UPPSALA
UNIVERSITET

Nucleotide diversity and Linkage disequilibrium in Norway spruce (*Picea abies*)

Venkata Raghava Pavankumar Thunga

Degree project in applied biotechnology, Master of Science (2 years), 2013
Examensarbete i tillämpad bioteknik 45 hp till masterexamen, 2013
Biology Education Centre and Department of Ecology and Genetics, Plant Ecology and Evolution, Uppsala University
Supervisor: Thomas Källman
External opponent: Jun Chen, Phd

Table of contents

Abstract

Pattern of Linkage Disequilibrium (LD) is a major factor largely determining the power of association mapping studies. Along with nucleotide diversities and DNA polymorphism, knowledge of patterns of LD along the genome needs to be to known to effectively design association mapping studies. In this study, patterns of nucleotide diversity, population structure, LD was estimated in Norway spruce (*Picea abies*). The data used for this were 23 nuclear loci sequenced in around 90 individuals originating from natural populations of Norway spruce throughout the current distribution range in Sweden and Finland. The observed levels of nucleotide diversity are variable among loci varying between 0.002 and 0.008 if measured by average pairwise nucleotide diversity. Despite the samples stretching large part of Finland and Sweden there were no evidence for strong population structure. As in earlier studies LD decays fast with distance and the average pattern of the squared correlation of allele frequencies drops to less than 0.2 within 100bp. In order to put the data in perspective previously generated data sets were re-analyzed and compared to the inferred results.

1. INTRODUCTION

Over the last 10 years our ability to collect genome wide genotype data from not only one, but thousands of individuals from a species has made it possible to disentangle the genetic variants underlying phenotypic variation among individuals (Weir, 2008; Stranger, *et al*., 2011). Studies in model systems now on a daily basis use full genome data to perform whole genome association to identify genetic variants controlling or at least co-varying with phenotype of interest (for example from plants see Seren, *et al.*, 2012). These approaches do however rely on detailed knowledge of a number of basic population genetic properties that will determine not only how many markers is needed to capture relevant parts of the genome, but also how large sample size is needed to have enough power in the association (McCarthy, *et al*., 2008; Spencer, *et al*., 2009).

Besides shedding lights upon the relationship between genotype and phenotype, multilocus and whole genome sequence and/or genotype data have also made it possible for scientist to infer demographic histories of species and in detail look at the pattern of divergence along the genome following speciation events (Pool, *et al*., 2010; Ellegren, *et al.*, 2012). One of the more striking results from these type of studies is the finding that the humans outside Africa have a small fraction of their genome that can be traced back to an admixture event between Neanderthals and humans that took place after the human migration out of Africa (Sanchez-Quinto, *et al*., 2012). For plants, the results have not been equally astonishing, but both the demographic and phylogeographic history of the model plant Arabidopsis has been rewritten and updated with the emergence of new genome wide data (François, *et al*., 2008). Similarly, demographic history and timing of domestication as well genetic variants that have been selected during the domestication process has been described in both maize (*Zea mays*) and rice (*Oryza sativa*) (Chia, *et al*., 2012; Huang, *et al*., 2012; Gross, 2012; Hufford, *et al*., 2012).

The results from the aforementioned studies highlight the importance of proper sampling and the benefits coming with detailed knowledge about key population genetic parameters (Flint-Garcia, *et al*., 2003; Hartl and Clark, 2007). Many of these basic properties are still largely unknown in non-model systems making it hard to plan and properly design association studies. In the current project the main goal is to increase our knowledge about basic population genetic parameters in the gymnosperm Norway spruce (*Picea abies*).

Norway spruce is one of around 56 species of spruce recognized in the world today (Kjällgren and Kullman, 2002). Many of the spruce (*Picea*) species have large effective sizes and can be found over large geographic areas. They belong to the family *Pinaceae*, which is the largest family within order Pinales (Wright, 1995). Spruce species are widely located in Asia, Europe, Canada and other parts of the world (Vendramin, *et al*., 2000). The genome size of *Picea* species varies from 15 to just over 20 GB, which is larger than any currently sequenced genome (Murray, 1998). This large genome size can be mainly attributed to a very large quantity of repetitive sequence in the form of retro elements. This is in contrast to what is seen in flowering plants where species with large genomes in most cases have experiences whole genome duplications (Bennetzen, 2002). The size and highly complex structure with 70-90% of the genome being repetitive has so far hampered full genome sequencing and outside the gene regions there is in essence no sequence data available. Due to the interest from forestry massive EST sequencing efforts and RNA-sequencing projects have lead to the identification of a majority of expressed genes (Ralph, *et al*., 2008; Chen, *et al*., 2012).

Norway spruce (*Picea abies*) is the most common spruce species found in western parts Europe today. The natural distribution range can be divided into two main groups; a northern group, including north Western Russia, the Baltic countries as well as Scandinavia and southern group, including the Alpine region and the Carpathian Mountains (Figure 1). Based on genetic data the Carpathian populations is the most divergent population, even though this is based on limited sample size from this area (Heuertz, *et al*., 2006). As large part of Europe were covered by ice during the last glacial maxima, the current distribution of spruce is the result of a recent migration into the north and extensive studies of pollen fossil data suggest that large parts of the northern range were re-colonized from a refugia located in Russia (Tollefsrud, *et al*., 2008). This re-colonization of Sweden does further seem to have mainly been from the north via Finland and middle and southern parts of Sweden were not reached until only a few thousand years ago (Tollefsrud, *et al*., 2008; Parducci, *et al*., 2012). More recently it was discovered that spruce (and possible also other species) likely were still present at high latitudes on the coast of Norway even during the last glacial maxima (Parducci, *et al*., 2012). However, these populations or plants does not seem to have been a large part of the re-colonization of Sweden as both fossil pollen data and genetic data is compatible with a more eastern origin of most populations in Sweden (Lagercrantz and Ryman, 1990; Tollefsrud, *et al*., 2008; Källman, 2009).

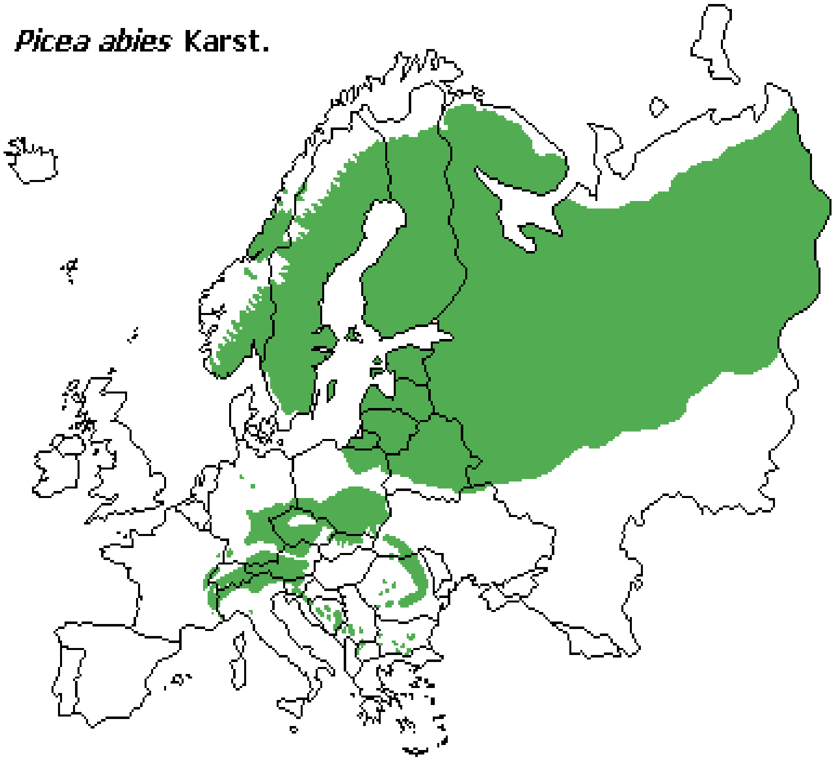

Figure 1: Geographical map showing the distribution of Norway spruce. In the most eastern part of the range *Picea abies* meets the close relative *Picea obovata*. (Source, Källman, 2009).

