UPPSALA
UNIVERSITET

Signatures of ancient and recent selection and demographic events in spruce (*Picea*) genomes

Jing Huang

Degree project in biology, 2008
Examensarbete i biologi, 20 p, 2008
Biology Education Center and Department of Evolutionary Functional Genomics, Uppsala University
Supervisors: Martin Lascoux, Thomas Källman and Niclas Gyllenstrand

1 Introduction

1.1 Population genetics background

Patterns of DNA sequence variation within and between species reflect the evolutionary history of genes and species. The evolutionary forces influencing genetic variation are natural selection, mutation, recombination, genetic drift and gene flow. Both natural selection and genetic drift have an impact on the standing genetic variation, however the degree to which allele frequencies are affected depends on population size. Small populations experience stronger genetic drift than large ones and, consequently, selection is less efficient in small populations than in large ones. Mutation introduces new variations into a population. Mutations that result in a substitution of an amino acid (non-synonymous mutations) are often deleterious for the organism and are removed by natural selection. Favorable non-synonymous mutations, on the other hand, tend to accumulate, resulting in evolutionary change. The rate of synonymous substitution is much higher than that of nonsynonymous substitution and is similar for many different genes. Therefore investigating the number of synonymous and non-synonymous mutations may provide information about the degree of selection operating on a system (Nei and Gojobori 1986). The great majority of mutations, according to Kimura's neutral theory (1983), are selectively neutral. That means, these molecular changes do not influence the fitness of individual organisms, they are thus neither subject to, nor explicable by, natural selection. The neutral theory was, and to the large extent still is, used as the null hypothesis in population genetic studies.

The neutral theory was later extended by Kingman's coalescent theory (1982). The coalescent, a gene genealogy, represents the inheritance relationships between a set of alleles. The standard coalescent, based on the Wright-Fisher model, is a retrospective model that traces all alleles of a gene in a sample from a population back in time to a single ancestral copy shared by all members of the population (known as the most recent common ancestor MRCA). It assumes that the population size has been constant through time, that mating is random and that the number of alleles in the sample is small compared to the effective population size. Coalescent-based estimation methods allow us to estimate the expected time to coalescence and establish the relationships of coalescent times to population size, age of the MRCA, and other population genetic parameters.

The standard coalescent shows that gene frequencies of a genome, which are only affected by mutation and genetic drift, will reach equilibrium in the end. The equilibrium is described by $\theta = 4N_e\mu$ for a diploid, where θ is the scaled mutation rate, N_e is the effective population size and μ is mutation rate per sequence per generation. The expected value of θ can be estimated from the number of segregating sites (θ_w) and the nucleotide diversity (π). θ_w is corrected for sequence length and π is the average number of pairwise nucleotide differences between two randomly chosen sequences in a sample. θ_w and π are equal under the standard neutral model. But selection, demography and other violations of the standard neutral model will change the expected values of θ_w and π, so that they are no longer expected to be equal.

Neutrality tests are therefore used to compare observed DNA sequence variation to that expected under the standard neutral model and to identify sequences which do not fit the model. For instance, Tajima's D test (1989) calculates a statistic D based on the standardized difference between π and θ_w. Under the standard neutral model, D tends to be close to zero. The direction of Tajima's D, to certain extent, infers the evolutionary and demographic forces a population went through. Positive D statistics reflect an excess of intermediate-frequency alleles in a population, which may result from balancing selection or population bottlenecks. Negative D statistics reflect an excess of rare alleles in a population, which may result from positive selection or an increase in population size. D test is unfortunately not very powerful and depends strongly on the equilibrium hypothesis. For instance, an excess of rare alleles can result from a recent founder event, the current variation corresponding to recent mutations that have not yet reached equilibrium. Fay and Wu's H test (2000) measures the difference between π and θ_H, an estimate of θ that puts more emphasis on high frequency derived variants, and specially tests for hitchhiking. The main effect of hitchhiking is that the sites linked to the favoured allele (known as linkage disequilibrium) "escape" selective sweep and are overrepresented in the population. A positive H statistic is indicative of an excess of high frequency derived mutations.

Recent studies have shown that the standard neutral model can globally be rejected for organisms as diverse as humans (Voight *et al.* 2006), *Drosophila* (e.g. Haddrill *et al.* 2005) or *Arabidopsis* (Nordborg *et al.* 2005) and that both demography and natural selection played a major role in the recent past of these species. To distinguish between past demographic events and selection one can examine many loci. Because all loci have been subject to the same demographic history, they should exhibit the same pattern if they evolved neutrally. The basic principle is that demographic events affect all genes in a genome in a similar way, whereas natural selection has only local effects. With a large sample of random sequences, it is possible to estimate the overall effect of demography on genetic variation, and thereby to identify genes that in addition have been affected by natural selection. A large number of independent loci are needed because the coalescent is inherently extremely variable (Hudson 2002).

1.2 Previous studies in spruce (*Picea*) genomes

Spruce refers to the trees of the genus *Picea*, a genus in the pine family (*Pinaceae*) including 28 to 50 species, depending on the taxonomic authority (Wright 1955; Everett 1981; Schmidt 1989). They grow in the northern temperate and boreal regions. They are large evergreen coniferous trees, having needle-like leaves and bearing cones. Most conifers including spruce species are characterized by large population sizes, a predominantly outcrossing mating system and the potential for long-distance gene flow (Hamrick *et al.* 1992). In the long run, these factors should retard the fixation of neutral or nearly neutral polymorphism and thus increase genetic variation. This was supported by high levels of isozyme diversity observed in conifers and trees in general, as compared to most other plants and organisms (e.g., Hamrick *et al.* 1992; Hamrick and Godt 1996; Ledig 1998). But the estimates of nucleotide diversity

reported so far in conifers (e.g. Dvornyk *et al.* 2002; García-Gil *et al.* 2003; Brown *et al.* 2004; Heuertz *et al.* 2006; Pyhäjärvi *et al.* 2007) have been much lower than expected on the basis of their life-history traits and high levels of isozyme diversity observed for these species (Hamrick and Godt 1996). For example, the average level of silent nucleotide diversity was 0.00399 in Norway spruce (Heuertz *et al.* 2006) and 0.00648 in Scots pine (Pyhäjärvi *et al.* 2007), which were quite lower than the estimate in *A. thaliana* (π_s=0.0083, Schmid *et al.* 2005) and an order of magnitude lower than estimates in aspen (π_s=0.0160, Ingvarsson 2005) and wild relatives of maize (π_s=0.012-0.013, Tiffin and Gaut 2001). On the other hand, little differentiation in nuclear DNA markers has been observed among spruce populations, which is consistent with the results in isozyme studies (Boyle and Morgenstern 1987; Furnier *et al.* 1991; Isabel *et al.* 1995; Müller-Starck 1995; Jaramillo-Correa *et al.* 2001; Perry and Bousquet 2001; Collignon *et al.* 2002; Gamache *et al.* 2003).

