Conifer Evolution, from Demography and Local Adaptation to Evolutionary Rates

Examples from the Picea genus

JUN CHEN
Dissertation presented at Uppsala University to be publicly examined in Lindahlsalen, EBC, Norbyvägen 18A, Uppsala, Friday, September 14, 2012 at 10:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

Evolutionary process can be inferred at three different levels: the species level, the population level and the molecular level. In this thesis, I applied approaches at these three levels and aimed to get a comprehensive picture of conifer evolution, from speciation and demography to geographic variation and local adaptation, and then to the molecular evolution of proteins and small regulatory RNAs.

Spruce species have been observed to possess a large number of trans-species shared polymorphisms. Using an “Isolation with migration” model, we found that the large effective population size of spruce retained these shared polymorphisms, inheriting them from the common ancestor. Post-divergence gene flow only existed between *Picea abies* and *P. glauca*, and between *P. wilsonii* and *P. schrenkiana*. The combination of Tajima’s *D* and Fay & Wu’s *H* at most of loci suggested an ancient and severe bottleneck for most species except *P. breviana*.

Furthermore, I investigated the effect of local selection in two parallel clines, which is one of the major forces that can cause divergence or even speciation. The timing of bud set and growth cessation was found correlated with latitude in populations of *P. abies* and *P. obovata*. Using allele frequency spectrum analyses we identified three genes under local selection in both species including two circadian-clock genes *GI* and *PRR7*, and one photoperiodic gene *FTL2*. This indicated that parallel evolution could occur through groups of genes within related pathways. Clinal variation at expression level provided stronger evidence of selection in *FTL2*, which has previously been associated with bud set in *P. abies*.

Finally we focused on the molecular evolution of mRNA and small regulatory RNAs in *P. abies*. With the help of Next-Generation sequencing, we have achieved in spruce the first *de novo* assembly of the needle transcriptome and a preliminary characterization of sRNA populations. Along with features common in plants, spruce also exhibited novelities in many aspects including lower substitution rate and protein evolutionary rate, dominance of 21-nt sRNA, and a large proportion of *TIR-NBS-LRR* genes as sRNA sources and targets.

Keywords: Speciation, Demographics, clinal variation, convergent evolution, transcriptome, small regulatory RNA

*Jun Chen, Uppsala University, Department of Ecology and Genetics, Plant Ecology and Evolution, Norbyvägen 18 D, SE-752 36 Uppsala, Sweden.*

© Jun Chen 2012

ISSN 1651-6214
ISBN 978-91-554-8411-8
urn:nbn:se:uu:diva-177482 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-177482)
"The way was long, and wrapped in gloom did seem,  
As I urged on to seek the hidden truth."

Qu, Yuan
List of papers

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


Reprints were made with permission from the publishers.
Contents

1 Introduction .................................................................................................... 9
  1.1 Speciation, Geographic Variation and Clines ..................................... 10
     1.1.1 Ecology and phylogeny of spruce ............................................. 10
     1.1.2 Speciation and population history ............................................. 12
     1.1.3 Identifying local adaptation along a cline ................................ 16
  1.2 Molecular evolution of Norway spruce ............................................... 18
     1.2.1 Transcriptome ............................................................................. 18
     1.2.2 Small noncoding RNAs and their regulatory role in plants ... 20

2 Results and discussion ................................................................................. 23
  2.1 Estimate of speciation history of spruce based on multilocus patterns - Paper I and II ................................................................. 23
     2.1.1 Nucleotide diversity and gene flow ........................................... 24
     2.1.2 Demographic history of eight spruce species ......................... 24
  2.2 Clinal variation and local selection - Paper III and IV ....................... 25
     2.2.1 Latitudinal gradient patterns in phenology ......................... 26
     2.2.2 Population genetic structures .................................................... 26
     2.2.3 Allele frequency spectrum analysis for local adaptation .......... 27
     2.2.4 Differentiation in gene expression ............................................ 29
     2.2.5 Parallel evolution and the selection on Gigantea .................... 29
  2.3 Norway spruce transcriptome and regulatory small RNAs - Paper V and VI ................................................................. 31
     2.3.1 Norway spruce needle transcriptome sequencing .................... 31
     2.3.2 Gene expression differences ...................................................... 31
     2.3.3 Comparative molecular evolution in conifers ......................... 32
     2.3.4 Norway spruce small regulatory RNA sequencing .................. 34
     2.3.5 MicroRNA prediction ............................................................... 34
     2.3.6 NBS-LRR gene family and relevant siRNAs ......................... 35

3 Conclusion ................................................................................................... 37

4 Svensk sammanfattning ............................................................................... 39

5 Acknowledgements ..................................................................................... 41

References ........................................................................................................ 42
1. Introduction

"Nothing in evolution makes sense except in the light of population genetics"
M. Lynch (2007)

The evolutionary process can be inferred at three different levels: at the species level, at the population level or at the molecular level. The first level is the domain of phylogenetics and systematics, while the second and third levels are the domains of population genetics and molecular evolution, respectively. All three levels have often been apprehended separately and led to multifaceted, and sometimes divergent, perceptions of the evolutionary process. For example, the inherent stochasticity of the evolutionary process which is one of the great obsessions of population geneticists, and with which we will grapple in the work presented here, was until recently largely neglected in systematics. And while molecular biologists have tended to embrace gene regulation – rather than structural changes – as the backbone of evolutionary changes, population geneticists have been more reserved (HOEKSTRA AND COYNE, 2007; STERN AND ORGOGOZO, 2009). In spite of obvious differences in time and spatial scales between the three levels, it is also evident that the three levels are highly interconnected. A first unifying factor is the environment since, at all three levels, the outcome of the evolutionary process is strongly contingent to the biotic and physical environments in which the organisms live. Incidentally, the organisms thereby also define the environment in which they live. A second unifying factor is of course genes, and perhaps more importantly genes in populations. Notwithstanding recent findings of an heritable contribution of epigenetic processes to quantitative traits (e.g. JOHANNES et al., 2009), genes remain the main evolutionary unit.

The present thesis is an attempt to look at the evolutionary process in spruce species at all three levels and, when possible, to try to highlight the connection between the different levels. The first two studies (Papers I and II) tackle aspects of speciation with population genetics tools. As we shall see spruce species sometimes still exchange genes and when they do not they still look surprisingly young if their age is measured on the right timescale. The next two studies (Papers III and IV) are related to adaptive variation and firmly rooted in population genetics. Yet they focus on specific genes and compare two species thereby constituting a natural bridge with the last two manuscripts (Papers V and VI) that constitute some initial forays in the genomic world. Yet again, even in this last group of studies, there is a connection with the species and population levels as divergence data are used to infer genome–wide patterns of selection.
1.1 Speciation, Geographic Variation and Clines

While it is generally accepted that evolution occurs by "descent with modification" and species originate from other species, the details of the speciation process are still hotly debated and, altogether, remain poorly understood. For example, can species evolve in sympatry? How are species maintained in the face of gene flow? Is natural selection necessary for speciation to take place or can speciation occur simply through mutation and random drift? If natural selection does indeed play an overwhelming role in evolution as argued by HAHN (2008), how then is genetic diversity maintained? In the present thesis I am going to address some of these questions using spruce species as an example.

1.1.1 Ecology and phylogeny of spruce

Distribution and ecology

The genus Picea currently consists of around 34 species. They are usually tall, sometimes large, evergreen trees with a monopodial, straight, and columnar trunk (FARJON, 1990). Spruce grows from the polar circle down to just south of the Tropic of Cancer (see Figure 1.1). Some species are widely distributed across the northern conifer forest biome: P. abies and its close relative P. obovata grow from the Atlantic coast of Norway eastwards to the Sea of Okhotsk, and the combined ranges of P. glauca and P. mariana cover most of Canada and Alaska. At southern latitudes spruce species are generally confined to small and scattered populations at high altitude: P. martinezii and P. chihuahuana in Mexico or P. morissonicola in Taiwan. The Tertiary relict P. breweriana also grows in scattered montane populations in Northern California. Between these two extremes we find species with intermediate distribution areas. Some are coastal, like P. sitchensis, or montane to subalpine, like P. schrenkiana, P. likiangensis, P. willsonii, and P. purpurea in the Sino-Himalayan area. Northern and southern range limits could be caused primarily by the deleterious effects of cold and hot air temperatures, respectively, on individual tree growth – although their importance has been disputed in P. mariana (BONAN AND SIROIS, 1992). Likely a combination of biotic and abiotic factors, such as the frequency of outbreaks of insect defoliators or pathogenic fungi, the duration of the growing season and moisture supply, all play a role in setting these limits.

