TRANSCRIPT PROFILING AND GENOME MAPPING IN BLACK SPRUCE (***Picea mariana***)

by

Ishminder K. Mann

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
December 2006

Your file Votre référence
ISBN: 978-0-494-27182-7
Our file Notre référence
ISBN: 978-0-494-27182-7

TO GOD.......

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Conifers represent an ecologically and economically important group of extant gymnosperms. They are primarily distributed in forests of the Northern-Hemisphere, supplying timber, pulp and resin resources. *Picea* (spruce) and *Pinus* (pine) are the two major genera in the family Pinaceae of conifers. Spruce genomics research is challenging because of the large genome size, the abundance of highly repetitive elements in its genome, paucity of sequence data and lack of advanced generation pedigrees. The results reported are a part of an active research program on structural and functional genomics of black spruce (*Picea mariana*), red spruce (*Picea rubens*) and their hybrids related to growth and adaptation to climate change being conducted under the leadership of my supervisor Dr. Om Rajora. The research objectives were to (1) develop an expressed sequence tag (EST) resource, (2) identify differentially-expressed genes in needle versus cambium tissues, and (3) construct a high density genetic linkage map, which could be used in mapping of traits related to growth and adaptation in black spruce.

An EST resource has been generated in black spruce by partially sequencing cDNA clones from a standard cDNA library developed from needles. Approximately 13% of the ESTs from black spruce showed no identity with sequences in the available EST database at the National Centre for Biotechnology Information (NCBI). A total of 586 transcripts coding for proteins with known functions have been annotated. The ESTs generated were compared with the EST databases from representative species belonging to conifers, monocots and eudicots to determine the extent of EST similarities. The ESTs from black spruce showed higher sequence similarities with conifers followed by eudicots than with monocots.

Transcriptome profiling of needle and cambium tissues was conducted. Differentially-expressed transcripts were identified in needle versus cambium tissues of black spruce via analysis of ESTs and dot-blot hybridization of cDNA clones from subtracted cDNA libraries. Transcripts coding for proteins involved in disease and stress response mechanisms, and cell structure were predominantly expressed in the cambium tissues, whereas those involved in photosynthesis, energy and metabolism were predominantly expressed in the needle tissues.

A near-saturated genetic linkage map of black spruce was developed using a three-generation outbred pedigree and amplified fragment length polymorphism (AFLP), selectively amplified microsatellite polymorphic loci (SAMPL), expressed sequence tag polymorphism (ESTP) and microsatellite (mostly cDNA based) markers. The consensus map had 941 markers distributed over 12 linkage groups, and covered more than 96% of black spruce genome. The mapped markers included 695 AFLPs, 213 SAMPL, 30 microsatellites and 3 ESTPs. Total estimated map length was 1898 cM, with an average of one marker every 2 cM. This is the first genetic linkage map of *Picea mariana*, and the first of its kind in the genus *Picea* based on a three-generation outbred pedigree.

LIST OF ABBREVIATIONS AND SYMBOLS USED

1C	Haploid genome
^{o}C	Degrees Celsius
µg	Micrograms
µL	Microlitres
µM	Micromolar
AFLPs	Amplified fragment length polymorphisms
BLAST	Basic local alignment search tool
bp	Base pairs
cDNA	Complementary deoxyribonucleic acid
cM	Centimorgan
dbEST	Database for expressed sequence tags
DNA	Deoxyribonucleic acid
dNTP	Deoxy-nucleoside tri-phosphate
EST	Expressed sequence tag
ESTPs	Expressed sequence tag polymorphisms
E	Expect value
et al.	*et alia*
F_1	First filial generation
F_2	Second filial generation
g	Gram(s)
K	Kosambi
LG	Linkage group

MAS	Marker assisted selection (or marker aided selection)
min	Minutes
mM	Millimolar
mRNA	Messenger ribonucleic acid
MT	Metallothionein-like
NCBI	National Centre for Biotechnology Information
ng	Nanogram
NR	Non-redundant
ORF	Open reading frame
PCR	Polymerase chain reaction
pg	Picogram
QTL	Quantitative trait loci
RAPDs	Random amplified polymorphic DNA
RNA	Ribonucleic acid
Rubisco	Ribulose-1,5-biphosphate carboxylase
s	Seconds
SAMPL	Selectively amplified microsatellite polymorphic loci
SNP	Single nucleotide polymorphism
SSH	Suppression subtractive hybridization
SSR	Simple sequence repeat (alternative for microsatellite)
TGOP	Three-generation outbred pedigree
UTR	Untranslated region

ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude to my supervisor Professor Om P. Rajora for his strong support, guidance and invaluable advices. I am also grateful to my committee members, Drs. Jonathan M. Wright and Robert G. Latta for their time, valuable help and suggestions.

I greatly appreciate our project collaborators John Major and Dr. Alex Mosseler for fruitful discussions and serving as committee members in the initial years of my Ph.D. program. I am also indebted to Dr. Kris Morgenstern for making F_1 crosses; Dr. Tim Boyle for making F_2 controlled crosses; Debby Barsi, Moira Campbell, and Stephanie West for raising, cloning, and maintaining the mapping population and plant material collection and sending the material as and when needed. I would also like to thank Dr. Daoquan Xiang for constructing the cDNA library from needles, which made me confident to develop more cDNA libraries in the black and red spruce structural and functional genomics project. I am also thankful to summer students, Taralynn Clunney, Waleed Abousamak and Brent Higgins for their help with EST sequencing in our lab. I would also like to thank Dr. Joy K. Roy, Nicholas Buckley and Betony Rowland for assisting in genotyping of the mapping population using microsatellite and ESTP markers and Dr B. Y. Kang for partly assisting in data scoring of SAMPL markers.

I am thankful to Dr. Rama Singh, Genome Atlantic Sequencing Platform, Institute for Marine Biosciences, NRC, Halifax, Canada for sequencing the clones from the suppression subtractive hybridization cDNA libraries. I am also grateful to Drs. D. S. Brar, M. Johnston and T. Platt for their suggestions and constant encouragement.

My special thanks, "a big thank you" to Dr. Paul Stothard, Canadian Bioinformatics Help Desk, for writing the Perl scripts to meet my research needs. Thanks also go to my friends Alan Bland, Alok Gupta, Anitha Nair, Christopher Kozela, Christina Popescu, Lisa O'Connell, Michelle Connolly, Natascha Forneris, Olga Hrytsenko and Stan Bashahkhanov for their support and friendships.

This work would not have been possible without the financial support of a Dalhousie University Graduate Scholarship, Natural Sciences and Engineering Research Council Strategic Project Grant and Genome Canada funding to Professor Om P. Rajora and Dr. Patrick Lett fund in the Department of Biology.

Lastly, I want to express my gratitude to my entire family especially my husband, Lakhvinder, and my mom and dad, whose understanding, encouragement and patience have made this work possible.

CHAPTER 1

INTRODUCTION

Knowledge of the genome structure, function and evolution at various levels of biological organization is necessary to understand the tremendous structural and functional diversity of living organisms, especially higher plants. This knowledge is also essential for understanding the genomic basis of evolution and genetic basis of many traits. The genome sizes vary enormously among eukaryotic species by a factor of about 3,300 and 1,000-fold, in animals and land plants respectively (Gregory 2005; Bennett and Leitch 2005). Recent advances in molecular and cellular biology have opened up new opportunities to understand the genome structure, function and evolution at various levels of biological organization. Characterization of genomes, genomic and expressed sequence tag (EST) sequences, have discovered numerous species-specific features, such as genome size, gene number, gene duplication, transposable elements and interspecific syntenic relationships (Stirling *et al.* 2003; Ku *et al.* 2000). As many of the functional regions of the genome are well conserved and these regions have been successfully used for annotation in eukaryotic genomes (Miller *et al.* 2004).

Comparative analysis of genomic, cDNA and protein sequences among different organisms provides valuable insights into organismal function, adaptation and evolution, and can assist in the construction of the so-called "tree of life". In addition, using EST sequences from a specific tissue for expression profiling helps us to identify the candidate genes underlying biological processes specific to that tissue. Genome maps provide an invaluable basic and applied resource. The occurrence of segregating polymorphic DNA loci has been used for the development of linkage maps in numerous species. These linkage maps serve as valuable tools for comparative genome mapping, and identification of genes and genomic regions underlying traits of interest, and can be used for marker-assisted selection (MAS). This is especially important for genetic improvement of certain species in which conventional breeding approaches are time consuming and resource intensive, owing primarily to long generation cycles and difficulty in achieving significant improvements in complex traits. This is particularly true for the genetic improvement of commercially important forest trees.

Forest trees are exceptionally important both ecologically and economically, as they are a major component of forest ecosystems and provide raw materials for a diverse array of products. In Canada, forest and other wooded lands cover approximately 46% of the landmass, of which softwoods constitute 66% and the rest are mixed woods (22%) and hardwoods (12%) (The State of Canada's Forests 2005–2006). Among softwoods/conifers, the major tree species are spruce (*Picea*), pine (*Pinus*), and fir (*Abies*) (The State of Canada's Forests 2005–2006).

The genus *Picea* is the second largest genus after *Pinus* in the family Pinaceae of conifers (Farjon 1990). Spruce species are ecologically important as they are major component of boreal and temperate forests in Canada. Spruce is the most widely harvested conifer species in Canada and is the bread and butter for the Canadian forest industry. Five species of spruce are native to Canada; Sitka spruce (*Picea sitchensis*) and Engelmann spruce (*Picea engelmannii*) in the west, two transcontinental species, white spruce (*Picea glauca*) and black spruce, (*Picea mariana*) and red spruce in the east (*Picea rubens*) (Farrar 1995).

Black spruce is a widespread transcontinental species of the North American boreal and temperate forests (Viereck and Johnston 1990). It is a diploid species with a haploid chromosome number (n) of 12 (2n=24), like most other Pinaceae members. The estimated genome size of black spruce is large (1C = 15.8 pg; Ohri and Khoshoo 1986; accessed on October 30th, 2006; http://www.rbgkew.org.uk/cval/ homepage.html), with an approximate 2C genome length of about 31,000 mega base pairs (accessed on October 30, 2006; http://www.rbgkew.org.uk/cval/homepage.html), which is approximately 100 times larger than that of the *Arabidopsis thaliana* genome (1C = 0.16 pg, Bennett *et al.* 2003). Black spruce is an early successional species, characterized by longevity and a mixed mating system, with high outcrossing rates (Boyle and Morgenstern 1986). It is one of the most important trees in Canada for the production of pulp and paper, and is a prime reforestation species in Canada (Morgenstern and Wang 2001). Increasing use of black spruce and rapid climate change necessitates acceleration of its genetic improvement programs, particularly for productivity, health and adaptation under changed climatic conditions. Genomics tools and technologies can assist in accelerating the otherwise slow conventional genetic improvement of black spruce.

Genomics research in conifers has lagged behind that in agricultural crops, since they represent a difficult experimental material. The long generation times required to reach sexual maturity challenge the development of advanced generation pedigrees required for mapping of traits of interest in conifers. Additionally, genetic transformation and regeneration is generally difficult. Even, relatively simple molecular procedures, such as isolation of good quality deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), have been challenging and time-consuming until early 1990's. These long generation plants need MAS to enhance genetic gains in breeding programs for complex traits, such as wood properties, disease and pest resistance and tolerance to abiotic stresses (Neale and Williams 1991).

The construction of a genetic linkage map, an important framework for genetic improvement of the traits of interest through MAS, requires the development of markers. Highly informative markers from the expressed regions of the genome are preferably required for linkage mapping, which can also be used in other forest genetics improvement programs. This requires sequence data from the expressed regions of the genome.

Rapid progress made in sequencing technology and collaborative research efforts at the international level have made available full genome sequences for a number of eukaryotic species. For higher plants, the complete genome of the model plants *Arabidopsis* (Arabidopsis genome initiative 2000), rice (*Oryza sativa*) (Yu *et al.* 2002; Goff *et al.* 2002) and poplar (*Populus trichocarpa*) (Tuskan *et al.* 2006) have been sequenced, providing invaluable information on the composition of plant genomes. Several studies focusing on comparative genomics among different plant species (Sasaki *et al.* 2002; Vandepoele *et al.* 2002) have shown comparative genomics approaches to be very useful to determine conservation of gene content and order among different species and to identify orthologous loci, which would be important for understanding the genomic evolution (Stirling *et al.* 2003; Ku *et al.* 2000; Goff *et al.* 2002; Yu *et al.* 2002). Therefore, to compare coniferous species with other plant species at genetic and molecular levels, sequence data is required. However, obtaining the whole genome sequence data from a coniferous tree is prohibitive at this time owing to cost, given that coniferous genomes are amongst the largest of all known plant genomes (Leitch *et al.* 2001; Murray 1998) and contain a large proportion of repetitive DNA (Kriebl 1985; Elsik and Williams 2000). Because sequencing an entire genome of a conifer is currently out of the realm of feasibility, alternative, more cost-effective approaches have been used to identify the transcriptome, which represents a small proportion of the total genome.

The most widely used approach for surveying transcribed portions of the genome is expressed sequence tag (EST) sequencing. An EST is a complementary DNA (cDNA) sequence derived from messenger ribonucleic acid (mRNA) and is usually 400-500 base pairs (bp) in length (Adams *et al.* 1991). Thus, for the same length of DNA sequenced, ESTs would detect a greater number of genes compared to arbitrary genomic sequences of similar length (Rudd 2003). EST sequencing is a rapid and cost-effective method for gene discovery and annotation, especially in conifers, given the size and complexity of the conifer genome. For example, in the family Pinaceae, the mean value of the genome size (1C) is approximately 23.37 pg (Murray 1998; Leitch *et al.* 2001; accessed on October 30th, 2006; http://www.rbgkew.org.uk/cval/homepage.html).

Another advantage of EST sequencing is that, cDNA clones are sequenced randomly, making it possible to compare the sequence data and assemble contigs after sequencing and can yield the entire sequence for full length cDNAs, representing the entire coding sequence of the corresponding gene. The number of ESTs that represent the same gene in a given library is a rough indication of the expression level of the gene in the tissue from which the library was derived. Therefore, EST data set can also provide information on gene expression in different plant tissues by comparing the EST frequencies in different cDNA libraries (Fernández *et al.* 2003). The needle and cambium tissues are sites for important functional pathways underlying growth, and adaptation in forest trees. Sequencing of ESTs from the needle and cambium tissues can identify the genes underlying growth and adaptation in conifers. As such, sequencing of ESTs from different tissues is as an efficient method for gene discovery in species with large genome size and provides a preliminary insight into gene expression profiles in a given tissue. However, to further advance our understanding of the genome organization, the genetic linkage map is required in a species.

For construction of a genetic linkage map, highly informative markers are required. Several types of polymerase chain reaction (PCR)-based DNA markers are available, each having various relative advantages and disadvantages. Highly informative markers can be developed from ESTs, because they target expressed regions of the genome, are codominant, often multiallelic and highly reproducible among different laboratories and across different genetic backgrounds. However, due to the occurrence of multigene families (Kinlaw and Neale 1997), single locus EST-based markers such as ESTPs are difficult to develop in conifers. The amplified fragment length polymorphism (AFLP; Vos *et al.* 1995) and selective amplification of microsatellite polymorphic loci (SAMPL; Witsenboer *et al.* 1997) markers have a high multiplex ratio and reproducibility, and therefore can be used to achieve sufficient coverage of the large genome of black spruce.

1.1 Research Objectives

As part of a major research program, being conducted by my supervisor Dr O. P. Rajora and his collaborators, on structural and functional genomics of black spruce and red spruce related to growth and adaptation to climate change, the work described in this thesis was undertaken with the following objectives –

1. Development of an EST resource in black spruce. The first objective was to generate ESTs from a standard needle cDNA library in black spruce. To achieve this goal, a standard cDNA library was constructed using RNA extracted from needles of three greenhouse-grown seedlings of black spruce. Sequence comparisons of ESTs from black spruce were conducted with publicly available databases and with representative species from monocots, eudicots and conifers to determine functional sequence similarity among extant seed plant groups. In addition, ESTs were also used for the development of expressed sequence tag polymorphism (ESTP) markers, and microsatellite markers for the construction of a genetic linkage map.

2. Identification of differentially-expressed genes between needle and cambium tissues. The second objective was to identify differentially-expressed genes between needle and cambium tissues using suppression subtractive hybridization (SSH) method and dot-blot hybridization of cDNA clones from the subtracted cDNA libraries. Needle and cambium tissues are sites for vital functional pathways in trees. To identify the differentially-expressed genes in needle and cambium tissues, four SSH cDNA libraries were constructed using RNA extracted from needle and cambium tissues obtained from the parents of the mapping population.

3. Development of AFLP, SAMPL and ESTP markers and construction of a genetic linkage map using polymorphic AFLP, SAMPL, ESTP and microsatellite markers. The third objective was to develop markers from expressed and anonymous genomic regions of black spruce, which can be used to construct a high density genetic linkage map in black spruce. These markers were tested for polymorphism between parents of the mapping population. The polymorphic AFLP, SAMPL, ESTP and microsatellite markers were used for genotyping the progeny followed by the construction of a genetic linkage map.

The thesis is organized in three major chapters according to the above three research objectives. The first deals with the development of the EST resource; the second describes the identification of differentially-expressed genes between needle and cambium tissues; and the third chapter presents the results on the development of AFLP, SAMPL and ESTP markers followed by construction of a genetic linkage map using polymorphic AFLP, SAMPL, ESTP and microsatellite markers.

CHAPTER 2

DEVELOPMENT OF AN EST RESOURCE IN BLACK SPRUCE

2.1 INTRODUCTION

Gene discovery has become an indispensable part of genomics research in species with large sized genomes. Sequencing of randomly selected cDNA clones (expressed sequence tags, ESTs) has been used as an efficient method for gene discovery (Adams *et al.* 1991). As of October 6th, 2006, there were approximately 38.95 million (38,953,178) ESTs from 1207 species available in the database for ESTs (dbEST) of GenBank at the National Centre for Biotechnology Information (NCBI; http://www.ncbi.nlm.-nih.gov/dbEST/). The number of ESTs is continuously increasing in dbEST, creating an important resource for development of molecular markers, generation of tags for map-based cloning and providing insights into genome evolution.

With the availability of sequences from several angiosperm and gymnosperm species in public databases, comparative sequence analysis among different species has been conducted (Pavy *et al.* 2005a; Kirst *et al.* 2003; Stirling *et al.* 2003). Such sequence-based comparisons facilitate unraveling the sequence similarity between different species, identify sequences unique to a particular species, and suggest evolutionary relationships at different taxonomic levels. In addition, sequence-based comparisons help in the identification of coding sequences and conserved non-coding sequences with regulatory functions, and provide an understanding of genome composition (Frazer *et al.* 2003; Grant *et al.* 2000; Ku *et al.* 2000).

Although many more ESTs from several spruce species have been deposited in the dbEST at NCBI, no EST resource has been reported for black spruce. Among conifers, the major EST contributing species are loblolly pine (*Pinus taeda*; 329,517), white spruce (160,818) and Sitka spruce (80,789) (accessed on October 6th, 2006; http://www.ncbi.nlm.-nih.gov/dbEST/).

Few studies have been conducted in conifers in which comparisons of their ESTs with those from different groups of seed plants, for example *Arabidopsis,* rice (*Oryza* species) and poplar (*Populus* species) have been performed (Kirst *et al.* 2003; Pavy *et al.* 2005a). The significant differences in genome size (Leitch *et al.* 2001; Murray 1998) and phenotypic diversity between gymnosperms and angiosperms raises the question as to the extent to which they share the same genes. Yet to date, there is no published report in which ESTs from black spruce have been compared with the three major groups of the extant seed plants *i.e.* conifers, monocots and eudicots. The EST sequencing studies in conifers have focused mostly on transcripts derived from wood-forming tissues and xylem with the expectation that they would include informative sequences related to wood formation owing to the economic importance of wood (Allona *et al.* 1998; Kirst *et al.* 2003).

In this chapter, I describe the production of an EST database in black spruce and a comparison of this collection with EST databases from representative angiosperm and gymnosperm species. I have conducted comparisons between ESTs from black spruce with total dbEST available at NCBI to estimate sequence similarity, followed by functional annotation of ESTs from black spruce. In addition, to study the extent of conservation of the functional genome across seed plants, ESTs from black spruce were compared with species-specific EST databases, taking two or three major EST contributing species from each of three major extant seed plant groups, monocots [rice and wheat (*Triticum* species)], eudicots (*Arabidopsis* and poplar) and conifers (loblolly pine, white spruce and Sitka spruce).

2.2 MATERIALS AND METHODS

2.2.1 Plant Material and RNA Isolation

Three black spruce seedlings were established in the green house at Dalhousie University, Halifax. For RNA extraction, the needles were collected during day time from the green house grown seedlings. Total RNA was extracted from 2 g of fresh needles as described by Chang *et al.* (1993). Quality and quantity of the isolated RNA were determined using a spectrophotometer (SPECTRAmax PLUS, Molecular Devices Corporation, Sunnyvale, CA, USA) and extracted RNA was found to be of high quality (OD_{260}/OD_{280}=1.82). The quantity of isolated RNA was approximately 120 μg per g of the needle tissue used. The polyA RNA was purified using RNeasy Mini Kit (Qiagen Inc., Mississauga, ON, Canada) following the manufacturer's instructions.

2.2.2 Construction of a cDNA Library

The cDNA library was constructed using approximately 1 μg of polyA RNA isolated from needles. The Creator Smart cDNA Library Construction Kit (Clontech Laboratories Inc., Mountain View, CA, USA) was used for the construction of a cDNA library using polyA RNA from needles following the manufacturer's instructions. The oligodT primed cDNA inserts were directionally cloned in the pDNR-LIB vector

(Clontech Laboratories Inc., Mountain View, CA, USA) and XL-10 gold ultra-competent cells of *Escherichia coli* were transformed with vector containing cDNA inserts (Invitrogen Canada Inc., Burlington, ON, Canada). The transformed cells were plated on Luria Broth agar medium containing chloroamphenicol (30 µg/ml) and the colonies containing inserts were identified by blue/white screening. Transformed white colonies were handpicked and cultured overnight at 37°C in Luria Broth medium. Plasmid DNA was isolated from *E. coli* cells using the QIA Prep® Spin Miniprep kit (Qiagen Inc., Mississauga, ON, Canada). The quantity of each isolated plasmid DNA was determined by gel-electrophoresis on 0.8% (w/v) agarose with known amounts of undigested Lambda DNA (GIBCO BRL, Burlington, ON, Canada) as a standard.

2.2.3 Sequencing of cDNA Clones

The sequencing reactions were performed in a PTC-200 thermal cycler (MJ Research, Reno, NV, USA) using Thermosequenase Fluorescent Labeled Primer Cycle Sequencing Kit with 7-deaza dGTP (Amersham Pharmacia Biotech, Freiburg, Germany) according to the manufacturer's instructions using 1 µM of IRD labeled M13F (5′-AAA CAG CTA TGA CCA TGT TCA-3′) and M13R (5′-GTA AAA CGA CGG CCA GT-3′) sequencing primers. Sequencing reactions were standardized to obtain good quality sequences with larger read length by using different concentrations of plasmid DNA. Finally, the sequencing reaction containing 650 ng of template plasmid DNA provided the best sequencing results in terms of length and quality of the sequence reads. The PCR profile consisted of initial denaturation at 94°C for 3 min, followed by 30 cycles each of denaturation at 94°C for 30 s, annealing at 55°C for 30 s and extension at 72°C for 1 min, followed by a final soak at 10°C. The sequencing products were resolved using a LI-COR 4200L sequencing system which can sequence from both ends of a cDNA clone in a single sequencing reaction (LI-COR Biosciences, Lincoln, NE, USA). The sequences from 2486 cDNA clones were obtained from both directions. Out of 2486 clones, a total of 2122 cDNA clones produced 4244 ESTs representing 5′ and 3′ ends of cDNA sequences, and the remaining 364 clones yielded 364 ESTs represented by either 5′ or 3′ ends of cDNA sequences. Therefore, a total of 4608 ESTs were obtained from the black spruce needle cDNA library. The other-end sequences of the 364 clones were excluded from the analyses due to their low sequence quality values and/or their sequence read length below the selected minimum criterion of 100 bp.

2.2.4 Sequence Processing of ESTs

Before analysis, 4608 ESTs were checked manually to remove ambiguous base assignments using e-Seq base caller software (LI-COR Biosciences, Lincoln, NE, USA). Only 3124 out of 4608 ESTs were analyzed using the phred, cross_match and phrap programs (Ewing *et al.* 1998; Ewing and Green 1998). Vector sequences were trimmed using the cross_match program with settings of minmatch 12 and minscore 20. All 3124 ESTs were assembled together using the phrap program to obtain contigs and singletons. Quality analysis of the remaining 1484 ESTs was not possible due to loss of their standard chromatogram files; however, the vector sequences were clipped manually. For BLAST (Basic Local Alignment Search Tool) analysis, 4608 ESTs and contigs longer than 100 bp were used.

2.2.5 Sequence Similarity Analyses and Gene Annotation

Nucleotide (BLASTN) and protein (BLASTX) sequence similarities of ESTs from black spruce and their contigs were determined with the total dbEST (accessed on October 6th, 2006) and the non-redundant (NR) protein database (accessed on August 5th, 2005) available at NCBI using custom designed Perl scripts (Altschul *et al.* 1990). The Perl scripts were designed so that the BLASTN and BLASTX similarity results were obtained from only those sequences meeting the criteria of minimum bit score of 100 and alignment length of 100 bp. This stringent criterion was applied because the results from the BLASTX similarity analysis were further used for gene annotation.

Transcripts from needle cDNA library were annotated based on their BLASTX similarities. The putative genes were identified and then grouped into functional categories based on the literature search, and following the grouping of Kyoto Encyclopedia of Genes and Genomes website (http://www.genome.ad.jp/kegg/; Kanehisa and Goto 2000). The length of contigs in bp was plotted against the fraction of contigs showing BLASTX similarity.

2.2.6 Similarities of ESTs from Black Spruce with Species-Specific dbEST

The EST datasets from species representing conifers [white spruce (160,818), Sitka spruce (80,789) and loblolly pine (329,517)], monocots [rice (1,217,899) and wheat (876,068)], and eudicots [poplar (376,600) and *Arabidopsis* species (737,418)] were downloaded from dbEST at the NCBI. Species-specific BLASTN similarity searches for 5′ and 3′ end sequences and contigs were conducted against white spruce, Sitka spruce, loblolly pine, rice, wheat, poplar and *Arabidopsis* dbEST using blastall client of BLAST, with e value cutoff of 10^{-5}. To avoid biasing the similarity results with species-specific dbEST, given that Sitka spruce has the lowest number of ESTs (80,789) available in dbEST as compared to other species used in the present study, two more types of analyses were performed, in one case random pick and in other using their split data. For random picking and splitting of EST datasets from species-specific dbEST, custom designed perl programs were used. In the first case, *i.e.* random picking, 25,000 ESTs from each of the above species were selected randomly ten times from their total ESTs and BLASTN sequence similarity analysis of ESTs and contigs from black spruce were conducted with sequences randomly selected from the species-specific dbEST. In case of split data, BLASTN sequence similarity analysis of ESTs and contigs from black spruce were conducted two times with white spruce, ten times with wheat (876,078), fifteen times with rice (1,217,899), nine times with *Arabidopsis* (737,418), four times with poplar (376,600) and loblolly pine (329,517) by splitting equally into 80,789 ESTs (as Sitka spruce dbEST has 80,789 ESTs) of their respective species-specific dbEST.

2.2.7 Redundancy of ESTs

The redundancy of ESTs was estimated using the following formula: (Total number of ESTs – number of singletons) x 100/total number of ESTs. While the contig redundancy was calculated as (Total number of ESTs – number of singletons)/number of contigs (Kirst *et al.* 2003).

2.2.8 Identification of Open Reading Frames (ORFs) and Microsatellites

ORFs in ESTs from black spruce were identified based on BLASTX similarity analyses with the NR protein database. The ESTs from black spruce showed ≥ 80% similarity with available sequences in the NR protein database at NCBI and with the initiation codon at the 5′end were classified as putative ORFs. The ESTs containing microsatellite repeat motifs were identified manually and using RepeatMasker with *A. thaliana* as the DNA sequence source (accessed on October 30th 2005; http://www.repeatmasker. org/cgi-bin/WEB RepeatMasker).