Here is the intriguing question: why are levels of nucleotide diversity observed in spruce species much lower than expected on the basis of their life-history traits and observed high levels of isozyme diversity? The insensitivity of isozyme markers probably can account for the conflict between low levels of nucleotide diversity and high levels of isozyme diversity observed in spruce species (Shaw 1970). However, outbreeding species with large population sizes, such as widely distributed spruce species, are expected to have more genetic variation. This inconsistency between low levels of nucleotide diversity observed in spruce species and their life-history traits still need to be addressed. From $\theta = 4N_e\mu$ we know that a particularly low mutation rate combined with a low effective population size could result in low nucleotide diversity. *P. taeda* might be such an example (Brown *et al.* 2004). It had a particularly low mutation rate (on the order of 1.7×10^{-10}/bp/year, i.e., an order of magnitude lower than that in angiosperms) and a low effective population size (5.6×10^{5}) probably resulting from the population fluctuations during the late Pleistocene and the Holocene. An alternative explanation of the low nucleotide diversity could be the presence of repeated selective sweeps, which result in an excess of linkage disequilibrium. But this seems unlikely in conifers since current estimates of LD do not extend beyond a few hundred or thousand base pairs (Neale and Savolainen 2004).

From recent studies on humans (Voight *et al.* 2005), *Drosophila* (e.g. Haddrill *et al.* 2005) and *Arabidopsis* (Nordborg *et al.* 2005), we know that both demography and natural selection played a major role in shaping DNA sequence variation patterns. In these species, past demographic events were reconstructed based on a large number of loci in order to detect genomic areas that are under recent selection (e.g. Akey *et al.* 2002; Schaffner *et al.* 2005; Wright *et al.* 2005). Although such fine-tuned reconstruction is still out of reach in other organism, limited surveys of nucleotide diversity together with coalescent simulations still do allow the evaluation of different demographic models on spruce species.

For instance, Heuertz *et al.* (2006) used multi-locus neutrality tests and coalescent simulations to show that a single and rather severe bottleneck, a few hundred of thousand years ago, i.e. predating the Last Glacial Maximum (LGM), would be sufficient to account for the pattern of polymorphism at 22 nuclear loci in Norway spruce (*Picea abies* [L.] Karst). The presence of a

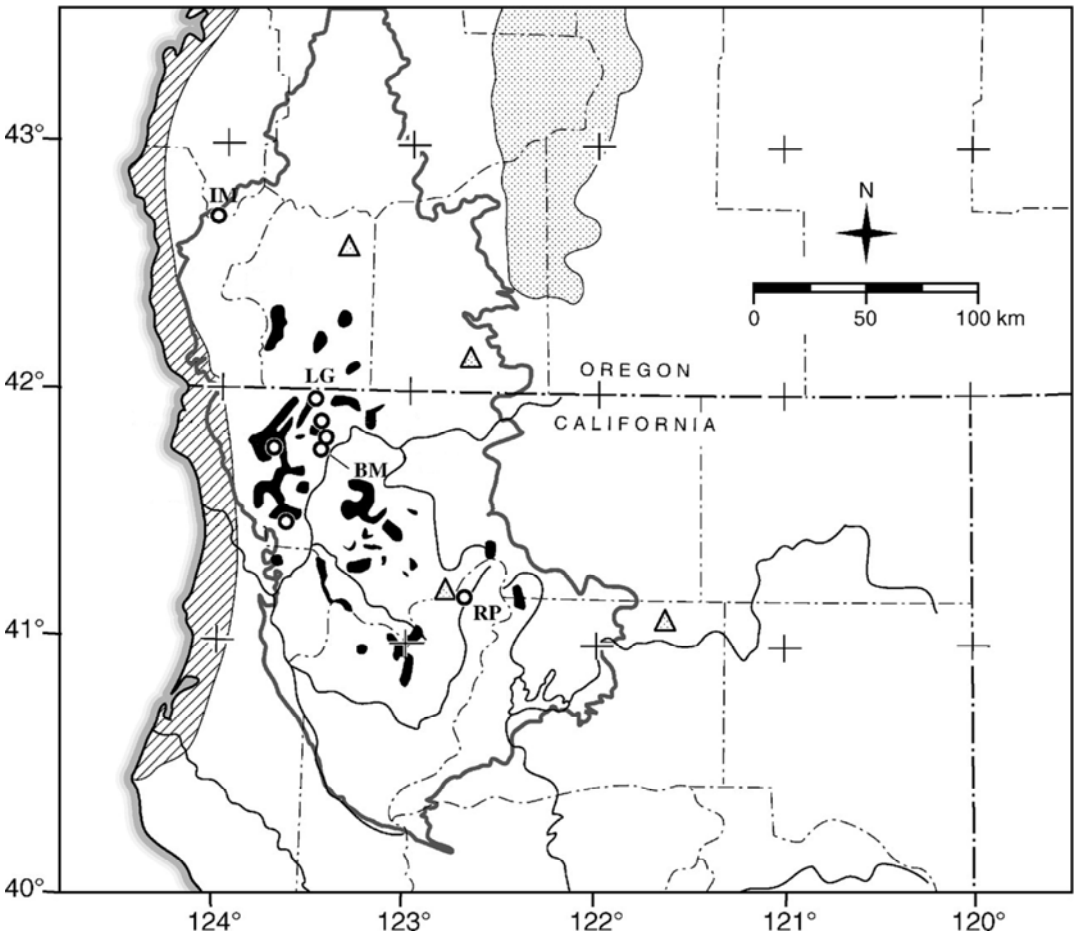
bottleneck helps explain the particularly low level of nucleotide diversity observed in Norway spruce (π_s=0.0037). Norway spruce natural range is genetically subdivided in at least three major domains: the Baltico-Nordic domain, the Alpine domain and the Carpathian domain. This distribution could be the result of the rapid postglacial re-colonization from three refuge populations located in western Russia, the Alps and the Carpathian Mountains after the Quaternary cold periods in Europe (Lagercrantz and Ryman 1990; Vendramin *et al.* 2000; Heuertz *et al.* 2006). Although demography alone is unlikely to explain the low nucleotide diversity in all coniferous species, it provides a simple explanation, at least in Norway spruce.

Genetic variation is generally assessed at two levels, namely, polymorphism within species and divergence among species, allowing inferences on different time scales. Polymorphism data allow inferences on a coalescent time scale (the time to MRCA is on the average $4N_e$ generations for a diploid). Divergence data, involving orthologous genes from different species, provide information on a much deeper time scale and have been very informative on past selection. For example, estimates of the ratio of the number of nonsynonymous mutations per nonsynonymous site (d_N) to the number of synonymous mutations per synonymous sites (d_S) based on human and chimpanzee data have indicated that positive selection ($d_N/d_S>1$) has been quite frequent in human (e.g. Nielsen *et al.* 2005). In Norway spruce, using the same approach on individual candidate genes for timing of budset, Källman *et al.* have also found the evidence of positive selection on a few of those (unpublished data). Divergence data and polymorphism data can be used jointly as a powerful approach to infer the past selection (Andolfatto 2005).