Phylogeny

As for other conifer genus, there is not yet any satisfying phylogeny of Picea. The best available phylogenies of the Picea genus are based on chloroplast and mitochondrial markers (e.g. RAN et al., 2006; BOUILLÉ et al., 2010) that both are non-recombining and therefore equivalent to single loci. These phylogenies based on cytoplasmic markers neither match morphological classifications nor are congruent with each other. These discrepancies could reflect the different inheritance and dispersal modes of the two cytoplasmic genomes as well as incomplete lineage sorting. Even if the effective population size of cytoplasmic markers is one half that of nuclear markers in a monoecious species, there are good grounds to believe that incomplete lineage sorting also matters.
in this case as it seems pervading for nuclear loci. For instance, Bouillé and Bousquet (2005) recorded a vast number of trans-species shared polymorphisms. They suggested that this could be the result of incomplete lineage sorting, which is consistent with the large effective population size of spruce (in a simple split model it takes $\sim 9 - 12N_e$ generations for a pair of taxa to reach reciprocal monophyly in 95% nuclear loci, Hudson and Coyne, 2002), and/or from inter-species gene flow which is indeed reflected by the presence of hybrid zones, for instance between Norway and Siberian spruces (Tollefsrud, 2008). Because of their smaller effective population size organelle DNA will also become monophyletic much more rapidly and fail to recover information on ancient speciation events (Hudson and Coyne, 2002). Nonetheless the phylogenies based on cytoplasmic markers provide a first glimpse at the general history of the genus. Both markers from the mitochondria and chloroplast, for instance, indicate that *P. breweriana* is basal and that other species belong to three clades: a glauca clade, an Asian clade and an abies–marianna clade. But eventually, a multilocus phylogeny will be required. A multilocus approach does not insure the retrieval of the true species tree (Degnan and Rosenberg, 2006, 2009; Degnan et al., 2009), especially if the number of nuclear loci used is limited, but it will certainly reduce the level of uncertainty. Computational problems associated with large multiloci trees remain formidable but new methods are emerging and seem promising (e.g. Bryant et al., 2012).
1.1.2 Speciation and population history

The main goal of phylogenetics is to retrieve the species tree linking the different taxa under study. In genus such as spruces, where incomplete lineage sorting is the rule rather than the exception and where hybridization does occur, going from gene tree to species tree is not an easy task and will require numerous independent loci. While the presence of a large proportion of shared polymorphisms is a cause for worries for species tree inference, it can also be viewed as a positive thing if one is instead interested in inferring ancient population demographics. Indeed the fact that ancient polymorphisms that coalesce in the ancestral population still segregate in current populations implies that the SNP variation detected in current data also contains information on the ancestral population. Hence, one can use polymorphism and divergence data to estimate key demographic parameters associated to the divergence of populations or species. Since the seminal paper of Wakeley and Hey (1997) and the introduction of "Isolation by migration" model, different approaches have been developed to estimate those demographic parameters. It should be noted that this large body of work is a direct evidence of the convergence of methods in population genetics and phylogenetics.

"Isolation with migration" model

In its generic form, the "Isolation with migration" (IM) model describes an ancestral population that splits into two descendant populations that are connected by gene flow. The IM model assumes neutrality and has six parameters: the scaled mutation rates of the ancestral and descendant populations ($\theta_A$, $\theta_1$ and $\theta_2$) where $\theta = 4 \times N_e \times \mu$, $N_e$ is the effective population size and $\mu$ is the mutation rate, the split time ($T$) between the descendant populations and the migration rates between the descendant populations after the split ($m_{12}$ and $m_{21}$) (Figure 1.2).

\[
\theta_A = 4N_e\mu, \quad \theta_1 = 4N_1\mu, \quad \theta_2 = 4N_2\mu
\]

![Figure 1.2. "Isolation with migration" model](image)

Different statistical methods have been developed to estimate those parameters (Hey and Nielsen, 2004; Becquet and Przeworski, 2007), all of which are based on the coalescent model presented in Nielsen and Wakeley (2001) for a single locus and Hey and Nielsen (2004) for multiple loci.
Briefly, the methods generally rely on a comparison between polymorphism and divergence and use summary statistics initially proposed by Wakeley and Hey (1997), or a slightly modified version of those (Becquet and Przeworski, 2007). Following Wakeley and Hey (1997) segregating sites (S) are classified into four categories, \( S_1, S_2, S_s, \) and \( S_f \). For each locus, \( S_1 \) and \( S_2 \) are the number of polymorphic sites unique to populations 1 and 2, respectively; \( S_s \) is the number of sites with shared alleles between the two samples, and \( S_f \) is the number of sites where fixed alleles are found in one sample and no polymorphisms are found in the other sample (Figure 1.3). These summary statistics contain information about the demographic parameters of the IM model. For example, a large number of shared polymorphisms are expected when the split time is short and the effective population sizes are large, as most coalescent events would then take place in the ancestral population. The presence of gene flow complicates things a bit and makes them less intuitive since gene flow will also lead to shared polymorphisms. However, gene flow will also lead to an excess of variation in coalescence time among loci compared to a pure split model (Becquet and Przeworski, 2009).

The two descendent populations can be two populations of the same species, two incipient species or two well-defined species. In contrast to the prevailing allopatric model of speciation a meta-analysis of Pinho and Hey (2010) suggests that speciation often occurs in the presence of gene flow. These results should, however, be interpreted carefully. For example, as shown by Becquet and Przeworski (2009), current methods to estimate gene flow under the IM models, will have a limited power to estimate gene flow early on in the divergence process, since it may not increase the variation in coalescence time to a detectable level. Unfortunately it is precisely gene flow occurring early on in the divergence process that one would like to detect in order to tell apart sympatric and allopatric speciations. Finally, we note that neutral models, even...
though they remain irreplaceable as null models, are unlikely to capture all the complexity of the speciation process that is the result of interactions between natural selection, genetic drift, gene flow and recombination, all of which can cause accumulation of fixed mutations and finally lead to genetic incompatibility between the two closely related populations (see reviews of COYNE AND ORR, 2004; HEY, 2006; PINHO AND HEY, 2010). So eventually, one would have to understand better the interplay of demography and selection within each species.

**Demographic history of spruce species**

Like most other plant and animal species, spruces experienced cycles of contractions and expansions in population size in response to climate changes during the late Quaternary (DAVIS AND SHAW, 2001). These fluctuations were particularly strong for species living at high latitudes in North America and Eurasia, but species at lower latitudes were also affected, though probably to a lesser degree. This is supported by a combination of palynological data, macrofossils and surveys of genetic diversity. The available data, as well as the data presented in the present thesis, however, suggest a large variation in the way individual species reacted to climatic fluctuations.

In North America paleoecological studies of pollen fossil showed that *P. mariana* and *P. glauca* went through a rapid population expansion that lasted over 3000 years since the Last Glacial Maximum (LGM) (McLeod and MacDonal, 1997) and *P. mariana* replaced the dominant position of *P. glauca* at ~ 14,000 cal yr BP (Lindbladh et al., 2007). *P. sitchensis* is also thought to have experienced a rapid expansion from southern refugia in coastal California (DAUBENMIRE, 1968; SOLTIS et al., 1997; Mimura and Aitken, 2007). However, the Californian spruce species *P. breweriana* had a fairly stable populations size during the same period (LEDIG AND HODGSKISS, 2005).

A similar variation in plant reactions to climatic changes is observed in Eurasian spruce species. Pollen data (GIESECKE AND BENNETT, 2004; Knaap, 2006) indicate that *P. abies* expanded strongly in Northwest Europe. After the LGM Norway spruce recolonized Scandinavia from populations located in central Russia and most likely also from cryptic refugia in the Northern part of European Russia (Välaranta et al., 2011; Chen et al., 2012). Spruces seem to have also survived in small glacial refugia on the Norwegian coast, but their contribution to current populations appears to have been limited (Parducci et al., 2012). The spread westwards followed at least two main recolonization routes (Tollefsrud et al., 2009), today reflected in a complex population genetic structure despite of extensive gene flow (Heuertz et al., 2006; Chen et al., 2012). Norway spruce reached eastern Finland about 6,500 years ago and eastern central Sweden about 2,700 years ago (Giesecke and Bennett, 2004; Seppä et al., 2010). Thus the current Scandinavian latitudinal clines in bud set (see below) were established during a short evolutionary period and from different source populations. Interestingly, multiloci studies in both conifers (*Picea abies*, *Pinus sylvestris*) and angiosperm trees (*Populus tremula*) found that the three species experienced a severe bottleneck long before the LGM (Pyhäjärvi et al., 2007; Ingvarsson, 2008). In the three species,
inferences were based on the intensive use of coalescent simulations and summary statistics of the site frequency spectrum (Tajima's $D$ and Fay & Wu's $H$). In Norway spruce, Heuertz et al. (2006) used 22 loci from a sample of seven populations representing Baltico-Nordic domain, Alpine domain and Romania domain. It should be noted that Namroud et al. (2010) using a smaller dataset but longer sequences detected a bottleneck corresponding roughly to the LGM.

As one moves eastwards into the Siberian plain and onto the Qinghai Tibetan Plateau (QTP), paleoecological data become more scarce. In both places the available data suggest a different impact of climate changes compared to western Europe. In Siberia, glaciation has been much less extensive than in Northwestern Europe (MacDonald et al., 2008; Stauch and Gualtieri, 2008). While the treeline certainly moved during the Holocene some conifer stands remained at high latitude in the Northeast (Mueller et al., 2010). Much less is known about central Siberia. A recent study indicates that most of the area was a cold desert (Velichko et al., 2011) as westerly moisture, which is today the source of both rain and snow in this region, was blocked by the Scandinavian Ice Sheet during the Glacial time. The depth of permafrost during the LGM was also much greater than it is today. However, macrofossil remains collected on lower river terrace and high floodplains suggest that while most of the region consisted of Aeolian dunes the river valleys were a separate habitat where spruces were able to survive (Binney et al., 2009). Altogether, paleoecological and genetic data hence suggest that the impact of climate changes on the population size of the main components of the boreal forest has probably been more limited than in their Scandinavian counterparts.