2.3 RESULTS

2.3.1 Sequence Analysis

Out of a total of 4608 quality ESTs from black spruce, the numbers of 5′ and 3′ end sequences were 2465 and 2143, respectively. The average length of the 5′ end sequences was 600 bp ranging from 116 to 1263 bp, whereas that of the 3′ end sequences was 466 bp ranging from 100 to 980 bp. The quality of sequences obtained from cDNA libraries is of utmost importance when sequencing of randomly selected cDNA clones is conducted. The average length of total ESTs was 537 bp with an average quality value of 30.2 per base after phred analysis. From the total ESTs, 3124 sequences assembled to form 264 contigs and 916 singletons. The number of sequences assembled to generate a contig varied from 2 to 135 ESTs. The sequence length of contigs and singletons ranged from 149 to 1691 and 114 to 1463 bp, with an average value of 783, and 640 bp, respectively.

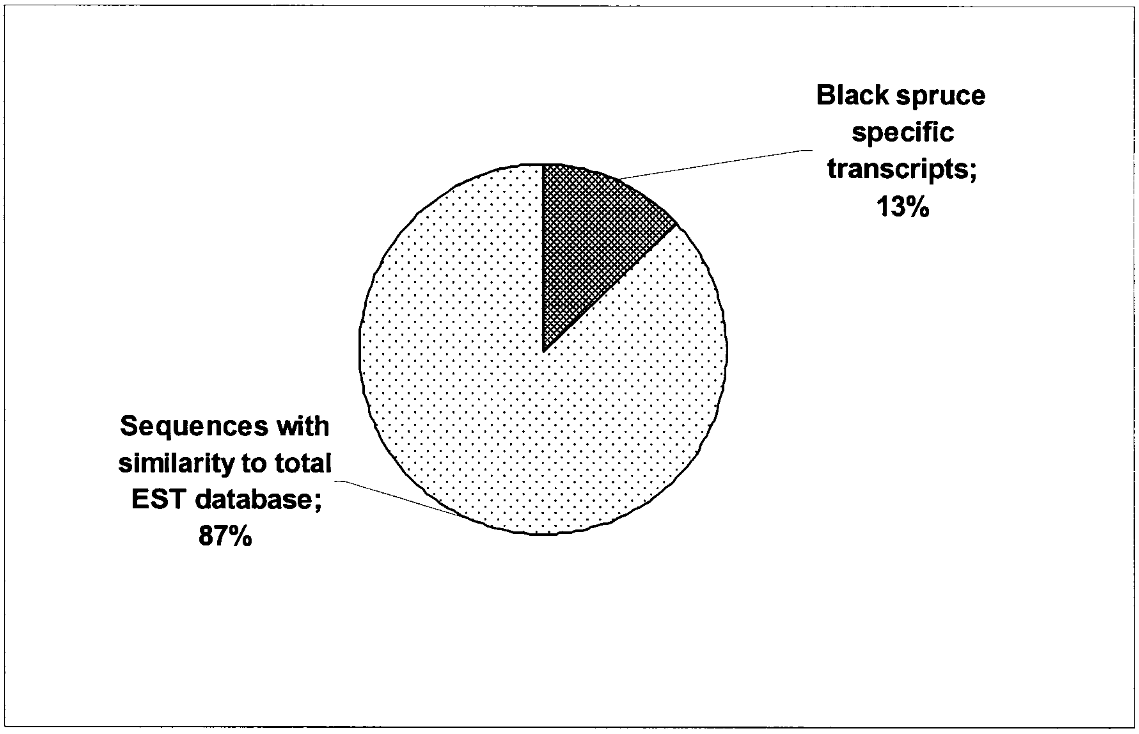
2.3.2 Sequence Similarity of ESTs from Black Spruce with Total dbEST

Out of 4608 ESTs, 4008 (87%) showed sequence similarity to publicly available dbEST at NCBI (Figure 2.1a). As might be expected, the majority of these 4008 ESTs showed sequence similarity to ESTs from *Picea* and *Pinus* species present in the total dbEST based on their first best hit (Figure 2.1b). Therefore, 600 (13%) ESTs from black spruce exhibited no sequence similarity with the available dbEST at NCBI, at the cutoff value of 100 for bit score and alignment length. The 5′ end sequences of ESTs showed higher sequence similarity (91.3%) than 3′ end sequences (82%) with total dbEST. The alignment length of the sequences showing similarity for 5′ end was higher (100-943 bp) than with 3′ end sequences (100-890 bp).

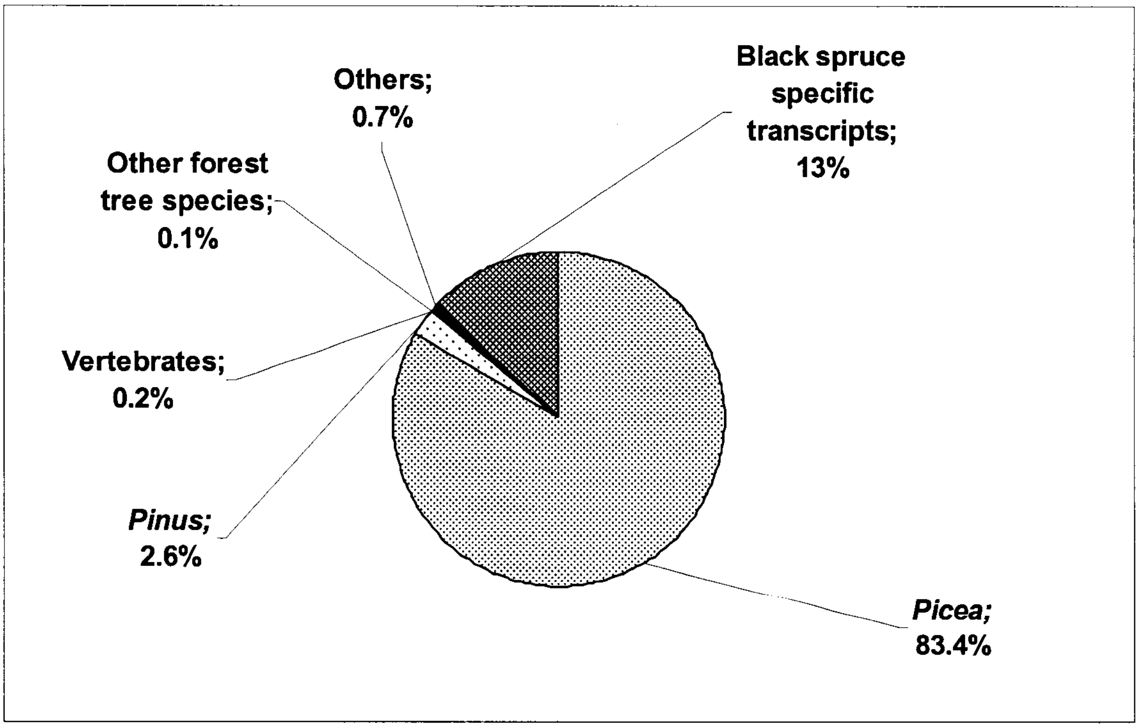
2.3.3 Functional Analysis of ESTs from Black Spruce

To identify the putative functions of ESTs from black spruce, sequences were compared with the NR protein database at NCBI. The BLASTX similarity analysis allowed annotation of 31.6% (1456) of black spruce transcripts. The alignment length of ESTs from black spruce showing similarity to NR protein database was higher with 5′ end sequences (150 to 945 bp) than with 3′ end sequences (138 to 771 bp). The distribution of e-values obtained in the BLASTX similarity analysis showed that the peaks of expect value centered on e^{-20} to e^{-40}. This value represents stringent criteria as most previously published studies of EST similarity adopt an expect value of less than

Figure 2.1 Sequence similarity of ESTs from black spruce derived from a needle cDNA library based on a BLASTN search of the dbEST at NCBI. (A) Showing percentage black spruce specific transcripts and, (B) showing similarity with different species based on the first best hit with the total dbEST.


Black spruce specific transcripts; 13%
Sequences with similarity to total EST database; 87%

(A)


Others; 0.7%
Black spruce specific transcripts; 13%
Other forest tree species; 0.1%
Vertebrates; 0.2%
Pinus; 2.6%
Picea; 83.4%

(B)

Table 2.1 Functional classification of black spruce transcripts coding for proteins with known functions and the number of ORFs identified.

Functional category	No. of transcripts coding for proteins with known functions	No. of ORFs
Protein synthesis	81	29
Energy and metabolism	71	6
Disease and stress response	57	8
Photosynthesis	41	7
Transcription & post-transcription	40	7
Protein destination and storage	37	8
Cell structure	36	6
Lipid biosynthesis and metabolism	32	0
Secondary Metabolism	22	2
Signal transduction	19	4
Transporters	13	3
Growth and development	11	1
Transposons	3	0
Expressed and unknown protein	123	16
Total	**586**	**97**

10^{-5}. From 1456 black spruce ESTs, 586 transcripts coding for proteins with known functions were identified (Table 2.1). Based on putative function(s), ESTs were further characterized by sorting into 14 functional categories. These functional categories provide a general overview of genes expressed in needles. They include genes involved in protein synthesis, energy and metabolism, disease and stress response, photosynthesis, transcription and post-transcription, protein destination and storage, cell structure, lipid biosynthesis and metabolism, secondary metabolism, signal transduction, transporters, transposons, expressed and unknown proteins (Table 2.1). Out of the 586 transcripts coding for proteins with known functions each of 367 was targeted by a single EST, whereas the remaining 219 were targeted by more than one EST, varying from 2 to 45 ESTs. For example, non-specific lipid transfer protein was targeted by 45 ESTs. The five predominant categories of transcripts coding for proteins expressed in needles according to this functional classification were: protein synthesis, energy and metabolism, disease and defense, photosynthesis and transcription and post-transcription (Table 2.1).

2.3.4 Similarities of ESTs from Black Spruce with Representative Gymnosperms and Angiosperms Species-Specific dbEST

Based on expect value cutoff of 10^{-5}, the ESTs from black spruce showing sequence similarity with species-specific dbEST from Sitka spruce, white spruce, loblolly pine, rice, wheat, poplar, and *Arabidopsis* were 82.6%, 88.5%, 83.6%, 25.2%, 21.6%, 33.4% and 31.8%, respectively. The 5′ end sequences showing similarity to species-specific dbEST varied from 27.6 % to 90.2% and those of 3′ end sequences varied from 14.8% to 86.6% (Figure 2.2). On the basis of ten times random pick of 25,000 ESTs from each species-specific dbEST the percentage sequence similarity of ESTs from black spruce with different species varied from 16.5% to 77.9%, the highest being with spruce species, as might be expected (Figure 2.3). On the basis of split databases in multiples of ~80,000 ESTs, the percentage sequence similarity for different species varied from 17.2% to 83.5%, the highest being with spruce species again (Figure 2.4). In all of the cases (total, random pick and split data from species-specific dbEST), the percent similarity values followed the same trend, being highest with conifers followed by eudicots and then monocots.

2.3.5 Similarities of Contigs with Total dbEST and Species-Specific dbEST

Out of a total of 264 contigs, 258 (97.7%) contigs showed sequence similarity to total dbEST based on BLASTN. Sequence comparisons of contigs with ESTs from Sitka spruce, white spruce, loblolly pine, rice, wheat, poplar and *Arabidopsis* showed 91.3%, 95.1%, 96.6%, 38.6%, 35.9%, 46.6%, and 41.7% were similar, respectively (Figure 2.2). Based on the ten times random pick of 25,000 ESTs from each species dbEST, the sequence similarity of contigs with Sitka spruce, white spruce, loblolly pine, rice, wheat, poplar, and *Arabidopsis* were 86.7%, 87.1%, 80.4%, 31.6%, 29.8%, 38.3%, and 38.4%, respectively (Figure 2.3). On the basis of split database in multiples of 80,789 ESTs from each species dbEST, the sequence similarity of black spruce contigs to that of Sitka spruce, white spruce, loblolly pine, rice, wheat, poplar, and *Arabidopsis* were 91.2%, 92.4%, 83.0%, 30.2%, 32.0%, 42.1%, and 36.4%, respectively (Figure 2.4). Contigs also followed the same trend for sequence similarity as unassembled ESTs with the total and species-specific dbEST, but showed higher percentages.

Figure 2.2 Percentage of ESTs from black spruce (5′ and 3′ end sequences) and contigs showing sequence similarity based on BLASTN searches with the species-specific dbEST from Sitka spruce, white spruce, loblolly pine, poplar, *Arabidopsis*, wheat and rice.

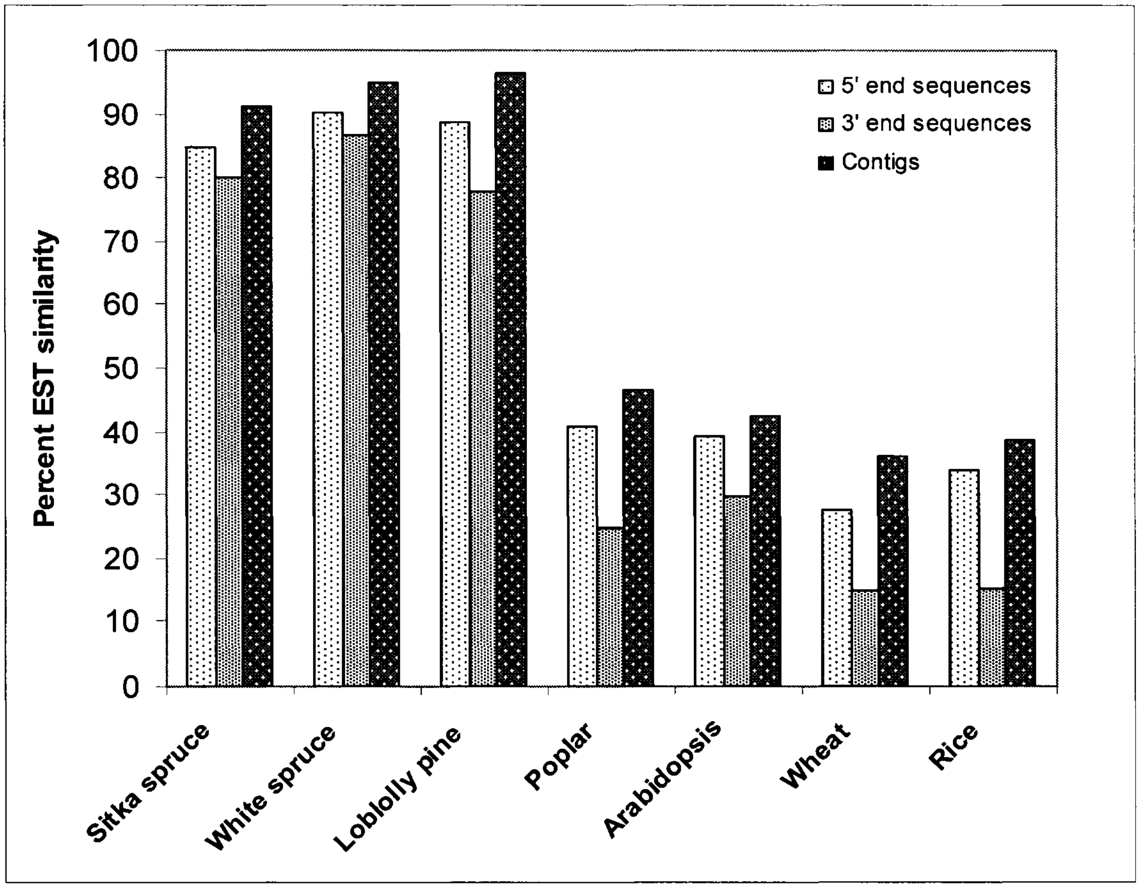

100
90
80
70
60
50
40
30
20
10
0
Percent EST similarity
5' end sequences
3' end sequences
Contigs
Sitka spruce
White spruce
Loblolly pine
Poplar
Arabidopsis
Wheat
Rice

Figure 2.3 Percentage of ESTs from black spruce (5′ and 3′ end sequences) and contigs showing sequence similarity based on BLASTN searches with the average of random picks of 25,000 from Sitka spruce, white spruce, loblolly pine, poplar, *Arabidopsis*, wheat and rice dbEST.

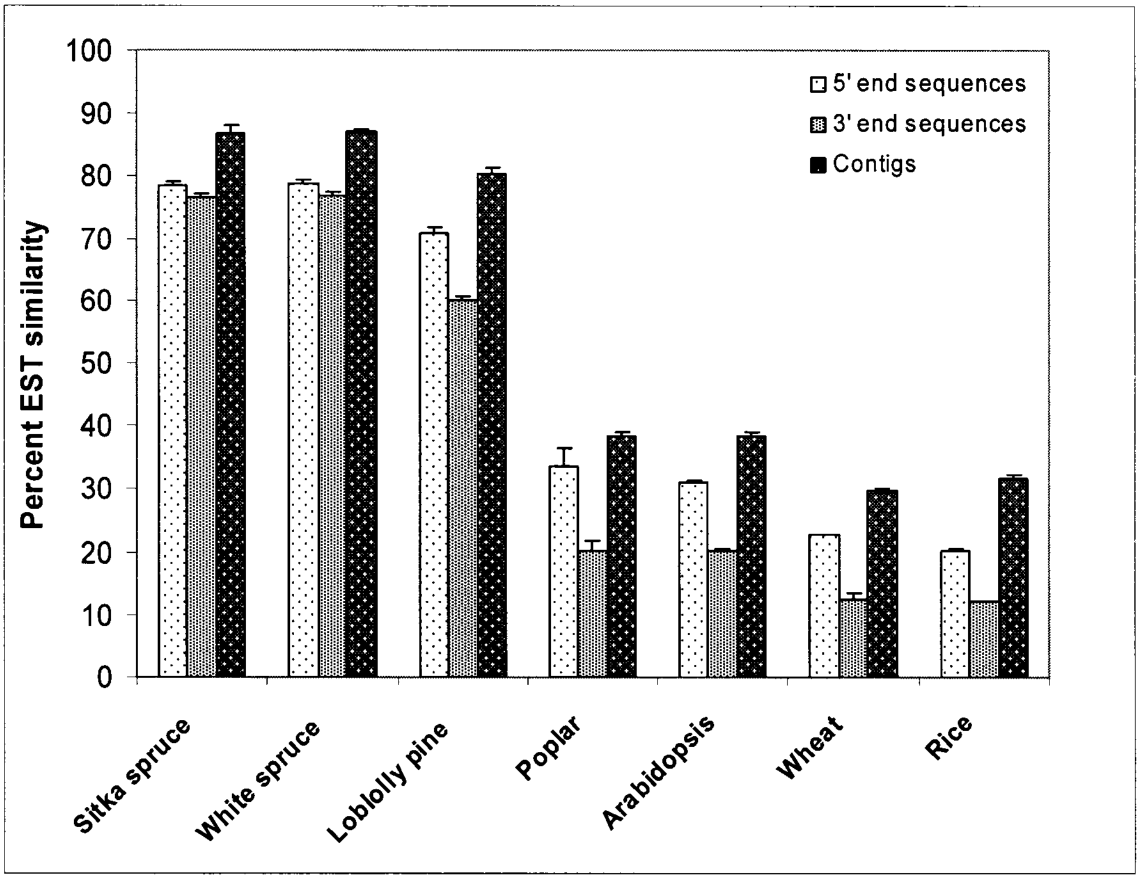
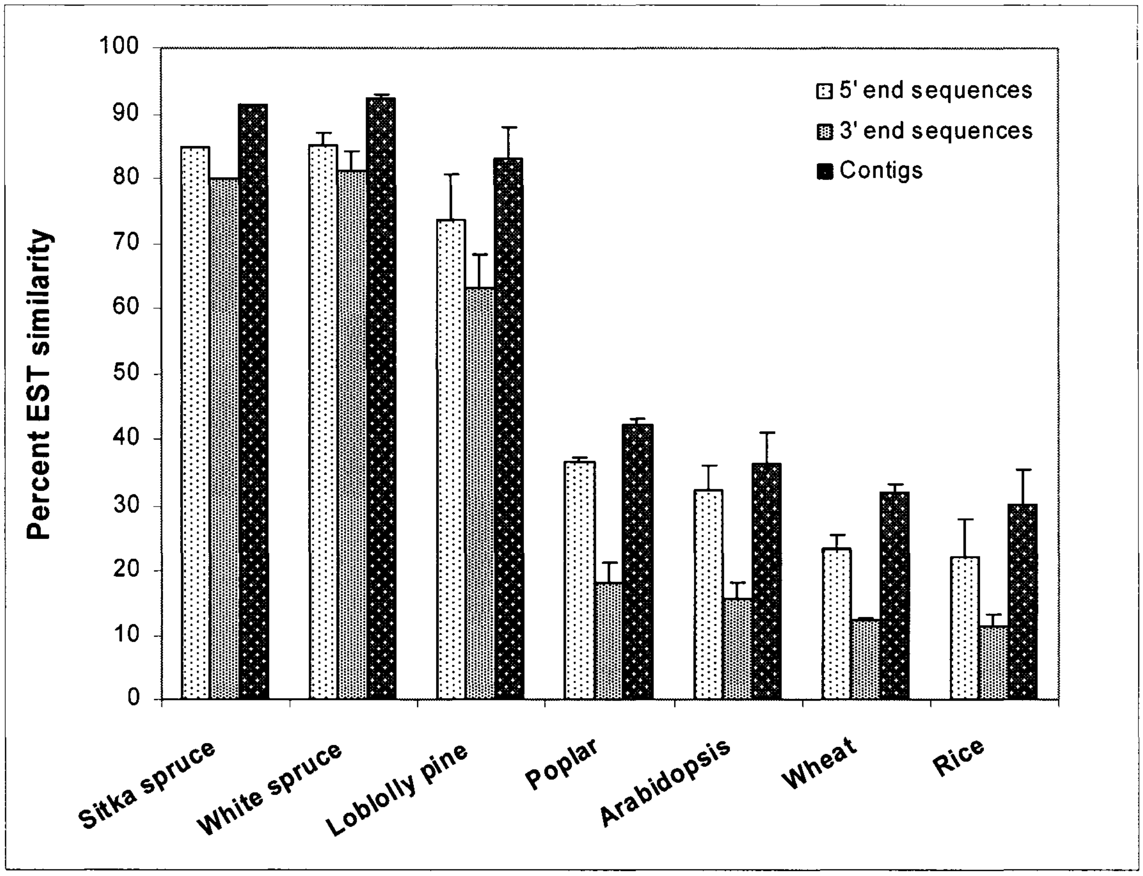

100
90
80
70
60
50
40
30
20
10
0
Percent EST similarity
□ 5' end sequences
▨ 3' end sequences
■ Contigs
Sitka spruce
White spruce
Loblolly pine
Poplar
Arabidopsis
Wheat
Rice

Figure 2.4 Percentage of ESTs from black spruce (5′ and 3′ end sequences) and contigs showing sequence similarity based on BLASTN searches with the average of split database of multiples of 80,789 from Sitka spruce, white spruce, loblolly pine, poplar, *Arabidopsis*, wheat and rice dbEST.


Percent EST similarity
100
90
80
70
60
50
40
30
20
10
0
5' end sequences
3' end sequences
Contigs
Sitka spruce
White spruce
Loblolly pine
Poplar
Arabidopsis
Wheat
Rice

2.3.6 Similarities of Contigs with the NR Protein Database

Based on BLASTX similarities with NR protein database at NCBI, 150 (56.8%) contigs were annotated. The percentage BLASTX similarity of contigs (56.8%) was higher than of unassembled ESTs (31.6%). Redundancy of contigs was 8.3%. With the increase in sequence length of contigs, the percentage of sequence similarity increased (Figure 2.5). The contigs being larger in size than unassembled ESTs and therefore, contain more information, hence showed higher sequence similarity with the NR protein database available at NCBI.

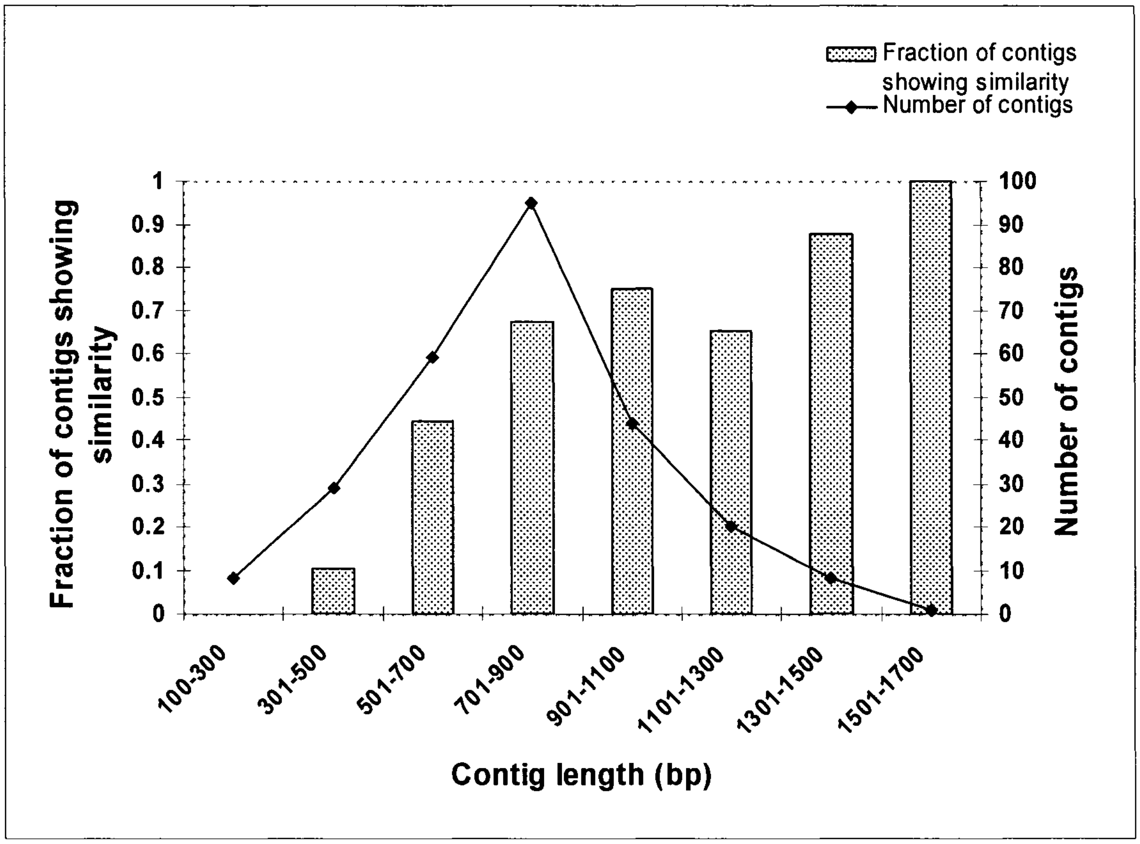
2.3.7 Functional Comparison of 5′ and 3′ End Sequences of ESTs

A total of 1059 (43.0%) out of 2465 5′ end sequences, and 397 (18.5%) out of 2143 3′ end sequences, showed similarity to the NR protein database and thus revealed that 5′ end sequences are more similar to the NR protein database than 3′ end sequences. Similarly, based on BLASTN, 5′ end ESTs (91.3%) showed higher similarity than 3′ end sequences (82%) with dbEST at NCBI. The alignment length of the sequences showing similarity with the dbEST at NCBI was also higher for 5′ end (100-943 bp) than with 3′ end sequences (100-890bp).

2.3.8 Redundancy of ESTs

The percent EST redundancy was high (70.6%), as random sequencing of cDNA clones was conducted from non-normalized cDNA library. This redundancy can also be predicted from BLASTX analysis. In BLASTX analysis, each of 367 transcripts coding for proteins with known functions were targeted by only one black spruce EST i.e. one

Figure 2.5 Fraction of contigs showing sequence similarity with non-redundant protein database.


Fraction of contigs showing similarity
Number of contigs
Fraction of contigs showing similarity
Number of contigs
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
100
90
80
70
60
50
40
30
20
10
0
100-300
301-500
501-700
701-900
901-1100
1101-1300
1301-1500
1501-1700
Contig length (bp)

EST represents one protein coding sequence of gene, whereas remaining 219 were targeted by 1089 ESTs, with the number of ESTs varying from 2 to 45.