Bouillé and Bousquet (2005) detected trans-species shared polymorphism at three orthologous nuclear gene loci (amounting to a total of around 2000 bp) across Norway spruce, *P. mariana* and *P. glauca*. Because these species do not cross naturally, differ in several morphological characters and are present in distinct lineages of phylogenetic tree in the genus, the trans-species shared polymorphisms are likely to be of shared ancestry. Such an investigation of the genealogy of orthologous alleles at nuclear gene loci among distant spruce species could reveal deep coalescence, perhaps preceding species and lineage divergence. The mutation rate μ for nuclear gene loci could be also estimated by calculating the average divergence per site between species (substitution only) and then dividing by estimated divergence time.

In this project nuclear DNA sequences from Brewer spruce (*Picea breweriana* Wats.) will be collected and compared with those from corresponding loci in Norway spruce. In order to distinguish ancestral from derived variants, sequences from other genera of the *Pinaceae* will be used as outgroup. Since these two species ranges separated by thousand of kilometres, we can safely assume that both species have been isolated for a long period of time, possibly since the divergence between *Picea* species (estimated to have occurred 13-20 million years ago, i.e. 500,000-800,000 generations ago, if we assume a generation time of 25 years).

Fig. 1 The range of Brewer Spruce (in black) after Griffin and Critchfield (1972) and Waring *et al.* (1975), the locations of Brewer spruce populations (open circles) sampled for isozyme analysis (Ledig *et al.* 2005). IM=Iron Mountain; LG=Little Grayback; BM=Baldy Mountain; RP=Russian Peak.



Brewer spruce, endemic to Western North America, is a relict of the Arcto-Tertiary forest. According to the fossil record, Brewer spruce had a wide distribution in the Miocene and Pliocene (Wolfe 1964). Accelerated mountain-building and increasingly dry climates at the close of Miocene forced Brewer spruce to shrink toward the coast and higher elevations (Whittaker 1961). Brewer spruce has survived since then with a highly fragmented range. It is only found in the mountains of northwestern California and southwestern Oregon near the pacific coast (Fig. 1). The varied climate in its natural habitats indicates that Brewer spruce has an ecological amplitude that should enable it to obtain a wider and more contiguous distribution. Its sensitivity to fire seems to have restricted its range (Sawyer and Thornburgh 1977). In spite of its limited distribution, Brewer spruce has similar heterozygosity at isozyme level as Norway spruce. For example, mean expected heterozygosity (H_e) in Brewer spruce averaged 0.129 compared to 0.115 in Norway spruce (Lagercrantz and Ryman 1990; Ledig *et al.* 2005).

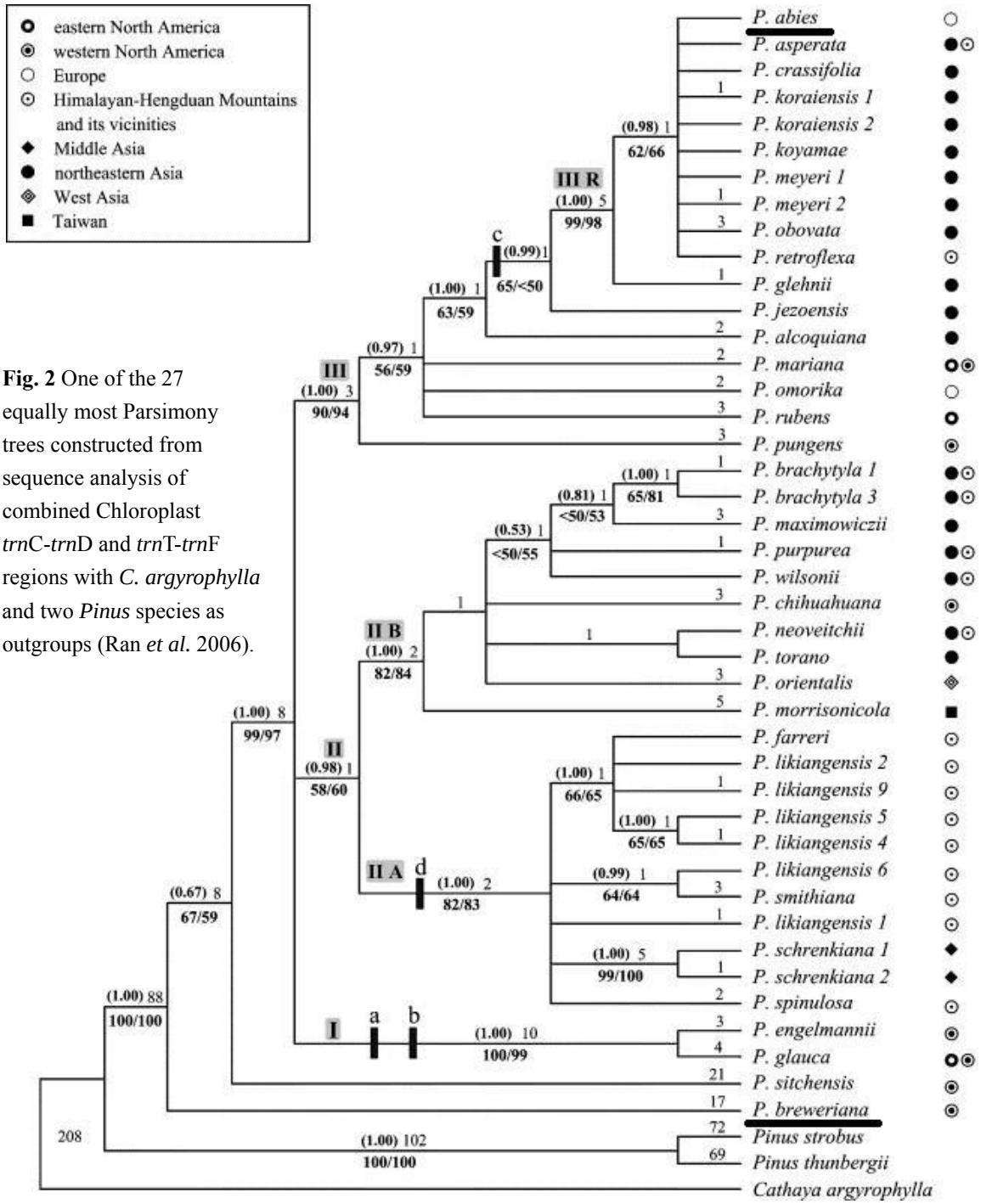


Fig. 2 One of the 27 equally most Parsimony trees constructed from sequence analysis of combined Chloroplast *trn*C-*trn*D and *trn*T-*trn*F regions with *C. argyrophylla* and two *Pinus* species as outgroups (Ran *et al.* 2006).