1.1. Linkage Disequilibrium

Linkage Disequilibrium (LD) or gametic phase disequilibrium, the non-random association between the alleles of two loci is one of the key properties in population genetics. There are several factors that affect LD; mating systems, genetic drift, selection, demographic history, population structure, and recombination (Lewontin, 1964; Flint-Garcia, *et al.*, 2003; Gaut and Long, 2003). Among those, recombination lowers LD (Brown, *et al.*, 2004; Reich, *et al.*, 2001; Hartl and Clark, 2007). If one assume that recombination rate is equal over the genome and a gene has two single nucleotide polymorphisms (SNPs) in close proximity to each other the likelihood of recombination between the two SNP's is lower than if they were located further apart and they are said to be in strong linkage (= high LD). Similarly, if there are SNPs on different chromosomes, they are physically linked to each other, but still they can be in LD (Flint-Garcia, *et al*., 2003; Hartl and Clark, 2007).

To measure the presence of linkage disequilibrium in a population there are number of statistics to estimate. Some of them are D, D' and r^2. The most commonly used estimate is r^2, which is squared correlation of allele frequency (Flint-Garcia, *et al*., 2003; Hartl and Clark, 2007).

1.2. Nucleotide Diversity and Population structure

Genetic variation in a population is an effect of various factors such as selection and demographic events. Nucleotide diversity is a measure of genetic variation in a species. The amount of variation in a population is determined from basically two parameters; the mutation rate and the effective size of the population (Hartl and Clark, 2007). Population mutation parameter is denoted as $\theta = 4N_e\mu$, where N_e is the size of the population and μ is mutation rate. Similar terminology is used for the population recombination rate that is denoted as $\rho = 4N_e r$, where r is the recombination rate (Lewis-Rogers, *et al*., 2004; Hartl and Clark, 2007). To determine level of nucleotide diversity, different estimates of population mutation parameters are used. Among the more common is Watterson's estimate (*θ_w)* and (*π),* which uses allele frequencies to estimate $4N_e\mu$. θ_w is based on polymorphic sites and π is based on pairwise nucleotide differences among the sequences (Wattersson, 1975; Nei and Li, 1979) and are both unbiased estimates of θ under a standard neutral model. To compare the variation of genetic patterns several neutrality tests were suggested and one of the most considered statistic is Tajima's *D*, which compares both θ_w and π (Tajima, 1989). On the other hand, population structure is also a key parameter to consider in population genetics. It determines whether the individuals in the populations originate from a single largely panmictic population or if they come from several differentiated populations. Many of the estimated populations will be affected by population structure, for example, LD increases as a result of population structure (Hartl and Clark, 2007).

1.3. Empirical data from Model organisms

Genome wide patterns of LD analyses have been done in several model organisms and in humans, for example, this information has been used to disentangle the genetics behind complex diseases and traits through association studies (Pritchard and Przeworski, 2001). The extent of LD over long genomic regions in human populations varies greatly as the demographic history of the populations varies. In North European populations it can reach up to 60kb something that largely is an effect of a bottleneck that happened about 27 thousand to 53 thousand years ago where the size of population was small and for time only harbored a limited set of haplotypes. Likely it has also been affected by population admixture in North European populations (Reich, *et al*., 2001). In contrast to these, there are populations of African origin that have a very different demographic history and due to this, they have much shorter average LD (less than 5kb has been observed in Nigerian populations) (Reich, *et al*., 2001; Flint-Garcia, *et al*., 2003). Harding, *et al*., 1997, mentioned that nucleotide diversity in African populations is higher than in Asian populations and average level of nucleotide diversity in a 11kb region from the introns of a

X-linked loci was quite low π = 0.063%, while other genes like, β-globin and Lipoprotein lipase gene (*Lpl*), had a π of 0.18% and 0.20%, respectively resulting higher diversity (Harding, *et al*., 1997; Nachman, *et al*., 1998; Nickerson, *et al*., 1998).

A Similar pattern as in humans has also been seen in the fruitfly (*Drosophila melanogaster*) where LD decays within 1kb in non-African populations (Andolfatto and Wall, 2003; Flint-Garcia, *et al*., 2003). The average level of nucleotide diversity θ_w and π in X-linked loci of African populations ranges from 0.025 and 0.024, while in non-African populations it is 0.01 and 0.01 which clearly shows that the diversity levels in African populations of *D. melanogaster* is reduced (Andolfatto, 2001).

In model plants, like Maize (*Zea Mays*) and Arabidopsis (*Arabidopsis thaliana*) LD has also been estimated of the complete length of the genome. In study on Maize, LD decays within a very short range in 1500bp and another study showed an interesting result in LD analyses where the decay in a gene named *su1* extended up to 12kb and this locus is a target of selection for domestication recently (Flint-Garcia, *et al*., 2003). In another study Rafalski and Morgante (2004) stated that *adh1* gene extends the decay of LD greater than 100kb in elite maize populations makes a clear impression that the variation in the patterns clearly depends on the populations chosen. The average silent nucleotide diversity level in six starch production genes is no so high and two genes *sugary1* (*su1*) and *brittle2* (*bt2*) even showed low diversity 0.002, which possible can be explained by hitchhiking (Whitt, *et al*., 2002). The reduced nucleotide diversities in maize can be due to domestication and population bottlenecks (Tenaillon, *et al*., 2004). In Arabidopsis, LD extends longer than any other plant species (Flint-Garcia, *et al*., 2003; Kim *et al*., 2007). Hagenblad and Nordborg (2002) mentioned that LD decays over 400kb in *FRI* locus, which controls the flowering time. In another study Kim *et al*. (2007) sampled 19 Arabidopsis accessions with 341,602 non-singleton SNPs and the decay of LD within 10kb which is faster compared to others in Arabidopsis studies. This long range decay of LD is mainly due to the selfing mating system of Arabidopsis which results in reduced effect of recombination events eventually leading to an increase in LD (Flint-Garcia, *et al*., 2003).

1.4. Patterns of LD and Nucleotide diversity in Conifers

As the genome size of conifers is in general very large (Murray, 1998; Morse, *et al.*, 2009; Mackay, *et al*., 2012) there is currently no single species with a fully sequenced genome and hence no genome-wide estimates of patterns of linkage disequilibrium available. However, genetic maps have been created and by using the extensive EST sequencing efforts, studies have started to look at the patterns of nucleotide diversity and LD within genes (Heuertz, *et al*., 2006; Weir, 2008; Brown, *et al.*, 2004). Nucleotide diversity is quite low in conifer species. The average diversity θ_w and π in Norway spruce ranges from 0.002 to 0.004 (Heuertz, *et al*., 2006). In *Pinus taeda* the average silent diversity (π_S) is 0.006, while in *Pinus sylvestris* average total diversity is 0.006 and 0.004 and in *Picea glauca* it is 0.005 and 0.004 (Brown, *et al*., 2004; Heuertz, *et al*., 2006; Pavy, *et al*., 2011). From these studies, which often use short fragments and limited number of genes, it is clear that LD in conifers decays very fast and even within a single 1 kb gene r^2 decays to lower than 0.2 (Heuertz, *et al*., 2006; Ingvarsson, 2005a; Pavy, *et al*., 2011).. Generally LD estimates has been done in the coding regions of conifers,there has been no sequence data available from non-coding regions, but in a recent paper of (Moritsuka, *et al*., 2012) using a bacterial artificial chromosome to obtain large pieces of noncoding regions from *Cryptomeria japonica.* Sequencing regions as far apart as 100kb in multiple individuals showed that LD could extend over large distances in those non-coding regions. In *Pinus taeda* LD decays within 2000bp when 32 samples are studied in 19 loci (Brown *et al*., 2004). In *Pinus sylvestris* LD decays at 250bp in central European samples and in Northern European samples the decay of LD is extended up to 1400bp (Pyhäjärvi, *et al*., 2007). And in an allozyme coding loci *aco* LD extended over several kilobases resulting in complete LD (Pyhäjärvi, *et al*., 2011). In a recent study of Lepoittevien, *et al.* (2012) mentioned that three genes showed a complete LD over 1304bp in *Pinus pinaster*. In *Picea mariana* the recombination rate is quite low when compared to other species and hence a high LD is estimated and LD extends up to 2000 bp (Namroud, *et al*., 2010). In *Picea glauca* LD decays with 50% after around 600 bp (Pavy, *et al*., 2011). In Norway spruce (*Picea abies*) LD is on the average low and decays within a few 100 bp when 22 loci are studied (Heuertz, *et al*., 2006). In summary patterns of LD in conifer genomes seem to vary along the genome.