2.3.9 Identification of ORFs

The ORFs in 4608 ESTs were identified based on BLASTX similarity with the available NR protein database at NCBI taking into consideration the initiation codon of the available gene sequence. A total of 97 ORFs were identified in ESTs from black spruce. These ORFs showed >80% similarity with the available full length sequence of genes in the NR protein database (Appendix).

2.3.10 Occurrence of Microsatellites in ESTs

The ESTs from black spruce provide an excellent source of codominant PCR-based markers for mapping and population-based analyses. Among the ESTs, 58 microsatellite repeat motifs were found. These are represented by 3 dinucleotide, 17 trinucleotide, 9 tetranucleotide, 20 pentanucleotide and 9 hexanucleotide repeats (Table 2.2). A large proportion of microsatellite loci were pentanucleotide and trinucleotide. The ESTs containing repeats have been used for the development of microsatellite DNA markers. These markers were used to construct the genetic linkage map for black spruce in a three-generation outbred pedigree as described in Chapter 4 of this thesis.

Table 2.2 Types and distribution of microsatellites in 4608 ESTs.

Type of repeat	Repeat motif(s)	Percentage of ESTs having repeat motifs
Hexanucleotide repeats	(TAAAAA)2	0.13
	(AGGGGG)2	0.2
	(GGAGAA)2or3	0.07 or 0.04
	(CGGGGG)2	0.07
	(CAAAAA)2	0.07
	(GGGAGA)2	0.07
	(TTAGGG)2	0.07
	(CCCTAA)3	0.02
Pentanucleotide repeats	(GACTG)3	0.02
	(TTTTC)2	0.76
	(GGAGA)2	0.09
	(GGGGA)2	0.11
	(TAAAA)2	1.02
	(GGGAA)2	0.04
	(CAGAG)2or3	0.17 or 0.09
	(CAGAA)2	0.24
	(GAGAA)2	2
	(CAGCG)3	0.02
	(CAGAT)2or3	0.09 or 0.02
	(TTTTA)3	0.02
	(GTCTG)2or3or4	0.07 or 0.07 or 0.02
	(TTATA)2or3	0.89 or 0.02
	(GGTCT)5	0.02
Tetranucleotide repeats	(CATG)2or3	1.3 or 0.02
	(TTTA)3	0.76
	(CCTG)4	0.04
	(GGGA)2	0.98
	(CCCA)2	0.61
	(CAGG)2	0.76
	(GAAA)2or3	3.23 or 0.09
Trinucleotide repeats	(TGG)3	0.56
	(GGA)4or5or8	0.5 or 0.02 or 0.02
	(CGA)8	0.02
	(CAG)4or5or6	0.26 or 0.11 or 0.02
	(TTA)4or5	0.02 or 0.04
	(TCG)5	0.11
	(CCA)4or5or6	0.02 or 0.04 or 0.13
	(TTC)4or5or6	0.2 or 0.07 or 0.04
Dinucleotide repeats	(TA)5or6or7	0.52 or 0.04 or 0.17

2.4 Discussion

2.4.1 Many ESTs from Black Spruce Lack Similarity with the Total dbEST Available at NCBI

No identity for 13% of ESTs from black spruce was found by BLASTN similarity searches (Figure 2.1), even though 38.95 million ESTs from 1207 different species were present in the total dbEST at NCBI at the time of analysis (October, 2006). Either these 13% black spruce transcripts are less conserved across species or a similar transcript has not been sequenced so far in other species. The lack of similarity for these ESTs from black spruce to the sequences in the dbEST may also be attributed to the source of tissues used. In the present study, the cDNA library was constructed using RNA extracted from needles whereas most of the cDNA libraries constructed in conifers are derived from RNA extracted from wood forming tissues (Allona *et al.* 1998; Kirst *et al.* 2003). The transcripts with no identity with available dbEST may represent the members of known multigene families as in the present study bidirectional sequencing has been conducted and non-coding regions are expected to be more extensive in sequences retrieved from polyA tail (3′ end) regions than the 5′ end regions of genes. These 13% transcripts, identified in black spruce represent interesting targets for detailed functional analysis.

2.4.2 Sequence Similarity Increases with Increased Sequence Length

Contigs are consensus sequences of longer length than single pass ESTs as they are derived from multiple and putatively overlapping ESTs. Contigs displayed higher sequence similarity values (Figures 2.2, 2.3, 2.4) than unassembled ESTs. This is due to increased length of contigs which leads to higher likelihood of finding the sequence similarity with different taxa and potential lack of conserved regions in shorter sequences (Kirst *et al.* 2003). With the increase in length of contigs, the percentage of sequence similarity increased with the NR protein database (Figure 2.5). A similar trend for sequence similarity for contigs has been reported for loblolly pine (Kirst *et al.* 2003). The longer sequence length of contigs aids in identification of ORFs and full length cDNAs including 5′ and 3′ untranslated regions (UTRs). These UTRs play an important role in post-transcriptional regulation and may be well conserved among different species (Jackson 1993; Duret *et al.* 1993).

2.4.3 Higher Redundancy Among ESTs from the Black Spruce cDNA Library

The redundancy of ESTs in the present study was high (71%) and is comparable to the EST redundancy obtained from six different cDNA libraries in loblolly pine (53-77%; Kirst *et al.* 2003). The high percentage of redundancy is common in non-normalized cDNA libraries (Cooke *et al.* 1996). One inevitable cause of redundancy, *i.e.* non-uniform abundance of mRNAs from different genes in cDNA libraries, is over representation of some genes, for example, housekeeping genes (Rounsley *et al.* 1996). The redundancy can also be due to the presence of paralogous genes and members of multigene families in conifers (Kinlaw and Neale 1997). This bias can be further enhanced by amplification or differences in cloning efficiency during construction of cDNA libraries. The problem of redundancy can be solved by assembling the ESTs into contigs (Rounsley *et al.* 1996; Adams *et al.* 1991) or by developing normalized or SSH cDNA libraries (Diatchenko *et al.* 1996; Bonaldo *et al.* 1996). On the other hand, the redundancy inherent in standard cDNA libraries helps in the generation of assemblies because overlapping ESTs from a single gene can be aligned and compiled to generate consensus sequence *in silico*. Redundancy of EST collections can be exploited for identification of single nucleotide polymorphisms (SNPs) (Buetow *et al.* 1999), alternative splice variants (Burke *et al.* 1998) and for digital Northerns (Ohlrogge and Benning 2000).

2.4.4 More Sequence Data is Required in Conifers for Gene Annotation

Based on BLASTX analysis, only 32% of the ESTs from black spruce have been annotated in the present study (Table 2.1), and is comparable to 27% reported for poplar (Stirling *et al.* 2003). However, the percentage of ESTs from black spruce annotated in the present study is lower than that reported in loblolly pine (Allona *et al.* 1998) and rice (Reddy *et al.* 2002), where ESTs showing similarity to sequences with known function were 59% and 65%, respectively. About 68% of ESTs from black spruce could not be annotated. Some of these ESTs in the black spruce cDNA library might represent non-conserved regions of known genes whose sequence has diverged to the extent that similar sequences are undetectable in related species. These transcripts may show similarity with the ever increasing NR protein database or may be black spruce specific transcripts. As only 32% of ESTs from black spruce have been annotated, sequence divergence among gymnosperms and angiosperms may be a limiting factor for gene annotation in conifers. Consequently, the sequence data derived from angiosperms is not sufficient for gene annotation of conifer ESTs. Future efforts should, therefore, be focused on increasing the number of sequences in the database for conifers.

2.4.5 Spruce ESTs Showed Greater Similarity with Gymnosperms than Angiosperms

As might be expected, the ESTs from black spruce exhibited greater sequence similarity to gymnosperms (conifers) than angiosperms (monocots and eudicots) (Figure 2.2, 2.3, 2.4). The genus *Picea* and *Pinus* diverged ~140 millions year ago (Florin 1963; Miller 1988), much later than the earlier split of 300 millions year ago between the angiosperm and gymnosperm lineages (Bowe *et al.* 2000). Hence, ESTs from black spruce showed higher sequence similarity to other members of Pinaceae than to monocots and eudicots. Among monocots and eudicots, similarity of ESTs from black spruce to monocots (rice and wheat) is lower than that to eudicots (poplar and *Arabidopsis*), even though rice and wheat are well represented by ESTs in dbEST at NCBI. This could be a consequence of a faster rate of evolution of gene sequences in the monocot lineage than in the eudicot lineage. A similar trend for sequence similarity analysis of ESTs from black spruce and contigs was observed using random picking and splitting of the database from the species-specific dbEST (Figures 2.3, 2. 4).

The sequence comparisons based on BLASTX suggest sharing of genes across the plant kingdom as 32% of ESTs from black spruce have been annotated. In loblolly pine, 50% of the ESTs have been annotated (Kirst *et al.* 2003). However, similarity of loblolly pine contigs (>1.2 kb) with *Arabidopsis* (eudicot) was reported to be approximately 92.0% (Kirst *et al.* 2003), which is higher than the similarity percentage of black spruce contigs (41.7%, present study) and white spruce contigs (68%) with *Arabidopsis* (Pavy *et al.* 2005b). The gymnosperm and angiosperm plants differ significantly in their genome size, ecological niche and phenotypic appearance. Nevertheless, they show substantial conservation in coding sequences of gene. This suggests that these genes have been derived from ancestral genes during evolution or that differential regulation of similar sets of genes has occurred.

Future alignment with ESTs from loblolly pine or other gymnosperm species will allow detection of sequences which are conserved between species and thus allow tags to be identified for genes encoding highly conserved proteins. These genes can be used as candidates to develop common framework markers in genome mapping projects in different species. In addition, alignment of nucleotide sequences for a given gene from different species can be used for detection of putative SNPs, which can be used as markers for genetic linkage mapping in conifers.

2.4.6 5′ end Sequences of ESTs are More Conserved in Plants than 3′ End Sequences

The sequence similarity of 5′ end sequences were found to be higher than for 3′ end sequences, implying greater conservation of 5′ end sequences of genes than their 3′end sequences in plants (Figure 2.2, 2.3, 2.4). In conifers and poplar, most studies have been conducted using 5′ ESTs (Kirst *et al.* 2003; Allona *et al.* 1998; Sterky *et al.* 1998). In only a few published studies, ESTs have been produced by sequencing both 5′ and 3′ ends of each clone; for example, barley (Michalek *et al.* 2002) and *Arabidopsis* (Cooke *et al.* 1996). The 5′ end sequences allow identification of putative protein products and 3′ end sequences help to define gene specific probes to be used in gene expression and genetic mapping studies (Cooke *et al.* 1996). As only a limited number of clones of a standard cDNA library are full length, contains portion of a transcript that usually codes for a protein and may contain part of 5′ untranslated region. These regions tend to be conserved across species and do not change much within a gene family. Therefore, contrary to the rather conserved 5′ end sequences, the ESTs generated from the 3′ end of a transcript from oligodT primed cDNA library contain 3′ UTRs, and therefore tend to exhibit less cross-species sequence conservation than do coding sequences.

CHAPTER 3

IDENTIFICATION OF DIFFERENTIALLY-EXPRESSED GENES IN NEEDLE AND CAMBIUM TISSUES

3.1 INTRODUCTION

Leaves and cambium play a vital role in growth, development and survival of plants. Leaves (needles in conifers) are the site of photosynthesis, respiration and transpiration (Hopkins and Hunter 2004). They are the principal photosynthetic machinery and provide energy to the plant by fixing atmospheric carbon-dioxide. The fixed carbon is then converted to chemical energy, which is used for growth, development and the maintenance of cellular homeostasis (Hopkins and Hunter 2004). The cambium is a lateral meristem producing xylem, phloem and bark. In perennial woody plants, such as conifer and other trees, perennial cambial activity produces xylem and phloem resulting in the production of wood (Kozlowski and Pallardy 1997). A major portion of the earth's biomass is produced by the cellular and metabolic activities in leaves (needles) and cambium. Despite leaves and cambium are sites of major functional pathways described above, very little is known about the similarities and differences in genes expressed in these tissues or organs. Transcriptome profiling of leaves versus cambium can help in identification of genes underlying growth, development and abiotic and biotic stress response in forest trees. Also, the sequencing of the transcripts from these tissues can provide an important genomic resource. Based on their cellular functions, majority of the genes are expected to be expressed differentially between leaves and cambium.

Tissue and organ-specific gene expression is a well known phenomenon in plants (Sakuta and Satoh 2000). For example, in several woody and non-woody plant species, tissue-specific genes have been identified during various morphological and physiological developments, such as embryogenesis (Nuccio and Thomas 1999), flowering (Sablowski and Meyerowitz 1998), fruiting (Boss *et al.* 2001), and various growth and reproductive stages (Opsahl-Ferstad *et al.* 1997; Woo *et al.* 1995; Walden *et al.* 1999). Most of the previous studies on organ and tissue specific gene expression are based on single genes, including those in conifers (e.g., Hutchinson *et al.* 1999). However there is no report on differential gene expression between needle and cambium tissues although ESTs have been sequenced from cDNA libraries constructed from needle and cambium tissues in white spruce (Pavy *et al.* 2005b).

Several methods have been used for the identification of differentially expressed genes. These include differential analysis of library expression (DALE) (Li *et al.* 2004), differential display (DD) (Liang and Pardee 1992), representational difference analysis (RDA) (Lisitsyn *et al.* 1993), physical removal of common sequences (PR) (Akopian and Wood 1995), serial analysis of gene expression (SAGE) (Velculescu *et al.* 1995), suppression subtractive hybridization (SSH) (Diatchenko *et al.* 1996) and microarrays (Schena *et al.* 1995). Except for the microarray and SSH, all other methods have a drawback of maintaining disproportionate concentrations of differentially expressed transcripts during subtraction. The microarray and SSH methods are powerful tools to identify differentially-expressed genes. The microarray method requires availability of microarray chips. Recently, microarray chips have been developed and used for gene expression profiling in response to wounding and insect feeding in Sitka spruce (Ralph *et* *al.* 2006), and during adventitious root development in lodgepole pine (*Pinus contorta*) (Brinker *et al.* 2004). However, these microarrays are restricted to specific laboratories and are not available for common use. In fact, no conifer microarrays are publicly available.

The SSH method generates an equalized representation of differentially-expressed genes irrespective of their relative abundance in the genome (Diatchenko *et al.* 1996). It enables the construction of subtracted cDNA libraries using hybridization and suppression PCR in a single procedure, and enriches the differentially-expressed mRNA in one transcript population (tester) as compared to the other transcript population (driver) (Diatchenko *et al.* 1996). This method has been successfully used to identify genes expressed differentially and involved in important biological and phenological processes in woody and non-woody plants, such as transcripts involved in early and late bud flushing in Norway spruce (Yakovlev *et al.* 2006), genes expressed during xylogenesis in *Eucalyptus* (Foucart *et al.* 2006), disease and stress responsive genes in rice (Xiong *et al.* 2001, Wang *et al.* 2005) and *Arabidopsis* (Mahalingam *et al.* 2003). The ESTs from different cDNA libraries provide information about the organ, or tissue specificity of expressed genes.

As in the case with other conifers, sequencing of the complete genome of black spruce is not feasible at present merely because of their large genome size (1C= 15.8 pg; Ohri and Khoshoo 1986; http://www.rbgkew.org.uk/cval/homepage.html). Transcriptomics seems to be a reasonable way to identify genes involved in growth, adaptation and disease resistance against pathogens. Identifying genes expressed differentially between needles and cambium in black spruce may help to improve understanding of the mechanisms underlying growth and adaptation in addition to providing an important genomic resource.

The objective of this study was to identify differentially-expressed genes in needle and cambium tissues in black spruce In this chapter, I have conducted transcript profiling of needle versus cambium tissues in black spruce with the snapshot of differentially-expressed transcripts, based on EST analysis and dot blot hybridization of SSH cDNA libraries constructed using RNA extracted from needle and cambium tissues of the parents of a mapping population.

3.2 MATERIAL AND METHODS

3.2.1 Plant Material and RNA Isolation

The female (P32) and male (P40) parents of a well characterized black spruce mapping population (described later in Chapter 4, see Figure 4.1), established in a genetic test at the Petawawa Research Forest located in Chalk River, Ontario (46° N, 77° 30′ W), were used as RNA source. These parents were of approximate 30 years of age. The small shoot cuttings and twigs with needles were collected from these parents during the day (June 25th, 2002), at the time of fresh growth of needle and cambium tissues. The collected material was immediately frozen in liquid nitrogen and remained frozen for two days during transportation to Dalhousie University, Halifax. Then the plant material was transferred to a -85°C freezer. Cambium was removed with forceps after peeling the bark and phloem from the frozen stem cuttings with a scalpel. Needles were harvested from frozen shoot cuttings. Total RNA was extracted from the needle and cambium tissues stored at -85°C as described by Chang *et al.* (1993). Quality and quantity of the isolated RNA were determined using a spectrophotometer (SPECTRAmax PLUS, Molecular Devices Corporation, Sunnyvale, CA, USA) and extracted RNA was of high quality (OD_{260}/OD_{280}=1.79-1.87). Quantity of the isolated RNA was higher for needles (~100 μg/gm) than for cambium (~50 μg/gm). The polyA RNA was purified using RNeasy Mini Kit (Qiagen Inc., Mississauga, ON, Canada).

3.2.2 Construction of SSH cDNA Libraries

Four SSH cDNA libraries were constructed using needle and cambium cDNA from the parents of the mapping population (Table 3.1). Two SSH cDNA libraries were constructed each from needle and cambium cDNA (one each from P32 and P40) to be considered as biological replication. First-strand of cDNA and double stranded cDNA were synthesized from 2 μg of polyA needle (tester population) and cambium (driver population) RNA following Smart PCR cDNA Synthesis Kit (Clontech Laboratories Inc., Mount View, CA, USA). The cDNA was then size-fractioned followed by digestion with restriction enzyme *Rsa*I. SSH was performed using the PCR-Select cDNA Subtraction Kit (Clontech Laboratories Inc., Mount View, CA, USA) according to the manufacturer's instructions. The final subtractive product is supposed to contain cDNA population predominantly expressed in needles. The same procedure was repeated by taking cambium cDNA as tester and needle cDNA as driver to obtain the PCR product containing cDNA population expressed in cambium (Table 3.1). The subtracted cDNA populations were cloned in T/A cloning vector pCR2.1 (Invitrogen Canada Inc., Mississauga, ON, Canada) and TOP10 chemically competent *E. coli* cells were

Table 3.1 Suppression subtractive hybridization (SSH) cDNA libraries constructed using RNA extracted from needle and cambium tissues of black spruce.

Library Name	**Tissue cDNA enriched/ Tester cDNA used**	**Driver cDNA used/ cDNA probe used in dot blot hybridization**
P32N	Needle cDNA from P32	cDNA from P32 cambium
P32C	Cambium cDNA from P32	cDNA from P32 needles
P40N	Needle cDNA from P40	cDNA from P40 cambium
P40C	Cambium cDNA from P40	cDNA from P40 needles

transformed with vector containing cDNA inserts (Invitrogen Canada Inc., Mississauga, ON, Canada) according to the manufacturer's instructions. Plasmid DNA was isolated from the transformed *E. coli* cells as described in section 2.2.2.

Transformation efficiency was confirmed by amplifying the inserts from randomly chosen 10 to 15 clones from each SSH cDNA library. Each amplification reaction mixture contained hand picked clone and 2.5 µL 10X Taq buffer, 2 µL $MgCl_2$ (25 mM), 2 µL dNTP (2.5 mM each), 1 µL of nested primer 1 and nested primer 2R (10 µM) provided in PCR-Select cDNA Subtraction Kit, 17.875 µL of autoclaved water, and 0.05 units of Taq (MBI fermentas Inc., Burlington, ON, Canada). PCR was performed according to the following parameters: 95°C for 5 min and 25 cycles of 95°C for 10 s and 68°C for 2 min, and final extension at 68°C for 5 min. All four SSH cDNA libraries displayed 100% transformation efficiency. The cDNA insert size was determined by gel-electrophoresis on 1.2% agarose (w/v) containing ethidium bromide. The inserts ranged in size from 300 bp to >1000 bp with an average of 700 bp (Figure 3.1).

3.2.3 Sequencing and Sequence Analysis

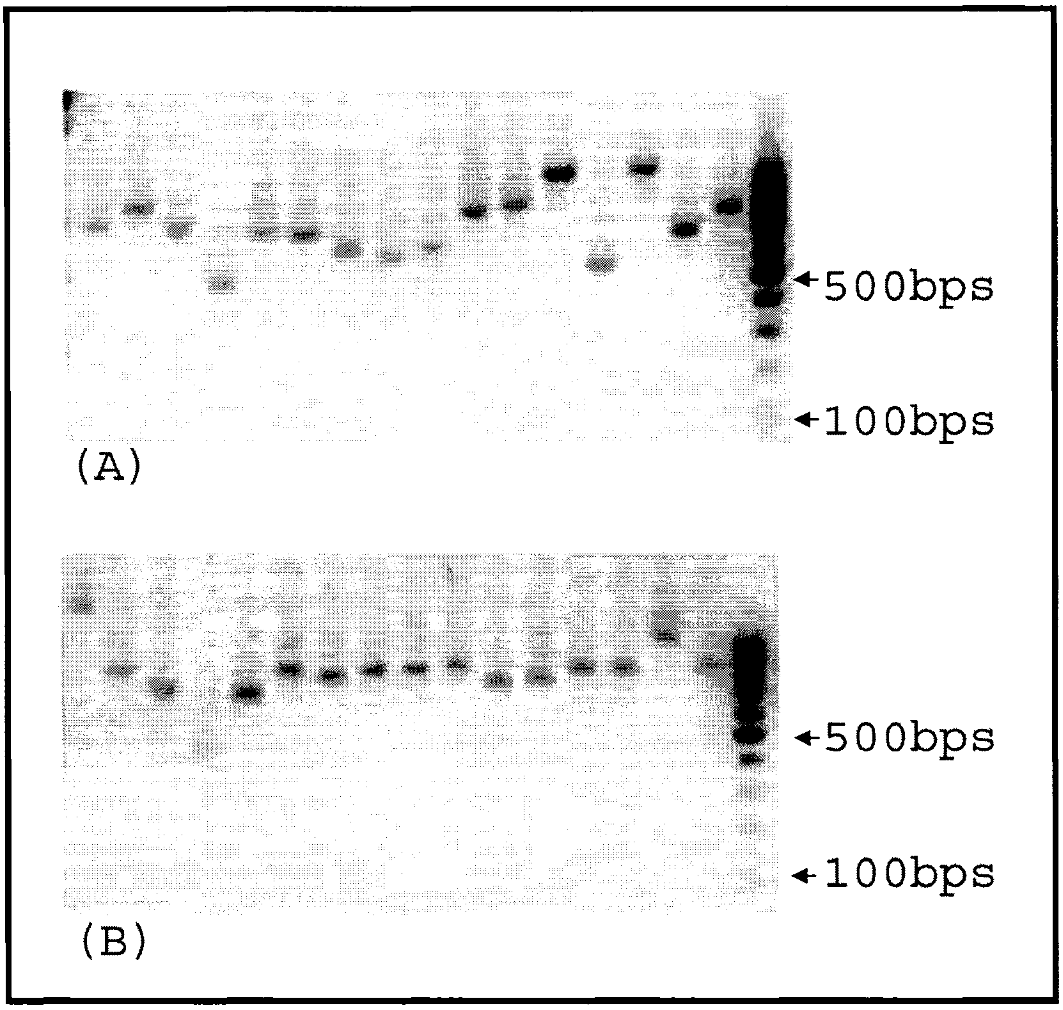
A total of 7872 clones were hand picked from four SSH cDNA libraries for sequencing (Table 3.2). This included 21 plates (96 clones per plate) for each of the P32C, P40N and P40C SSH cDNA libraries and 19 plates from P32N SSH cDNA library. The sequencing was done at the Genome Atlantic Sequencing Platform located at the Institute for Marine Biosciences, NRC, Halifax, Canada. A total of 7232 clones were successfully sequenced (Table 3.2). These cDNA sequences were analyzed for their

Table 3.2 Number of ESTs obtained from the four SSH cDNA libraries in black spruce, their similarities with total dbEST and non-redundant protein database available at NCBI, and the number of transcripts coding for proteins with known functions identified.

Library Name	Total number of sequences	Number of good quality sequences*	Average sequence length (bp)	Number of sequences showed similarity based on		No. of transcripts coding for proteins with known functions
				BLASTN No. (%)	BLASTX No. (%)	
P32N	1628	1451	529	1205 (83.05)	776 (53.48)	383
P32C	1829	1298	469	883 (68.03)	428 (32.97)	95
P40N	1926	1303	461	1231 (94.47)	510 (39.14)	126
P40C	1849	1688	505	1286 (76.18)	449 (26.90)	157
Subtotal (Needle)	3554	2754		2436 (88.76)	1286 (46.3)	509
Subtotal (Cambium)	3678	2986		2169 (71.01)	877 (29.7)	252
Total	**7232**	**5740**		**4605 (80.43)**	**2163 (38.05)**	**761**

* Good quality sequences are with phred quality >20 and sequence length of >100bp

Figure 3.1 Amplification of cDNA clones from the SSH cDNA libraries showing insert size variation. A 100bp DNA ladder (MBI Fermentas Inc., Burlington, ON, Canada) was used as the molecular weight markers in the far right lane. (A) Amplification of inserts from cDNA clones derived from the P40N SSH cDNA library, (B) amplification of inserts from cDNA clones derived from the P40C SSH cDNA library.


500bps
100bps
(A)
500bps
100bps
(B)

quality assessment using phred and for vector clipping using cross_match as described in section 2.2.4.

Only the sequences with phred quality score value of >20 and with the sequence length of greater than 100 bp were used for further analysis. The number of good quality sequences obtained after phred and cross_match ranged from 1298 to 1688 with a total of 5740 sequences for the four SSH cDNA libraries (Table 3.2).

3.2.4 Differential Gene Expression Analysis

In spruce, DNA microarray chip was not available. Therefore, EST analysis of SSH cDNA libraries and dot blot hybridization were used to examine differential gene expression between the needle and cambium tissues. The two methods are complementary as each one compensates for limitations in the other. The EST analysis method is based on similarities of ESTs between the needle and cambium SSH cDNA libraries, similarities of ESTs with the available dbEST and NR protein database, followed by annotation of transcripts based on BLASTX similarities and to determine differential level of representation of annotated transcripts in the cambium versus needle SSH cDNA libraries. These sequence similarities are based on the length of the sequences available in the database and the test SSH cDNA libraries. Because of limitation of length of sequence available in similarity analysis, the percentage of differentially represented transcripts is likely to be biased on the basis of similarity analysis. Since the individual cDNA clones and whole cDNA population from a tissue are used in dot blot hybridization, this method will be more realistic to identify differentially-expressed genes in terms of transcript abundance. However, the dot blot hybridization method is time and labor intensive, and thus limited by the number of cDNA clones that can be used. A combination of EST analysis and dot blot hybridization methods used in the present study can provide a more accurate estimate of differential gene expression.

3.2.5 Sequence Similarity Analysis

Sequence similarity analysis was performed for the ESTs from four SSH cDNA libraries to (1) examine the percentage of similar ESTs among the four SSH cDNA libraries, (2) determine the similarities of ESTs with the total dbEST and *Picea* species dbEST at NCBI, and with the ESTs from black spruce from a standard needle cDNA library (2.2.2), and (3) annotation of transcripts based on similarity of ESTs with the NR protein database.

The sequence similarity search based on BLASTN was performed using blastall client from NCBI to determine the number and percentage of similar ESTs among the four SSH cDNA libraries. A lower threshold criterion of e^{-1} was used so that the number of non-identical sequences is not overestimated. BLASTN analysis using blastall was also performed to determine similarity of ESTs from SSH cDNA libraries with the ESTs developed from a needle cDNA library (2.2.2), and dbEST available from all the *Picea* species at NCBI (249,705 ESTs, April 24, 2006), using cutoff of e^{-5}.

Nucleotide (BLASTN) and protein (BLASTX) sequence similarities of ESTs from four SSH cDNA libraries were determined with the total dbEST and NR protein database available at NCBI using custom designed Perl scripts (Altschul *et al.* 1990). The Perl scripts were designed so that the BLASTN and BLASTX results were obtained from only those sequences meeting the criteria of minimum bit score of 100 and alignment length of 100. This stringent criterion was applied because the results from the BLASTX analysis were further used for gene annotation.