The mountain ranges surrounding the QTP are home to a very large diversity of plant species (López Pujol et al., 2011). In particular, a third of the spruce species are growing there, a proportion strikingly different to that of boreal species, which occupied a much larger area. This high species diversity is probably a consequence of the orogenesis of the highly rugged and dissected topography of the area. In the words of López Pujol et al. (2011) the eastern fringe of the Tibetan Plateau clearly constitutes the "evolutionary front" of China. Palynological studies suggest that conifers were present in the region some 50 mya but that spruce species only started to be an important fraction of the pollen record on the QTP around 38 mya (Wang et al., 1990; Dupont Nivet et al., 2008). Spruce species remained common until ca. 20 mya at which time their contribution to the pollen record started to decline and stayed low until 17 mya. The ensuing years are unfortunately poorly documented, but the continuing elevation and climatic changes through which the QTP went during this period certainly led to further fluctuations in abundance. In particular, the significant uplift of southern Tibet some 7 mya inferred by some authors (e.g. Wang et al., 2006, 2008) would certainly have affected spruce distribution. As for Siberia, the impact of the Quaternary full-glacial periods on species genetic diversity seems to have been more limited than in Western Europe even though there is indirect evidence that species also retreated to glacial refugia. Altogether the Quaternary history of conifers on the fringes of the QTP appears very dynamic with major population shifts but also
species staying put, secondary contacts and the presence of homoploid hybrids (MA et al., 2006; LI et al., 2010b, 2011).

In summary, paleoecological and genetic data suggest that spruce species experienced very different demography histories in various parts of the world. A comprehensive description of the history of spruce species should account for this diversity. This is especially true for reconstruction based on standing genetic variation which is going to be influenced by the recent demographic changes reviewed above. Further, as we noted in the Phylogeny section, the presence of incomplete lineage sorting and hybridization means that retrieving species trees in such group of species is going to be hard. A multilocus population genetic analysis of smaller groups of species therefore seems more promising, at least in the short term. When this thesis was initiated the only available study BOUILLÉ AND BOUSQUET (2005) considered three distantly related species and gave estimates of $N_e$ varying from 96,000 to 182,000 and a divergence time ranging from 10 to 18 million years, depending on the species pair considered (assuming mutation rate $\mu = 1 \sim 2 \times 10^{-8}$ and generation time of 50 years). These estimates were simply derived from pairwise nucleotide divergence ($\pi$), without any consideration of the demographic history of the pair of species analyzed.

The first two studies of the thesis (Paper I and II) constitute a first attempt in spruce species to use multilocus data and IM methods to:

i Estimate divergence time and gene flow between different species
ii Identify putative hybrids
iii Resolve the phylogenetic relationships between groups of spruce species

1.1.3 Identifying local adaptation along a cline

Geographic variation and clines

Speciation, like all evolutionary processes, depends in the end on a balance between evolutionary forces. As a matter of fact, populations can diverge genetically and yet not lead to new species. Instead the various evolutionary forces acting on the populations, like natural selection, mutation, gene flow and genetic drift might reach an equilibrium which is reflected by different forms of population genetic structure. Geographic variation is found in nearly every group of organisms, and a given species may exhibit within its range any of a large array of geographical patterns such as "Disjunction", "Overlap", "Hybrid zone", "Conjunction" and "Gradation" (ENDLER, 1977). The latter two are called "clines" and have since long fascinated biologists (ENDLER, 1977) as they can provide striking examples of natural selection. At a simpler level, a cline can be regarded as a gradient of characters controlled by either a single major gene or groups of linked or interacting genes, the dispersal of whose frequencies could be used as a quantitative description of a cline itself. The correlation of genotype or phenotype frequency with climatic variables (such as temperature and precipitation) or physical variables (such as latitude, altitude and longitude) has been used as a powerful method of detecting spatially varying natural selection and has been reported in a vast number of genes.
across species, for example the pigmentation in humans (ROBERTS, 1977), Adh gene in Drosophila (BERRY AND KREITMAN, 1993), FRI and FLC genes in Arabidopsis (CAICEDO et al., 2004). The large distribution ranges of forest trees such as spruces, span a broad range of environments and thereby populations are under a highly diverse set of selection environments. SAVOLAINEN et al. (2007) and ERIKSSON (2008) have summarized a large number of studies showing that phenotypic variation is often associated to environmental variables in forest trees. Classical example of clines in forest trees are traits related to phenology such as bud burst, bud set, growth cessation or cold tolerance. While there are certainly good grounds to believe that these clines are adaptive, proving it at the gene level has turned out to be hard. This is in part because gene flow and population history can also create clines in allele frequency but also because these traits are inherently quantitative traits.

Detecting local adaptation in a cline

A cline can be generated by selection (INGVARSSON et al., 2006) as well as many other forces like population demography (HOLLIDAY et al., 2010) and random genetic drift (POLECHOVÁ AND BARTON, 2011). Various groups of plant and animal species recolonized Scandinavia from different LGM refugia (e.g. JAAROLA et al., 1999; DE CARVALHO et al., 2010; CHEN et al., 2012), resulting in a contact zone between the different ancestral lineages of varying breadth and location or in a more gradual change in allele frequency across Scandinavia. In wind-pollinated forest trees, in general, and in Norway spruce, in particular, another difficulty comes from extensive gene flow that results in low population genetic differentiation (average within population $F_{ST}$ ∼ 0.025) and limited linkage disequilibrium (LD). Therefore only a slight shift of allele frequency could be observed at most loci even across a wide geographical range. To address these issues requires careful sampling and choice of statistics. It is also necessary to combine patterns from multilocus genetic markers since it can be difficult to detect clinal variation at individual loci if the clines in phenotypic traits are due to small frequency differences at many loci (BARTON, 1983, 1999; LE CORRE AND KREMER, 2003; LE CORRE AND KREMER, 2012; KREMER AND LE CORRE, 2012).

Parallel evolution can provide even more compelling evidence of local adaptation if similar geographic variation is observed at the same genetic markers along parallel clines with different population history. ORR (2005) pointed out that in a single-step substitution model parallel evolution occurs twice as often under natural selection as under neutrality. However, evolution in multi-step substitution becomes more complex and unpredictable since many factors such as pleiotropy, epistasis or population history (STERN AND ORGOGOZO, 2009; UNCKLESS AND ORR, 2009; CHEVIN et al., 2010) can influence the distribution of beneficial mutants. Empirical data suggest that adaptive convergence may be more common that previously accepted at the level of genes and function complexes (e.g. TENAILLON et al., 2012; DOMINGUES et al., 2012; JONES et al., 2012) although there are also examples of studies that failed to find common genes in different parts of the range (e.g. FOURNIER-LEVEL et al., 2011).
Comparison of molecular and quantitative genetics data suggests that recent selection, occurring after postglacial recolonization is the predominant factor that shapes standing quantitative trait variation (Collignon et al., 2002; KREMER et al., 2002). While there is a large number of studies at the phenotypic level, studies at the genotypic are less frequent and are mainly based on QTL or association mapping. This, in part, may be due to the fact that differentiation in nucleotide markers is much lower than for morphological or other polygenic traits among populations of forest trees ($F_{ST} < Q_{ST}$) and also to the dearth of physiological studies highlighting putative candidate genes. The increasing ease to obtain sequence data has, however, led to a renewed focus on approaches to identify local adaptation at the gene level in forest trees (Eckert et al., 2010; Hall et al., 2011; Keller et al., 2012).

In Norway spruce, Gyllenstrand et al. (2007) showed that the expression of the FT-homologue paFT4 (more correctly named paFTL2 after Karl-Gren et al. 2011) is significantly correlated with bud set and that populations from latitude 66.7°N and 47.3°N are differentiated. These first results are a promising beginning and we can now start to address more detailed questions:

i Can clinal variation be observed at Single Nucleotide Polymorphism (SNP)?

ii How can we disentangle the effect of natural selection from that of population history when analyzing clinal variation at SNP?

iii Can we observe adaptive convergence in another parallel cline with similar geographic range in a phylogenetically close species P. obovata? And more importantly, at which level: genes or single substitutions?

1.2 Molecular evolution of Norway spruce

Had contemporaries of Darwin been aware that human beings share on average more than 95 ~ 98% genome sequence with chimpanzees (Human Genome Project: http://genomics.energy.gov/), they would probably had found it less provocative that we descend from apes. The main question today has shifted and is more focused on explaining how this limited 2 ~ 5% nucleotide divergence between the two species could lead to such dramatic anatomical and cognitive differences. Or stated differently: what makes a given species appearance and abilities so specific? The question was first brought up by King and Wilson (1975) who highlighted two possible places to search for answers: the proteome and the regulatory mechanisms controlling gene expression (see Carroll, 2005 for a review). In attempts to look at this in spruce we carried out comparative studies (Paper V and VI) of the protein-coding transcriptome and of a group of small noncoding regulatory RNAs, respectively.

1.2.1 Transcriptome

Proteins are ultimately responsible for an organism’s anatomical, physiological, and behavioral characteristics. A fully characterized transcriptome is therefore a key step to understand life diversity. It is also crucial for genome annotation.
and to gain new insights on the temporal and spatial patterns of gene expression. It has recently become evident that species transcriptomes are highly diverse and complex in many organisms (e.g. Mortazavi et al., 2008; Pan et al., 2008; Filichkin et al., 2010; Zhang et al., 2010).