Phylogenies based on several molecular markers have shown that Brewer spruce was basal to the other spruces and stand alone in the genus with no close relatives (Fig. 2, Ran *et al.* 2006). Brewer spruce is so different from all other spruces that the attempts to cross Brewer spruce with other spruce species have failed (Fig. 3, Ledig *et al.* 2004). Note that, interestingly, there is some congruence between the two figures.

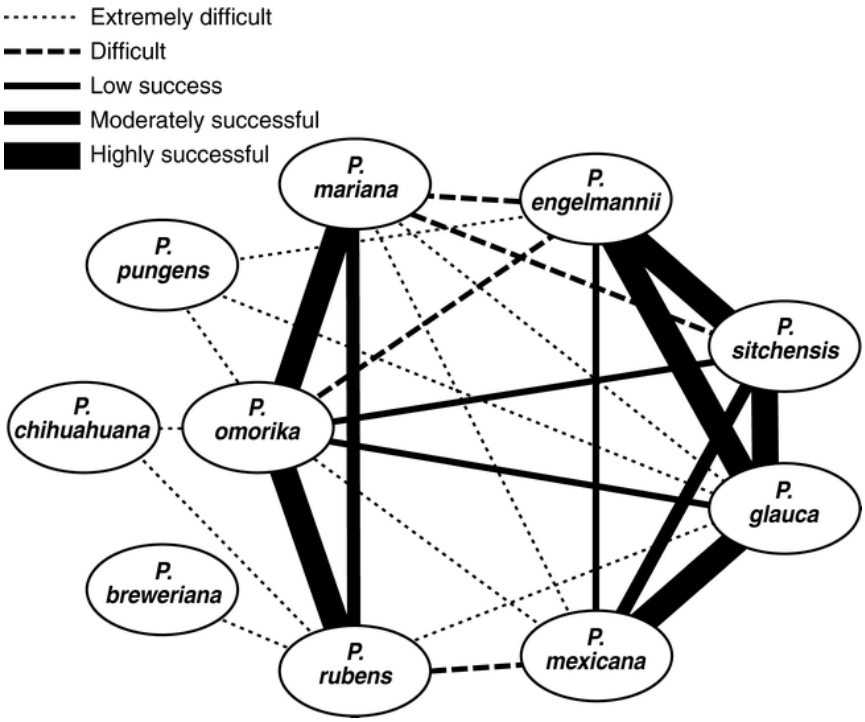


Fig. 3 Crossability among spruces of North America and *P. omorika*. Crosses that have been attempted and have consistently failed are (not necessarily with maternal parent listed first): *P. breweriana* with *P. mariana, P. rubens, P. omorika* and *P. mexicana* (Ledig *et al.* 2004).

1.3 The aim of study

The primary aim of this project is to use DNA polymorphism and divergence to reconstruct the demographic history of Norway spruce and Brewer spruce and test for the presence of ancient and recent selection at a set of candidate genes. The data generated by the project will allow us to produce less biased estimates of nucleotide polymorphism and linkage disequilibrium, two key parameters for association mapping studies, and to obtain new estimates of the mutation rate.

2 Materials and Methods

2.1. Plant Materials

Seven populations of Norway spruce and four populations of Brewer spruce were compared in this study. Norway spruce populations were sampled along a latitudinal cline within the natural distribution of the species. They were from southern Sweden, northern Sweden, Russia, Romania, Germany, Switzerland and Italy. Brewer spruce populations were from Iron Mountain, Little Grayback, Baldy Mountain and Russian Peak, which represented the natural range of the species (Fig. 1). Seed samples of each population were previously harvested on an average of 20 half-sib families and stored in a cold room.

2.2. DNA extraction, PCR amplification and sequencing

Twelve individuals were randomly selected from each population of Brewer spruce (48 individuals in total). The seed samples were soaked in water for 20 hours. The megagametophytes and embryos were dissected and separated. DNA was extracted from the megagametophyte using a modified cetyltrimethylammonium bromide (CTAB) method (Doyle & Doyle, 1990). Since the megagametophyte is maternally derived and haploid, the DNA samples were haploid and the identification of haplotypes could be unambiguous.

A total of 11 genes were tested (Table 1). Three seasonal growth cessation candidate genes were chose in Norway spruce, showing similarity with the *A. thaliana genes* GI (GIGANTEA, Fowler *et al.* 1999), ZTL (ZEITLUPE, Somers *et al.* 2000) and APRR3 (Matsushika *et al.* 2000). *Sb*16 is part of a putative gene for the 60S ribosomal protein L13a. *Sb*29 is part of a putative gene for the "No Apical Meristem" (NAM) protein fragment. *Sb*62 is part of a putative gene for the ribosomal protein L15 (Perry and Bousquet, 1998). Control loci *a priori* not involved in the photoperiod or vernalization pathways (se1358; se1364; se1390; xy225; xy1420) were selected from a pilot resequencing survey of 21 EST-based loci across 12 Norway spruce individuals (S. Degli Ivanissevich and M. Morgante, unpublished data).

The PCR amplification was conducted in a PTC-100 Programmable Thermal Controller (MJ Research, Watertown, MA) and an iCycler (Bio-Rad, Hercules, CA) with a reaction volume of 2μL of DNA template, 1.6 μL of dNTPs mix (2.5 mM each nucleotide), 0.6μM of each of the primer pair (1μM /μL each primer), 0.1 μL of Fusion DNA polymerase (Finnzyme, Espoo, Finland). The protocols of PCR amplification and the sequences of primers can be found in Appendix. The PCR products were purified using ExoSAP-IT (USB, Cleveland, OH) according to manufacturer's instructions. The reads of each individual for each locus were base-called with program PHRED and assembled together with program PHRAP, producing one contig (Ewing and Green 1998, Ewing *et al.* 1998). The sequences were then visualized and edited with program CONSED 13.0 (Gordon *et al.* 1998). All chromatograms were checked visually. A putative sequence variant was only accepted when the phred quality score exceeded 25 at that site. Re-sequencing was performed as needed to maintain the quality criterion.

The DNA sequences from corresponding loci in Norway spruce were available from Genbank and already summarized in Heuertz *et al.* (2006).

2.3 Data analysis

All the sequences of Brewer spruce were first assigned coding and non-coding regions based on the verified corresponding sequences of Norway spruce. Sequence statistics for all loci were calculated using DnaSP version 4.0 (Rozas and Rozas 1999). Insertion-deletion mutations and sites with missing data were excluded from all estimates. For each locus, haplotype and nucleotide diversity and the genetic structure of populations were estimated. Tajima's D test and Fay and Wu's H test were used to investigate whether the loci were evolving neutrally. The loci were also checked for linkage disequilibrium. The global fixation index (F_{st}) was used to compare the level of differentiation among populations.

