

Chamary JV, Parmley JL & Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Rev Genet 7: 98-108.

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2(4): 0379-0384.

Cheddadi R, de Beaulieu JL, Jouzel J, Andrieu-Ponel V, Laurent JM, Reille M, Raynaud D & Bar-Hen A (2005) Similarity of vegetation dynamics during interglacial periods. Proc Natl Acad Sci U.S.A. 102(39): 13939-13943.

Cheddadi R, Vendramin GG, Litt T, Francois L, Kageyama M, Lorentz S, Laurent JM, de Beaulieu JL, Sadori L, Jost A & Lunt D (2006) Imprints of glacial refugia in the modern genetic diversity of *Pinus sylvestris*. Global Ecol Biogeogr 15(3): 271-282.

Corander J, Waldmann P & Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. Genetics 163(1): 367-374.

Creevey CJ & McInerney JO (2002) An algorithm for detecting directional and non-directional positive selection, neutrality and negative selection in protein coding DNA sequences. Gene 300(1): 43-51.

Darwin C (1859) The Origin of Species. New York: Bantam Books.

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

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(4): 2123-2133.

Dobzhansky T (1955) A review of some fundamental concepts and problems of population genetics. Cold Spring Harbor Symp Quant Biol 20: 1-15.

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(2): 179-188.

Eckert AJ & Hall BD (2006) Phylogeny, historical biogeography, and patterns of diversification for *Pinus* (Pinaceae): phylogenetic tests of fossil-based hypotheses. Mol Phylogenet Evol 40(1): 166-182.

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

Fay JC & Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155(3): 1405-1413.

Fay JC, Wyckoff GJ & Wu CI (2002) Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature 415(6875): 1024-1026.

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

Franco M & Silvertown J (1996) Life history variation in plants: an exploration of the fast-slow continuum hypothesis. Philos Trans R Soc Lond, B 351(1345): 1341-1348.

Friesen N, Brandes A & Heslop-Harrison JS (2001) Diversity, origin, and distribution of retrotransposons (*gypsy* and *copia*) in conifers. Mol Biol Evol 18(7): 1176-1188.

Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J & Di Rienzo A (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet 69(4): 831-843.

Garcia MV & Ingvarsson PK (2007) An excess of nonsynonymous polymorphism and extensive haplotype structure at the *PtABI1B* locus in European aspen (*Populus tremula*): a case of balancing selection in an obligately outcrossing plant? Heredity 99(4): 381-388.

García-Gil MR, Mikkonen M & Savolainen O (2003) Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. Mol Ecol 12(5): 1195-1206.

Gaut BS, Wright SI, Rizzon C, Dvorak J & Anderson LK (2007) Opinion - Recombination: an underappreciated factor in the evolution of plant genomes. Nature Rev Genet 8(1): 77-84.

Gernandt DS, Lopez GG, Garcia SO & Liston A (2005) Phylogeny and classification of *Pinus*. Taxon 54: 29-42.

Gillespie JH (2000) Genetic drift in an infinite population: The pseudohitchhiking model. Genetics 155(2): 909-919.

Gillespie JH (1991) The Causes of Molecular Evolution. New York: Oxford University Press.

Goldman N & Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11(5): 725-736.

González-Martínez SC, Ersoz E, Brown GR, Wheeler NC & Neale DB (2006) DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. Genetics 172(3): 1915-1926.

Grotkopp E, Rejmanek M, Sanderson MJ & Rost TL (2004) Evolution of genome size in pines (*Pinus*) and its life-history correlates: Supertree analyses. Evolution 58(8): 1705-1729.

Gullberg U, Yazdani R, Rudin D & Ryman N (1985) Allozyme variation in Scots pine (*Pinus sylvestris* L) in Sweden. Silvae Genet 34(6): 193-201.

Hamblin MT, Casa AM, Sun H, Murray SC, Paterson AH, Aquadro CF & Kresovich S (2006) Challenges of detecting directional delection after a bottleneck: lessons from *Sorghum bicolor*. Genetics 173(2): 953-964.