1.5. Aim of the study

The main goals and aims with the present study was to extend previous attempt to estimate pattern of nucleotide diversity and linkage disequilibrium in the economically and ecologically important tree species Norway spruce (*Picea abies*). Previous studies have estimated these parameters by sampling few individuals in multiple populations and merging these when estimating population genetic parameters. However, recent advances in analyzing population genetic data from natural populations as well as of simulated non-equilibrium data have highlighted problems with this approach and pooling data from the complete distribution range of a species can under certain demographic and phylogeographic scenarios bias the results (Städler, *et al*., 2009; St. Onge, *et al*., 2012). This has also earlier been noted by more theoretical attempts to understand the effect of for example, population structure on LD and other parameters. The sampling used here allows us to not only estimate “species”-wide patterns of LD and nucleotide diversity, but also to obtain estimates within individual populations and to study any detrimental effect that pooling might have.

2. Material and Methods

2.1. DNA extraction and PCR amplification

Haploid DNA was extracted from megagametophytes from the seeds of Norway spruce (*Picea abies*) using QIAGEN DNeasy plant mini kit (Hilden, Germany). Two sets of DNA extractions were eluted; one is high concentration DNA and the other elution more diluted suitable for PCR amplification without further dilution. The PCR protocol for the loci used is described in Table 1.

Table 1. Table of primer sequences and PCR conditions for the genes amplified in this study. Multiple primer sequences are given in case where the gene was amplified in overlapping fragments. All PCR's were run for 35 cycles with an initial heating at 98° for 30 seconds and ending with a 5 minute Extension at 72°.

Gene	Fwd primer seq	Rev Primer seq	Denat	Annl	Elng
PaAP2L3	GGAAACAGGTTTATCTGG	AAGTGACCAAAAGAAAGG	98° 10s	60° 20s	72° 3min
PaCDF1	TGTAGAACGGGGTGAGT	CTGAACCCTGCTCTTGTAAT	98° 10s	60° 20s	72° 3min
PaCOL1	CAGCAGTGGAGAATGGT	CTGCATCCACATCCAATGA	98° 10s	60° 30s	72° 30s
PaFT2	TGAGGACCTTCGCAACTT	TGTCTGATTCATTCATGGCT	98° 10s	63° 15s	72° 3min
PaPRR7	TATAAGGTTAATGAAGGG	ATAAGATGTGAATGAGAAT	98° 10s	59° 30s	72° 1min
PaPRR	GGCCAGTCATCCTGAGTG	GGGCAATAAATAGTTTGTGA	98° 10s	60° 20s	72° 3min
PaWS02746	CAAGGCGGAGGATATTTC	TATTTGGCTTGGGATTGAGC	98° 10s	63° 15s	72° 3min
PaWS02749	GCATATCTGAATTCACTT	AAGACAACTTTATTTGATTT	98° 10s	60° 20s	72° 3min
PaZIP	CTATGGTTCGGGCGTCTAA	CAGCACAGGGAGTTCAGGT	98° 10s	63° 30s	72° 3min

2.2. Sequence editing and Alignment

All the purified PCR products were sent to Macrogen Sequencing Facility (Macrogen, Korea) to get chromatogram files from all the individuals. Once all the sequence files are received for particular loci, a reference sequence was imported in Phred, Phrap and Consed (Version 13.26) (Ewing *et al*. 1998 a, b; Gordon *et al*. 1998) software suit and all the sequence files were aligned to this reference sequence. Low quality sequences were removed and indels and variable sites (eg. SNPs) were checked manually and ends of the sequences were trimmed to only retain high quality sequence. Once all the editing was done the saved ace file an in-house Python script were used to create an alignment file of all the sequences. Alignment file was opened in ebiox (Version 1.5) (http://www.ebioinformatics.org/) and all the variable sites in the alignment were visually inspected on chromatograms before further analyses.

All other data files analysed were obtained as fasta alignment files that were directly used for analysis.

2.3. Nucleotide Diversity Analyses

For all the available sequence data Nucleotide diversity is estimated. Watterson's estimate of population mutation parameter θ_w (Watterson, 1975) and average number of pairwise nucleotide per site between sequences π (Nei, 1987) was calculated. Also, statistical neutrality tests like Tajima's *D* was estimated which calculates the difference between π and θ_w (Tajima, 1989). All these analyses were calculated in DNASP (Version 5.10.01) (Rozas, *et al*., 2003). Nucleotide diversity estimates and statistical neutrality tests were performed on both sub-populations and pooled populations.

2.4. Population structure Analyses

Data sets of current and previous sequences were studied to know the population structure of *Picea abies* in a model based clustering algorithm applied in a software STRUCTURE (Version 2.2) (Pritchard, *et al*. 2000; Falush, *et al*. 2003). The input file for STRUCTURE is created by taking all the SNPs into account from all the Fasta sequences. A total of 356 SNPs were observed for the current dataset and 394 for previous data set and changed the nucleotides as A = 1, C = 2, G = 3, T = 4, missing data = -9, Indels = 0 and 1 respectively. Indels longer than 1bp having SNPs were excluded. All the sequences were arranged and edited in BIOEDIT SEQUENCE ALIGNMENT EDITOR (Vesion 7.0.9.0) (http://www.mbio.ncsu.edu/bioedit/bioedit.html).

2.5. Linkage disequilibrium Analyses

The level of LD was estimated between parsimony informative sites by one of the parameters r^2 which was defined as correlation coefficient (Hill and Robertson, 1968). The decay of LD was estimated with distance over the SNPs *vs.* non-linear regression of r^2 between polymorphic sites, which was done in R package (http://www.r-project.org/). To estimate the population recombination parameter ($\rho = 4Ner$) a software package LDHAT (Version 2.2) (McVean, *et al*. 2002), was used. Haplotype diversity (*Hd*), and number of Haplotypes were calculated in DNASP.

2.6. Demographic Inference

An approximate Bayesian computational approach was used to estimate demographic parameters associated with three demographic scenarios. These scenarios are very simple and the goal with these were not to propose a new demographic "null" model for Norway spruce, but rather to investigate if the results from these type of analysis were influenced by sampling strategy. Three demographic scenarios were evaluated; Standard neutral model (SNM), Population Expansion model (PEM) and Bottleneck model (BNM) (Figure 2). The program used for this analysis is called EGGLIB (De Mita and Siol, 2012) and the summary statistics used in the estimating model parameters for all the three populations FUL, HOG, SOD were as follows.

The models used was same for all the three populations, where it computes, θ_w, π, Average H_e, and the number of points choose to sample was 100000 from a prior distribution file. SNM model has two parameters and the prior values for those two are θ (0; 0.1) and ρ (0; 0.05). And PEM model has three parameters with prior values of θ (0; 0.1), α (0; 20) and ρ (0; 0.05). Whereas BNM model has five parameters with prior values: θ (0; 0.1), t (0; 2), d (0; 0.5), f (0; 1), size of ancestral population (0; 1) and ρ (0; 0.05).

The methods used in EGGLIB were, *ABC- sample*; which takes the fasta files as an input and simulates set of points randomly for posterior estimation. *ABC-fit*; it takes the output file from *ABC-sample* and extracts the data points simulated for posterior estimation. The output file from *ABC-fit* thus contains selected values and can be considered as a posterior distribution also allows choosing the suitable demographic model for the analyses. *ABC-bin*; It analyses the posterior distribution values generated from ABC-fit. Another method *ABC-psimul*; is used to generate samples from the inferred model and hence make it possible to obtain distributions of summary statistic from the model. *ABC-compare*; It compares the given models, In this case we compared all the three models SNM, PEM, BNM in all the three populations FUL, HOG, SOD with a tolerance of 0.01.

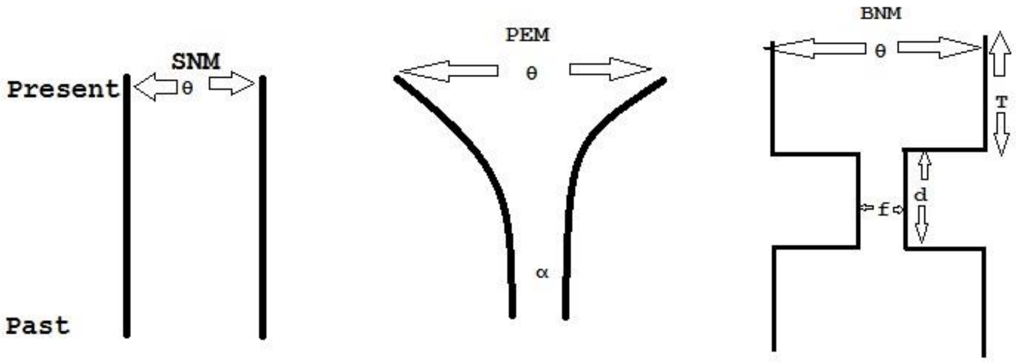

Figure 2. The above figure explains the three demographic models, which were used to analyze by Approximate Bayesian Computational approach in EGGLIB. The Standard Neutral Model estimates two parameters, Population mutation rate θ and Recombination ρ. Population Expansion model has three estimates θ, ρ, and exponential growth rate α. While Bottleneck Model estimates five parameters, along with θ, ρ, the time of the bottleneck (T) and population size during the period of bottleneck (f), and duration of the bottleneck (d). Modified from Källman, 2009.