3 Results

3.1 Nucleotide variation in Brewer spruce

Sequence variation was obtained for all 11 loci in an average of 42 megagametophytes. For ZTL and APRR3, high quality sequence data were obtained from only 25 and 26 individuals, respectively. A total of 7310 bp were aligned over the 11 genes, of which a half was coding sequence. A total of 22 segregating sites were identified, of which 5 were singletons and 17 were parsimony-informative sites. This corresponds to 1 SNP every 332 bp. Loci se1364, *Sb*16, GI and APRR3 were non-polymorphic. Statistics of sequence variation are summarized in Table 1. Total nucleotide diversity π_t varied from 0 to 3.88×10^{-3} (average $\pi_t=0.78\times10^{-3}$) and nonsynonymous nucleotide diversity π_a varied from 0 to 1.59×10^{-3} (average $\pi_a=0.25\times10^{-3}$). Silent nucleotide diversity, including synonymous and non-coding positions, was between 0 and 9.45×10^{-3} (average $\pi_s=2.01\times10^{-3}$).

Table 1

Nucleotide variation and neutrality tests in 11 *Picea breweriana* loci sequenced across four populations

		Nucleotide diversity													
		Total				Non-synonymous sites				Silent sites				Neutrality tests	
Locus	***n***	***L***	***S* (singl.)**	θ_{wt}	π_t^a	***L***	***S***	θ_{wa}	π_a	***L***	***S***	θ_{ws}	π_s	***D***	***H***
se1358	48	485	4 (1)	1.86	1.39	381	1	0.59	0.21	102	3	6.65	5.81	-0.58	-1.55
se1364	46	547	0	0	0	221	0	0	0	319	0	0	0	—	—
se1390	48	559	4 (3)	1.61	0.95	373	2	1.21	0.22	185	2	2.43	2.40	-0.95	0.34
xy225	47	267	1(1)	0.85	0.16	50	0	0	0	155	1	1.05	0.20	-1.09	0.04
xy1420	46	565	6 (0)	2.42	3.88	391	2	1.17	1.59	166	4	5.47	9.45	1.57	0.07
***Sb*16**	45	795	0	0	0	190	0	0	0	611	0	0	0	—	—
***Sb*29**	44	556	2 (0)	0.83	0.73	379	1	0.61	0.72	174	1	1.32	0.75	-0.22	0.34
***Sb*62**	43	576	2 (0)	0.80	0.70	217	0	0	0	356	2	1.30	1.13	-0.25	-2.82
GI	48	760	0	0	0	252	0	0	0	508	0	0	0	—	—
ZTL	25	1274	3 (0)	0.62	0.78	846	0	0	0	426	3	1.87	2.35	0.64	0.75
APRR3	26	926	0	0	0	384	0	0	0	542	0	0	0	—	—
Total	—	7310	22 (5)	—	—	3685	6	—	—	3543	16	—	—	—	—
Average	42	665	—	0.82	0.78	—	—	0.33	0.25	—	—	1.83	2.01	-0.13	-0.40

n, sample size; *L*, length in base pairs, no gaps; *S* (singl.), number of segregating sites (number of singletons); *D*, Tajima's D-statistic; *H*, Fay and Wu's H-statistic. Nucleotide variation estimates (θ_w and π) are $\times10^3$.

3.2 Statistic tests of neutrality in Brewer spruce

Tajima's D-value varied between -1.09 and 1.57. Fay and Wu's H-value was between -2.82 and 0.75. Mean values of both D- and H-statistics were negative, with values of -0.13 and -0.40, respectively. All D- and H-values were nonsignificant. (Table 1)

3.3 Population structure and linkage disequilibrium in Brewer spruce

The global fixation index (F_{st}) was used to indicate the level of differences among populations. For 7 of 11 loci, F_{st} had a value of 0. The F_{st} values of loci se1358, se1420, *Sb*29 and *Sb*62 varied from 0.020 to 0.138 and none was significant. Linkage disequilibrium, the non-random pattern of association between variants of different polymorphic sites within a population, was not estimated due to few polymorphic sites, low level of variation and short DNA sequences.

3.4 Comparison of nucleotide variation between Norway spruce and Brewer spruce

Table 2

Nucleotide variation and Tajima's D statistics in 11 loci sequenced across 7 populations of *Picea abies* and 4 populations of *Picea breweriana.*

S (singl.), number of segregating sites (number of singletons).Nucleotide diversity estimates (θ_w and π) are $\times 10^3$.

Locus	Species	*S* (singl.)	θ_{wt}	π_t^a	Tajima's D
se1358	*P. abies*	8 (3)	4.01	2.88	-0.78
	P. breweriana	4 (1)	1.86	1.39	-0.58
se1364	*P. abies*	4 (1)	1.64	1.32	-0.44
	P. breweriana	0	0	0	—
se1390	*P. abies*	13 (4)	5.89	4.83	-0.54
	P. breweriana	4 (3)	1.61	0.95	-0.95
xy225	*P. abies*	6 (3)	6.47	3.42	-1.21
	P. breweriana	1	0.85	0.16	-1.09
xy1420	*P. abies*	20 (4)	7.86	6.81	-0.43
	P. breweriana	6 (0)	2.42	3.88	1.57
***Sb*16**	*P. abies*	25 (14)	7.45	4.21	-1.43
	P. breweriana	0	0	0	—
***Sb*29**	*P. abies*	19 (3)	8.44	7.24	-0.45
	P. breweriana	2 (0)	0.83	0.73	-0.22
***Sb*62**	*P. abies*	15 (3)	6.77	3.14	-1.70
	P. breweriana	2 (0)	0.80	0.70	-0.25
GI	*P .abies*	7 (3)	2.04	1.28	-1.00
	P. breweriana	0	0	0	—
ZTL	*P. abies*	33 (13)	6.33	4.16	-1.19
	P. breweriana	3 (0)	0.62	0.78	0.64
APRR3	*P. abies*	10 (1)	2.61	3.88	1.44
	P. breweriana	0	0	0	—
Total	*P. abies*	164(55)	—	—	—
	P. breweriana	22 (5)	—	—	—
Average	*P.abies*	—	5.09	3.68	-0.72
	P.breweriana	—	0.82	0.78	-0.13

The comparison of nucleotide variation between Norway spruce and Brewer spruce is summarized in Table 2. A total of 164 segregating sites were identified in Norway spruce for 11 loci, of which 55 were singletons and 109 were parsimony-informative sites. Comparing the number of segregating sites, Norway spruce had 7.45 times more segregating sites than Brewer spruce. All 11 loci were polymorphic in Norway spruce but only 7 of them were polymorphic in Brewer spruce. Loci se1364, GI and APRR3 were non-polymorphic in Brewer spruce and had relatively low level of variation in Norway spruce. Locus *Sb*16 was also non-polymorphic in Brewer spruce but it had 25 segregating sites (the second highest) in Norway spruce. Locus ZTL had 33 segregating sites (the highest) in Norway spruce but only 3 in Brewer spruce. Similar to loci *Sb*16 and ZTL, loci *Sb*29 and *Sb*62 also had little variation in Brewer spruce but relatively high level of polymorphism in Norway spruce.