Transcripts from four SSH cDNA libraries were annotated based on their BLASTX similarities as described in section 2.2.5. The enrichment of transcripts coding for proteins with known functions was calculated by comparing their number in the needle versus cambium SSH cDNA libraries. The enrichment factor is the ratio of the percentage of ESTs coding for a particular protein in the needle to the percentage of ESTs coding for a same protein in the cambium SSH cDNA libraries, where the percentage of ESTs coding for a particular protein is the number of ESTs coding for a same protein in the library to the total number of ESTs in that library or vice-versa.

The transcripts encoding for the most abundantly expressed proteins (Ribulose-1,5-biphosphate carboxylase (Rubisco) and its small subunit (RbcS) in the needle SSH libraries and metallothionein-like (MT) protein in the cambium SSH libraries) were further analyzed (1) to check whether the transcripts coding for the same protein are the sequences from the same transcript or the transcripts represent different members of a multigene family, and (2) putatively to identify the groups or classes of the multigene family to which the identified transcripts belong. The ESTs available from plant species coding for Rubisco and its small subunit and MT protein (family 15; http://www.expasy.org/cgi-bin/lists?metallo.txt) family members were downloaded from the dbEST available at NCBI. The similarities of the ESTs in the black spruce SSH cDNA libraries coding for Rubisco, RbcS and MT protein were determined with the downloaded dbEST.

Gene transcripts from different members of a multigene family present in the black spruce SSH cDNA libraries and their putative classes were identified.

3.2.6 Dot Blot Hybridization

Ninety-six transformed colonies from each of the four SSH cDNA libraries were used for dot blot hybridization experiment. The cDNA clones were selected on the basis of BLASTN and/or BLASTX similarity results for P40N library and randomly for other three libraries. The 384 clones represented 157 transcripts coding for proteins with known functions belonging to 12 functional groups including protein synthesis, energy and metabolism, disease and stress response, photosynthesis, protein destination and storage, cell structure, lipid biosynthesis and metabolism, secondary metabolism, signal transduction, transporters, growth and development, expressed and unknown protein and 227 transcripts specific to black spruce. The plasmid DNA was isolated from transformed colonies as described in section 2.2.2. The PCR amplification of inserts from the isolated plasmid DNA was carried out as follows. Each amplification reaction mixture contained 1.5 µL (~20 ng) of plasmid DNA, 2.5 µL 10X Taq buffer, 2 µL $MgCl_2$ (25 mM), 2 µL dNTP (2.5 mM each), 1 µL of nested primer 1 and nested primer 2R (10 µM), 16.375 µL of autoclaved Millique water, and 0.05 units of Taq (MBI Fermentas Inc., Burlington, ON, Canada). PCR was performed according to the following parameters: 95°C for 30 s and 25 cycles of 95°C for 10 s and 68°C for 2 min, and final extension at 68°C for 5 min. The presence of a single PCR product was confirmed by gel-electrophoresis of the amplified PCR products on 1.2% (w/v) agarose. Five µL of PCR product from each positive clone was mixed with five µL of 0.6N NaOH. Then, one µL of the mixture was applied on positively charged nylon membrane (Roche Diagnostics, Laval, Quebec, Canada) made in duplicates for each library (using one blot as control as it is hybridized with the cDNA population from same tissue). The nylon membranes were placed on Whatmann 3MM paper presoaked with 10X saline sodium citrate. These membranes were exposed to UV light for 10 min to cross-link the DNA to the membrane. Fifteen µL (20 ng) of the cDNA population was labeled with digoxigenin-deoxy-uridine tri-phosphate according to the manufacturer's instructions using the Digoxigenin High Prime DNA Labeling and Detection Starter Kit II (Roche Diagnostics, Laval, Quebec, Canada). The hybridization and detection were performed according to the manufacturer's instructions. The membranes were pre-hybridized for 30 min and then hybridized for 12 h with denatured labeled cDNA probes. The membranes were washed with 0.5X saline sodium citrate and 0.1% sodium dodecyl sulphate. After incubation in blocking solution, antibody solution, washing solution, and detection solution, the membranes were subjected to immunological detection. Detection was performed by chemiluminescence with Chemiluminescent Substrate for Alkaline Phosphatase (CSPD) as a substrate. CSPD was applied to the membranes followed by exposure to X-ray film at 20°C for 15-20 min to visualize the hybridization intensity of transcripts in needle and cambium SSH cDNA libraries. Differences in the intensity of dot blot hybridization signals of the same clone with needle and cambium cDNA population used as probe are indicative of the relative transcript abundance. The number of clones showing differential expression in P32N, P32C, P40N and P40C SSH cDNA libraries were determined.

3.3 RESULTS

3.3.1 Similarities Among ESTs from Needle and Cambium SSH cDNA Libraries

Based on the BLASTN similarity analysis among four SSH cDNA libraries, 31 to 88% of the ESTs were similar using the cutoff of $e<10^{-1}$ (Table 3.3). This high variability in the percentage of sequences showing similarity may be, among other factors, due to the differences in the cDNA subtraction efficiency. The percentage of similar ESTs was higher between either two needle or cambium SSH cDNA libraries than between the needle and cambium SSH cDNA libraries (Table 3.3).

3.3.2 Similarities of ESTs with the *Picea* Species dbEST at NCBI

Based on a standard cutoff value of e^{-5}, ESTs from the SSH cDNA libraries showed high similarity with the *Picea* dbEST available at NCBI (Table 3.4). About 0.4 to 14% ESTs, with a total of 256 transcripts from all four SSH cDNA libraries showed no identity with *Picea* species dbEST at NCBI. Sequence similarity of ESTs from the SSH cDNA libraries with those from a standard black spruce cDNA library varied from 46-90% (Table 3.4). A total of 1961 black spruce transcripts were identified in the four SSH cDNA libraries in addition to the transcripts present in the standard needle cDNA library (2.2.2).

Table 3.3 The percentage of similar ESTs among four SSH cDNA libraries of black spruce based on BLASTN.

D I	P32N	P32C	P40N	P40C
P32N				
P32C	45			
P40N	_77_*	38		
P40C	31	_88_*	33	

I = ESTs from SSH cDNA library were used as an input file for BLASTN similarity analysis

D = ESTs from SSH cDNA library were used to create database for BLASTN similarity analysis

*Values in italics represent similarities between either two cambium or two needle tissue SSH cDNA libraries

Table 3.4 Similarities of ESTs from SSH cDNA libraries with the *Picea* species dbEST available at NCBI (249,705 ESTs, April 24, 2006), and with 4608 ESTs from black spruce from a standard needle cDNA library.

		Similarity of ESTs (BLASTN, e^{-5})			
Library Name	**Number of good quality sequences**	***Picea* sp. dbEST at NCBI**		**ESTs from Black spruce standard cDNA library from needles**	
		Number of sequences	**%**	**Number of sequences**	**%**
P32N	1451	1242	85.60	664	45.76
P32C	1298	1275	98.23	850	65.49
P40N	1303	1297	99.54	1169	89.72
P40C	1688	1670	98.93	1096	64.93

3.3.3 Similarities of ESTs with the Total dbEST and NR Protein Database at NCBI

The percentage of ESTs from black spruce from four SSH cDNA libraries showing similarities with the available total dbEST at NCBI varied from 68 to 95%, (Table 3.2). The remaining 5 to 32% of the ESTs, with a total of 1135 ESTs showed no identity with the available dbEST at NCBI at a cutoff value of bit score of 100 and alignment length of 100. The percentage of ESTs showing BLASTX similarities with the NR protein database ranged from 26.6 to 53.5, and the number of transcripts coding for proteins with known functions ranged from 95 to 383 from the four SSH cDNA libraries (Tables 3.2, 3.5). The criterion used for BLASTN and BLASTX similarity analyses here was more stringent than and different from that used for the EST similarity analysis with the *Picea* species dbEST at NCBI or ESTs black spruce from standard cDNA library from needles. This was done to minimize errors in gene annotation based on the above results.

3.3.4 Gene Annotation

A total of 570 different transcripts coding for proteins with known functions were identified in all four SSH cDNA libraries (Table 3.5). According to their putative functional roles, ESTs were further characterized by sorting into 13 functional categories (Table 3.5). The five most abundant functional categories of the transcripts included energy and metabolism, protein synthesis, cell structure, photosynthesis, disease and stress response (Table 3.5). The transcripts coding for proteins involved in lignin biosynthesis (secondary metabolism) were also identified in both needle and cambium

Table 3.5 Functional classification of transcripts coding for proteins with known functions identified from individual and all needle and cambium SSH cDNA libraries, and the number of new transcripts identified in addition to those previously identified from EST sequencing of a standard needle cDNA library in black spruce (Chapter 2).

Functional categories	**Number of transcripts**					
	P32N	**P32C**	**P40N**	**P40C**	**All 4 libraries**	**New transcripts**
Protein synthesis	48	11	10	27	76	8
Energy and metabolism	97	6	24	25	108	53
Disease and stress response	12	16	7	12	35	14
Photosynthesis	42	5	27	14	39	10
Transcription & post-transcription	7	4	3	5	17	9
Protein destination and storage	17	8	9	8	33	6
Cell structure	25	11	7	15	46	17
Lipid biosynthesis and metabolism	20	5	9	11	31	10
Secondary Metabolism	15	1	5	10	24	13
Signal transduction	11	4	1	1	16	11
Transporters	16	2	6	10	22	15
Growth and development	11	5	3	5	21	13
Expressed and Unknown protein	62	17	15	14	102	21
Total	**383**	**95**	**126**	**157**	**570**	**200**

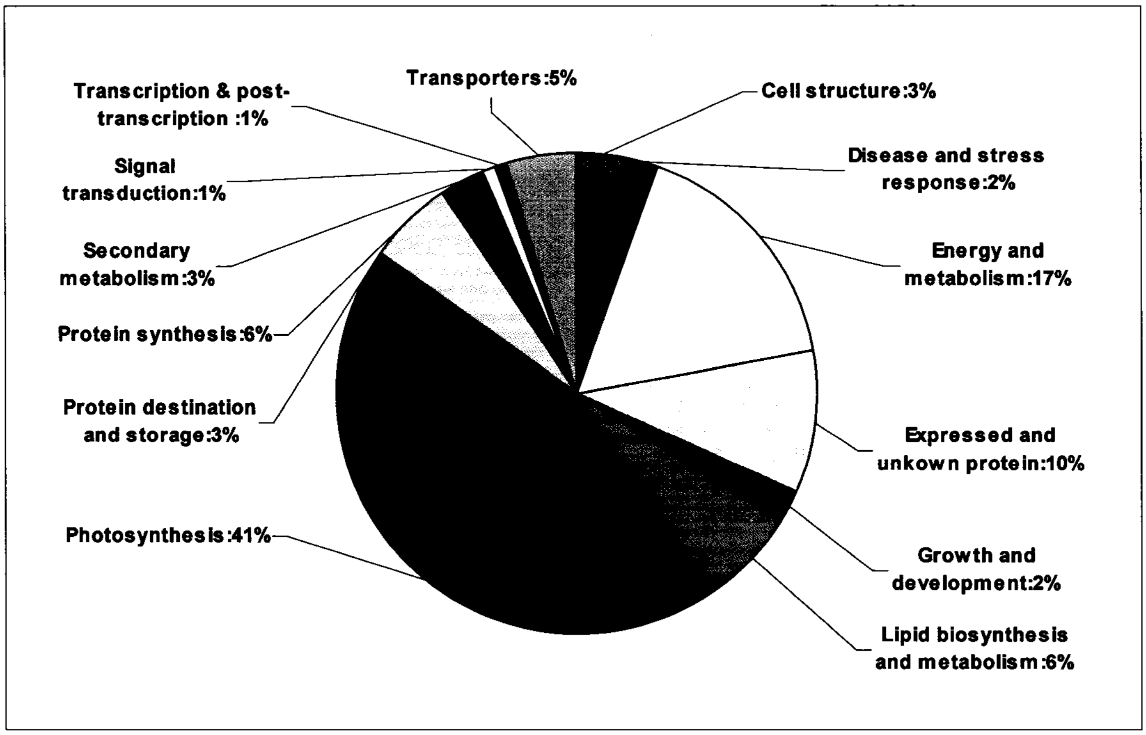
SSH cDNA libraries. These include 4-coumarate-CoA ligase family protein (4CL), cinnamoyl-CoA reductase (CCR), cinnamoyl-CoA reductase family, caffeoyl-CoA O-methyltransferase (CCoA-OMT), phenylalanine ammonia-lyase (PAL), p-coumarate 3-hydroxylase. From these SSH cDNA libraries, 200 transcripts coding for proteins with known functions were identified in black spruce in addition to those identified previously from a standard needle cDNA library thus increasing the total number of transcripts coding for proteins with known functions to 786 in black spruce (Chapter 2).

3.3.5 Distribution of Transcripts in Needle and Cambium SSH cDNA Libraries Based on their Functions

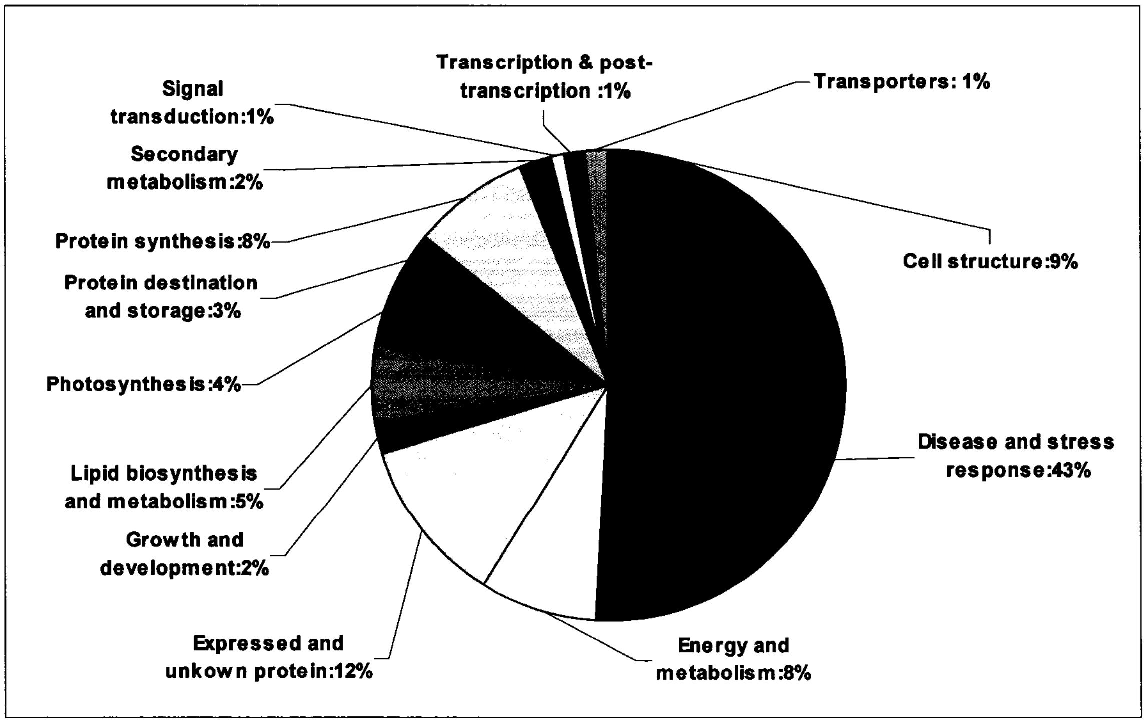
The distribution of transcripts (ESTs) based on their functional categories was substantially different between the needle and cambium tissue SSH cDNA libraries (Figure 3.2a, b). Among the ESTs obtained from the needle SSH cDNA libraries, transcripts coding for proteins involved in photosynthesis (41%) constituted the most abundant class, followed by those involved in energy and metabolism and lipid biosynthesis and metabolism (Figure 3.2a). In contrast, in the cambium SSH cDNA libraries, transcripts coding for proteins involved in the disease and stress response (43%) were the most abundant class of ESTs, followed by those involved in cell structure, energy and metabolism, and protein synthesis (Figure 3.2b). The most abundant transcripts encoding for proteins involved in the photosynthesis category was represented Rubisco and its small subunit RbcS, whereas in the cambium SSH cDNA libraries MT protein was expressed abundantly. Both Rubisco and MT protein belong to large multigene families.

Figure 3.2 Functional distribution of transcripts identified in SSH cDNA libraries according to their putative functions. (A) Distribution of 1286 transcripts identified from the ESTs sequenced from the needle SSH cDNA libraries (P32N and P40N), (B) distribution of 877 transcripts identified from the ESTs sequenced from the cambium SSH cDNA libraries (P32C and P40C).

(A) Needle SSH cDNA libraries: 1286 transcripts



(B) Cambium SSH cDNA libraries: 877 transcripts



Of the 337 ESTs encoding Rubisco and RbcS in the needle SSH cDNA libraries, 317 showed similarities with the larch (*Larix laricina*), and 20 with the Japanese black pine (*Pinus thurbergii*) Rubisco sequences. The sequence similarities ranged from 65-98% over a large sequence length (49-167 amino acids). These ESTs also showed similarities with Rubisco and RbcS sequences from other plants but over a small sequence length (10 – 24 amino acids).

Of the 315 transcripts encoding MT proteins in the cambium SSH cDNA libraries, 310 showed 73-100% similarities with the white spruce MT sequence, and five showed 82.5-97.9% similarities with the Norway spruce MT sequence, over a sequence length of 45 to 61 amino acids. The conifer MT proteins belong to type 3 of the family 15 MT multigene family (accessed on November 1st, 2006; http://www.expasy.org/cgi-bin/lists?metallo.txt). The ESTs coding for MT proteins in cambium SSH cDNA libraries showed similarities to other classes of MT proteins from plants by excluding spruce MT sequences for the similarity analysis. The black spruce MT-ESTs showed similarities with four other subdivisions of the MT multigene family: MT1 from wheat, MT2 from *Brassica napus*, *Cicer arietinum* and *Musa accuminata*, MT2A and MT2B from *A. thaliana*. However, the sequence length showing similarities was low (4 to 15 amino acids).

3.3.6 Differential Enrichment of Transcripts in Needle and Cambium SSH cDNA Libraries

Comparison of the number of transcripts coding for same protein in the needle versus cambium SSH cDNA libraries provided preliminary estimates for the enrichment of transcripts in these libraries. The enrichment values for ESTs based on 20 most abundant gene products varied from 1.04 to 24.72 for the needle (Table 3.6) and 0.06 to 23.98 for the cambium (Table 3.7) SSH cDNA libraries. Of the 20 most abundant transcripts found in the needle SSH cDNA libraries, 12 represented transcripts coding for proteins involved in photosynthesis, energy and metabolism functional categories (Table 3.6). In cambium SSH cDNA library, 13 out of 20 abundant transcripts encoded proteins underlying disease and stress, energy and metabolism, cell structure and protein synthesis functional categories (Table 3.7).

3.3.7 Dot Blot Hybridization

All 384 cDNA clones (96 from each SSH cDNA library) showed positive hybridizations with the needle and cambium cDNA populations. However, about 10% of the clones showed differences in hybridization intensities with the needle and cambium cDNA probes (Figure 3.3, Table 3.8). The differences in intensity were most likely due to the differences in the abundance of specific transcripts in needle versus cambium tissues. The majority of these differentially-expressed transcripts did not show any BLASTX similarities. The transcripts coding for proteins with known functions based on BLASTX similarities that were found differentially expressed in terms of their abundance in cambium versus needle tissues are in Table 3.8. The dot-blot hybridization results for these transcripts correspond with the abundance observed from the transcript enrichment analysis.

Figure 3.3 Dot blot hybridization results showing differential-expression of transcripts from P32N SSH cDNA library P32, (A) hybridized to needle cDNA probe, (B) hybridized to cambium cDNA probe (clones in squares show differential expression in terms of transcript abundance: c07,e02, f02, f03 clones have similarity to Ribulose-1, 5- bisphosphate carboxylase, plasma membrane intrinsic protein 2A, Lipase and Ribulose-1, 5- bisphosphate carboxylase respectively and clones f06, f07, f09, f10, g06, g08 and h08 did not show BLASTX similarity).

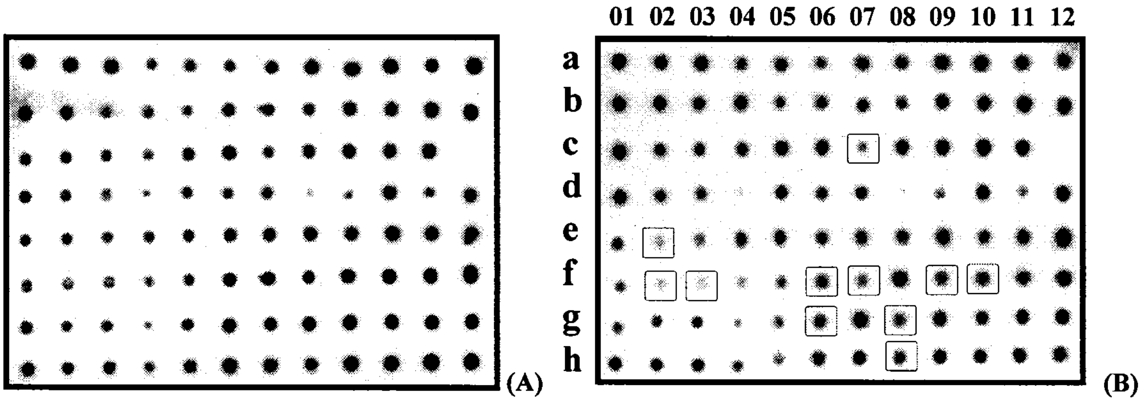

01 02 03 04 05 06 07 08 09 10 11 12
a
b
c
d
e
f
g
h
(A)
(B)

Table 3.6 The 20 most abundant transcripts found in the needle SSH cDNA libraries.

Predicted gene product	Number of transcripts coding for predicted gene product		Enrichmen t in needle SSH cDNA library*
	Needle SSH cDNA library	Cambium SSH cDNA library	
Ribulose-1,5-bisphosphate carboxylase/oxygenase	223	15	16.12
Ribulose bisphosphate carboxylase small chain	114	5	24.72
Chlorophyll a/b-binding protein Lhcb5	37	2	20.06
Photosystem I P700 apoprotein A2	34	3	12.29
Lipid transfer protein	26	27	1.04
Plasma membrane intrinsic protein 2A	16	1	17.35
Rubisco activase	12	1	13.01
Enolase (2-phosphoglycerate dehydratase)	10	0	-
Pinene synthase	9	1	9.76
Flavonol synthase	8	1	8.67
Fructose-1,6-bisphosphate aldolase	21	1	22.77
Glyceraldehyde 3-phosphate dehydrogenase (GAPDH)	16	1	17.35
Probable aquaporin	15	0	-
Glycolate oxidase	11	2	5.96
Galactosyl transferase GMA12/MNN10 family protein	8	1	8.67
Aquaporin	7	1	7.59
Actin	7	0	-
Phosphoribulokinase; ribulose-5-phosphate kinase	7	0	-
p-Coumarate 3-hydroxylase	6	0	-
Sedoheptulose-1,7-bisphosphatase precursor	4	0	-

* The enrichment factor is the ratio of the percentage of ESTs in needle SSH libraries to the percentage of ESTs in the cambium SSH libraries.

Table 3.7 The 20 most abundant transcripts found in the cambium SSH cDNA libraries

Predicted gene product	**Number of transcripts coding for predicted gene product**		**Enrichment in cambium SSH cDNA library***
	Cambium SSH cDNA library	**Needle SSH cDNA library**	
Metallothionein-like protein	315	0	-
Class IV chitinase	31	0	-
Lipid transfer protein	27	26	0.96
Thioredoxin H-type (TRX-H)	26	1	23.98
Ribulose bisphosphate carboxylase	15	223	0.06
Translationally controlled tumor protein homolog	14	0	-
Pschi4	12	0	-
R40g3 protein	12	0	-
Translation elongation factor-1 alpha	9	1	8.30
Peptidyl-prolyl cis-trans isomerase 1	7	1	6.46
SEC14 cytosolic factor family protein	6	0	-
Hydroxyproline-rich glycoprotein family protein	6	2	2.77
Pathogenesis-related protein 1	5	0	-
40S ribosomal protein S6	5	0	-
60S ribosomal protein L34 (RPL34A)	5	0	-
60S ribosomal protein L9 (RPL90B)	4	0	-
Asparaginyl-tRNA synthetase 1, cytoplasmic / asparagine-tRNA ligase 1 (SYNC1)	4	0	-
Cellulase (EC 3.2.1.4) 2 precursor	4	0	-
Embryonic abundant protein EMB32	4	0	-
Fiber protein Fb34	4	0	-

* The enrichment factor is the ratio of the percentage of ESTs in cambium SSH libraries to the percentage of ESTs in the needle SSH libraries.

Table 3.8 The number of transcripts showing differential expression in four SSH cDNA libraries based on dot blot hybridization.

Library Name	**cDNA population used as probe**	**No. of differentially expressed transcripts (no. of transcripts with BLASTX similarities)**	**Transcripts coding for proteins with known functions showing differential-expression**
P32N	cDNA from P32 cambium	11(4)	Rubisco (2); Plasma membrane intrinsic protein 2A (1); Lipase (1)
P32C	cDNA from P32 needles	9(3)	Class IV chitinase (1); Class II chitinase (2)
P40N	cDNA from P40 cambium	9(4)	Rubisco (1); Oxidoreductase (1); Lipid transfer protein (1); Unknown protein (1)
P40C	cDNA from P32 needles	10(2)	MT protein (1); Glyceraldehyde-3-phospahte dehydrogenase (1)

3.4 DISCUSSION

3.4.1 Differential Representation of Transcripts in Needle and Cambium SSH cDNA Libraries

The distribution of the ESTs based on their predicted functions showed that the transcripts coding for proteins involved in photosynthesis, energy and metabolism are abundantly expressed in the needle tissues, whereas the transcripts encoding proteins involved in disease and stress response, and cell structure are expressed more in the cambium tissues of black spruce. The results from differential representation of transcripts are consistent with the respective functional roles of needles (photosynthesis, respiration and transpiration) and cambium (growth, wood and bark formation) in plants. The dot blot hybridization results, albeit for a small number of genes (Table 3.8), are also consistent with the enrichment of differentially-expressed transcripts between the needle and cambium SSH cDNA libraries (Tables 3.6, 3.7). The transcripts coding for MT proteins and chitinases, were down-regulated in the needle tissues whereas those of Rubisco, lipase, lipid transfer protein, and plasma membrane intrinsic protein 2A were down-regulated in the cambium tissues (Table 3.8).

The transcripts encoding proteins involved in photosynthesis constituted 41% and 4% of the total ESTs, respectively, from the needle and cambium SSH cDNA libraries, compared with 13% from the standard cDNA library prepared from RNA extracted from needles of black spruce (Table 2.1). Hence, these transcripts have been enriched in needle SSH cDNA libraries by more than ten fold compared with the cambium SSH cDNA libraries and by more than three fold compared with the standard cDNA library from needles (Table2.1). Transcripts involved in energy and metabolism were also enriched in the needle cDNA libraries by more than two fold compared with the cambium SSH cDNA libraries (Figure 3.2) and the standard needle cDNA library (Table 2.1).

The higher abundance of transcripts coding for proteins involved in the photosynthesis category in black spruce needles is quite expected as needles are the primary site of photosynthesis. These results are consistent with observations for sunflower and petunia leaves (Fernández *et al.* 2003; Nagy *et al.* 1986). As might be expected, among the transcripts encoding proteins involved in photosynthesis, Rubisco was the most abundant transcript in the needle SSH cDNA libraries (26%), compared to the cambium SSH cDNA libraries (2.3%). These results are consistent with those in petunia, where transcripts encoding Rubisco were found to be most abundant in leaves (Nagy *et al.* 1986). Rubisco is the most abundant protein in the leaves of higher plants (Gray and Kekwick 1974) and is composed of eight large and eight small subunits (Spreitzer 2002). The small subunits of Rubisco are encoded by a small multigene family of nuclear genes whereas the large subunits are coded by chloroplast genes (Spreitzer 2002). The ESTs from black spruce coding for RbcS belong to three members of the multigene family compared to six members identified in petunia (Dean *et al.* 1985) and common ice plant (*Mesembryanthemum crystallinum*) (DeRocher *et al.* 1993). Since the samples for cDNA libraries were collected during the day, higher abundance of Rubisco is not surprising given that this enzyme is regulated by light (Dedonder *et al.* 1993). The other abundantly expressed trancripts in the needle SSH cDNA libraries were coding for photosystem I P700 apoprotein A2, chlorophyll a/b-binding protein Lhcb5, and Rubisco activase, which are also involved in photosynthesis (Table 3.6).