2.7. Geographical location showing all the populations

The Geographical locations of *Picea abies* which are used in our current study are described in Table 2.

Table 2. Location and sample size of populations from Sweden and Finland.

Population	Name	Latitude	Longitude	No. sampled individuals
Saleby	SE-58	58° 36’N	13° 12’E	8
SörAmsberg	SE-60	60° 45’N	15° 42’E	8
Fulufjället	SE-61	61° 57’N	12° 78’E	24
Strängsund	SE-62	62° 63’N	15° 12’E	8
Höglunda	SE-64	64° 08’N	18° 74’E	24
Jock/Erkinvinsa	SE-66	66° 58’N	22° 70’E	8
Punkaharju	FI-61	61° 72’N	29° 39’E	8
Vilpuula	FI-62	62° 02’N	24° 63’E	8
St2	FI-66	66° 24’N	26° 53’E	8
Sodankylä	FI-67	67°41N	26°62’E	24

3. Results

3.1 Nucleotide diversity

In total more than 11,947 bp aligned nucleotides were analyzed and 186 bp were found in indels and removed from all downstream analysis. The total number of segregating sites and haplotypes were 243 and 310. The average population mutation rate was 0.004 and 0.002 as estimated by θ_w (Watterson 1975) and θ_π (Nei 1978). For many of the longer sequences the number of observed haplotypes was close to the number of collected individuals yielding a very high haplotype diversity of 0.552. Summary of all the estimates of current data are shown in Table 3.

Table 3: Nucleotide diversity estimates θw, π, Number of individuals; *n*, length of the gene; *L*, number of segregating sites; *S*, indels; *I*, Singletons; Singl, No.of Mutations; nMut, No.of Haplotypes; *h*, *D* states the Tajima's *D* which shows evidence of any deviation from neutrality in 23 loci.

Gene	**n**	***L***	***I***	***S***	**Singl**	**nMut**	***h***	***Hd***	**θw**	**π**	***D***
Col1F	110	607	1	20	7	20	14	0.672	0.0062	0.002	-1.93
Col1R	102	697	16	18	7	18	16	0.62	0.005	0.005	-1.71
Ap2L3F	45	227	2	1	1	1	2	0.04	0.001	0.0001	-1.11
Ap2L3R	39	462	3	1	1	1	2	0.051	0.0005	0.0001	-1.12
CDF1F	102	556	0	9	3	9	10	0.54	0.003	0.001	-1.11
CDF1R	101	613	3	10	2	10	11	0.77	0.003	0.002	-0.28
FT2F	96	583	26	11	4	11	11	0.74	0.003	0.003	-0.0009
FT2R	89	497	6	7	3	7	7	0.64	0.002	0.001	-0.75
PaMYB2F	84	436	4	9	3	9	10	0.51	0.004	0.001	-1.54
PaMYB2R	90	486	9	14	4	15	11	0.79	0.006	0.004	-0.9
PRR7_1F	100	495	1	6	2	6	7	0.53	0.002	0.001	-1.05
PRR7_1R	103	521	1	7	3	7	7	0.54	0.002	0.001	-0.76
PRR7_2F	102	434	2	5	3	5	6	0.4	0.002	0.001	-1.1
PRR7_2R	98	446	10	4	3	4	4	0.11	0.001	0.0003	-1.6
TOC2F2	109	603	6	15	3	15	11	0.84	0.004	0.004	-0.02
TOC2R	100	750	0	8	3	8	6	0.62	0.002	0.002	1.05
Ws02745F	97	480	6	9	3	9	9	0.46	0.003	0.001	-1.53
Ws02745R	85	471	18	2	1	2	3	0.17	0.0008	0.0003	-0.88
Ws02746F	74	542	15	22	6	22	14	0.86	0.008	0.008	0.08
Ws02746R	66	494	8	11	2	12	17	0.89	0.005	0.004	-0.6
Ws02749R	76	404	14	26	7	27	15	0.85	0.04	0.008	-1.14
ZIPF	95	435	2	8	0	8	6	0.5	0.003	0.003	0.26

ZIPR	101	678	30	10	5	10	11	0.57	0.002	0.001	-1.51
Total	2064	11947	186	243	76	235	310	12.713	----	----	----
Average	90	519	--	--	--	--	--	0.552	0.0047	0.0023	-0.86

In order to identify the variation in Tajima's *D* and also in population mutation rate samples were divided into subpopulations and same summary statistics were estimated. Due to less sample size in some populations only three populations with more sample size were considered. The results of average nucleotide diversity values from the three subpopulations are described in the Table 4.

Table 4. Nucleotide diversity and Tajima's *D* in 21 loci for 3 sub populations. Average mean values of Nucleotide diversity and statistical neutrality tests were mentioned. Total values without mean were denoted as (T).

Population	$n_{(T)}$	$I_{(T)}$	$S_{(T)}$	**Singl**$_{(T)}$	**nMut**$_{(T)}$	$h_{(T)}$	H_d	**θw**	**π**	D
Fulufjället	379	91	157	62	155	119	0.59	0.0034	0.0028	-0.57
Höglunda	349	134	137	58	138	113	0.65	0.0033	0.0031	-0.21
Sodankylä	337	51	137	60	140	117	0.66	0.0032	0.0028	-0.26

3.2 Population structure

Grouping the population according to geographic sampling locations revealed an average low *Fst* and only a few values were significant. In total, 128 individuals were considered from 10 populations containing 356 SNPs and performed analyses for number of clusters from $K = 1$ to $K = 5$ with a burn in period of 100,000 and a run length of 1000000 iterations by choosing no admixture model. The results of the clusters from $K = 2$ to $K = 5$ is shown below (Figure 3; 4; 5; 6). The most likely number of cluster is $K = 2$. This low level of population structure is further evident also from the results of the model based clustering approach implemented in the program STRUCTURE. Moreover, the results were not stable over runs and no evident easy to interpret pattern of population structure were obtained (Figure 3; 4; 5; 6).

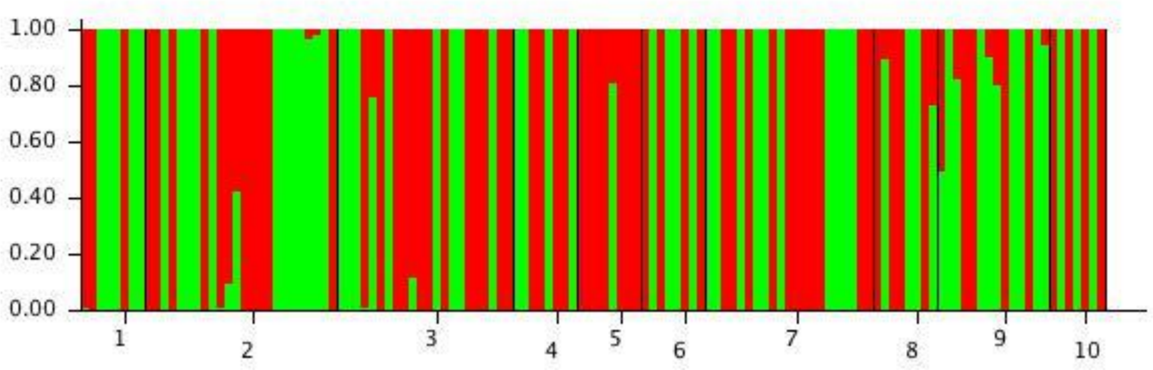

Figure 3. The structure results by choosing number of clusters $K = 2$. The values on the left represent the likelihood values and the numbers in the bottom is population number.