Four polymorphic sites were shared by Norway spruce and Brewer spruce. They were one at locus se1358 (position 365), one at locus ZTL (position 834) and two at locus *Sb*62 (position 110 and 231).

The range of total nucleotide diversity π_t was (1.28×10^{-3}, 7.24×10^{-3}) in Norway spruce and (0, 3.88×10^{-3}) in Brewer spruce; the range of θ_w was (1.64×10^{-3}, 8.44×10^{-3}) in Norway spruce and (0, 2.42×10^{-3}) in Brewer spruce. Comparing the average of π_t and θ_w, Norway spruce had 4.71 and 6.21 times more than Brewer spruce, respectively.

Mean values of Tajima's D for Norway spruce and Brewer spruce were -0.72 and -0.13, respectively. Both were negative. The comparison of Tajima's D values of 11 loci between Norway spruce and Brewer spruce is shown in Fig. 4. The figure shows that the plots of Brewer spruce are distributed near 0 on both sides of origo and the plots of Norway spruce are more gathered on the left-side axis.

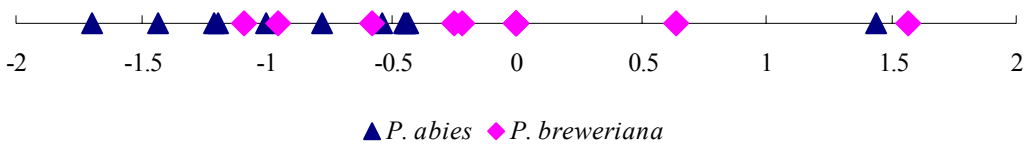


Fig. 4 Plot of Tajima's D values of 11 loci for 7 populations of *Picea abies* and 4 populations of *Picea breweriana*

4 Discussion

4.1 Population genetics parameters in Brewer spruce

This study showed that Brewer spruce had a very low level of nucleotide variation (π_s=0.0020, θ_w=0.0018) and little population differentiation. Linkage disequilibrium could not even be estimated due to too few polymorphic sites and short DNA sequences.

The average level of silent nucleotide diversity in Brewer spruce is among the lowest in conifers (π_s=0.0020), compared to the average π_s in *P. sylvestris* (π_s≈0.00648 across 16 genes, Pyhäjärvi *et al.* 2007), *P. taeda* (π_s=0.0064 across 19 wood-production candidate genes, Brown *et al.* 2004; π_s=0.0079 across 18 drought stress candidate genes, González-Martínez *et al.* 2006) and *P. abies* (π_s=0.0039 across 21 genes, Heuertz *et al.* 2006). This low level of nucleotide diversity in Brewer spruce is consistent with those observed in conifers, including other spruces. It also supports the contention that Brewer spruce is the least diverse spruce species in western North America (Ledig *et al.* 2005).

The isozyme diversity in Brewer spruce is similar to that in Norway spruce (Ledig *et al.* 2005; Lagercrantz and Ryman 1990) but the average π_s in Brewer Spruce is two-fold lower than that in Norway spruce. This is all the most striking as isozyme markers are not as sensitive as DNA markers. The isozymes are variants of the same enzyme with different amino acid substitutions. Amino acid substitutions change the electric charge of the enzyme so isozymes can be separated by gel electrophoresis. However, in theory only ca. 30% of the possible amino acid substitutions result in a difference in charge (Shaw 1970). So the isozyme markers may not detect all genetic variation. And yet the level of variation was much higher at the isozymes than at the nucleotide levels. At that stage we do not have any satisfying explanation for the discrepancy in variability at the isozymes and nucleotide levels.

Little population differentiation among four Brewer spruce populations was shown in this study. This is conflict with that Brewer spruce may be characterized as moderately structured (Ledig *et al.* 2005). Ledig *et al.* (2005) did the isozyme study on ten populations covering the entire species distribution ranges. Samples used in this study were only from four of them so lack of samples can partly account for not being able to detect the population structure in Brewer spruce. Another possible reason is that the DNA nuclear markers used in this study might not be very informative about the population structure but the inaccuracy of isozyme markers should always call for caution. More samples from more populations are required for further investigation in population structure in Brewer spruce.

4.2 Shared polymorphism among spruce species

Bouillé and Bousquet (2005) detected trans-species shared polymorphisms at loci *Sb*16, *Sb*29 and *Sb*62 across Norway spruce, *P. mariana* and *P. glauca*, which were likely to be of shared ancestry. In this study trans-species shared polymorphisms across Norway spruce and Brewer spruce were also observed at locus *Sb*62 (position 110 and 231), and loci se1358 (position 365) and ZTL (position 834). Since recent lateral gene flow between Brewer spruce and Norway spruce is almost impossible, the trans-species shared polymorphisms detected in this study are very likely to be of shared ancestry. This strongly supports the contention that allele coalescence preceded species divergence and suggests that the ancestral population size of Brewer spruce was large. In general, divergence among spruce species seems very low: most genes analyzed in the study showed very few segregating sites between Brewer spruce and Norway spruce.

Table 3 shows nucleotide diversity π for loci *Sb*16, *Sb*29 and *Sb*62 in Norway spruce from this study and the study by Bouillé and Bousquet (2005). The results for the same loci were quite different. This is probably due to differences in individual sampling strategies. But they all showed that locus *Sb*16 harboured the lowest nucleotide diversity and locus *Sb*62 harboured the highest. More loci will need to be investigated in order to compare adequately the diversity residing in the various species.

Table 3

Nucleotide diversity π for three nuclear gene loci in *P. abies* and *P. breweriana*

Species	*Sb*16 (n)	*Sb*29 (n)	*Sb*62 (n)
P.abies *	0 (7)	0.0134 (5)	0.0064 (5)
P.abies	0.0042 (48)	0.0072 (52)	0.0031 (41)
P.breweriana	0 (45)	0.0007 (44)	0.0007 (43)

Note: * means data from the study of Bouillé and Bousquet (2005); n, haploid sampling size.

4.3 Evolutionary history and demographic events in Brewer spruce

The mean values of Tajima's D and Fay and Wu's H for Brewer spruce were close to zero so no departure from the standard neutral model was detected. Hence, the particularly low level of nucleotide variation in Brewer spruce might simply be the consequence of a long-term limited population size. Interestingly, these results are in contrast to those obtained in Norway spruce where mean Tajima's D and Fay and Wu's H values across loci suggested the presence of a severe and ancient bottleneck. The distribution of Brewer spruce was wide in the Pliocene and Miocene, at least as far east as Idaho and Nevada, north to central Oregon, and south to central California (Wolfe 1964). It was reduced by increasingly dry climates in the interior West, especially when mountain building accelerated at the close of Miocene, and then shrunk toward the coast and higher elevations (Whittaker 1961). The regions where Brewer spruce has survived escaped submergence, extensive glaciation and volcanism in their paleohistory and have been relatively stable for at least 100 million years (Whittaker 1961; Irwin 1966; Villa-Lobos 2003). Ledig *et al*. (2005) showed that Brewer spruce populations had not been bottlenecked recently and our data concurred to that conclusion.

In summary, these results, together with those obtained in Norway spruce, show the potential of population history reconstruction based on nucleotide variation.

5 Acknowledgements

I would like to thank my supervisor Martin Lascoux for the opportunity of doing my degree project in the department of Evolutionary Functional Genomics. I am most grateful to Niclas Gyllenstrand and Thomas Källman for helping me with practical and theoretical problems in both molecular and computer lab. Thank everyone who has helped me with this project.

6 References

Akey, J. M.; Zhang, G.; Zhang, K.; Jin, L. and Shriver, M. D. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Research* 12: 1805-1814.

Andolfatto, P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437: 1149-1152.

Bouillé, M. and Bousquet, J. 2005. Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer *Picea* (Pinaceae): implications for the long-term maintenance of genetic diversity in trees. *American Journal of Botany* 92(1): 63-73.

Boyle, T. B. and Morgenstern, E. K. 1987. Some aspects of the population structure of black spruce in central New Brunswick. *Silvae Genetica* 36: 53-60.

Brown, G. R.; Gill, G. P.; Kuntz, R. J.; Langley, C. H. and Neale, D. B. 2004. Nucleotide variation and linkage disequilibrium in loblolly pine. *Proc. Natl. Acad. Sci*. USA 101: 15255-15260.

Collignon, A. M.; Van de Sype, H. and Favre, J. M. 2002. Geographical variation in random amplified DNA and quantitative traits in Norway spruce. *Canadian Journal of Forest Research* 32: 266-282.

Doyle, J. J. and Doyle, J. L. 1990. Isolation of plant DNA from fresh tissue. *Focus* 12: 13-15.

Dvornyk, V.; Sirviö, A.; Mikkonen, M. and Savolainen, O. 2002. Low nucleotide diversity at the pal1 locus in the widely distributed *Pinus sylvestris. Molecular Biology and Evolution* 19: 179-188.

Everett, T. H. 1981. The New York Botanical Garden illustrated encyclopedia of horticulture. Vol. 9. Par-Py. New York: Garland Publishing.

Ewing, B.; Hillier, L. and Wendl, M. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research* 8: 175-185.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using Ptued. Ⅱ. Error probabilities. *Genome Research* 8: 186-194.

Fay, J. and Wu, C. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405-1413.

Fowler, S.; Lee, K.; Onouchi, H.; Samach, A.; Richardson, K. *et al.* 1999. GIGANTEA: a circadian clock-controlled gene that regulates photoperiodic flowering in *Arabidopsis* and encodes a protein with several possible membrane-spanning domains. *The EMBO Journal* 18: 4679-4688.

Furnier, G. R.; Stine, M.; Mohn, C. A. and Clyde, M. A. 1991. Geographic patterns of variation in allozymes and height growth in white spruce. *Canadian Journal of Forest Research* 21: 707-712.

Gamache, I.; Jaramillo-Corea, J. P.; Payette, S. and Bousquet, J. 2003. Diverging patterns of mitochondrial and nuclear DNA diversity in subarctic black spruce: imprint of a founder effect associated with postglacial colonization. *Molecular Ecology* 12: 891-901.

García-Gil, M. R.; Mikkonen, M. and Savolainen, O. 2003. Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris. Molecular Ecology* 12: 1195-1206.

González-Martínez, S. C.; Ersoz, E.; Brown, G. R.; Wheeler, N. C. and Neale, D. B. 2006. DNA sequence

variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. *Genetics* 172: 1915-1926.

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

Griffin, J. R. and Critchfield, W. B. 1972. The distribution of forest trees in California. USDA Forest Service Research Paper PSW-82. Pacific Southwest Forest and Range Experiment Station, Berkeley, California, USA

Haddrill, P. R.; Thornton, K. R.; Charlesworth, B. and Andolfatto, P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Research* 15: 790-799.

Hamrick J. L. and Godt, M. J. W. 1996. Effects of life history traits on genetic diversity in plant species. *Philosophical Transactions of the Royal Society of London, B, Biological Sciences* 351: 1291-1298.

Hamrick J. L.; Godt, M. J. W. and Sherman-Broyles, S. L. 1992. Factors influencing levels of genetic diversity in woody plant species. *New Forests* 6: 95-124.

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

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

Ingvarsson, P. K. 2005. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). *Genetics* 169: 945-953.

Irwin, W. P. 1966. Geology of Klamath Mountains Province. Geology of northern California. *California Division of Mines and Geology Bulletin* 190: 19-38.

Isabel, N.; Beaulieu, J. and Bousquet, J. 1995. Complete congruence between gene diversity estimates derived from genotypic data at enzyme and random amplified polymorphic DNA loci in black spruce. *Proceedings of the National Academy of Sciences, USA* 92: 6369-6373.

Jaramillo-Correa, J. P.; Beaulieu, J. and Bousquet, J. 2001. Contrasting evolutionary forces driving population structure at expressed sequence tag polymorphisms, allozymes and quantitative traits in white spruce. *Molecular Ecology* 10: 2729-2740.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Kingman, J. F. C. 1982. The coalescent. *Stochastic Processes and their application* 13: 235-248.

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

Ledig, F. T. 1998. Genetic variation in *Pinus*. *In* D. M. Richardson [ed.], Ecology and biogeography of *Pinus*, 251-280. Cambridge University Press, Cambridge, UK

Ledig, F. T.; Hodgskiss, P. and Johnson, D. 2005. Genetic diversity, genetic structure, and mating system of brewer spruce (Pinaceae), a relict of the Arcro-Tertiary forest. *American Journal of Botany* 92 (12): 1975-1986.

Ledig, F. T.; Hodgskiss, P.; Krutovskii, K.; Neale, D. and Eguiluz-Piedra, T. 2004. Relationships among the spruces (*Picea*, Pinaceae) of southwestern North America. *Systematic Botany* 29 (2): 275-295.

Matsushika, A.; Makino, S.; Kojima, M. and Mizuno, T. 2000. Circadian waves of expression of the APRR1/TOC1 family of pseudo-response regulators in *Arabidopsis thaliana*: Insight into the plant circadian clock. *Plant and Cell Physiology* 41: 1002-1012.