Conversely, in the cambium SSH cDNA libraries transcripts coding for proteins involved in disease and stress response and cell structure were most abundant, accounting for 52% of the total ESTs. In comparison with the needle SSH cDNA libraries, enrichment for transcripts coding for proteins involved in disease and stress response was more than 21 fold, and that of cell structure was three fold in the cambium SSH cDNA libraries (Figure 3.2a, b). The abundant expression of transcripts coding for proteins underlying cell structure in cambium (vascular and cork) is expected because cambium is a lateral meristem and produces secondary growth by cell division activities. Higher abundance of transcripts encoding disease and stress responsive proteins in cambium is a novel finding and suggests that cambium cells and tissues play a role in disease and stress response in black spruce. These findings seem to be consistent with the role of cork cambium in the formation of bark in forest trees. Bark is an effective physical barrier that protects against entry of pathogen into stem.

The transcripts coding for MT proteins were more abundantly expressed in the cambium SSH cDNA libraries than in needle SSH cDNA libraries in black spruce. This suggests that the cellular activities occurring in the cambium tissue demands higher levels of expression of MT proteins. The results are consistent with the abundance of MT-proteins reported in two other spruce species. The MT protein encoding mRNAs were abundantly expressed in the vegetative bud SSH cDNA libraries of the early- and late flushing Norway spruce families (Yakovlev et al. 2006), and during somatic embryogenesis of white spruce (Dong and Dunstan 1996). The MT proteins are known to be involved in metal detoxification in plants (Palmiter 1998), and are encoded by a multigene family (Robinson *et al.* 1993; Cobbett and Goldsbrough 2002). Transcript profiling in rice seedlings suggested that the MT proteins might contribute to plant growth in addition to metal detoxification (Maatsumura *et al.* 1999). Other predicted functions of MT proteins include cell death (Bhalerao *et al.* 2003), environmental stress (Etscheid *et al.* 1999), drought stress (Reddy *et al.* 2002), senescence (Yakovlev et al. 2006), cell wall lignification and cell elongation (Omann *et al.* 1994; Yu *et al.* 1998). High abundance of transcripts encoding MT proteins in black spruce cambium, Norway spruce vegetative buds and white spruce somatic embryogenesis suggest that MT proteins may be involved in growth and development in spruce and possibly in other conifers in addition to their known function in stress response and metal detoxification. In white spruce, the MT protein has been fully sequenced and is encoded by 60 amino acids (Dong and Dunstan 1996). All of the ESTs encoding MT proteins in black spruce showed very high (45 to 60 amino acid) similarities with the white spruce or Norway spruce MT proteins, however, similarities were much lower with MT proteins from other plants (4-15 amino acids). Therefore, the ESTs encoding MT proteins in black spruce likely belong to plant MT-Type 3 of MT multigene family.

3.4.2 Meristematic Activity in Needle and Cambium tissues

Although 39 (10%) of the 384 transcripts showed differential expression based on dot blot hybridization between the needle and cambium SSH cDNA libraries (Figure 3.3), none of the transcripts was found to be specific to either of these tissues. These results, although limited to 384 transcripts, suggest that the same or similar genes were expressed in the needle and cambium tissues of black spruce at the time of sampling of needles and cambium tissues. This is most likely to be due to meristematic activity taking place in both the needle and cambium tissues. At the Petawawa Forest Research where the sampled black spruce trees are located, the active growth period for black spruce needles (after bud burst) is considered to be from mid-May to mid-July, whereas that for cambium from mid-June to mid-September (John Major, personal communication). The stem cuttings and needles were collected in the last week of June in 2002 when active growth occurs. It is worth noting that in conifers, mature needles have been found to have a cambial zone of 2-3 cell layers wide, which seasonally produces secondary phloem in needles (Ewers and Aloni 1987). Meristematic activity has also been detected in leaf blades in *Arabidopsis* (Ha *et al.* 2003). Most likely, the presence of meristematic activity in both the tissues explains the similar expression pattern of transcripts present in needle and cambium tissues. Another possible reason for the absence of gene transcripts expressed only in needles or cambium tissue could be inefficiency of subtraction during SSH cDNA library construction. The differences in the distribution of transcripts based on predicted gene product functions (Figure 3.2), enrichment of transcripts (Tables 3.6, 3.7), and higher EST similarities between the needle and cambium SSH cDNA libraries (Table 3.3) suggest that the enrichment of transcripts has occurred in needle and cambium SSH cDNA libraries.

3.4.3 Spruce Genomic Resource

A total of 256 (4.5%) transcripts showed no identity with the spruce dbEST at NCBI, in spite of the fact that 249,705 ESTs were available from various spruce species (April 24, 2006). These 256 transcripts provide an important genomic resource in spruce. The information generated from ESTs, gene annotation, and differentially expressed transcripts coding for proteins with known functions in this study provides an important resource for various genomics studies and applications. These include gene discovery, identification of candidate genes involved in growth and development, SNP detection, candidate genes for association mapping, and microsatellite markers.

CHAPTER 4

DEVELOPMENT OF AFLP, SAMPL AND ESTP MARKERS AND CONSTRUCTION OF THE GENETIC LINKAGE MAP USING POLYMORPHIC AFLP, SAMPL, ESTP AND MICROSATELLITE MARKERS

4.1 INTRODUCTION

Genetic maps provide an extremely important genomic resource, especially for understanding genome organization and evolution, comparative genomics, mapping of genes and QTL, associating genomic segments with phenotypic traits, positional cloning of genes/genomic segments of interest, and MAS for desired traits. To dissect the genetic architecture of traits, genetic maps with high levels of genome coverage and confidence in the marker order are required. High-density genetic maps and identification of genes or genetic factors governing traits related to productivity, health, and adaptation to climatic change, and marker-assisted selection at an early stage could accelerate otherwise slowly progressing forest tree improvement programs. Conifers are economically and ecologically important, and are the dominant tree species of the boreal and temperate forests. Genetic mapping and other genomics research is challenging in conifers, mainly because of their large genome size (~25,000–30,000 mbp, Neale and Williams 1991), the long time required to reach sexual maturity, inbreeding depression, and general lack of advanced generation pedigrees.

Black spruce is an important transcontinental conifer species of the North American boreal and temperate forests (Viereck and Johnston 1990). The estimated genome size of black spruce is large (1C = 15.8 pg) and is distributed over 12 chromosomes (Ohri and Khoshoo 1986; http://www.rbgkew.org.uk/cval/homepage .html).

Although the first genetic linkage map in conifers was constructed for white spruce using mega-gametophytes from a single tree (Tulsieram *et al.* 1992), the progress in the spruce genome mapping has been rather slow, particularly compared with the genus *Pinus*. Genetic linkage maps have been constructed for Norway spruce (Binelli and Bucci 1994; Bucci *et al.* 1997; Paglia *et al.* 1998; Acheré *et al.* 2004), white spruce (Tulsieram *et al.* 1992; Gosselin *et al.* 2002), and a black x red spruce hybrid complex with an unknown proportion of the black spruce and red spruce genetic contribution to this hybrid (Pelgas *et al.* 2005). A parentage test with species-specific DNA markers revealed that the crosses used in Peglas *et al.* (2005) harbored red spruce genetic background. The markers used in the above genome mapping studies were random amplification of polymorphic DNA (RAPD) or a combination of RAPD, microsatellite, ESTP, SAMPL, and/or 5S rDNA. With the exception of the map constructed for Norway spruce from F_1 mapping population (Acheré *et al.* 2004) and the map constructed for putative black x red spruce hybrid from F_1 and BC_1 mapping populations (Pelgas *et al.* 2005), all other maps were constructed for single trees from the segregation of a small number of markers in haploid megagametophytes. Single-tree genetic maps are of limited value. Most of the genetic linkage maps have not coalesced into 12 linkage groups corresponding to the haploid chromosome number of *Picea*. The linkage groups have ranged from 12 to 29. The first single tree genetic linkage map of white spruce developed from 47 RAPD markers coalesced into 12 linkage groups. Although the consensus map of Norway spruce (Acheré *et al.* 2004) and the composite map of putative black X red spruce hybrid (Pelgas *et al.* 2005) coalesced into 12 linkage groups, the maternal and/or paternal maps in these species coalesced into 13–23 linkage groups. There is no information published on genetic linkage maps in pure black spruce. In this chapter, a high-density genetic linkage map has been developed using a three-generation outbred pedigree (TGOP) of black spruce. The maternal, paternal, and near-saturated consensus genetic linkage maps were developed in pure black spruce using AFLP, SAMPL, ESTP, and microsatellite markers.

4.2 MATERIALS AND METHODS

4.2.1 Mapping Population

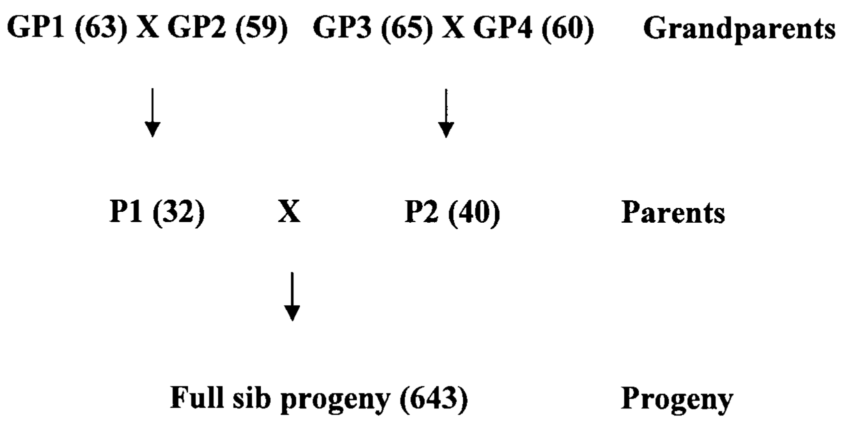
A three-generation outbred pedigree (TGOP), the grandparents, parents, and F_2 progeny (Figure 4.1), was used to construct the black spruce genetic linkage map. The grandparents of this pedigree were part of a 7 x 7 diallel F_1 controlled-cross experiment, performed by Dr. E.K. Morgenstern in the early 1970s at the Petawawa National Forestry Institute, in Chalk River, Ontario, Canada (46° N, 77° 30′ W) (Morgenstern 1974). The seven parental trees used for the diallel cross were from a plantation established at the Petawawa Research Forest, but the exact origin of the trees is unknown, other than that they were grown from seeds collected from the Lake Simcoe–Rideau region in Ontario (Morgenstern 1974). The F_1 seedlings from the full-sib families of this diallel cross were planted in genetic tests at three sites at Petawawa Research Forest in 1973 (Morgenstern 1974). The parents of the mapping pedigree were crossed in 1987 and 1988 by Dr. Tim Boyle at Petawawa National Forestry Institute to produce F_2 controlled crosses (Boyle

1987). The F_2 family 643 (P32 x P40) was selected for genetic mapping purposes based on near-top and bottom ranking of its parents for growth and ^{13}C discrimination rate and availability of sufficient number of F_2 seeds. Grandparents, parents, and 90 F_2 individuals from this family were used as the mapping population. The F_2 progeny were raised and grown at the Canadian Forest Service-Atlantic Forestry Centre, Fredericton, New Brunswick, Canada (45° 52' N, 66° 31' W).

4.2.2 DNA Extraction

Genomic DNA was extracted from the needle tissues of female and megagametophtye of male grandparents, and needle tissues from the parents and their progenies, using the Qiagen DNeasy Plant® Mini kit, following the manufacturer's protocol (Qiagen Inc., Mississauga, ON, Canada). Needle tissues from the paternal grandparent were not available as the tree was harvested from the plantation, but its open-pollinated seeds were stored at the Atlantic Forestry Centre. Therefore, to genotype this grandparent, DNA extracted from pooled megagametophyte tissues from 20 to 30 seeds was used. The quality and quantity of DNA preparations were determined by subjecting the DNA samples to gel-electrophoresis along with a standard of undigested Lambda DNA (GIBCO BRL, Burlington, ON, Canada) on 0.8% agarose followed by staining with ethidium bromide.

Figure 4.1 Three-generation outbred pedigree used for genetic linkage mapping in black spruce.



GP1 (63) X GP2 (59)
GP3 (65) X GP4 (60)
Grandparents
P1 (32)
X
P2 (40)
Parents
Full sib progeny (643)
Progeny

4.2.3 Marker Systems

Four different marker systems were used to genotype the grandparents, parents, and F_2 progeny of the TGOP: AFLPs, SAMPL, microsatellites or simple sequence repeats (SSRs), and ESTPs. Four different marker systems were used in order to achieve better genome coverage because different marker types target different regions of the genome.

4.2.3.1 AFLP Markers

As the genome size of the black spruce is extremely large, a standard AFLP protocol, based on *Eco*RI-*Mse*I digestion, *Eco*RI, and *Mse*I primer extension by one base extension in the preamplification step and three base extension to the *Eco*RI and *Mse*I primers in the selective amplification step (Vos *et al.* 1995), produced complex AFLP fragment patterns. Methods were developed for high throughput resolution of high-quality and clearly scorable AFLP markers for black spruce, using LI-COR 4200L® (LI-COR Biosciences, Lincoln, NE, USA) or Beckmann Coulter CEQ 8000 Genetic Analysis System® (Beckmann Coulter, Fullerton, CA, USA), by evaluating a variety of conditions, including *Eco*RI and *Mse*I restriction digestion time of the template DNA, and the number of selective nucleotides used in the preamplification and selective amplification steps. The primer combinations producing consistent, clear, and easily scorable polymorphic AFLP markers were identified and used for genotyping the progeny.

The AFLP method followed was that described by Vos *et al.* (1995), with some modifications. Black spruce genomic DNA (500 ng) was digested with 2U each of *Eco*RI and *Mse*I (New England Biolab Inc., Ipswich, MA) for 3 h at 37°C, followed by incubation at 70°C for 20 min. The digested DNA was ligated overnight with the *Eco*RI and *Mse*I adapters in a total volume of 20 µl at 25°C, followed by incubation at 70°C for 20 min. This restriction-ligation mixture was diluted 1:5 with autoclaved Millique water before using it in the preamplification step.

A 3 µl aliquot of the restriction-ligation mixture was preamplified using *Eco*RI (E) and *Mse*I (M) preamplification primers with an extension of one or two selective nucleotides at the 3′ end. *Eco*RI preamp primers (+1/+2):

(+1) 5′- GAC TGC GTA CCA ATT C**A** – 3′

(+2) 5′- GAC TGC GTA CCA ATT C**AC** -3′

*Mse*I preamp primers (+1/+2):

(+1) 5′- GAT GAG TCC TGA GTA **AC** – 3′

(+2) 5′- GAT GAG TCC TGA GTA A**CC** – 3′

The PCR profile for the preamplification step consisted of 20 cycles each of denaturation at 94°C for 30 s, annealing at 56°C for 1 min, and extension at 72°C for 1 min, followed by a final soak at 10°C using a PTC-200 thermal cycler (MJ Research, Reno, NV, USA). After the preamplification step, the reaction mixture was diluted 1:50 with autoclaved Millique water. A total of 54 different *Eco*RI and *Mse*I primer pairs were tested with one or two selective nucleotides at the preamplification step and three to five selective nucleotides at the selective amplification step. From these, 40 AFLP primer combinations were selected for further use in mapping. Selective amplifications were performed using these primer combinations with various selective nucleotide extensions (E+3/M+3, E+3/M+4, E+3/M+5) (Table 4.1). The reaction mixture for the selective amplification consisted of 2 µl of diluted preamplified template DNA, 1 U *Taq* polymerase, 2.5 ng of *Eco*RI labeled primer (IRD 700 label for LI-COR and D2 or D3 label for Beckmann Coulter CEQ 8000 Genetic Analysis System), 12.5 ng *Mse*I primer, 10X PCR buffer (MBI Fermentas Inc., Burlington, ON, Canada), 1.5 mM $MgCl_2$ (MBI Fermentas Inc., Burlington, ON, Canada), 0.2 mM each of all four dNTPs (MBI Fermentas Inc., Burlington, ON, Canada), and BSA (1 μg/μl) (Sigma-Aldrich, Oakville, ON, Canada). The PCR amplification profile consisted of 12 cycles each of denaturation at 94°C for 30 s, annealing at 65°C for 30 s (with lowering of 0.7°C per cycle) and extension at 72°C for 1 min, followed by 23 cycles each of denaturation at 94°C for 30 s, annealing at 56°C for 60 s and extension at 72°C for 1 min, followed by a final soak at 10°C.

Reaction products following selective amplification were resolved either on a LI-COR 4200L or a Beckmann Coulter CEQ 8000 Genetic Analysis System. For LI-COR, selective amplification products were resolved on 6.5% denaturing Long Ranger polyacrylamide gels (LI-COR Biosciences, Lincoln, NE, USA). Approximately 0.5 μl of each sample (10 μl of PCR product and 15 μl of loading dye) was loaded on the gel. IRD-labelled molecular-weight markers were loaded in three lanes as a size-standard. Electrophoresis was carried out using 1X TBE running buffer, with run parameters of 1500 V, 35 mA, 70 W, signal channel 3, motor speed 3, 50°C plate temperature and 16-bit pixel depth for collection of TIFF image files. Polymorphic fragments were visually scored in the TIFF image files. Only those markers that were segregating in a Mendelian ratio (χ^2- test, $P < 0.05$) were scored. For Beckmann Coulter CEQ 8000 Genetic Analysis System, 2 μl of the selective amplification product was added to 27.5 μl of sample loading solution and 0.5 μl of CEQ DNA size standard-600 (Beckmann Coulter,

Table 4.1 AFLP primer combinations used, and the number and size of polymorphic fragments, and their segregation ratios.

Marker	Primer combinations	Size of fragments (bp)	Total no. of polymorphic markers	Markers segregating 1:1	Markers segregating 3:1
A01	E-AAC/M-CCAC	54-340	53	30	23
A02	E-AAC/M-CCACC	54-134	2	2	0
A03	E-AAC/M-CCAG	65-502	17	17	0
A04	E-AAC/M-CCATC	95-604	37	26	11
A07	E-AAG/M-CCAG	80-236	24	13	11
A08	E-AAG/M-CCATC	66-324	26	21	5
A10	E-ACA/M-CCAT	56-264	8	7	1
A16	E-ACG/M-CAT	52-198	26	18	8
A18	E-ACG/M-CCAG	30-245	13	7	6
A20	E-ACG/M-CTG	73-275	14	3	11
A25	E-ACT/M-CCTA	54-321	31	22	9
A23	E-ACT/M-CCAA	54-475	36	27	9
A27	E-AGC/M-CTC	60-547	35	9	26
A28	E-AGC/M-CTG	45-660	49	26	23
A29	E-ACG/M-CCAA	126-275	14	7	7
A30	E-ACG/M-CCTA	60-430	30	12	18
A31	E-ACG/M-CCGC	56-622	10	8	2
A32	E-ACT/M-CCAC	59-100	5	1	4
A33	E-ACT/M-CCAG	56-70	7	3	4
A34	E-ACT/M-CTA	74-327	33	19	14
A35	E-ACT/M-CAC	61-397	33	22	11
A36	E-ACT/M-CAT	80-262	28	17	11
A38	E-ACG/M-CCAT	66-148	8	4	4
A39	E-ACG/M-CCAC	67-327	11	4	7
A40	E-ACG/M-CTC	90-280	9	5	4
A41	E-ACG/M-CCAGC	78-210	13	12	1
A44	E-AAC/M-CCTA	62-240	27	13	14
A45	E-AAC/M-CCAT	69-414	23	13	10
A46	E-AAC/M-CCAA	67-271	25	14	11
A47	E-AAC/M-CCGA	66-194	11	6	5
A49	E-ACA/M-CAC	67-292	25	15	10
A50	E-ACA/M-CCACC	69-391	22	16	6
A51	E-ACA/M-CCGA	67-262	20	11	9
A52	E-ACA/M-CTA	60-261	34	27	7
A54	E-ACA/M-CCAC	67-178	9	7	2
A57	E-AAG/M-CCAT	66-276	18	12	6
A58	E-AAG/M-CCACC	65-121	9	5	4
A59	E-AAG/M-CCAC	64-311	18	9	9
A60	E-AAG/M-CCAA	69-249	11	7	4
A61	E-AAG/M-CCTA	66-153	17	7	10
Total			**841**	**504**	**337**

Fullerton, CA, USA), followed by overlaying a drop of mineral oil. Samples were injected into a 33 cm capillary at 2.0 KV for 90 s and subjected to electrophoresis at 7.5 KV for 70 min at 35°C. The AFLP fragments were exported to an Excel® file using fragment analysis software for further analysis of genetic linkage parameters.

4.2.3.2 SAMPL Markers

Selectively amplified microsatellite polymorphic loci, based on a combination of AFLP and microsatellite technology, can combine features of both AFLP and microsatellite markers, and can reduce the marker complexity of AFLPs in spruce. The SAMPL technology is a modified AFLP technique, in which a compound microsatellite sequence is used as one of the two AFLP primers in selective amplification, generally in place of *Eco*RI primers (Gupta *et al.* 2005). SAMPL markers were developed using the compound microsatellite repeats from *Lactuca* species (Witsenboer *et al.* 1997) and SAMPL primer (Table 4.2) was used in place of the *Eco*RI primer in the selective amplification step. The SAMPL markers were analyzed on the LI-COR and Beckman CEQ 8000 systems, using the protocol described above for AFLP analysis as well as in Gupta *et al.* (2005).

Sixteen combinations of four SAMPL and four *Mse*I primers (with an extension with three selective nucleotides) were tested to screen SAMPL marker polymorphisms between the parents of the mapping population. Of these, 12 primer combinations were selected for genotyping of the mapping population based on the quality and polymorphism of the markers resolved (Table 4.3). The SAMPL marker data were scored as described above for AFLP markers.

Table 4.2 SAMPL primers developed from *Lactuca* species[1] compound microsatellite repeats and used for SAMPL marker mapping in the black spruce mapping population.

Primer name	Primer sequence (5′→3′)	Compound repeats
SL3	ACA CAC ACA CAC ACA TAT AA	$A(CA)_7\,(TA)_2A$
SL4	TGT GTG TGT GTG TGT ATA	$T\,(GT)_7\,(AT)_2$
SL5	CTC TCT CTC ACA CAC ACA CA	$C(TC)_4\,(AC)_4A$
SL6	CTC TCT CTC GTG TGT GTG	$C(TC)_4\,(GT)_4G$

[1] From Witsenboer *et al.* (1997)

Table 4.3 SAMPL primer and *Mse*I primer extension combinations used, and the number and size of polymorphic fragments, and their segregation ratios.

SAMPL marker	**Primer combinations SAMPL/MseI**	**Size of fragments (bp)**	**Total number of polymorphic markers**	**Markers segregating 1:1**	**Markers segregating 3:1**
S31	SL3/M-CTT	91-315	21	18	3
S32	SL3/M-CAC	32-661	40	18	22
S33	SL3/M-CCG	40-416	8	8	0
S41	SL4/M-CTT	40-612	44	23	21
S42	SL4/M-CAC	57-440	23	10	13
S43	SL4/M-CCG	60-640	23	4	19
S51	SL5/M-CTT	88-223	8	8	0
S52	SL5/M-CAC	48-239	3	3	0
S53	SL5/M-CCG	50-437	14	13	1
S61	SL6/M-CTT	66-277	40	35	5
S62	SL6/M-CAC	66-272	18	4	14
S63	SL6/M-CCG	66-248	20	10	10
Total			**262**	**154**	**108**

4.2.3.3 Microsatellite/SSR Markers

Seventy-eight microsatellites developed from black spruce cDNA (EST) or genomic DNA sequences and white spruce ESTs in our laboratory (to be published elsewhere) were used to screen for polymorphisms between the parents of the mapping population (Table 4.4). Forty-five of these microsatellite loci showed inter-parental polymorphisms, and were used to genotype the mapping population. Out of these, 21 were from the white spruce ESTs, 6 from ESTs from black spruce, and 18 from black spruce genomic sequences (SSR-enriched and AFLP-SSR libraries). The microsatellite markers were resolved on the LI-COR system and data were scored as described in Rajora *et al.* (2001, 2005).

4.2.3.4 ESTP Markers

Primer pairs for 198 ESTs, obtained from sequencing of a black spruce cDNA library prepared from needle tissue (Ishminder Mann, unpublished data), were designed using Primer 3.0 software (Rozen and Skaletsky 2000). The primer testing was conducted at three different annealing temperatures (50°C, 55°C, 60°C). The parents of the mapping population were screened for ESTPs (length polymorphism). The PCR amplification profile consisted of initial denaturation at 94°C for 5 min, 40 cycles each of denaturation at 94°C for 1 min, annealing temperature (see Table 4.5) for 1 min, and extension at 72°C for 1.3 min, followed by final extension at 72°C for 10 min. The ESTP markers were resolved by electrophoresis on either 2% agarose or 6% polyacrylamide gels as described in Rajora *et al.* (2001).

Table 4.4 Microsatellite DNA markers used for genetic linkage mapping in black spruce.

Source	No. markers screened	No. markers polymorphic between the parents			No. of markers mapped		
		Heterozygous in the maternal parent	**Heterozygous in the paternal parent**	**Heterozygous in both parents**	**Consensus map**	**Maternal map**	**Paternal map**
Picea glauca EST database	44	7	4	10	13	10	12
Picea mariana ESTs	10	4	2		4	3	1
Picea mariana genomic sequences	24	8	5	5	13	7	7
Total	78	19	11	15	30	20	20

Table 4.5 Expressed sequence tag polymorphic loci, number of alleles, primer sequences, and optimum annealing temperatures

Name of the ESTP locus	Annealing temperature (°C)	No. of alleles	Forward and reverse primer sequences (5' - 3')
RPMEP 622	60	2	F-CACGGACGATTCCACTGTC R-CGGCATCAGCATTAGCCCGT
RPMEP682 A **RPMEP682 B**	55 55	2 2	F-CGGTCTCTCCTTCGACTCAC R-CAGAAAAGATCTTCAGCCCC
RPMEP 638	55	2	F-AGATCTCAGAGTCTGTGCTTTGC R-ACAATCCTGCCAAGTCCCC
RPMEP 687	60	2	F-CAGAAATGGCAAGAAAGGGA R-CTATATCACAAAGAAAAATCTAGC

4.2.4 Nomenclature of Markers on the Genetic Linkage Map

The AFLP and SAMPL markers were named, starting with letters A, and S, respectively, followed by the primer number, and then the size of the fragment. The SSR markers were named with a prefix of five letters. The first letter represents the laboratory (R = Rajora laboratory), the next two letters the species name (PG = *Picea glauca,* PM = *Picea mariana*) from which the markers were developed, the next letter S representing SSR, and the last letter denoting the source of sequences or library type (E = EST; G = Genomic; A = AFLP-SSR genomic). These prefix letters were followed by the marker number. Thus, SSR markers developed from the white spruce EST database have a prefix of RPGSE, SSR markers developed from ESTs from black spruce have a prefix of RPMSE, SSR markers developed from the genomic library have a prefix of RPMSG, and SSR markers developed from the black spruce SSR-enriched AFLP sequences have a prefix of RPMSA. The ESTP markers developed from the ESTs from cDNA library were named starting with RPMEP, followed by the marker number.

4.2.5 Statistical Analysis

4.2.5.1 Segregation Analysis and Map Construction

Individual paternal and maternal maps were constructed according to two-way pseudo-testcross mapping strategy (Grattapaglia and Sederoff 1994). All linkage analysis and genetic map construction, including marker order and map length estimations, were performed using JOINMAP® 3.0 software (Van Ooijen and Voorrips 2001). The Kosambi (1944) mapping function was used for map length estimations. The marker data set for genetic linkage mapping included three different segregation patterns: 1:1 for markers heterozygous in one parent and homozygous or null in the other, 3:1 for dominant markers heterozygous in both parents, and 1:2:1 or 1:1:1:1 for co-dominant markers heterozygous in both parents. The JOINMAP command “similarity of loci” was used to identify the similar loci. Only one of the markers was kept from the similar loci for linkage mapping analysis. The two parental maps based on segregating markers were grouped and ordered using a minimum LOD (log of odds) score of 3.0 and recombination fraction of 0.4 as the grouping criterion. The marker order obtained from the third round of analysis was retained with the JOINMAP command “calculate map”. This order was fixed to allow positioning of additional markers. The parental maps were aligned based on the intercross (3:1 segregation) markers, with at least three such markers per linkage group. The two data sets from the parental linkage maps were merged. Finally, a consensus map was constructed using all 1:1 and selected intercross markers using the JOINMAP function “combine groups for map integration”. The individual linkage groups were drawn using Mapchart® version 2.0 software (Voorrips 2002).