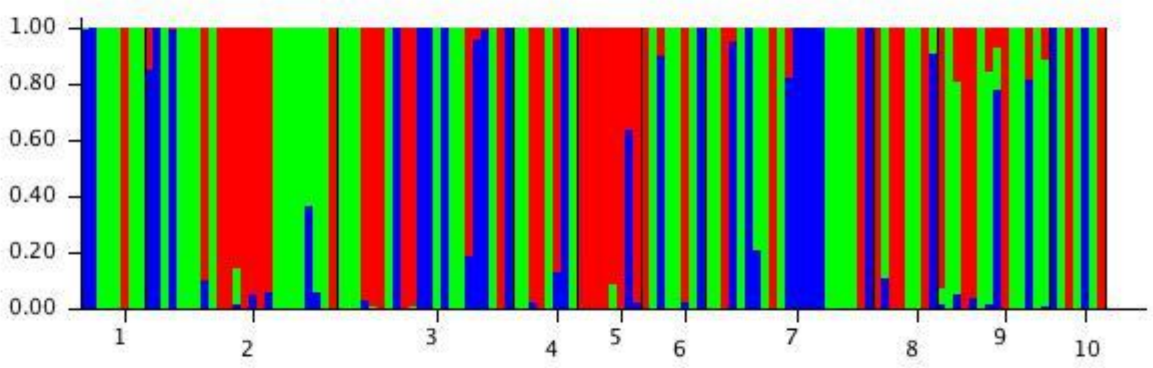

Figure 4. Likelihood values of population with number of clusters $K = 3$.

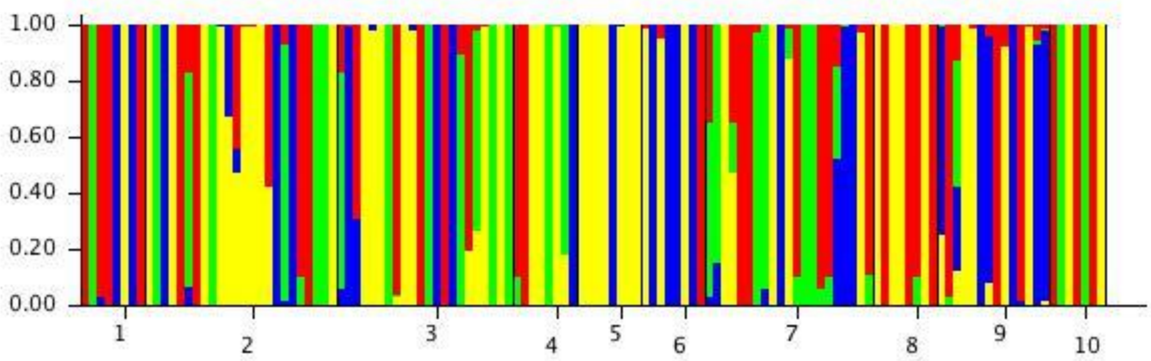

Figure 5. Likelihood values of populations with number of clusters *K* = 4.

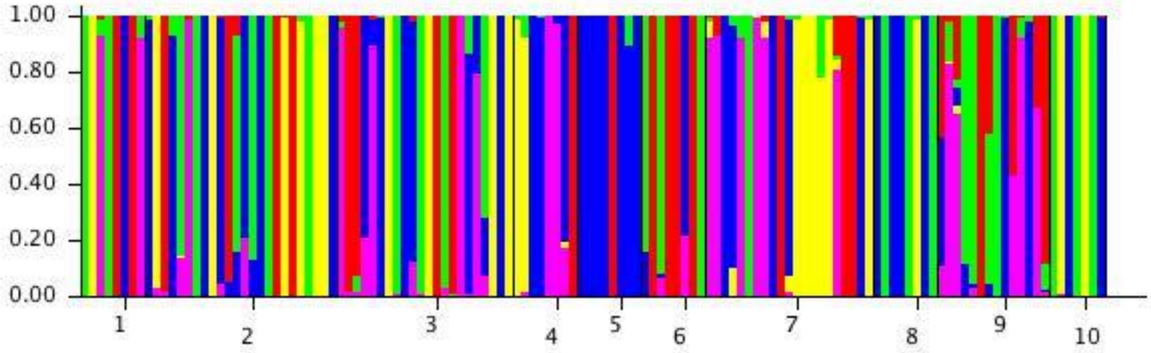

Figure 6. Likelihood values of populations with number of clusters *K* = 5.

Similarly, For previous sequence data population structure analyses were made to check the difference in allele frequency globally across 14 populations containing 394 SNPs in 126 individuals were analyzed for clusters *K* = 1 to *K* = 5 with a burn-in period of 100,000 and a run length of 500,000 iterations by choosing no admixture model. The most likely cluster is *K* = 3 (Figure 7).

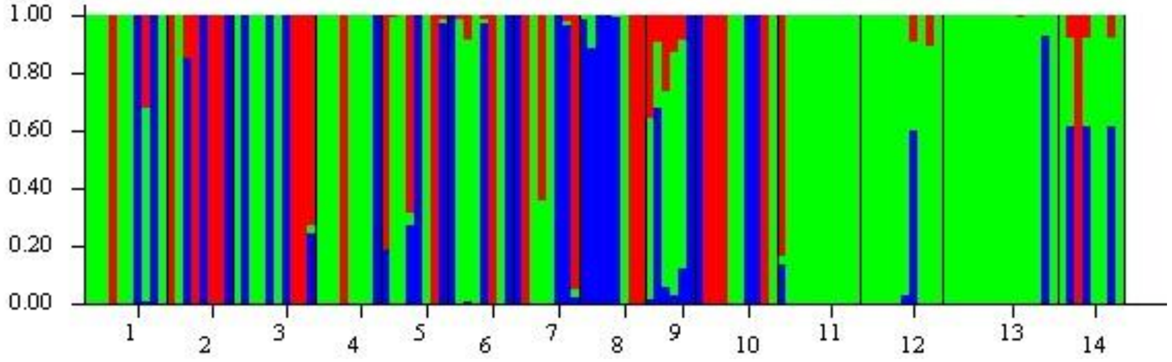

Figure 7. Structure results from the data set covering both the northern and southern populations. Columns 1-10 are northern populations and 11-14 represents the southern populations. Population 11-13 in the southern range is clearly differentiated from the rest, whereas the northern range show limited differentiation among each other.

3.3 Demographic inference

The demographic analyses were performed on the total data set with 128 individuals over 21 loci and also separately for the three populations FUL, HOG, SOD were within sample size of 24 individuals per loci were analyzed. Two loci *Ap2L3F* and *AP2L3R* were excluded for this analysis from nucleotide analysis due to bad sequence quality. All the three populations with all the three models were examined and performed a comparison between the acceptance rates for choosing the best demographic model for our data set. The results are shown in Table 5.

Table 5. Comparing three demographic models, Standard Neutral Model (SNM), Population Expansion Model (PEM), Bottleneck Model (BNM).

Population	**Acceptance rates**			**Tolerance**
	SNM	**PEM**	**BNM**	
FUL	0.22	0.55	0.24	0.01
HOG	0.22	0.53	0.26	0.01
SOD	0.24	0.50	0.26	0.01
Total	0.22	0.53	0.26	0.01

In order to look on the variations in the Tajima's *D*, *ABC-psimuls* step is used and most of the loci are well fitted with the observed values against the simulated results. But *TOC2R* slightly deviated from the observed value (Figure 8).

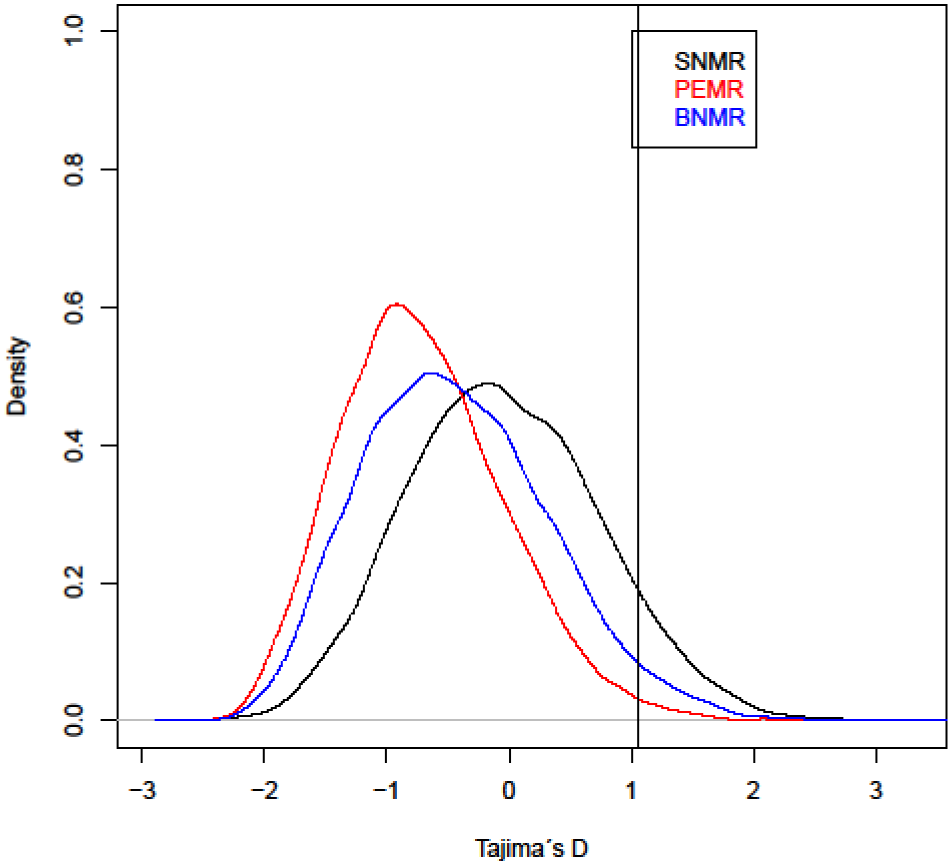

Figure 8. Graphical representation of *TOC2R* gene, which is plotted against Tajima's *D* and density. The black, red, blue lines represent the values expected from the three different inferred demographic models SNM, PEM, BNM and the straight line represents the original observed value of Tajima's D which is 1.05 (see Table 4).