Müller-Starck, G. 1995 Genetic variation in high elevated populations of Norway spruce (*Picea abies* (L.) Karst.) in Switzerland. *Silvae Genetica* 44: 356-362.

Neale, D. B. and Savolainen, O. 2004. Association genetics of complex traits in conifers. *Trends in Plant Science* 9: 325-330.

Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* 3(5): 418-426.

Nielsen, R.; Bustamante, C.; Clark, A. G.; Glanowski, S.; Sackton, T. B. *et al.* 2005. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. *PLoS Biology* 3(6): e170.

Nordborg, M.; Hu, T.-T.; Ishino, Y.; Jhaveri, J.; Toomajian, C. *et al.* 2005. The Pattern of Polymorphism in *Arabidopsis thaliana*. *PLoS Biology* 3(7): e196.

Perry, D. J. and Bousquet, J. 1998. Sequence-tagged-site (STS) markers of arbitrary genes: development, characterization and analysis of linkage in black spruce. *Genetics* 149: 1089-1098.

Perry, D. J. and Bousquet, J. 2001. Genetic diversity and mating system of post-fire and post-harvest black spruce: an investigation using codominant sequence-tagged-site (STS) markers. *Canadian Journal of Forest Research* 31: 32-40.

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.

Ran, J.-H.; Wei, X.-X. and Wang, X.-Q. 2006. Molecular phylogeny and biogeography of *Picea* (Pinaceae): implications for phylogeographical studies using cytoplasmic haplotypes. *Molecular Phylogenetics and Evolution* 41: 405-419.

Rozas, J. and Rozas, R. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* 15: 174-175.

Sawyer, J. O. and Thornburgh, D. A. 1977. Montane and subalpine vegetation of the Klamath Mountains. *In* Barbour, M. G. and Major, J. [ed.], Terrestrial vegetation of California, 699-732. John Wiley, New York.

Schaffner, S. F.; Foo, C.; Gabriel, S.; Reich, D.; Daly, M. J. *et al.* 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research* 15: 1576-1583.

Schmid, K. J.; Ramos-Onsins, S.; Ringys-Beckstein, H.; Weisshaar, B. and 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: 1601-1615.

Schmidt, P. A. 1989. Beitrag zur Systematik und Evolution der Gattung *Picea* A. Dietr. *Flora* (*Jena*) 182: 435-461.

Shaw, C. R. 1970. How many genes evolve? *Biomedical Genetics* 4: 275-283.

Somers, D. E.; Schultz, T. F.; Milnamow, M. and Kay, S. A. 2000. ZEITLUPE encodes a novel clock-associated PAS protein from Arabidopsis. *Cell* 101: 319-329.

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

Tiffin, P. and Gaut, B. S. 2001. Sequence diversity in the tetraploid *Zea perennis* and the closely related diploid *Z. diploperennis*: insights from four nuclear loci. *Genetics* 158: 401-412.

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. *Genome* 43: 68-78.

Villa-Lobos, J. 2003. Klamath-Siskiyou: jewel of the Pacific Coast. *Plant Talk* 31: 29-33.

Voight, B. F.; Kudaravalli, S.; Wen, X. and Pritchard, J. K. 2006. A Map of Recent Positive Selection in the Human Genome. *PLoS Biology* 4(3): e72.

Waring, R. H.; Emmingham, W. H. and Running, S. W. 1975. Environmental limits of an endemic spruce, *Picea breweriana*. *Canadian Journal of Botany* 53: 1599-1613.

Whittaker, R. H. 1961. Vegetation history of the Pacific Coast states and the "central" significance of the Klamath Region. *Madroño* 16: 5-23.

Wolfe, J. A. 1964. Miocene floras from Fingerrock Wash southwestern Nevada. United States Geological Survey Professional Paper no. 454-N

Wright, J. W. 1955. Species crossability in spruce in relation to distribution and taxonomy. *Forest Science* 1: 319-349.

Wright, S. I.; Bi, I. V.; Schroeder, S. G.; Yamasaki, M.; Doebley, J. F.; McMullen, M. D. and Gaut, B. S. 2005. The effects of artificial selection of the maize genome. *Science* 308: 1310-1314.

Appendix

Table 1 Protocols used in amplification of different loci in the genome of Brewer spruce

Locus	PCR Protocol							
	Start	Times of Cycle	Cycle				End	
xy225, se1364		30		60°C	10s		1 min	10 min
se1358, se1390	98°C	35	98°C	60°C	15s	72°C	1 min 30s	72°C 7 min
xy1420	30 s	30	10 s	55°C	15s		30s	5 min
ZTL		35		58°C	10s		2 min	5 min

Locus	PCR Protocol											
	Start	Times	Touch-down			Times	Cycle				End	
GI	98°C	4	98°C	65°C 10s -1°C/cycle	72°C	1 min 30 s	30	98°C	60°C 10 s	72°C	1 min 30 s	72°C 5 min
*Sb*16,*Sb*29 *Sb*62	30 s	5	10 s	62°C 10s -1°C/cycle		30s	29	10 s	57°C 10 s		30 s	2 min

Table 2 Primers used in amplification and sequencing of different loci in the genome of Brewer Spruce

Locus	**Primer name**	**Primer sequence 5'-3'**
se1358		TGGCAGCTCACGGACTATGA
		CAAAACTGGTGCAACTGCCG
se1364	1364f	CCGGAACAGATGGAAGTGCT
	1364r	CCTTCAGTTGCTGTCCCCACC
se1390		GCAAGGATTAATGCCCACCAC
		AGATCCGTCCAGCACAAAGC
xy225	225f	AAGGAGGCTGGGCTTTACAG
	225r	CGGGGCAAGACGAATACAT
xy1420	1420U*	CAAGTCGTTCGCGCCGCTGGTGA
	1420R*	AGCACAATTACAGTGGCGTCGTG
***Sb*16**	sb16fwd	GTTCCGCCACCATATGAC
	sb16rev	GCTCATTCAGCTACAAAAGC
***Sb*29**	sb29fwd	AGCGGCATTGAACAGAGTAAC
	sb29rev	AATGGAAATGAAGGCAGACTC
***Sb*62**	sb62fwd	TGAGATCCGTGGCTGAAGAG
	sb62rev	GATAACGCCGGAGAGATAGAG
GI	G5	ATACAAGTCCCGCATGGCTGTTAT
	G6	CAGGCAAGGCAATGGCAGAAGGGCTAT
ZTL	Z1466$^{\nabla}$	TTACAGTTGGAGGGGCAGTTGAACC
	Z1248	TCACGGTTGACACCTAGAGA

	Z2492	AGGGCGTAATCAAGCAGCACAATCT

Note: * means amplification primer only, ▽ means sequencing primer only, primers without markings have been used for both amplification and sequencing.