4.2.5.2 Estimation of Genome Length and Map Coverage

The length of the black spruce genome was estimated using the Method 4 of Chakravarti *et al.* (1991), as well as the method described by Fishmann *et al.* (2001). The observed genome length was calculated by summing up the map lengths of the twelve individual linkage groups. The map coverage was calculated as the ratio of the observed to the estimated genome length. The number of markers required to cover the whole genome and to saturate the genetic map of black spruce was calculated according to Lange and Boehnke (1982).

4.2.5.3 Marker Distribution Analysis

To evaluate whether the mapped markers were randomly distributed on the linkage map, the linkage groups were divided into 2.5, 5, 10, 20, and 40 centimorgan (cM) blocks, and the number of markers per block were counted. Observed frequencies of the number of markers per block were compared with the expected ones by performing a Chi-square test (Remington *et al.* 1999; Cervera *et al.* 2001; Yin *et al.* 2003), using a Poisson distribution function, $P(x) = e^{-\mu}\mu^{x}/x!$, where x is the number of markers per block and μ is the average marker density in the consensus map. The average marker density was used to calculate the expected binomial frequencies for each marker class per block interval for all the linkage groups. The distribution of markers on the linkage groups was also evaluated separately for the AFLP and SAMPL markers. The microsatellite and ESTP markers could not be considered independently for this analysis because of their small numbers or low frequencies on each linkage group.

4.3 RESULTS

4.3.1 AFLP Markers

Forty AFLP primer combinations generated 841 markers segregating according to expected Mendelian ratios. The number of polymorphic fragments ranged from 2-53, with an average of 21 polymorphic fragments per primer combination (Table 4.1). The average number of polymorphic fragments obtained per primer combination was 28, 19, and 18 with the use of three, four, and five selective nucleotides (at the selective amplification step), respectively. The size of the segregating polymorphic fragments ranged from 30 to 660 bps. Of the 841 markers, 504 segregated in a ratio of 1:1 and 337 in a ratio of 3:1. The number of markers segregating in the 1:1 ratio was 258 in the maternal parent and 246 in the paternal parent.

4.3.2 SAMPL Markers

A total of 262 SAMPL markers, segregating according to Mendelian ratios, were obtained from the 12 SAMPL-*Mse*I primer combinations (Table 4.3). The fragment size ranged from 32 to 661 bp (Table 4.3). The number of polymorphic fragments ranged from 3 to 44, with an average of 22 polymorphic fragments per SAMPL primer combination. Of the 262 SAMPL markers, 154 segregated in ratio of 1:1, whereas 108 SAMPL markers segregated in ratio of 3:1. The number of SAMPL markers segregating in the 1:1 ratio was 96 in the maternal parent and 58 in the paternal parent.

4.3.3 Microsatellite Markers

Twenty-one of the 45 white spruce EST-based SSR loci were polymorphic between the parents of the mapping population (Table 4.4). The primer pairs for RPGSE45 resolved two loci and both were polymorphic between the parents. Seven of the 45 SSR loci screened were monomorphic and three showed distorted segregations. The remaining 15 SSR primer pairs did not amplify any DNA fragments. Seven of the 21 polymorphic loci were heterozygous in the female parent, four in the male parent and ten in both parents.

Two alleles at a SSR locus heterozygous in the male or female parent, segregated in a 1:1 ratio in the progeny. Where both the parents were heterozygous, the progeny segregated either in a 1:2:1 or a 1:1:1:1 ratio for their parental alleles (2-4). Only 13 of the 21 polymorphic SSR loci could be mapped on the consensus map, the remaining eight did not group with any of the 12 linkage groups.

Six of the ten black spruce EST-based SSR loci were polymorphic between the parents; four were heterozygous in the female and two in the male parent (Table 4.4). These markers segregated in a 1:1 ratio in the progeny. However, only four of these SSR loci could be mapped on the consensus map. Eighteen black spruce genomic sequence-based SSR loci were polymorphic between the parents (Table 4.4). Markers for the eight loci heterozygous in the female parent and five loci heterozygous in the male parent segregated in a 1:1 ratio. Markers for the five loci heterozygous in both the parents segregated in a 1:2:1 or a 1:1:1:1 ratio for the co-dominant and in a 3:1 ratio for the dominant SSR loci. Thirteen of the 18 genomic sequence-based SSRs could only be mapped on the consensus map, five did not group with any of the 12 linkage groups.

4.3.4 ESTP Markers

Only 12 of the 198 EST primer pairs screened revealed polymorphisms between the parents. However, only five of these ESTP markers showed Mendelian segregation in the progeny, and could be used for mapping. The remaining seven ESTP markers showed distorted segregation patterns. Of the five ESTP loci, three were heterozygous in the female parent, and two were heterozygous in the male parent (Figure 4.2). These markers segregated in a 1:1 ratio. However, only three ESTP markers could be mapped, one each on three linkage groups 1, 5, and 10 (Figure 4.3). Two ESTP markers remained ungrouped.

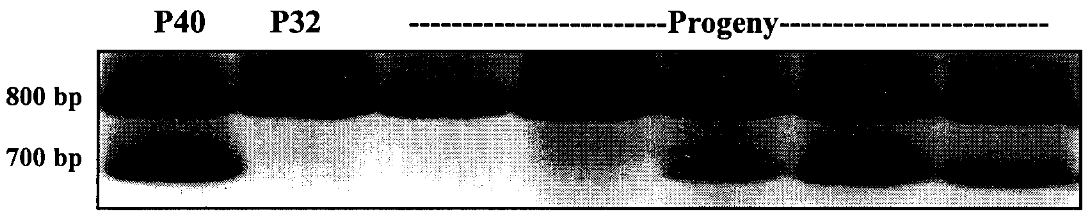
4.3.4 Genetic Linkage Maps

The maternal map consisted of 626 markers distributed on 12 linkage groups covering 1530 cM (Tables 4.6, 4.7). The number of mapped markers ranged from 28 to 64, with an average of 52 markers per linkage group. The length of the linkage groups ranged from 107 to 154 cM, with an average of 127 cM per linkage group (Table 4.7). The paternal map consisted of 634 markers assigned on 12 linkage groups, which covered 1641 cM (Tables 4.6, 4.7). The number of mapped markers ranged from 45 to 74, with an average of 53 markers per linkage group. The length of the linkage groups ranged from 103 to171 cM with an average of 136 cM per linkage group (Table 4.7).

The homologous linkage groups between the parents were identified on the basis of segregating intercross AFLP, SAMPL, and SSR markers, in the maternal and paternal maps. At least three intercross markers per linkage group were used. The integrated data set from the maternal and paternal maps allowed construction of a consensus linkage map. The consensus linkage map composed of 941 markers (Tables 4.6, 4.7) mapped to 12 linkage groups (Figure 4.3). The linkage groups correspond to the haploid chromosome number (n = 12) of black spruce. It is worth noting that I have consistently obtained 12 linkage groups for the maternal, paternal, and consensus linkage maps, unlike the case in other mapping studies where parental and/or consensus maps did not coalesce into 12 linkage groups. The consensus map covered 1898 cM, with an average of 78 markers per linkage group and an

Figure 4.2 Segregation of expressed sequence tag polymorphic (ESTP) markers within a black spruce mapping population. Sizes of the segregating bands (in bp) are indicated on left side. (A) Polyacrylamide-gel showing segregation of ESTP marker fragments between parents (P32 and P40) and selected individuals from the mapping population amplified using primer pair RPMEP622 (B) Showing segregation of ESTP marker fragments between parents (P32 and P40) and selected individuals from the mapping population amplified using primer pair RPMEP 687 resolved on 2% (w/v) agarose gel. A 100bp DNA ladder Plus (MBI Fermentas Inc., Burlington, ON, Canada) was used as the molecular weight markers in the far left lane.

(A) RPMEP622



(B) RPMEP687

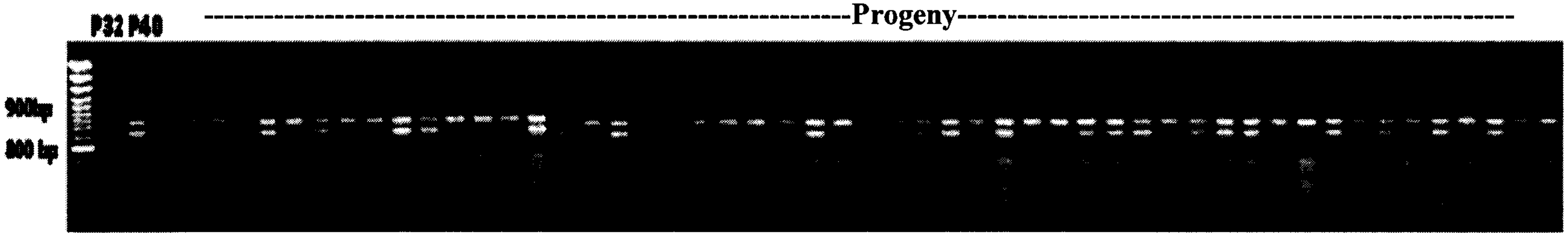


Figure 4.3 Consensus genetic linkage map of black spruce (*Picea mariana*) constructed using 695 AFLPs (*A), 213 SAMPL (*S), 3 ESTPs (bold and underlined) and 30 SSR (bold and italicized) markers. Names of the markers are indicated to the right of the linkage groups (LGs), with the fragment size indicated in bp. Genetic distances, in cM, are indicated on the left.

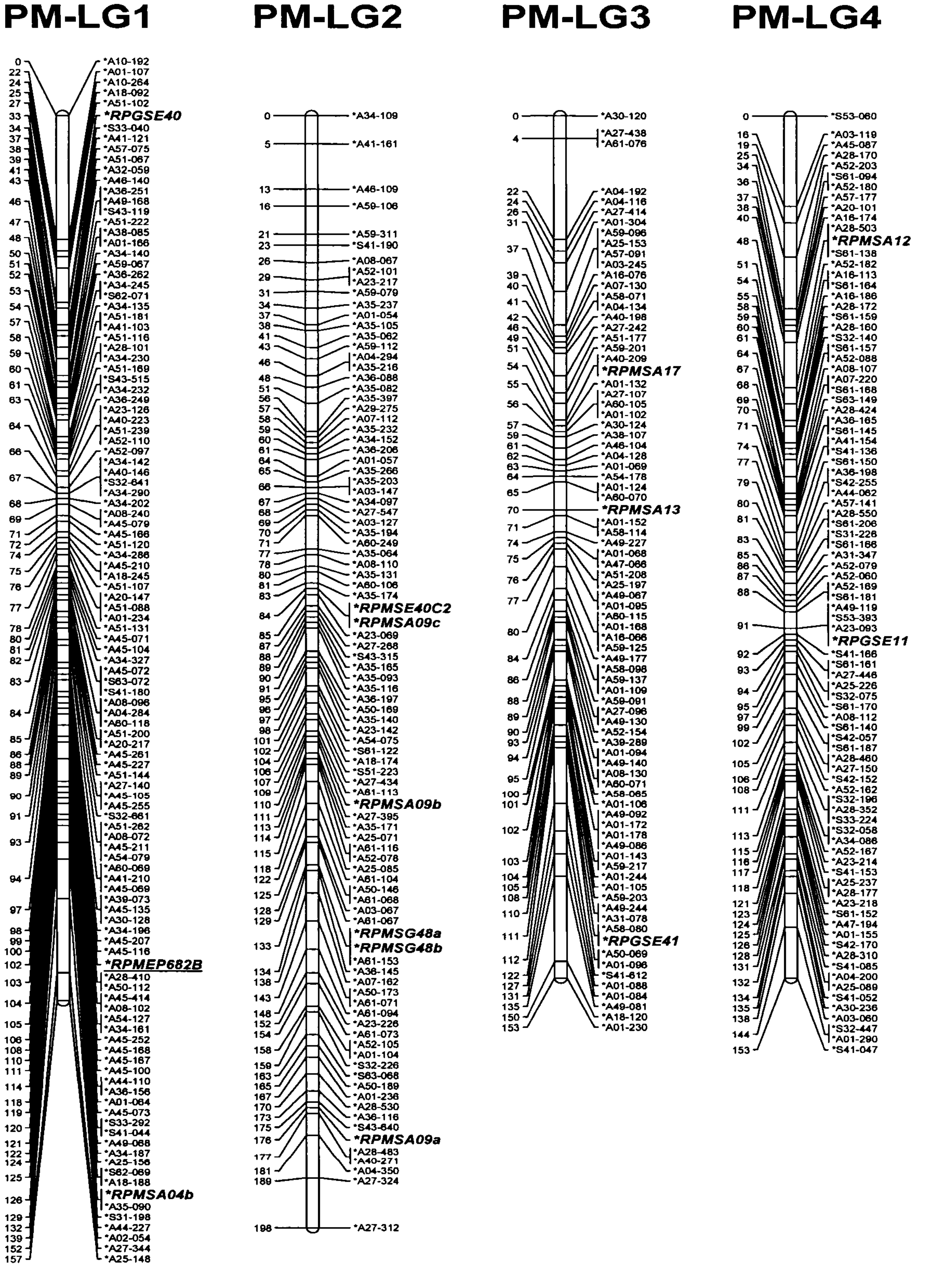
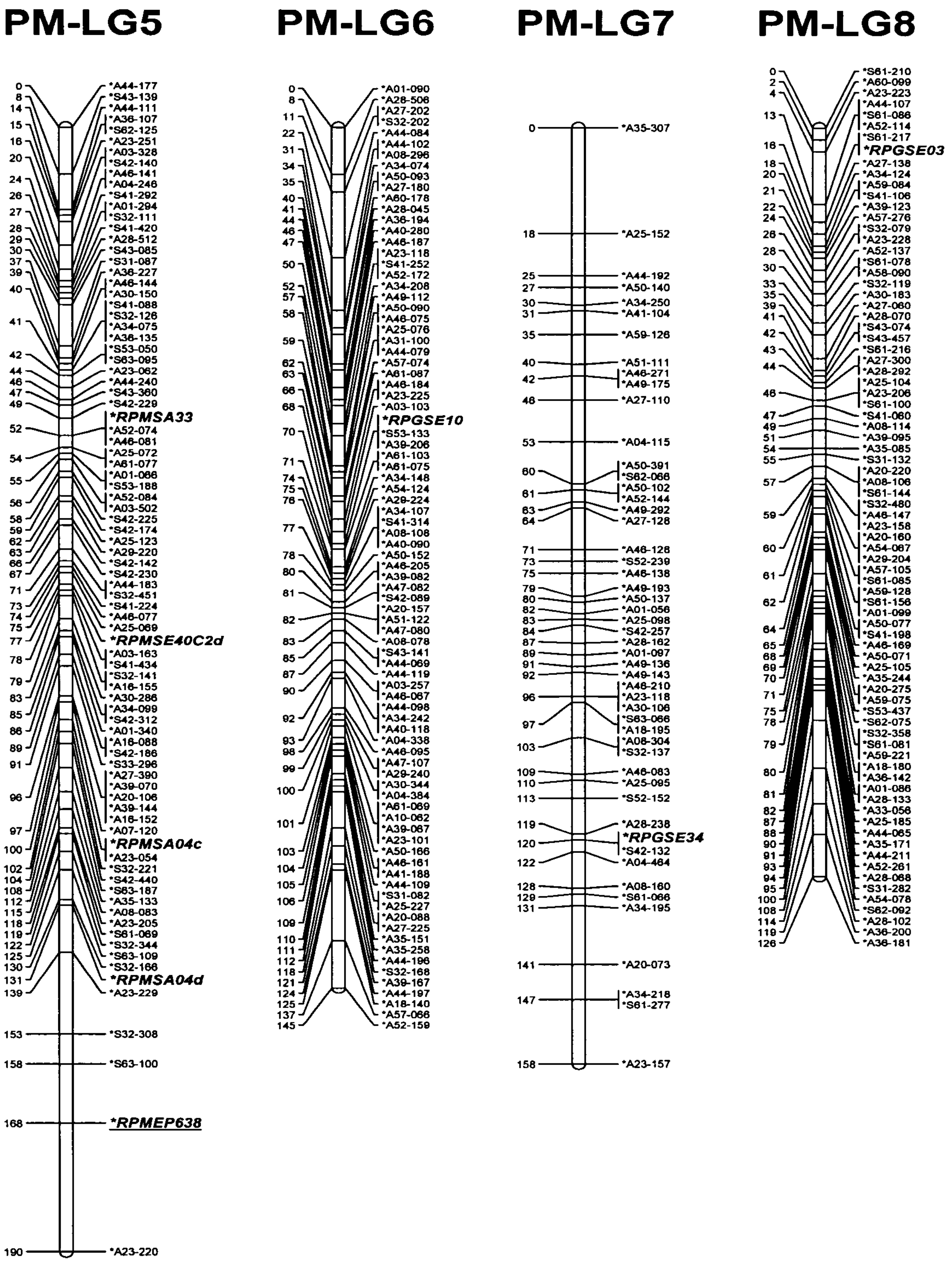


Figure 4.3 continued



PM-LG5

PM-LG6

PM-LG7

PM-LG8

Figure 4.3 continued

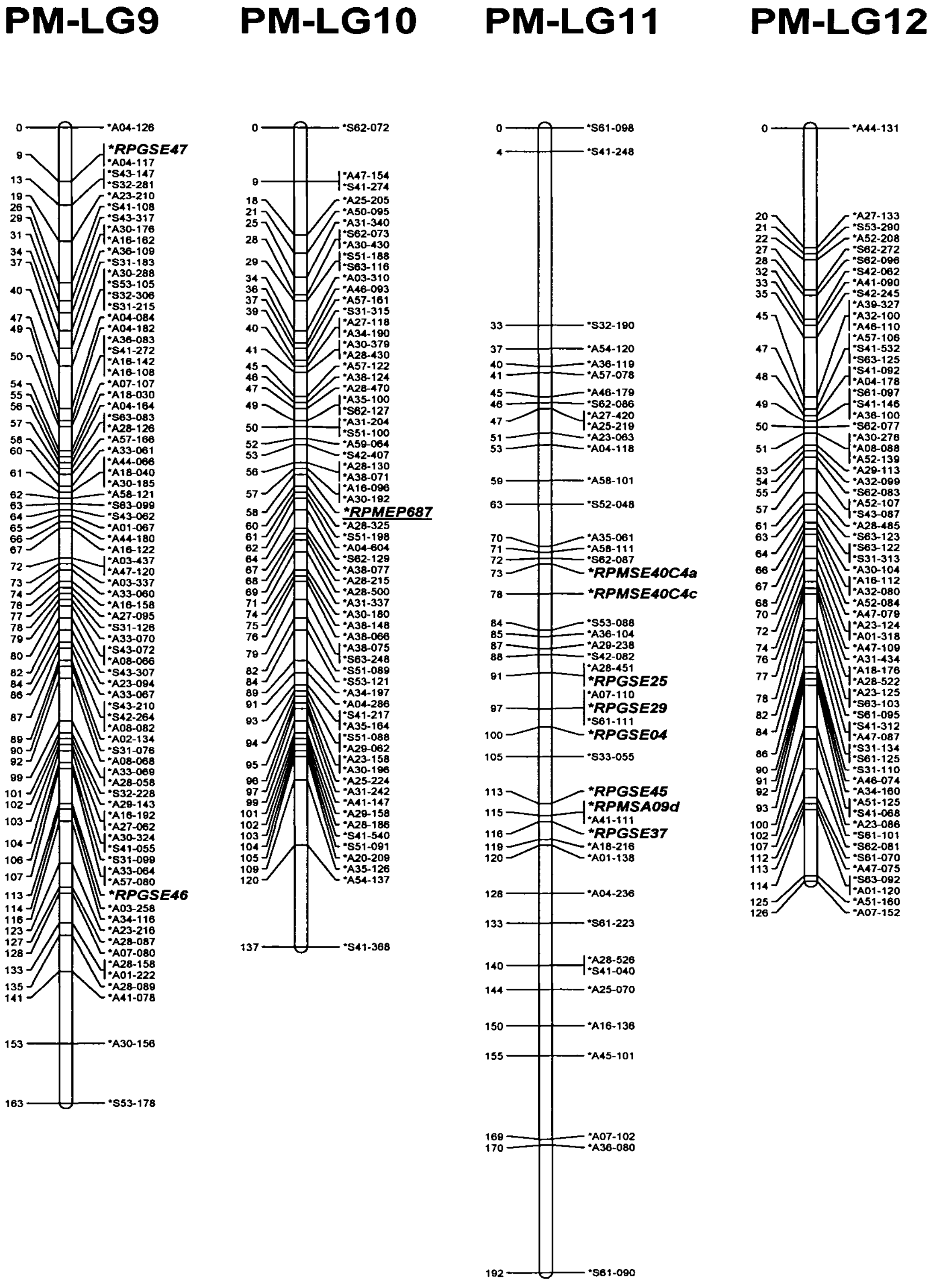

PM-LG9
0 *A04-126
9 *RPGSE47 *A04-117
13 *S43-147 *S32-281
19 *A23-210
26 *S41-108
29 *S43-317
31 *A30-176 *A16-162
34 *A36-109
37 *S31-183
40 *A30-288 *S53-105 *S32-306 *S31-215
47 *A04-084
49 *A04-182
50 *A36-083 *S41-272 *A16-142 *A16-108
54 *A07-107
55 *A18-030
56 *A04-164
57 *S63-083 *A28-126
58 *A57-166
60 *A33-061
61 *A44-066 *A18-040 *A30-185
62 *A58-121
63 *S63-099
64 *S43-062
65 *A01-067
66 *A44-180
67 *A16-122
72 *A03-437 *A47-120
73 *A03-337
74 *A33-060
76 *A16-158
77 *A27-095
78 *S31-126
79 *A33-070
80 *S43-072 *A08-066
82 *S43-307
84 *A23-094
86 *A33-067
87 *S43-210 *S42-264 *A08-082
89 *A02-134
90 *S31-076
92 *A08-068
99 *A33-069 *A28-058
101 *S32-228
102 *A29-143
103 *A16-192 *A27-062
104 *A30-324 *S41-055
106 *S31-099
107 *A33-064 *A57-080
113 *RPGSE46
114 *A03-258
116 *A34-116
123 *A23-216
127 *A28-087
128 *A07-080
133 *A28-158 *A01-222
135 *A28-089
141 *A41-078
153 *A30-156
163 *S53-178
PM-LG10
0 *S62-072
9 *A47-154 *S41-274
18 *A25-205
21 *A50-095
25 *A31-340
28 *S62-073 *A30-430
29 *S51-188 *S63-116
34 *A03-310
36 *A46-093
37 *A57-161
39 *S31-315
40 *A27-118 *A34-190
41 *A30-379 *A28-430
45 *A57-122
46 *A38-124
47 *A28-470
49 *A35-100 *S62-127
50 *A31-204 *S51-100
52 *A59-064
53 *S42-407
56 *A28-130 *A38-071
57 *A16-096 *A30-192
58 *RPMEP687
60 *A28-325
61 *S51-198
62 *A04-604
64 *S62-129
67 *A38-077
68 *A28-215
69 *A28-500
71 *A31-337
74 *A30-180
75 *A38-148
76 *A38-066
79 *A38-075 *S63-248
82 *S51-089
84 *S53-121
89 *A34-197
91 *A04-286
93 *S41-217 *A35-164
94 *S51-088 *A29-062
95 *A23-158 *A30-196
96 *A25-224
97 *A31-242
99 *A41-147
101 *A29-158
102 *A28-186
103 *S41-540
104 *S51-091
105 *A20-209
109 *A35-126
120 *A54-137
137 *S41-368
PM-LG11
0 *S61-098
4 *S41-248
33 *S32-190
37 *A54-120
40 *A36-119
41 *A57-078
45 *A46-179
46 *S62-086
47 *A27-420 *A25-219
51 *A23-063
53 *A04-118
59 *A58-101
63 *S52-048
70 *A35-061
71 *A58-111
72 *S62-087
73 *RPMSE40C4a
78 *RPMSE40C4c
84 *S53-088
85 *A36-104
87 *A29-238
88 *S42-082
91 *A28-451 *RPGSE25
97 *A07-110 *RPGSE29 *S61-111
100 *RPGSE04
105 *S33-055
113 *RPGSE45
115 *RPMSA09d *A41-111
116 *RPGSE37
119 *A18-216
120 *A01-138
128 *A04-236
133 *S61-223
140 *A28-526 *S41-040
144 *A25-070
150 *A16-136
155 *A45-101
169 *A07-102
170 *A36-080
192 *S61-090
PM-LG12
0 *A44-131
20 *A27-133
21 *S53-290
22 *A52-208
27 *S62-272
28 *S62-096
32 *S42-062
33 *A41-090
35 *S42-245
45 *A39-327 *A32-100 *A46-110
47 *A57-106 *S41-532 *S63-125
48 *S41-092 *A04-178
49 *S61-097 *S41-146 *A36-100
50 *S62-077
51 *A30-276 *A08-088 *A52-139
53 *A29-113
54 *A32-099
55 *S62-083
57 *A52-107 *S43-087
61 *A28-485
63 *S63-123
64 *S63-122 *S31-313
66 *A30-104
67 *A16-112 *A32-080
68 *A52-084
70 *A47-079
72 *A23-124 *A01-318
74 *A47-109
76 *A31-434
77 *A18-176 *A28-522
78 *A23-125 *S63-103
82 *S61-095
84 *S41-312 *A47-087
86 *S31-134 *S61-125
90 *S31-110
91 *A46-074
92 *A34-160
93 *A51-125 *S41-068
100 *A23-086
102 *S61-101
107 *S62-081
112 *S61-070
113 *A47-075
114 *S63-092 *A01-120
125 *A51-160
126 *A07-152

Table 4.6 Marker systems used for the construction of genetic linkage map and the number of markers mapped.

Marker Type	Total number of markers analyzed	Total number of markers mapped		
		Consensus map	**Maternal map**	**Paternal map**
AFLPs	841	695	472	476
SAMPL	262	213	132	136
SSR	45	30	20	20
ESTPs	5	3	2	2
Total	**1143**	**941**	**626**	**634**

Table 4.7 Linkage groups, markers mapped, and marker density for the maternal, paternal, and consensus linkage maps in black spruce.

LG	Maternal map			Paternal map			Consensus map		
	Length (cM)	Markers		Length (cM)	Markers		Length (cM)	Markers	
		Total	Average /cM		Total	Average /cM		Total	Average /cM
1	107	61	1.8	136	74	1.8	157	115	1.4
2	137	55	2.5	171	48	3.6	198	94	2.1
3	120	64	1.9	147	48	3.1	153	83	1.8
4	154	54	2.9	111	63	1.8	153	89	1.7
5	120	48	2.5	134	54	2.5	190	86	2.2
6	111	61	1.8	137	56	2.4	145	86	1.7
7	115	41	2.8	128	40	3.2	158	51	3.1
8	126	50	2.5	103	50	2.1	126	80	1.6
9	133	60	2.2	166	48	3.5	163	80	2.0
10	144	47	3.1	148	45	3.3	137	66	2.1
11	141	28	5.0	120	45	2.7	192	46	4.2
12	122	57	2.1	140	63	2.2	126	65	1.9
Total	**1530**	**626**	**2.6**	**1641**	**634**	**2.7**	**1898**	**941**	**2.2**

average of one marker every 2.2 cM. The size of linkage groups varied from 126 to 192cM, with an average of 158cM (Table 4.7; Figure 4.3).

4.3.5 Genome Length and Map Coverage

The estimated length of the black spruce genome was 1947 cM based on the method of Chakravarti *et al.* (1991), and 1976 cM according to the method of Fishmann *et al.* (2001). The observed length of the black spruce genome obtained from the consensus genetic linkage map was 1898 cM (Kosambi, K). Thus, the consensus genetic map constructed in the current study covered more than 96% of the estimated genome size of black spruce. Taking an estimated genome size of 2000 cM in black spruce, a total of 1000 randomly distributed markers at an average inter-marker distance of 1 cM is required to cover the whole black spruce genome and to saturate the map. The corresponding number of markers for an estimated genome length of 1900 cM is 950.

4.3.6 Distribution of Markers Along Linkage Groups

Significant deviations from the Poisson distribution of markers were observed for marker intervals of 2.5cM, 5cM, 10 cM, 20cM, and 40cM. For a 10 cM interval, the significant deviation (df = 10; χ^2=244; $P < 0.05$) is shown in Figure 4.4, indicating that the markers were not randomly distributed on the black spruce linkage groups. Marker distribution for other intervals (2.5cM, 5cM, 20cM, and 40cM) also showed clustering of markers ($P < 0.05$) along the linkage groups. Separate analyses were done for the AFLP and SAMPL markers to test the random distribution of markers. The independent analysis for testing the random distribution of AFLP (df = 8; χ^2 = 144; $P < 0.05$) and SAMPL (df =

6; χ^2 = 64; P < 0.05) markers indicated deviations from the random distribution. No correlation was observed between the number of markers and the size of linkage groups. These correlations further support the clustering of markers in the linkage map.

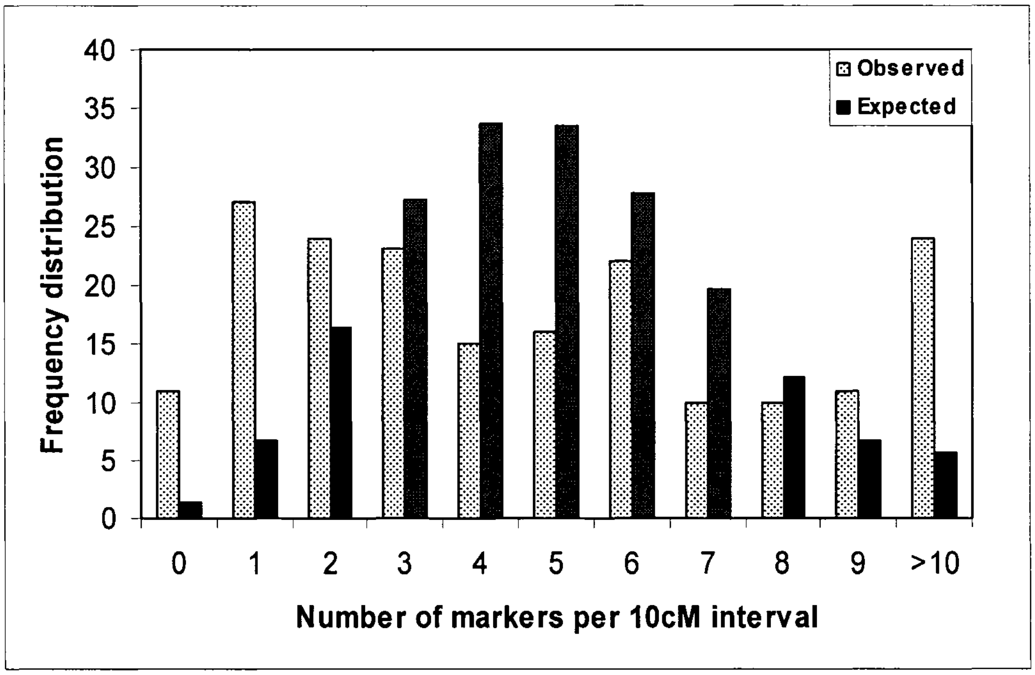
The distance between two adjacent markers on the linkage groups varied from 0 to 28.8 cM, with an average distance of 2.2 cM between any two adjacent markers (Figure 4.5; Table 4.7). This distance distribution reveals a strong skewness (P < 0.05) indicating the non-random distribution of markers along the linkage groups (Figure 4.5). Among the 929 intervals on 12 different linkage groups, 639 intervals were smaller than 2 cM (68.7%), and 88 intervals were larger than 5 cM (9.5%).

4.4 DISCUSSION

4.4.1 Genetic Linkage Map

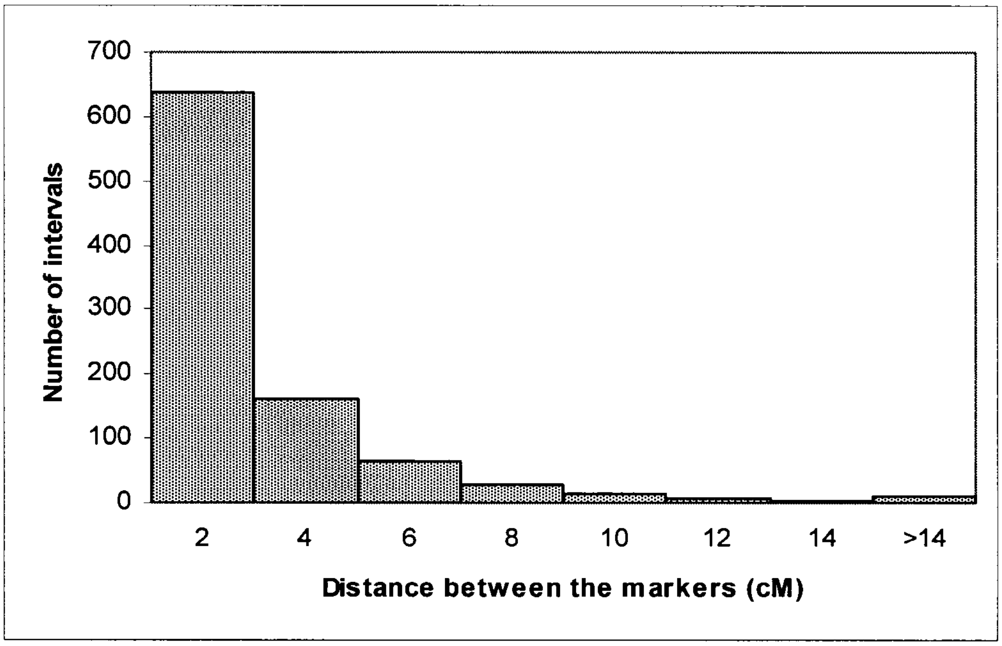
A near-saturated genetic linkage map of black spruce has been developed. This is the first genetic linkage map for black spruce, although a composite genetic linkage map has recently been reported for a black spruce x red spruce hybrid complex (Pelgas *et al.* 2005). Except in the southern part of its range, red spruce is largely sympatric with black spruce. These two species hybridize in nature, although interspecific hybrids represent a substantial but imperfect reproductive barrier for maintaining the separation of the species (Major *et al.* 2005). The differentiation of black and red spruce and their interspecific hybrids based on DNA markers, as used in Pelgas *et al.* (2005), is tenuous. The parents of the mapping pedigree in the present study represent pure black spruce (Morgenstern 1974; Boyle 1987).

Figure 4.4 Poisson distribution function for the observed and expected frequencies of the markers distributed at 10 cM interval.



40
35
30
25
20
15
10
5
0
Frequency distribution
▨ Observed
■ Expected
0
1
2
3
4
5
6
7
8
9
>10
Number of markers per 10cM interval

Figure 4.5 Distribution of the distance between two adjacent markers.


700
600
500
400
300
200
100
0
Number of intervals
2
4
6
8
10
12
14
>14
Distance between the markers (cM)

The genetic linkage map developed for black spruce is the first map in the genus *Picea* based on a TGOP. There are only two other genetic maps in the genus *Picea* that are based on pedigreed material: one in Norway spruce based on an F_1 mapping population (Acheré *et al.* 2004) and another in the black x red spruce natural hybrid complex based on F_1 and BC_1 mapping populations (Pelgas *et al.* 2005). Almost all other reported maps are based on single trees (Table 4.8). The single-tree maps are based on segregation of markers in haploid megagametophytes of maternal trees and do not take into account the segregation or recombination of markers in the paternal trees. Also, in conifers, the recombination rates appear to be lower in female gametes than those in male gametes (Groover *et al.* 1995; Plomion and O'Malley 1996). Thus, single-tree linkage maps are not as informative as genetic maps developed from diploid segregating, pedigreed populations, are specific to a single tree (genotype), and thus have limited application for QTL mapping. Black spruce, like other spruce or pine species, is highly outbred (Boyle and Morgenstern 1986; Sproule and Dancik 1996). For genome and QTL mapping in outbred plants, TGOP is more informative than any other pedigree used so far in the genus *Picea* (Liu 1998). Indeed, TGOP allows not only differentiation of up to four segregating alleles at a locus but also establishment of linkage phase among alleles in the mapping population (Liu 1998). This information is required to use genetic linkage maps for QTL mapping. Furthermore, the outbred pedigrees are more representative of natural populations in an outbred plant. In *Pinus*, where the genome and QTL mapping work has been more advanced than in its sister genus *Picea*, high-density genetic and QTL maps have been prepared using TGOP in loblolly and maritime (*Pinus pinaster*) pine (Devey *et al.* 1994; Sewell *et al.* 1999; Chagne *et al.* 2003).

4.4.2 Genome Length and Map Coverage

The black spruce genome length estimated in the present study is 1898 cM (K). This is comparable with the genome length of 1845.5 cM (K) reported for putative black x red spruce hybrids based on a composite map of F_1 and BC_1 mapping populations (Pelgas *et al.* 2005; Table 4.8) and that of 1865 cM estimated for black x red spruce controlled cross hybrids from a BC_1 mapping population (Om Rajora, unpublished data). These results support the conclusion that the length of the genetic map for black spruce is likely to be between 1900 and 2000 cM. The genetic map length of black spruce observed in this study is about 7% lower than that reported for Norway spruce (2035 cM; Acheré *et al.* 2004) and white spruce (2007 or 2059 cM; Gosselin *et al.* 2002) maps. The genome size (1C nuclear DNA contents) of Norway spruce (18.6 pg; Siljak-Yakovlev *et al.* 2002) and white spruce (20.2 pg; Ohri and Khoshoo 1986) is 17.8% and 27.8% higher, respectively, than that of black spruce (15.8 pg; Ohri and Khoshoo 1986). Although no direct relationships between the nuclear DNA contents and genetic map lengths apparently exist, the shorter genetic map length in black spruce than in Norway spruce or white spruce is apparently consistent with its comparatively smaller genome size. The results suggest that the observed genome length covers more than 96% of the estimated genome length of black spruce. This is the highest map coverage so far for any *Picea* species. The estimated length and the extent of coverage of genetic maps in different species could vary owing to differences in the mapping populations used, variation in recombination rates of the parents, and the number and types of markers used in linkage map construction (Liu 1998). The mapping pedigree and markers used in this study are different from those used in Norway spruce and white spruce (Table 4.8). The number of markers mapped in black spruce (941) is very close to the number (1000) required to completely saturate the genetic map based on an assumed genome size of 2000 cM and an average inter-marker distance of 1 cM. Thus, the black spruce genetic map reported here can certainly be considered as almost saturated.

4.4.3 Linkage Groups and Marker Density

The paternal, maternal, and consensus linkage maps consistently coalesced into 12 linkage groups, corresponding to the haploid chromosome number (n = 12) in black spruce. By contrast, in all other studies reported on genome mapping in the genus *Picea* (Table 4.8), with one exception, maternal, paternal, and/or consensus map did not coalesce into 12 linkage groups (see Table 4.8). The consensus map reported here has 941 markers distributed over 12 linkage groups, which represents an almost complete coverage of the black spruce genome, as the number of linkage groups corresponds to the haploid chromosome number. The average distance observed among adjacent markers mapped for the genetic map of black spruce (2.2 cM) is comparable with that reported for two other genetic maps in the genus *Picea* constructed from pedigreed populations (Acheré *et al.* 2004; Pelgas *et al.* 2005), although it is lower than that reported for Norway spruce (2.6 cM) and higher than that reported for the black x red spruce hybrid complex genetic maps (Table 4.8). However, the marker density of the genetic map for black spruce is the highest for any genetic map based on a single cross in the genus *Picea*, as well for the maternal and paternal genetic maps (see Table 4.8 for comparisons).

4.4.4 Marker Systems

The genetic linkage map of black spruce was constructed using AFLP, SAMPL, SSR, and ESTP marker systems. The AFLP and SAMPL systems provided a sufficient number of polymorphic and informative markers to construct a near-saturated genetic map, whereas the SSR and ESTP systems provided highly informative and co-dominant markers. Being highly informative and co-dominant, SSR and ESTP markers would be preferred for genome mapping, availability of limited numbers of these markers precludes constructing a high density to saturated genetic map using only these markers. A total of 695 AFLP markers were resolved by 40 primer combinations, showing an average multiplex ratio of 17 markers per primer combination. This multiplex ratio is comparable to that observed in Norway spruce (14; Acheré *et al.* 2004) and loblolly pine (21; Remington *et al.* 1998).

The SAMPL markers (213) were also mapped on 12 different linkage groups. The multiplex ratio for SAMPL markers (18 polymorphic mapped markers per primer combination) was comparable with that observed for AFLP markers. The SAMPL markers combine features of both AFLP and SSR markers (Gupta *et al.* 2005). The only other report where SAMPL markers have been used for genetic linkage mapping in conifers is for Norway spruce (Paglia *et al.* 1998), where 20 SAMPL markers were mapped using two primer combinations. A large number of AFLP and SAMPL markers segregated in the 3:1 ratio, which suggests that the parental genomes are heterozygous. As AFLP and SAMPL markers were dominant, the 3:1 segregating (intercross) markers may, at first, appear uninformative, but these intercross markers are useful for aligning the paternal maps to construct a consensus map, which cannot be established directly. Also the intercross

Table 4.8 Comparison of the genetic linkage maps constructed in black spruce with those for other species in the genus *Picea*.

Species	Mapping population	Mapping population size	Map type	No. markers mapped	Marker systems	No. linkage groups	Map length in cM (K)	Average distance between markers (cM)	Reference
Picea mariana	TGOP	90 F_2 progeny	Maternal	627	AFLP, SAMPL, SSR, ESTP	12	1530	2.6	This study
			Paternal	635		12	1641	2.7	
			Consensus	941		12	1898	2.2	
Picea abies	1 single tree	72 mega-gametophytes	Single tree	185	RAPD	17	3584	22.0	Binelli and Bucci 1994
	48 single trees	384(48 x 8) mega-gametophytes	Population	70	RAPD	15			Bucci *et al* 1997
	1 single tree	72 mega-gametophytes	Single tree	413	AFLP, SSR SAMPL,	29	2198.3	9.3	Paglia *et al* 1998
	F_1	73 F_1 progeny	Maternal	461	AFLP, SSR, ESTP, 5srDNA	12	1920	4.0	Acheré *et al* 2004
			Paternal	360		16	1792	4.9	
			Consensus	755		12	2035	2.6	
Picea glauca	1 single tree	47 mega-gametophytes	Single tree	47	RAPD	12	873.8	18.5	Tulsieram *et al.* 1992
	2 single trees	92	Single tree	165	RAPD, SC-AR, ESTP	23	2059.4		Gosselin *et al.* 2002
		96	Single tree	144		19	2007.7		
Picea mariana x P. rubens complex	F_1	80 F_1 progeny	Maternal	326	AFLP, SSR, ESTP, RAPD	15	1489.3	4.6	Pelgas *et al.* 2005
			Paternal	303		20	1724.6	5.6	
	BC_1	109 F_1 progeny	Maternal	313		14	1819.5	5.8	
			Paternal	281		17	1573.6	5.6	
			Male consensus	626		13	1704.8	2.8	
	Composite-F_1 and BC_1	Composite-F_1 and BC_1	Composite	1124		12	1845.5	1.6	

markers can help to identify additional linkage groups that were not represented in the parental maps (Hanley *et al.* 2002).

Thirty SSR loci were mapped onto 11 linkage groups. Because of their co-dominant behavior, the SSRs were highly informative for integrating the maternal and paternal maps to construct the consensus map. Out of 30 SSR loci, 17 SSR were derived from cDNA sequences, thus are targeting genes. The mapped cDNA-based SSR markers are good candidates for comparative mapping among various spruce species and other conifer species.

Most EST primer pairs resolved multilocus patterns, which is not surprising given the occurrence of multigene families in conifers (Kinlaw and Neale 1997). Also, the rate of polymorphism observed in ESTPs was very low, which suggests that more powerful methods, such as methods to detect point mutation (Plomion *et al.* 1999; Temesgen *et al.* 2000) or the prospecting of variation in the non-coding regions flanking the ESTs, can be used to increase the resolution of polymorphism. Nevertheless, the mapped three ESTP loci along with cDNA-based SSR markers provide good candidates for comparative mapping in *Picea*, Pinaceae, or other conifers (Komulainen *et al.* 2003; Krutovsky *et al.* 2004; Pelgas *et al.* 2005).

4.4.5 Clustering of Markers

Even though only markers segregating in Mendelian ratios and not those showing a distorted segregation were used for the linkage analysis, clustering of AFLP and SAMPL markers was detected in the linkage group of black spruce. These results agree with the clustering of AFLP markers reported for genetic linkage maps of *Picea abies* (Scotti *et al.*

2005), *Pinus taeda* (Remington *et al.* 1999), and *Pinus sylvestris* (Yin *et al.* 2003), but in contrast to random distribution of AFLP markers in the genetic maps reported for Norway spruce (Acheré *et al.* 2004) and putative black X red spruce hybrid complex (Pelgas *et al.* 2005). It should be noted that the study by Scotti *et al.* (2005) was performed specifically to examine the distribution of marker classes in a genetic linkage map of Norway spruce.

The non-random distribution of markers may be caused by non-random and unequal crossing over and recombination along the chromosome length. The recombination is suppressed in the centromeric and heterochromatic pericentromeric regions (Tanksley *et al.* 1992), and the presence of heterochromatin in pericentromeric regions is a general feature of plant chromosomes. Assuming a random distribution of markers, low levels of meiotic recombination may well cause markers that are physically well separated, to cluster on a linkage map.

CHAPTER 5

CONCLUSIONS

The work described in this thesis was undertaken with the objectives to develop an EST resource in black spruce, identify genes differentially-expressed in needle and cambium tissues and finally to construct a genetic linkage map with medium to high saturation for black spruce.

In chapter two, EST resource has been developed by sequencing ESTs from a cDNA library constructed from RNA extracted from needles. Although, only 4608 ESTs were generated for black spruce in this study, compared to the thousands of ESTs generated for *Arabidopsis*, rice, loblolly pine and other species, this is the first attempt for comparative sequence analysis involving three different groups of extant seed plants i.e. conifers, monocots and eudicots. Approximately 13% of ESTs from black spruce were not identified by BLAST searches of the dbEST at NCBI. The black spruce EST resource generated in the present study represents at least 586 coding sequences of genes with known functions, including some putative transcripts specific to black spruce and transcripts which code for proteins of unknown function. These 586 ESTs with known functions have been grouped into functional categories, such as protein synthesis, energy and metabolism, disease and stress response, photosynthesis, transcription and post-transcription, protein destination and storage, cell structure, lipid biosynthesis and metabolism, secondary metabolism, signal transduction, transporters, transposons, expressed and unknown proteins (Table 2.1). These functional categories provide a general overview of genes involved in different functional pathways in needles of black spruce.

The sequence similarity of ESTs from black spruce and contigs with different species from conifers (white spruce, Sitka spruce and loblolly pine), eudicots (poplar and *Arabidopsis*) and monocots (rice and wheat) showed greater similarity to conifers than to eudicots followed by monocots. The 5′ end sequences were found to be more similar to dbEST than 3′ end sequences of black spruce transcripts as has been reported for other species. Fifty-eight different microsatellite repeat motifs were identified in the black spruce EST resource. The sequences containing repeats have been used for the development of microsatellite markers for use in genetic linkage mapping and population genetic analysis.

In chapter three, I have identified differentially-expressed transcripts in needle and cambium tissues, the primary sites for key functional pathways in plants. The needle and cambium SSH cDNA libraries uncovered transcripts involved in almost all major functional pathways. The SSH cDNA libraries revealed 256 transcripts that showed no identity to the total spruce EST database at NCBI. Annotation of ESTs identified 200 additional coding sequences of genes with known functions, thus raising the total number of coding sequences with known functions in black spruce to 786. The distribution of transcripts coding for proteins with known functions in needles was different from that of cambium tissue. Genes involved in cell structure, disease and stress response mechanisms were predominantly expressed in the cambium tissues. Conversely, genes involved in photosynthesis and energy metabolism were predominantly expressed in the needle tissues. Transcripts coding for ribulose-1, 5-biphosphate carboxylase and MT proteins were expressed more in the needle and cambium SSH cDNA libraries respectively (Tables 3.6, 3.7).

In chapter four, the first near-saturated genetic linkage map of black spruce, with above 96% genome coverage has been constructed. The maternal map had 627 and the paternal map had 636 markers distributed over 12 linkage groups each (Tables 4.6, 4.7). The consensus genetic linkage map consisted of 941 markers distributed over 12 linkage groups (Figure 4.3, Tables 4.6, 4.7), and covered almost the entire black spruce genome. The mapped markers included 695 AFLPs, 213 SAMPL, 30 microsatellites, and 3 ESTPs (Table 4.6). Total estimated map length was 1898 cM, with an average of one marker every 2.2 cM (Table 4.7).Unlike previous studies, the maternal, paternal, and consensus maps consistently coalesced into 12 linkage groups, corresponding to the haploid chromosome number of 12 in the genus *Picea*. For a single cross/mapping pedigree, the consensus, maternal, or paternal maps are of the highest density in any spruce species. Also, this is the first genetic map based on a three-generation outbred pedigree in the genus *Picea*. The genome length in *P. mariana* is likely between 1900-2000 cM. This genetic map can serve as a reference map for mapping quantitative trait loci and determining the genetic basis of complex quantitative traits of interest, as well as for comparative genomics.

REFERENCES

Acheré V, Faivre Rampant P, Jeandroz S, Besnard G, Markussen T, Aragones A, Fladung M, Ritter E, Favre JM (2004) A full saturated linkage map of *Picea abies* including AFLP, SSR, ESTP, 5S rDNA and morphological markers. Theor Appl Genet 108:1602–1613

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno R, Kerlavage AR, McCombie WR, Venter JC (1991) Complementary DNA sequencing: "expressed sequence tags" and the human genome project. Science 252: 1651-1656

Akopian AN, Wood JN (1995) Peripheral nervous system-specific genes identified by subtractive cDNA cloning. J Biol Chem 270:21264-21270

Allona I, Quinn M, Shoop E, Swope K, Cyr SS, Carlis J, Riedl J, Retzel E, Campbell MM, Sederoff R, Whetten RW (1998) Analysis of xylem formation in pine by cDNA sequencing. Proc Natl Acad Sci 95:9693-9698

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410

Arabidopsis genome initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* Nature 408:796-815

Bennett MD, Leitch IJ (2005) Genome size evolution in plants. In: Gregory TR (ed) The Evolution of the Genome. Elsevier, San Diego, pp 89-162

Bennett MD, Leitch IJ, Price HJ, Johnston JS (2003) Comparisons with *Caenorhabditis* (~ 100 Mb) and *Drosophila* (~175Mb) using flow cytometry show genome size in *Arabidopsis* to be ~157 Mb and thus ~25 % larger than the *Arabidopsis* Genome Initiative of ~125 Mb. Ann Bot 91:1–11

Bhalerao R, Keskitalo J, Sterky F Erlandsson R, Bjorkbacka H, Birve SJ, Karlsson J, Gardestrom P, Gustafsson P, Lundeberg J, Jansson S (2003) Gene expression in autumn leaves. Plant Physiol 131:430–442

Binelli G, Bucci G (1994) A genetic linkage map of *Picea abies* Karst., based on RAPD markers, as a tool in population genetics. Theor Appl Genet 88:283–288

Bonaldo MF, Lennon G, Soares MB (1996) Normalisation and subtraction: Two approaches to facilitate gene discovery. Genome Res 6:791-806

Boss PK, Vivier M, Matsumoto S, Dry IB, Thomas MR (2001) A cDNA from grapevine (*Vitis vinifera* L.), which shows homology to AGAMOUS and SHATTERPROOF, is not only expressed in flowers but also throughout berry development. Plant Mol Biol 45:541–553

Bowe LM, Coat G and dePamphilis CW (2000) Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Proc Natl Acad Sci USA 97:4092-4097

Boyle TJB (1987) A diallel cross in black spruce. Genome 29:180-186

Boyle TJB, Morgenstern EK (1986) Estimates of outcrossing rates in six populations of black spruce in central New Brunswick. Silvae Genet 35:102–106

Brinker M, Zyl LV, Liu W, Craig D, Sederoff RR, Clapham DH, Arnold SV (2004) Microarray Analyses of Gene Expression during Adventitious Root Development in *Pinus contorta*. Plant Physiol 135:1526–1539

Bucci G, Kubisiak TL, Nance WL, Menozzi P (1997) A population consensus partial linkage map of *Picea abies* Karst. based on RAPD markers. Theor Appl Genet 95:643–654

Buetow KH, Edmonson MN, Cassidy DB (1999) Reliable identification of large numbers of candidate SNPs from public EST data. Nat Genet 21:323–325

Burke J, Wang H, Hide W, Davison DB (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. Genome Res 8:276–290

Cervera MT, Storme V, Ivens B, Gusmao J, Liu BH, Hostyn V, Slycken JV, Montagu MV, Boerjan W (2001) Dense genetic linkage maps of three *Populus* species (*Populus deltoides*, *P. nigra* and *P. trichocarpa*) based on AFLP and microsatellite markers. Genetics 158:787–809

Chagne D, Brown GR, Lalanne C, Madur D, Pot D, Neale DB, Plomion C (2003) Comparative genome and QTL mapping between maritime and loblolly pines. Mol Breeding 12:185–195

Chakravarti A, Lasher LK, Reefer JE (1991) A maximum-likelihood method for estimating genome length using genetic linkage data. Genetics 128:175–182

Chang S, Puryear J, Cairney J (1993) A simple and efficient method for isolating RNA from pine trees. Plant Mol Biol Rep 11:113–116

Cobbett C, Goldsbrough P (2002) Phytochelatins and metallothioneins: roles in heavy metal detoxification and homeostasis. Ann Rev Plant Biol 53:159–182

Cooke R, Raynal M, Laudie M, Grellet F, Delseny M, Morris PC, Guerrier D, Giraudat J, Quigley F, Clabault G, Li YF, Mache R, Krivitzky M, Gy IJ, Kreis M, Lecharny A, Parmentier Y, Marbach J, Fleck J, Clement B, Philipps G, Herve C, Bardet C, Tremousaygue D, Hofte H, (1996) Further progress towards a catalogue of all *Arabidopsis* genes: analysis of a set of 5000 non-redundant ESTs. Plant J 9:101-124

Dean C, van den Elzen P, Tamaki S, Dunsmuir P, Bedbrook J (1985) Differential expression of the eight genes of the petunia ribulose bisphosphate carboxylase small subunit multi-gene family. EMBO J 4:3055-3061

Dedonder A, Rethy R, Fredericq H, Van Montagu M, Krebbers E (1993) Arabidopsis rbcS genes are differentially regulated by light. Plant Physiol 101:801-808

DeRocher EJ, Quigley F, Mache R, Bohnert HJ (1993) The six genes of the Rubisco small subunit multigene family from *Mesembryanthemum crystallinum*, a facultative CAM plant. Mol Gen Genet 239:450–462

Devey ME, Fiddler TA, Liu BH, Knapp SJ, Neale BD (1994) An RFLP linkage map for loblolly pine based on a three-generation outbred pedigree. Theor Appl Genet 88:273–278

Diatchenko L, Lau Y-FC, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Pukyanov K, Gurskaya N, Sverdlov ED, Siebert PD (1996) Suppression subtractive hybridization: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc Natl Acad Sci USA 93:6025–6030

Dong J-Z, Dunstan DI (1996) Expression of abundant mRNAs during somatic embryogenesis of white spruce [*Picea glauca* (Moench) Voss]. Planta 199:459 466

Duret L, Dorkeld F, Gautier C (1993) Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. Nucleic Acids Res 21:2315-2322

Elsik CG, Williams CG (2000) Retroelements contribute to excess low-copy number DNA in pine. Mol Gen Genet 264:47-55.

Etscheid M, Klümper S, Riesner D (1999) Accumulation of metallothionein-like mRNA in Norway spruce under environmental stress. J Phytopathol 147:207-213

Ewers FW, Aloni R (1987) Seasonal secondary growth in needle leaves of *Pinus strobus* and *Pinus brutia*. Am J Bot 74:980-987

Ewing B, Green P (1998) Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Res 8:186-194

Ewing B, Hillier L, Wendl M, Green P (1998) Base-calling of automated sequencer traces using *Phred.* I. Accuracy assessment. Genome Res 8:175-185

Farjon A (1990) Pinaceae. Koleltz Scientific Books, Konigstein, Germany.

Farrar JL (1995) Trees in Canada. Fitzhenry and Whiteside, Markham ON, and the Canadian Forest Services, Ottawa, ON.

Fernández P, Paniego N, Lew S, Hopp HE, Heinz RA (2003) Differential representation of sunflower ESTs in enriched organ-specific cDNA libraries in a small scale sequencing project. BMC Genom 4:40

Fishman L, Kelly AJ, Morgan E, Willis JH (2001) A genetic map in the *Mimulus guttatus* species complex reveals transmission ratio distortion due to heterospecific interactions. Genetics 159:1701–1716

Florin R (1963) The distribution of conifer and taxad genera in time and space. Acta Horti Bergiani 20:121-312

Foucart C, Paux E, Ladouce N, San-Clemente H, Grima-Pettenati J, Sivadon P (2006) Transcript profiling of a xylem vs phloem cDNA subtractive library identifies new genes expressed during xylogenesis in *Eucalyptus*. New Phytolog 170:739-752

Frazer KA, Elnitski L, Church, DM, Dubchak I, Hardison RC (2003) Cross-species Sequence Comparisons: A Review of Methods and Available Resources. Genome Res 13:1-12

Goff SA *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296: 92–100

Gosselin I, Zhou Y, Bousquet J, Isabel N (2002) Megagametophyte-derived linkage maps of white spruce (*Picea glauca*) based on RAPD, SCAR and ESTP markers. Theor Appl Genet 104:987–997

Grant D, Cregan P, Shoemaker RC (2000) Genome organization in dicots: Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. Proc Natl Acad Sci USA 97:4168-4173

Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo testcross: mapping strategy and RAPD markers. Genetics 137:1121–1137

Gray JC, Kekwick RG (1974) The synthesis of the small subunit of ribulose 1,5-bisphosphate carboxylase in the french bean *Phaseolus vulgaris*. Eur J Biochem 44:491–500

Gregory, T.R. 2005. Genome size evolution in animals. In: Gregory TR (ed) The Evolution of the Genome. Elsevier, San Diego, pp 3-87

Groover A, Williams CG, Devey ME, Lee JM, Neale DB (1995) Sex-related differences in meiotic recombination frequency in *Pinus taeda.* J Hered 86:157–158

Gupta AK, Kang BY, Roy JK, Rajora OP (2005) Large scale development of selective amplification of microsatellite polymorphic loci (SAMPL) markers in spruce (*Picea*). Mol Eco Not 5:481-483

Ha CM, Kim G-T, Kim BC, Jun JH, Soh MS, Ueno Y, Machida Y, Tsukaya H, Nam HG (2003) The BLADE-ON-PETIOLE 1 gene controls leaf pattern formation through the modulation of meristematic activity in *Arabidopsis*. Development 130: 161-172

Hanley S, Barker A, Van Ooijen W, Aldam C, Harris L, Ahman I, Larsson S, Karp A. (2002) A genetic linkage map of willow (*Salix viminalis*) based on AFLP and microsatellite markers. Theor Appl Genet 105:1087–1096

Hopkins WG, Hunter NPA (2004) Introduction to Plant Physiology. John Wiley and Sons, Inc. Hoboken, NJ

Hutchison KW, Singer PB, McInnis S, Diaz-Sala C, Greenwood MS (1999) Expansins Are Conserved in Conifers and Expressed in Hypocotyls in Response to Exogenous Auxin. Plant Physiol 120: 827-832

Jackson RJ (1993) Cytoplasmic regulation of mRNA function: The importance of the 3′ untranslated region. Cell 74:9-14

Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28:27-30

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

Kirst M, Johnson AF, Baucom C, Ulrich E, Hubbard K, Staggs R, Paule C, Retzel E, Whetten R, Sederoff R. (2003) Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana.* Proc Natl Acad Sci USA 100:7383-7388

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

Kosambi DD (1944) The estimation of map distances from recombination values. Ann Eugen 12:172-175

Kozlowski TT, Pallardy SG eds (1997) Physiology of woody plants. Academic Press, New York.

Kriebl (1985) DNA sequence component in *Pinus strobus* nuclear genome. Can J Forest Res 15:1-4

Krutovsky KV, Troggio M, Brown GR, Jermstad KD, Neale DB (2004) Comparative mapping in the Pinaceae. Genetics 168:447–461

Ku HM, Vision T, Liu J, Tanksley SD (2000) Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. Proc Natl Acad Sci USA 97:9121-9126.

Lange K, Boehnke M (1982) How many polymorphic genes will it take to span the human genome? Am J Hum Genet 34:842–845

Leitch IJ, Hanson L, Winfield M, Parker J, Bennett MD (2001) Nuclear DNA C-values complete familial representation in Gymnosperms. Ann Bot 88:843-849.

Li H, Gu X, Dawson VL, Dawson TM (2004) Identification of calcium- and nitric oxide-regulated genes by differential analysis of library expression (DAZLE). Proc Natl Acad Sci USA 101:647-652

Liang P, Pardee AB (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257:967-971

Lisitsyn NA, Lisitsyn N, Wigler M (1993) Cloning the differences between two complex genomes. Science 259:946-951

Liu BH (1998) Statistical genomics: linkage mapping and QTL analysis. CRC Press, New York.

Mahalingam R, Gomez-Buitrago A, Eckardt N, Nigam Shah N, Guevara-Garcia A, Day P, Raina R, Fedoroff NV (2003) Characterizing the stress/defense transcriptome of Arabidopsis. Genome Biol 4:R20

Major JE, Mosseler A, Johnsen K, Rajora OP, Barsi DC, Kim K-H, Park J-M, Campbell M (2005) Reproductive barriers and hybridity in two spruces, *Picea rubens* and *Picea mariana*, sympatric in eastern North America. Can J Bot 83: 163-175

Matsumura H, Nirasawa S, Terauchi R (1999) Transcript profiling in rice (*Oryza sativa* L.) seedlings using serial analysis of gene expression (SAGE). Plant J 20: 719–726

Michalek W, Weschke W, Pleissner KP, Graner A (2002) EST analysis in barley defines a unigene set comprising 4,000 genes. Theor Appl Genet 104:97-103

Miller CN (1988) The origin of modern conifer families. In: Beck CB (ed) Origin and evolution of Gymnosperms. Columbia University Press, New York, pp 448–486

Miller W, Makova KD, Nekrutenko A, Hardison RC (2004) Comparative genomics. Annu Rev Genomics Hum Genet 5: 15–56

Morgenstern EK (1974) A diallel cross in black spruce *Picea mariana* (Mill.) B.S.P. Silvae Genet 23:67–70

Morgenstern EK, Wang BSP (2001) Trends in forest depletion, seed supply, and reforestation in Canada during the past four decades. Forest Chron 6:1014–1021

Murray BG (1998) Nuclear DNA amount in gymnosperms Ann Bot 82: 3-15

Nagy F, Fluhr R, Morelli G, Kuhlemeier C, Poulsen C, Keith B, Boutry M, Chua N-H. (1986) The Rubisco small subunit gene as a paradigm for studies on differential gene expression during plant development. Phil Trans R Soc Lon 313:409-417

Neale DB, Williams CG (1991) Restriction fragment length polymorphism mapping in conifers and applications to forest genetics and tree improvement. Can J For Res 21:545-554

Nuccio ML, Thomas TL (1999) ATS1 and ATS3: two novel embryo-specific genes in *Arabidopsis thaliana.* Plant Mol Biol 39: 1153–1163

Ohlrogge J, Benning C (2000) Unraveling Plant Metabolism by EST Analysis. Curr Opin Plant Biol 3: 224-228

Ohri D, Khoshoo TN (1986) Genome size in gymnosperms. Plant Syst Evol 153:119–132

Omann F, Beaulieu N, Tyson H (1994) cDNA sequence and tissue-specific expression of an anionic flax peroxidase. Genome 37: 137–147

Opsahl-Ferstad HG, Le Deunff E, Dumas C, Rogowsky PM (1997) ZmEsr, a novel endosperm-specific gene expressed in a restricted region around the maize embryo. Plant J 12: 235–246

Paglia GP, Olivieri AM, Morgante M (1998) Towards second generation STS (sequence-tagged sites) linkage maps in conifers: a genetic map of Norway spruce (*Picea abies* K.). Mol Gen Genet 258:466–478

Palmiter RD (1998) The elusive function of metallothioneins. Proc Natl Acad Sci USA 95: 8428-8430

Pavy N, Laroche J, Bousquet J, Mackay J (2005a) Large-scale statistical analysis of secondary xylem ESTs in pine. Plant Mol Biol 57:203–224

Pavy N, Paule C, Parsons L, Crow JA, Morency MJ, Cooke J, Johnson JE, Noumen E, Guillet-Claude C, Butterfield Y, Barber S, Yang G, Liu J, Stott J, Kirkpatrick R,Siddiqui A, Holt R, Marra M, Seguin A, Retzel E, Bousquet J and MacKay J (2005b) Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters. BMC Genom 6:144

Pelgas B, Bousquet J, Beauseigle S, Isabel N (2005) A composite linkage map from two crosses for the species complex *Picea mariana* X *Picea rubens* and analysis of synteny with other Pinaceae. Theor Appl Genet 111:1466–1488

Plomion C, Hurme P, Frigerio J-M, Ridolfi M, Pot D, Pionneau C, Avila C, Gallardo F, David H, Neutelings G, Campbell M, Canovas FM, Savolainen O, Bodénès C, Kremer A (1999) Developing SSCP markers in two *Pinus* species. Mol Breeding 5:21–31

Plomion C, O'Malley DM (1996) Recombination rate differences for pollen parents and seed parents in pine. Heredity 77:341–350

Rajora OP, Mann IK, Shi Y-Z (2005) Genetic diversity and population structure of boreal white spruce (*Picea glauca*) in pristine conifer-dominated and mixedwood forest stands. Can J Bot 83:1096-1105

Rajora OP, Rahman MH, Dayanandan S, Mosseler A (2001) Isolation, characterization, inheritance and linkage of microsatellite DNA markers in white spruce (*Picea glauca*) and their usefulness in other spruce species. Mol Gen Genet 264:871-882

Ralph SG, Yueh H, Friedmann M, Aeschliman D, Zeznik JA, Nelson CC, Butterfield YS, Kirkpatrick R, Liu J, Jones SJ, Marra MA, Douglas CJ, Ritland K, Bohlmann J (2006) Conifer defence against insects: microarray gene expression profiling of Sitka spruce (*Picea sitchensis*) induced by mechanical wounding or feeding by spruce budworms (*Choristoneura occidentalis*) or white pine weevils (Pissodes strobi) reveals large-scale changes of the host transcriptome. Plant Cell Environ 29:1545-1570

Reddy AR, Ramakrishna W, Chandrasekhar A, Nagabhushana I, Ravindra Babu P, Bonaldo MF, Soarrese MB, Bennetzen JL (2002) Novel genes are enriched in normalized cDNA libraries from drought stressed seedlings of indica rice (*Oryza sativa* L.cv.Nagina22). Genome 45: 204-211

Remington DL, Whetten RW, Liu BH, O'Malley DM (1999) Construction of an AFLP genetic map with nearly complete genome coverage in *Pinus taeda*. Theor Appl Genet 98:1279–1292

Robinson NJ, Tommey AM, Kuske C, Jackson PJ (1993) Plant metallothioneins. Biochem J 295: 1–10

Rounsley SD, Glodeck A, Sutton G, Adams MD, Somerville CR, Venter JC, Kerlavage AR (1996) The construction of *Arabidopsis* expressed sequence tag assemblies. Plant Physiol 112:1177-1183

Rozen S, Skaletsky HJ (2000) Primer 3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa NJ, pp 365–386

Rudd S (2003) Expressed sequence tags: alternative or complement to whole genome sequences? Trends Plant Sci 8: 321–329

Sablowski RW, Meyerowitz EM (1998) A homolog of NO APICAL MERISTEM is an immediate target of the floral homeotic genes APETALA3/PISTILLATA. Cell 92: 93–103

Sakuta C, Satoh S (2000) Vascular tissue-specific gene expression of xylem sap glycine-rich proteins in root and their localization in the walls of metaxylem vessels in cucumber. Plant Cell Physiol 41:627-638

Sasaki T *et al.* (2002) The genome sequence and structure of rice chromosome 1. Nature 420:312-316

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470

Scotti I, Burelli A, Cattonaro F,Chagné D, Fuller J, Hedley PE, Jansson G, Lalanne C, Madur D, Neale D, Plomion C, Powell W, Troggio M, Morgante M (2005) Analysis of the distribution of marker classes in a genetic linkage map: a case study in Norway spruce (*Picea abies* Karst). Tree Genet Genom 1:93–102

Sewell MM, Sherman BK, Neale DB (1999) A consensus map for loblolly pine (*Pinus taeda* L.). I. Construction and integration of individual linkage maps from two outbred three-generation pedigrees. Genetics 151:321–330

Siljak-Yakovlev S, Cerbah M, Coulaud J, Stoian V, Brown SC, Zoldos V, Jelenic S, Papes D (2002) Nuclear DNA content, base composition, heterochromatin and rDNA in *Picea omorika* and *Picea abies*. Theor Appl Genet 104:505–512

Spreitzer RJ (2002) RUBISCO: Structure, Regulatory Interactions, and Possibilities for a Better Enzyme. Ann Rev Plant Biol 53: 449-475

Sproule AT, Dancik BP (1996) The mating system of black spruce in north-central Alberta, Canada. Silvae Genet 45:159–164

Sterky *et al.* (1998) Gene discovery in the wood-forming tissues of poplar: analysis of 5,692 expressed sequence tags. Proc Natl Acad Sci USA 95:13330-13335

Stirling B, Yang Z K, Gunter LE, Tuskan GA, Bradshaw HD (20003) Comparative sequence analysis between orthologous regions of the *Arabidopsis* and *Populus* genomes reveals substantial synteny and microcollinearity. Can J For Res 33: 2245–2251

Temesgen B, Neale DB, Harry DE (2000) Use of haploid mixtures and heteroduplex analysis enhance polymorphism revealed by denaturing gradient gel electrophoresis. BioTechniques 20:114–122

The State of Canada's Forests 2005–2006. NRC Press. Available from http://www.nrcan .gc.ca/cfs-scf/national/what-quoi/sof/latest_e.html (accessed on September 25, 2006)

Tulsieram LK, Glaubitz JC, Kiss G, Carlson JE (1992) Single tree genetic linkage analysis in conifers using haploid DNA from megagametophytes. BioTechnology 10:686-690

Tuskan GA *et al.* (2006) The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313: 1596 – 1604

Van Ooijen JW, Voorrips RE (2001) JOINMAP 3.0, Software for the calculation of genetic linkage maps. Plant Research International, Wageningen, the Netherlands. http://www.plant.wageningen-ur.nl

Vandepoele K, Simillion C, Van de Peer Y (2002) Detecting the undetectable: uncovering duplicated segments in *Arabidopsis* by comparison with rice. Trends Genet 18:606-608

Viereck LA, Johnston WF (1990) *Picea mariana* (Mill.) B.S.P. - Black spruce. In: Burns RM, Honkala BH (eds) Vol 1 Conifers. Silvics of North America. USDA, Forest Service, Agriculture Handbook 654, Washington DC, pp 227-237

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484-487

Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 93:77–78

Vos P, Hogers R, Bleeker M, Reijans M, Van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprints. Nucleic Acids Res 23:4407–4414

Walden AR, Walter C, Gardner RC (1999) Genes expressed in *Pinus radiata* male cones include homologs to anther-specific and pathogenesis response genes. Plant Physiol 121:1103–1116

Wang X-L, He R-F, He G-C (2005) Construction of suppression subtractive hybridization libraries and identification of brown planthopper-induced gene. J Plant Physiol 162:1254-1262

Witsenboer H, Vogel J, Michelmore R W (1997) Identification, genetic localization, and allelic diversity of selectively amplified microsatellite polymorphic loci in lettuce and wild relatives (*Lactuca* spp.). Genome 40:923-936

Woo HH, Brigham LA, Hawes MC (1995) Molecular cloning and expression of mRNAs encoding H1 histone and an H1 histone-like sequences in root tips of pea (*Psium sativum* L.). Plant Mol Biol 28:1143–1147

Wu H, Michler CH, LaRussa L, Davis JM (1999) The pine Pschi4 promoter directs wound-induced transcription. Plant Sci 142:199-207

Xiong L, Lee MW, Qi M, Yang Y (2001) Identification of defense-related rice genes by suppression subtractive hybridization and differential screening. Mol Plant Microbe Int 14:685–692

Yakovlev IA. Carl-Gunnar Fossdal, C-G, Johnsen Ø, Junttila O, Skrøppa T (2006) Analysis of gene expression during bud burst initiation in Norway spruce via ESTs from subtracted cDNA libraries. Tree Genet Gen 2:39–52

Yu LH, Umeda M, Liu JY, Zhao NM, Uchimiya H (1998) A novel MT gene of rice plants is strongly expressed in the node portion of the stem. Gene 206:29–35

Yu *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. Indica). Science 296:79-91

Yin TM, Wang XR, Andersson B, Lerceteau-Kohler E (2003) Nearly complete genetic maps of *Pinus sylvestris* L. (Scots pine) constructed by AFLP marker analysis in a full-sib family. Theor Appl Genet 106:1075–1083

APPENDIX

OPEN READING FRAMES (ORFs) IDENTIFIED IN THE PREDICTED GENE PRODUCTS IN BLACK SPRUCE

Identified ORFs in the predicted gene product in black spruce	Genbank Accession number	Black spruce clone number
14-3-3-like protein	gi|6752903|gb|AAF27931.1|	2463-700
60S ribosomal protein L10A	gi|10444370|gb|AAG17879.1|	1267-700
60S ribosomal protein L12	gi|40287508|gb|AAR83868.1|	2141-700
60S ribosomal protein L13E	gi|12580867|emb|CAC27142.1|	2471-700
60S ribosomal protein L17	gi|31432581|gb|AAP54196.1|	402-700
60S ribosomal protein L5	gi|34484312|gb|AAQ72789.1|	1489-700
ADP-ribosylation factor	gi|50511366|gb|AAT77289.1|	196-700
Alpha subunit of translation elongation factor 1	gi|1321656|dbj|BAA08249.1|	2683-700
Antimicrobial peptide 1	gi|15638607|gb|AAL05052.1|	2219-700
Ascorbate peroxidase	gi|39939493|gb|AAR32786.1|	890-700
At1g22780	gi|56236126|gb|AAV84519.1|	1717-700
At1g50010/F2J10_12	gi|23506129|gb|AAN31076.1|	1928-700
At3g59540/T16L24_90	gi|23507773|gb|AAN38690.1|	1617-800
At4g13350	gi|30725432|gb|AAP37738.1|	2114-800
At5g19770/T29J13_190	gi|23505949|gb|AAN28834.1|	1195-700
ATP-dependent protease subunit	gi|29469692|gb|AAO74020.1|	1973-800
Beta-tubulin 1	gi|1743277|emb|CAA70891.1|	629-700
Beta-tubulin 5	gi|37038246|gb|AAQ88118.1|	425-700
Calmodulin	gi|41072353|gb|AAR99412.1|	1843-700
Chalcone synthase	gi|20689|emb|CAA43166.1|	413-700
Chitinase	gi|1161165|gb|AAA85364.1|	2117-700
Chlorophyll a/b-binding protein	gi|9719392|gb|AAF97781.1|	1845-700
Copper chaperone	gi|47176684|gb|AAT12488.1|	2847-800
Cyclophilin	gi|4454307|emb|CAA10766.1|	2973-700
Cytochrome c	gi|20137614|sp|O22642|	2584-700
Defender against cell death 1	gi|6014902|sp|O65085|	2012-700
Disease resistance gene	gi|40018852|gb|AAR36911.1|	2883-700
E2, ubiquitin-conjugating enzyme, putative	gi|21555312|gb|AAM63831.1|	709-700
Elongation factor 1A	gi|18339|emb|CAA42843.1|	1424-700
Eukaryotic initiation factor 5A4	gi|2225883|dbj|BAA20878.1|	358-700
Eukaryotic translation initiation factor 5A isoform IV	gi|33325123|gb|AAQ08194.1|	1981-700
Eukaryotic translation initiation factor 5A isoform VI	gi|33325127|gb|AAQ08196.1|	2789-700
Glutaredoxin	gi|1076561|pir||S54825	1809-700
Glycine-rich RNA-binding protein	gi|4704605|gb|AAD28176.1|	1085-700
GTP binding protein	gi|12311684|emb|CAC24477.1|	617-700
H2A homolog	gi|2317760|gb|AAB66346.1|	1262-700

<table>
<tr><td>Histone H3</td><td>gi|10732809|gb|AAG22548.1|</td><td>324-700</td></tr>
<tr><td>Histone H3.2 protein</td><td>gi|37991915|gb|AAR06361.1|</td><td>1936-700</td></tr>
<tr><td>Lhcb5 protein</td><td>gi|22750|emb|CAA78900.1|</td><td>2138-700</td></tr>
<tr><td>Light harvesting chlorophyll a /b-binding protein Lhcb1*2-2</td><td>gi|607152|emb|CAA57409.1|</td><td>2403-700</td></tr>
<tr><td>Low molecular weight heat-shock protein</td><td>gi|1213118|emb|CAA63571.1|</td><td>1084-700</td></tr>
<tr><td>Malate dehydrogenase</td><td>gi|39939491|gb|AAR32785.1|</td><td>961-700</td></tr>
<tr><td>Metallothionein-like protein</td><td>gi|4099917|gb|AAD00709.1|</td><td>1888-700</td></tr>
<tr><td>NADP specific isocitrate dehydrogenase</td><td>gi|3811007|dbj|BAA34112.1|</td><td>1097-700</td></tr>
<tr><td>Nonspecific lipid transfer protein</td><td>gi|1076240|pir||S51816</td><td>418-700</td></tr>
<tr><td>OJ000126_13.9</td><td>gi|50924099|ref|XP_472410.1|</td><td>1716-700</td></tr>
<tr><td>ORF107</td><td>gi|7524707|ref|NP_042461.1|</td><td>2115-700</td></tr>
<tr><td>ORF64c</td><td>gi|29469812|gb|AAO74140.1|</td><td>2830-700</td></tr>
<tr><td>ORF68b</td><td>gi|29469783|gb|AAO74111.1|</td><td>1841-800</td></tr>
<tr><td>ORF75</td><td>gi|29469782|gb|AAO74110.1|</td><td>1841-700</td></tr>
<tr><td>OSJNBa0089K21.2</td><td>gi|50924966|ref|XP_472822.1|</td><td>2997-700</td></tr>
<tr><td>Photosystem II protein K</td><td>gi|7524599|ref|NP_042353.1|</td><td>1105-700</td></tr>
<tr><td>Pinocembrin chalcone synthase</td><td>gi|7576362|dbj|BAA94594.1|</td><td>686-700</td></tr>
<tr><td>Polyubiquitin</td><td>gi|1332579|emb|CAA66667.1|</td><td>2779-700</td></tr>
<tr><td>Porin</td><td>gi|2258135|emb|CAB06080.1|</td><td>417-700</td></tr>
<tr><td>Probable 40S ribosomal protein S15</td><td>gi|2982268|gb|AAC32121.1|</td><td>1823-700</td></tr>
<tr><td>Probable 60s ribosomal protein L13a</td><td>gi|2982259|gb|AAC32117.1|</td><td>2938-700</td></tr>
<tr><td>Probable 60S ribosomal protein L31</td><td>gi|2982295|gb|AAC32133.1|</td><td>2256-700</td></tr>
<tr><td>Probable thioredoxin H</td><td>gi|2982247|gb|AAC32111.1|</td><td>2524-700</td></tr>
<tr><td>Profilin</td><td>gi|3694872|gb|AAC62482.1|</td><td>1135-700</td></tr>
<tr><td>Putative 40S ribosomal protein s12</td><td>gi|643074|gb|AAA79921.1|</td><td>2725-800</td></tr>
<tr><td>Putative 40S ribosomal protein S19</td><td>gi|6513924|gb|AAF14828.1|</td><td>2624-700</td></tr>
<tr><td>Putative 60S ribosomal protein L37</td><td>gi|50904689|ref|XP_463833.1|</td><td>1140-700</td></tr>
<tr><td>Putative 60S ribosomal protein L44</td><td>gi|34901296|ref|NP_911994.1|</td><td>1474-700</td></tr>
<tr><td>Putative alpha7 proteasome subunit</td><td>gi|14594925|emb|CAC43323.1|</td><td>577-700</td></tr>
<tr><td>Putative ethylene-responsive transcriptional coactivator</td><td>gi|50944921|ref|XP_481988.1|</td><td>696-700</td></tr>
<tr><td>Putative gamma-thionin protein</td><td>gi|1360108|emb|CAA62761.1|</td><td>607-700</td></tr>
<tr><td>Putative L24 ribosomal protein</td><td>gi|64420807|gb|AAY41426.1|</td><td>789-700</td></tr>
<tr><td>Putative protein [<i>Arabidopsis thaliana</i>]</td><td>gi|6911856|emb|CAB72156.1|</td><td>1636-800</td></tr>
<tr><td>Putative ribosomal protein</td><td>gi|21554085|gb|AAM63166.1|</td><td>1355-700</td></tr>
<tr><td>Putative small nuclear ribonucleoprotein polypeptide F</td><td>gi|50252055|dbj|BAD27986.1|</td><td>2792-700</td></tr>
<tr><td>Putative translation factor</td><td>gi|20218809|emb|CAC84489.1|</td><td>1826-700</td></tr>
<tr><td>Ribosomal protein L11</td><td>gi|21592421|gb|AAM64372.1|</td><td>2840-700</td></tr>
<tr><td>Ribosomal protein L2</td><td>gi|19343|emb|CAA45863.1|</td><td>1435-700</td></tr>
<tr><td>Ribosomal protein L23</td><td>gi|2959593|gb|AAC95507.1|</td><td>2386-800</td></tr>
<tr><td>Ribosomal protein L34</td><td>gi|1076636|pir||S48027</td><td>2423-700</td></tr>
<tr><td>Ribosomal protein L37A</td><td>gi|4090257|emb|CAA10493.1|</td><td>61</td></tr>
<tr><td>Ribosomal protein S14</td><td>gi|29469774|gb|AAO74102.1|</td><td>2445-800</td></tr>
<tr><td>Ribosomal protein S14</td><td>gi|7524712|ref|NP_042466.1|</td><td>1585-800</td></tr>
<tr><td>Ribosomal protein S26</td><td>gi|5706704|gb|AAD47346.1|</td><td>2093-800</td></tr>
<tr><td>Ribosomal protein S27</td><td>gi|8388627|emb|CAB94147.1|</td><td>1324-700</td></tr>
</table>

Ribosomal protein S7	gi|33318768|gb|AAQ05289.1|	181-700
Ribulose-1,5-carboxylase/oxygenase	gi|295822|emb|CAA34161.1|	1264-700
Ring-box protein-like	gi|21592528|gb|AAM64477.1|	1576-700
TCTP-like protein	gi|3850844|emb|CAA10048.1|	1932-700
TFIIA-S	gi|39545876|gb|AAR28001.1|	839-700
Translation initiation factor 5A	gi|14193249|gb|AAK55848.1|	2431-700
Type 1 chlorophyll a /b-binding protein	gi|20788|emb|CAA41404.1|	1934-700
Type III chlorophyll a /b-binding protein	gi|20794|emb|CAA41407.1|	2828-700
Ubiquitin	gi|30523391|gb|AAP31578.1|	390-700
Ubiquitin extension protein	gi|1771780|emb|CAA71132.1|	936-700
Ubiquitin-conjugating enzyme UBC2	gi|5762457|gb|AAD51109.1|	1316-700
Unknown protein [*Arabidopsis thaliana*]	gi|19310765|gb|AAL85113.1|	2491-700
Unknown protein [*Oryza sativa (japonica cultivar-group)*]	gi|55296600|dbj|BAD69198.1|	2781-700
Vacuolar ATPase subunit c	gi|12659320|gb|AAK01292.1|	3025-700
YGL010w-like protein	gi|2982301|gb|AAC32136.1|	2940-700
Zinc finger protein-like	gi|53792608|dbj|BAD53623.1|	2917-700