3.4 Linkage Disequilibrium

To analyze the decay of LD in our current data set, the nonlinear regression model is used (Figure 9). The squared correlation of allele frequencies r^2 is used as an estimate to analyze the level of LD in 23 loci of *Picea abies*. The results explain that r^2 declined so rapidly within 0.2 among genes over 100 bp.

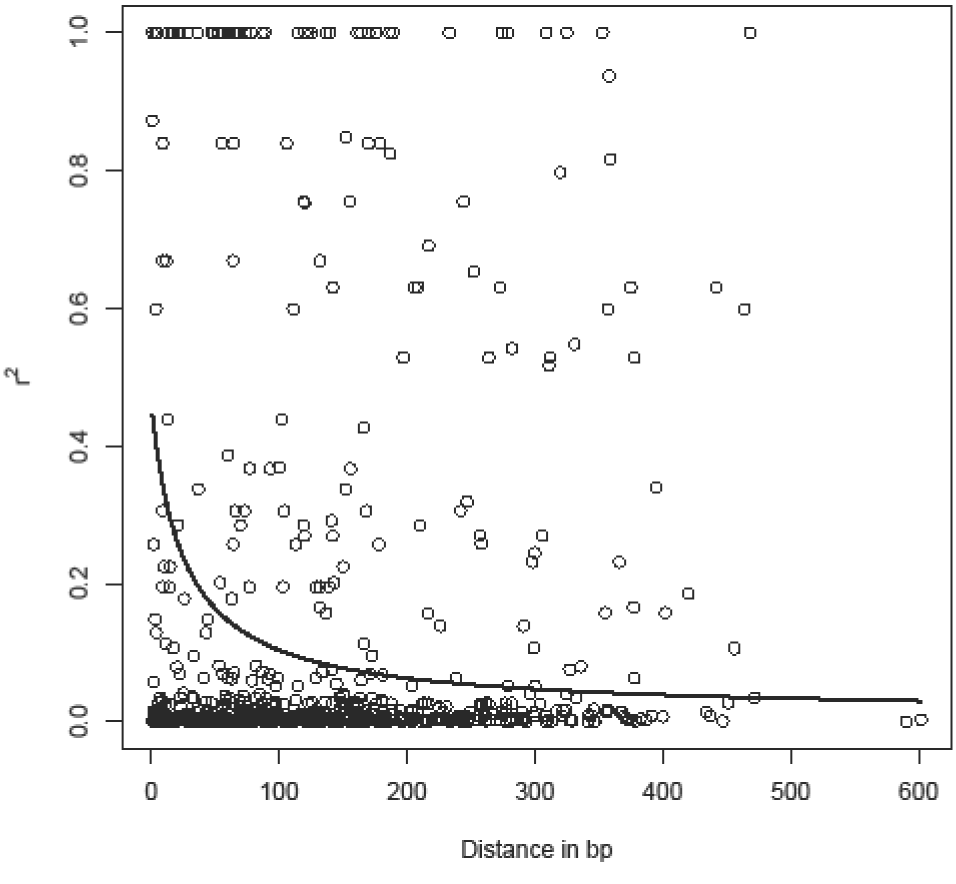

Figure 9. Graphical representation of decay of LD is shown by plotting r^2 of allele frequencies against distance between the informative sites in 23 loci. The nonlinear regression line shows the decay of LD among the genes.

4. Discussion

Outside model species there is still limited information on several basic population genetic parameters. This makes it impossible to efficiently design association mapping studies and makes it difficult to plan field sampling. Compared to many previous studies more extensive sampling allowing us to identify differences between populations for many basic population genetic summary statistics despite an inferred low population structure. Below I will try to put these data into perspective, but also discuss some of the caveats and problems that still exist when working with natural populations with large distribution ranges that perhaps does not strictly conform to assumptions commonly used population genetics models.

In both data sets the mean level of genetic diversity in Norway spruce is lower than one would expect from the very large population size it has today. Standard statistical neutrality tests such as Tajima's *D* resulted in negative values. Both of this observation supports a growing population that is recovering from a smaller size in the past. Like many other temperate plant species in Europe the glacial history of the continent likely have pushed distribution ranges south. This has led to conclusions that the majority of species were confounded to different refugia mainly in southern and Eastern Europe during the last glacial maxima and only very recently obtained its current distribution range (Petit, *et al*., 2003; Parducci, *et al*., 2012). More recent analysis of the same data and additional data from nuclear data have however highlighted the fact that they may not have been restricted to refugia, but rather had a more southern distribution range that may not for all species have been very restrictive (Lascoux, *et al*., 2004, Heuertz, *et al*., 2006). This concept of southern refugia has for Norway spruce been even more questioned as studies of ancient DNA from lake sediment cores in northern and southern Norway has indicated that trees might have survived much further north than previously assumed (Parducci, *et al*., 2012). The pollen fossil data from Sweden do however still support a re-colonization mainly from the north as the trees migrated in from Sweden north of the Baltics and reached central and southern Sweden as late as for only a few thousand years ago (Giesecke and Bennet, 2004).

4.1. Nucleotide diversity and Statistical neutrality tests

The estimated average level of population mutation rate for the current data is θ_W = 0.004 and π = 0.002 which consistent with earlier results from Heuretz, *et al*. (2006). The average silent nucleotide diversity π_S = 0.0029 a value that is close to Heuertz data (0.0039). The level of nucleotide diversity estimates makes the point clear that conifer carries a low level of diversity when compare to other species. In other conifers, like *Pinus sylvestris* the level of average π_s across 14 sequenced genes is 0.0041 (Dvornyk *et al*. 2002), while in *Pinus taeda* when 19 genes were sequenced estimates were even lower i.e., π_S = 0.0064 (Brown, *et al*. 2004). While in *Populus tremula* the level of polymorphic data and level of gene diversity is even lower than previously estimated values, which can be due to different, sequence strategies (Ingvarsson, 2008). The average population mutation rate ranges from 0.0048 and 0.0042, and the mean π in noncoding regions is 0.016 in the earlier study of Ingvarsson (2005b) and reduced to 0.0048 (Ingvarsson, 2008). In outcrossing species like *Arabidopsis lyrata* and *Arabidopsis halleri* the average silent nucleotide diversity is 0.023 and 0.015 (Ingvarsson, 2005a). While in *Arabidopsis thaliana* π_S = 0.0083 which is lower when compared to the other two Arabidopsis species (Heuertz, *et al*. 2006).

The average Tajima's *D* value of -0.86 suggests that the data comes from an expanding population. The Tajima's *D* in Heuertz, *et al*. (2006) across 22 loci in seven populations is -0.92 and this is quite similar and supports our data that *D* is more or less similar over different *Picea abies* populations. In order to make sure that pattern of population structure did not influence our results, we did the analyses also on individual subpopulations where we considered 3 subpopulations. The observed negative values also in the 3 subpopulations confirm that sampling did not have any effect on our data set.

4.2 Population Structure

Population structure analyses from the current data set were performed by choosing different runs and with varying different *K* values. The most likely cluster from the current data set is *K* = 2 (Figure 3). Though the current data set does not show any clear structure in populations the cluster *K* = 2 is likely good enough to choose as an example to describe that no clear population structure is seen in the populations of Sweden and Finland. The results with previous data set were different as it contains the samples globally from a whole distribution range (Figure 7). In Figure 7 we can see that the populations from northern part of Europe does not show clear structure, but re-analysis of the previous data show that the northern group is distinguishable from the southern range of populations originating from Germany, Switzerland and Romania.

4.3 Linkage Disequilibrium

The estimation of LD in *Picea abies* is low when plotted against distance of bp and squared correlation of allele frequency r^2. LD decay to r^2 lower than 0.2 already at distances of around 100bp (Figure 9). Generally, the decay of LD is rapid in conifers (Heuertz, *et al*., 2006; Ingvarsson, 2005a; Pavy, *et al*., 2011) but most studies are restricted to short genie fragments that might not be representative for the complete genome. A number of recent publications looking at longer fragments and non-coding DNA in conifers do suggests that decay of LD is heterogenous and extrapolation from genes should be avoided (Larsson, *et al*., 2013). As mentioned already in the introduction, the decay of LD is much faster in outcrossing species like conifers when compared to other selfing species like *Arabidopsis* (Heuertz, *et al*., 2006; Ingvarsson, 2005a; Pavy *et al*., 2011). Even the theory predicts a clear impact of population structure on estimates of LD the level of population structure observed here does not seem to be strong enough to affect estimate of LD strongly. The low level of LD in Norway spruce is consistent with a growing outcrossing population with a fairly large effective population size.

4.4 Approximate Bayesian Computation (ABC) Analysis

In order to understand the influence past demographic events have had on present day nucleotide diversity we performed an ABC analysis to compare three simple demographic scenarios. As we observed over all negative Tajima´s D values it is not surprising that the ABC analysis gave support for a population growth model compared to both a bottleneck and standard neutral model. This support was found both for individual populations and in analysis where all data was merged and treated as a single population. Taken together the results hence support expanding populations of Norway spruce and highlight the fact that one need to take this into account if one wants to pinpoint loci that might be subjected to selection. In attempts to pinpoint loci subjected to selection we to the conservative approach of testing observed values of summary statistics to values simulated from the posterior simulation of the three different models. The vast majority of the loci fell well within the expectations from one or more of the three models. One fragment, *TOC2R* gene did with a positive Tajima´s D value fall in the tail of the inferred distributions and does compared to expectations have few singletons and might be a gene that is subjected to some kind of selection.

5. Conclusion

In this study I tried to estimate the patterns of nucleotide diversities, LD, population structure, and demographic model that is compatible with the data collected. All the results in this study, like the variation in nucleotide diversities, increased level of average heterozygosity, and rapid decay of LD over 0.2 within 100bp, no clear population structure, the departure of overall data from standard neutral model were in agreement with earlier studies on Norway spruce. As theoretical work suggest that one should be careful in pooling individuals from multiple populations to estimate demographic scenario it is important to estimate not only species-wide demographic history, but also to look at local populations with denser sampling. This study lends support for previous inferred demographic history even at local populations suggesting that the effect of population structure in Norway spruce is not biasing demographic inferences strongly. Further, studies on Norway spruce with longer fragments and more sample size in large number of loci is needed to fully understand the genome structure of spruce and also help in design of association mapping attempts.

6. Acknowledgments

I sincerely thank Prof. Martin Lascoux for giving me the opportunity to do my thesis project in the Department of Plant Ecology and Genetics. I thank Dr. Thomas Kallman in all the aspects right from the day I entered this Department. I would not have finished my project without his support, guidance, patience, motivation and encouragement. I am so happy to do my thesis under his supervision. I also thank Jun and Yoshiaki for being helpful in lot of things and Nagarjun for helping me when I am stuck with programming language R. I also thank my thesis Coordinator Anna for helping me with my thesis dates and report corrections. I thank all the people of the department for being so nice with me; I will definitely miss those late night working lab hours. I thank my family and friends for their enormous love, and support all the time. I thank my girlfriend Asha for being so patient and supportive in the times I needed the most.

I dedicate this thesis to my supervisor Dr. Thomas Kallman and to my family.

7. Abbreviations

LD	Linkage disequilibrium
EST	Expressed sequence tags
SNP	Single nucleotide polymorphism
PCR	Polymerase chain reaction
ABC	Approximate Bayesian Computation
SNM	Standard neutral model
PEM	Population expansion model
BNM	Bottleneck model
FUL	Fulufjället
HOG	Höglunda
SOD	Sodankylä

8. References

Andolfatto, P. (2001). Contrasting Patterns of X-Linked and Autosomal Nucleotide Variation in *Drosophila melanogaster* and *Drosophila simulans*. *Mol. Biol. Evol*, 18: 279-290.

Andolfatto P., and Wall J. D., (2003). Linkage disequilibrium patterns across a recombination gradient in African Drosophila melanogaster. *Genetics*, 165: 1289-1305.

Bennetzen, J. L. (2002). Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica*, 115: 29-36.

Brown, G. R., Gill, G. P., Kuntz, R. J., Langley, C. H., and Neale, D. B. (2004). Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences of the United States of America*, 101:15255-15260.

Chia J. M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., Doebley, J., Elshire, R. J., *et al*. (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet*, 44: 803-807.

Chen, J., Uebbing, S., Gyllenstrand, N., Lagercrantz, U., Lascoux, M., and Källman, T. (2012). Sequencing of the needle transcriptome from Norway spruce (Picea abies Karst L.) reveals lower substitution rates, but similar selective constraints in gymnosperms and angiosperms. *BMC Genomics*, 13: 589.

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

De Mita, S., and Siol, M. (2012). EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC *Genetics*, 13:27.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., Kunstner, A., Mäkinen, H., Nadachowska-Brzysja, K., Qvarnström, A., Uebbing, S., and Wolf, J. B. (2012). The genomic landscape of species divergence in Ficedula flycatchers. *Nature*, 491:756-760.

Ewing, B., and Green, P. (1998a). Base-Calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Research*, 8:186-194.

Ewing, B., Hillier, L., Wendl, C. M., and Green, P. (1998b). Base-Calling automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research*, 8:175-185.

Falush, D., Stephens, M., and Pritchard, K. J. (2003).Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics*, 164:1567-1587.

Flint-Garcia, S. A., Thornsberry, J. M., and Buckler, E. S. IV. (2003). Structure of Linkage Disequilibrium in plants. *Annu.Rev.Plant Biol*, 54:357-374.

François, O., Blum, M. G. B., Jakobsson, M., Rosenberg, N. A. (2008). Demographic History of European Populations of Arabidopsis thaliana. *PLoS Genet*, 4:e1000075.

Gaut, B. S. and Long, A. D. (2003). The lowdown on Linkage Disequilibrium. *The Plant Cell*, 15: 1502-1506.

Giesecke, T., and Bennett, K. D. (2004). The Holocene spread of Picea abies (L.) Karst. in Fennoscandia and adjacent areas. *Journal of Biogeography,* 31:1523-1548.

Gordon, D., Abajian, C., and Green, P. (1998). Consed: A graphical tool for sequence finishing. *Genome Research*, 8:195-202.

Gross, B. L. (2012). Rice domestication: histories and mysteries. *Molecular Ecology*, 21: 4412-4413.

Hagenblad, J., and Nordborg, M. (2002). Sequence variation and haplotype structure surrounding the flowering time locus FRI in Arabidopsis thaliana. *Genetics*, 161:289-298.

Harding, R. M., Fullerton, S. M., Griffiths, R. C., Bond J., *et al*. (1997). Archaic African and Asian Lineages in the Genetic Ancestry of Modern Humans. *Ann. J. Hum. Genet*, 60: 772-789.

Hartl, L. D., and Clark G. A. (2007). Principles of Population Genetics. 4th ed, Sinauer Associates,Inc,Sunderland,U.S.A.

Heuertz, M., De paoli, E., Kallman, T., Larsson, H., Jurman, I., Morgante, M., Lascoux, M., and Gyllenstrand, N. (2006). Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [Picea abies (L.) Karst]. *Genetics*, 174:2095-105.

Hill, W. G, and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet*, 38:226-331.

Huang, P., Molina, J., Flowers J. M., Rubinstein, s., Jackson, S. A., Purugganan, M. D., and Schaal, B. A. (2012). Phylogeography of Asian wild rice, Oryza rufipogon: a genome-wide view. *Molecular Ecology*, 21: 4593-4604.

Hufford, M. B., Xu, X., Heerwaarden, J. V., Pyhäjärvi, T., Chia, J. M., *et al*. (2012). Comparative population genomics of maize domestication and improvement. *Nature Genetics*, 44: 808-811.

Ingvarsson, K. P. (2005a). Nucleotide Polymorphism and Linkage Disequilibrium Within and Among Natural Populations of European Aspen (Populus tremula L., Salicaceae). *Genetics*, 169:945-953.

Ingvarsson, P. K. (2005b). Molecular population genetics of herbivore-induced protease inhibitor genes in European Aspen (Populus tremula L., Salicaceae). *Mol Biol Evol*, 22:1802-1812.

Ingvarsson, P. k. (2008). Multilocus Patterns of Nucleotide Polymorphism and the Demographic History of Populus tremula. *Genetics*, 180: 329-340.

Källman, T. (2009). Adaptive evolution and demographic history of Norway spruce (*Picea abies*). *Digital Comprehensive summaries of Uppsala Dissertations from the Faculty of Science and Technology*. Uppsala.

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R.M., Ossowski, S., Ecker, J.R., weigel, D., and Nordborg, M. (2007). Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics*, 39:1151-1155.

Kjällgren, L., and Kullman, L. (2002). Geographical patterns of tree-limits of Norway spruce and Scots pine in the southern Swedish Scandes. *Norwegian Journal of Geography*, 56: 237-245.

Lagercrantz, U., and Ryman, N. (1990). Genetic structure of Norway spruce (Picea abies): Concordance of morphological and allozymic variation. *Evolution*, 44: 38-53.

Lascoux, M., Palme, A. E., Cheddadi, R., and Latta, R. (2004). Impact of the Ice Ages on the genetic structure of trees and shrubs. Phil. Trans. Roy. Soc. 359:197-207.

Larsson, H., Kallman, T., Gyllenstrand, N., and Lascoux, M. (2013). Distribution of long-range Linkage disequilibrium and Tajima's D values in Scandinavian populations of Norway spruce (*Picea abies*). *G3*, 3:795-806.

Lepoittevin, C., Harvengt, L., Plomion, C., Garnier-Géré, P.(2012). Association mapping for growth, straightness and wood chemistry traits in the Pinus pinaster Aquitaine breeding population. *Tree Genetics & Genomes*, 8:113-126.

Lewis-Rogers, N., Crandall, K. A., and Posada, A. (2004). Evolutionary analyses of genetic recombination. *Dynam. Genet*, p.408.

Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, 49: 49-67.

Mackay, J., Dean, J. F., Plomion, C., Peterson, D. G., Canovas, F. M., Pavy, N., Ingvarsson, P. K., Savolainen, O., Guevara, M. A., Fluch, S., Vinceti, B., Abarca, D., Díaz-Sala, C., and Cervera, M. T. (2012). Towards decoding the conifer giga-genome. *Plant Mol Biol*, 80:555-569.

McCarthy, I. M., and Hirschhorn, N. J. (2008). Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet*, 17:R156-R165.

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

Moritsuka, E., Hisataka, Y., Tamura, M., Uchiyama, K., Watanabe, A., Tsumura, Y., and Tachida, H. (2012). Extended linkage disequilibrium in noncoding regions in a conifer, *Cryptomeria japonica*. *Genetics*, 190:1145-8.

Morse A. M., Peterson, D. G., Islam-Faridi, M. N., Smith, K. E., Magbanua, Z., *et al*. (2009). Evolution of Genome Size and Complexity in Pinus. *PLoS ONE*, 4: e4332.

Murray, B. G. (1998). Nuclear DNA amounts in Gymnosperms. *Annuals of Botany*, 82:3-15.

Namroud, M. C., Guillet-Claude, C., Mackay, J., Isabel, N., and Bousquet, J. (2010). Molecular evolution of regulatory genes in spruces from different species and continent: heterogeneous patterns of linkage disequilibrium and selection but correlated recent demographic changes. *J Mol Evol*, 70:371-386.

Nachman, M. W., Bauer, V. L., Crowell, S. L. and Aquadro, C. F. (1998). DNA variability and recombination rates at X-linked loci in humans. *Genetics*, 150: 1133–1141.

Nickerson, D. A., Taylor, S. L., Weiss, K. M., Clark, A. G., Hutchinson R. G., *et al*., (1998). DNA sequence diversity in a 9.7 kb region of the human lipoprotein lipase gene. *Nat. Genet*. 19: 233–240.

Nei, M., and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci*, 76: 5269-5273.

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

Parducci, L., Jorgensen, T., Tollefsrud, M. M., Elverland, E., Alm, T., Fontana, S. L., Bennett, K. D., Haile, J., Matetovici, I., Suyama, Y., Edwards, M. E., Andersen, K., Rasmussen, M., Boessenkool, S., Coissac, E., Brochmann, C., Taberlet, P., Houmark-Nielsen, M., Krog Larsen, N., Orlando, L., Gilbert, M. T. P., Kjær, K. H., Greve-Alsos, I., and Willerslev, E. (2012). Glacial Survival of Boreal Trees in Northern Scandinavia. *Science*, 335:1083-1086.

Pavy, N., Namroud, M. C., Gagnon, F., Isabel, N., and Bousquet, J. (2011). The heterogeneous levels of linkage disequilibrium in white spruce genes and comparative analysis with other conifers. *Heredity (Edinb)*, 108:273-84.

Petit, R. J., Aguinagalde, I., de Beaulieu, J. L., Bittkau, C., Brewer, S., Cheddadi, R., Ennos, R., Fineschi, S., Grivet, D., Lascoux, M. et al. (2003). Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, 300:1563-1565.50.

Pool, J. E., Hellmann, I., Jensen, J. D., and Nielsen, R. (2010). Population genetic inference from genomic sequence variation. *Genome Res*, 20:291-300.

Pritchard, K. J., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155:945-959.

Pritchard, K. J., and Przeworski, M. (2001). Linkage Disequilibrium in humans: Models and data. *Am J Hum Genet*, 69:1-14.

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

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

Rafalski, A., and Morgante, M. (2004). Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. *Trends in Genetics*, 20:103-111.

Ralph, S. G., Chun, H. J., Kolosova, N., Cooper, D., Oddy, C., Ritland, C. E., *et al*. (2008b). A conifer genomics resource of 200,000 spruce (Picea spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (Picea sitchensis). *BMC Genomics*, 9: 484.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001).Linkage disequilibrium in the human genome, *Nature*, 411:199-204.

Rozas, J., Sanchez-Delbarria, C. J., Messeguer, X., and Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, 29:2496-2497.

Sanchez-Quinto, F., Botigue, L. R., Civit, S., Arenas, C., Avila-Arcos, M. C, *et al*. (2012). North African Populations Carry the Signature of Admixture with Neandertals. *PLoS ONE*, 7: e47765.

Seren, Ü., Vilhjálmsson, B. J., Horton, M. W., Meng, D., Forai, P., Huang, Y. S., Long, Q., Segura, V., and Nordborg, M. (2012). GWAPP: A Web Application for Genome-Wide Association Mapping in Arabidopsis. *The Plant Cell Online*, 24: 4793-4805.

Segerström, U., and von Stedingk, H. (2003). Early-Holocene spruce, Picea abies (L.) Karst., in west central Sweden as revealed by pollen analysis. *Holocene*, 13: 897.

Spencer, C. C. A., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLoS Genetics*, 5:e1000477.

St. Onge, K. R., Palme, A. E., Wright, S. I., and Lascoux, M. (2012). Impact of Sampling Schemes on Demographic Inference: An Empirical Study in Two Species with Different Mating Systems and Demographic Histories. *G3(Bethesda)*, 2:803-814.

Städler, T., Haubold, B., Merino, C., Stephan, W., and Pfaffelhuber, P. (2009). The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics*, 182:205-216.

Stranger, B. E., Stahl, E. A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187:367-383.

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

Tenaillon, M. I., U'Ren, J., Tenaillon, O., Gaut, B. S. (2004). Selection versus demography: A multilocus investigation of the domestication process in maize. *Mol Biol Evol*, 21: 1214-1225.

Tollefsrud, M. M., Kissling, R., Gugerli, F., Johnsen, O. et al. (2008). Genetic consequences of glacial survival and postglacial colonization in Norway spruce: Combined analyses of mitochondrial DNA and fossil pollen. *Mol Ecol*, 17: 4134-41.

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

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

Weir, S. B. (2008). Linkage Disequilibrium and Association Mapping. *Annu Rev Genomics Hum Genet*, 9:129-142.

Whitt, S. R., Wilson, L. M., Tenaillon, M. I., Gaut, B. S., Buckler, E. S. (2002). Genetic diversity and selection in the maize starch pathway. Proc. *Natl. Acad. Sci*, 99: 12959–62.

Wright, J. W. (1995). Species crossability in spruce in relation to distribution and taxonomy. *Forest science*, 1:319-340.