In conifers, however, enormous genome sizes and a very large fraction of repetitive elements have so far hindered genomic studies. The transcriptome studies are mainly based on re-sequencing or microarrays. In spruce, EST sequences based on Sanger sequencing are still the main sources for genetic studies. Large EST libraries have been constructed in *P. glauca* (Ralph et al. 2008) and in *P. sitchensis* (Rigault et al. 2011). 27,720 unique transcribed genes were sequenced in glauca and 23,589 were annotated as full length cDNA. In sitchensis, the number of annotated full length cDNA decreases to around 8,000. Compared to those two species, *P. abies* still requires intensive efforts that only 8,715 Putative Unique Transcripts (PUT) are available and are poorly annotated (http://www.plantgdb.org).

mRNA-Seq and gene expression profiling

Application of next-generation sequencing (NGS) provides a full access to a species’ transcriptome. For example, mRNA-sequencing (mRNA-Seq) leads to deep coverage and base-level resolution (Nagalakshmi et al., 2008). Compared to conventional methods such as microarrays, mRNA-Seq gives a much finer characterization of key aspects of the transcriptome and has already been instrumental for full annotation of protein-coding gene, gene expression profiling, noncoding RNA discovery and detection, transcript rearrangement discovery and single-nucleotide variation profiling in many model species (e.g. *Zea mays*, Emrich et al., 2007; *Medicago truncatula*, Cheung et al., 2006; *Arabidopsis thaliana*, Weber et al., 2007).

Gene expression profiling has been the main method for gene functional studies ever since the development of Northern blot, microarray (or gene chip) and other hybridization-based technology. The appearance of mRNA-Seq offers both sequence data and estimates of gene expression level in a single experiment. Compared to microarray, mRNA-Seq requires no prior annotation of the genome for probe selection, avoids biases introduced during hybridization of microarray, has a relatively low cost for the amount of data it produces and is the method of choice in projects using non-model organisms. Thus mRNA-Seq has become more and more popular for characterization of transcriptomes in a number of organisms and in various tissues (e.g. Mortazavi et al., 2008; Zhang et al., 2010; Gan et al., 2011).

A circadian rhythm is the reaction to the light-dark cycle in most of living organisms. It is one of the most important and complicated biological processes in plants and affects thousands of physiological processes including the timing to flowering, leaf movement, germination, gas exchange and so on. In angiosperms, about 20% of the genes exhibit different diurnal patterns at expression level (estimates of percentage varied due to differences in experimental treatments, species and tissues, see e.g. Jiao et al., 2005; Ma et al., 2005; Filichkin et al., 2011). However due to lack of data in both genetics and phys-
iology, no clear diurnal patterns have been reported yet in key photosynthesis genes in gymnosperms (ALOSI et al., 1990; GUSTAFSSON et al., 1991).

**Comparative evolution of coding sequences**

Comparative genomics exploits both similarities and differences in the proteins, RNA and regulatory regions of different organisms to compare orthologues and test for selection (e.g. CLARK et al., 2003 for primates; LI et al., 2010a for vascular plants). To identify selection and its directionality, one can compare subsets of key genes between closely related species by assuming that they could explain the evolution of species in the context of their phylogenetic order. Therefore, most of the first genome-wide studies of natural selection have focused on coding sequence and estimates of non-synonymous versus synonymous ratio ($dN/dS$) as the signal of selection intensity (e.g. YANG, 2006).

Gymnosperms separated from angiosperms about 300 million years ago (RENNER, 2009) and they differ from the latter in many aspects, such as recombination (JARAMILLO-CORREA et al., 2010; MORITSUKA et al., 2011), genome duplication (AHUJA AND NEALE, 2005) and genome size (MORGANTE AND DE PAOLI, 2011). We would therefore expect different mode and tempo of evolution between these two groups. Early studies (e.g. PALMÉ et al., 2008a; BUSCHIAZZO et al., 2012) indeed found lower synonymous mutation rate in gymnosperms than in angiosperms, particularly in annual angiosperms. BUSCHIAZZO et al. (2012) also reported a different pattern in average evolutionary rates (measured by $dN/dS$ ratio) between gymnosperms (0.314, between *P. sitchensis* and *P. taeda*) and between angiosperms (0.092, between *Arabidopsis thaliana* and *Populus trichocarpa*).

Our assembly is the first release of massive parallel sequencing applied to the needle transcriptome of *P. abies*, even though it is limited by its small sample size. Still together with the extant PUT sequences, the *P. abies* transcriptome now has reached a size of the same order of magnitude as the North American species. It becomes possible to do genome-wide protein-coding gene annotation, gene expression profiling, orthologue comparison and test for selection by comparing to *P. glauca* and *P. sitchensis* using *Pinus taeda* or *Taxus mairei* as outgroup. The goal of Paper V is to address the following questions using full transcriptome data for the first time:

i How much of the transcriptome is differentially expressed between dark and light treatments?

ii How reproducible are the estimates of mutation and evolutionary rates previously reported?

iii Are gene evolutionary rate and expression level correlated?

### 1.2.2 Small noncoding RNAs and their regulatory role in plants

Small noncoding RNA (sRNA) of 20-30 nucleotides (nt) guides regulatory processes at the DNA or RNA level in a wide range of eukaryotes and prokaryotes. Although small-RNA-mediated silencing had already been observed in the late 1980s, the underlying mechanism remained unknown until the breakthrough
study of FIRE et al. (1998) in Caenorhabditis elegans. Small RNAs can cause either transcriptional gene silencing by guiding heterochromatin formation at homologous loci, or lead to post-transcriptional gene silencing through mRNA degradation or translational inhibition. In plants, a large portion of them has important roles in regulation of diverse biological processes, including regulation of patterning and development, response to the environment, defense against pathogens and silencing of endogenous transposable elements (BONNET et al., 2006).

Biogenesis

Small RNAs can either be endogenous or pathogen-derived. They can be classified according to their different modes of biogenesis into three major types, short interfering RNA (siRNA), microRNA (miRNA) and piwi-interacting RNA (piRNA). miRNAs and siRNAs are processed from precursors by the RNase III endonuclease dicer, which acts on double-stranded substrates to release small RNA duplexes with 2-nt 3’ overhangs.

Dicer-like proteins have a distinct, hierarchical, and overlapping function in small RNA biogenesis. In Arabidopsis, four such Dicer-like proteins (DCL1-4) are responsible for processing specifically sized sRNA duplexes: DCL1 synthesizes 18 ∼ 21-nt-long sRNA including most of miRNAs (though a few miRNAs appear to be DCL4 dependent, see RAJAGOPALAN et al., 2006; FAHLGREN et al., 2007), whereas the products of DCL2, DCL3 and DCL4 are 22 nt, 24-nt and 21-nt long, respectively (XIE et al., 2004). Numerous studies of model and non-model plants have shown that the most abundant class of plant sRNA are siRNAs, which are typically 24-nt long in angiosperms and guide DNA methylation and heterochromatin formation of repetitive and transposable elements (XIE et al., 2004; RAJAGOPALAN et al., 2006). Gymnosperms seem to lack both 24-nt siRNAs and DCL3 (MORIN et al., 2008). DCL4 and DCL2 conduct a phased reaction dicing of trans-acting siRNA (ta-siRNA) precursors that are produced upon miRNA-guided cleavage of non-coding primary transcripts (reviewed in CHAPMAN AND CARRINGTON 2007).

Response to biotic stress

Besides their roles in developmental patterning and maintenance of genome integrity, many sRNAs are also components of plant responses to adverse environmental conditions, including biotic stress (reviewed by RUIZ-FERRER AND VOINNET, 2009 and CHEN, 2009). All four Arabidopsis DCLs play critical and redundant antiviral roles. 21-nt-long siRNAs produced by DCL2 and DCL4 move between cells and likely immunize tissues just ahead of the infection by amplifying a systemic silencing response (MOISSIARD et al., 2007; DONAIRE et al., 2008). DCL1-dependent miRNAs can target DCL4 transcripts for degradation or negatively control transcription factors required for DCL4 expression (QU et al., 2008). Finally DCL3-dependent, 24-nt-long sRNA might dampen viral transcription by inducing chromatin condensation of nuclear viral episomes and mini-chromosomes (RAJ et al., 2008).

Recent researches in Medicago truncatula and Solanum lycopersicum have shown that high levels of siRNA match to a large number of defense-related
nucleotide binding site-leucine-rich repeats (NBS-LRR) coding genes, in a way characterized by spaced in 21-nt "phased" intervals (ZHAI et al., 2011; SHIVAPRASAD et al., 2012). The production of these phased siRNAs are triggered by three groups of 22-nt miRNA (miRNA2118, miRNA173, and miRNA390). No similar patterns have been reported in model species like Arabidopsis thaliana or Oryza sativa. This observation might reflect the evolutionary divergence in strategies of pathogen defense between species.

Characterization of sRNA using NGS and Bioinformatics

Next-generation sequencing has been developed to help characterize sRNA in many species (e.g. AXTELL et al., 2006). Conventional methods like direct cloning and sequencing can only provide sequences and microarray is extensively used for expression profiling but restricted to known miRNA and siRNA genes. While NGS can provide both at one experiment and has been highly efficient for the discovery of novel miRNAs, siRNAs and piRNAs genes on a genome-wide scale (e.g. BARAKAT et al., 2007; AXTELL et al., 2006; GIRARD et al., 2006).