Hamrick JL & Godt MJW (1990) Allozyme diversity in plant species. In: Brown AHD, Clegg MT, Kahler AL & Weir BS (eds) Plant population genetics, breeding and genetic resources. Sunderland, Massachusetts: Sinauer, 43-63.

Hamrick JL, Mitton JB & Linhart YB (1981) Levels of genetic variation in trees: influence of life history characteristics. In: Conkle MT (ed) Proceedings of the symposium on isozymes of North American forest trees and forest insects, July 27. US Department of Agriculture, Forest Service, Pacific Southwest Forest and Range Experiment Station, Berkeley, CA: 35-41.

Haudry A, Cenci A, Ravel C, Bataillon T, Brunel D, Poncet C, Hochu I, Poirier S, Santoni S, Glemin S & David J (2007) Grinding up wheat: a massive loss of nucleotide diversity since domestication. Mol Biol Evol 24(7): 1506-1517.

Heuertz M, De Paoli E, Källman T, Larsson H, Jurman I, Morgante M, Lascoux M & Gyllenstrand N (2006) Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce *Picea abies* (L.) Karst. Genetics 174(4): 2095-2105.

Hill WG & Robertson AV (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38: 226-231.

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(12): 1247-1266.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18(2): 337-338.

Hudson RR (2001) Two-locus sampling distributions and their application. Genetics 159(4): 1805-1817.

Hudson RR (2000) A new statistic for detecting genetic differentiation. Genetics 155(4): 2011-2014.

Hudson RR (1983a) Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23(2): 183-201.

Hudson RR (1983b) Testing the constant-rate neutral allele model with protein sequence data. Evolution 37(1): 203-217.

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(4): 1329-1340.

Hudson RR, Boos DD & Kaplan NL (1992) A statistical test for detecting geographic subdivision. Mol Biol Evol 9(1): 138-151.

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

Hudson RR, Slatkin M & Maddison WP (1992) Estimation of levels of gene flow from DNA-sequence data. Genetics 132(2): 583-589.

Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity 99(4): 364-373.

Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ & Yeager M (2003) Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc Natl Acad Sci U.S.A. 100(26): 15754-15757.

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

Hurme P, Repo T, Savolainen O & Pääkkonen T (1997) Climatic adaptation of bud set and frost hardiness in Scots pine (*Pinus sylvestris*). Can J Bot 27(5): 716-723.

Innan H, Zhang K, Marjoram P, Tavare S & Rosenberg NA (2005) Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. Genetics 169(3): 1763-1777.

Järvinen P, Lemmetyinen J, Savolainen O & Sopanen T (2003) DNA sequence variation in BpMADS2 gene in two populations of *Betula pendula*. Mol Ecol 12(2): 369-384.

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.

Kado T, Matsumoto A, Ujino-Ihara T & Tsumura Y (2008) Amounts and patterns of nucleotide variation within and between two Japanese conifers, sugi (*Cryptomeria japonica*) and hinoki (*Chamaecyparis obtusa*) (Cupressaceae *sensu lato*). Tree Genet Gen 4: 133-141.

Kane NC & Rieseberg LH (2007) Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in Common Sunflower, *Helianthus annuus*. Genetics 175(4): 1823-1834.

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(1): 215-221.

Kärkkäinen K, Koski V & Savolainen O (1996) Geographical variation in the inbreeding depression of Scots pine. Evolution 50(1): 111-119.

Kimchi-Sarfaty C, Oh JM, Kim I, Sauna ZE, Calcagno AM, Ambudkar SV & Gottesman MM (2007) A "silent" polymorphism in the *MDR1* gene changes substrate specificity. Science 315(5811): 525-528.

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

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61(4): 893-903.

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624-626.

Kimura M & Ohta T (1969) The average number of generations until fixation of a mutant gene in a finite population. Genetics 61(3): 763-771.

Kingman JFC (1982) On the genealogy of large populations. J Appl Probab 19: 27-43.

Kingman JFC (2000) Origins of the coalescent: 1974-1982. Genetics 156(4): 1461-1463.

Kinlaw CS & Neale DB (1997) Complex gene families in pine genomes. Trends Plant Sci 2(9): 356-359.

Komulainen P, Brown GR, Mikkonen M, Karhu A, García-Gil MR, O'Malley D, Lee B, Neale DB & Savolainen O (2003) Comparing EST-based genetic maps between *Pinus sylvestris* and *Pinus taeda*. Theor Appl Genet 107(4): 667-678.

Kreitman M (1996) The neutral theory is dead. Long live the neutral theory. Bioessays 18(8): 678-683.

Kreitman M & Di Rienzo A (2004) Balancing claims for balancing selection. Trends Genet 20(7): 300-304.

Kremenetski CV, Liu K & MacDonald GM (1998) The late Quaternary dynamics of pines in northern Asia. In: Richardson DM (ed) Ecology and Biogeography of *Pinus*. Cambridge: Cambridge University Press, 95-106.

Krutovsky KV & Neale DB (2005) Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir. Genetics 171(4): 2029-2041.

Langlet O (1936) Studier över tallens fysiologiska variabilitet och dess samband med klimatet. Medd Statens Skogsforskningsinst 29: 219-470.

Ledig FT (1998) Genetic variation in *Pinus*. In: Richardson DM (ed) Ecology and biogeography of *Pinus*. Cambridge: Cambridge University Press, 251-280.

Lercher MJ & Hurst LD (2002) Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet 18(7): 337-340.

Lewontin RC (1974) The Genetic Basis of Evolutionary Change. New York: Columbia University Press.

Linhart YB & Grant MC (1996) Evolutionary significance of local genetic differentiation in plants. Annu Rev Ecol Syst 27(1): 237-277.

MacCallum C & Hill E (2006) Being positive about selection. Plos Biol 4(3): 293-295.

Magri D, Fineschi S, Bellarosa R, Buonamici A, Sebastiani F, Schirone B, Simeone MC & Vendramin GG (2007) The distribution of *Quercus suber* chloroplast haplotypes matches the palaeogeographical history of the western Mediterranean. Mol Ecol

Maynard Smith J & Haigh J (1974) The hitch-hiking effect of a favorable gene. Genet Res 23(23-35)

Mayr E (1963) Animal species and evolution. Cambridge: Harvard University Press.

McDonald JH & Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351(6328): 652-654.

McVean G (2007) The structure of linkage disequilibrium around a selective sweep. Genetics 175(3): 1395-1406.

McVean G, Awadalla P & Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160(3): 1231-1241.

Mikola J (1982) Bud-set phenology as an indicator of climatic adaptation of Scots pine in Finland. Silva Fenn 16: 221-228.

Morgan T (1932) The scientific basis of evolution. New York: WW Norton.

Morgan T (1925) Evolution and genetics. Princeton, N.J.: Princeton University Press.

Morgenstern EK (1996) Geographic Variation in Forest Trees, Genetic Basis and Application of Knowledge in Silviculture. Vancouver: UBC Press.

Müller UC, Pross J & Bibus E (2003) Vegetation response to rapid climate change in Central Europe during the past 140,000 yr based on evidence from the Furamoos pollen record. Quatern Res 59(2): 235-245.

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(5): 221-228.

Muona O, Harju A & Kärkkäinen K (1988) Genetic comparison of natural and nursery grown seedlings of *Pinus sylvestris* using allozymes. Scand J For Res 3: 37-46.

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(1): 233.

Neale DB & Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9(7): 325-330.

Neale DB & Sederoff RR (1989) Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in loblolly pine. Theor Appl Genet 77(2): 212-216.

Nei M (2005) Selectionism and neutralism in molecular evolution. Mol Biol Evol 22(12): 2318-2342.

Nei M (1987) Molecular Evolutionary Genetics. New York, NY, USA: Columbia University Press.

Neuhauser C & Krone SM (1997) The genealogy of samples in models with selection. Genetics 145(2): 519-534.

Nordborg M (2001) Coalescent Theory. In: Nordborg M (ed) Handbook of Statistical Genetics. Chichester: Wiley, 179-212.

Nordborg M (2000) Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. Genetics 154(2): 923-929.

Nordborg M & Innan H (2003) The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. Genetics 163(3): 1201-1213.

Nybom H (2004) Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. Mol Ecol 13(5): 1143-1155.

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

Otto SP & Whitlock MC (1997) The probability of fixation in populations of changing size. Genetics 146(2): 723-733.

Pakendorf B & Stoneking M (2005) Mitochondrial DNA and human evolution. Annu Rev Genomics Hum Genet 6(1): 165-183.

Pal C & Hurst LD (2003) Evidence for co-evolution of gene order and recombination rate. Nature Genet 33(3): 392-395.

Palme AE, Su Q, Rautenberg A, Manni F & Lascoux M (2003) Postglacial recolonization and cpDNA variation of silver birch, *Betula pendula*. Mol Ecol 12(1): 201-212.

Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, Cheddadi R, Ennos R, Fineschi S, Grivet D, Lascoux M, Mohanty A, Muller-Starck GM, Demesure-Musch B, Palme A, Martin JP, Rendell S & Vendramin GG (2003) Glacial refugia: Hotspots but not melting pots of genetic diversity. Science 300(5625): 1563-1565.

Petit RJ & Hampe A (2006) Some evolutionary consequences of being a tree. Annu Rev Evol Ecol Syst 37: 187-214.

Plagnol V, Padhukasahasram B, Wall JD, Marjoram P & Nordborg M (2006) Relative influences of crossing over and gene conversion on the pattern of linkage disequilibrium in *Arabidopsis thaliana*. Genetics 172(4): 2441-2448.

Pons O & Petit RJ (1996) Measuring and testing genetic differentiation with ordered versus unordered alleles. Genetics 144(3): 1237-1245.

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

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(1): 13-22.

Rosenberg NA & Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Rev Genet 3(5): 380-390.

Roth C & Liberles D (2006) A systematic search for positive selection in higher plants (Embryophytes). BMC Plant Biol 6(1): 12.

Sarvas R (1962) Investigations on the flowering and seed crop of *Pinus sylvestris*. Helsinki: Valtion painatuskeskus.

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

Schlötterer C (2002) A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics 160(2): 753-763.

Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B & Mitchell-Olds T (2005) A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics 169(3): 1601-1615.

Seppänen T & Norokorpi Y (1998) The location of the coniferous timberline in the Pallas-Ounastunturi National Park. In: Tasanen T (ed) Research and management of the northern timberline region. Wilderness Center Inari, September 4.-5. 1997 Saarijärvi: Gummerus Kirjapaino, 67-73.

Shutyaev AM & Giertych M (1998) Height growth variation in a comprehensive Eurasian provenance experiment of (*Pinus sylvestris* L.). Silvae Genet 46(6): 332-349.

Sinclair WT, Morman J & Ennos RA (1998) Multiple origin of Scots pine (*Pinus sylvestris* L.) in Scotland: evidence from mitochondrial DNA variation. Heredity 80: 233-240.

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(1): 83-88.

Skibinski DOF & Ward RD (2004) Average allozyme heterozygosity in vertebrates correlates with Ka/Ks measured in the human-mouse lineage. Mol Biol Evol 21(9): 1753-1759.

Smith NG & Eyre-Walker A (2002) Adaptive protein evolution in *Drosophila*. Nature 415(6875): 1022-1024.

Somero GN (1978) Temperature adaptation of enzymes: biological optimization through structure-function compromises. Annu Rev Ecol Syst 9(1): 1-29.

Soranzo N (1999) Genetic variation in native European populations of *Pinus sylvestris* (L.).

Soranzo N, Alía R, Provan J & Powell W (2000) Patterns of variation at a mitochondrial sequence-tagged-site locus provides new insights into the postglacial history of European *Pinus sylvestris* populations. Mol Ecol 9(9): 1205-1211.

Sperisen C, Büchler U & Mátyás G (1998) Genetic variation of mitochondrial DNA reveals subdivision of Norway Spruce (*Picea abies* (L.) Karst.). In: Karp A, Isaac PG & Ingram DS (eds) Molecular tools for screening biodiversity: Plants and animals. London: Chapman & Hall, 413-417.

Storey JD & Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U.S.A. 100(16): 9440-9445.

Szmidt AE & Muona O (1989) Linkage relationships of allozyme loci in *Pinus sylvestris*. Hereditas 111(2): 91-97.

Szmidt AE & Muona O (1985) Genetic effects of Scots pine (*Pinus sylvestris*) domestication. Population genetics in forestry. Lect Notes in Biomath 60: 241-252.

Taberlet P, Fumagalli L, Wust-Saucy AG & Cosson JF (1998) Comparative phylogeography and postglacial colonization routes in Europe. Mol Ecol 7(4): 453-464.

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

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

Toomajian C, Hu TT, Aranzana MJ, Lister C, Tang CL, Zheng HG, Zhao KY, Calabrese P, Dean C & Nordborg M (2006) A nonparametric test reveals selection for rapid flowering in the *Arabidopsis* genome. Plos Biol 4(5): 732-738.

Valkonen JPT, Nygren M, Ylönen A & Mannonen L (1994) Nuclear DNA content of *Pinus sylvestris* (L.) as determined by laser flow cytometry. Genetica 92: 203-207.

Vendramin GG, Anzidei M, Madaghiele A, Sperisen C & Bucci G (2000) Chloroplast microsatellite analysis reveals the presence of population subdivision in Norway spruce (*Picea abies* K.). Genome 43(1): 68-78.

Wakeley J (2008) Coalescent Theory, an Introduction. Greenwood Village, Colorado: Roberts and Company Publishers.

Wakeley J & Aliacar N (2001) Gene genealogies in a metapopulation. Genetics 159(2): 893-905.

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

Watterson GA (1978) The homozygosity test of neutrallity. Genetics 88(405)

Weir BS & Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38: 1358-1370.

Willis KJ, Bennett KD & Birks JB (1998) The late Quaternary dynamics of pines in Europe. In: Richardson DM (ed) Ecology and Biogeography of *Pinus*. Cambridge University Press, 107-121.

Willis KJ & van Andel TH (2004) Trees or no trees? The environments of central and eastern Europe during the Last Glaciation. Quatern Sci Rev 23(23): 2369-2387.

Willyard A, Syring J, Gernandt DS, Liston A & Cronn R (2007) Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. Mol Biol Evol 24(1): 90-101.

Wright S (1943) Isolation by distance. Genetics 28(2): 114-138.

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

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD & Gaut BS (2005) The effects of artificial selection on the maize genome. Science 308(5726): 1310-1314.

Zeng K, Shi S & Wu CI (2007) Compound tests for the detection of hitchhiking under positive selection. Mol Biol Evol 24: 1898-1908.

Original articles

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

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

III Pyhäjärvi T, Kujala ST & Savolainen O What accounts for high protein heterozygosity and low nucleotide diversity in trees? (Manuscript).

IV Savolainen O & Pyhäjärvi T (2007) Genomic diversity in forest trees. Curr Opin Plant Biol 10: 162-167.

V Palmé A, Pyhäjärvi T, Wachowiak W & Savolainen O Selection on nuclear genes along a *Pinus* phylogeny. (Manuscript).

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

Errata

In paper I, Figure 1 the pie diagram in Hungary should instead be in Austria. In paper IV, Figure 1 references should be [3, 16-24, 26, 27, 30, 31] and labels *Populus trichocarpa* and *Populus tremula* have changed places. In Table 1 references for *Populus tremula* should be [25, 30, 42].

Book orders:
OULU UNIVERSITY PRESS
P.O. Box 8200, FI-90014
University of Oulu, Finland

Distributed by
OULU UNIVERSITY LIBRARY
P.O. Box 7500, FI-90014
University of Oulu, Finland

UNIVERSITY OF OULU P.O. Box 7500 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

ISBN 978-951-42-8767-1 (Paperback)
ISBN 978-951-42-8768-8 (PDF)
ISSN 0355-3191 (Print)
ISSN 1796-220X (Online)