The aim of most sRNA studies is to profile miRNAs and their targets and thereby characterize miRNAs biogenesis and miRNA-target-recognition mechanisms. miRNAs comprise a large family of small ~ 21-nucleotide-long non-coding RNAs that emerge as key post-transcriptional regulators of gene expression in metazoan animals, plants and protozoa (JONES-RHIOADES et al., 2006). In plants, microRNA gene (MIR) is transcribed into a capped and polyadenylated primary microRNA (pri-miRNA). The pri-miRNA is processed into a stem-loop precursor (pre-miRNA), which is further processed into a duplex of miRNA and miRNA*. Plant miRNAs regulate target mRNAs through two major mechanisms: transcript cleavage and translation inhibition. They are high-level regulators of gene expression that affect numerous aspects of plant biology, especially developmental patterning. Mutants impaired in miRNA biogenesis exhibit severe, pleiotropic abnormalities and over-expression of particular miRNAs or their targets exhibit a wide array of unusual phenotype in plants (see a review of JONES-RHIOADES et al., 2006). Many miRNAs and targets have been successfully identified by computational approaches based on empirical rules (e.g. DAI et al., 2011).

In spruce, YAKOVLEV et al. (2010) performed direct cloning and sequencing and reported 44 miRNAs, 24 of which were novel in Norway spruce. 25 target groups were also identified and the majority were annotated as TIR-NBS-LRR genes or transcription factors. All their predictions were based on hairpin stability and sequence complementarity. For the description of whole sRNA populations, little has been accomplished even in model systems. Our goal in this part of the thesis is to:

i Obtain a detailed characterization of the sRNA populations from P. abies.
ii Identify putative miRNAs and their targets.
iii Compare sRNA populations between different plant species with emphasis on features identified in P. abies
2. Results and discussion

2.1 Estimate of speciation history of spruce based on multilocus patterns - Paper I and II

In Paper I and II we investigated the multilocus pattern of polymorphism and divergence in two groups of spruce species. We sequenced fragments of 10 and 12 genes covering over 8,000 – 10,000 base pair sequences in boreal and montane species, respectively. The first group included one Eurasian species, *P. abies* and three North American species, *P. glauca*, *P. mariana* and *P. breweri*ana (Paper I) and the second group comprised four species from the Qinghai-Tibetan Plateau: *P. shrenkiana*, *P. likiangensis*, *P. purpurea* and *P. wilsonii* (Paper II). Average numbers of 35 and 44 individuals were studied across genes and species. In both cases, nucleotide diversity and divergence between pairs of species were summarized and used to estimate the speciation parameters (effective population size $N_e$, divergence time $T$ and migration rate $M$) under IM model. We examined the standard neutral model as well as three other demographic models (see Figure 2.1) and the goodness-of-fit for the models was examined using a set of summary statistics including nucleotide diversity ($\pi$), private, shared and fixed polymorphism ($S_1$, $S_2$, $S_s$ and $S_f$), population differentiation ($F_{ST}$) and Tajima’s $D$.

![Figure 2.1. Diagram of the different demographic models evaluated in Paper I and Paper II. The SNM is determined by population mutation rate $\theta$ alone, whereas PEM has two parameters $\theta$ and $\alpha$, the exponential growth rate. The BNM is characterized by four parameters $\theta$, $f$ (size of bottleneck), $d$ (duration of bottleneck), $T$ (time after bottleneck). The BEM has four parameters $\theta$, $\alpha$, $T$ and $f$.](image-url)
2.1.1 Nucleotide diversity and gene flow

In agreement with Bouillé and Bousquet (2005), we observed a strikingly large number of shared polymorphisms \((S_s)\). The mean values are generally higher between Chinese species (61.5) than between Eurasian and North American species (21). Conversely, the average number of fixed sites between these species \((S_f)\) is low (5) although it is much higher (34) when they are paired with *P. breweriana*.

The analysis of polymorphism and divergence in these eight species led to interesting results. First, *Picea* species generally have fairly large effective population sizes \((85,000 – 206,000\) depending on species) with two notable exceptions: *P. schrenkiana* effective population size is about 30,000 and with a value around 12,000, *P. breweriana* has an even smaller effective population size. Our estimates reflected quite well the differences of nucleotide diversity between *P. breweriana* and the other three North American and Eurasian spruce studied here. Interestingly, this low effective population size of breweriana spruce is not observed for allozyme loci that have similar diversity levels (Ledig et al., 1997). Such a discrepancy between levels of polymorphism at allozyme and nucleotide has been previously noted in *P. abies* (Heuertz et al., 2006) and *Pinus sylvestris* (Pyhäjärvi et al., 2007). The latter suggested that selection could be the cause. Second, the three spruce species from the QTP have a surprisingly larger nucleotide diversity \((> 0.005)\) considering their relatively small distribution compared to the continent-wide distributed species such as *P. glauca* and *P. mariana*. The low diversity in the boreal species could be due to a stronger impact of quaternary glaciations. Third, gene flow occurs between species, even species today found on different continents. We found evidence of gene flow from *P. glauca* to *P. abies* and from *P. schrenkiana* to *P. wilsonii*, although most shared polymorphisms likely reflect incomplete lineage sorting. Because today these species cannot be crossed easily, our estimate of gene flow should be regarded as an average value dating back to the time when new species formed. The presence of gene flow between *P. glauca* and *P. abies* could trace back to a warm period where the two species where sympatric on Greenland (De Vernal and Hillaire-Marcel, 2008).

Overall, our results are consistent with previous estimates of \(N_e\) and \(T\) based on nucleotide diversity (Bouillé and Bousquet, 2005). Additionally, our studies highlighted the possibility of shared ancestral polymorphisms between species that separated earlier than 15 mya. This result is actually not too surprising if one considers an effective population size timescale. Considering the large population sizes of conifer species and assuming a generation time of 50 years, 15 million years translate into mere \(4N_e\) generations, a value much smaller than the \(9 – 12N_e\) generations that is required for 95% of loci to become reciprocally monophyletic (Hudson and Coyne, 2002).

2.1.2 Demographic history of eight spruce species

Despite an overall fit to the standard IM model, negative values in both Tajima’s \(D\) and Fay & Wu’s \(H\) suggested a relatively ancient bottleneck (Heuertz et al., 2006; Pyhäjärvi et al., 2007; Ingvarsson, 2008) for most spruce species.
that have been studied. However, we expected a recent population expansion in *P. likiangensis* and a rather stable population history of *P. breweriana*. Therefore, we examined the possibility of two other different demographic scenarios including the exponential growth model (PEM) and bottleneck model (BNM) in Paper II and we estimated parameters under a model of bottleneck with exponential growth (BEM) in Paper I (all models are illustrated in Figure 2.1). For most pairwise comparisons, SNM could not be rejected and the demographic models only had a marginal preference based on approximate Bayesian computation (Paper II). But simulations under the growth scenario did significantly improve the model fit to the observed data (Paper I, Figure 2.2). Our estimates implied an ancient bottleneck dating back to 2 mya followed by a slow population expansion. It is stark contrast to the traditional view of a recent bottleneck and quick expansion at the end of Last Glacial Maximum (LGM) based on palynology and fossil study (Mcleod and Macdonald, 1997; Knaap, 2006; Lindbladh *et al.*, 2007). The latter is heavily limited by its data type and could only reveal the effect of quite recent history. Our results of population demography are close to those obtained by Heuertz *et al.* (2006) (0.15 – 0.3 mya) which were also based on multilocus data but considered only one species.

In general, the researches in Paper I and II have highlighted the importance of using multilocus sequences and species comparisons when inferring population history. Multilocus data should have an obvious advantage over cytoplasmic DNA in inferring demographics further back in time. More and more phylogenetic researches have adopted multilocus data, population size, recombination rate and other population genetic concepts to help solving incongruence between species tree and gene trees (e.g. Liu 2008; Liu *et al.* 2008). Additionally with the development of Approximate Bayesian Computing (ABC) approaches, we now have the ability to implement and assess much more complex demographic models using data from a larger number of genes and species (see a review by Beaumont, 2010). However, questions still remain as pointed out by Pinho and Hey (2010): what are the details of the interaction between disruptive selection and linkage and how often disruptive selection itself can be the initial trigger for divergence with gene flow? Answering these question require identification of the genes that are the target of selection, a topic that we are going to discuss in the following sections.

### 2.2 Clinal variation and local selection - Paper III and IV

After obtaining a global picture of spruce speciation history, we focused on more recent history in Papers III and IV. Also, while Papers I and II focused on neutral processes both Papers III and IV deal with adaptation. More specifically we addressed the following questions: how did conifer trees adapt to the local environment during their postglacial recolonization? And, does parallel evolution occur in Norway and Siberian spruces, two species with different recent histories? In Norway spruce, data on latitudinal clinal variation at the levels of phenotype, genotype and gene expression were combined and a three-step ap-
proach was applied to disentangle the effect of local selection with population history. The study in Siberian spruce had, at that stage, a more restricted scope and focused on allele frequencies at candidate genes and on a more limited set of phenotypic measurements. Together the two datasets are the first evidence of parallel evolution in conifer species.

2.2.1 Latitudinal gradient patterns in phenology

While clinal variation in bud set has been amply demonstrated previously (e.g. DORMLING, 1973; GYLLENSTRAND et al., 2007) we confirmed it in a subset of populations in both species. Seedlings were exposed to photoperiodic treatments of increasing night length. Bud set percentages of Norway spruce and growth cessation of Siberian spruce were summarized in 5 and 3 populations from 47°N or 54°N south to 68°N north, respectively. As would be expected in the natural environment, northern populations were more sensitive to the increase of night length under experimental conditions (see Figure 2.3A). In Norway spruce, nearly 50% individuals from northernmost populations started to set bud under 6.5-h darkness treatment while in the southernmost populations 9.5h-darkness treatment was required to reach a similar percentage. We also observed similar clinal variation in growth cessation in P. obovata: the average growing period was significantly shorter in northernmost population Igarka (14.3 days) compared to that of two southern populations (27.4 and 28.8, see Figure 2.3B).

2.2.2 Population genetic structures

We genotyped 137 SNPs from 19 candidate genes as well as 308 control SNPs from genes a priori unrelated to bud set in 18 Norway spruce populations. The

Figure 2.2. An example of goodness-of-fit plots from P. abies × P. glauca under both the neutral (left) and growth (right) models.
candidate genes are putatively involved in the photoperiodic pathway, circadian clock and shoot apical development. The population genetic structure inferred using the program STRUCTURE (Figure 2.4A, B) confirmed the differentiation of populations into three main clusters corresponding to the Alpine domain (Germany and the Saleby population), populations from central Sweden and Finland showing a high level of admixture and populations from Northern Finland (latitude > 66°N). These data may reflect the existence of at least two recolonization routes into Scandinavia after the Last Glacial Maximum (GIESECKE AND BENNETT, 2004; TOLLEFSRUD et al., 2009) and the presence of high latitude refugia in Northwestern Russia (VÄLIRANTA et al., 2011).

In Siberian spruce, post-glacial history did not generate any hierarchical structure among the populations investigated. The overall $F_{ST}$ value was 0.0152 and STRUCTURE results based on 14 SSR markers failed to delineate any meaningful clusters: all individuals appeared admixed reflecting the lack of population genetic structure (see Figure 2.4C, D). "Isolation by distance" (IBD) was significant but weak. Estimate based on 80 independent silent SNPs gave slightly different results: each population was divided into two clusters and the proportions of the two clusters were rather similar across populations. This indicated again a lack of latitudinal population genetic structure, but might also suggest the possibility of some longitudinal population genetic structure.

### 2.2.3 Allele frequency spectrum analysis for local adaptation

The admixture zone detected in Norway spruce certainly complicates the detection of clinal variation at candidate SNPs as even neutral polymorphisms could show clines in allele frequencies simply because of population history (NOVEMBRE AND DI RENZO, 2009). Therefore to tell selection apart from demography we performed a three-step approach that combined linear regression,
Figure 2.4. Clustering analysis conducted in STRUCTURE. Results were plotted when $K = 2$ and $K = 3$ in Norway spruce (A, B) based on 308 control SNPs, and in Siberian spruce (C, D) based on 14 SSR loci.
Bayesian generalized linear mixed model and $F_{ST}$ outlier tests. Significance thresholds were generated from empirical distributions based on control SNPs in order to limit the number of false positives. Enrichment of candidate SNPs at each significant level was used as evidence of selection. When only the linear regression method was considered no significant enrichment was detected. Furthermore, the number of SNPs showing clinal variation sharply decreased when the Northern Finland populations were excluded, which reflected the effect of population structure. However, when population structure was accounted for by using BayEnv, a method similar to the mixed linear model in association studies, enrichment of candidate SNPs became evident. Finally, we used an $F_{ST}$-based method BayeScan to test whether the clinal pattern could be due to local adaptation and we also obtained significant enrichment of candidate SNPs. Combining the results from the three methods hence yielded solid evidence of departure from neutrality at SNPs from $PaGI$, $PaPhyP$, $PaPhyN$, $PaPRR7$ and $PaFTL2$ (Figure 2.5A).

To be consistent, we applied the same three-step approach to detect local adaptation in Siberian spruce as well, although we had already pointed out the effect of population structure should be negligible in generating any latitudinal variations. Similar results pointed to two circadian clock genes $PoGI$ and $PoPRR7$, and the promoter of photoperiodic gene $PoFTL2$ that showed significant evidence of enrichment of non-synonymous LD groups. Notably, the same mutations were detected to be under local selection in Siberian spruce and Norway spruce: non-synonymous mutation $PoGI_F2_605$ in $PoGI$ and mutation $PoFTL2_1567$ in the promoter region of $PoFTL2$ (named $PaGI_F2_987$ and $PaFTL2_1950$ respectively in Norway spruce, Paper III). The former causes a replacement of Histidine to Tyrosine at the position 78 of GI peptide.

### 2.2.4 Differentiation in gene expression

_Gyllenstrand et al. (2007)_ reported that the expression level of $PaFT4$ (named $PaFTL2$ in Paper III) was significantly differentiated between a northern (66.7°N) and a southern (47.3°N) populations. We therefore examined the gene expression patterns of the photoperiodic pathway gene $PaFTL2$ as well as three additional circadian clock genes including $PaPRR7$, $PaGI$ and $PaCCA1$ under different photoperiodic treatments (Paper III). The mean population expression level of $PaFTL2$ had a significant linear relationship with latitude ($P$-value < 0.05, adjusted $R^2 = 0.75$, Figure 2.5B), being lowest in the south (58.3°N) and highest in the north (66.4°N). The other three genes showed slight non-significant differences in expression between populations and $PaPRR7$ and $PaGI$ exhibited a diurnal expression rhythm as expected from circadian clock genes.

### 2.2.5 Parallel evolution and the selection on *Gigantea*

By combining results of Table 2 in Paper III and Table 3 in Paper IV, it is obvious to notice that many outlier mutations were specific to one of the species
although *P. abies* and *P. obovata* are closely related and share quite a few polymorphic sites. However, the same group of genes under local selection have been identified in both species (e.g. *GI*, *PRR7* and *FTL2*). This could be an indication that parallel evolution occurs more often at the levels of gene or gene group within related pathways than at the level of single mutation.

The *Gigantea* gene (*GI*) offers probably the most intriguing case as it seems to be highly pleiotropic and associated with fitness-related traits in both *A. thaliana* (BROCK et al., 2007) and forest trees (ROHDE et al., 2011; KELLER et al., 2012). *GI* displays a high level of divergence but a low level of shared polymorphism, in contrast to the general pattern observed in spruce (Paper I). The structure of its nucleotide diversity is also remarkable: the ratio of $\pi$ at non-synonymous sites to synonymous sites is 11.6 while it is under one in all other loci except *PRR3* in *P. obovata*. This pattern, compared with Norway spruce, might indicate recurrent selective sweeps that have been suggested in *GI* in *Populus balsamifera* and genes from the photoperiodic pathway in *Populus tremula* (HALL et al., 2011; KELLER et al., 2012). However, because of its overall low polymorphism, it has been difficult to detect significant departure from neutrality in *GI*. The signature of selection found in Paper III and IV might alternatively reflect recent local selection related to life at high latitudes.

Physiology and protein structure studies have shown that the first 394 amino acids in the N-terminal helps mediate the interaction of *GI* with, for example, F-box family proteins like ZTL or FKF1 (KIM et al., 2007; SAWA et al., 2007; BLACK et al., 2011). This region is highly conserved among orthologue sequences from 17 plant genera. Although the exact location of the receptor is not known, the H78Y mutation might cause a change in *GI* protein configuration due to amino acid property change in electronic charges and hydrophobicity. Analyses based on 3D-structure prediction showed a small decrease in distance...
of H-bonds from His78 to Phe74, and to Glu80 when mutated to Tyr78 (from 2.88Å to 2.65Å and from 3.13Å to 2.92Å, respectively).

2.3 Norway spruce transcriptome and regulatory small RNAs - Paper V and VI

In Paper V and VI, we used a comparative approach to study molecular evolution at transcriptome scale. By sequencing both the mRNA and the small RNA populations we made it possible to identify features specific to \textit{P. abies}, but also to look at patterns common between different groups of plant species.

2.3.1 Norway spruce needle transcriptome sequencing

Needles were collected from a single adult individual of Norway spruce at two different time points, 1pm during day light and 1am during the following night in the dark. mRNA was selected from the total RNA extraction and its cDNA product was deeply sequenced using Illumina. Our final assembly contained 59,556 putative unique transcripts with a N50 size of 551bp. Over 50% of the assembled PUTs showed high similarity to either \textit{P. glauca} or \textit{P. sitchensis}. Using bioinformatic prediction, we identified 6,194 potential full length open read frames (ORF) from PUTs longer than 150 nucleotides. Compared with the length distribution of both untranslated regions (UTR) and the ORF to the validated full length mRNA library in \textit{P. sitchensis} and \textit{P. glauca}, a fraction of our assembly should include the real full length transcripts, even though it was enriched for short sequences.

2.3.2 Gene expression differences

Gene expression patterns under dark and daylight condition were identified using an empirical Bayesian approach. 2,076 transcripts were detected as significantly differentially expressed based on posterior probabilities but only a few showed a difference larger than 2 fold (Figure 2.6). Studies in \textit{Arabidopsis}, \textit{Oryza}, and \textit{Populus} (Filichkin et al., 2011; Jiao et al., 2005; Ma et al., 2005) suggest that on average around 20% of the transcriptome is differentially expressed between dark and light conditions, although the number of genes varies depending on species, tissues and actual treatment. Compared to angiosperms, our results could be an indication of weak difference in diurnal expression patterns at the key photosynthesis pathway of gymnosperms, which was first suggested by earlier studies of Aloisi et al. (1990) and Gustafsson et al. (1991). To validate this finding, more studies are required in multiple tissues and species.
2.3.3 Comparative molecular evolution in conifers

By comparing EST databases from three spruce species *P. abies*, *P. glauca* and *P. sitchensis*, one pine *Pinus taeda* and one yew *Taxus mairei*, we were able to estimate the phylogenetic trees and study the rate of molecular evolution in conifers. Orthologues between pairs of species were identified by reciprocal best blast hit on nucleotide similarity while sequence alignments were based on amino acids.

The ratios of nonsynonymous to synonymous substitution ($dN/dS$) were calculated from an average size of 4,737 ORFs aligned in 6 pairwise comparisons between spruce and pine, or spruce and yew. Based on the mean synonymous divergence between spruce and pine (0.175) and between spruce and yew (0.6), we obtained an average synonymous substitution rate of $0.6 \times 10^{-09}$ and $1.1 \times 10^{-09}$ bp/year, values that were in line with previous estimates of $0.7 - 1.31 \times 10^{-09}$ bp/year (WILLYARD et al., 2007). Our study lends substantial support to a lower annual substitution rate in gymnosperms compared to many angiosperms (PALMÉ et al., 2008b; BUSCHIAZZO et al., 2012; GERNANDT et al., 2008; WILLYARD et al., 2007). Although the mechanism underlying this lower substitution rate is still unclear, it should be noted that this substitution rate in perennial gymnosperms does not seem very low if one accounts for their long generation time.

Estimating evolutionary rate using $dN/dS$ ratio remains difficult as the ratio depends on the identification of the true orthologs and on the inherently stochastic nature of the substitution process. Our estimates of $dN/dS$ ratios between spruce and either pine (0.236) or yew (0.167) (Figure 2.7) lay between the values of 0.12 – 0.15 reported by PALMÉ et al. (2008b), and 0.314 by BUSCHIAZZO et al. (2012). The small number of sequences (138) considered by PALMÉ et al. (2008b) probably entails a bias towards highly conserved
gene groups across the four conifers they considered. They have therefore likely underestimated \(dN/dS\). BUSCHIAZZO et al. (2012) used a much larger dataset. The discrepancy between their estimate and ours seems to stem from the way they predicted ORFs and orthologues. For some reasons they have shorter orthologue alignments that in quite a few instance have been aligned to wrong coding frames. This leads to overestimates of \(dN\) and more importantly underestimates of \(dS\), which is indicated by the fact that in their study high \(dN/dS\) values are strongly correlated with short alignments where number of synonymous changes are also small. As suggested by recent work of WOLF et al. (2009), RATNAKUMAR et al. (2010), and EYRE-WALKER (2011), high \(dN/dS\) ratios can easily be due to low \(dS\) values and should not be taken as indication of positive selection without further work.

**Figure 2.7.** Histograms showing the distribution of \(dN/dS\) values from pairwise comparisons of species. The left plot shows patterns for all potential orthologues whereas the right one shows the patterns restricted to putative full length PUTs

We also investigated the correlation between the selective constraint \(dN/dS\) and the gene expression levels in needle but no significant patterns could be detected. In animals, fungi and plants, highly-expressed genes tend to have lower \(dN/dS\) values (SUBRAMANIAN AND KUMAR, 2004; PÁL et al., 2001; YANG AND GAUT, 2011). The lack of correlation in our study could simply be due to the incomplete transcriptome sequencing (missing lowly-expressed genes) and limited sample size (needle tissue of one individual).
2.3.4 Norway spruce small regulatory RNA sequencing

In Paper VI, we sequenced Norway spruce small regulatory RNA libraries from newly flushed bud tissues using Illumina sequencing. In total, 59,616 unique short sequences were obtained with a length varying from 18 to 29 nucleotides. Constraining the length between 18 to 24 nt, the most abundant class is 21-nt long, constituting 72% of all reads. The 22-nt class is the second most abundant (22%), while the 24-nt class which is reported as the most abundant in angiosperms is only 1% in Norway spruce (Figure 2.8). Our data support the previous observation of Morin et al. (2008) that conifer sRNA is devoid of 24-nt size class as well as the study of Yakovlev et al. (2010). Further study reveals that only a rare fraction (< 1%) of other size sRNA can be mapped to Picea and Pinus repetitive and transposable elements. Compared to the distribution obtained in Populus, this percentage lies in the lower 10%. It might explain the less efficient mechanisms for silencing transposable elements and thus could lead to the large genome sizes in conifers.

Figure 2.8. The observed length distribution of sRNA sequences from P. abies. The total number of reads in each size class is plotted.

2.3.5 MicroRNA prediction

In total only a small fraction (1.6%) of short sequences have properties suggesting they are either miRNA or miRNA* sequences. These include 51 families that showed high similarity to already described miRNAs and contain majority (39 out of 44) of spruce miRNA families previously described by Yakovlev et al. (2010). Compared to their lacking of conserved miRNAs, we identified 20 conserved families that had best homologue hits in various species including Arabidopsis, Oryza, Glycine, Zea, Vitis, Citrus, and Arachis, as well as 30 families matched homologues in Pinus or Picea. This could simply be explained by the improved sequencing technology. We also identified 20 miRNA families that matched none of the miRNAs ever recorded, but were well supported by the precursor hairpin structures, miRNA positions and frequencies and at least 68 other selected features specific to plant. We considered these as novel miRNA candidates possibly specific to Norway spruce.
Identification of miRNA target genes provided further evidence to our identification of miRNAs in half of the families, especially for the conserved ones. We are surprised to find out that 11 of conserved miRNA families in spruce have similar targets as in many angiosperm species (KHRAIWESH et al., 2012), such as SBP for miRNA156, MYB for miRNA159 and miRNA858, ARF for miRNA160 and so on. In a total number of 143 EST target genes, we found miRNA159 and miRNA858 collectively targeted up to 41 MYB genes. These 41 MYB genes all contain a signature domain that belongs to MYB R2R3 group, the largest MYB group in plants (STRACKE et al., 2001). According to the functions of target homologues in A. thaliana, these two miRNA families might regulate a vast range of biological processes, including response to chemical stimulus and salt stress, seed germination, flower development and etc. This is concordant with the observation of recent study by XIA et al. (2012) in apple. Another big share of miRNA targets is TIR-NBS-LRR disease resistance genes. Eight miRNA families in spruce targeted at proteins containing any of TIR, NBS or LRR domains. This phenomenon was first noticed by YAKOVLEV et al. (2010) and probably is specific to spruce. Some of these 22-nt miRNAs were later found to be involved in the TIR-NBS-LRR gene dependent generation of phased siRNA in spruce (e.g. miRNA951 in Figure 2.9).

![Figure 2.9. An example of 22-nt miRNA triggered phased 22-nt siRNA production in P. abies.](image)

The top graph shows the phasing score distribution of sRNA. Below is a magnification around a predicted miRNA cleavage site.

2.3.6 NBS-LRR gene family and relevant siRNAs

Annotation of EST clusters characterized by high numbers of sRNA sequences indicates that putative NBS-LRR genes could be a gene family that is degraded in a phased pattern. In total, 44% of 4.7 million reads mapped to Picea EST clusters could be mapped to clusters that showed high similarity to NBS-LRR
genes. Out of 100 genes with highest number of different sRNA sequences mapped to them, 73 are classified as NBS-LRR sequences. Restricting sRNA reads to those of 21-nt long strengthened the enrichment pattern of NBS-LRR genes. A large number of NBS-LRR genes might be subjected to phased siRNA degradation, especially for those of 21-nt long. Our data further suggest that those phased siRNA were initiated from miRNA guided cleavage, often by an abundant 22-nt miRNA (see an example of miRNA951 in Figure 2.9). Encouraged by similar findings in Medicago and Solanum (Zhai et al., 2011; Shivaprasad et al., 2012), we investigated this phenomenon in additional plant species. By mapping 21-nt sRNA sequences and focusing only on genes containing NBS-LRR domain, highly variable patterns were identified across species. Amobrella, Medicago, Cotton, Populus and Vitis have a high percentage of NBS-LRR genes mapped by 21-nt sRNA while a low percentage of genes hit by 21-nt sRNA were found in monocots and non-angiosperm species except Picea where over 90% of targeting genes hit by more than 10 unique sRNA sequences were annotated to contain TIR, NBS or BED domains (Figure 2.10).

![Figure 2.10. Percentage of NBS-LRR type genes hit by more than 10 sRNA reads in different plant species.](image)

The comparison of sRNA populations across four angiosperm species shows that transposable element genes are the main sources of sRNA in annual species (Arabidopsis thaliana and Oryza sativa) and few genes were identified with high read counts of 21-nt size class. On the other hand, top ranking genes with the highest numbers of sRNA hits were annotated as NBS-LRR resistance genes and retroelements in perennial species (Populus trichocarpa and Vitis vinifera), with a predominant role in 21-nt sRNA. This divergent pattern might reflect a mechanism to limit runaway transcription of these genes in species with rapidly expanding NBS-LRR gene families. Alternatively it might reflect variation in a counter-counter defense mechanism partly caused by their differences in life history traits, for instance perennials have longer generation time and will certainly encounter a larger array of pathogens than annuals.
3. Conclusion

Our studies on multilocus polymorphism patterns in spruce are among the first to apply model-based simulation and multi-species comparison to infer the speciation and demographic history. We confirmed that spruces share quite extensive trans-species polymorphism. One source is the shared ancestry as the split time is relatively short if scaled by the generally large effective population sizes of spruce species. Migration could be another possible source as estimate of gene flow is larger than one from *P. glauca* to *P. abies* and from *P. schrenkiana* to *P. wilsonii*. Though standard isolation with migration model could not be rejected, statistics such as Tajima’s $D$ and Fay & Wu’s $H$ imply various possible demographic scenarios for most species. Simulations based on population expansion models improved the performance of goodness-of-fit tests. The growth of spruce suggested by nuclear DNA could trace back to millions of years ago, much more ancient than estimates based on fossils or organelle DNA. It is possible since nuclear DNA should keep more information that had been distorted by random drift or selective sweeps in organelle DNA.

In addition to divergence history, we also studied the geographic variation of bud set (or growth cessation) in spruce by sampling two latitudinal clines, one centered on Scandinavia and the other along the Yenisei river in Siberia. After correcting for the effect of population genetic structures, we successfully identified genes under local selection in both species, such as *GI*, *PRR7* and *FTL2*. This could be an indication that parallel evolution occurs through natural selection on groups of genes within related pathways. Interestingly, two mutations in *GI* protein coding area and *FTL2* promoter are shared by both species and exhibit the strongest signals of selection. Clinal variation at gene expression level was also detected in *FTL2* and provided even stronger evidence of local adaptation.

Using new sequencing technology, we are able to get the first assembly of needle transcriptome and build a collection of small regulatory RNA in Norway spruce. From the point of view of comparative genomics, spruce differs from angiosperms in gene expression differentiation, synonymous mutation rate, evolutionary rate in protein coding sequences, sources and targets of sRNA. At that stage, although our research in spruce genome is merely a piece of a jigsaw, however, it has gave us a glance of spruce transcriptome evolution and transcripts regulation, which cannot be simply extrapolated from those of model plant species. To get a comprehensive understanding would require a full characterization of the genome as well as physiological and functional studies combined with ecological ones.

In summary, in this thesis we have attempted to combine studies of multilocus DNA, transcriptome and regulatory RNAs, in single or multiple spruce species, in order to obtain a multi-faceted view of the way species arise and
evolve. Our efforts should contribute to the understanding of evolution in non-model systems and with the fast development of new sequencing technology and theoretical models in population genetics we should soon be able to have a more comprehensive picture of the effects of demographic history and natural selection at the genomic scale.
Evolution är den process genom vilka nya arter bildas och andra dör ut. Ur ett genetiskt perspektiv omfattar evolutionen två huvudsakliga processer: hur innehållet i den genetiska koden är organiserat och hur denna kod realiseras. Ur ett populationsgenetiskt perspektiv kan evolution betraktas som resultatet av naturlig selektion och slumpmässig genetisk drift. Även om selektionen inte verkar direkt på DNA, så lämnar den spår i form av ett specifikt allelfrekvensspektrum; till exempel så indikerar ett överskott av sällsynta genetiska varianter negativ selektion, medan ungefär lika frekvenser av olika genetiska varianter antyder balanserande selektion. Men de flesta av dessa klassiska frekvensmönster kan också vara resultatet av olika demografiska händelser, ensamt eller i kombination med selektion. Så kan ett överskott av sällsynta alleler eller lika allelfrekvenser också förklaras av en historiskt expanderande population, respektive perioder av kraftig reduktion av populationsstorleken. För att särskilja mellan effekter orsakade av selektion och demografiska händelser är det nödvändigt att studera fler genetiska locin. Effekten av demografiska händelser kommer då att visa sig i en likartad signal över större delar av genomet, medan spår av selektion i förstone kan förvántas i mindre grupper av gener som kan associeras till specifika fenotyper av adaptiv betydelse. Denna avhandling avser att studera den evolutionära processen i gran (P. abies), från arbildning till lokal anpassning, på såväl DNA- som RNA-nivå. Gran utgör en av de viktigaste boreala trädarterna och utgör den dominerandearten över stora skogsytor på det norra halvklotet. Den långa generationstiden och förekomsten av väldiga populationer med stort utbyte av genetiskt material gör granen till ett unikt system för det populationsgenetiska studiet av demografisk historia och naturlig selektion. Ett flertal granarters fylogenetiska släktspark och historia har tidigare studerats med hjälp av organell-DNA. I artiklar I och II använde vi istället multipla nukleära locin och modelbaserade simulationer i syfte att försöka bestämma tidpunkter för arbildning och graden av genflöde mellan olika granarter. Våra resultat visar att flertalet genetiska varianter som återfinns i fler arter kan spåras tillbaka till den gemensamma stamfaderen till två granarter, även om också genflöde mellan specifika par av arter förekommit. Vår jämförelse mellan olika modeller visade också att mer komplexa demografiska modeller bättre förklarade vårt data. Studiet av multipla locin antyder att många granarter genomgått en långsam populationsökning som startat långt tidigare än vad som uppskattats genom studiet av fossil material och organell-DNA. En förklaring till detta är att de stora populationerna bidragit till bevarandet av mer information i nukleärt DNA, medan slumpmässig genetisk drift och selektiva svep kan ha reducerat och förvrängt informationen i organell-DNA.
Artbildning kan betraktas som det slutgiltiga resultatet av genetisk diversitet mellan populationer, medan reducerat genflöde orsakat av geografiskt avstånd kan utgöra en av orsakerna till att diversiteten i det förstone uppstått och bibehållits. Geografisk variation i genetisk diversitet har dokumenterats i en mängd djur- och växarter. Granens stora utbredningsområde och sentida postglaciala rekolonisering gör arten till en ideal modell för studiet av geografisk variation och lokal anpassning. I våra studier kunde latitudinell variation i knoppsättning bekräftas i parallelle kliner hos två euroasiatiska granarter (artikel III och IV). Analysen av variationen i allelfrekvenser identifierade ett fåtal gemensamma kandidatgener under lokal selektion, bland annat i de cirkadiska- och fotoperiodiska nätverken. Mutationer i promotorregionen hos GI, PRR7 och FTL2 antyder att parallel evolution tycks förekomma oftare mellan gener involverade i relaterade metaboliska nätverk, än mellan mutationer inom enskilda specifika kandidatgener. Expressionsstudier av dessa kandidatgener bekräftade deras variation också på transkriptionsnivå.

I jämförelse med DNA, så utgör proteinkodande och regulatoriskt RNA intermedierande fenotyper, som befinner sig närmare de fenotypiska karaktärer som den naturliga selektionen direkt verkar på. Med hjälp av modern sekvenseringssteknologi genomförde vi i artikel V och VI, en så kallad de novo sammanställning av grantranskriptomet i barr, samt en preliminär karakterisering av små regulatoriska RNA, så kallade siRNA. Genom att använda molekylära evolutionära metoder och komparativ genomik upptäckte vi att gran särskiljer sig på en mängd viktiga punkter: uttrycksmönster i cirkadiska klockgener, synonyma mutationshastigheter, graden av evolutionär förändring i proteinkodande RNA, samt i mål- och ursprungssekvenser för siRNA och miRNA. Dessa karaktärsdrag skiljer sig från vad som observerats i blommande modellväxter som A. thaliana, men återfinns delvis i andra vindpollinerade arter med långa generationstider.

Det arbete som presenteras här utgör ett litet steg mot en ökad förståelse av evolutionen av gran. Mycket arbete, bland annat i form av detaljerad genomnotering, funktionella studier och jämförande analyser av fler arter är nödvändigt för att skapa en mer fullständig bild.
5. Acknowledgements

This thesis and the papers included cover all the research work done during my four-year Ph.D. study. I would not have been able achieve this without the help of many people and I would like to give my sincerest appreciation here.

First I would like to thank my supervisor Martin Lascoux for all his support and encouragement. He introduced population genetics to me and gives me enormous help and guidance, especially for manuscript preparation. His optimism, wisdom, patience and life philosophy win my highest respect. I also want to thank Thomas Källman who acts both as a teacher and a friend of mine. He has always been helpful to me.

Second, I want to thank all the people who helped with sampling, lab work and analyses. Many thanks to my co-supervisors Ulf Lagercrantz and Niclas Gyllenstrand for their brilliant work in experimental design, measurement, discussion and analysis. I would also like to give my special thanks to the three "bad boys": Michael Stocks, Yoshiaki Tsuda and Ma Xiaofei. My research life would have been much more boring without them. Thanks to Kate, Hanna, Sofia, Johan and Anna for making a great study atmosphere in our group. Thank Kalle for his elegant translation of "Svenska sammanfattning" in this thesis. Thank Tanja for helping me get started in the lab and teaching me to keep organized. Thank Kerstin for helping with chemical, equipment and technique in the lab and Rose-Marie for keeping the program going. Thanks to all the administrators and teachers at IBG for all the help even before my first arrival at Uppsala.

Special thanks to all my friends, Chuan, Ren and Lin for all the traveling, gyms, Friday badminton, buffé, games and jokes. You were all my after-class life during my stay in Sweden. Finally, thank my family for your love and support, my Papa and three most important ladies, my Mama, my wife and my little princess.
References


ERIKSSON, G., 2008 Picea abies Recent Genetic Research. Uppsala: Department of Plant Biology and Forest Genetics, SLU.


SEPPÄ, H., H. J. B. BIRKS, A. E. BJUNE, AND A. NESJE, 2010 Current continental palaeoclimatic research in the Nordic region (100 years since Gunnar Andersson


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 947

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology.