Structural and functional evolution of genes in conifers

Thèse

Juliana Stival Sena

Doctorat en sciences forestières

Philosophiae doctor (Ph. D.)

Québec, Canada

Structural and functional evolution of genes in conifers

Thèse

Juliana Stival Sena

Sous la direction de :

Jean Bousquet, directeur de recherche

John Mackay, codirecteur de recherche

RÉSUMÉ

Le développement de nouvelles techniques a accéléré l'exploration structurale et fonctionnelle des génomes des conifères et contribué à l'étude de leur physiologie et leur adaptation aux conditions environnementales. Cette thèse s'intéresse à l'évolution des gènes chez les conifères et (i) fait le point sur les facteurs génomiques qui ont influencé la structure des gènes et (ii) analyse une grande famille de gènes impliqués dans la tolérance à la sécheresse, les déhydrines. Notre étude de la structure génique s'est fait à partir de diverses séquences de l'épinette blanche (*Picea glauca* [Moench] Voss) provenant de clones BAC, de l'assemblage du génome et de l'espace génique obtenu à partir de la technologie de «sequence capture». Par le biais d'analyses comparatives, nous avons observé que les conifères présentent plus de séquences introniques par gène que la plupart des plantes à fleurs (angiospermes) et que la longueur moyenne des introns n'était pas directement corrélée à la taille du génome. Nous avons constaté que les éléments répétitifs qui sont responsables de la très grande taille des génomes des conifères affectent également l'évolution des exons et des introns. Dans la deuxième partie de la thèse, nous avons entrepris la première analyse exhaustive de la famille des gènes des déhydrines chez les conifères. Les analyses phylogénétiques ont indiqué l'apparition d'une série de duplications de gènes dont une duplication qui a provoqué l'expansion de la famille génique spécifiquement au sein du genre *Picea*. L'analyse démontre que les déhydrines ont une structure modulaire et présentent chez les conifères des agencements variés de différents motifs d'acides aminés. Ces structures sont particulièrement diverses chez l'épinette et sont associées à différents patrons d'expression en réponse à la sècheresse. Dans l'ensemble, nos résultats suggèrent que l'évolution de la structure génique est dynamique chez les conifères alors que l'évolution des chromosomes est largement reconnue comme étant lente chez ceux-ci. Ils indiquent aussi que l'expansion et la diversification des familles de gènes liés à l'adaptation, comme les déhydrines, pourraient conférer de la plasticité phénotypique permettant de répondre aux changements environnementaux au cours du long cycle de vie qui est typique de plusieurs conifères.

ABSTRACT

Technical advances have accelerated the structural and functional exploration of conifer genomes and opened up new approaches to study their physiology and adaptation to environmental conditions. This thesis focuses on the evolution of conifer genes and explores (i) the genomic factors that have impacted the evolution of gene structure and (ii) the evolution of a large gene family involved in drought tolerance, the dehydrins. The analysis of gene structure was based on white spruce (*Picea glauca* [Moench] Voss) sequence data from BAC clones, the genome assembly and the gene space obtained from sequence capture. Through comparative analyses, we found that conifers presented more intronic sequence per gene than most flowering plants (angiosperms) and that the average intron length was not directly correlated to genome size. We found that repetitive elements, which are responsible for the very large size of conifer genomes, also affect the evolution of exons and introns. In the second part of the thesis, we undertook the first exhaustive analysis of the dehydrin gene family in conifers. The phylogenetic analyses indicated the occurrence of a series of gene duplications in conifers and a major lineage duplication, which caused the expansion of the dehydrin family in the genus *Picea*. Conifer dehydrins have an array of modular amino acid structures, and in spruce, these structures are particularly diverse and are associated with different expression patterns in response to dehydration stress. Taken together, our findings suggest that the evolution of gene structure is dynamic in conifers, which contrast with a widely accepted slow rate of chromosome evolution. They further indicate that the expansion and diversification of adaptation-related genes, like the dehydrins in spruce, may confer the phenotypic plasticity to respond to the environmental changes during their long life span.

TABLE OF CONTENT

RÉSUMÉ ..... **iii**
ABSTRACT ..... **iv**
TABLE OF CONTENT ..... **v**
LIST OF TABLES ..... **viii**
LIST OF FIGURES ..... **x**
LIST OF ABBREVIATIONS ..... **xiii**
ACKNOWLEDGEMENTS ..... **xv**
FOREWORD ..... **xvi**
Chapter 1: General introduction **1**
1.1 Preamble 1
1.2 Genome analysis and evolution 3
1.2.1 Genome sequencing and assembly 4
1.3 Genome evolution 6
1.3.1 Transposable elements 6
1.3.2 Gene content 8
1.3.3 Impacts of whole genome duplication and single duplications 9
1.3.4 Gene structure 11
1.4. Evolution at the functional level 13
1.4.1 Evolution of gene families 13
1.4.2. Gene expression 17
1.5. Project context, research objectives and hypotheses 20
1.5.1 Gene structure evolution in conifers 20
1.5.2 Molecular structure and evolution of dehydrins in *Picea glauca* 21
1.6 References 22

Chapter 2: Evolution of gene structure in the conifer *Picea glauca*: a comparative analysis of the impact of intron size **31**
2.1 Abstract 31
2.2 Résumé 33
2.3 Introduction 35
2.4 Results 37
2.4.1 Genomic sequences 37
2.4.2 Gene expression profiles 38
2.4.3 Gene structures and comparative analysis with angiosperms 40
2.4.4 Comparative analysis of gene structures between *Picea glauca* and *Pinus taeda* 45
2.4.5 Repeat elements in *Picea glauca* genes 47
2.5 Discussion 48
2.5.1 Evolution of gene structure in plants 48
2.5.2 Repetitive sequences in gene evolution 51
2.5.3 Slow evolution of conifer genes 52
2.5.4 Costs and benefits associated with intron size 53
2.6 Conclusions 54

2.7 Methods 55
2.7.1 *Picea glauca* BAC isolation and validation 55
2.7.2 *Pinus taeda* orthologous sequences 57
2.7.3 Screening for highly expressed genes in the whole genome shotgun assembly 58
2.7.4 Identification of closest homologs in angiosperms 58
2.7.5 Statistical analyses of introns 59
2.7.6 Gene space obtained from sequence capture technology 59
2.7.7 *Picea glauca* repetitive library and identification of repeat elements 60
2.8 Acknowledgements 62
2.9 References 62
2.10 Supplementary information 68
2.10.1 Additional experimental procedures for BAC isolation and sequence capture. 68
2.10.2 Supplementary tables 70
2.10.3 Supplementary figures 79

Chapter 3: Expansion of the dehydrin gene family in conifers is associated with considerable structural diversity and drought responsive expression 86
3.1 Abstract 86
3.2 Résumé 88
3.3 Introduction 89
3.4 Materials and Methods 91
3.4.1 Dehydrin sequences 91
3.4.2 Phylogenetic analysis 91
3.4.3 Identification of conserved amino acid motifs and classification of dehydrins 92
3.4.4 Plant material 93
3.4.5 RNA extraction and cDNA synthesis 94
3.4.6 Primer design and quantitative RT-PCR 94
3.4.7 Sequence analysis of amplicons 95
3.5 Results 96
3.5.1 Identification of abundant dehydrins sequences in spruce 96
3.5.2 Dehydrins are highly divergent between angiosperms and conifers 97
3.5.3 A major duplication is uniquely detected in the genus *Picea* 100
3.5.4 Degenerate K-segments and structural variations in conifer dehydrins 100
3.5.5 Dehydrin expression varies between different tissues and conditions 102
3.5.6 Members of the dehydrin family respond differently to water stress 103
3.6 Discussion 107
3.6.1 Structural diversity in dehydrin protein sequences 107
3.6.2 Dehydrin gene family evolution and expansion in spruce 109
3.6.3 Expression of dehydrin genes in developmental and stress responses 111
3.7 Conclusions 114
3.8 Acknowledgements 114
3.9 References 115
3.10 Supplementary information 120
3.10.1 Supplementary figures 120
3.10.1 Supplementary tables 122

Chapter 4: Conclusions 144
4.1 Major findings and conclusions 144
4.1.1 Evolution of gene structure 144
4.1.2 Gene family evolution: a case study of dehydrins 146
4.2 Critical overview and contributions 147
4.3 Perspectives 150
4.3.1 Large scale gene structure evolution and comparative analyses of intergenic regions 150
4.3.2 Dehydrins: multi-function proteins 151
4.3.3 Linking gene structure and gene family evolution. 152
4.4 References 152

LIST OF TABLES

Table 1.1- Characteristics of sequence assemblies published for forest trees. 3

Table 2.1- Average number and length of exons in genes used for comparative analyses . 40

Table 2.2- Abundance of repetitive elements in *P. glauca* genes obtained from sequence capture 48

Table S2.1- Gene structure data of orthologs of *Picea glauca* and *Pinus taeda*. 70

Table S2.2- Genes associated with secondary cell-wall formation or with nitrogen metabolism in *P. glauca* targeted for BAC isolations. 71

Table S2.3- Primer information and sequences used for BAC screening and sequencing validation 72

Table S2.4- Accession numbers of *P. taeda* orthologs and sequence similarity to *P. glauca*. 73

Table S2.5- Accession numbers for the closest homologous sequences between *P. glauca*, *Arabidopsis thaliana, Populus trichocarpa* and *Zea mays*..... 74

Table S2.6- Summary of sequencing results of *P. glauca* BAC clones isolated each containing a different single copy gene associated with cell- wall formation or with nitrogen metabolism. 76

Table S2.7- GenBank accessions of complete cDNA utilized for gene structure definition when the cDNA in *Picea glauca* gene catalogue was incomplete. 77

Table S2.8- Repetitive elements detected within gene structure of the 35 *P. glauca* genes[1]. 78

Table S3.1- Gene specific primers utilized to determine RNA transcript levels from drought stress and tissue comparison experiments by using quantitative RT-PCR. 122

Table S3.2- A total of 144 conifer dehydrins were clustered on the basis of at least 97% of sequence similarity, 78 clusters were formed using CD-hit. The sequences indicated by asterisks were used as the representative sequence of the cluster. 123

Table S3.3- A total of 76 angiosperm dehydrins were clustered by sequence similarity (97%), 57 clusters were formed using CD-hit. Sequences indicated by an asterisk were used as representative sequence of the cluster. 131

Table S3.4- Classification of angiosperm and conifer dehydrins based on their conserved amino-acid segments (segment-K, A, E, S, Y and N1). The graphical representation of all possible classifications (models) is in Fig.3. Sequences indicated by //, *, ** presented one degenerate A, K or Y segment, respectively. 135

Table S3.5- The one-way ANOVA tested if the expression levels between the three tissues (phelloderm, xylem and young foliage) were different. 140

Table S3.6- A three-way ANOVA with water potential as a function of type of treatment (watering regimes), genotype, sampling dates and their interaction. 140

Table S3.7- A three-way ANOVA with expression as a function of type of treatment (watering regimes), genotype, sampling dates and their interaction. 141

LIST OF FIGURES

Figure 1.1- A. White spruce is a large conifer forest tree. B. It is widely produced in Canada both in natural forests (as seen here) and plantations and is used in multipurpose manufacture of wood products including boards as well as pulp and paper. C. White spruce is also used in Christmas tree production. 2

Figure 1.2- White spruce is native of North America. It has a transcontinental distribution covering all of Canada and several northern American states. 2

Figure 2.1- Transcript accumulation profiles from the PiceaGenExpress database (Raherison et al. [30]**)** of the *P. glauca* genes. The transcript abundance data are classified from 1 to 10, from lowest to highest microarray hybridization intensities detected within a given tissue. The profiles of highly expressed genes (top) (according to Raherison et al. [30]; class 8 to 10 are contrasted with most of the genes associated with secondary cell wall formation and nitrogen metabolism (bottom, names in bold). NA: Not detected. Tissues: B (Vegetative buds), F (Foliage), X-M (Xylem – from mature trees), X-J (Xylem –juvenile trees), P (Phelloderm), R (Adventitious roots), M (Megagametophytes), E (Embryogenic cells) 39

Figure 2.2- Comparative analysis of individual intron length in *P. glauca*, *A. thaliana P. trichocarpa* and *Z. mays*. Box plots represent intron length data for all of the introns of the 35 genes used in comparative analyses. Intron lengths were compared among the four species by Kruskal-Wallis test with post-test analysis by Dunn's multiple comparisons: NS, not significant ($P \geq 0.06$); *P = 0.06; **$P < 0.01$; ***$P < 0.001$. 41

Figure 2.3- Comparative analysis of total intron length in *P. glauca, A. thaliana, P. trichocarpa* and *Z. mays*. Average ratio of total length of intron sequences in pair-wise comparisons in: A- all genes; B- highly expressed genes and genes involved in secondary cell-wall formation and nitrogen metabolism (For individual ratios, see Figure 2.4). The total intron lengths were compared among the four species by Kruskal-Wallis test with post-test analysis by Dunn's multiple comparisons: NS, not significant ($P \geq 0.05$); **$P <$ 0.01; ***$P < 0.001$. 43

Figure 2.4- Gene by gene pair-wise comparisons of total length of intronic sequences in *P. glauca, A. thaliana, Populus trichocarpa* and *Z. mays.* (A) highly expressed genes and (B) genes associated with secondary cell-wall formation and nitrogen metabolism. 44

Figure 2.5- Gene structures of six genes from different angiosperm and gymnosperm species. The first three genes are associated with secondary cell-wall formation and nitrogen metabolism; and highly expressed genes are bolded. 45

Figure 2.6 Relationship between intron size and sequence similarity of introns from *P. glauca* and *P. taeda*. A total of 138 introns were obtained from 22 genes and sequence alignments were produced with the Needle software (see Methods). 46

Figure 2.7- Variation in intron length and genome size in 35 target genes. Average intron size for *Arabidopis*, *P. trichocarpa*, *Z. mays* and *P. glauca* determined from the analysis of 35 homologous genes. Note that Y- axes are in log 10 scale. 50

Figure S2.1- Content of repetitive elements in 21 different BAC clones. The analysis used the RepeatMasker software and a *P. glauca* repetitive sequence library (see Methods). Repetitive elements were classified as LTR (long terminal repeat) and unclassified (no hit in RepBase)..... 79

Figure S2.2- Comparative analysis of individual intron length in *P. glauca, A. thaliana, P. trichocarpa* and *Z. mays*. A. Average and median length of individual introns in all genes. B Average and median length of individual introns in highly expressed genes and genes associated with secondary cell-wall formation and nitrogen metabolism in four species. Intron lengths were compared among the four species by Kruskal-Wallis test with post-test analysis by Dunn's multiple comparisons: NS, not significant (P > 0.06); * P < 0.06; **P < 0.01; ***P < 0.001..... 80

Figure S2.3- Boxplot of the 35 homologous genes in *P. glauca*, *A. thaliana*, *P. trichocarpa* and *Z. mays.* 81

Figure 3.1- Phylogeny of the conifer dehydrin gene family represented by a consensus tree from Bayesian analysis, with threshold support equal or superior to 0.75. We used 41 white spruce dehydrins and 37 other conifer sequences; see details of sequence clusters in Table S3.2. A dehydrin from *Physcomitrella* was used as the root. The phylogeny was obtained with MrBayes after protein alignment with MAFFT, and visualized with FigTree..... 98

Figure 3.2- A) Phylogeny of the angiosperm and conifer dehydrin gene family represented by a consensus tree from Bayesian analysis, with threshold support equal or superior to 0.75. A dehydrin from the moss *Physcomitrella* was used to root the tree. The phylogeny was obtained with MrBayes after protein alignment with MAFFT. B) Synthesis of speciation and gene duplication events with N1 Skn type as ancestor; D - duplication, S - speciation. 99

Figure 3.3- Conifer and angiosperm dehydrins classification based on their amino-acid motifs. A) Sequences were grouped by similarity and classified by motif composition. B) Each dehydrin type was represented showing the variation in number of motifs. 101

Figure 3.4- Transcript accumulation profiles from F (Foliage), X (Xylem) and P (Phelloderm) measured by qPCR. Significant differences between tissue expression levels are indicated on the right side, ANOVA, Tukey's HSD (*P* <0.05; ns indicates no significant difference between the expression level among the three tissues). 103

Figure 3.5- Midday water potential in needles of well-watered plants (dashed line) and unwatered plants (solid line) in three different genotypes (clones 8, 11 and 95). The water potential of water-stressed plants was compared with that of control plants for each sampling date in all three genotypes (ANOVA, Tukey's HSD, *** *P* <0.001)..... 104

Figure 3.6- Expression profile of dehydrin genes during 22 days of treatment. The gene expression of water-stressed plants (solid lines) was compared with that of control plants (dashed lines) for each sampling date in all three genotypes (clones 8, 11 and 95). ANOVA, Tukey test, *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$. 106

Figure S3.1- Logo of motifs discovered in angiosperms and conifers by MEME. 120

Figure S3.2- Phylogeny of the angiosperm dehydrin gene family represented by a consensus tree from Bayesian analysis, threshold support equal or superior to 0.75. We used 57 angiosperm dehydrins; see details of sequence clusters in Table S3.2. A dehydrin from *Physcomitrella* was used as the root. The phylogeny was created with MrBayes after protein alignment with MAFFT. 121

LIST OF ABBREVIATIONS

BAC; paired-end bacterial artificial chromosome

cDNA; complementary DNA

CDS; coding sequence

DNA; deoxyribonucleic acid

EST; expressed sequence tag

FL-cDNA; full-length-cDNA

LEA; late embryogenesis abundant

LTR-RT; long terminal repeat retrotransposons

NGS; next generation sequencing

Nt; nucleotide

ORF; open reading frame

PCR; polymerase chain reaction

qPCR; quantitative real-time PCR

RNA-seq; ribonucleic acid –sequencing

RT-qPCR; reverse transcription–qPCR

SSR; simple sequence repeat

TEs; transposable elements

UTR; untranslated transcribed region

WGD; whole genome duplications

Aos meus pais

ACKNOWLEDGEMENTS

I am so happy to get here, to complete this step. When I look back, I realize how much the new country, the new language, the new friends, and all the effort invested along this academic journey have had a great influence on me. I am so grateful for what I have learned these last years and I keep looking forward for new challenges with the same excitement of the beginning.

I would like to thank John Mackay and Jean Bousquet for the patience and commitment during the realization of this thesis. Thank you for all the advices, support and encouragement.

My sincere thanks to François Belzile, Armand Séguin and Maria Teresa Cervera for accepting to evaluate this thesis.

Thank you to all the colleagues and professionals of the Mackay's lab group for the critiques, comments and friendship. Especially Isabelle Giguère and Sébastien Caron for the guidance in the laboratory work. Thank you to Jérôme Laroche and Brian Boyle for the advices in bioinformatics.

Thank you to all my multicultural friends and family. Merci d'être là, d'apporter de la joie et pour vos mots d'encouragement. Obrigada aos que estão longe, mas que permanecem próximos, é sempre bom receber o apoio e carinho de vocês. Obrigada Brice por ser o meu braço direito (ou esquerdo).

Thank you to my parents who have always encouraged me in my projects. Graças à vocês eu continuo a realizer meus sonhos.

FOREWORD

This thesis includes two articles, one peer-reviewed and published one (Chapter 2) and another one to be submitted for review (Chapter 3). Part of the general introduction of this thesis (Chapter 1) was published in a book chapter, in which J. Stival Sena and J. Mackay were among the authors. The sections 1.2 and 1.3 of Chapter 1 were written by J. Stival Sena for the book chapter published in 2015 and were updated for their inclusion in the general introduction of this thesis.

Part of the sections 1.2 and 1.3 in Chapter 1 were published in:

Parent GJ, Raherison E, Sena J, MacKay JJ. Forest tree genomics: review of progress. Advances in Botanical Research. Elsevier; 2015. p. 39–92.

Chapter 2 is based on the published peer-reviewed article:

Stival Sena J, Giguère I, Boyle B, Rigault P, Birol I, Zuccolo A, et al. Evolution of gene structure in the conifer *Picea glauca*: A comparative analysis of the impact of intron size. BMC Plant Biology. 2014; 14:95.

Chapter 3 of this thesis represents a manuscript to be submitted as:

Stival Sena J, Giguère I, Rigault P, Bousquet J, Mackay J. Expansion of the dehydrin gene family in conifers is associated with considerable structural diversity and drought responsive expression.

In both studies, J. Stival Sena planned and executed the experiments, conducted lab manipulations, the bioinformatics and statistical analyses, interpretation of results, and drafted the manuscripts with the supervision of J. Mackay and J. Bousquet.

Chapter 1: General introduction

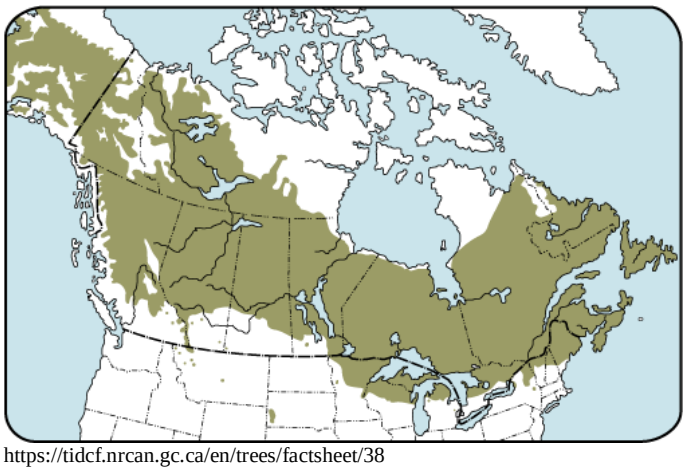
1.1 Preamble

Boreal forests cover approximately 30% of the global forest area and provide major ecosystem services but their health is being negatively impacted by climate change (Aitken et al., 2008; Alberto et al., 2013). It has been proposed that mitigating the effects of droughts and higher temperatures on boreal forests requires a better understanding of biological processes and a re-thinking of forest management practices (Gauthier et al., 2015). In the last few years, developments in genomics have contributed new knowledge on the function and evolution of trees that is expected to help adopt new practices. Understanding how organisms evolve and adapt to environmental change is a complex task that involves a wide range of investigations including structural and functional analyses which are being enriched by genome studies at the individual and population levels (Neale and Ingvarsson, 2008; Soltis and Soltis, 2013).

This thesis reports on the evolution of gene structure and gene families in conifers with a primary focus on white spruce (*Picea glauca* (Moench) Voss) (Fig.1.1), a member of the family Pinaceae. Conifer trees present a set of genome features that make them unique (MacKay et al., 2012). Most notorious among these features is their uniformly large genome size, varying from 18 to 35 Gpb in size (Murray et al., 2012). They also contain a large fraction of repetitive sequences, have a gene space representing less than 1% of their genome, a unique evolutionary trajectory in many gene families compared to angiosperms, and low mutation rates (reviewed in Parent et al., 2015), which are thought to have contributed to their adaptation to diverse environments (Fig.1.2). One example of evolution that is relevant for adaptation to climate change is that some gene families involved in abiotic stress responses, including dehydrin encoding genes, have expanded in white spruce compared to angiosperm plants (Rigault et al., 2011).

This chapter presents recent developments in genome analysis and evolution in forest trees, with a particular emphasis on genes – the functional unit of heredity. Genome evolution invariably results in changes in gene content, which in turn may produce functional changes and are accompanied by changes in gene expression. These various aspects are explored and developments are compared between conifers, which belong to the gymnosperms, and hardwood trees, which belong to the angiosperms. Research hypotheses and objectives are described at the end of the chapter.

Figure 1.1- A. White spruce is a large conifer forest tree. B. It is widely produced in Canada both in natural forests (as seen here) and plantations and is used in multipurpose manufacture of wood products including boards as well as pulp and paper. C. White spruce is also used in Christmas tree production.



https://tidcf.nrcan.gc.ca/en/trees/factsheet/38

Figure 1.2- White spruce is native of North America. It has a transcontinental distribution covering all of Canada and several northern American states.

1.2 Genome analysis and evolution

Forest tree genome sequencing has accelerated significantly very recently. With the development of Next Generation Sequencing (NGS) technologies, most forest tree genomes have been reported recently since 2013. To date, published forest tree genomes span both hardwood and conifer trees distributed among several genera including *Populus* (Tuskan et al., 2006), *Salix* (Dai et al., 2014), *Eucalyptus* (Myburg et al., 2014), *Betula* (Wang et al., 2013), *Fraxinus* (http://www.ashgenome.org), *Castanea* (http://www.hardwoodgenomics.org/chinese-chestnut-genome), *Quercus* (Plomion et al,. 2016), *Picea* (Birol et al., 2013; Nystedt et al., 2013; Warren et al., 2015) and *Pinus* (Neale et al., 2014; Stevens et al., 2016) (See table 1.1). In this section, we focus on the most fully characterized hardwood genomes which are *Populus* and *Eucalyptus* and, on recently available conifer genomes.

Table 1.1- Characteristics of sequence assemblies published for forest trees**.**

Specie	Genome assembly	No. of scaffolds	Contig N50 (Kbp)	Scaffold N50
Populus trichocarpa [a]	422.9 Mbp	1.4 K	552.8	19.5 Mb
Eucalyptus grandis [b]	691 Mbp	4.9 K	67.2	57.5 Mb
Salix suchowensis [c]	303.8 /425 Mbp	103.1 K	17.4	925 Kbp
Betula nana [d]	450 Mbp	551.9 K	5.1	18.7 Kbp
Quercus robur [e]	1.5 Gbp	17.9 K	NA	260 Kbp
Picea glauca [f] *	22.4 Gbp	4.3 M	6.8	19.9 Kbp
Picea abies [g]	19.6 Gbp	10.2 M	0.6	0.72 Kbp
Pinus taeda [g]	22 Gbp	14.4 M	8.2	66.9 Kbp

NA : information not available

* : No. of scaffolds ≥ 500 bp

[a] release of Phytozome assembly v3, Tuskan et al., 2006;

[b] release of Phytozome assembly v2, Myburg et al., 2014;

[c] Dai et al., 2014; [d] Wang et al., 2013; [e] Plomion et al., 2016; [f] Warren et al., 2015; [g] Zimin et al., 2014

1.2.1 Genome sequencing and assembly

Hardwoods – *Populus and Eucalyptus*

The first forest trees genome sequenced was that of a *P. trichocarpa* female tree (Nisqually 1). It was obtained by using a hybrid strategy that combined whole genome shotgun sequencing, construction of a physical map based on bacterial artificial chromosome (BAC) restriction fragment fingerprints, BAC-end sequencing, and extensive genetic mapping based on simple sequence repeat (SSR) length polymorphisms that allowed chromosome reconstruction with the assembled genome (Tuskan et al., 2006).

The improved version of the *Populus* genome assembly (v 3) has approximately 422.9 Mb. It includes 81 Mb of finished clone sequences combined with a high density physical map, which resulted in a genome assembly of 422.9 Mb. This assembly can be accessed in the JGI comparative plant genomics portal at: http://phytozome.jgi.doe.gov.

For *Eucalyptus grandis*, the genome assembly was based on whole-genome Sanger shotgun sequencing, paired-end bacterial artificial chromosome sequencing and a high-density genetic linkage mapping, resulting in the first non-redundant chromosome-scale reference (V1.0) sequence (Myburg et al., 2014). A recent comparison between new high-resolution genetic maps for *E. grandis* and *E. urophylla* (Bartholomé et al., 2015) with the reference genome highlighted non collinear and non syntenic regions. These regions were corrected in the latest version (V2.0) which is available on Phytozome 11 (https://phytozome.jgi.doe.gov/pz/portal.html). The *E. grandis* assembly (V2.0) is approximately 691 Mb arranged in 4,943 scaffolds.

Conifers

Genome sequences were recently reported for Norway spruce (Nystedt et al., 2013), white spruce (Birol et al., 2013) and loblolly pine (Neale et al., 2014). In addition, assemblies were released for sugar pine (Stevens et al., 2016) and Douglas-fir (http://pinegenome.org/pinerefseq/) and, reduced depth sequencing was reported for six other species (Nystedt et al., 2013). These developments are driven by progress in shotgun

genome sequencing and associated bioinformatics methods (Simpson et al., 2009; Birol et al., 2013; Nystedt et al., 2013; Zimin et al., 2013) which have been applied to analyzing both haploid (Norway spruce and loblolly pine) and diploid conifer DNA. Different strategies were explored to assemble the genomes into contigs and scaffolds. Despite these efforts, assemblies reported to date remain highly fragmented, presenting millions of unordered scaffolds.

The conifer sequences and assemblies are shedding new light into conifer genome evolution (Soltis & Soltis, 2013; De la Torre et al., 2014); however, the very large size and the highly repetitive content of conifer genomes continue to represent a challenge for achieving contiguous assemblies (Warren et al. 2015). We may also expect that pseudogenes, which appear to be abundant in conifer genomes (Bautista et al., 2007; Kovach et al., 2010; Magbanua et al., 2011; Nystedt et al. 2013), will complicate further analyses and finishing of assemblies. To overcome the fragmentation challenge, the conifer sequencing projects have developed bioinformatics tools and make use of sequenced fosmids (Nystedt et al., 2013), BAC clones, RNA-Seq data (Warren et al., 2015) and linking libraries. In parallel to the improvement of reference sequences, another long term objective is to place the scaffolds onto chromosomes, by anchoring the genome assembly to a genetic map (De la Torre et al., 2014).

In contrast to conifers, genome size was not the most important challenge in the genome assembly of *Populus* and *Eucalyptus,* but these projects also faced the challenges of assembling repeat sequences represented by transposable elements and duplicated regions originated from whole genome and/or tandem duplications (Tuskan et al., 2006; Myburg et al., 2014). These projects have improved the quality of their genome assembly by increasing contiguity and minimizing haplotypes within the assembly, besides improving linkage maps and scaffold anchoring (https://phytozome.jgi.doe.gov/pz/portal.html).

All of the genomic resources available for forest trees have permitted a better understanding of genome organization, composition and functionality and, through

comparative analyses with other sequenced plant genomes; insights have been gained into forest tree genome evolution.

1.3 Genome evolution

Given the very large difference in genome sizes, it is not surprising that genome structure and evolution differ greatly between *Eucalyptus* and *Populus* on the one hand, and conifers on the other. The conifers stand out as having the largest average genome sizes among plant orders, which have been estimated between 18 to over 35 Gbp for major conifers (Murray et al., 2012). In contrast, the genomes of *Populus* (450 Mbp) and *Eucalyptus* (640 Mbp) are much more compact. For example, at 20 Gbp, the *Picea glauca* genome is 31 and 44 times larger than the *Populus* and *Eucalyptus* genomes, respectively. It is well known that among angiosperms large genomes are the consequence of multiple genome duplications and polyploidization events with intense periods of transposable element activity and multiplication (Bennetzen 2002). In conifer genomes analyzed to date, retrotransposons are abundant and widespread (Nystedt et al., 2013; Neale et al., 2014, Wegrzyn et al., 2014) and recent evidence indicates that whole genome duplications (WGD) were also involved in the dynamics of conifer genome evolution (Li et al., 2015).

1.3.1 Transposable elements

Transposable elements (TEs) are widespread in plant genomes, exceptionally abundant in species with large genomes, and they play a major role in their evolution. In conifer trees, transposable elements can represent a large portion of the genomes, estimated at 69% in *Picea abies* (Nystedt et al., 2013) and up to 80% in *Pinus taeda* (Wegrzyn et al., 2014). Class I TEs (retrotransposons) are by far the most abundant and are primarily represented by long terminal repeat retrotransposons (LTR-RT). The LTR-RT sequences were estimated to represent 58% of the genome both in *Picea abies* and in *Pinus taeda* (Nystedt et al., 2013; Neale et al., 2014, Wegrzyn et al., 2014). Only three families, the *Ty3/Gypsy*, *Ty1/Copia* and *Gymny* superfamilies make up the bulk of LTR-RTs in conifers as shown by recent genome annotations (Morse et al., 2009; Nystedt et al., 2013; Neale et al., 2014,

Wegrzyn et al., 2014) and BAC sequencing (Kovach et al., 2010; Magbanua et al., 2011; Stival Sena et al., 2014).

Hardwood tree genomes also comprise significant but variable TE content. As in many plant species, retrotransposons account for a major portion of the *Eucalyptus* genome (44.5%), with LTR-RT sequences being the most abundant (21.9%) (Myburg et al., 2014).

The DNA transposons (class II TEs) represent only 5.6% of the genome and *Helitron* elements were found to be the most abundant with an estimated 15,000 copies (3.8 % the genome) (Myburg et al., 2014). *Populus trichocarpa* has approximately 40% of repetitive elements; however, a small fraction seems to be transposable elements as described in RepPop (Zhou and Xu 2009). The most abundant classes of transposable elements are LTR *Gypsy* and *Copia* (Douglas and DiFazio 2010).

Transposable elements have variable roles in the evolution of tree genomes. In *Populus*, it was suggested that very few TEs are transcriptionally active. Their estimated insertion date indicated that *Copia* and *Gypsy* elements have both been active after the separation of the different poplar sections but with different time courses (Cossu et al., 2012). A comparison of *E. globulus* (530 Mbp) and *E. grandis* (640 Mbp) indicated that recent TE activity only accounts for 2 Mbp of the genome size difference and that a very large number of small non-active transposable elements account for most of the difference. A parallel may be drawn to comparisons between the congeneric *A. thaliana* (125 Mbp) and *A. lyrata* (~200 Mbp) genomes but in the case of Arabidopsis most of the difference in genome size could be accounted for by hundreds of thousands of small deletions, mostly in noncoding DNA (Hu et al., 2011).

By comparison, conifers present a very different evolutionary history. The accumulation of TEs in conifers is ancient and has occurred over a very long timeframe spanning tens to hundreds of millions of years (Nystedt et al., 2013). The lack of removal of replicated LTR-RTs appears to be responsible for their massive accumulation rather than a higher rate of multiplication (Morgante and De Poali, 2011; Nystedt et al., 2013).

1.3.2 Gene content

Gene content, i.e. the number of predicted genes, was estimated to be in the same range for *Populus* and *Eucalyptus*, but could be slightly higher in conifers. In *Populus*, Tuskan et al. (2006) identified a first draft reference set of 45,555 protein-coding gene loci in the nuclear genome using a variety of *ab initio*, homology-based and expressed sequence tag (EST). Since then, the gene models have been improved by using RNA-seq transcript assemblies. Phytozome v10.1 (http://phytozome.jgi.doe.gov) contains 41,335 loci containing protein-coding transcripts for poplar. In *Eucalyptus grandis*, 36,349 protein-coding transcripts were predicted based on EST and cDNA data. The gene models are also available in Phytozome v10.1 (http://phytozome.jgi.doe.gov).

Gene content estimates range from 50,174 in loblolly pine (Wegrzyn et al., 2014) to 70,968 in Norway spruce (Nystedt et al., 2013), but in the latter case, only about one third of the gene models were reported as high confidence, i.e. supported by expressed sequences. The white spruce gene catalog which is based on transcribed sequences predicted approximately 33,000 genes, fewer than estimated in loblolly pine and Norway spruce (Rigault et al., 2011; Warren et al., 2015). Conifer genome annotations have revealed a surprisingly large fraction of sequences classified as genes or gene-like fragments. Gene-like sequences represented 2.4% and 2.9% of the *P. abies* and *P. taeda* genome, respectively, (Nystedt et al., 2013; Neale et al., 2014) and as high as 4% from earlier analyses (Morgante and De Paoli, 2011). This is far larger than would be expected for the number of predicted genes. This discrepancy may be explained by the abundance of pseudogenes reported in conifers (Bautista et al., 2007; Kovach et al., 2010; Magbanua et al., 2011) and truncated and misassembled genes as reported in different gene families by Warren et al. (2015).

One factor that may explain the difference in gene number between *Populus*, *Eucalyptus* and conifer species is their different polyploidization histories. Other factors which may have had an influence are tandem duplication frequency and the evolutionary forces that influence the fate of duplicated copies.

1.3.3 Impacts of whole genome duplication and single duplications

Hardwoods

Single gene and whole-genome duplications (WGD) have played a major role in the evolution of angiosperm plants. The genome sequences of *Populus* and *Eucalyptus* provided evidence of two whole-genome duplications, an ancient paleohexaploidy event shared with many dicotyledonous plants, and a more recent and lineage-specific WGD. The recent WGD detected in *Populus* was specific of the *Salicaceae* family and occurred 60-65 Myr ago (Tuskan et al., 2006) whereas, in *Eucalyptus*, the lineage-specific WGD occurred about 106-114 Myr ago. Interestingly, the *Eucalyptus* WGD is older than those detected in other rosids and could have played an important role in the origin of Myrtales, since it is estimated to have occurred around the same time as the origin of this plant order (Myburg et al., 2014).

Over the course of evolution, duplicated gene copies resulting from WGD events may be retained, as indicated by the 8,000 pairs of duplicated genes in *Populus*. Duplicated genes may retain the same set of functions as the ancestral copy (Davis and Petrov, 2004), retain only a subset of the original set of functions (subfunctionalization) (Lynch and Force, 2000; Guillet-Claude et al., 2004), acquire a new function (neofunctionalization), or degrade into a nonfunctional gene (nonfunctionalization) (Ohno, 1970). Rodgers-Melnick et al., (2012) used microarray expression analyses of a diverse set of tissues in *Populus* and functional annotation to evaluate the factors that are associated with the retention of duplicate genes. They hypothesized that duplicate gene retention from WGD in *Populus* is driven by a combination of subfunctionalization of duplicate pairs and purifying selection favoring retention of genes encoding proteins with large numbers of interactions, as proposed by the gene balance hypothesis. This hypothesis posits that genes encoding components of multi-subunit complexes are more likely to evolve in concert because the dosage change in the quantities of subunits affects the interaction and function of the whole complex (Birchler et al., 2007).

Gene loss in *Populus* after the salicoid genome duplication has been less extensive than following the previous whole genome duplication (c. 120 Myr), suggesting that the *Populus* genome reorganization is a dynamic process in progress. In contrast to *Populus*, most of the *Eucalyptus* duplicates have been lost after the most recent WGD. The extensive loss of duplicates in *Eucalyptus* has been shown by a pairwise comparison of syntenic segments with *Vitis*, which was selected for comparison as outgroup because it is a basal rosid lineage that is a paleohexaploid and without evidence of more recent WGD events, as were detected in *Populus* and *Eucalyptus* (Jaillon et al. 2007).

In contrast to genes encoding proteins with large numbers of interactions, genes with poorly connected products in a network would have an elevated probability of retention following tandem duplication (Ren et al., 2014). A study of the gene family of Class III peroxidases (PRX) in *Populus* identified other mechanisms that play a role on gene retention such as protein subcellular relocalization associated with a new function. Class III PRX are involved in stress responses in plants but some PRX duplicates have been recruited to cell wall metabolism, including lignin polymerization, or to the vacuole as part of defense responses to abiotic and biotic stresses (Ren et al., 2014). Although the *Eucalyptus grandis* genome has lost many paralogous genes that appeared following the recent WGD, it has retained genes in tandem duplications (34% of the total genes) at a much higher frequency than observed in the *Populus* genome (Tuskan et al., 2006, Myburg et al., 2014). Some of the expanded gene families are related to lignocellulosic biomass production, secondary metabolites and oils (e.g. phenylpropanoid biosynthesis, terpene synthase and phenylpropanoid gene families). It was proposed that tandem duplication has a significant role in shaping functional diversity in *Eucalyptus* (Myburg et al., 2014).

Conifers

Polyploidy has had a large impact in the evolution of angiosperm genomes but in gymnosperms, only a few natural polyploids have been described, among them two conifer species: alerce (*Fitzroya cupressoides*) and coastal redwood (*Sequoia sempervirens*) (Ahuja, 2005).

Although just two conifer species have been identified as polyploids, it was suggested that an older WGD would have occurred before the split of angiosperms and gymnosperms, being restricted to seed plants and not shared with ferns and relatives (Li et al., 2015b). These authors also reported that two independent WGDs occurred in the ancestor of the Pinaceae and that of cupressophyte conifers (Cupressaceae and Taxaceae), about 200 to 342 million years ago and 210 to 275 million years ago, respectively (Li et al., 2015b).

Nystedt et al. (2013) proposed a model for conifer genome evolution with no WGDs and low chromosomal rearrangements in which the 12 ancestral Pinaceae chromosomes expanded at slow rate due the activity of transposable elements. It is well established that transposable elements had an impact in genome size evolution but findings from other authors suggest that the chromosome number is likely to have varied rather than remained static over time since the inception of conifers. Based on very recent work, a hypothesis was supported where more substantial chromosome rearrangements have occurred between conifer families. Within the Pinaceae family, comparative mapping revealed high levels of synteny and collinearity (Pavy et al., 2012; Miguel et al., 2015), suggesting a lack of WGD. However, a comparison of genome structure between Pinaceae and Cupressaceae suggests rearrangements of ancestral conserved blocks with conservation of gene order in the interior of the blocks (de Miguel et al., 2015). The two independent WGDs in Pinaceae and cupressophytes reported by Li et al. (2015) may explain these rearrangements. Additional comparative genomic analyses among different conifer species will be necessary to delineate the impact of these WGDs more fully on the chromosomal organization in different conifer families.

1.3.4 Gene structure

In eukaryotic genes, amino-acid coding sequences are often interrupted by intervening non-coding sequences called introns, which may vary in size and number, and which are removed after transcription. The intron-exon structure varies considerably between genes and its evolution seems to be affected by several factors such as genome size, recombination rate, expression level and even effective population size (Lynch and

Connery, 2003; Deutsch and Long, 1999; Comeron and Kreiman, 2000; Castillo-Davis et al., 2002). As a consequence, the gene structure of homologous genes may vary between different species as a result of their genome architectures and evolutionary histories.

Exon size seems to be more conserved than intron size in forest trees. For example, similar exon lengths have been reported when comparing homologous genes and their exons between *Picea glauca* and *Populus trichocarpa* (Stival Sena et al., 2014) and *Eucalyptus grandis* (Myburg et al., 2014). In contrast, intron lengths are more variable among these species. Conifer genes tend to accumulate long introns with the largest introns surpassing 60 kb in spruce (Nystedt et al., 2013) and 120 kb in pine (Wegrzyn et al., 2014). On average, the *P. abies* introns are 1000 bp in length, *Populus* 380 bp and, *Eucalyptus* approximately 425 bp (Tuskan et al., 2006; Nystedt et al., 2013; Myburg et al., 2014). The intron average length is higher in conifer genes, which typically accumulate one or a few very long introns although the majority of introns are in the 100 – 200 bp range and are comparable in size to those found in angiosperms (Stival Sena et al., 2014).

Comparative analyses have shown conserved gene structure among conifers. Selected orthologous genes between *P. glauca* and *P. taeda* clearly showed the conservation of gene structure and the distribution of intron sizes despite a divergence time of 100 to 140 Myr (Stival Sena et al., 2014). The conservation of long introns was also observed across gymnosperm taxa, where a group of long introns in *P. abies* was identified as orthologous to long introns in *P. sylvestris* and *Gnetum gnemon* (Nystedt et al., 2013). These observations pointed to the slow evolution of conifer gene structure and suggest that the long introns observed in conifers likely date back to the divergence of major conifer groups.

1.4. Evolution at the functional level

1.4.1 Evolution of gene families

A gene family is a group of genes that have descended from a common ancestral gene. The size of a gene family is variable across different species and may have an impact on adaptation and speciation. The main factors that affect gene family size are gene duplication and gene deletion.

Gene duplications can occur as consequence of WGDs or at a smaller scale by segmental duplications or tandem duplications. Unequal crossing over and transposable element activity are the most common molecular mechanisms causing segmental duplication. The duplicated genes can be preserved by neo-functionalization or sub-functionalization, which means to evolve a new beneficial function or to partition an ancestral function among duplicated copies (Lynch and Force, 2000; Guillet-Claude et al., 2004). Another possible scenario is that the duplicate gene becomes non-functional (Lynch and Force, 2000). Gene losses can be the consequence of an abrupt mutational event with physical removal of the gene from the organism's genome or the consequence of a slow accumulation of mutations with loss of function (Albalat and Ca ñestro, 2016). The process of expansion or contraction of gene number is influenced by the molecular processes involved in gene duplication and also the population genetic forces that impact the dynamics of newly arisen genes (Lynch and Force, 2000).

Many gene families are shared among plants as a "core" proteome (Guo, 2013). Comparative analyses and phylogenetic studies of gene families have shown different evolutionary trajectories between conifers and angiosperms. These different trajectories can lead to expansion or contraction of gene families, facilitating functional specializations and structure variation in protein sequences. In this section, we examine the different evolutionary trajectories of a few gene families in angiosperms and conifers, highlighting the variation in the number of gene family members and possible variations in protein structure.

The expansion and contraction of many gene families has contributed to the evolution and biology of plants. A few gene families involved in lignocellulosic biomass production, secondary metabolites, reproduction and response to biotic/abiotic stresses have been investigated because of their relevance to the growth habit and longevity of many forest tree species. They include transcription factors such as R2R3-MYB NAC and knox-1, terpene synthase genes, late embryogenesis abundant family (LEA) gene families, which we discuss here. Other gene families that have been discussed elsewhere include enzymes involved in the biosynthesis of lignin precursors, peroxidases and laccases, auxin transporters (Aux/IAA), GRAS family of transcription factors, among others (Tuskan et al., 2006; Myburg et al., 2014; Li et al., 2015a).

Phylogenetic analyses have been useful to compare the different evolutionary trajectories of gene families among various species. Let us take as examples the NAC and MYB gene families that are part of the network of transcription factors regulating secondary cell wall biosynthesis. *Eucalyptus grandis* has the largest NAC domain family (189 members) (Hussey et al., 2015) when compared to *Arabidopsis thaliana* (117 members) (Naruzzaman et al., 2010), *Populus trichocarpa* (163 members) (Hu et al., 2010), *Vitis vinifera* (74 members) (Wang et al., 2013), *Oryza sativa* (151 members) (Naruzzaman et al., 2010) and *Picea glauca* (36 members) (Duval et al., 2014). The NAC expansion in *Eucalyptus* is mainly explained by tandem duplications in several clades (Hussey et al., 2015). Most NAC subfamilies are represented in angiosperms with apparently no specific lineages, with the exception of one subfamily in the *Solanaceae* (Singh et al., 2013). The NAC genes in conifers are not distributed among all clades (Pascual et al., 2015).

This gene family is defined by a NAC DNA-binding domain in the N-terminal region. The protein domains are compact and conserved regions of a protein that often are related to a distinct function (Bagowski, Bruins and Velthuis, 2010). In the case of NAC, it has been associated with DNA binding and the formation of homo- or heterodimers with other NAC domains (Olsen et al., 2005). Beyond domain functional classification, conserved motifs have been identified in many protein families. NAC genes present a modular structure with conserved motifs shared among angiosperms and gymnosperms. In the five motifs of the

N-terminal regions are conserved among angiosperms and gymnosperms while the C-terminal region is more variable, presenting lineage-specific motifs (Pascual et al., 2015; Hussey et al., 2015).

For the MYB gene family, *Populus* harbored the larger R2R3-MYB family (192 members) as a consequence of gene retention after a quite recent WGD in the Salicaceae lineage and tandem gene duplications (Wilkins et al., 2009). *Eucalyptus* and *Picea glauca* possess a total of 141 (Soler et al., 2015) and 122 (Rigault et al., 2011) MYB genes, respectively. Unequal expansion of particular clades has been observed in the R2R3-MYB gene family. The subgroups WPS-I to WPS-V are only present in woody species and some of them seem to be implicated in the regulation of wood development (Soler et al., 2015). Another subgroup (Sg4C) has expanded mainly in conifers with evidence of gene family expansion after the split of angiosperms and gymnosperms (Bedon et al., 2010). The MYB gene family also presents a modular structure. As the NAC gene family, it has a conserved DNA binding domain in the N-terminal region and a more variable C-terminal region. The MYB DNA binding domain is composed of sequence repeats, and as the name suggests, R2R3-MYB are composed by repeats R2 and R3 (Ambawat et al., 2013). This region is highly conserved between angiosperms and gymnosperms while in the C-terminal region, lineage-specific motifs were identified in conifers (Bedon et al., 2007).

The terpene synthase (TPS) gene family has also attracted much recent attention for its expansion in *Eucalyptus* (113 genes) (Külheim et al., 2015) and in *Picea glauca* (83 genes) (Warren et al., 2015), when compared to other angiosperms such as *Populus* (32 genes) (Tuskan et al., 2006) and *Arabidopsis* (32 genes) (Aubourg et al., 2002). This gene family is important in the synthesis of terpenes, which are compounds with protective functions against pests, which are highly relevant in ecological interactions. Four of the seven subfamilies are lineage-specific: TPS-a, TPS-b and TPS-g are angiosperm-specific and the subfamily TPS-d is gymnosperm-specific (Keeling et al; 2011). Lineage-specific expansion occurred in the TPS gene family. In flowering plants, TPS-a usually account for more than half of their TPS genes (Chen et al., 2011) while in conifers, the majority are TPS-d (Keeling et al; 2011; Warren et al., 2015). In *Eucalyptus*, the expansion of the gene family

was mainly due to single duplications, since they are mainly organized in tandem arrays (Külheim et al., 2015). The vast diversity found in the TPS gene family is reflected in their functional diversification. Several functional characterizations have been reported, however many genes remain to be characterized (Bohlmann et al., 1997; Keeling et al., 2011; Chen et al., 2011). For example, in *Eucalyptus*, the role of TPS genes in non-green tissues remains unknown (Külheim et al., 2015).

The Late Embryogenesis Abundant (LEA) gene family is a multigene family involved in the response to abiotic stresses such as drought, cold, salinity and heat. They are not only associated with environmental changes but also to water limitation produced under normal conditions during plant development such as the development of seeds (Battaglia et al., 2008). The LEA proteins can be classified in seven different groups based on their specific amino-acid motifs. The number of LEA genes varies among *Eucalyptus* (129 genes) (Li et al., 2015a), *Populus* (53) (Lan, Gao and Zeng, 2012) and *Picea glauca* (122) (Rigault et al., 2011), and there is even a greater variation between the different classes. For example, in *Picea glauca,* almost half of LEA genes are dehydrins (Rigault et al., 2011), showing an expansion of this class when compared to *Eucalyptus* (14 dehydrins) (Li et al., 2015a), and *Populus* (10 dehydrins) (Lan, Gao and Zeng, 2013). On the other hand, the high number of LEA genes in *Eucalyptus* is due to the many tandem duplications that occurred in another LEA class that has been called LEA2 (85 members) (Li et al., 2015a).

The most studied LEA genes are the dehydrins. They are composed of many amino-acid motifs (K, S, A, E, Y segments) presenting a variable composition and rearrangements of motifs. The K-segment defines the dehydrins and is present in all dehydrins studied to date with one exception in *Pinus taeda* (Perdiguero et al., 2014). Other motifs are lineage-specific such as the Y-segment present only in angiosperms (Tunnacliffe and Wise, 2007) and the A- and E-segments present only in conifers (Perdiguero et al., 2012). The different dehydrins have diversified expression profiles under drought, cold or salinity stresses. The relationship between classification and expression profiles is still unclear (Perdiguero et al., 2012; Falavigna et al., 2015; Hundemark and Hincha, 2008), though Perdiguero et al. (2012) suggested that A- and E- segments are related to divergent expression profiles.

1.4.2. Gene expression

1.4.2.1 Response to abiotic stress with emphasis on dehydration

As sessile and perennial organisms, trees have to endure environmental stresses such as drought, salinity and high and low temperatures. These environmental changes affect tree physiology and metabolism, impacting their growth and development as consequence (Cramer et al., 2011; Mackay and Dean, 2011). Stresses like drought, salinity and cold are interrelated, causing loss of cell water and decrease of cell osmotic potential, disrupting the osmotic and ionic homeostasis. Cell turgor also decreases, and as consequence, leaf expansion rate and plant growth rate also decrease (Duque et al., 2013). To adapt to adverse conditions, plants have developed cascades of molecular networks. The initial stress signals trigger downstream signaling processes and transcriptional regulation, which activate genes and proteins implicated in stress defensive mechanisms to re-establish cellular homeostasis and repair damaged proteins and membranes (Vinocur and Altman, 2005; Gong, Rao and Yu, 2013).

In recent years with the advance of sequencing technologies, several studies in forest trees have focused on understanding the transcriptomic network involved in stress response, especially drought stress (Harfouche, Meilan and Altman, 2014). Microarray and RNA-seq based transcriptomic studies have shown that environmental stresses affect gene expression of several groups of genes (Harfouche, Meilan and Altman, 2014). The expression patterns of genes involved in the response can be conserved but can also vary depending on the species, the genotype, the tissue and the type of stress. For example, the comparison of two genotypes of *Populus* showed differences in transcript abundance in response to drought (Wilkins et al., 2009). Moreover, comparisons of the transcriptome of leaves and roots have identified organ specificity in response to water deficit (Bogeat-Triboulot et al., 2007). In another example, transcriptome comparisons were made between two conifers (*Pinus contorta* and the natural hybrid *Picea engelmannii x Picea glauca*) that diverged over 100 million years ago (Savard et al., 1994). The analyses involved conditions varying in temperature, humidity and day length, and it showed that despite the time of divergence

between the two species, they have conserved expression patterns in response to stress in several orthologous genes (Yeaman et al., 2014).

Substantial knowledge has accumulated about proteins induced under dehydration stress (Ramanjulu and Bartels, 2002). One of the groups of proteins that accumulate during dehydration is the LEA. The LEA proteins have been described in a range of plants, from ferns to angiosperms. The LEA proteins are hydrophilic and accumulate in vegetative tissues under normal conditions and seeds at the last stages of embryogenesis (Tunnacliffe and Wise, 2007). LEA genes have been largely studied in many angiosperms and their expression profiles indicate that not all LEA genes are implicated in dehydration stress response (Tunnacliffe and Wise, 2007; Wang et al., 2007; Bies-Ethève et al., 2008; Lan, Gao and Zeng, 2013). In *Populus*, nine LEA genes from three different classes (4, 7 and 8) were differentially expressed under salt and drought stress. The LEA 4 class, which is the largest LEA group in *Populus*, presented the highest number of genes (6 genes) with differential expression under stress. Only two dehydrins (called LEA 8) responded to drought and salt stress (Lan, Gao and Zeng, 2013). In gymnosperms, many LEA genes have been identified (Gonzalez-Martinez et al., 2006; Lorenz et al., 2011; Rigault et al., 2011; Reid et al., 2013), but only a few of them have been studied structurally and functionally. Among conifer LEA genes, mainly dehydrins have been implicated in drought response (Perdiguero et al., 2012; Eldhuset et al., 2013; Perdiguero, Soto and Collada, 2015), in cold stress (Joosen et al., 2006) and in timing of bud burst (Yakovlev et al., 2008).

1.4.2.2 The dehydrin family

The dehydrin proteins have been shown to be implicated in biochemical processes such as buffering water, sequestering ions, stabilizing membranes, and also acting as chaperones (Kovacs et al., 2008). Their unstructured nature, lacking defined secondary and tertiary structure, confers the inability to denature during desiccation or at freezing temperatures. The unfolded state of dehydrins, their higher accumulation in various compartments inside the cells, and their capability to bind water, confers the capability of keeping the original

cell volume and preventing cellular collapse under dehydration conditions (Hanin et al., 2011).

In accordance with the cellular function of their gene products, the accumulation of dehydrins transcripts has been correlated with the effects of dehydrating conditions (Graether and Boddington et al., 2014). Several of these proteins are important in growth and plant development under permissive conditions since they accumulate in several tissues such as stem, leaves, flowers, fruit, phloem and xylem under normal conditions (Nylander et al., 2001; Bies-Ethève et al., 2008).

In *Populus* under normal growth conditions, different dehydrins were expressed in at least one of the following tissues: young leaves, roots, xylem and female and male catkins; presenting a diversified expression profile (Liu et al., 2012). In the PiceaGenExpress database (Raherison et al., 2012) the *Picea glauca* dehydrins showed also diversified expression profiles. Eight dehydrins were expressed in all tested tissues: vegetative buds, foliage, xylem, phelloderm, roots, megagametophytes and embryogenic cells, while the other 42 dehydrins were expressed in at least one of these tissues, wherein the great majority was expressed in more than four of the tissues listed above (Raherison et al., 2012).

Dehydrin expression can change in response to dehydration following drought, salinity or cold, but it is not a rule for all dehydrins. From the eight dehydrins analyzed in *Pinus pinaster* and *Pinus pinea*, three of them showed notable expression increase in roots and needles with the increase of water deficit. In the stem only, *Pinus pinea* showed an expression increase of these three dehydrins (Perdiguero, Soto and Collada, 2015). In *Picea obovata*, eight out of nine dehydrins increased their expression during acclimation to low temperature and decreased expression during deacclimation, suggesting an implication of dehydrins in protecting cells and tissues against freezing (Kjellsen et al., 2013), as suggested in other woody species. Similar patterns were observed in *Malus domestica*, *Populus nigra*, and *Prunus persica*, among others (Wisniewski et al., 1996). Under stress

conditions, only one dehydrin (*PgDhn1*) could be characterized in *Picea glauca*, revealing induction upon drought and cold stress (Richard et al., 2000).

1.5. Project context, research objectives and hypotheses

This thesis addresses two main subjects related to the evolution of genes in conifers. First, the evolution of gene structure is examined in the context of specific genome features that may have an impact on their evolution. Second, the evolution of the dehydrin gene family was analyzed to follow up on the observation of Rigault et al. (2011) that the dehydrin family was larger in conifers by reconstructing phylogenetic relationships and analyzing expression in response to dehydration stress.

This research was initiated before any of the conifer genome sequences were available, but different projects were underway to sequence the genomes of white spruce, Norway spruce and loblolly pine (Birol et al. 2013, Nystedt et al. 2013, Neale et al. 2014). A small number of BAC sequences had become available and other genomic sequence data was being produced including exome capture data. It thus became possible to catch a first glimpse of gene structure on a larger scale than previously possible in conifers. The accumulation of RNA-Seq data (Verta et al. 2016) and large-scale gene expression profiles (Raherison et al. 2012, 2015) complementing the gene catalogue available in white spruce (Rigault et al. 2011) also translated into opportunities to investigate entire gene families efficiently. Within this context of developing white spruce genomic resources, specific objectives and hypotheses were developed.

1.5.1 Gene structure evolution in conifers

We aimed to develop an understanding of gene structure in conifers based on interspecific comparisons in conifer trees and angiosperm plants by examining exon and intron sequences. We explored three main hypothesis: (1) Intron length is the major type of variation affecting gene structure in conifers compared to other plant species; (2) there is a positive relationship between genome size and intron length when comparing *Picea glauca* to *Arabidopsis thaliana*, *Zea mays* and *Populus trichocarpa*; (3) *Picea glauca* and *Pinus*

taeda present a conserved gene structure despite the fact that they diverged over 100 million years ago in keeping with their low rate of genome evolution.

The analyses initially focused on sequences isolated from a white spruce BAC library through comparisons with loblolly pine (*Pinus taeda*) BAC sequences and genome sequences from angiosperm plants. We then expanded the analyses to include sequences from the first genome draft assembly and sequence capture data from white spruce.

Our objectives were: (1) to define the gene structure (exons and introns) of 35 genes of *Picea glauca* taking into account gene expression profiles, i.e. highly expressed ubiquitous genes as well as more specialized genes with tissue preferential expression; (2) perform comparative structural analyses of the 35 *Picea glauca* genes and their close homologues in *Arabidopsis thaliana*, *Zea mays* and *Populus trichocarpa;* (3) perform pairwise comparisons of introns and exons between *Picea glauca* and *Pinus taeda;* (4) develop a *Picea glauca* repetitive library and screen nearly 2000 gene sequences obtained from sequence capture aiming to explore the potential impact of repetitive sequences on intron size.

1.5.2 Molecular structure and evolution of dehydrins in *Picea glauca*

A preliminary identification of 53 dehydrin genes in the white spruce transcriptome database (Raherison et al., 2012) suggested a possible expansion of this gene family in conifers. Taking into consideration the putative function of dehydrins, a premise of this work was that the expansion of this gene family could be linked to adaptation to dehydration stress resulting from arid or very cold conditions. As a basis to understand the role and evolution of dehydrin genes in conifers and to evaluate their involvement in water stress responses, we explored the following hypothesis: (1) the dehydrins containing the A- and E-segments exclusive to conifers, classified as A2E2SKn and ESKn, harbor profiles of increased expression under water deficit conditions: (2) *Picea glauca* dehydrins are classified in diverse groups presenting divergent patterns from angiosperm dehydrins.

Our objectives were to: (1) assess the extent of dehydrin gene family expansion in conifers starting from full length gene sequences identified in *P. glauca*; (2) trace the evolutionary origin of dehydrins in both conifers and angiosperms by studying phylogenetic relationships; (3) classify these genes based on conserved amino acids motifs such as the A, E, S, K segments; and (4) evaluate the expression profile of dehydrins in different tissues under normal conditions and in response to water stress.

1.6 References

Ahuja MR. Polyploidy in gymnosperms: revisited. Silvae Genetica, 2005:54: 59–69

Albalat R, Cañestro C. Evolution by gene loss. Nature Reviews Genetics. 2016;17:379–91.

Aitken SN, Yeaman S, Holliday JA, Wang T, Curtis-McLane S. Adaptation, migration or extirpation: climate change outcomes for tree populations. Evolutionary Applications. 2008;1:95–111.

Alberto FJ, Aitken SN, Alía R, González-Martínez SC, Hänninen H, Kremer A, et al. Potential for evolutionary responses to climate change – evidence from tree populations. Global Change Biology. 2013;19:1645–61.

Ambawat S, Sharma P, Yadav NR, Yadav RC. *MYB* transcription factor genes as regulators for plant responses: an overview. Physiology and Molecular Biology of Plants. 2013;19:307–21.

Aubourg S, Lecharny A, Bohlmann J. Genomic analysis of the terpenoid synthase (*AtTPS*) gene family of *Arabidopsis thaliana*. Molecular Genetics and Genomics. 2002;267:730–45.

Bagowski CP, Bruins W, Te Velthuis AJW. The nature of protein domain evolution: shaping the interaction network. Current Genomics. 2010;11:368–76.

Bartholomé J, Mandrou E, Mabiala A, Jenkins J, Nabihoudine I, Klopp C, et al. High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. New Phytologist. 2015;206:1283–1296.

Battaglia M, Olvera-Carrillo Y, Garciarrubio A, Campos F, Covarrubias AA. The Enigmatic LEA Proteins and Other Hydrophilins. Plant Physiology. 2008;148:6–24.

Bautista R, Villalobos DP, Díaz-Moreno S, Cantón FR, Cánovas FM, Claros MG. Toward a *Pinus pinaster* bacterial artificial chromosome library. Annals of Forest Science. 2007;64:855–864.

Bedon F, Grima-Pettenati J, Mackay J. Conifer *R2R3-MYB* transcription factors: sequence analyses and gene expression in wood-forming tissues of white spruce (*Picea glauca*). BMC Plant Biology. 2007;7:17.

Bedon F, Bomal C, Caron S, Levasseur C, Boyle B, Mansfield SD, et al. Subgroup 4 R2R3-MYBs in conifer trees: gene family expansion and contribution to the isoprenoid- and flavonoid-oriented responses. Journal of Experimental Botany 2010;61:3847–64.

Bennetzen JL. Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica. 2002;115:29–36.

Bies-Ethève N, Gaubier-Comella P, Debures A, Lasserre E, Jobet E, Raynal M, et al. Inventory, evolution and expression profiling diversity of the LEA (late embryogenesis abundant) protein gene family in *Arabidopsis thaliana*. Plant Molecular Biology. 2008;67:107–24.

Birchler JA, Veitia RA. The gene balance hypothesis: from classical genetics to modern genomics. The Plant Cell. 2007;19:395–402.

Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, et al. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. Bioinformatics. 2013;29:1492–1497.

Bogeat-Triboulot M-B, Brosché M, Renaut J, Jouve L, Le Thiec D, Fayyaz P, et al. Gradual soil water depletion results in reversible changes of gene expression, protein profiles, ecophysiology, and growth performance in *Populus euphratica*, a poplar growing in arid regions. Plant Physiology. 2007;143:876–92.

Bohlmann J, Steele CL, Croteau R. Monoterpene synthases from grand fir (*Abies grandis*). cDNA isolation, characterization, and functional expression of myrcene synthase, (-)-(4S)-limonene synthase, and (-)-(1S,5S)-pinene synthase. The Journal of Biological Chemistry. 1997;272:21784–92.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. Selection for short introns in highly expressed genes. Nature Genetics. 2002;31:415–8.

Chen F, Tholl D, Bohlmann J, Pichersky E. The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. The Plant Journal. 2011;66:212–29.

Comeron JM, Kreitman M. The correlation between intron length and recombination in drosophila. Dynamic equilibrium between mutational and selective forces. Genetics. 2000;156:1175–90.

Cossu RM, Buti M, Giordani T, Natali L, Cavallini A. A computational study of the dynamics of LTR retrotransposons in the *Populus trichocarpa* genome. Tree Genetics and Genomes. 2012;8:61–75.

Cramer GR, Urano K, Delrot S, Pezzotti M, Shinozaki K. Effects of abiotic stress on plants: a systems biology perspective. BMC Plant Biology. 2011;11:163.

Dai X, Hu Q, Cai Q, Feng K, Ye N, Tuskan GA, et al. The willow genome and divergent evolution from poplar after the common genome duplication. Cell Research. 2014;24:1274–7.

Davis JC, Petrov DA, Ken H. Wolfe. Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biology. 2004;2:e55.

De La Torre, A. R., Birol I, Bousquet J, Ingvarsson PK, Jansson S, Jones SJM, et al. Insights into conifer giga-genomes. Plant Physiology. 2014;166:1724–1732.

De Miguel M, Bartholomé J, Ehrenmann F, Murat F, Moriguchi Y, Uchiyama K, et al. Evidence of intense chromosomal shuffling during conifer evolution. Genome Biology and Evolution 2015;7:2799–809.

Deutsch M, Long M. Intron-exon structures of eukaryotic model organisms. Nucleic Acids Research. 1999;27:3219–28.

Douglas CJ, DiFazio SP. The *Populus* genome and comparative genomics. In: Jansson S, Bhalerao R, Groover A, Editors. Genetics and Genomics of *Populus*. New York, NY: Springer New York; 2010. pp. 67–90.

Duque AS, de Almeida AM, da Silva AB, da Silva JM, Farinha AP, Santos D, et al. Abiotic stress responses in plants: unraveling the complexity of genes and networks to survive. In: Vahdati K, Editor. Abiotic Stress - Plant Responses and Applications in Agriculture. In Tech 2013. pp. 49-101.

Duval I, Lachance D, Giguère I, Bomal C, Morency M-J, Pelletier G, et al. Large-scale screening of transcription factor–promoter interactions in spruce reveals a transcriptional network involved in vascular development. Journal of Experimental Botany 2014;65:2319–33.

Eldhuset TD, Nagy NE, Volařík D, Børja I, Gebauer R, Yakovlev IA, et al. Drought affects tracheid structure, dehydrin expression, and above- and belowground growth in 5-year-old Norway spruce. Plant and Soil. 2012;366:305–20.

Falavigna V da S, Miotto YE, Porto DD, Anzanello R, Santos HP dos, Fialho FB, et al. Functional diversification of the dehydrin gene family in apple and its contribution to cold acclimation during dormancy. Physiologia Plantarum. 2015;155:315–29.

Gauthier S, Bernier P, Kuuluvainen T, Shvidenko AZ, Schepaschenko DG. Boreal forest health and global change. Science. 2015;349:819–22.

Gong Y, Rao L, Yu D. Abiotic stress in plants. In: Stoytcheva M, Editor. Agricultural Chemistry. In Tech; 2013; pp.113-52.

González-Martínez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB. DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response *in Pinus taeda L*. Genetics. 2006;172:1915–26.

Graether SP, Boddington KF. Disorder and function: a review of the dehydrin protein family. Frontiers in Plant Science 2014;5:576.

Guillet-Claude C., Isabel N, Pelgas B, Bousquet J. The evolutionary implications of *knox-I* gene duplications in conifers: correlated evidence from phylogeny, gene mapping, and analysis of functional divergence. Molecular Biology and Evolution. 2004:21:2232-2245.

Guo Y-L. Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes. The Plant Journal. 2013;73:941–51.

Hanin M, Brini F, Ebel C, Toda Y, Takeda S, Masmoudi K. Plant dehydrins and stress tolerance. Plant Signal and Behaviour 2011;6:1503–9.

Harfouche A, Meilan R, Altman A. Molecular and physiological responses to abiotic stress in forest trees and their relevance to tree improvement. Tree Physiology. 2014;34:1181–98.

Hu R, Qi G, Kong Y, Kong D, Gao Q, Zhou G. Comprehensive Analysis of NAC Domain Transcription Factor Gene Family in *Populus trichocarpa*. BMC Plant Biology. 2010;10:145.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nature Genetics. 2011;43:476–481.

Hundertmark M, Hincha DK. LEA (late embryogenesis abundant) proteins and their encoding genes in *Arabidopsis thaliana*. BMC Genomics. 2008;9:118.

Hussey SG, Saïdi MN, Hefer CA, Myburg AA, Grima-Pettenati J. Structural, evolutionary and functional analysis of the NAC domain protein family in *Eucalyptus*. New Phytologist. 2015;206:1337–50.

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449:463–467.

Joosen RVL, Lammers M, Balk PA, Brønnum P, Konings MCJM, Perks M, et al. Correlating gene expression to physiological parameters and environmental conditions during cold acclimation of *Pinus sylvestris*, identification of molecular markers using cDNA microarrays. Tree Physiology 2006;26:1297–313.

Keeling CI, Weisshaar S, Ralph SG, Jancsik S, Hamberger B, Dullat HK, et al. Transcriptome mining, functional characterization, and phylogeny of a large terpene synthase gene family in spruce (*Picea spp.)*. BMC Plant Biology. 2011;11:43.

Kovach A, Wegrzyn JL, Parra G, Holt C, Bruening GE, Loopstra CA, et al. The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics. 2010;11:420.

Kovacs D, Kalmar E, Torok Z, Tompa P. Chaperone activity of *ERD10* and *ERD14*, two Disordered stress-related plant proteins. Plant Physiology. 2008;147:381–90.

Külheim C, Padovan A, Hefer C, Krause ST, Köllner TG, Myburg AA, et al. The *Eucalyptus* terpene synthase gene family. BMC Genomics. 2015;16:450.

Lan T, Gao J, Zeng Q-Y. Genome-wide analysis of the LEA (late embryogenesis abundant) protein gene family in *Populus trichocarpa*. Tree Genetics and Genomes. 2012;9:253–64.

Li Q, Yu H, Cao PB, Fawal N, Mathé C, Azar S, et al. Explosive tandem and segmental duplications of multigenic families in *Eucalyptus grandis*. Genome Biology and Evolution. 2015;7:1068–81.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, et al. Early genome duplications in conifers and other seed plants. Science Advances. 2015;1:e1501084.

Liu C-C, Li C-M, Liu B-G, Ge S-J, Dong X-M, Li W, et al. Genome-wide identification and characterization of a dehydrin gene family in poplar (*Populus trichocarpa*). Plant Molecular Biology Report. 2012;30:848–59.

Lorenz WW, Alba R, Yu Y-S, Bordeaux JM, Simões M, Dean JF. Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (*P. taeda L.*). BMC Genomics. 2011;12:264.

Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000;154:459–473.

Lynch M, Conery JS. The Origins of Genome Complexity. Science. 2003;302:1401–4.

Mackay J, Dean JFD. Transcriptomics. In: Plomion C, Bousquet J, Kole C, Editors. Genetics, Genomics and Breeding of Conifers. .CRC Press and Science Publishers, New York; 2011, pp.323-357.

Mackay J, Dean JFD, Plomion C, Peterson DG, Cánovas FM, Pavy N, et al. Towards decoding the conifer giga-genome. Plant Molecular Biology. 2012;80:555–69.

Magbanua ZV, Ozkan S, Bartlett BD, Chouvarine P, Saski CA, Liston A, et al. Adventures in the enormous: a 1.8 Million clone BAC library for the 21.7 Gb genome of loblolly pine. PLoS ONE. 2011;6:e16214.

Morgante M, Paoli E. Toward the conifer genome sequence. In: Plomion C, Bousquet J, Kole C, Editors. Genetics, Genomics and Breeding of Conifers. .CRC Press and Science Publishers, New York; 2011, pp.389-403.

Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, Garcia SA, et al. Evolution of genome size and complexity in *Pinus*. PLoS ONE. 2009;4:e4332.

Murray, B. G., Leitch, I. J., Bennett, M. D. Gymnosperm DNA C-values Database 2012, http://www.kew.org/cvalues.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, et al. The genome of *Eucalyptus grandis*. Nature, 2014; 510:356-62.

Neale DB, Ingvarsson PK. Population, quantitative and comparative genomics of adaptation in forest trees. Current Opinion in Plant Biology. 2008;11:149–55.

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biology. 2014;15:R59.

Nuruzzaman M, Manimekalai R, Sharoni AM, Satoh K, Kondoh H, Ooka H, et al. Genome-wide analysis of NAC transcription factor family in rice. Gene. 2010;465:30–44.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, et al. The Norway spruce genome sequence and conifer genome evolution. Nature. 2013;497:579–84.

Nylander M, Svensson J, Palva ET, Welin BV. Stress-induced accumulation and tissue-specific localization of dehydrins in *Arabidopsis thaliana*. Plant Molecular Biology. 2001;45:263–79.

Ohno S. Evolution by Gene Duplication. Berlin and New York: Springer-Verlag; 1970.

Olsen AN, Ernst HA, Leggio LL, Skriver K. NAC transcription factors: structurally distinct, functionally diverse. Trends in Plant Science. 2005;10:79–87.

Parent GJ, Raherison E, Sena J, MacKay JJ. Forest tree genomics: review of progress. Advances in Botanical Research. Elsevier; 2015. p. 39–92.

Pascual MB, Cánovas FM, Ávila C. The NAC transcription factor family in maritime pine (*Pinus Pinaster*): molecular regulation of two genes involved in stress responses. BMC Plant Biology, 2015; 15:254.

Pavy N, Pelgas B, Laroche J, Rigault P, Isabel N, Bousquet J. A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. BMC Biology. 2012;10:84.

Perdiguero P, Barbero MC, Cervera MT, Soto Á, Collada C. Novel conserved segments are associated with differential expression patterns for *Pinaceae* dehydrins. Planta. 2012;236:1863–74.

Perdiguero P, Collada C, Soto Á. Novel dehydrins lacking complete K-segments in Pinaceae. The exception rather than the rule. Frontiers in Plant Science. 2014;5:682.

Perdiguero P, Soto Á, Collada C. Comparative analysis of Pinus pinea and Pinus pinaster dehydrins under drought stress. Tree Genetics and Genomes. 2015;11:70.

Plomion C, Aury J-M, Amselem J, Alaeitabar T, Barbe V, Belser C, et al. Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. Molecular Ecology Resources. 2016;16:254–65.

Raherison E, Rigault P, Caron S, Poulin P-L, Boyle B, Verta J-P, et al. Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression. BMC Genomics. 2012;13:434.

Raherison ESM, Giguère I, Caron S, Lamara M, MacKay JJ. Modular organization of the white spruce (*Picea glauca*) transcriptome reveals functional organization and evolutionary signatures. New Phytologist. 2015;207:172–87.

Ramanjulu S, Bartels D. Drought- and desiccation-induced modulation of gene expression in plants. Plant, Cell and Environment. 2002;25:141–51.

Reid KE, Holliday JA, Yuen M, Nguyen A, Aitken SN, Bohlmann J. Sequencing of Sitka spruce (*Picea sitchensis*) cDNA libraries constructed from autumn buds and foliage reveals autumn-specific spruce transcripts. Tree Genetics and Genomes. 2013;9:683–91.

Ren L-L, Liu Y-J, Liu H-J, Qian T-T, Qi L-W, Wang X-R, et al. Subcellular relocalization and positive selection play key roles in the retention of duplicate genes of *Populus* class III Peroxidase family. The Plant Cell. 2014;26:2404–2419.

Richard S, Morency M-J, Drevet C, Jouanin L, Séguin A. Isolation and characterization of a dehydrin gene from white spruce induced upon wounding, drought and cold stresses. Plant Molecular Biology. 2000;43:1–10.

Rigault P, Boyle B, Lepage P, Cooke JEK, Bousquet J, MacKay JJ. A white spruce gene catalog for conifer genome analyses. Plant Physiology. 2011;157:14–28.

Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, et al. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. Genome Research. 2012;22:95–105.

Savard L, Li P, Strauss SH, Chase MW, Michaud M, Bousquet J. Chloroplast and nuclear gene sequences indicate Late Pennsylvanian time for the last common ancestor of extant seed plants. Proceedings of the National Academy of Sciences of the U.S.A. 1994:91:5163-5167.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: A parallel assembler for short read sequence data. Genome Research. 2009;19:1117–1123.

Singh AK, Sharma V, Pal AK, Acharya V, Ahuja PS. Genome-wide organization and expression profiling of the *NAC* transcription factor family in potato (*Solanum tuberosum L.*). DNA Research 2013;20:403–23.

Soler M, Camargo ELO, Carocha V, Cassan-Wang H, San Clemente H, Savelli B, et al. The *Eucalyptus grandis R2R3-MYB* transcription factor family: evidence for woody growth-related evolution and function. New Phytologist. 2015;206:1364–77.

Soltis PS, Soltis DE. A conifer genome spruces up plant phylogenomics. Genome Biology. 2013;14:122.

Stevens KA, Wegrzyn J, Zimin A, Puiu D, Crepeau M, Cardeno C, et al. Sequence of the sugar pine megagenome. Genetics. 2016;genetics.116.193227.

Stival Sena J, Giguère I, Boyle B, Rigault P, Birol I, Zuccolo A, et al. Evolution of gene structure in the conifer *Picea glauca*: A comparative analysis of the impact of intron size. BMC Plant Biology. 2014;14:95.

Tunnacliffe A, Wise MJ. The continuing conundrum of the LEA proteins. Naturwissenschaften. 2007;94:791–812.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science. 2006;313:1596–1604.

Verta J-P, Landry CR, MacKay J. Dissection of expression-quantitative trait locus and allele specificity using a haploid/diploid plant system – insights into compensatory evolution of transcriptional regulation within populations. New Phytologist. 2016;211:159–71.

Vinocur B, Altman A. Recent advances in engineering plant tolerance to abiotic stress: achievements and limitations. Current Opinion in Biotechnology. 2005;16:123–32.

Wang N, Thomson M, Bodles WJA, Crawford RMM, Hunt HV, Featherstone AW, et al. Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. Molecular Ecology 2013;22:3098–111.

Wang N, Zheng Y, Xin H, Fang L, Li S. Comprehensive analysis of *NAC* domain transcription factor gene family in *Vitis vinifera*. Plant Cell Reports 2013;32:61–75.

Wang X-S, Zhu H-B, Jin G-L, Liu H-L, Wu W-R, Zhu J. Genome-scale identification and analysis of LEA genes in rice (*Oryza sativa L.).* Plant Science. 2007;172:414–20.

Warren RL, Keeling CI, Yuen MMS, Raymond A, Taylor GA, Vandervalk BP, et al. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. The Plant Journal. 2015;83:189–212.

Wegrzyn JL, Liechty JD, Stevens KA, Wu L-S, Loopstra CA, Vasquez-Gross HA, et al. Unique features of the loblolly pine (*Pinus taeda L.*) megagenome revealed through sequence annotation. Genetics. 2014;196:891–909.

Wilkins O, Waldron L, Nahal H, Provart NJ, Campbell MM. Genotype and time of day shape the *Populus* drought response. The Plant Journal. 2009;60:703–15.

Wisniewski M, Close TJ, Artlip T, Arora R. Seasonal patterns of dehydrins and 70-kDa heat-shock proteins in bark tissues of eight species of woody plants. Physiologia Plantarum. 1996;96:496–505.

Yakovlev IA, Asante DKA, Fossdal CG, Partanen J, Junttila O, Johnsen O. Dehydrins expression related to timing of bud burst in Norway spruce. Planta. 2008;228:459–72.

Yeaman S, Hodgins KA, Suren H, Nurkowski KA, Rieseberg LH, Holliday JA, et al. Conservation and divergence of gene expression plasticity following c. 140 million years of evolution in lodgepole pine (*Pinus contorta*) and interior spruce (*Picea glauca*×*Picea engelmannii*). New Phytologist. 2014;203:578–91.

Zhou F, Xu Y. RepPop: A database for repetitive elements in *Populus trichocarpa*. BMC Genomics. 2009;10:14.

Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, et al. Sequencing and assembly of the 22-Gb loblolly pine genome. Genetics. 2014;196:875–90.

Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29:2669–2677.

Chapter 2: Evolution of gene structure in the conifer *Picea glauca*: a comparative analysis of the impact of intron size

[Stival Sena J, Giguère I, Boyle B, Rigault P, Birol I, Zuccolo A, Ritland K, Ritland C, Bohlman J, Jones S, Bousquet J, Mackay J. Evolution of gene structure in the conifer *Picea glauca*: A comparative analysis of the impact of intron size. BMC Plant Biology. 2014;14:95]

2.1 Abstract

A positive relationship between genome size and intron length is observed across eukaryotes including angiosperms plants, indicating a co-evolution of genome size and gene structure. Conifers have very large genomes and longer introns on average than most plants, but the impacts of their large genome and longer introns on gene structure have not been described.

Gene structure was analyzed for 35 genes of *Picea glauca* obtained from BAC sequencing and genome assembly, including comparisons with *A. thaliana*, *P. trichocarpa* and *Z. mays*. We aimed to develop an understanding of the impact of long introns on the structure of individual genes. The number and length of exons was well conserved among the species compared but on average, *P. glauca* introns were longer and genes had four times more intronic sequence than *Arabidopsis*, and 2 times more than poplar and maize. However, pairwise comparisons of individual genes gave variable results and not all contrasts were statistically significant. Genes generally accumulated one or a few longer introns in species with larger genomes but the position of long introns was variable between plant lineages. In *P. glauca*, highly expressed genes generally had more intronic sequence than tissue preferential genes. Comparisons with the *Pinus taeda* BACs and genome scaffolds showed a high conservation for position of long introns and for sequence of short introns. A survey

of 1836 *P. glauca* genes obtained by sequence capture mostly containing introns <1 Kbp showed that repeated sequences were 10× more abundant in introns than in exons.

Conifers have large amounts of intronic sequence per gene for seed plants due to the presence of few long introns and repetitive element sequences are ubiquitous in their introns. Results indicate a complex landscape of intron sizes and distribution across taxa and between genes with different expression profiles.

2.2 Résumé

Une relation positive entre la taille du génome et la longueur des introns est observée chez les eucaryotes, y compris les plantes du groupe des angiospermes, indiquant une co-évolution de la taille du génome et de la structure des gènes. Les conifères présentent des génomes très grands et des introns plus longs en moyenne que la plupart des plantes, mais les impacts de leur grand génome et des longs introns sur la structure génique n'ont pas été décrits.

La structure génique de 35 gènes obtenus à partir du séquençage de BAC et de l'assemblage du génome a été analysée chez *Picea glauca*, comprenant des analyses comparatives avec *Arabidopsis thaliana, Populus trichocarpa* et *Zea mays*. Notre objectif était de comprendre l'impact des longs introns sur la structure des gènes, individuellement. Le nombre et la longueur des exons se sont vus bien conservés parmi les espèces comparées, mais en moyenne les introns chez *Picea glauca* étaient plus longs et les gènes avaient quatre fois plus de séquence intronique qu'Arabidopsis, et deux fois plus que *Populus* et *Zea mays*. Cependant, le résultat des comparaisons entre les gènes homologues a été variable, et les contrastes, pas tous statistiquement significatifs. Généralement, les gènes ont accumulé un ou quelques introns plus longs chez les espèces avec des génomes plus grands, toutefois la position des longs introns était variable entre les espèces. Chez *Picea glauca*, généralement, les gènes fortement exprimés dans plusieurs tissus ont présenté plus de séquence intronique que les gènes plus spécialisés. Les comparaisons des séquences de BAC et des séquences génomiques entre *Picea glauca* et *Pinus taeda* ont montré que la position des introns longs et que la séquence des introns courts sont conservés entre les deux espèces. Une étude de 1836 gènes de *Picea glauca* obtenus par «sequence capture» contenant principalement des introns <1 Kbp a montré que les séquences répétées étaient dix fois plus abondantes dans les introns que dans les exons.

Les conifères présentent des quantités importantes de séquences introniques par gène due à la présence de quelques introns longs et des séquences d'éléments répétitifs. Les résultats

montrent que la taille des introns, leur distribution entre les taxons et les différents profils d'expression entre les gènes font partie d'un scenario évolutif complexe.

2.3 Introduction

Many factors related to genome size, recombination rate, expression level, and effective population size, among others, have been proposed to affect the evolution of gene structure [1, 2, 3, 4]. At the molecular level, genome size variations may result from mobile or transposable elements (TEs), whole genome duplication events, and polyploidization events, among others. Comparative studies have shown that intron lengths and the abundance of mobile elements directly correlate with genome size, such that large genomes have longer introns and a higher proportion of mobile elements [1]. Mobile elements also impact gene structure and function as they can insert into genes, including introns and exons, and thus contribute to the evolution of genes.

Conifer trees have very large genomes ranging from 18 to 35 Gbp [5] that are composed of a large fraction of repetitive sequences [6, 7]. New insight into plant genome evolution are expected from the unique structure and history of conifer genomes [8], which may contribute to a broader understanding of the relationships between gene structure and genome architecture. Draft genome assemblies were recently reported for the European *Picea abies* (Norway spruce) [9] as well as the North American species *Picea glauca* (white spruce) [10] and *Pinus taeda* (loblolly pine) [11, 12]. Nystedt et al. [9] reported that Norway spruce and other conifers accumulate long introns and showed that some introns can be very long (>10 Kbp) compared to other plant species.

A positive relationship between genome size and intron length has been observed in broad phylogenetic studies [2, 13, 14] including between recently diverged *Drosophila* species harboring considerable difference in genome size, where *D. viliris* had longer introns than *D. melanogaster*[15]. In plants, a few studies investigated this question within angiosperms, indicating that genome size is not necessarily a good predictor of intron length [16, 17] although a general trend is observed. For instance, *Arabidopsis thaliana*, *Populus trichocarpa*, *Zea mays* have well characterized genomes that range in size from 125 Mbp to 2.3 Gbp; their average exons sizes are between 250 and 259, whereas their introns sizes are 168 bp, 380 bp and 607 bp on average, respectively [18, 19, 20]. The

length of introns may depend upon gene function and expression level; however, there is considerable debate surrounding this issue when it comes to plant genomes. In *Oryza sativa* and *A. thaliana* it was found that highly expressed genes contained more and longer introns than genes expressed at a low level [21], which is in contrast to findings in *Caenorhabditis elegans* and *Homo sapiens* [4].

Transposable elements are among the factors that may influence the evolution of intron size, as they represent the major component of plant genomes [22]. In *Vitis vinifera*, transposable elements comprise 80% of long introns [17]. In many plants, LTR-RT represent a large fraction of the genome but are more abundant in gene poor regions of the genome; therefore, their impact on the evolution of gene structure may actually be lesser than other classes of transposable elements such as MITEs [23] and helitrons, both of which are known to insert into or close to genes [24].

To date, studies related to genome size and the evolution of plant introns have primarily involved angiosperms (flowering plants), many of which have genomes under 1 Gbp. More recently, the *Picea abies* and *Pinus taeda* genomes were shown to have among the largest average introns size [9, 12]. We aimed to develop an understanding of the gene structure in conifers through a detailed analysis of individual genes with a particular emphasis on the potential impact of long introns on gene structure through comparative analyses. An underlying question relates to potential impacts on gene expression; therefore, our analyses took into account their expression profiles. Gene structure was analyzed in two conifers (*P. glauca* and *P. taeda*) and three angiosperms. We explored three main hypothesis: (1) Intron length is the major type of variation affecting gene structure in conifers compared to other plant species; (2) there is a positive relationship between genome size and intron length in *P. glauca* compared to *A. thaliana*, *Z. mays* and *P. trichocarpa*; (3) *P. glauca* and *P. taeda* present a conserved gene structure despite the fact that they diverged over 100 MYA in keeping with their low rate of genome evolution [8].

We present a detailed analysis of gene structure for 35 genes from the conifer *Picea glauca* obtained from BAC sequencing and genome assembly and comparative analyses

with *A. thaliana*, *P. trichocarpa* and *Z. mays*. Our study also included the analysis of nearly 6000 gene sequences obtained from sequence capture aiming to explore the potential impact of repetitive sequences on intron size in *P. glauca*. Our findings show that intron size and the position of long introns within genes is variable between plant lineages but highly conserved in conifers.

2.4 Results

2.4.1 Genomic sequences

Genomic sequences were analyzed for several *P. glauca* genes. The sequences were obtained either by targeted BAC isolations, from an early assembly of the *P. glauca* genome [10], or from a sequence capture experiment (for details, see Methods).

A total of 21 BAC clones were isolated each containing a different single copy gene associated with secondary cell-wall formation or with nitrogen metabolism. Following shotgun sequencing by GS-FLX and assembly with the Newbler software, the integrity and identity of each gene was verified. Estimated size of BAC clones was 131 Kb on average and coverage was 144× (for Summary statistics, see Supplementary information: Table S2.6). Twenty of the 21 targeted genes were complete as determined by sequence alignment indicating full coverage of FL cDNA sequences from spruces and pines (*P. glauca, P. sitchensis, P. taeda* and *Pinus sylvestris*) [25, 26, 27,28] (Supplementary information: Table S2.7). Nearly all genes were contained within a single contig, except the LIM gene which lacked one exon, and the Susy gene which was complete cDNA sequence but spanned two contigs. None of BACs contained other genes as determined by BLAST searches against the *P. glauca* gene catalog [29] and the Swiss Prot database.

Sequences were also isolated from a whole genome shotgun assembly of *P. glauca* [10]. Sequences with ubiquitous expression were targeted in order to complement the set of more specialized genes which had been selected for BAC isolation. The *P. glauca* genome shotgun assembly was screened with the complete CDS derived from cDNA sequences (according to Rigault et al. [29]) that were highly expressed in most tissues (according to

Raherison et al. [30]). A total of 18 genomic sequences were randomly selected among those that spanned the entire coding region of the targeted gene.

2.4.2 Gene expression profiles

Transcript accumulation profiles from eight different tissues were obtained from the PiceaGenExpress database [30] for each of the gene sequences described above (Figure 2.1). The transcript data indicated that the group of highly expressed genes was detected in all tissues and with average abundance class above 9.7 (out of 10) across all tissues (Figure 2.1, top). In contrast, the genes associated with wood formation and nitrogen metabolism nearly all had tissue preferential expression patterns; they were detected in six tissues on average (range of two to eight tissues) and had an average transcript abundance class of 5.8 in those tissues where the genes were expressed (Figure 2.1, bottom).

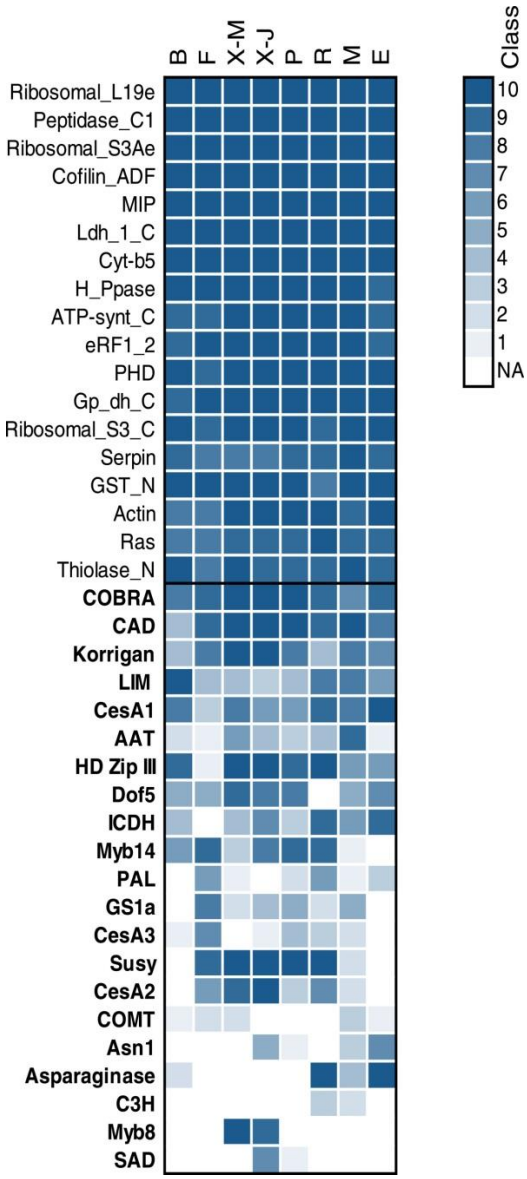


Figure 2.1- Transcript accumulation profiles from the PiceaGenExpress database (Raherison et al. [30]**)** of the *P. glauca* genes. The transcript abundance data are classified from 1 to 10, from lowest to highest microarray hybridization intensities detected within a given tissue. The profiles of highly expressed genes (top) (according to Raherison et al. [30]; class 8 to 10 are contrasted with most of the genes associated with secondary cell wall formation and nitrogen metabolism (bottom, names in bold). NA: Not detected. Tissues: B (Vegetative buds), F (Foliage), X-M (Xylem – from mature trees), X-J (Xylem –juvenile trees), P (Phelloderm), R (Adventitious roots), M (Megagametophytes), E (Embryogenic cells).

2.4.3 Gene structures and comparative analysis with angiosperms

The gene structure (exon and introns regions) of *P. glauca* genes was determined by mapping the complete cDNA onto the genomic sequence (BACs or shotgun contigs) for 35 genes. Homologs were retrieved from three well-characterized angiosperm genomes, *Arabidopsis thaliana* [19], *Populus trichocarpa* [18] and *Zea mays* [20]. The comparative analyses considered all of the genes together and also as two separate groups, i.e. genes highly expressed and genes related to secondary cell-wall formation and nitrogen metabolism. On average, the protein coding sequence similarity between *P. glauca* and *A. thaliana* was 76%, 78% with *P. trichocarpa* and 75% with *Z. mays*.

The number of exons and introns was well conserved between homologous genes among the different species (Table 2.1). The average length of exons was also well conserved between homologs among species (average of 240 bp, median of 155 bp) and varied only slightly between the two sub-groups genes (Table 2.1 and Supplementary information: Figure S2.2). Pairwise comparisons of matching exons also indicated conservation of length among the species considered (not shown). These observations indicate that exon structure is generally well conserved.

Table 2.1- Average number and length of exons in genes used for comparative analyses

	Highly expressed genes[1]			**Secondary cell-wall formation and nitrogen metabolism genes[2]**		
	Exon number	**Exon length**	**Standard deviation**	**Exon number**	**Exon length**	**Standard deviation**
Arabidopsis thaliana	5.9	220.8	215.0	9.1	228.9	189.8
Populus trichocarpa	6.2	241.5	253.3	9.4	261.1	263.5
Zea mays	6.1	244.5	236.7	9.0	257.6	274.8
Picea glauca	6.2	236.5	226.0	9.5	223.9	217.8

[1]Data were obtained from 18 different genes and an average total of 109 exons per species.
[2]Data were obtained from 17 different genes and an average total of 157 exons per species.

In contrast, introns revealed much more variation between species. Our analyses included comparisons of individual introns and of total intronic sequences in each gene. The average length of individual introns (in bp) was 144, 295, 454, and 532 for *A. thaliana, P. trichocarpa, Z. mays and P. glauca*, respectively (Figure 2.2 and Supplementary information: Figure S2.2). The average intron length varied significantly among *P. glauca*

and the three species; pairwise contrasts were significant with *A. thaliana* and *Z. mays*, and nearly significant with *P. trichocarpa* (Figure 2.2). In *P. glauca*, *P. trichocarpa* and *Z. mays*, we also observed that intron lengths were more heterogeneous as shown by differences between low and upper quartiles, minimum and maximum lengths and outliers of large size (Figure 2.2). The average length of the longest intron per gene was 382 bp in *A. thaliana*, 806 bp in *P. trichocarpa*, 1652 bp in *Z. mays* and 2022 bp in *P. glauca*.

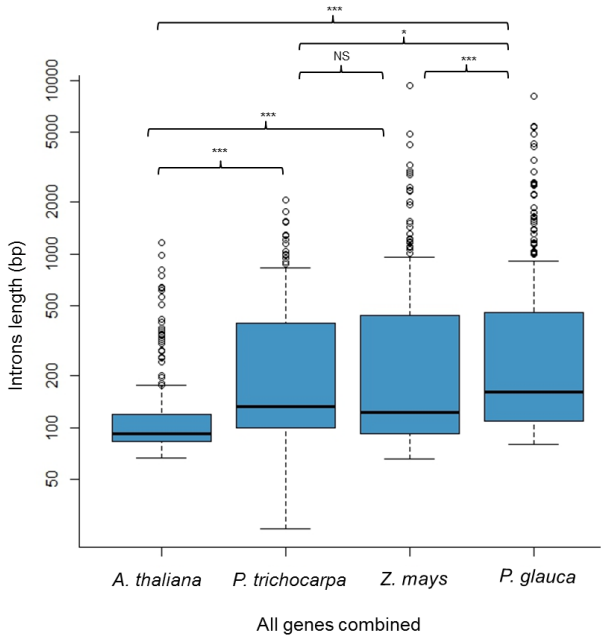


Figure 2.2- Comparative analysis of individual intron length in *P. glauca*, *A. thaliana P. trichocarpa* and *Z. mays*. Box plots represent intron length data for all of the introns of the 35 genes used in comparative analyses. Intron lengths were compared among the four species by Kruskal-Wallis test with post-test analysis by Dunn's multiple comparisons: NS, not significant ($P \geq 0.06$); *P = 0.06; **$P < 0.01$; ***$P < 0.001$.

Comparison of the total length of intronic sequences on a gene-by-gene basis showed that on average, *P. glauca* genes had 4.1 times more intronic sequences than *A. thaliana*, 2.2 times more than *P. trichocarpa* and 1.8 times more than *Z. mays* (Figure 2.3A). The total length of intron sequences and length ratio was calculated for each gene in pairwise comparisons between all of the species. Comparisons between *P. glauca* and *A. thaliana* gene sets were statistically significant (Figure 2.3); the ratios were close to five on

average in highly expressed genes and three in genes associated with secondary cell-wall formation and nitrogen metabolism (Figure 2.3B). In contrast, the ratio of total intron lengths between *P. glauca* compared to *P. trichocarpa* and *Z. mays* was constant at around two-fold and the total length of intronic sequence per gene was not statistically different. Results also indicated that *A. thaliana* has significantly less intronic sequence than *P. trichocarpa* and *Z. mays* and that their ratios were most different for the highly expressed genes and more similar for the genes involved in secondary cell-wall formation and nitrogen metabolism (Figure 2.3B). A significant difference of intron lengths was also observed between the two expression groups within *P. glauca* ($p < 0.05$).

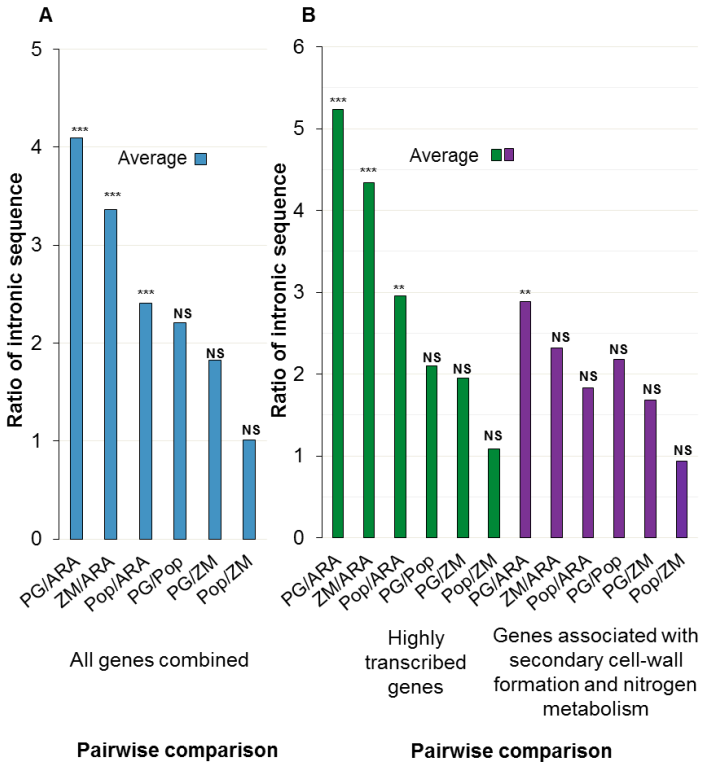


Figure 2.3- Comparative analysis of total intron length in *P. glauca*, *A. thaliana*, *P. trichocarpa* and *Z. mays*. Average ratio of total length of intron sequences in pair-wise comparisons in: A- all genes; B- highly expressed genes and genes involved in secondary cell-wall formation and nitrogen metabolism (For individual ratios, see Figure 2.4). The total intron lengths were compared among the four species by Kruskal-Wallis test with post-test analysis by Dunn's multiple comparisons: NS, not significant ($P \geq 0.05$); $^{**}P < 0.01$; $^{***}P < 0.001$.

The variation in the ratios of total intron sequence per gene was quite striking, for both of the gene expression groups (Figure 2.4). For instance, depending on the gene, the ratios ranged from 0.2 to 10. This high level of heterogeneity in pairwise comparisons is likely to account for the lack of statistically significant differences. In addition, the intron length ratios were not consistent across species (Figure 2.4A and B).

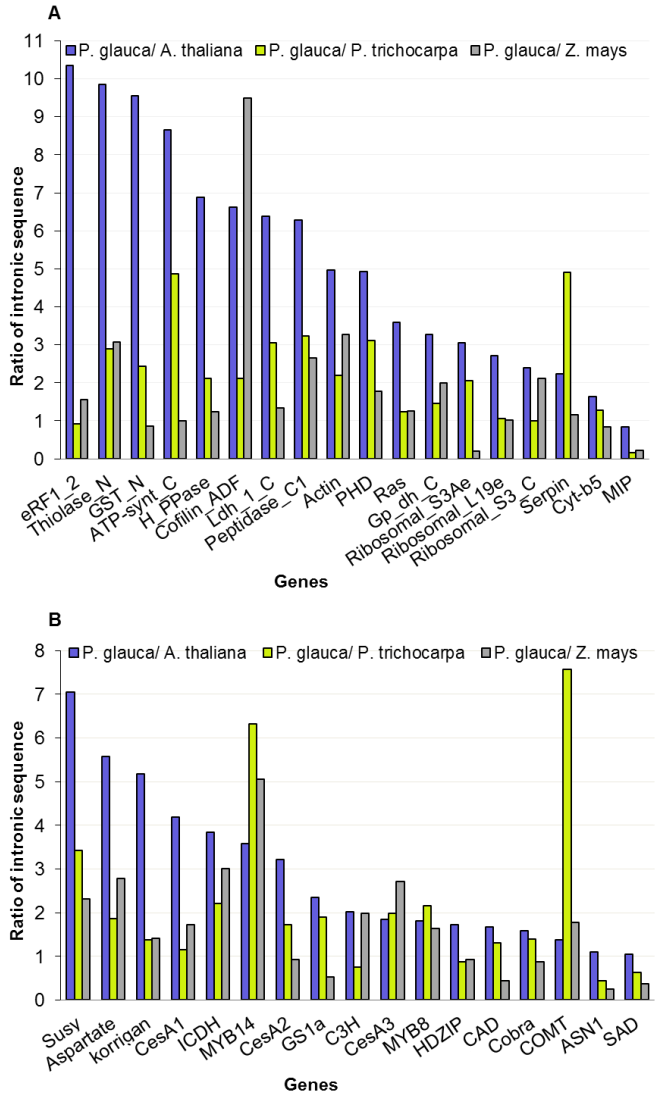


Figure 2.4- Gene by gene pair-wise comparisons of total length of intronic sequences in *P. glauca*, *A. thaliana*, *Populus trichocarpa* and *Z. mays.* (A) highly expressed genes and (B) genes associated with secondary cell-wall formation and nitrogen metabolism.

In this study, we show that much of the divergence in the total length of intron sequences per gene was related to a few long introns. Very long introns were observed in a few *P. glauca* genes such as PHD, Peptidase_C1 and Thiolase. Structure plots showed that introns in *A. thaliana* generally had uniform lengths whereas the other species had introns that were highly heterogeneous in length (Figure 2.5 and Supplementary information: Figure S2.3). While most of the *P. glauca* genes only had a few (1–3) very long introns (>1000

bp), gene sequences such as those for sucrose synthase (Susy) had many introns of moderate size (Figure 2.5). The longest introns in *P. glauca* were most often in a different position than long *Z. mays* and *P. trichocarpa* introns. In addition, we did not observe a trend of increased length in first introns in 5′ UTRs as reported for several eukaryotes [31], as the long introns in *P. glauca* appeared to be randomly distributed.

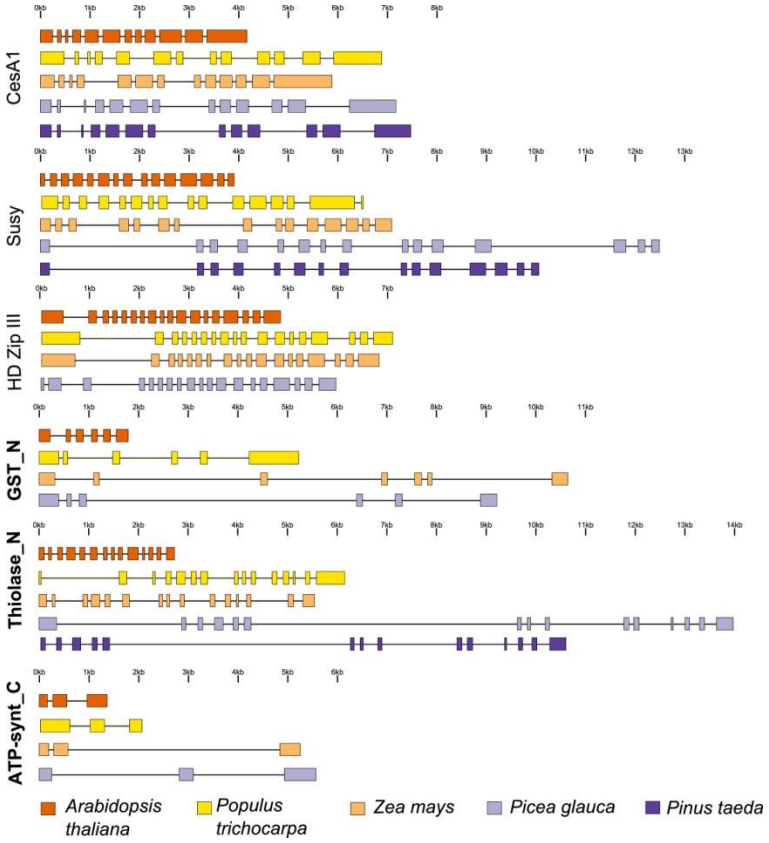


Figure 2.5- Gene structures of six genes from different angiosperm and gymnosperm species. The first three genes are associated with secondary cell-wall formation and nitrogen metabolism; and highly expressed genes are bolded.

2.4.4 Comparative analysis of gene structures between *Picea glauca* and *Pinus taeda*

A total of 23 different genes were submitted to pairwise comparisons between *Picea glauca and Pinus taeda*, which are both of the Pinaceae (for details, see Methods). A high level of similarity was observed for coding sequences (91% on average) indicating that they were

likely orthologous genes (Supplementary information: Table S2.4), and gene structure was conserved between the two conifers, with almost identical numbers of exons. The total intronic sequences per gene did not vary significantly at 3.13 and 3.17 Kbp for *P. glauca* and *P. taeda*, respectively (Supplementary information: Table S2.1). Pairwise comparison of introns indicated that the majority of individual introns were similar in length in the two species, despite the fact that the two genera diverged ca. 140 million years ago [32, 33] (Figure 2.5). Although these observations are based on a set of only 23 genes, they provide an indication that intron length is mostly conserved between these two conifer genera.

The 138 intron sequences of the 22 genes (PAL gene does not have introns) were aligned between spruce and pine; sequence similarity ranged quite broadly among homologous introns (Figure 2.6). We observed that highly conserved introns generally were short, and that longer introns had highly variable levels of sequence similarity, except for two introns that were both long and highly conserved.

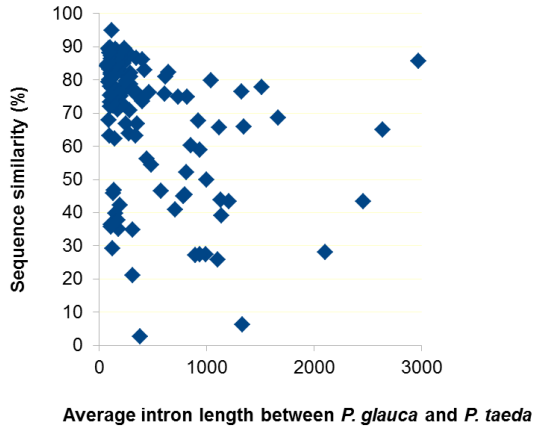


Figure 2.6 Relationship between intron size and sequence similarity of introns from *P. glauca* and *P. taeda*. A total of 138 introns were obtained from 22 genes and sequence alignments were produced with the Needle software (see Methods).

2.4.5 Repeat elements in *Picea glauca* genes

The possible origin of long introns as observed in conifer genomes was investigated by searching for the presence of repeated sequences including transposable elements.

First, the repetitive element content of the BACs was estimated based on a repetitive library constructed with *P. glauca* data (see Methods) as a baseline. It was 55% on average, but it varied considerably among the BAC clones, ranging from 18% to 83%. Supplementary information: Figure S2.1 shows that around half of repetitive sequences were classified as LTR-RT elements and the other half as unknown elements (without significant hits in Repbase and nr genbank).

We then analyzed the sequences of the 35 *P. glauca* genes described above including those identified in BACs, representing a total of 238 introns. The gene structures of these genes were screened for repeat elements using a *P. glauca* repeat library (see Methods). We found repetitive elements in 10 of the genes for a total of 24 unclassified fragments with no significant hits in RepBase; 22 of the fragments produced no hits in genbank and were 179 bp on average and only two had significant hits in nr genbank (Supplementary information: Table S2.8).

We also extended our analysis to include an additional set of genomic sequences obtained by targeted gene space sequencing based on sequence capture (see Methods, for details). Complete genomic sequences spanning the entire known mRNA sequence were recovered for 5970 complete genes, 1836 of which contained one or more introns. The different repetitive elements identified in introns and exons were then estimated. The proportion of genes harboring repetitive elements in their introns was 32.4% and was only 3.2% in exons. The repetitive elements represented 2.94% and 0.74% of the intronic and exonic sequences, respectively (Table 2.2). The repetitive sequences that were identified ranged from 31 to 1142 bp (median 117 bp) in exons and from 17 to 1189 bp (median 114bp) in introns. The unclassified elements were the most numerous, representing on average 80% of the hits in both introns and exons (Table 2.2). Class I LTR transposons were the most abundant group of classified repetitive elements and were only represented by incomplete elements. The

LTRs accounted for the higher repetitive element sequence representation in introns; however, on average, the sequences identified as *Copia* and *Gypsy* elements were longer in exons than in introns.

Table 2.2- Abundance of repetitive elements in *P. glauca* genes obtained from sequence capture

Class	Exons (%)	Introns (%)
Copia	0.09	0.24
Gypsy	0.09	0.19
LINE	0.03	0.15
UNK[1]	0.03	0.07
NHF[2]	0.49	2.29

A total of 5970 genes were analyzed, 1836 contained one or more introns.
[1] no significant hit in RepBase but significant hits in nr genbank.
[2] no significant hit in RepBase and nr genbank.

2.5 Discussion

This study reports on the detailed gene structure analysis of 35 genes from the conifer *Picea glauca* obtained from BAC sequencing and genome assembly. Recent analyses of the *Picea abies* and *Pinus taeda* genomes have analyzed individual introns and reported among the highest average intron lengths, the longest introns and highest average among long introns [9, 12]. We aimed to develop an understanding of the gene structure in conifers through a detailed analysis of entire genes taking into account gene expression profiles, with a particular emphasis on the potential impact of longer introns on gene structure through comparative analyses. Our findings were also derived from the analysis of nearly 6000 gene sequences obtained from sequence capture sequencing. We present an interpretation of our findings in regard to the evolution of gene structure.

2.5.1 Evolution of gene structure in plants

Analyses over a broad phylogenetic spectrum in eukaryotes showed that increases in genome size correlate with increases in the average intron length [2, 13]. A strong relationship between intron length and genome size was observed from studies in humans

and pufferfish [14], species of *Drosophilla* [15], and from studies of plants with small genomes [2, 13].

Our study compared the gene structure (introns and exons) of 35 homologous genes between four seed plant species with very different genome sizes. The conifer *P. glauca* has the largest genome with 19.8 Gbp [34]; among angiosperms, the monocot *Z. mays* has a genome of 2.3 Gbp [24], and dicots represent smaller plant genomes in this set, i.e. *P. trichocarpa* with genome of 484 Mbp [18] and *A. thaliana* with the smallest genome of 125 Mbp [19]. In the present study, the average exon length was similar between the four species, but the overall length of genes varied owing to longer introns in *P. glauca*, *P. trichocarpa* and *Z. mays*. For the set of sequences analyzed, *P. glauca* had 4.1 times more intron sequence per gene than *Arabidopsis*, 2.2 times more than poplar and 1.8 times more than maize (Figures 2.3 and 2.4); however, the statistical significance of these differences was variable.

The landscape of intron sizes in plants appears rather complex. A significant number of *Vitis vinifera* introns were shown to be uncommonly large for its genome size of 416 Mbp, compared to other plants [17]. In *Gossypium*, after multiple inferred rounds of genome expansion and contraction, intron size remained unchanged [16]. Such a pattern may be expected, given that genome size increase by polyploidy is sudden and fundamentally different than other types of genome size variation such as the gradual accumulation or loss of repeat elements over time. Taken together, observations from different plants indicated that events resulting in the expansion or contraction of intergenic regions are not clearly reflected by shifts in intron length. It thus appears that the evolution of intron length and genome size may be uncoupled in plants or alternatively, that the evolution of intron length is lineage specific (Figure 2.7).

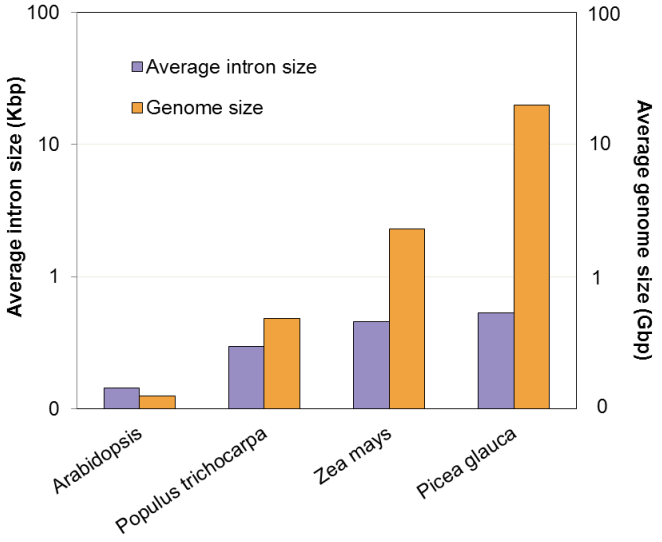


Figure 2.7- Variation in intron length and genome size in 35 target genes. Average intron size for *Arabidopis*, *P. trichocarpa*, *Z. mays* and *P. glauca* determined from the analysis of 35 homologous genes. Note that Y- axes are in log 10 scale.

Even though our study was based on 35 genes, our results are consistent with variations of intron size reported for *A. thaliana*, *P. trichocarpa* and *Z. mays* genomes [9, 12, 18, 19, 20]. We concluded that the increased intron length in *P. glauca, P. trichocarpa* and *Z. mays* was heterogeneous compared to *A. thaliana*. Even in genes with many introns, only a few introns were very long, whereas in *Arabidopsis*, genes exhibited a more uniform intron length, suggesting that intron expansion or contraction within a gene may be independent across species.

Comparisons between the *A. thaliana* (125 Mbp) and *A. lyrata* (~200 Mbp) genomes, which diverged about 10 million years ago, showed that most of the difference in genome size was due to hundreds of thousands of small deletions, mostly in noncoding DNA [35]. The authors concluded that evolution toward genome compaction is occurring in *Arabidopsis*. Conifers such as species of *Picea* and *Pinus* have large amounts of repetitive elements in intergenic regions and apparently more intronic sequence per gene in comparison to many angiosperms. Our results do not reveal whether the *P. glauca* genome and introns are expanding, or alternatively evolving at slower pace, than other plant

genomes which are contracting. Some evidence like the presence of very ancient retrotransposon elements [9, 36] and the lack of gene rearrangements since before their split from extant angiosperms [8] lend credence to the paradigm that conifer genomes are slowly evolving.

2.5.2 Repetitive sequences in gene evolution

Transposable elements play a role in plant genes as was shown by the abundance of TE-gene chimeras in *Arabidopsis* which was reported as 7.8% of expressed genes [37]. The abundance of TEs may be especially high in long introns as recently shown in *Picea abies* where most of the introns were longer than 5 Kbp, representing 5% of the total intron count [9]. This trend was also observed in other repeat-rich genomes as *V. vinifera* and *Z. mays* [20, 21, 38].

We isolated *P. glauca* BAC clones each containing a different complete transcription unit for 21 target genes. In each of the BACs (average 131 Kb), only one intact gene sequence was identified, which is indicative of large intergenic regions as reported for other conifers [39, 40, 41]. Previous studies on conifer trees have considered only two targeted genes (from terpenoid biosynthesis) isolated from *P. glauca* BAC clones [40] and only a few other intact genes with complete coding sequence isolated from BACs in pines [7, 39, 41].

Complete sequencing of the *P. glauca* BACs showed that the repetitive element content is not distributed uniformly in proximal intergenic regions, as indicated by the variable proportion of repetitive elements among the different BACs. A study in 10 *P. taeda* BACs, sequences similar to eukaryote repeat elements (according to Repbase) represented 23% of the sequence on average, and ranged from 19% to 33% [7]. In *P. glauca*, 26% of BAC sequences were classified as LTR-RT repetitive elements on average and ranged from 8% to 47%, while *P. taeda* had an average of 18.8% of LTR-RT [7]. Furthermore, an average 26% of the *P. glauca* BAC sequences were unknown repeat elements. Results in spruce and pine indicate a relatively low abundance of TEs in gene-proximal sequences compared to whole genomes, at 70% in the *Picea abies* genome [9] and around 80% in *Pinus taeda* [12].

Picea and *Pinus* genomes are reported to have among the highest average for the longest intron per gene, when compared to angiosperms of diverse genome sizes [9]. We verified whether insertions of repetitive elements could be responsible for the length of introns in *P. glauca* in a set of more than 1800 genes sequences, and found that genes harboring repetitive elements in introns were 10 times more frequent than genes harboring repetitive elements in exons, i.e. 29.8% vs 3.2%. The vast majority of the repetitive elements were short fragments, suggesting that they were remnants or fragments of TE insertions that have not persisted and could represent ancient insertion events. Importantly, interpretation of our findings in *P. glauca* must take into account the fact that the sequences were derived from a sequence capture study and that nearly all of the introns in the data set were <1 Kbp. Thus we show that TE sequences are ubiquitous even in genes that do not harbor long introns, suggesting that their presence has been very widespread during the evolution of conifer genes. An analysis of intact LTR TEs in *Picea* genomic sequences showed that most insertions date back to 10 MYA or more, with a maximum around 20–25 MYA [9]. The TE remnants that we detected in *P. glauca* indicate that many genes introns contained TE in a more or less distant past. In this report and in recent analyses of conifer genomes, an emphasis has been placed on long introns; however the median intron length in conifers is very similar to other plant species, most of which have a median between 100 bp and 200 bp. Therefore our findings on intron are relevant for a large majority of introns rather than a small fraction represented by large or very large introns.

2.5.3 Slow evolution of conifer genes

Analyses of the gene structure of 23 orthologous genes between *P. glauca* and *P. taeda* clearly showed the conservation of gene structure and the distribution of intron sizes in spite of a divergence time of 100 to 140 MYA [32, 33]. The conservation of long introns was also observed across gymnosperm taxa, where a group of long introns in *P. abies* was identified as orthologous to long introns in *P. sylvestris* and *Gnetum gnemon* [9]. We suggest that the long introns observed in *P. glauca* likely date back to a period predating the divergence of major conifer groups. As more conifer genomes become available

[9, 10, 11] and assembly contiguities are improved it will be possible to extend this analysis of orthologous gene structures among conifers.

We also observed that the sequence of many introns was highly similar between spruce and pine, and that shorter introns were more conserved on average. Between humans and chimpanzee, a strong positive correlation was found between intron length and divergence [42]. The pattern found in conifers as well as observations in primates lead to the hypothesis that shorter introns could be under stronger selection pressure than longer introns, which could be explained by factors such as the maintenance of functional regulatory elements in shorter introns or impacts on RNA transcript processing and stability. In our analysis of sequence similarity between *Picea* and *Pinus*, 20 of the introns were longer than 1 Kbp and only two of them had high sequence similarity. Future studies with more long introns are required to confirm the hypothesis that shorter introns are more conserved in conifers. Despite the fact that introns are non-protein-coding sequences by definition, conserved introns may play a functional role related to gene expression.

2.5.4 Costs and benefits associated with intron size

There is also considerable debate about other factors that may impact the evolution of introns, aside from transposable elements. Lynch [43] stated that the reduced efficiency of selection in regions of low recombination may lead to an increase in intron size if small introns provide a slightly improved transcription efficiency or splicing accuracy. On the other hand, Comeron and Kreitman [3] proposed that there might be situations in which a longer intron is selectively advantageous as an explanation for intron persistence and increased lengths. If so, there would be indirect selection for large introns in regions of low recombination because they can reduce the load caused by deleterious mutations by increasing the recombination rate. It was proposed that conifers have low recombination rates at both the genome and within-gene scales [44]. Their low recombination rates may explain at least in part, the accumulation of longer introns.

The high degree of sequence conservation that we observed in short introns between spruce and pine may also depend on the recombination rate within genes, where small introns

would be under stronger selection because of efficiency in transcription and splicing, and long introns in regions of low recombination diverge because of reduced selection pressure. Another factor underlying the evolution of intron size is that intron length would be constrained by energy use during transcription, given that large introns represent a higher cost of transcription, the so-called "economy" or low-cost transcription hypothesis [4]. In the present study, the 35 *P. glauca* genes analyzed were divided in two groups based on their expression profiles, i.e. 17 genes associated with secondary cell-wall formation or nitrogen metabolism, many of which had tissue preferential expression, and 18 genes that were highly transcribed in a large range of tissues (based on Raherison et al. [30]). The highly expressed genes had more intronic sequences per gene on average than the more specialized subset of genes (4,182 bp versus 3,013 bp). We also observed a large variation among genes in each group, i.e. from 446 to 12,009 bp in highly transcribed genes and 440 to 9,847 bp in the set of more specialized genes. These observations do not support the "economy" hypothesis in *P. glauca* as there appears to be no clear rule governing the relationship between intron size and expression levels or profiles. In humans, genes contained total intronic sequences are ~5,500 bp per gene on average [45], which more than any plant described so far. It was observed that intron length declines steadily as the expression level increases in humans, in agreement with the low-cost transcription hypothesis [4]. Considering the smaller amount of intron sequences in plant genes including conifers compared to humans, it may be that the economy rule does not impact their introns as strongly as in vertebrates and that other evolutionary forces are main drivers of intron size evolution. This interpretation is consistent with the findings reported for the *P. abies* genome [9]. Future studies with more genes are needed to confirm this hypothesis.

2.6 Conclusions

Our results indicate that *P. glauca* has longer introns than *Arabidopsis*, *P. trichocarpa* and *Z. mays* on average due the presence of few long introns. Intron size and the position of long introns within genes were variable between plant lineages but well conserved in conifers. Our findings are consistent with recent reports indicating that conifers accumulate very long introns but we point out that long introns represent a relatively small fraction of

the overall intronic content, which is reflected by the median length of similar size to other plants. We show that RE sequences are detected at a high frequency (32%) even in introns <1 Kbp, indicating their ubiquitous presence in conifer genes over the course of evolution.

Taken together, our observations and the recent literature suggest that the evolution of plant gene structure is determined by more interacting forces than classically expected. The pattern is reminiscent of the heterogeneity of rates of evolution at the genetic, genomic and morphological levels seen among seed plants including angiosperms, conifers, annual and perennial taxa. It stands to reason that the distinctive features of the conifer genome, such as its large size and relatively small occupancy of the gene space, its conserved macro-structure, the large numbers of repetitive elements, and long introns, represent the product of the intricate evolutionary history of conifers.

2.7 Methods

2.7.1 *Picea glauca* BAC isolation and validation

A BAC library developed from the single *Picea glauca* (Moench) Voss individual PG29 from the BC Ministry of Forests was utilized. The non-arrayed library consisted of approximately 1.1 million BAC clones with an average insert size of 140 kbp, representing approximately 3× coverage of the *P. glauca* genome [40]. The library was screened by quantitative PCR (qPCR) through successive steps using BAC super-pools and pools, serial dilutions and clone identity verification by amplicon sequencing. The BAC isolation and sequencing are reported here for the first time.

We isolated 21 BAC clones each containing a different single-copy gene from *P. glauca* (see list of genes and accession numbers in Supplementary information: Table S2.2). Each of the genes screened was represented by a unique FL-cDNA clone in *P. glauca* as described in Rigault et al. [29]. The selected genes encoded enzymes and transcriptional regulators involved in secondary cell-wall formation and nitrogen metabolism and were subject to manual curation. They were chosen as to facilitate comparison with BAC isolation studies conducted in other conifers species (e.g. [21]). Two

sets of gene-specific primers were designed for each gene based on the cDNA sequence available in the *P. glauca* gene catalogue [29]. The genomic sequence obtained was used to design two additional primers such that two small amplicons of 120–200 bp could be amplified by quantitative PCR (qPCR) (Supplementary information: Table S2.3). All of the primer sets were verified by PCR and qPCR using the genomic DNA from *P. glauca,* genotype PG-653, and then they were used to screen the BAC library in three steps. See PCR conditions in Supplementary information: additional experimental procedures.

The BAC library was subdivided into pools with a titer 1000 BACs on average, which were arrayed into ten 96 deep-well plates. Each plate was inoculated in 96-well culture plates with 1 ml of terrific broth (TB) and 20 μg/mL of chloramphenicol and grown in a 37°C shaker at 300 rpm overnight. The same TB medium and growing conditions were utilized to culture bacteria throughout the screening steps. Bacterial cultures from each of the columns and rows within a plate were combined in a total of 200 super-pools for DNA isolation as described in Osoegawa et al. [46].

The first step followed Jeukens et al. [47]. Briefly, the super-pool DNA was amplified by the two small amplicons for each target gene by qPCR using QuantiTect SYBR Green master mix as described in Boyle et al. [48]. The intersection of a positive row and a positive column was indicative of positive wells on the original plate. The presence of target genes in the positive super-pools was verified by qPCR in 30 μL reactions using QuantiTect SYBR Green master [48]. We performed PCR of the long amplicon and its purification for gene sequence validation by Sanger sequencing (Supplementary information: Table S2.3). The second step of the screening relied on serial dilutions of the super-pools to inoculate 50 bacteria from the positive well in a 96 deep-well plate. DNA super-pools were extracted and screened by qPCR using the same conditions as in the first step. Then, we extracted DNA from 1 μL of bacterial culture from each well of a positive column to test it by qPCR and determine the positive well in the column. From the positive well of the same bacterial culture plate, we proceeded with serial dilutions and we inoculated a 96 deep-well plate with one isolated colony per well. The third step of the screening consisted to pool columns and rows of bacterial cultures. We identified positive

wells by qPCR and plated the culture of each positive well on a different Petri dish and one colony per dish was inoculated in 5 mL TB with chloramphenicol. DNA was extracted from 2 mL of each culture. Positive isolated clones were validated by PCR, qPCR and resequencing of the long amplicon. The validation steps to confirm gene identity and integrity proved essential such as conifers contain many pseudogenes that reduce the efficiency of targeted BAC isolation [39].

The 21 isolated BAC clones, each identified by screening for a different gene (for accession numbers, see Supplementary information: Table S2.2) were sequenced by Roche 454 FLX pyrosequencing at McGill University and Genome Québec Innovation Centre, Montreal, Canada. Sequences were assembled *de novo* into contigs using the GS *De novo* Assembler module of Newbler version 2.3 (Roche) [49]. In this analysis, the BAC vector and *E. coli* genome were trimmed and the assembly parameters were a minimum overlap of 200 bp, minimum overlap identity of 98% and minimum contig length of 500 bp. In general, more than one contig per BAC was obtained; therefore, the order of the contigs within each BAC was tested by PCR.

To determine gene structure (introns and exons), cDNAs were mapped onto the BAC contigs containing the respective gene using est2genome incorporated in the annotation software MAKER [50]. Four of the genes were eliminated from the comparative gene structure analyses because they were either incomplete, lacked introns or identifiable homologs in the species targeted for comparative analyses.

2.7.2 *Pinus taeda* orthologous sequences

Seven BAC clones of *Pinus taeda* containing orthologs of *P. glauca* genes were identified by BLAST [51] using an e-value threshold of 1e-20 and sequence identity > 90% (Supplementary information: Table S2.4). An additional 16 sequences were identified by BLAST [51] using an e-value threshold of 1e-20 in the whole genome shotgun assembly of *Pinus taeda* [11]. Their gene structures were defined using est2genome [50] and *P. taeda* cDNA or *P. glauca* cDNA when *P. taeda* complete cDNA was not available. Accession numbers available in Supplementary information: Table S2.4. A pairwise

alignment of all corresponding intron and exon sequences of orthologous genes between *P. glauca* and *P. taeda* was conducted, followed by the estimation of their similarity with the software Needle, part of the analysis package EMBOSS [52]. The BAC clones containing the LIM gene in *P. glauca* and the Korrigan, Peptidase_C, Thiolase, Gp_dh_C and eRF1_2 genes in *P. taeda* lacked an intron and exon; these missing exons and introns were excluded from the comparison between *P. glauca* and *P. taeda*.

2.7.3 Screening for highly expressed genes in the whole genome shotgun assembly

Based on transcript profiles (PiceaGenExpress database [30]) a set of 500 gene sequences each representing a unique FL-cDNA clone that was highly expressed in all tissues, was identified from the *P. glauca* gene catalog [29]. A preliminary assembly of the *P. glauca* genome assembly described by Birol et al. [10] was screened with each of these sequences. The screening was performed using exonerate and the est2genome model [53], which considers intron/exon boundaries. The cDNA/genome alignments were further filtered based on the identity and length coverage to retain only complete alignments with entire cDNAs; i.e., genes with complete genomic sequence. We randomly selected 18 the genomic sequences thus identified as containing complete structures of highly expressed genes. As for genes contained the BACs, gene annotations were generated automatically and curated manually individual reciprocal BLAST and sequence alignments.

2.7.4 Identification of closest homologs in angiosperms

Homologous sequences to *P. glauca* genes were identified in *Arabidopsis thaliana and P. trichocarpa* using BLASTX [51] with a threshold e-value of 1e-10. Reciprocal analysis (BLASTX) between the *A. thaliana* and *P. trichocarpa* sequences and the *P. glauca* gene catalogue was used to verify that the genes were the closest homologs among known sequences. In *Zea mays*, the closest homolog was identified based on *A. thaliana* sequences using BLASTX, with a threshold e-value of 1e-10, and orthology was verified in the Maize Genome Project Sequencing database (Supplementary information: Table S2.5). We also performed a BLASTX of *P. glauca* against *Z. mays* sequences and we verified that the closest homologs identified between *Z. mays* and *A. thaliana* were among the best hits.

Gene structures were recovered from the following databases: TAIR 10 [54], Phytozome (JGI v3.0 gene annotation of assembly v3 of *P. trichocarpa*) [18, 55] and the Maize Genome Sequencing Project [56]. Accession numbers are available in Supplementary information: Table S2.5.

From the 21 genes contained the *P. glauca* BAC clones, four genes were eliminated from the gene structure comparative analysis between *P. glauca*, *A. thaliana*, *P. trichocarpa* and *Z. mays*: (1) Dof5 because a clear A. *thaliana* homolog could not be identified, (2) asparaginase because a clear *Z. mays* homolog could not be identified, (3) PAL because it lacked introns and (4) LIM because it presented an incomplete sequence (cDNA). A total of 35 genes had their gene structure compared with closest homologs in angiosperms: 17 genes related to secondary cell wall formation and nitrogen metabolism and 18 highly transcribed genes with little tissue-specific expression.

2.7.5 Statistical analyses of introns

Intron lengths were compared between *P. glauca*, *A. thaliana*, *P. trichocarpa* and *Z. mays* by nonparametric Kruskal-Wallis tests with post-test analysis by Dunn's multiple comparisons, because intron length did not follow a Gaussian distribution. Comparisons of two groups of genes by Wilcoxon rank sum test with continuity correction; was used to compare total intron sequences in *P. glauca* genes belonging to the two expression groups and to compare *P. glauca* with *P. taeda*. Data analyses were performed using the R packages coin and multcomp [57, 58, 59].

2.7.6 Gene space obtained from sequence capture technology

Sequences were obtained by using genomic DNA hybridizations on a custom *P. glauca* chip containing oligonucleotide baits for 23,864 genes. The method development, the DNA sequence isolation and analysis procedure and the resulting sequence data are reported in this manuscript for the first time.

DNA was extracted from needles of the *P. glauca* individual 77111 from the Canadian Forest Service as described in Pelgas et al. [60] using the DNeasy Plant mini kit according

to the manufacturer's instructions (QIAGEN). One microgram of high quality DNA was used to prepare a GS-FLX rapid library according to the manufacturer's instructions (Roche). The library was amplified by ligation-mediated PCR using 454 A and B primers as described in the NimbleGen SeqCap EZ Library LR User's guide.

Custom probes were designed by Nimblegen based on the cDNA sequences and ESTs from the *P. glauca* gene catalogue [29]. We used a Newbler (gsAssembler module v2.5.3) assembly of sequences from random genomic sequencing (0.15× of coverage) from *P. glauca*[29] to identify highly repetitive elements and to filter out probes representing such elements as they were expected to reduce the efficiency of the sequence capture. Next, a comparative genomic hybridization (Array CGH) experiment was conducted in collaboration with Nimblegen (Madison, WI, USA) to eliminate probes with abnormally high levels of hybridization that could not be identified with *in silico* approaches. Throughout the process, probes within genes harboring abnormally high capture levels were eliminated. The final design covered 23,864 genes. The target enrichment procedures including quantitative PCR assessments are described in Supplementary information: Additional experimental procedures.

Emulsion PCR and GS-FLX Titanium sequencing was performed according to the manufacturer's instructions at the Plateforme d'Analyses Génomiques of the Institut de Biologie Intégrative et des Systèmes (Université Laval, Quebec, Canada). Raw sequencing reads were *de novo* assembled using the gsAssembler module of Newbler v2.5.3. Contigs were screened for complete gene structures based on the *P. glauca* gene catalogue [29]. Technical details are available in Supplementary information: Additional experimental procedures.

2.7.7 *Picea glauca* repetitive library and identification of repeat elements

For repeat identification, a random sample of 100,000 *P. glauca* 454 reads from randomly sheared DNA was searched *de novo* for repeats using the software RepeatScout [61]. The results were filtered by removing low complexity sequences and sequences shorter than 100 nt, and retaining only repeats having at least 10 matches when mapped onto the original

454 set using RepeatMasker [62]. Since RepeatScout is tailored to analyze complete genomes or at least large scaffolds, its output is usually fragmented when the program is run on random sheared reads. In order to reduce fragmentation, we merged the repeats belonging to the same element running cap3 [63] under relaxed settings (-o 30, -p 80, -s 500) on the RepeatScout output. Finally, the entire set of repeated sequences was clustered using the software cd-hit-est [64] by collapsing all the repeats sharing at least 80% similarity in order to remove redundancies.

Repeat characterization proceeded by similarity searches were used to associate candidate repeats to known TE families and to remove repeats showing similarity to gene sequences and being possibly part of gene families. In particular, the repeat candidates from each species were searched against RepBase [65] using TBLASTX [57] and setting as significance threshold an e-value of 1e-5. Repeats that did not provide significant hits were used as queries in BLASTX searches against the non-redundant (nr) division of Genbank. Those having significant hits with genes were removed from the library while those having significant hits with TEs were labeled accordingly and the remaining repeats were considered as unclassified.

The search for repeat elements in all BAC contigs and in the gene space obtained from sequence capture was conducted using RepeatMasker [62] using the *Picea glauca* repetitive library and default parameters.

2.8 Acknowledgements

The authors thank D. Peterson (Mississippi State Univ., USA) and G. Claros and F. Canovas (Univ. de Málaga, Spain) for sharing information on gene targets and strategies for BAC isolation in pines. Technical assistance of S. Caron, É. Fortin, G. Tessier (Univ. Laval, Canada) is acknowledged for BAC screening. F. Belzile, R. Lévesque, L. Bernatchez (Univ. Laval, Canada) are acknowledged for valuable discussions and suggestions at the project planning stage. Funding for the project was received from Génome Québec for a Genome exploration grant (JM, JBou, PR, KR), from Genome Canada, Génome Québec and Genome British Columbia for the SmarTForests project (JM, JBoh, JBou, IB, KR, SJ) and NSERC of Canada (JM). JS received partial funding from Univ. Laval.

2.9 References

1. Lynch M, Conery JS: The origins of genome complexity. *Science* 2003, 302:1401–1404.

2. Deutsch M, Long M: Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res* 1999, 27:3219–3228.

3. Comeron JM, Kreitman M: The correlation between intron length and recombination in *Drosophila*: dynamic equilibrium between mutational and selective forces. *Genetics* 2000, 156:1175–1190.

4. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA: Selection for short introns in highly expressed genes. *Nat Genet* 2002, 31:415–418.

5. Murray BG, Leitch IJ, Bennett MD: *Gymnosperm DNA C-values database*. ; 2004. http://www.kew.org/cvalues/.

6. Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, Garcia SA, Kubisiak TL, Amerson HV, Carlson JE, Nelson CD, Davis JM: Evolution of genome size and complexity in *Pinus*. *PLoS One* 2009, 4:e4332.

7. Magbanua ZV, Ozkan S, Bartlett BD, Chouvarine P, Saski CA, Liston A, Cronn RC, Nelson CD, Peterson DG: Adventures in the enormous: A 1.8 million clone BAC library for the 21.7 Gb genome of loblolly pine. *PLoS One* 2011, 6:e16214.

8. Pavy N, Pelgas B, Laroche J, Rigault P, Isabel N, Bousquet J: A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. *BMC Biol* 2012, 10:84.

9. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hällman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Käller M, Luthman J, Lysholm F, Niittylä T, Olson Å, Rilakovic N, Ritland C, Rosselló JA, Sena J, *et al*: The Norway spruce genome sequence and conifer genome evolution. *Nature* 2013, 497:579–584.

10. Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Yuen MMS, Keeling CI, Brand D, Vandervalk BP, Kirk H, Pandoh P, Moore RA, Zhao Y, Mungall AJ, Jaquish B, Yanchuk A, Ritland C, Boyle B, Bousquet J, Ritland K, Mackay J, Bohlmann J, Jones SJM: Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics* 2013, 29:1492–1497.

11. Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, Martínez-García PJ, Vasquez-Gross HA, Lin BY, Zieve JJ, Dougherty WM, Fuentes-Soriano S, Wu L-S, Gilbert D, Marçais G, Roberts M, Holt C, Yandell M, Davis JM, Smith KE, Dean JF, Lorenz WW, Whetten RW, Sederoff R, Wheeler N, McGuire PE, *et al*: Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol* 2014, 15:R59.

12. Wegrzyn JL, Liechty JD, Stevens KA, Wu L-S, Loopstra CA, Vasquez-Gross HA, Dougherty WM, Lin BY, Zieve JJ, Martínez-García PJ, Holt C, Yandell M, Zimin AV, Yorke JA, Crepeau MW, Puiu D, Salzberg SL, Dejong PJ, Mockaitis K, Main D, Langley CH, Neale DB: Unique features of the Loblolly Pine (*Pinus taeda L.*) Megagenome revealed through sequence annotation. *Genetics* 2014, 196:891–909.

13. Vinogradov AE: Intron–genome size relationship on a large evolutionary scale. *J Mol Evol* 1999, 49:376–384.

14. McLysaght A, Enright AJ, Skrabanek L, Wolfe KH: Estimation of synteny conservation and genome compaction between pufferfish (*Fugu*) and human. *Yeast* 2000, 17:22–36.

15. Moriyama EN, Petrov DA, Hartl DL: Genome size and intron size in *Drosophila*. *Mol Biol Evol* 1998, 15:770–773.

16. Wendel JF, Cronn RC, Alvarez I, Liu B, Small RL, Senchina DS: Intron size and genome size in plants. *Mol Biol Evol* 2002, 19:2346–2352.

17. Jiang K, Goertzen LR: Spliceosomal intron size expansion in domesticated grapevine (*Vitis vinifera*). *BMC Res Notes* 2011, 4:52.

18. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen G-L, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, *et al*: The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 2006, 313:1596–1604.

19. Arabidopsis Genome Initiative: Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000, 408:796–815.

20. Haberer G, Young S, Bharti AK, Gundlach H, Raymond C, Fuks G, Butler E, Wing RA, Rounsley S, Birren B, Nusbaum C, Mayer KFX, Messing J: Structure and architecture of the maize genome. *Plant Physiol* 2005, 139:1612–1624.

21. Ren X-Y, Vorst O, Fiers MWEJ, Stiekema WJ, Nap J-P: In plants, highly expressed genes are the least compact. *Trends Genet* 2006, 22:528–532.

22. Kumar A, Bennetzen JL: Plant retrotransposons. *Annu Rev Genet* 1999, 33:479–532.

23. Feschotte C, Jiang N, Wessler SR: Plant transposable elements: where genetics meets genomics. *Nat Rev Genet* 2002, 3:329–341.

24. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, *et al*: The B73 maize genome: complexity, diversity, and dynamics. *Science* 2009, 326:1112–1115.

25. Ralph SG, Chun HJ, Kolosova N, Cooper D, Oddy C, Ritland CE, Kirkpatrick R, Moore R, Barber S, Holt RA, Jones SJ, Marra MA, Douglas CJ, Ritland K, Bohlmann J: A conifer genomics resource of 200,000 spruce (*Picea spp*.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (*Picea sitchensis*). *BMC Genomics* 2008, 9:484.

26. Bedon F, Grima-Pettenati J, Mackay J: Conifer R2R3-MYB transcription factors: sequence analyses and gene expression in wood-forming tissues of white spruce (*Picea glauca*). *BMC Plant Biol* 2007, 7:17.

27. Cañas RA, de la Torre F, Cánovas FM, Cantón FR: High levels of asparagine synthetase in hypocotyls of pine seedlings suggest a role of the enzyme in re-allocation of seed-stored nitrogen. *Planta* 2006, 224:83–95.

28. Nairn CJ, Lennon DM, Wood-Jones A, Nairn AV, Dean JFD: Carbohydrate-related genes and cell wall biosynthesis in vascular tissues of loblolly pine (*Pinus taeda*). *Tree Physiol* 2008, 28:1099–1110.

29. Rigault P, Boyle B, Lepage P, Cooke JEK, Bousquet J, MacKay JJ: A white spruce gene catalog for conifer genome analyses. *Plant Physiol* 2011, 157:14–28.

30. Raherison E, Rigault P, Caron S, Poulin P-L, Boyle B, Verta J-P, Giguère I, Bomal C, Bohlmann J, MacKay J: Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression. *BMC Genomics* 2012, 13:434.

31. Bradnam KR, Korf I: Longer first introns are a general property of eukaryotic gene structure. *PLoS One* 2008, 3:e3093.

32. Savard L, Li P, Strauss SH, Chase MW, Michaud M, Bousquet J: Chloroplast and nuclear gene sequences indicate late Pennsylvanian time for the last common ancestor of extant seed plants. *Proc Natl Acad Sci U S A* 1994, 91:5163–5167.

33. Wang X-Q, Tank DC, Sang T: Phylogeny and divergence times in Pinaceae: evidence from three genomes. *Mol Biol Evol* 2000, 17:773–781.

34. Ohri D, Khoshoo TN: Genome size in gymnosperms. *Plandt Syst Evol* 1986, 153:119–132.

35. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottilar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KFX, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D, Guo Y-L: The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet* 2011, 43:476–481.

36. Morgante M, De Poali E: Toward the conifer genome sequence. In *Genetics, Genomics and Breeding of Conifers Trees*. Edited by Plomion C, Bousquet J, Kole C. Enfield: Science Publishers; 2011:389–403.

37. Lockton S, Gaut BS: The contribution of transposable elements to expressed coding sequence in *Arabidopsis thaliana*. *J Mol Evol* 2009, 68:80–89.

38. Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Fabbro CD, Alaux M, Gaspero GD, Dumas V, *et al*: The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 2007, 449:463–467.

39. Kovach A, Wegrzyn JL, Parra G, Holt C, Bruening GE, Loopstra CA, Hartigan J, Yandell M, Langley CH, Korf I, Neale DB: The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. *BMC Genomics* 2010, 11:420.

40. Hamberger B, Hall D, Yuen M, Oddy C, Hamberger B, Keeling CI, Ritland C, Ritland K, Bohlmann J: Targeted isolation, sequence assembly and characterization of two white spruce (*Picea glauca*) BAC clones for terpenoid synthase and cytochrome P450 genes involved in conifer defence reveal insights into a conifer genome. *BMC Plant Biol* 2009, 9:106.

41. Bautista R, Villalobos DP, Díaz-Moreno S, Cantón FR, Cánovas FM, Claros MG: Toward a *Pinus pinaster* bacterial artificial chromosome library. *Ann For Sci* 2007, 64:855–864.

42. Gazave E, Marqués-Bonet T, Fernando O, Charlesworth B, Navarro A: Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biol* 2007, 8:R21.

43. Lynch M: Intron evolution as a population-genetic process. *Proc Natl Acad Sci U S A* 2002, 99:6118–6123.

44. Jaramillo-Correa JP, Verdú M, González-Martínez SC: The contribution of recombination to heterozygosity differs among plant evolutionary lineages and life-forms. *BMC Evol Biol* 2010, 10:22.

45. Sakharkar MK, Chow VTK, Kangueane P: Distributions of exons and introns in the human genome. *In Silico Biol (Gedrukt)* 2004, 4:387–393.

46. Osoegawa K, de Jong PJ, Frengen E, Ioannou PA: Construction of bacterial artificial chromosome (BAC/PAC) libraries. In *Current Protocols in Molecular Biology*. Edited by Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K. Hoboken, NJ, USA: John Wiley & Sons Inc; 2001.

47. Jeukens J, Boyle B, Kukavica-Ibrulj I, St-Cyr J, Lévesque RC, Bernatchez L: BAC library construction, screening and clone sequencing of lake whitefish (*Coregonus clupeaformis*, Salmonidae) towards the elucidation of adaptive species divergence. *Mol Ecol Resour* 2011, 11:541–549.

48. Boyle B, Dallaire N, MacKay J: Evaluation of the impact of single nucleotide polymorphisms and primer mismatches on quantitative PCR. *BMC Biotech* 2009, 9:75.

49. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Goodwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, *et al*: Genome sequencing in open microfabricated high density picoliter reactors. *Nature* 2005, 437:376–380.

50. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M: MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 2008, 18:188–196.

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

52. Rice P, Longden I, Bleasby A: EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000, 16:276–277.

53. Slater GS, Birney E: Automated generation of heuristics for biological sequence comparison. *BMC Bioinforma* 2005, 6:31.

54. *The Arabidopsis Information Resource.* http://arabidopsis.org.55. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS: Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 2012, 40(D1):D1178–D1186.

56. *Maize Genome Sequencing Project.* http://www.maizesequence.org.

57. Hothorn T, Bretz F, Westfall P: Simultaneous inference in general parametric models. *Biom J* 2008, 50:346–363.

58. Hothorn T, Hornik K, van de Wiel MA, Zeileis A: Implementing a class of permutation tests: the coin package. *J Stat Softw*, 28(8):1–23. URL http://www.jstatsoft.org/v28/i08/.

59. *R project.* http://www.r-project.org.

60. Pelgas B, Bousquet J, Meirmans PG, Ritland K, Isabel N: QTL mapping in white spruce: gene maps and genomic regions underlying adaptive traits across pedigrees, years and environments. *BMC Genomics* 2011, 12:145.

61. Price AL, Jones NC, Pevzner PA: De novo identification of repeat families in large genomes. *Bioinformatics* 2005, 21(Suppl 1):i351–i358.

62. Smit AFA, Hubley R, Green P: *RepeatMasker Open-3.0.* 1996–2010 http://www.repeatmasker.org.

63. Huang X, Madan A: CAP3: a DNA sequence assembly program. *Genome Res* 1999, 9:868–877.

64. Huang Y, Niu B, Gao Y, Fu L, Li W: CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010, 26:680–682.

65. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 2005, 110:462–467.

66. Haun WJ, Hyten DL, Xu WW, Gerhardt DJ, Albert TJ, Richmond T, Jeddeloh JA, Jia G, Springer NM, Vance CP, Stupar RM: The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. Plant Physiol 2011, 155:645–655.

67. Bolon Y-T, Haun WJ, Xu WW, Grant D, Stacey MG, Nelson RT, Gerhardt DJ, Jeddeloh JA, Stacey G, Muehlbauer GJ, Orf JH, Naeve SL, Stupar RM, Vance CP: Phenotypic and genomic analyses of a fast neutron mutant population resource in soybean. Plant Physiol 2011, 156:240–253.

2.10 Supplementary information

2.10.1 Additional experimental procedures for BAC isolation and sequence capture.

PCR conditions for *Picea glauca* BAC isolation and validation

The first set of primers was designed so as to obtain one long amplicon of 500-1000 bp by PCR. PCRs of the long amplicon were carried out in 50 µL reactions using Platinum ® Taq DNA polymerase High Fidelity (Life Technologies, Carlsbad, CA, USA) 0.2µM of each primer, 2 µl of genomic DNA] and carried out following the following conditions: 5 min activation at 95 °C followed by 35 cycles consisting of 15 s at 95 °C, 1 min at 62 °C and 1 min at 68°C; to finish 5 min at 68°C. The PCR products were purified using QIAquick PCR purification kit (Qiagen, Germantown, MD, USA) following manufacturer's instructions. The primers sequences are available in Supplementary information: Table S2.3.

Gene space obtained from sequence capture technology

Genomic DNA hybridization and target enrichment

Target enrichment was performed by using 2.1M developer arrays (Roche Nimblegen, Madison, WI, USA) with the SeqCap hybridization and wash kit (Roche Nimblegen, Madison, WI, USA) following the guidelines in the Sequence Capture Array delivery user's guide with the exception that we used a plant capture enhancer as previously described [66,67]. Briefly, a plant capture enhancer (Roche Nimblegen, Madison, WI, USA) and hyb enhancing A and B primers were added to one microgram of amplified *P.glauca* library and

dried. The mixture was resuspended in 1X SC hybridization buffer containing SC component A and heated to 70 °C for 10 minutes to rehydrate. The mixture was incubated at 95 °C for 10 minutes and brought to 42 °C prior to be loaded on the capture array. The hybridization was carried out at 42 °C and its duration was extended to 72 hours given the very large size of the P. glauca genome. Non-captured DNA was washed away according to the manufacturer's instructions. The captured DNA was amplified by ligation-mediated PCR using 454 A and B primers as described in the NimbleGen SeqCap EZ Library LR User's guide.

Target enrichment using SeqCap EZ developer (Roche Nimblegen, Madison, WI, USA) was performed according to the general guidelines provided in the NimbleGen SeqCap EZ Library LR User's guide. Briefly, 10 µl of plant capture enhancer (Roche Nimblegen, Madison, WI, USA) and 5 ul of 100 µM of hyb enhancing A and B primers were added to one microgram of amplified library and dried. The mixture was resuspended in 7.5 µl of 2X SC hybridization buffer and 3 µl of SC component A and heated to 70 °C for 10 minutes. After a quick spin, 4.5 µl of capture oligonucleotides solution in water were added and the hybridization mixture was incubated at 95 °C for 10 minutes followed by 72 hours at 47.5 °C. The hybridization mixture was put in contact with Streptavidin coated Dynabeads (Invitrogen, Carlsbad, CA, USA) and non-captured material was washed away according to the NimbleGen SeqCap EZ Library LR User's guide. The captured DNA was amplified by ligation-mediated PCR using 454 A and B primers as described in the NimbleGen SeqCap EZ Library LR User's guide. The quality of the captures was assessed by comparing pre- and post-capture libraries with quantitative PCR (qPCR) and primers designed against four spruce ESTs. The primer-pair efficiency brought to the power of the Cq difference between post- and pre-capture generated fold enrichment. These values varied from gene to gene but were around 100 times, on average.

2.10.2 Supplementary tables

Table S2.1- Gene structure data of orthologs of *Picea glauca* and *Pinus taeda*.

	Picea glauca		***Pinus taeda***		
Gene	**N° exons**	**Total introns length (bp)**	**N° exons**	**Total introns length (bp)**	**Ratio of intron length PG/PT**
LIM1	5	1877	5	2018	0,9
CesA1	13	3843	13	4340	0,9
CesA2	14	5462	14	7242	0,8
PAL	1		1		
Korrigan	5	1256	5	1183	1,1
Susy	15	9847	15	7424	1,3
MYB8	4	464	4	525	0,9
CAD	6	1159	6	2395	0,5
COBRA	6	1733	6	1777	1,0
COMT	3	1384	3	578	2,4
C3H	3	1446	3	1600	0,9
GS1a	14	3648	14	3680	1,0
H_PPase	8	5220	9	5239	1,0
Ldh_1_C	7	5433	7	3163	1,7
Cyt-b5	2	1613	2	1723	0,9
Gp_dh_C	11	3269	11	3951	0,8
Serpin	2	1380	2	1276	1,1
Peptidase_C1	9	1784	9	1798	1,0
eRF1_2	2	350	2	349	1,0
Ribosomal_S3_C	6	2618	7	3791	0,7
Cofilin_ADF	3	2687	3	2855	0,9
Ras	8	2964	8	3942	0,8
Thiolase_N	14	9491	14	9059	1,0
Average	7,0	3133,1	7,1	3177,6	1,0

Table S2.2- Genes associated with secondary cell-wall formation or with nitrogen metabolism in *P. glauca* targeted for BAC isolations.

Genes	**GenBank accession**	***Picea glauca* Reference ID (GCAT-pgl[1])**	***Picea glauca* BAC GenBank accessions**
Aspartate Aminotransferase (AAT)	BT117995*	GQ03919_P11*	KC860233
Asparagine synthetase (Asn1)	CO478951*	GQ0177_K02*	KC860234
Asparaginase	BT101939	GQ0133_M14	KC860235
Coumarate 3-hydroxylase (C3H)	BT106474	GQ03002_I06	KC860236
Cinnamyl alcohol dehydrogenase (CAD)	BT112280	GQ03312_O11	KC860237
Cellulose synthase (CesA1)	BT116636	GQ03803_L08	KC860238
Cellulose synthase (CesA2)	BT106827	GQ03011_H12	KC860239
Cellulose synthase (CesA3)	BT116976	GQ03810_K09	KC860240
COBRA	BT104865	GQ02816_A06	KC860241
Caffeate o-methyltransferase (COMT)	BT108042	GQ03116_D16	KC860242
Dof5	BT105779	GQ02828_F13	KC860243
Glutamine synthetase (GS1a)	BT114315	GQ03512_F02	KC860244
Homeobox-leucine zipper family protein (HD-ZIPIII)	BT117426*	GQ03819_E16*	KC860245
Isocitrate dehydrogenase (ICDH)	BT104232	GQ02808_B04	KC860246
Korrigan	EX433116*	GQ03912_H23*	KC860247
LIM 1	BT117230	GQ03815_F15	KC860253
Myb14	BT101254*	GQ0082_F08*	KC860248
Myb8	BT108136*	GQ03117_E18*	KC860249
Phenylalanine ammonia-lyase (PAL)	BT100475*	GQ0015_I17*	KC860250
Cinnamyl alcohol dehydrogenase (SAD)	BT112656	GQ03319_B08	KC860251
Sucrose synthase (Susy)	EX336506*	GQ03002_P04*	KC860252

* Incomplete cDNA in the white spruce gene catalogue (GCAT-pgl[1]) and GenBank. See Supplemental table 2.6

[1] Rigault et al. 2011

Table S2.3- Primer information and sequences used for BAC screening and sequencing validation

Gene name	Primer 1 [1]	Primer 2 [2]	Primer 3 [1]	Primer 4 [2]	Amplicon length (bp)[3]		
					1-4	1-2	3-4
COBRA	TACCAATACCAACATGAGGGTCCAG	TGTTATGTTGCCGTTTGGATCTAGTG	ACCAGGGACTCCTTACAACCAACAG	TGGTTCCTGCATTACCCACACTTAC	500	136	125
CAD	TTGCCATCTGCAAGCAATACAGTAG	TCCTGCATTTTTAGATGTACCTGAGAG	TGCATCATACCAGCAATGGGTATAG	ACAACGCCCAGCATAACTAGCTTTC	1750	176	197
Korrigan	ATGCTCAGGTTGGGAAAGGAGATAC	CAGGTCGGAACAAGAACTGCAATC	CCTGCCTTCCGTTCCTTCAATAG	CTTCCGTGGTTCAACTGAATCAAAC	894	128	152
CesA1	ATCGTCTATCCCCTCACGTCTCTTC	CACATTATGAATCAATCTGAGTTGGAG	GTTGGTGATCAACCTTGTTGGAATG	GCAGGACAGACCACAATATAACGATG	1463	127	194
AAT	ATTGATTCTTTGTTCCCCTTCCAAC	AATTTCAGCTTGCTGGATAGATCAG	CCAAAGCACTTTGTTACAGCTTGTG	TACTGCTAGCACCTGATGTGGTCTG	1022	225	285
HD Zip III	ACACAAGCTGTGAGTCTGTGGTGAC	CCGATAATTTGGAAAGGTAGACCAATG	ACACCTGTTTCCATTGATCCCTTTC	TGAGTACCACTCAGGGACCTTTCAC	996	122	130
Dof5	AGTGAAGCTGGGCTCCTTGAAATAC	CACTGGTGTAACCAAGTCAAGGCTAAG	TTCTTTTGTGTGTGCAAACAACCTG	GACAGTAATGGGGTTTTGCCTCTTC	603	139	143
ICDH	TGATGGTGATGTGCAGAGTGATTTC	TCTTACCAGTACAGACGTCATCAAG	TCTTGGATTTTGTTGTCTTGTCCAG	ATGCTGTTCGTGCTAGTTTCACCTC	1214	220	129
Myb14	CGGACAACGAGATAAAGAACCACTG	GAGGGCCAGGTAATGTTACTGTTATG	CAGAATGCGGGGTTTCTCTCATTAC	CTGGACCGCCAGCGATAGTAAG	589	124	168
PAL	AGCTGCCTTTAAGAGGAACCATCAC	CTCCGCTCATTTCTGTTCCATCTC	AGCGATCATGGAGTATGTCTTGGAC	GCAGATCTGATTACCTCGACCTGAG	529	126	156
GS1a	TTGCAAATCGAGGAGCTTCAGTTAG	ATATGTCCAATTTTGCAATCAACATC	ATAGGGCGATGGCTTTAAAAGACTG	CAGCTAAAAAGCGCAACAACATTTC	1853	139	182
CesA3	ATCCTCAGGTTGGGAGAAAAGTCTG	CAGTTCCCACATATACTGGCCCTTG	TGTGGATATGAAGACAAAACCGAATG	TCGTTTGGGCATACAGTAAATGGAC	1475	155	174
Susy	ATATTGATCCCTACCATGGGGTTTC	GGAATCAGGCCCTGTACTTACTTTTC	GGGTATTATAGTCATCGTGGCCTGTTC	CTTCAACCCCATTTACGCTTTCTTC	692	110	135
CesA2	CAGGGTCCAGTGTATGTAGGGACTG	GCTTTTTGGATGATTTCTTTGTTTTC	GGTTAATGTCGCAGAAGAGCTTTG	CTCGTATCCGCAGCTTATGACG	169	151	1227
COMT	GAGGGGTTTAAGTCTGGGGAGAAG	GCAATCACATGAGGAAGATCGAAG	GCGGAACCCAAGTCACAATATTG	GTAACCTCCAAAACGCTGAAATGTG	984	131	177
Asn1	GGGCTAATAAGTCGACATCTGCATG	TAAGGATTCTGCTCCTCGTCAAATG	TCTAGGTGTCCATGCCTCCGCATAC	CGGGTGCCTATACTTTCTTCCATTC	1013	265	120
Asparaginase	GGGTTGTGTTGTTGTTGACAGTGAG	GCATAAGTGCCTGCCCCTATTATTG	TACAAACAAGGGTTTGTCCTTGGTC	AATTGAACCCATAACGGGCAGATAC	648	120	139
C3H	TTTGTGGATGCATTGCTCACTCTAC	ATGGGTATGAGGAACCCCTTAATTC	GGCCTGAGAGATTTATTGAGGAAGATG	CATTCGAAATGATGAAGGAGGTGTC	572	152	154
Myb8	AGCTCCGAGTCGATCTGTAGGTTTC	GTACATGGATGGATTCGCCATTATC	GAAATGCTGCCTCCGTTTTCAAG	GGAGTACTGAGCATTGTGGTCCTG	696	148	218
SAD	TTAAAGACTCCTCCGCTCTTCGTTC	ACCGGCTTAATTCTCCTTCATTCAC	CTCCCACCTTTGGAGGTTACTCATC	GGAGGTATTTCATGGGGCTGTAGAC	933	141	233
LIM	CGAGGGCTCAAGACCGTTATTTTAC	TTGCAATGGTTACATCTGAAGCAAG	TGCCCTGAACTGGTTAATAAAGTTG	CTGCGTTTTCAAAGATAATGGCAAG	572	152	154

[1] 5' primer
[2] 3' primer
[3] Amplicons obtained from pairs 1-2 and 3-4 were used for PCR screening and verification of BAC; amplicons obtained from primer pair 1-4 were used for Sanger sequencing to verify the identity of the gene

Table S2.4- Accession numbers of *P. taeda* orthologs and sequence similarity to *P. glauca*.

A. BAC clones of *P. taeda.* Sequence identity with *P. glauca* was based on coding sequence.

Genes	***P. taeda* BAC GenBank accession**	**cDNA accession**	**Sequence identity with *P. glauca* (%)**
Cellulose synthase (CesA1)	AC241295.1	AY789650.1	93
Cellulose synthase (CesA2)	AC241331.1	AY789651.1	92
LIM 1	AC241349.1	BT117230*	93
Myb8	AC241314.1	DQ399057.1	93
Korrigan	AC241332.1	EF619968.1	90
Phenylalanine ammonia-lyase (PAL)	AC241300.1	PTU39792	90
Sucrose synthase (Susy)	AC241289.1	EF619967.1	93

* *P. glauca* cDNA was utilized because *P. taeda* cDNA was incomplete

B. Sequence from *P. taeda* shotgun assembly (Wegrzyn et al., 2014). Identity with *P. glauca* was based on coding sequence.

Genes	**cDNA accession**	**Sequence identity with *P. glauca* (%)**
Serpin	BT100637*	92
C3H	AY064170.1	93
COBRA	BT104865*	92
Peptidase_C1	BT107363*	91
eRF1_2	BT107692*	91
COMT	BT108042*	87
Gp_dh_C	BT111068*	89
Cofilin_ADF	BT111103*	93
Ras	BT111640*	93
Ribosomal_S3_C	BT112106*	86
CAD	Z37991.1	92
Gs1a	BT103460*	88
H_Ppase	BT115473*	90
Thiolase_N	BT115978*	89
Ldh_1_C	BT117871*	92
Cyt-b5	BT119045*	85

* *P. glauca* cDNA was utilized because *P. taeda* cDNA was incomplete

Table S2.5- Accession numbers for the closest homologous sequences between *P. glauca*, *Arabidopsis thaliana*, *Populus trichocarpa* and *Zea mays*.

Picea glauca		***Arabidopsis thaliana***		***Populus trichocarpa***		***Zea mays***	
GCAT accessions	**GenBank accessions**	**GenBank accessions**	**Identity %**	**JGI v3.0 gene name**	**Identity %**	**GenBank accessions**	**Identity %**
GQ0033_L20	BT100637	NP_175202.1	49	Potri.014G036000.1	55	NP_001167655.1	52
GQ0082_F08	BT101254	NP_195574.1	65	Potri.009G134000.1	60	NP_001106009.1	66
GQ0177_K02	CO478951	NP_196586.1	82	Potri.005G075700.1	81	ACF80883.1	76
GQ0182_H10	BT102359	NP_564098.2	95	Potri.009G125000.1	90	NP_001150274.1	94
GQ02808_B04	BT100366	NP_176768.1	85	Potri.004G074900.1	85	NP_001140324.1	77
GQ02810_I18	BT104452	NP_568336.1	70	Potri.017G125100.1	61	NP_001151414.1	60
GQ02816_A06	BT104865	NP_568930.1	83	Potri.015G060000.1	85	NP_001105970.1	71
GQ03002_I06	BT100373	NP_850337.1	76	Potri.006G033300.1	82	NP_001130442.1	69
GQ03002_P04	EX336506	NP_199730.1	78	Potri.002G202300.1	76	NP_001105194.1	78
GQ03011_H12	BT106827	NP_199216.2	80	Potri.004G059600.1	76	NP_001105672.1	85
GQ03012_N11	BT106885	NP_187818.1	96	Potri.006G192700.1	97	NP_001150462.1	97
GQ03104_C22	BT107197	NP_171777.1	91	Potri.015G029500.1	93	NP_001148813.1	93
GQ03106_H10	BT107363	NP_563648.1	61	Potri.002G184200.1	64	NP_001150152.1	68
GQ03109_L23	BT107587	NP_195193.1	89	Potri.005G051200.1	88	NP_001149096.1	86
GQ03111_J03.2	BT107692	NP_189295.3	93	Potri.014G141000.1	95	NP_001151538.1	93
GQ03116_D16	BT108042	NP_200227.1	42	Potri.019G102900.1	46	NP_001149617.1	36
GQ03117_E18	BT108136	NP_172425.2	65	Potri.010G004300.1	79	NP_001132070.1	83
GQ03210_A11	BT109562	NP_197551.2	74	Potri.018G050200.1	83	NP_001169011.1	64
GQ03232_E11	BT111068	NP_187062.1	89	Potri.001G335800.1	86	NP_001105385.1	90
GQ03232_K24	BT111103	NP_567182.1	76	Potri.001G106200.1	72	NP_001151716.1	73
GQ03301_J24	BT111640	NP_171715.1	77	Potri.004G226600.1	80	NP_001105441.1	77
GQ03310_B15	BT112106	NP_198403.1	85	Potri.006G222100.1	87	NP_001149150.1	91
GQ03312_O11	BT112280	NP_195149.1	68	Potri.009G095800.1	70	NP_001105654.1	69
GQ03319_B08	BT102039	NP_195643.1	67	Potri.009G062800.1	61	NP_001147726.1	60
GQ03512_F02	BT103460	NP_568335.1	81	Potri.017G131100.1	83	ACB06727.1	80
GQ03610_A06	BT115139	NP_567178.1	85	Potri.003G128600.1	84	NP_001104934.1	84
GQ03617_H21	BT115473	NP_173021.1	84	Potri.006G063000.1	85	NP_001105380.1	82
GQ03709_L23	BT115978	NP_199583.1	80	Potri.014G168700.1	83	NP_001148667.1	75
GQ03803_L08	BT106211	NP_199216.2	64	Potri.011G069600.1	73	NP_001105672.1	63
GQ03810_K09	BT116956	NP_197244.1	74	Potri.006G181900.1	75	NP_001105532.1	76

GQ03819_E16	BT117426	NP_174337.1	62	Potri.001G372300.1	73	NP_001142394.1	65
GQ03912_H23	EX433116	NP_199783.1	74	Potri.003G151700.1	79	NP_001183308.1	77
GQ03915_D23	BT117871	NP_171936.1	89	Potri.010G071000.1	89	NP_001147160.1	86
GQ03919_P11	BT100655	NP_850022.1	67	Potri.005G079200.	72	NP_001143769.1	74
GQ04013_M05	BT119045	NP_190458.1	63	Potri.012G137800.1	66	NP_001149328.1	65

Table S2.6- Summary of sequencing results of *P. glauca* BAC clones isolated each containing a different single copy gene associated with cell- wall formation or with nitrogen metabolism.

BAC	Total of contigs	Average coverage	Total size (bp)
Asparagine synthetase (Asn1)	2	135	130054
Asparaginase	5	497	39145
Aspartate Aminotransferase (AAT)	9	141	192670
Coumarate 3-hydroxylase (C3H)	14	70	161072
Cinnamyl alcohol dehydrogenase (CAD)	5	140	104305
Cellulose synthase (CesA1)	14	91	133796
Cellulose synthase (CesA2)	8	89	150078
Cellulose synthase (CesA3)	6	117	114058
Cobra	4	188	75244
Caffeate o-methyltransferase (COMT)	7	167	104865
Dof5	13	78	196708
Glutamine synthetase (GS1a)	6	67	141000
Homeobox-leucine zipper family protein (HD-ZIPIII)	14	287	101367
Isocitrate dehydrogenase (ICDH)	5	56	118357
Korrigan	5	119	83070
LIM1	11	89	137136
MYB14	14	204	160947
MYB8	4	73	92372
Phenylalanine ammonia-lyase (PAL)	4	111	148199
Cinnamyl alcohol dehydrogenase (SAD)	6	51	112140
Sucrose synthase (Susy)	12	227	135947
Average	8	143	125359

Table S2.7- GenBank accessions of complete cDNA utilized for gene structure definition when the cDNA in *Picea glauca* gene catalogue was incomplete.

Genes	***Picea glauca* Reference ID (GCAT-pgl[1])**	**Specie**	**GenBank accession**	**Reference**
Aspartate Aminotransferase (AAT)	GQ03919_P11	*P. sitchensis*	WS0284_A12	Ralph et al. 2008
Asparagine synthetase (Asn1)	GQ0177_K02	*P. sylvestris*	AJ496567	Canas et al. 2006
Homeobox-leucine zipper family protein (HD-ZIPIII)	GQ03819_E16	*P. glauca*	HQ391914	Cote et al. 2010
Korrigan	GQ03912_H23	*P. sitchensis*	WS02912_I08	Ralph et al. 2008
Myb14	GQ0082_F08	*P. glauca*	pending	Fortin et al. (unpublished)
Myb8	GQ03117_E18	*P. taeda*	DQ399057	Bedon et al. 2007
Phenylalanine ammonia-lyase (PAL)	GQ0015_I17	*P. glauca*	pending	This report
Sucrose synthase (Susy)	GQ03002_P04	*P. taeda*	EF619967	Nairn et al. 2008

[1] Rigault et al. 2011

Table S2.8- Repetitive elements detected within gene structure of the 35 *P. glauca* genes[1].

Gene	Number of Matches	Matching Class[2]	Average lenght (bp)
AAT	3	NHF	105
CAD	2	NHF	86
CesA1	3	NHF	152
CesA2	7	NHF	162
CesA2	1	UNK	103
Korrigan	1	NHF	243
H_Ppase	1	NHF	190
Thiolase_N	1	NHF	331
GST_N	2	NHF	325
Peptidase_C1	2	NHF	217
Ldh_1_C	1	UNK	121

[1]All of the Repetitive elements were detected in intron sequence.
[2]Repetitive elements are classified as NHF (no significant hit in RepBase and nr genbank) and UNK (significant hits in nr genbank only).

2.10.3 Supplementary figures

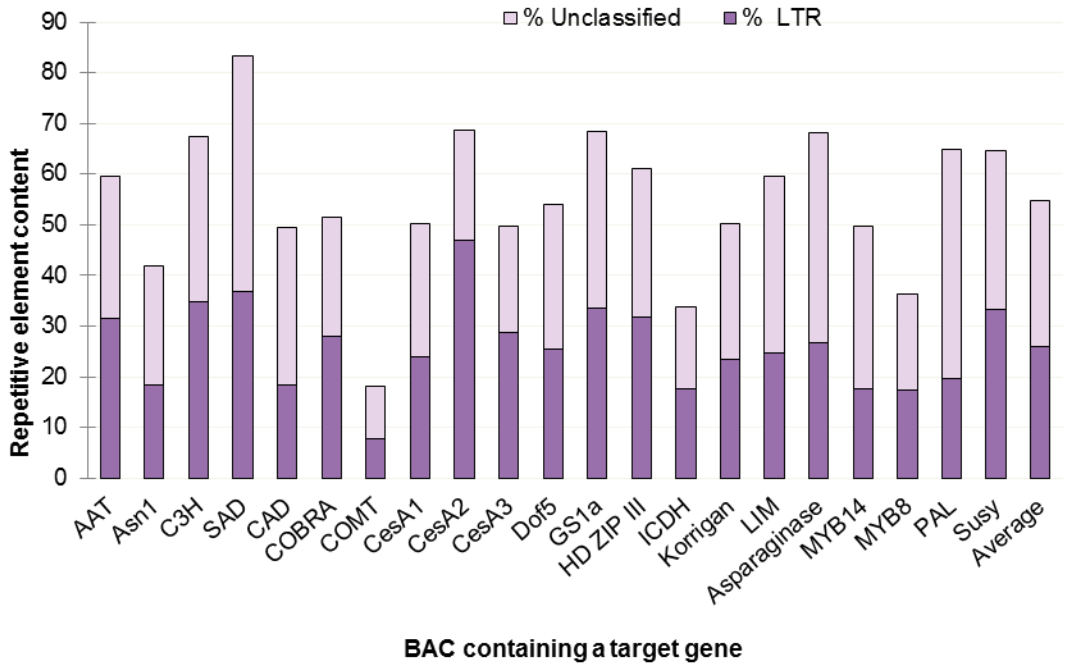


Figure S2.1- Content of repetitive elements in 21 different BAC clones. The analysis used the RepeatMasker software and a *P. glauca* repetitive sequence library (see Methods). Repetitive elements were classified as LTR (long terminal repeat) and unclassified (no hit in RepBase).

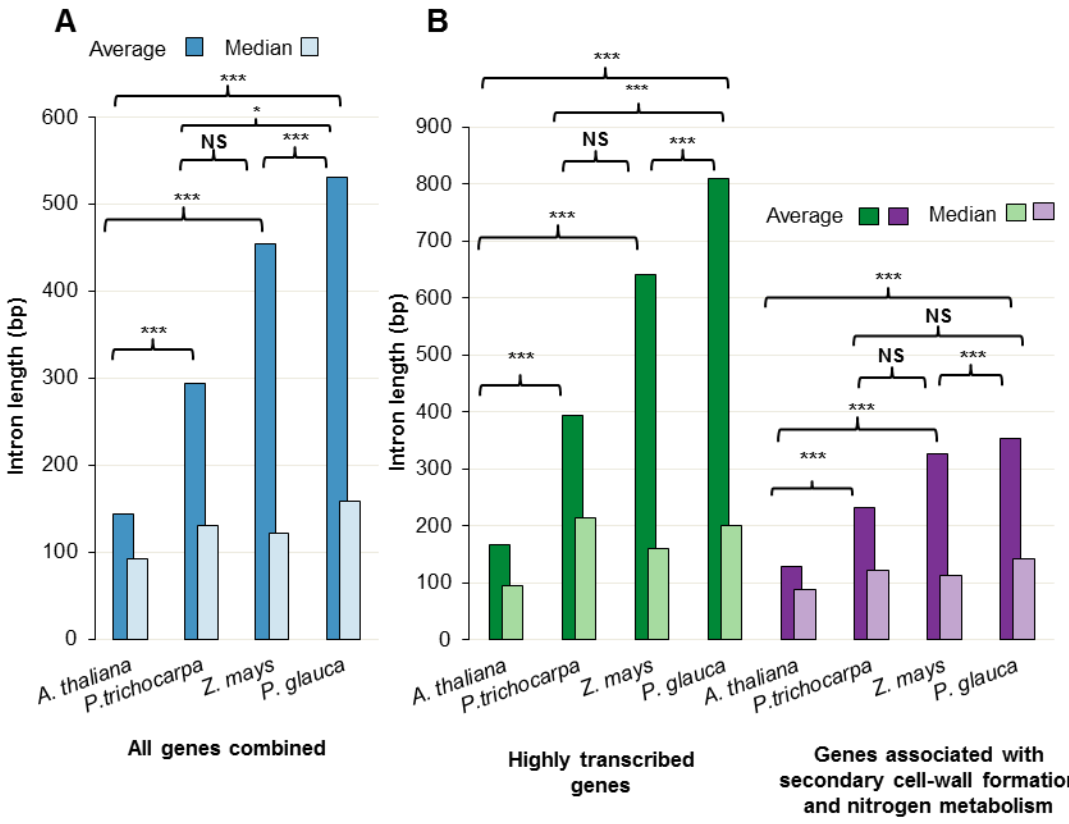


Figure S2.2- Comparative analysis of individual intron length in *P. glauca*, *A. thaliana*, *P. trichocarpa* and *Z. mays*. A. Average and median length of individual introns in all genes. B Average and median length of individual introns in highly expressed genes and genes associated with secondary cell-wall formation and nitrogen metabolism in four species. Intron lengths were compared among the four species by Kruskal-Wallis test with post-test analysis by Dunn's multiple comparisons: NS, not significant ($P > 0.06$); * $P < 0.06$; **$P < 0.01$; ***$P < 0.001$.

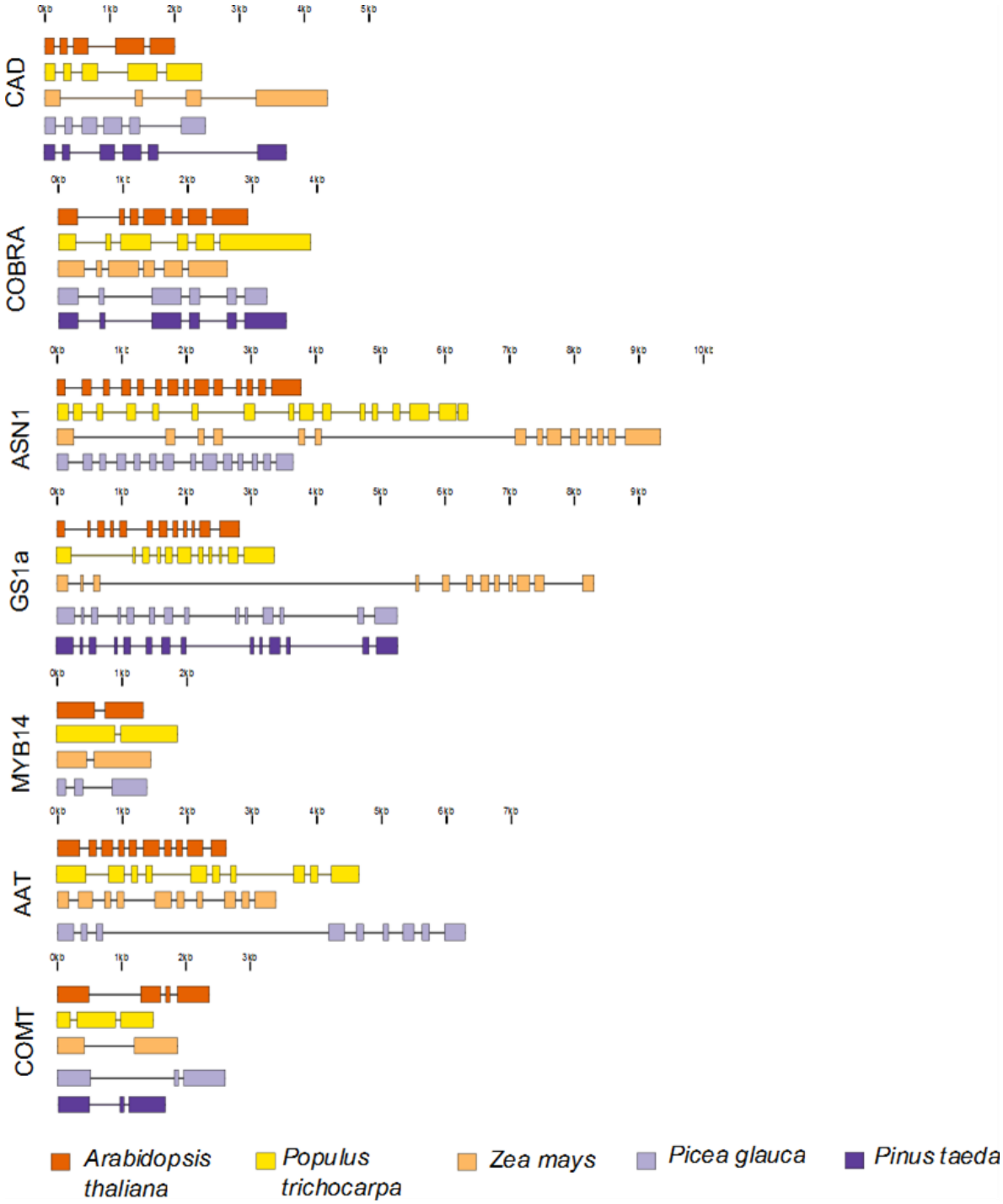


Figure S2.3- Boxplot of the 35 homologous genes in *P. glauca*, *A. thaliana*, *P. trichocarpa* and *Z. mays*.

Figure S2.3 continuation

Figure S2.3 continuation

Figure S2.3 continuation

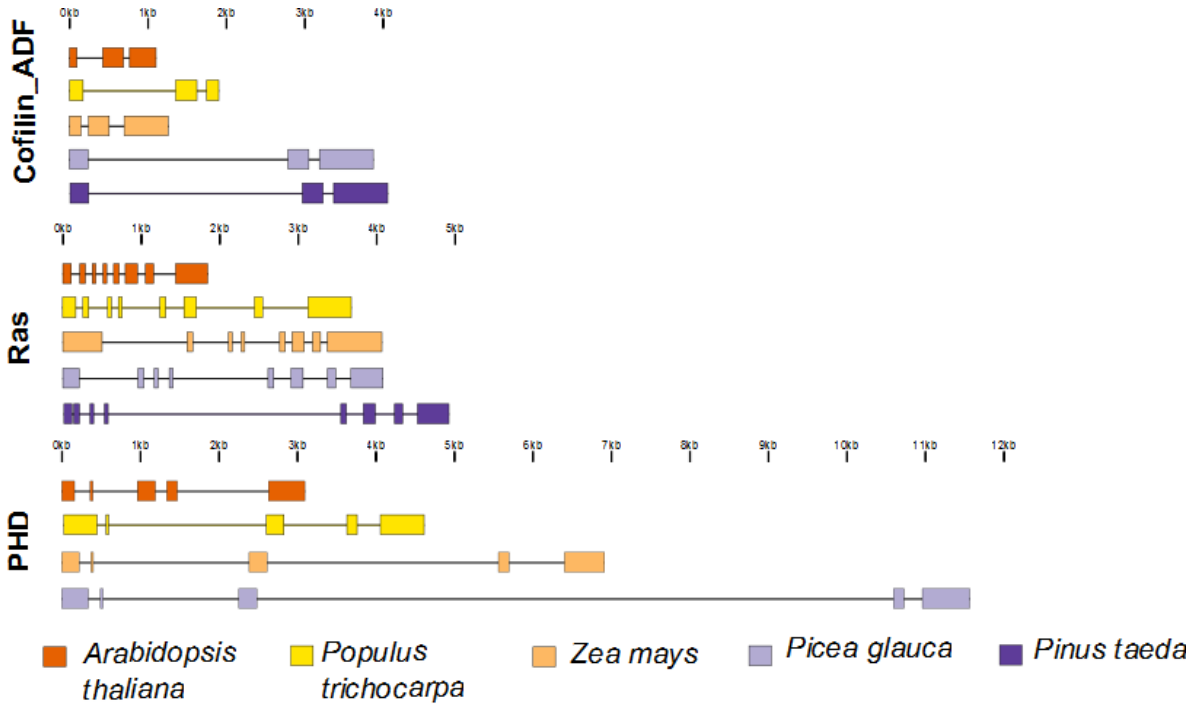


Figure S2.3 continuation

Chapter 3: Expansion of the dehydrin gene family in conifers is associated with considerable structural diversity and drought responsive expression

[Stival Sena J, Giguère I, Rigault P, Bousquet J, Mackay J. Expansion of the dehydrin gene family in conifers is associated with considerable structural diversity and drought responsive expression. *manuscript*]

3.1 Abstract

Temperatures are expected to increase over the next century in all terrestrial biomes and particularly in boreal forests, where drought-induced mortality has been predicted to rise. Understanding the molecular basis of drought tolerance will help to preserve the genetic diversity relevant for maintaining adaptability in managed forests. It was recently suggested that osmo-protecting dehydrin proteins formed a larger gene family in conifers than in flowering plants. The main objective of this study was to identify all of the putative members of the family, trace their evolutionary origin and examine their functional and structural diversity. We identified 41 complete dehydrin coding sequences in *Picea glauca*, which is four times more than in angiosperms studied to date on average, and more than in pines. Phylogenetic reconstructions indicated that the gene family has undergone an expansion in conifers, with parallel evolution implicating the sporadic resurgence of certain amino acid sequence motifs, and a major duplication giving rise to a clade specific to the genus *Picea* only. The delineated phylogenetic clades were also highly congruent with structural variation in dehydrins. No support was found for a major whole-genome duplication (WGD) common to all Pinaceae. A wide variety of dehydrin structures were identified across all plants with variable numbers and assemblages of the A-, E-, S- and K-segments and an N-terminal (N1) amino acid motif. In *Picea glauca*, gene-specific determinations of transcript level identified several sequences with tissue preferential expression and eight dehydrins that had increased expression following drought stress, with

N1-K2 and N1-AESK2 sequences being the most responsive. Altogether, these observations of family expansion, patterns of expression, and structural diversification implicating loss and gain of amino acid motifs, indicate that subfunctionalization would be the main driver for the diversity seen among gene duplicates. Dehydrins thus represent a potent gene family for adaptation to drought stress in long-lived spruces, likely providing them with more flexibility in the face of spatially and temporally variable environments.

3.2 Résumé

On s'attend à ce que les températures augmentent au cours du prochain siècle dans tous les biomes terrestres et particulièrement dans les forêts boréales, où il est prédit une augmentation de la mortalité due à la sécheresse. La compréhension des bases moléculaires impliquées dans la tolérance à la sécheresse pourra aider à préserver la diversité génétique pertinente pour l'adaptabilité dans les forêts aménagées. Il a été récemment suggéré que les protéines osmo-protectrices, les déhydrines, formaient une plus grande famille de gènes chez les conifères que chez les angiospermes. L'objectif principal de cette étude a été d'identifier tous les membres de la famille des déhydrines, de tracer leur origine évolutive et d'examiner leur diversité fonctionnelle et structurelle. Nous avons identifié 41 séquences codantes complètes de déhydrines chez *Picea glauca*, soit quatre fois plus que la moyenne chez les angiospermes étudiées à ce jour, et aussi chez les pins. Des reconstructions phylogénétiques ont indiqué que cette famille de gènes a subi une expansion chez les conifères, dont une évolution parallèle impliquant la résurgence sporadique de certains motifs ainsi qu'une duplication majeure ayant donné lieu à un clade spécifique au genre *Picea* uniquement. Les clades phylogénétiques délimités étaient hautement congruents avec la variation structurale observée au niveau des déhydrines. Nous n'avons pas observé d'evidences en faveur d'une duplication complète du génome qui aurait été commune aux Pinaceae. Une grande variété de structures des déhydrines a été observée chez toutes les plantes avec des variations dans le nombre et l'ordre des segments d'acides aminés A, E, S, K et N-terminal (N1). Chez *Picea glauca*, les profils d'expression ont montré que plusieurs gènes s'exprimaient préférentiellement dans certains tissus et que huit déhydrines augmentaient leur expression en réponse à la sècheresse, les séquences N1-K2 et N1-AESK2 étant les plus sensibles. Prises ensemble, ces observations d'expansion, de patterns d'expression, et de diversification structurelle de la famille impliquant des gains et pertes de motifs d'acides aminés, indiquent que la sous-fonctionnalisation serait la force principale favorisant les multiples duplications de gènes. En conséquence, les déhydrines représentent une famille de gènes avec une implication palpable dans l'adaptation à la sécheresse *Picea*, conférant à ces espèces longévives de meilleures habiletés pour affonter des conditions environnementales spatialement et temporellement hétérogènes.

3.3 Introduction

A major factor driving the evolution and diversification of vascular plants is the adaptation to water availability (Micco and Aronne, 2012). To face the various stresses that impact on water relations such as drought, heat, freezing and salinity, plants have developed mechanisms to prevent the loss of intracellular water (Farooq et al., 2012). The dehydrins are among the most studied proteins that are believed to have dehydration protective functions in plants (Hanin et al., 2011)

Dehydrin proteins have a modular structure comprised of a variable number of conserved motifs. A lysine-rich sequence motif named K-segment is the only motif that is present in every dehydrin described to date, with the exception of a unique dehydrin described in maritime pine (Perdiguero, Soto and Collada, 2014). Another conserved motif is the S-segment defined by five to seven consecutive serine residues. Other motifs are lineage-specific and include the Y-segment that is characterized by the presence of Tyrosine residues and only present in angiosperms (Campbell and Close, 1997), and the A- and E-segments exclusive to conifers and characterized by the presence of alanine and glutamine residues respectively (Perdiguero et al., 2012). Based on their motif composition, these proteins have been classified as Kn, SKn, YnSKn, YnKn (Close, 1996), EnSKn and AnEnSKn (Perdiguero et al., 2012).

Some dehydrins accumulate in maturing seeds or are induced in vegetative tissues after salinity conditions, dehydration, cold stress and frost (Close, 1996; Tunnacliffe and Wise, 2007). The expression of dehydrins varies in different tissues and according to the type and intensity of stress. For example, in grapevine, *Dhn1* was not expressed in vegetative tissues under normal conditions but was induced by drought, cold, heat and embryogenesis. In contrast, *Dhn3* was induced to low level during seed development and not responsive to stress treatments (Yang et al., 2012). In Norway spruce (*Picea abies)*, *Dhn1* and *Dhn6* were highly expressed in bark and leaves during drought stress while the others dehydrins were poorly induced or not responsive (Eldhuset et al., 2012).

Proteins from different classifications were shown to be upregulated by abiotic stresses including cold and/or salt and/or desiccation, and no clear relationship has been observed between the structural classification and the stress responsiveness profile. For example, the sequences *YnSKn DHR18* and *Kn XERO2* from mouse-ear cress, and *SKn EuglDhn2* from *Eucalyptus,* are upregulated in response to cold stress, but only *DHR18* and *XERO2* are induced by salt stress and *DHR18* and *EuglDhn2* by drought (Hundertmark and Hincha, 2008; Fernández et al., 2012).

The number of members of the dehydrin gene family is variable among different species. In angiosperms, it can range from a few members as in the primitive *Amborella*, which has two dehydrins (Pfam 30.0, Finn et al., 2014), to more than ten as in apple trees (*Malus domestica*), which have twelve dehydrins (Liang et al., 2012). In contrast, conifer trees appear to have significantly more dehydrin genes although only a few have been investigated in any detail (Perdiguero et al., 2012; Perdiguero, Soto and Collada, 2014; Yakovlev et al., 2008; Joosen et al., 2006; Kjellsen et al., 2013). A total of 53 distinct dehydrin genes have been identified in the white spruce (*Picea glauca* [Moench] Voss) transcriptome database (Rigault et al., 2011), which is many more than in herbaceous angiosperms studied to date (Liu et al., 2012; Liang et al., 2012; Hundertmark and Hincha, 2008). As a basis to understand the role and evolution of dehydrin genes in conifers and evaluate their involvement in water stress responses, we aimed to: (1) assess the extent of the dehydrin gene family and its expansion in conifers by using full-length gene sequences identified in white spruce; (2) trace the evolutionary origin of dehydrins in both conifers and angiosperms by studying phylogenetic relationships; (3) classify these genes based on conserved amino acids motifs such as the A-, E-, S- and K-segments; and (4) evaluate the expression profile of dehydrins in different tissues under normal conditions and in response to water stress.

3.4 Materials and Methods

3.4.1 Dehydrin sequences

Dehydrins were identified in the white spruce catalog of expressed genes (Rigault et al. 2011) by using the HMMER software (v 3.0) (Johnson et al., 2010) and the Pfam database, release 27.0 (Finn et al., 2014). We also performed sequence similarity searches (BLASTp; Altschul et al., 1990) with published conifer dehydrins. We utilized RNA-seq data from white spruce (Verta et al., 2016) to extend the incomplete dehydrin cDNA sequences identified in the gene catalog. These cDNA sequences were then translated into amino acids and the integrity of the Open Reading Frames (ORF) was verified by Blastp (e-value 1e-10) using complete dehydrin sequences from conifers and other plants.

We named the dehydrins as suggested by Richard et al. (2000), in which the white spruce dehydrin (*PgDhn1*) was first described. In this present study, dehydrins were named from *PgDhn2* to *PgDhn41*.

3.4.2 Phylogenetic analysis

We searched for conifer dehydrin proteins in the non-redundant protein database (nr) using Blastp, e-value threshold of $1e^{-20}$, and white spruce ORFs as query. We retained only conifer dehydrins with at least 70% of amino acid sequence similarity and coverage over a minimum of 80 amino acids with the white spruce ORFs.

The search for angiosperm dehydrins was performed by phmmer (e-value < 0.01) (https://www.ebi.ac.uk), which uses profile hidden Markov models and provides a more accurate and sensitive detection of remote homologs than BLAST. In this analysis, we used the white spruce ORF sequences to search in UniProt database (The UniProt Consortium, 2015) for *Arabidopsis thaliana***,** *Malus domestica, Eucalyptus grandis, Eucalyptus. globulus, Prunus persica, Prunus dulcis, Zea mays, Amborella trichopoda, Populus trichocarpa, Vitis vinifera, Oryza sativa* and *Physcomitrella patens* dehydrin proteins. The proteins were verified as dehydrins by hmmerscan (https://www.ebi.ac.uk) using the Pfam database.

The sequences were first selected by using the cd-hit program (Li and Godzik, 2006) to separately cluster the conifer and angiosperm dehydrin sequences based on a 97% similarity threshold at the amino acid sequence level. Each cluster was represented by one sequence (Table S3.2). We then aligned the representative sequences using MAFFT version 7.0 and FTT-NS-I (iterative refinement method; 1000 iterations) strategy (Katoh and Standley, 2013) found in Geneious R6 (http://www.geneious.com, Kearse et al., 2012).

Phylogenetic trees were constructed following a Bayesian framework using MrBayes 3.2.1 (Ronquist et al., 2011). Half a million generations of the Markov chain Monte Carlo (MCMC) method using four chains sampling every 10 generations were completed using the Wag model (Whelan and Goldman, 2001), with gamma-distributed rate variation across sites and a proportion of invariable sites. A dehydrin (*A9RQA9*) of the moss *Physcomitrella patens* was used as outgroup. The standard deviation of split frequencies was < 0.05 after 485,000 generations. The first 25% of the recovered topologies were discarded. We calculated the consensus tree with Bayesian posterior probability equal or superior to 0.75 and the resulting samples of best-fit trees. Trees were visualized with FigTree v1.4.2 (Rambaut 2009, http://tree.bio.ed.ac.uk).

We also constructed the phylogenetic trees utilizing the Maximum Likelihood (ML) approach implemented in MEGA6 (Tamura et al., 2013), and also using the WAG model and gamma-distributed rates among sites. We obtained similar but less resolved topologies when compared to results from MrBayes analysis. Thus, only results from MrBayes analysis will be shown.

3.4.3 Identification of conserved amino acid motifs and classification of dehydrins

The identification of amino acid motifs was performed by using MEME version 4.9.0 (Multiple expectation Maximization for motif Elicitation) (Bailey et al., 2015) with the following parameters: distribution of motif occurrences was any number of repetitions, maximum number of motifs was 10 and motif width between 6 and 20 amino acids. Motif scanning was performed by MAST (MEME suite, Bailey et al., 2015) and then, sequences were classified among the possible groups: Kn, SKn, YnSKn, YnKn, EnSKn, AnEnSKn,

and new groups described in the current study. We verified the motifs in the multiple sequence alignments and also identified the degenerate motifs.

3.4.4 Plant material

Drought experiment

We used young trees from three genetically unrelated *Picea glauca* genotypes (clones 8, 11 and 95) that were propagated *in vitro* by somatic embryogenesis and grown in containers for two years at the Vegetative Propagation Centre of the Saint-Modeste tree nursery of the Quebec Ministry of Forests, Wildlife and Parks of Québec (Saint-Modeste, Canada). The plants were 40 cm on average, were potted in pots of 5 liters containing a mix of peat, perlite and vermiculite (3:1:1, by weight) and grown in a greenhouse with day temperature of 23° C, night temperature of 20° C, 16/8 (day/night) photoperiod and watered three times per week, for two months prior to the experiment.

For the experiment, one half of the plants were watered and for the other half, water was withheld; the plants were arranged in a completely randomized design. Plants were destructively sampled and the newly formed foliage (needles) was collected at 0, 7, 14, 18 and 22 days from the beginning of the watering treatments. Five plants per genotype (replicates) in both watering treatments were sampled at each sampling point (a total of 150 plants). The watered plants were sampled 2 hours after the last watering. The sampling time was at midday.

At each sampling day, the midday water potential (branch) of four plants per genotype in both watering regimes was measured using a Scholander pressure chamber (Model 610, PMS instruments, Albany, OR, USA).

Foliage samples were frozen in liquid N_2 immediately after removal from the trees and stored at −80°C. The needles were ground to powder using a MixerMill 300 (Retsch, http://www.retsch.com/) and steel grinding balls cooled in nitrogen. Powdered foliage tissue was stored at −80°C until RNA extraction.

Shoot xylem, phelloderm and young foliage from plants under natural conditions

Three biological replicates of shoot secondary xylem, phelloderm and young foliage from an experience published by Raherison et al. (2015) were used to analyse tissue-preferential expression. Each replicate was a pool of samples from five genetically distinct young white spruce trees grown under non-limiting conditions in a glasshouse under natural light as described by Raherison et al. (2015). The samples were collected at 6 A.M., immediately frozen in liquid nitrogen and stored at −80°C until RNA isolation.

3.4.5 RNA extraction and cDNA synthesis

Total RNA was extracted from powdered frozen tissue utilizing the cetyltrimethyl ammonium bromide (CTAB) extraction method as described by Chang et al. (1993) with modifications (Pavy et al., 2008), and stored at −80°C. The total RNA concentration was determined using a NanoDrop 1000 (Thermo cientific, http://www.thermoscientific.com/) and assessed for quality with an Agilent 2100 Bioanalyzer and Agilent RNA 6000 Nano Kit LabChips (Agilent Technologies Inc.,http://www.agilent.com/). Complementary DNAs were prepared from 500 ng of total RNA using Quantitect Reverse Transcription Kit (Qiagen, X) and then diluted 1:4 in RNase-free water.

3.4.6 Primer design and quantitative RT-PCR

A pair of primers was designed for each dehydrin using the Primer3Plus software (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) (Table S3.1). The self-complementarity of the designed primers was verified using Oligo Calc (Oligonucleotide Properties Calculator software,http://www.basic.northwestern.edu/biotools/oligocalc.html) and specificity was verified against the *P. glauca* gene catalogue (Rigault et al., 2011) and the extended dehydrin cDNA sequences.

We used gene-specific dehydrin primers (Table S3.1) to determine RNA transcript levels from drought stress and the tissue comparison experiments (Raherison et al., 2015) by using quantitative RT-PCR (qPCR). We used QuantiFast® SYBR® Green PCR kit (Qiagen) as follows: 1× master mix, 300 nMof 5′ and 3′ primers and 5 µl of cDNA (5ng) in a final

volume of 15 µl. Amplifications were carried out in a LightCycler® 480 (Roche, http://www.roche.com/) as described in Boyle et al. (2009). We used the LRE method (Rutledge and Stewart, 2008) adapted for Excel (Boyle et al., 2009) to estimate the number of transcript molecules, which was normalized by the geometric mean of three reference genes: elongation factor 1a (EF1-α) (BT102965), cell division cycle 2 (CDC2) (BT106071) and ribosomal protein L3A (BT115036) as described in Beaulieu et al. (2013). PCR products were sequenced with the Sanger method to verify primer specificity.

3.4.7 Sequence analysis of amplicons

Statistical analysis

The transcript abundance data (as per normalized number of molecules) determined by RT-qPCR for each dehydrin was transformed to log2 for subsequent statistical analysis. We used the R software for statistical computing and the construction of graphs (package ggplot2) (http://www.r-project.org; Wickham et al., 2009).

Drought experiment

The transcript abundance data of each gene was analyzed separately using analyses of variance (ANOVA) as a function of the type of treatment (different watering regimes simulating different water potential), genotypes, sampling dates and their interactions. If significant differences were detected, a multiple comparison test (Tukey's honest significant difference, HSD) was performed.

Tissue preferential expression

To evaluate gene expression differences among the three tissues under non-limiting conditions, we performed gene-specific analysis of variance (ANOVA) with expression as a function of tissues. If significant differences were detected, a multiple comparison test was performed (Tukey's honest significant difference, HSD).

3.5 Results

3.5.1 Identification of abundant dehydrins sequences in spruce

This study was initiated based on 53 cDNA dehydrin sequences that were identified in the white spruce gene catalog (Rigault et al., 2011; Raherison et al., 2012). In total, 41 of the sequences contained an open reading frame (ORF) that was deemed to be complete based on sequence alignments.

Sequence similarity and HMM searches were used to find sequences in other species. We identified 108 amino acid sequences in the genus *Pinus* and 36 in *Picea* with at least 70% of similarity and coverage with the previously discovered white spruce dehydrins. Our search of dehydrins also extended to both monocots and dicots in order to represent the diversity of dehydrins in angiosperms as well as major conifers. We identified a total of 76 dehydrins distributed in mouse-ear cress (*Arabidopsis thaliana*), apple (*Malus domestica*), *Eucalyptus* (*Eucalyptus grandis* and *Eucalyptus. globulus*), peach (*Prunus persica*), almond (*Prunus dulcis*), maize (*Zea mays*), *Amborella trichopoda*, poplar (*Populus trichocarpa*), grape (*Vitis vinifera*) and rice (*Oryza sativa*).

Many of the sequences within both the angiosperms and conifers were very similar, therefore we clustered the sequences that were 97% identical or more. The conifer sequences thus formed 78 distinct clusters (Table S3.2) whereas the angiosperms formed 57 clusters (Table S3.3), suggesting a larger number and greater diversity of sequences in conifers than in angiosperms, as previously reported (Rigault et al., 2011). The representation of different taxa in the clusters strongly supported this hypothesis as the individual conifer genera were represented in more clusters, i.e. 56 clusters for *Picea* and 21 clusters for *Pinus* whereas major angiosperms were represented in only two to ten clusters. The data also showed that dehydrin sequences were particularly abundant in *Picea*, but this could also be the effect of a more complete reporting of such sequences in *Picea* than in other conifers.

3.5.2 Dehydrins are highly divergent between angiosperms and conifers

We used one representative sequence from each cluster to construct the phylogenetic trees and carried out an exhaustive *de novo* search for amino acid motifs in the dehydrin sequences. Five of the motifs had previously been identified in plant dehydrins (Campbell and Close, 1996; Zolotarov and Stromvik, 2015). They include the K-segment and S-segment common to both angiosperms and gymnosperms, the A-segment and E-segment found only in conifers (Perdiguero et al., 2012), and the Y-segment found only in angiosperms. Here, we also identified a conserved N1 motif located at the N-terminal region of sequences (Fig. S3.1). In the present study, the A-segment was found to be less conserved than described (Perdiguero et al., 2012), which we explain by the fact that we used a total of 78 conifer dehydrins in the motif-identification process. This result impacts on the classification of dehydrins. In a previous study, the maritime pine dehydrin *Ppter_dhn_ESK2* was reported to lack an A-segment (Perdiguero et al. (2012) but we found that it contains a less conserved A-segment.

We began by constructing separate phylogenetic trees with the dehydrin sequences from angiosperms and from conifers. The angiosperm tree was split into two main groups, with characteristic YnSKn and N1 SKn amino acid motif structures (Fig S3.2); the conifer tree showed two distinct groups with different amino acid motif structures, i.e. N1 Kn and N1 AnEnSKn (Fig. 3.1). Next, we constructed a combined angiosperm and conifer tree which also revealed the same two conifer groups and the two disjunct angiosperm groups; some of the branches indicating further phylogenetic structuring had weak statistical support and for this reason, we estimated a consensus tree with an *a posteriori* probability threshold support equal or superior to 0.75 (Fig. 3.2).

Within the conifers, two paraphyletic groups were observed for both the N1 Kn and Kn structures (shades of violet in Fig. 3.2), suggesting parallel evolution. We observed the presence of few isolated Kn dehydrins in the major groups YnSKn (orange) and N1 AnEnSKn (light blue in Fig. 3.2), also suggesting parallel evolution.

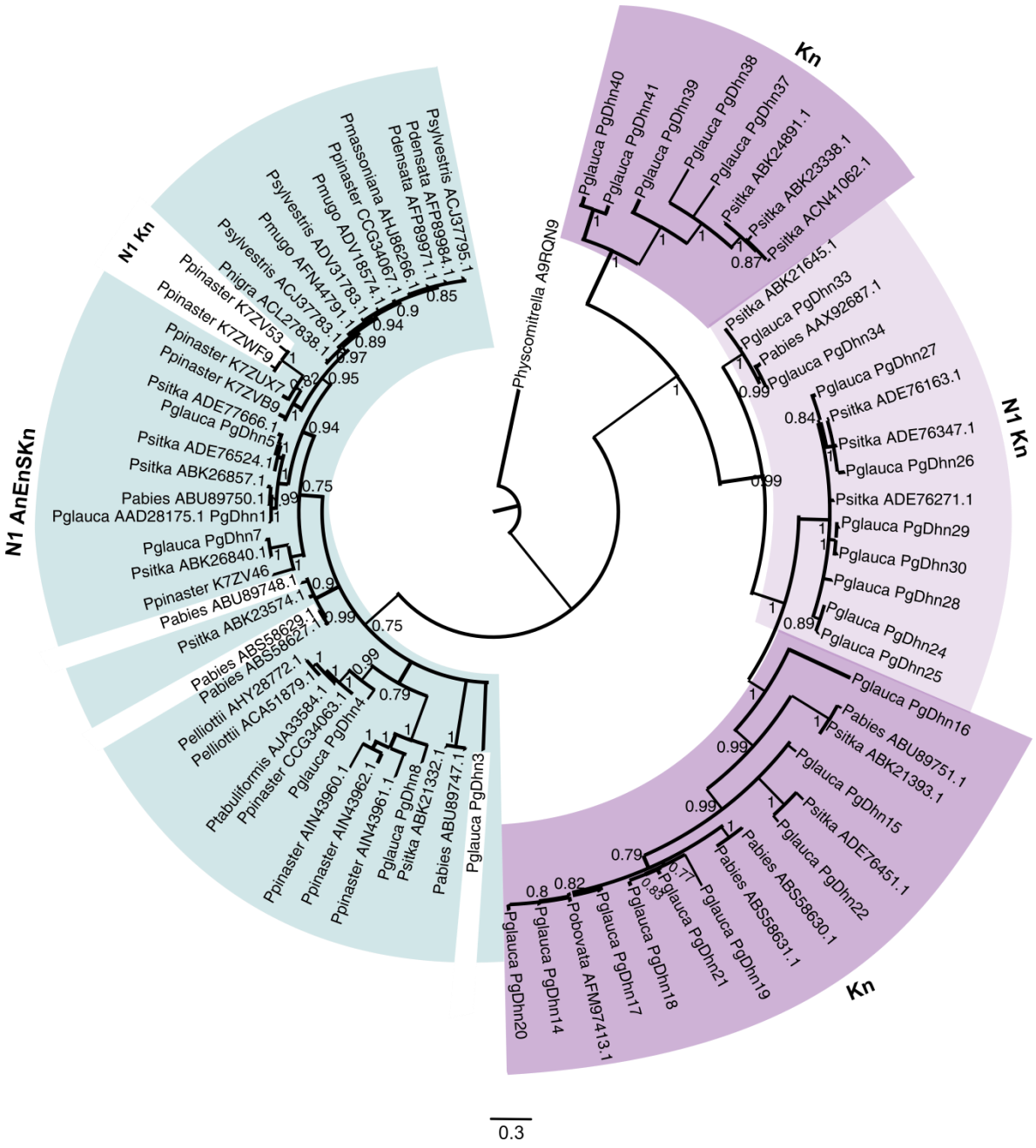


Figure 3.1- Phylogeny of the conifer dehydrin gene family represented by a consensus tree from Bayesian analysis, with threshold support equal or superior to 0.75. We used 41 white spruce dehydrins and 37 other conifer sequences; see details of sequence clusters in Table S3.2. A dehydrin from *Physcomitrella* was used as the root. The phylogeny was obtained with MrBayes after protein alignment with MAFFT, and visualized with FigTree.

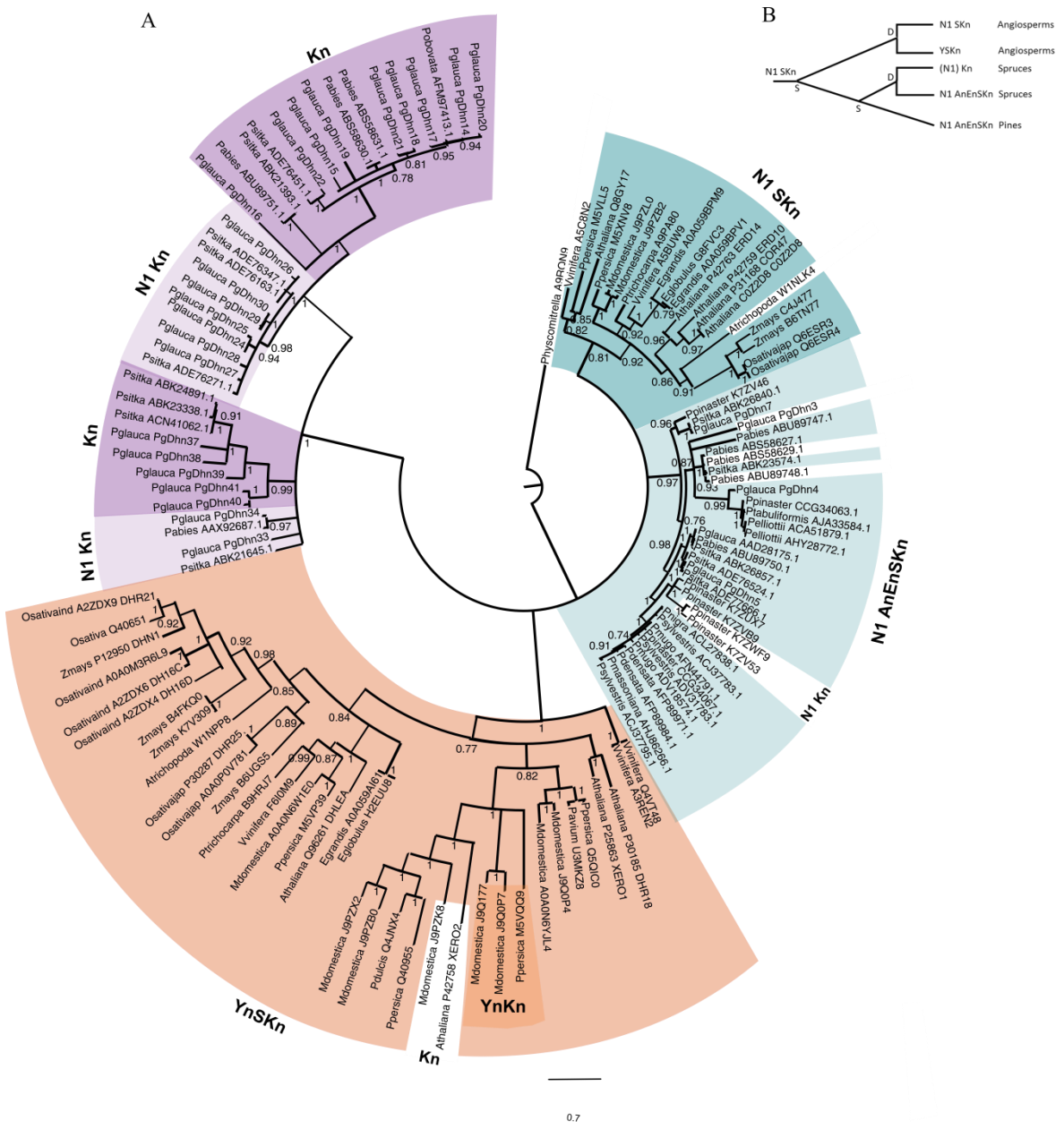


Figure 3.2- A) Phylogeny of the angiosperm and conifer dehydrin gene family represented by a consensus tree from Bayesian analysis, with threshold support equal or superior to 0.75. A dehydrin from the moss *Physcomitrella* was used to root the tree. The phylogeny was obtained with MrBayes after protein alignment with MAFFT. B) Synthesis of speciation and gene duplication events with N1 Skn type as ancestor; D - duplication, S - speciation.

3.5.3 A major duplication is uniquely detected in the genus *Picea*

The conifer phylogenetic tree revealed diversification patterns supported by high *a posteriori* probabilities and this led to the hypothesis that a major duplication event occurred within the lineage giving rise to the genus *Picea*, after the split between *Picea* and *Pinus* (Fig.3.1). The most parsimonious interpretation of the tree is that this event then gave rise to the Kn group which is specific to *Picea,* and that this group underwent several further duplications generating the several groups of dehydrins in white spruce.

3.5.4 Degenerate K-segments and structural variations in conifer dehydrins

Based on the amino acid motifs and in congruence with results from phylogenetic analyses, all 135 representative dehydrins from conifers and angiosperms were classified into four major amino acid structures: (1) N1 AnEnSKn with variations in the presence of A and E-segments; (2) N1 SKn and (3) N1 Kn with some variations in the presence of N1-segment and (4) YnSKn. Almost all dehydrins contained one or multiple K-segments with the exception of one sequence in maritime pine (*Pinus pinaster*) (model 10, Fig. 3.3, Table S3.4) as previously reported (Perdiguero et al., 2014).

The K-segment was degenerate (p-value < 1.1e-5) in many conifer dehydrins such as *PgDhn8* and *PgDhn9*, among others (see Table S3.4), and was much more conserved in angiosperms (Fig. 3.3 and Fig. S3.1). In some cases, the K-segment was highly degenerated to the point of not being identified by motif sequence similarity; in these cases we used the sequence alignment to identify these K-segments.

In the group N1 AnEnSKn, we identified 12 structural variations, some of which have not been described before. For example, *PgDhn8* and *PgDhn9* (model 9 and 11, Fig.3.3), which lack an E-segment. In the Kn group in spruces we observed a high number (>6) of K-segments never reported in other conifers.

We observed that sequence similarity and amino acid motif structure were not always congruent, i.e. some sequences were grouped together in the phylogenetic tree although their amino acid motif classification differed (Fig. 3.3). This could be explained by the

considerable variation in amino acid sequence that exists within the motifs (Fig. S3.1). Similarly, dehydrins that were not regrouped tightly on the phylogenetic tree may share the same motif structure. For instance, the model 1 from maritime pine (N1 K2) was not grouped with other N1 K2 sequences such as *PgDhn33* and *PgDhn34* (model 25). Similarly, angiosperm Kn dehydrins were not grouped with conifer Kn dehydrins (models 18 and 19) and, maritime pine sequences from model 12, classified as SK, were not grouped with the angiosperm sequences SKn.

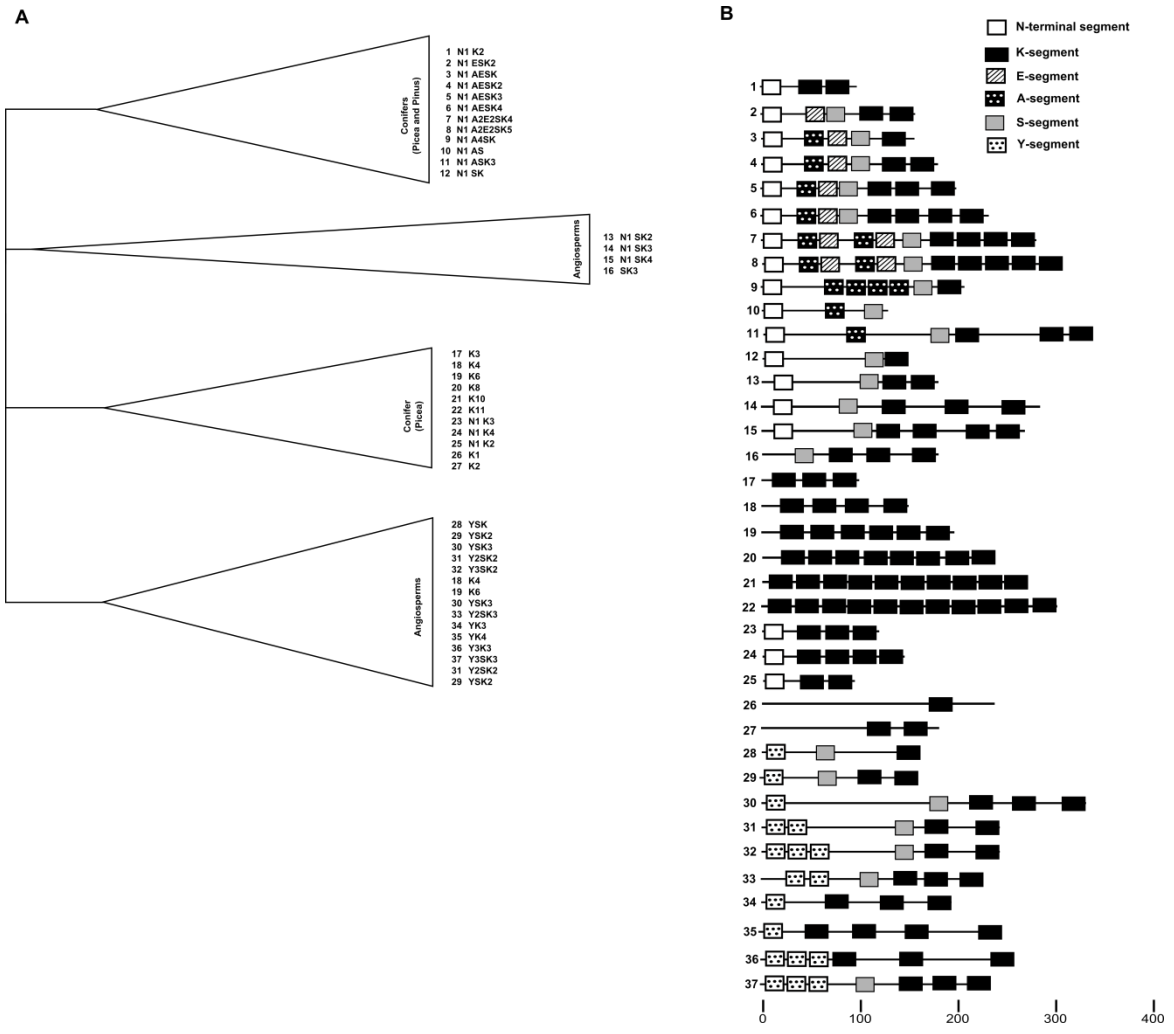


Figure 3.3- Conifer and angiosperm dehydrins classification based on their amino-acid motifs. A) Sequences were grouped by similarity and classified by motif composition. B) Each dehydrin type was represented showing the variation in number of motifs.

3.5.5 Dehydrin expression varies between different tissues and conditions

We designed gene-specific primer pairs for the white spruce dehydrins with complete ORFs to evaluate their RNA accumulation profile by RT-qPCR (Table S3.1). Given the large number of sequences and high levels of similarity, the assay specificity was verified by preliminary tests in which amplicons were sequenced for validation. Next, we surveyed RNA transcript levels in three different tissues, phelloderm, xylem and foliage from plants growing under non-limiting conditions. Reliable transcript detection was recorded for thirteen of the white spruce dehydrins in at least one of the tissues (Fig. 3.4). Amplifications that lacked specificity were eliminated from the analyses and dehydrin sequences producing no detectable product were considered to have specificity to other tissues or to other biological conditions.

The gene-by-gene analysis of variance with expression data as a function of the type of tissue identified nine genes with differential expression (Table S3.5). The multiple comparison tests showed that five of the sequences produced preferentially expressed transcripts in the phelloderm, four in the foliage and three in the secondary xylem, and the four remaining sequences did not vary between tissues (Fig. 3.4).

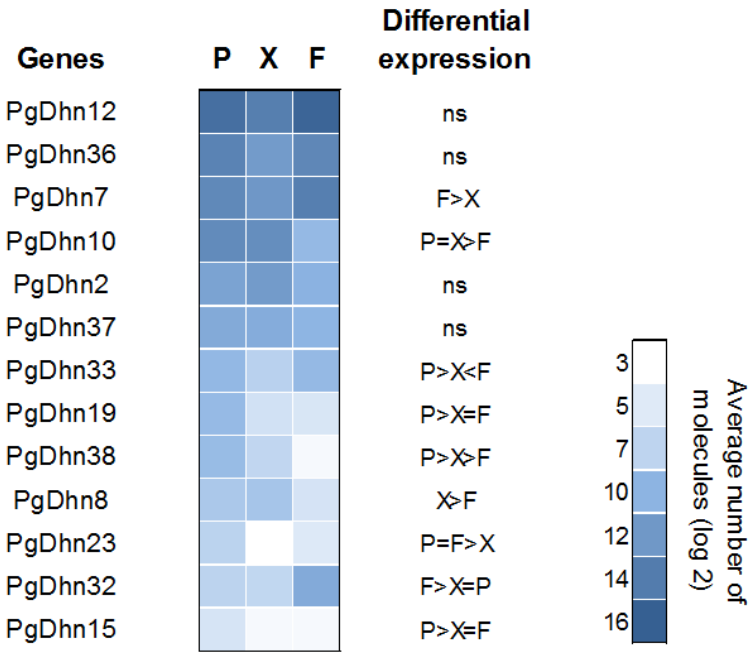


Figure 3.4- Transcript accumulation profiles from F (Foliage), X (Xylem) and P (Phelloderm) measured by qPCR. Significant differences between tissue expression levels are indicated on the right side, ANOVA, Tukey's HSD (P <0.05; ns indicates no significant difference between the expression level among the three tissues).

3.5.6 Members of the dehydrin family respond differently to water stress

We conducted a greenhouse experiment with three white spruce genotypes comparing dehydrin transcript accumulation profiles in well-watered and non-watered plants at several time points. Starting from 14 days of treatment, statistically significant differences in water potential were detected between the well-watered and non-watered plants (Fig. 3.5) and the response was similar among the three genotypes (Table S3.6).

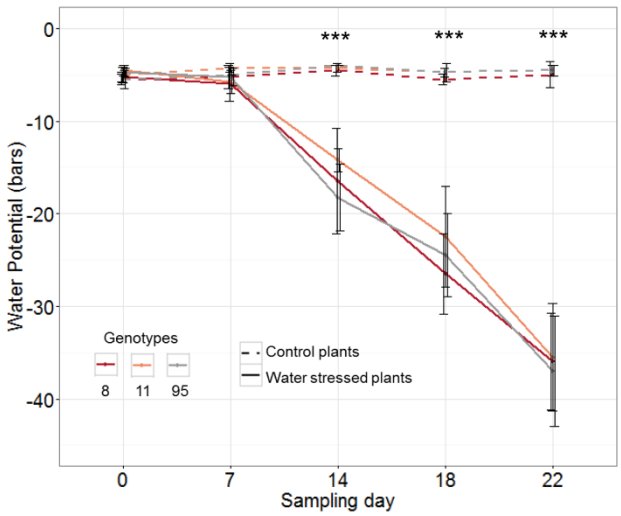


Figure 3.5- Midday water potential in needles of well-watered plants (dashed line) and unwatered plants (solid line) in three different genotypes (clones 8, 11 and 95). The water potential of water-stressed plants was compared with that of control plants for each sampling date in all three genotypes (ANOVA, Tukey's HSD, *** P <0.001).

We were able to reliably detect transcript abundance in the foliage for ten of the dehydrins in well-watered and non-watered plants. Gene-by-gene analysis of variance followed by multiple comparison tests showed that, under the same conditions, there was no statistical difference among the expression pattern of the three clones, except for the gene *PgDhn10* where after 7 days under water stress, a significant difference in gene expression level between clones 11 and 95 was detected (Table S3.7).

Eight genes had statistically different expression levels between watered and non-watered plants at least for one sampling date. The genes *PgDhn10*, *PgDhn16*, *PgDhn33* and *PgDhn35* showed a remarkable increase of gene expression in water-stressed plants compared to well-watered plants. Their increased transcript levels were statistically significant starting at day 14, which coincides with the changes in water potential. The genes *PgDhn7*, *PgDhn9* and *PgDhn12* showed a slight increase in expression in stressed plants compared to watered plants after 18 to 22 days of treatment. Only the gene *PgDhn36* showed a decrease in expression, which was slight and was observed only after 22 days without watering (Fig. 3.6).

We examined the transcript profiles by comparing the two major classes of dehydrins found in spruces, N1 AnEnSKn and N1 Kn. Among the six genes that had increased transcript levels in response to water stress, four were of the N1 AnEnSKn type (N1 AESK2; N1 A2E2SK4; N1 ASK3, N1 ASK2) and two were of the N1 Kn type (N1 K2; K4). The genes with slightly decreasing and those with no response to water stress, were classified as K1 (2 sequences) and K6. Taken together, these classifications indicate that diverse dehydrin sequences are water-stress responsive and that neither of the two major classes appears to have a clearly characteristic profile; however, the number of genes assayed only represents 27% of the white spruce dehydrin sequences identified.

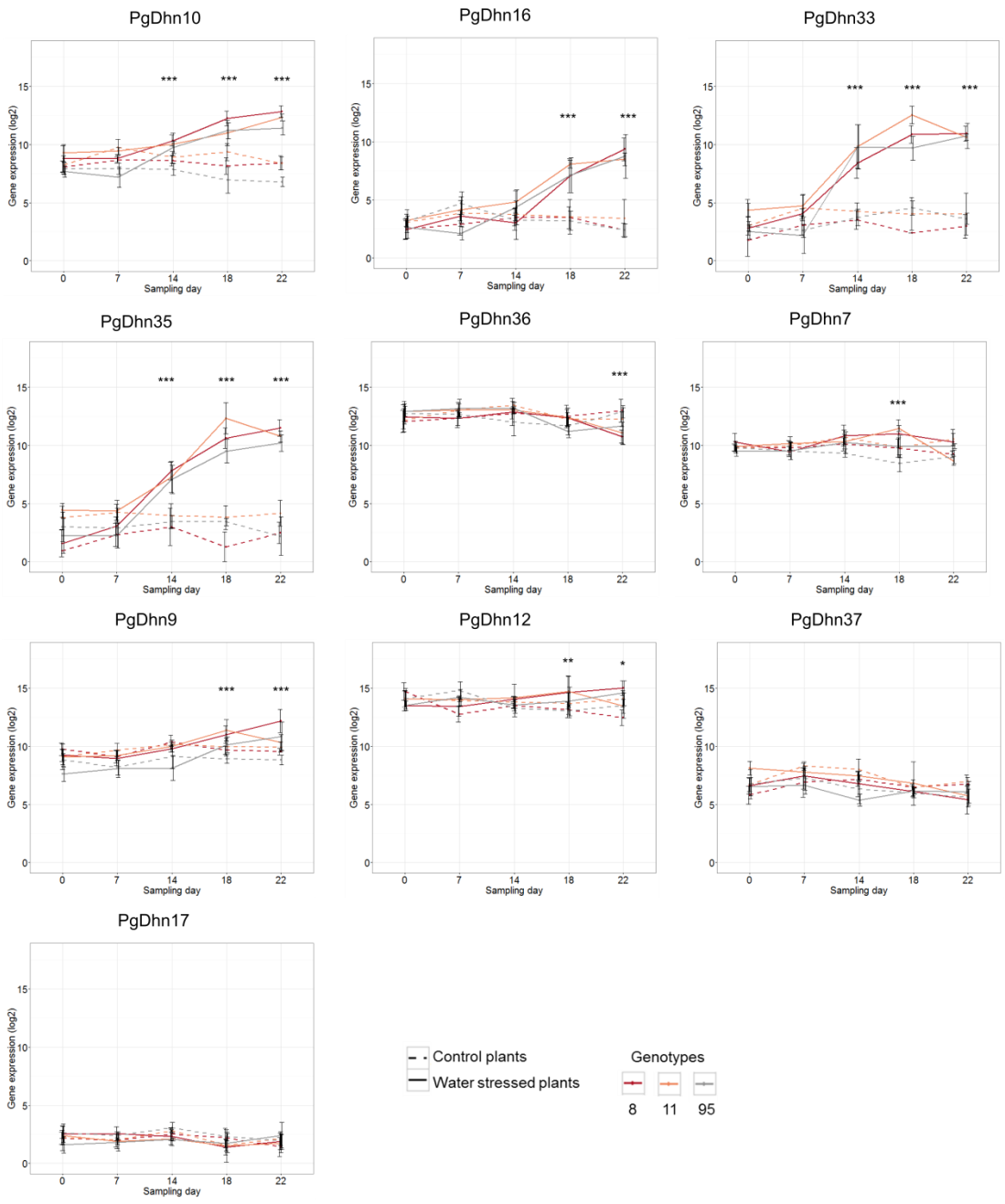


Figure 3.6- Expression profile of dehydrin genes during 22 days of treatment. The gene expression of water-stressed plants (solid lines) was compared with that of control plants (dashed lines) for each sampling date in all three genotypes (clones 8, 11 and 95). ANOVA, Tukey test, *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

3.6 Discussion

We identified and classified a large number of dehydrins and the results indicate that they form a large gene family in conifers and are particularly abundant in spruces. Conifer dehydrins appeared structurally diverse and notable differences were observed when compared to angiosperm dehydrins, including poorly conserved K-segments and conifer-specific segments. We identified nine dehydrins with differential tissue expression under normal conditions and eight dehydrins that respond to drought stress. The most strongly induced dehydrins were classified as Kn type. Below, we discuss these results against a backdrop of speciation and adaptation to abiotic factors.

3.6.1 Structural diversity in dehydrin protein sequences

We identified a large number of dehydrins detected in *Picea* but did not find any new conserved amino acid motifs. All of the white spruce dehydrins presented at least one K-segment and many of them contained A and E-segments together, as previously found in maritime pine (Perdiguero et al., 2012). Here, we showed that the A-segment may be less conserved than previously reported (Perdiguero et al., 2012). In contrast to angiosperms, no Y-segments were identified in conifers (Campbell and Close, 1997; Zolotarov and Stromvik, 2015). The previously reported conserved N-terminal sequence (N1) had not been considered as a *bona fide* protein motif (Perdiguero et al., 2012) but we have included it in our structural classification of protein sequence.

We classified all dehydrins represented in the phylogenetic trees based on the amino acid motifs (Fig. S3.1) (Fig. 3.3). The angiosperm and conifer sequences were classified among 37 different models showing a wide diversity of dehydrin protein structures varying in the composition and number of motifs (Fig. 3.3). White spruce dehydrins were classified mainly as AnEnSKn and (N1) Kn types. As was found in other conifers (Perdiguero et al., 2012), we observed a variation in the number of A, E and K-segments. However, for the first time we report conifer dehydrin genes containing more than six K-segments, including *PgDhn21*, *PgDhn22* and *PgDhn23* (Table S3.4), and dehydrins classified as AnSK, which harbor the A-segment but lack the E-segment (*PgDhn8* and *PgDhn9*). The structural

diversity is present not only between the different structure types but also within each type. Some conifer and angiosperm dehydrins were classified as Kn type but their amino acid sequences were highly divergent. We also observed that the amino-acid sequence of the K-segment was variable among conifer dehydrins, in contrast to angiosperm dehydrins which contain more conserved K-segment sequences (Fig. S3.1). *In vitro* assays suggested that the K-segments play a key role in the protective function of dehydrins in preventing deleterious changes in protein secondary and tertiary structure (Reyes et al., 2008). It remains to be elucidated whether the reduced conservation of K-segments that is observed in conifers has functional implications and whether the protective function is maintained. The S-segment also appears to be important in dehydrin function. It may be important as putative phosphorylation sites (Jensen et al., 1998), being involved in post-translational protein modifications impacting tolerance to drought and salt (Brini et al., 2007). It is also possible that the A, E and the N1-terminal motifs play important roles in protein conformation and function, but this remains to be tested. The modular conformation of dehydrins, the large variation seen in the number and position of the different motifs, and their patterns of expression not connected tightly to structural differences and phylogenetic grouping (see below) is consistent with subfunctionalization following duplication events, as also reported for conifer transcription factors (Guillet-Claude et al., 2004). Dehydrin genes were also found to diverge very rapidly in *P. glauca* at the nucleotide sequence level, the family showing among the highest ratios of nonsynonymous to synonymous substitutions (A/S) among more than 2000 gene families analysed (Pavy et al., 2013). Such highly positive ratios indicate a more rapid evolution at the amino acid sequence level than expected if genes were not under positive selection (neutrality). A similar pattern was also reported between *P. glauca* and *P. abies* (De La Torre et al., 2015). Gene families with high A/S ratio were also those with the highest heterogeneity of gene expression across white spruce tissues (Pavy et al., 2013), supporting the notion of rapid subfunctionalization with obvious implications for adaptive potential.

3.6.2 Dehydrin gene family evolution and expansion in spruce

We first identified 53 dehydrins in the white spruce gene catalog, of which 41 had a complete ORF. Further sequence discovery in RNA-Seq datasets (Verta et al. 2016) and clustering based on sequence similarity suggested that spruces contained up to 56 distinct dehydrin genes (Table S3.2). This relatively large number of dehydrins exceeds that observed in other plant species. For example, 10 genes were identified in mouse-ear cress and poplar (Hundertmark and Hincha, 2008; Liu et al., 2012), 12 were found in apple (Liang et al., 2012), eight in rice (Wang et al., 2007), six in each of peach and maize (Basset, Fisher and Ferrel, 2015; Pfam 30.0, Finn et al., 2014), four in grapevine (Yang et al., 2012) and only two in the primitive *Amborella trichopoda* (Pfam 30.0, Finn et al., 2014). Previous studies in other conifers, such as *Pinus pinaster*, *Picea obovata* and *Picea abies*, reported less than ten dehydrins per species (Perdiguero et al., 2012; Joosen et al., 2006; Yakovlev et al., 2008; Kjellsen et al., 2013), which could reflect sampling effects. We carried out an exhaustive search for dehydrin homologs in conifers and found more dehydrins in both *Picea glauca* and *Picea sitchensis* (Table S3.2) than in pine species, suggesting that the *Picea* genus may have more dehydrins than angiosperms and *Pinus*. On the other hand, this may be the consequence of sampling effects since full-length cDNA have been more extensively explored in *P. glauca* and *P. sitchensis* (Ralph et al., 2008; Rigault et al., 2011) than in *Pinus* spp. We performed separate and combined phylogenetic analysis of angiosperm and conifer dehydrins (Figs. 3.1 and 3.2). The dehydrins were distributed into four main groups paralleling structural differences: two angiosperm groups, with YnSKn and N1 SKn amino acid structures, and two conifer groups, with N1 Kn and AnEnSKn amino acid structures.

The combined angiosperm and conifer phylogenetic tree suggests an interesting evolutionary history for this gene family. The simplest hypothesis is that the most ancestral gene had a structure most similar to the N1 SKn and N1 AnENSKn sister group types, in which is reflected by their highest taxonomical representation including conifer, dicot and monocot sequences, and that sequences diverged through gene duplications as well as loss and acquisition of amino acid motifs in a parallel fashion, which occurred largely after the

split of angiosperms and gymnosperms, around 300 Myr (Savard et al., 1994). This most parsimonious interpretation assumes there were no major gene losses in the taxa analyzed. In line with this interpretation, a major duplication would have occurred very early in the angiosperm lineage, before the split between monocots and dicots (140-150 Myr) (Chaw et al., 2004), and giving rise to the Yn SKn structure type, which is present along with the N1 SKn type in all angiosperms tested, including monocots and dicots. This new gene duplicate would have lost the N1 motif and acquired a Yn motif. Within the conifers, a *Picea*-specific duplication, i.e. occurring after the split of *Picea* and *Pinus* around 120 to 140 Myr (Savard et al., 1994), likely gave rise to the Kn and N1 Kn groups found only in spruce with only two exceptions (Fig. 3.2). The topology was not well resolved at the root of this spruce-only group, but it suggests that more than one duplication may have occurred where the N1 motif was likely lost. The presence of the Kn motif in two *Pinus pinaster* sequences located within the conifer N1 type sequences suggests parallel evolution, which indicates that under certain environmental pressures, specific amino acid sequence motifs could re-emerge sporadically. In addition to the major gene duplication affecting *Picea* only and the sporadic resurgence of amino acid sequence motifs, several other duplications have been observed at various stages, most frequently in the conifers. These duplications have obviously impacted the size of the dehydrin gene family especially in *Picea*, with likely implications on adaptation. Taken together, these results suggest that the higher diversification rate of dehydrin genes seen in the conifers, compared to angiosperms, might be related to long-term genetic adaptation to a spatially and temporarily more heterogeneous environment throughout the evolution and diversification of the lineage. Conifers are long-lived species that often colonize extreme habitats, as seen for boreal conifers such as white spruce; therefore, it is likely that larger families of key genes related to adaptation could confer more plasticity and survival ability through sub-functionalization.

Duplicated genes may result either from whole-genome duplication (WGD) or from more localized segmental or single-gene duplications (Blanc and Wolfe 2004; Cannon et al., 2004). Many WGD events have been detected in angiosperms. For example, *Arabidopsis* has experienced at least three WGD including an event that was shared by all eudicots

(Bowers et al., 2003). Angiosperms have also experienced lineage-specific WGD, some of which were reported in forest trees, for instance in *Eucalyptus* (Myburg et al., 2014) and in the Salicaceae (Tuskan et al., 2006). The expansion of the dehydrin gene family in angiosperms is thus the consequence of both tandem duplication and WGD events (Liu et al., 2012; Hundertmark and Hincha, 2008; Wang et al., 2007). In the conifers, the Pinaceae were recently reported to have experienced two WGD events (Li et al., 2015). One ancient event would have been shared with all seed plants including angiosperms, as well as the Cupressaceae and the Taxaceae (which could not be sampled in the present study), while another WDG would have occurred in the common ancestor of Pinaceae only. Although the topology of the dehydrin phylogenetic tree lacks resolution near the origin (Fig. 3.2), no clear evidence was seen that could support either of these ancient WGD events. The angiosperm sequences were split into two large groups, which is likely the consequence of a major duplication event at least preceding the monocot-dicot divergence and likely involving WGD. However, the lack of intervening conifer sequences in each of these groups would indicate that the duplication event occurred after the angiosperm-gymnosperm split, or alternatively, that the ancient duplicated copy had been lost in the lineage leading to conifers if this event had occurred before the angiosperm-gymnosperm split, as previously reported. Similarly, the conifer sequences were split into two major groups (Fig. 3.2) without any intervening angiosperm sequences, and with one group represented by *Picea* sequences only (Kn and N1 Kn types). This pattern does not support either a WGD common to all Pinaceae because of the lack of pine sequences in this group. Rather, the dehydrin tree topology suggests a quite recent duplication event in the spruce common ancestor after the pine-spruce lineage split. More intensive sampling of dehydrins in other conifer genera and families should help ascertain these interpretations and better understand the evolutionary history of conifer dehydrins.

3.6.3 Expression of dehydrin genes in developmental and stress responses

Dehydrins have been reported as multifunction proteins that accumulate during seed formation and are present in vegetative tissues under normal conditions (Bies-Ethève et al., 2008; Campbell and Close, 1997). They have been linked to protective functions (Brini et

al., 2010; Reyes et al., 2008), chaperone activity (Kovacs et al., 2008), water-binding capacity (Rinne et al., 1999), and to an antioxidant role (Hara et al., 2004). The expression of many dehydrin genes changes in response to abiotic stress conditions such as drought, salt and cold (Close 1996), as well to biotic stress such as wounding and infection (Richard et al., 2000; Hundertmark and Hincha, 2008; Yang et al., 2012). Expression has also linked some dehydrins to growth processes such as spring bud burst in conifers (Yakovlev et al., 2008).

We identified nine dehydrins in white spruce with differential expression when comparing three different tissues (foliage, secondary xylem, and phelloderm) under normal conditions. However, these differences were not tightly linked to structural types or phylogenetic groups. Other dehydrins in angiosperms also showed differential expression between tissues types. *Arabidopsis* dehydrins *AtLEA2-5*, *AtLEA2-6* and *AtLEA2-7* were expressed in seeds while *AtLEA2-1*, *AtLEA2-2* and *AtLEA2-4* were strongly expressed in vegetative tissues (Bies-Ethève et al., 2008). In apple, five dehydrins were expressed in flowers, seeds, leaves, fruit, and roots and another four in a subset of these tissues (Liang et al., 2012). These observations indicate that different dehydrins may have acquired a degree of tissue specificity and suggests that some dehydrins are important to plant development.

Many studies showed angiosperms dehydrins were induced by drought, cold stress or both and many of them were classified as YnSKn and grouped together in our phylogenetic analysis (Fig. 3.2) They include between one and five dehydrin genes in rice (Wang et al., 2007), grape (Yang et al., 2012), *Arabidopsis* (one Kn sequence) (Hundertmark and Hincha 2008), peach (Basset et al., 2015), apple (Liang et al., 2012) and *Eucalyptus* (Fernández et al., 2012). However, some of the dehydrins in this group, like *PpDhn5* in peach and *AtLEA2-7* in *Arabidopsis* were neither induced by cold or drought stress (Basset et al., 2015; Hundertmark and Hincha 2008). The other group of angiosperm dehydrins in the phylogenetic tree has the typical N1 SKn structure and includes sequences that are cold responsive (*AtLEA2-1* and *2*; *EuglDhn2*) and sequences that had either no detectable transcripts or very low expression under drought and cold stress (Hundertmark and Hincha 2008; Liang et al., 2012; Basset et al., 2015; Yang et al., 2012).

Our analysis identified eight dehydrins that responded to water stress in white spruce. Four of them increased their transcript levels several fold after several days without watering. They were classified as N1 K2, K4, and N1 AESK2 (*PgDhn10*, *16*, *33* and *35*), indicating that the two main conifer groups in the phylogenetic tree comprise dehydrin sequences that are drought-stress responsive. In Norway spruce and maritime pine, N1 K2 dehydrins had among the highest transcript levels after a period of water stress (Eldhuset et al., 2012, Perdiguero et al., 2012). In maritime pine, a N1 AESK2 dehydrin presented a very similar transcript accumulation pattern (Perdiguero et al., 2012).

An interesting observation is that both N1 K2 dehydrins from maritime pine were grouped with the N1AnEnSKn cluster (light blue) in the phylogenetic tree, while N1 K2 dehydrins from white spruce fell in the (N1) Kn group (purple). The maritime pine genes harbored sequences that are more similar to N1 AnEnSKn dehydrins but their structure is closer to that of the N1 K2 cluster, suggesting parallel evolution in which selective forces may have shaped these proteins to carry out the same function. Similar observations of parallel evolution have been made from the identification of different adaptive genes to climatic factors among different Pinaceae taxa, but pertaining to same large gene families (Prunier et al., 2011).

The spruce dehydrins *PgDhn 7*, *9*, *12*, *17*, *36* and *37*, including both N1 AnEnSKn and N1 (Kn) types, were less responsive to drought conditions such as reported for N1 AESK (a and b) in maritime pine (Perdiguero et al., 2012) (Fig. 3.6). Considering the diverse roles attributed to dehydrins, they may be more responsive to other types of stress, as observed in Siberian spruce where *Dhn 2* (N1 AESK3) and *Dhn Cap1.1* (K6) were induced by cold conditions (Kjellsen et al., 2013). These observations and findings that some dehydrins are more strongly expressed in other organs including roots or stem in response to water stress (Perdiguero et al., 2012; Lorenz et al., 2011; Eldhuset et al., 2012) indicate that our analysis is likely to reveal a partial picture of their whole range of expression in conifers.

3.7 Conclusions

Our results indicate that the dehydrin gene family is larger in conifers than in angiosperms, and suggest that a major duplication contributed to a lineage-specific expansion in the genus *Picea*. The present results also suggest that subfonctionalization rather than neofunctionalization appears to be the main driver for the increased diversity of dehydrins in conifers, with diversification implicating loss and gain of structural motifs.

The dehydrin gene family has been well studied in angiosperms and has been linked to a variety of cellular processes. The diversity of dehydrin sequences, together with their tissue-preferential and drought-responsive expression, suggests that they are involved in a variety of physiological processes in spruce. Further experiments including additional assessments of stress responsiveness will likely be needed to shed more light onto the potential processes in which they are involved. The N1 K2 and N1 AESK2 dehydrins were very responsive to water stress in conifers. Studies involving diverse genotypes and genetic experiments could reveal the potential of these genes as molecular markers for tolerance to drought.

In the next decades, the boreal biome is expected to experience the largest increase in temperatures of all forest biomes and drought–induced mortality is predicted to increase (Gauthier et al., 2015). An improved understanding of the molecular response of conifers to drought will be highly useful to design diagnostic tools to help map and conserve the natural genetic diversity that is relevant for adaptation to drought stress in order to maintain a healthy boreal forest.

3.8 Acknowledgements

The authors thank François Larochelle and Marie-Andrée Paré (both of Université Laval) for assistance with plant materials and the drought stress experiment. Marie R. Coyea (Université Laval) for advice and assistance for the water potential measurements. Stéphane Daigle from Centre for Forest Research for statistical advices. Elie Raherison, Mebarek Lamara, Benjamin Dufils and Sébastien Caron helped with plant sampling. Funding was

received from Génome Québec and Genome Canada for the SmarTForests project (JB and JM), and form NSERC of Canada for a discovery grant (JM).

3.9 References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215:403–10.

Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Research. 2015;gkv416.

Bassett CL, Fisher KM, Jr REF. The complete peach dehydrin family: characterization of three recently recognized genes. Tree Genetics & Genomes. 2015;11:1–14.

Beaulieu J, Giguère I, Deslauriers M, Boyle B, MacKay J. Differential gene expression patterns in white spruce newly formed tissue on board the International Space Station. Advances in Space Research. 2013;52:760–72.

Bies-Ethève N, Gaubier-Comella P, Debures A, Lasserre E, Jobet E, Raynal M, et al. Inventory, evolution and expression profiling diversity of the LEA (late embryogenesis abundant) protein gene family in *Arabidopsis thaliana*. Plant Molecular Biology. 2008;67:107–24.

Blanc G, Wolfe KH. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell. 2004;16:1667–78.

Bowers JE, Chapman BA, Rong J, Paterson AH. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature. 2003;422:433–8.

Boyle B, Dallaire N, MacKay J. Evaluation of the impact of single nucleotide polymorphisms and primer mismatches on quantitative PCR. BMC Biotechnology. 2009;9:75.

Brini F, Hanin M, Lumbreras V, Irar S, Pagès M, Masmoudi K. Functional characterization of *DHN-5*, a dehydrin showing a differential phosphorylation pattern in two Tunisian durum wheat (*Triticum durum* Desf.) varieties with marked differences in salt and drought tolerance. Plant Science. 2007;172:20–8.

Brini F, Saibi W, Amara I, Gargouri A, Masmoudi K, Hanin M. Wheat Dehydrin *DHN-5* Exerts a Heat-Protective Effect on β-Glucosidase and Glucose Oxidase Activities. Bioscience, Biotechnology and Biochemistry. 2010;74:1050–4.

Campbell SA, Close TJ. Dehydrins: genes, proteins, and associations with phenotypic traits. New Phytologist. 1997;137:61–74.

Cannon SB, Mitra A, Baumgarten A, Young ND, May G. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. BMC Plant Biology. 2004;4:10.

Chang S, Puryear J, Cairney J. A simple and efficient method for isolating RNA from pine trees. Plant Molecular Biology Reporter. 1993;11:113–6.

Chaw S-M, Chang C-C, Chen H-L, Li W-H. Dating the monocot–dicot divergence and the origin of core eudicots using whole chloroplast genomes. Journal of Molecular Evolution. 2004;58:424–41.

Close TJ. Dehydrins: Emergence of a biochemical role of a family of plant dehydration proteins. Physiologia Plantarum. 1996;97:795-803.

De La Torre AR, Lin Y-C, Van de Peer Y, Ingvarsson P. Genome-wide analysis reveals diverged patterns of codon bias, gene expression, and rates of sequence evolution in *Picea* gene families. Genome Biology and Evolution. 2015;7:1002-1015.

Eldhuset TD, Nagy NE, Volařík D, Børja I, Gebauer R, Yakovlev IA, et al. Drought affects tracheid structure, dehydrin expression, and above- and belowground growth in 5-year-old Norway spruce. Plant and Soil. 2012;366:305–20.

Farooq M, Hussain M, Wahid A, Siddique KHM. Drought Stress in Plants: An Overview. In: Aroca R, editor. Plant Responses to Drought Stress. Springer Berlin Heidelberg; 2012. p. 1–33.

Fernández M, Valenzuela S, Barraza H, Latorre J, Neira V. Photoperiod, temperature and water deficit differentially regulate the expression of four dehydrin genes from *Eucalyptus globulus*. Trees. 2012;26:1483–93.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Research. 2014;42:D222–30.

Gauthier S, Bernier P, Kuuluvainen T, Shvidenko AZ, Schepaschenko DG. Boreal forest health and global change. Science. 2015;349:819–22.

Guillet-Claude C, Isabel N, Pelgas B, Bousquet J. The Evolutionary Implications of *knox-I* Gene Duplications in Conifers: Correlated Evidence from phylogeny, gene mapping, and analysis of functional divergence. Molecular Biology and Evolution. 2004;21:2232–45.

Hanin M, Brini F, Ebel C, Toda Y, Takeda S, Masmoudi K. Plant dehydrins and stress tolerance. Plant Signaling & Behavior. 2011;6:1503–9.

Hara M, Fujinaga M, Kuboi T. Radical scavenging activity and oxidative modification of citrus dehydrin. Plant Physiology and Biochemistry. 2004;42:657–62.

Hundertmark M, Hincha DK. LEA (late embryogenesis abundant) proteins and their encoding genes in *Arabidopsis thaliana*. BMC Genomics. 2008;9:118.

Jensen AB, Goday A, Figueras M, Jessop AC, Pagès M. Phosphorylation mediates the nuclear targeting of the maize Rab17 protein. The Plant Journal. 1998;13:691–7.

Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics. 2010;11:431.

Joosen RVL, Lammers M, Balk PA, Brønnum P, Konings MCJM, Perks M, et al. Correlating gene expression to physiological parameters and environmental conditions during cold acclimation of *Pinus sylvestris*, identification of molecular markers using cDNA microarrays. Tree Physiology. 2006;26:1297–313.

Katoh K, Standley DM. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution. 2013;30:772–80.

Kjellsen TD, Yakovlev IA, Fossdal CG, Strimbeck GR. Dehydrin accumulation and extreme low-temperature tolerance in Siberian spruce (*Picea obovata*). Tree Physiology. 2013;33:1354–66.

Kovacs D, Kalmar E, Torok Z, Tompa P. Chaperone Activity of *ERD10* and *ERD14*, two disordered stress-related plant proteins. Plant Physiology. 2008;147:381–90.

Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, et al. Early genome duplications in conifers and other seed plants. Science Advances. 2015;1:e1501084.

Liang D, Xia H, Wu S, Ma F. Genome-wide identification and expression profiling of dehydrin gene family in *Malus domestica*. Molecular Biology Reports. 2012;39:10759–68.

Liu C-C, Li C-M, Liu B-G, Ge S-J, Dong X-M, Li W, et al. Genome-wide identification and characterization of a dehydrin gene family in poplar (*Populus trichocarpa*). Plant Molecular Biology Reporter. 2012;30:848–59.

Lorenz WW, Alba R, Yu Y-S, Bordeaux JM, Simões M, Dean JF. Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (*P. taeda L.).* BMC Genomics. 2011;12:264.

Micco VD, Aronne G. Morpho-anatomical traits for plant adaptation to drought. In: Aroca R, editor. Plant responses to drought stress. Springer Berlin Heidelberg; 2012. p. 37–61.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, et al. The genome of *Eucalyptus grandis*. Nature, 2014; 510:356-62.

Pavy N, Boyle B, Nelson C, Paule C, Giguère I, Caron S, et al. Identification of conserved core xylem gene sets: conifer cDNA microarray development, transcript profiling and computational analyses. The New Phytologist. 2008;180:766–86.

Pavy N, Deschênes A, Blais S, Lavigne P, Beaulieu J, Isabel N, Mackay J, Bousquet J. The landscape of nucleotide polymorphism among 13,500 0 genes of the conifer *Picea glauc*a, relationships with functions, and comparison with *Medicago truncatula*. Genome Biology and Evolution. 2013;5:1910-1925.

Perdiguero P, Barbero MC, Cervera MT, Soto Á, Collada C. Novel conserved segments are associated with differential expression patterns for *Pinaceae* dehydrins. Planta. 2012;236:1863–74.

Perdiguero P, Collada C, Soto Á. Novel dehydrins lacking complete K-segments in *Pinaceae*. The exception rather than the rule. Frontiers in Plant Science. 2014 ;5:682.

Prunier J, Laroche J, Beaulieu J, Bousquet J. Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. Molecular Ecology. 2011;20:1702–16.

Raherison E, Rigault P, Caron S, Poulin P-L, Boyle B, Verta J-P, et al. Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression. BMC Genomics. 2012;13:434.

Raherison ESM, Giguère I, Caron S, Lamara M, MacKay JJ. Modular organization of the white spruce (*Picea glauca*) transcriptome reveals functional organization and evolutionary signatures. New Phytologist. 2015;207:172–87.

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

Reyes JL, Campos F, Wei H, Arora R, Yang Y, Karlson DT, et al. Functional dissection of hydrophilins during in vitro freeze protection. Plant Cell Environ. 2008;31:1781–90.

Richard S, Morency M-J, Drevet C, Jouanin L, Séguin A. Isolation and characterization of a dehydrin gene from white spruce induced upon wounding, drought and cold stresses. Plant Molecular Biology. 2000;43:1–10.

Rigault P, Boyle B, Lepage P, Cooke JEK, Bousquet J, MacKay JJ. A white spruce gene catalog for conifer genome analyses. Plant Physiology. 2011;157:14–28.

Rinne PLH, Kaikuranta PLM, Plas LHW van der, Schoot C van der. Dehydrins in cold-acclimated apices of birch (*Betula pubescens Ehrh*.): production, localization and potential role in rescuing enzyme function during dehydration. Planta. 209:377–88.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology. 2012;61:539–42.

Rutledge RG, Stewart D. A kinetic-based sigmoidal model for the polymerase chain reaction and its application to high-capacity absolute quantitative real-time PCR. BMC Biotechnology. 2008;8:47.

Savard L, Li P, Strauss SH, Chase MW, Michaud M, Bousquet J. Chloroplast and nuclear gene sequences indicate late Pennsylvanian time for the last common ancestor of extant seed plants. Proceedings of the National Academy of Sciences of the U.S.A. 1994;91:5163–7.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution. 2013;30:2725–9.

The uniprot Consortium. UniProt: a hub for protein information. Nucleic Acids Research 2015;43:D204–12.

Tunnacliffe A, Wise MJ. The continuing conundrum of the LEA proteins. Naturwissenschaften. 2007;94:791–812.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). Science. 2006;313:1596–1604.

Verta J-P, Landry CR, MacKay J. Dissection of expression-quantitative trait locus and allele specificity using a haploid/diploid plant system – insights into compensatory evolution of transcriptional regulation within populations. New Phytologist. 2016;211:159–71.

Wang X-S, Zhu H-B, Jin G-L, Liu H-L, Wu W-R, Zhu J. Genome-scale identification and analysis of LEA genes in rice (*Oryza sativa L.).* Plant Science. 2007;172:414–20.

Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution. 2001;18:691–9.

Wickham H. ggplot2. New York, NY: Springer New York; 2009.

Yakovlev IA, Asante DKA, Fossdal CG, Partanen J, Junttila O, Johnsen O. Dehydrins expression related to timing of bud burst in Norway spruce. Planta. 2008;228:459–72.

Yang Y, He M, Zhu Z, Li S, Xu Y, Zhang C, et al. Identification of the dehydrin gene family from grapevine species and analysis of their responsiveness to various forms of abiotic and biotic stress. BMC Plant Biology. 2012;12:140.

Zolotarov Y, Strömvik M. *De novo* regulatory motif discovery identifies significant motifs in promoters of five classes of plant dehydrin genes. PLos ONE. 2015;10:e0129016.

3.10 Supplementary information

3.10.1 Supplementary figures

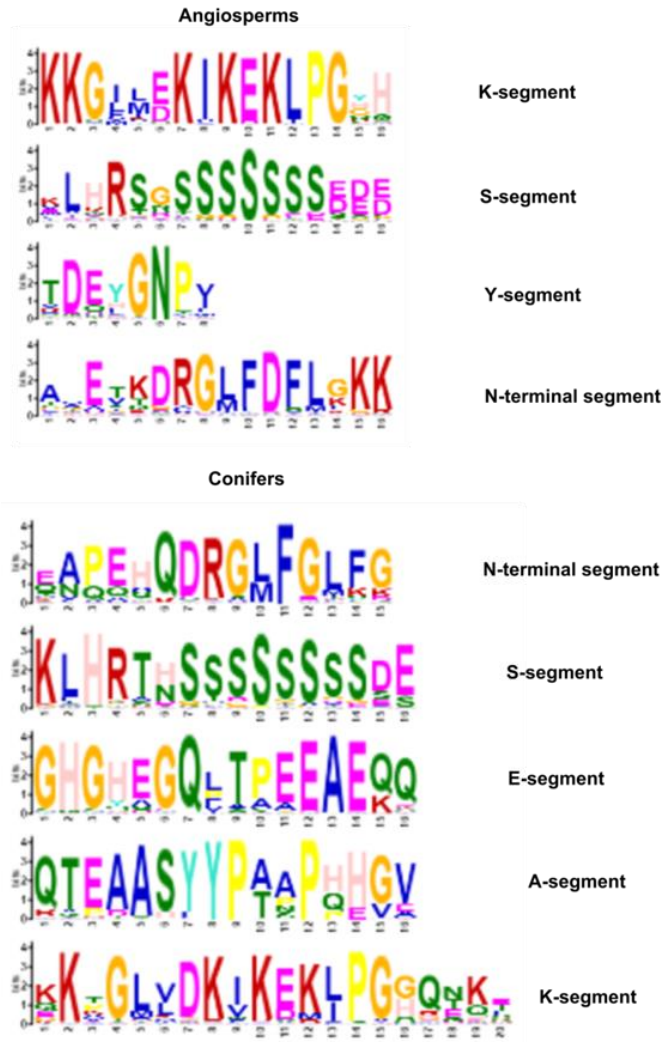


Figure S3.1- Logo of motifs discovered in angiosperms and conifers by MEME.

Figure S3.2- Phylogeny of the angiosperm dehydrin gene family represented by a consensus tree from Bayesian analysis, threshold support equal or superior to 0.75. We used 57 angiosperm dehydrins; see details of sequence clusters in Table S3.2. A dehydrin from *Physcomitrella* was used as the root. The phylogeny was created with MrBayes after protein alignment with MAFFT.

3.10.1 Supplementary tables

Table S3.1- Gene specific primers utilized to determine RNA transcript levels from drought stress and tissue comparison experiments by using quantitative RT-PCR.

Gene	5' primer	3' primer
PgDhn10	ACTCCTGTGTGTTTCACTTGTCGTTC	AGGAAACTCCAAATCCGAACTAATCTG
PgDhn36	GTTGGATGATGTGATGATGGAGAATC	CTGCCTTCTCTCGTCCTTCTGC
PgDhn17	ATAGCATCCCTTCGGACTGGTG	TATCCATCAGCCCCGTTTTTTG
PgDhn33	TGCAATTTTACGAAGTTGTTTGTCTCAC	CTGCCTTCCCTCGTCCTTCTT
PgDhn9	ATGGCGAAGAGAAGAAGAAAGAAAATG	CTTCCTCACTCCTTTCTCTTCTTCAGC
PgDhn37	GCTGATGGTGATTCTATCGGTCTTG	CGAATGCCCAACCTAACTGCTC
PgDhn12	AGGAAAGAAGAAAGAAGGGGGGAG	CATCATTCCCTCCTTCTTCACACC
PgDhn7	GAAGGAAGAAGAGGAAGAGGAAGTGG	ATTTTCTTCCACTTCCCCTTCCTTC
PgDhn35	AGTATTTGGCTGGGAGGAGATTTGTAC	TTTTTCCACATTAATGCATGCCC
PgDhn16	AAGGAAAAGGAAGAAAATGGAAGGC	GCATCTCTTTGATTTTATCCACCAGC
PgDhn19	GCGGGGATGGTAGATAAAGTCAAAG	GGAAGCGTGTCTTTGAATTTATCCAC
PgDhn38	TGGACAAGAAGGACGAGGGAAG	GCAGCTTCTCTTTCACTTCATCCAC
PgDhn23	CCACACCCAAGGTCACACTCAG	TTCTATCCATCAGCCCCGTTTG
PgDhn2	CTGTGTTTAACTCTTCCTGGGAATCTG	TGGTGCCCTGAACCCTATCAAG
PgDhn8	AGGAAGTGAAGAAGGAGAATAGGCATG	TGAGCCTGATTGTGGGTTTGTG
PgDhn32	TGCAAATCAGGAGTGTCAGGACC	TCTCTTTGATTTTATCCGCCATTCC
PgDhn15	AATAAAAAAACGGGGCTGATGGATAG	GTCTTTGACTTTATCCACAACCCCAG

Table S3.2- A total of 144 conifer dehydrins were clustered on the basis of at least 97% of sequence similarity, 78 clusters were formed using CD-hit. The sequences indicated by asterisks were used as the representative sequence of the cluster.

Cluster	Sequence information	
>Cluster 0		
0	347aa, >Psitka|ADE77238.1|... *	
1	346aa, >Pglauca|GQ03719_O20.F77.2_1... at 98.84%	PgDhn9
>Cluster 1		
0	329aa, >Pglauca|GQ03208_F17.2_1... *	PgDhn23
>Cluster 2		
0	326aa, >Ppinaster|K7ZV46|... *	
>Cluster 3		
0	295aa, >Pglauca|GQ03806_L21.4_1... *	PgDhn37
>Cluster 4		
0	284aa, >Psitka|ADE76451.1|... *	
>Cluster 5		
0	284aa, >Pglauca|GQ03812_G05.F77.2_1... *	PgDhn7
>Cluster 6		
0	280aa, >Psitka|ACN41062.1|... *	
>Cluster 7		
0	272aa, >Psitka|ABK26840.1|... *	
>Cluster 8		
0	270aa, >Psitka|ABK23338.1|... *	
1	270aa, >Pglauca|GQ02818_G19.F77.1_1... at 97.04%	PgDhn36
>Cluster 9		
0	270aa, >Psitka|ABK24891.1|... *	
>Cluster 10		
0	268aa, >Pglauca|GQ03507_E05.F77.2_1... *	PgDhn22
>Cluster 11		
0	245aa, >Psitka|ADE76524.1|... *	
1	238aa, >Pglauca|GQ03813_H24.F77.2_1... at 97.06%	PgDhn11
>Cluster 12		
0	245aa, >Pglauca|AAD28175.1|AF109916_1... *	PgDhn1
1	245aa, >Pglauca|GQ03612_M06.F77.2_1... at 98.78%	PgDhn2
>Cluster 13		
0	240aa, >Pglauca|GQ03126_L06.3_2... at 97.08%	PgDhn6
1	245aa, >Pglauca|GQ0044_A18.F77.2_1... *	PgDhn5

>Cluster 14
0 244aa, >Psitka|ABK26857.1|... *
>Cluster 15
0 244aa, >Pglauca|GQ03614_C12.2_2... * PgDhn21
>Cluster 16
0 176aa, >Ptaeda|AAW59241.1|... at 98.30%
1 176aa, >Ptaeda|AAW59253.1|... at 97.73%
2 220aa, >Ptaeda|AHY28768.1|... at 98.18%
3 211aa, >Pcontorta|ACL27840.1|... at 97.16%
4 211aa, >Pbanksiana|ACL27841.1|... at 97.16%
5 132aa, >Pdensata|ABB54920.1|... at 98.48%
6 192aa, >Pdensata|AFP89980.1|... at 97.92%
7 192aa, >Pdensata|AFP89983.1|... at 97.40%
8 192aa, >Pdensata|AFP89986.1|... at 98.44%
9 192aa, >Pdensata|AFP89991.1|... at 97.92%
10 192aa, >Pdensata|AFP89992.1|... at 97.92%
11 192aa, >Pdensata|AFP89993.1|... at 97.92%
12 192aa, >Pdensata|AFP89995.1|... at 97.92%
13 192aa, >Pdensata|AFP90012.1|... at 99.48%
14 192aa, >Pdensata|AFP90017.1|... at 98.96%
15 192aa, >Pdensata|AFP90019.1|... at 98.96%
16 235aa, >Pechinata|AHY28767.1|... at 97.87%
17 120aa, >Phalepensis|ACO57100.1|... at 97.50%
18 207aa, >Phwangshanensis|AIF75721.1|... at 99.03%
19 207aa, >Phwangshanensis|AIF75722.1|... at 97.58%
20 207aa, >Phwangshanensis|AIF75725.1|... at 99.03%
21 238aa, >Pmassoniana|AHJ86266.1|... *
22 207aa, >Pmassoniana|AIF75716.1|... at 100.00%
23 207aa, >Pmassoniana|AIF75717.1|... at 99.52%
24 211aa, >Pmugo|ADV18578.1|... at 98.10%
25 211aa, >Pmugo|ADV18580.1|... at 97.16%
26 208aa, >Pmugo|ADV18582.1|... at 98.08%
27 211aa, >Pmugo|ADV18583.1|... at 97.63%
28 159aa, >Pmugo|AFN44792.1|... at 97.48%
29 211aa, >Pmugoxrotundata|ADV18588.1|... at 98.58%
30 203aa, >Ppinaster|CAM58808.1|... at 97.04%
31 211aa, >Pponderosa|ACL27842.1|... at 97.16%
32 192aa, >Ptabuliformis|AFP89982.1|... at 97.92%
33 192aa, >Ptabuliformis|AFP90004.1|... at 98.44%

34 192aa, >Ptabuliformis|AFP90006.1|... at 97.92%
35 192aa, >Ptabuliformis|AFP90009.1|... at 97.92%
36 192aa, >Ptabuliformis|AFP90011.1|... at 98.96%
37 192aa, >Ptabuliformis|AFP90014.1|... at 98.44%
38 192aa, >Ptabuliformis|AFP90016.1|... at 98.44%
39 192aa, >Ptabuliformis|AFP90018.1|... at 98.44%
40 192aa, >Ptabuliformis|AFP90023.1|... at 99.48%
41 238aa, >Ptabuliformis|AJA33586.1|... at 99.16%
42 192aa, >Pyunnanensis|AFP89981.1|... at 98.44%
43 192aa, >Pyunnanensis|AFP89989.1|... at 97.92%
44 192aa, >Pyunnanensis|AFP90015.1|... at 97.92%
45 192aa, >Pyunnanensis|AFP90021.1|... at 97.40%
46 192aa, >Pyunnanensis|AFP90024.1|... at 98.44%
47 192aa, >Pyunnanensis|AFP90025.1|... at 97.92%
48 208aa, >Psylvestris|ACJ37785.1|... at 98.08%
49 208aa, >Psylvestris|ACJ37802.1|... at 98.56%
50 208aa, >Psylvestris|ACJ37819.1|... at 98.08%
51 202aa, >Psylvestris|ADV31707.1|... at 98.02%
52 202aa, >Psylvestris|ADV31708.1|... at 97.52%
>Cluster 17
0 238aa, >Ppinaster|CCG34067.1|... *
1 238aa, >Ppinaster|K7ZW68|... at 100.00%
>Cluster 18
0 140aa, >Pabies|AAX92688.1|... at 97.14%
1 171aa, >Pabies|AAX92689.1|... at 98.25%
2 234aa, >Pabies|ABU89747.1|... *
3 171aa, >Pobovata|AFM97415.1|... at 98.83%
4 170aa, >Psitka|ABK21213.1|... at 97.06%
>Cluster 19
0 232aa, >Psitka|ADE77666.1|... *
>Cluster 20
0 218aa, >Pabies|ABU89750.1|... *
>Cluster 21
0 212aa, >Pmugo|ADV18574.1|... *
1 212aa, >Presinosa|ACL27839.1|... at 97.17%
2 212aa, >Psylvestris|ACJ37796.1|... at 99.53%
>Cluster 22
0 211aa, >Psylvestris|ACJ37795.1|... *

>Cluster 23
0 210aa, >Pnigra|ACL27838.1|... *
>Cluster 24
0 210aa, >Psylvestris|ACJ37783.1|... *
1 210aa, >Psylvestris|ACJ37812.1|... at 98.10%
2 210aa, >Psylvestris|ACJ37816.1|... at 99.52%
3 204aa, >Psylvestris|ADV31725.1|... at 100.00%
4 204aa, >Psylvestris|ADV31766.1|... at 98.53%
5 204aa, >Psylvestris|ADV31785.1|... at 98.04%
>Cluster 25
0 209aa, >Pglauca|GQ04103_G22.F77.2_1... * PgDhn16
>Cluster 26
0 206aa, >Pglauca|GQ03904_P15.F77.2_1... * PgDhn8
>Cluster 27
0 205aa, >Pabies|ABS58631.1|... *
>Cluster 28
0 202aa, >Psylvestris|ADV31783.1|... *
>Cluster 29
0 193aa, >Ppinaster|K7ZUX7|... *
>Cluster 30
0 193aa, >Ppinaster|K7ZVB9|... *
1 193aa, >Ppinaster|K7ZV88|... at 97.41%
>Cluster 31
0 192aa, >Pdensata|AFP89971.1|... *
1 192aa, >Pdensata|AFP89972.1|... at 98.96%
2 192aa, >Pdensata|AFP89973.1|... at 98.44%
3 192aa, >Pdensata|AFP89974.1|... at 98.96%
4 192aa, >Pdensata|AFP89975.1|... at 98.44%
5 192aa, >Ptabuliformis|AFP89969.1|... at 99.48%
6 192aa, >Ptabuliformis|AFP89970.1|... at 98.96%
>Cluster 32
0 192aa, >Pdensata|AFP89984.1|... *
1 192aa, >Ptabuliformis|AFP89985.1|... at 98.96%
>Cluster 33
0 188aa, >Pglauca|GQ03326_D07.1_1... * PgDhn17
>Cluster 34
0 188aa, >Pglauca|GQ04112_D12.1_1... * PgDhn18
>Cluster 35

0 187aa, >Pglauca|GQ03515_G02.F77.2_1... * PgDhn39
>Cluster 36
0 183aa, >Pabies|ABS58630.1|... *
>Cluster 37
0 183aa, >Pglauca|GQ02010_J18.1_1... * PgDhn19
>Cluster 38
0 178aa, >Pglauca|GQ02828_E08.F77.2_1... * PgDhn20
>Cluster 39
0 177aa, >Pobovata|AFM97413.1|... *
>Cluster 40
0 177aa, >Ppinaster|CCG34063.1|... *
1 177aa, >Ppinaster|K7ZW95|... at 100.00%
2 177aa, >PPinea|AIN43955.1|... at 97.74%
>Cluster 41
0 144aa, >Phwangshanensis|AIF75741.1|... at 97.92%
1 144aa, >Phwangshanensis|AIF75742.1|... at 99.31%
2 142aa, >Phwangshanensis|AIF75744.1|... at 100.00%
3 144aa, >Pmassoniana|AIF75736.1|... at 98.61%
4 144aa, >Pmassoniana|AIF75738.1|... at 99.31%
5 177aa, >Ptabuliformis|AJA33584.1|... *
6 144aa, >Psylvestris|ACA51876.1|... at 100.00%
7 144aa, >Psylvestris|ACA51877.1|... at 99.31%
8 127aa, >Psylvestris|ADA85539.1|... at 99.21%
>Cluster 42
0 173aa, >Pglauca|GQ03616_G15.F77.2_1... * PgDhn4
>Cluster 43
0 170aa, >Psitka|ABK21332.1|... *
1 170aa, >Pglauca|GQ03808_I16.F77.2_2... at 100.00% PgDhn12
>Cluster 44
0 169aa, >Pabies|ABS58627.1|... *
1 167aa, >Pabies|ABS58628.1|... at 97.01%
>Cluster 45
0 169aa, >Pabies|ABU89748.1|... *
>Cluster 46
0 164aa, >Psitka|ABK22729.1|... at 98.78%
1 166aa, >Psitka|ABK23574.1|... *
2 166aa, >Psitka|ABK24884.1|... at 98.19%
3 165aa, >Pglauca|GQ0067_P11.F77.2_1... at 98.18% PgDhn10
>Cluster 47

0 159aa, >Pmugo|AFN44791.1|... *
>Cluster 48
0 158aa, >Pabies|ABS58629.1|... *
1 154aa, >Pglauca|WS02628_O05.F77.2_2... at 98.05% PgDhn13
>Cluster 49
0 150aa, >Pglauca|GQ03607_L02.F77.2_1... * PgDhn26
>Cluster 50
0 149aa, >Pglauca|GQ03602_G21.F77.2_1... * PgDhn25
>Cluster 51
0 148aa, >Pglauca|GQ03901_J22.F77.2_2... * PgDhn3
>Cluster 52
0 146aa, >Psitka|ADE76163.1|... *
>Cluster 53
0 143aa, >Ptaeda|AAW59164.1|... at 100.00%
1 144aa, >Ptaeda|AAW59168.1|... at 97.92%
2 143aa, >Ptaeda|AAW59174.1|... at 99.30%
3 143aa, >Ptaeda|AAW59176.1|... at 98.60%
4 143aa, >Ptaeda|AAW59180.1|... at 99.30%
5 139aa, >Ptaeda|AHY28775.1|... at 99.28%
6 146aa, >Pelliottii|ACA51879.1|... *
7 139aa, >Pelliottii|AHY28771.1|... at 100.00%
>Cluster 54
0 141aa, >Ppalustris|AHY28773.1|... at 100.00%
1 145aa, >Pelliottii|AHY28772.1|... *
>Cluster 55
0 143aa, >Ppinaster|AIN43962.1|... *
>Cluster 56
0 142aa, >Ppinaster|AIN43961.1|... *
>Cluster 57
0 136aa, >Pglauca|GQ03201_C14.1_1... * PgDhn14
>Cluster 58
0 135aa, >Pglauca|GQ03511_K03.F77.2_1... * PgDhn28
>Cluster 59
0 135aa, >Pglauca|GQ03610_H22.F77.2_1... * PgDhn29
>Cluster 60
0 135aa, >Pglauca|GQ03912_I07.F77.2_1... * PgDhn30
>Cluster 61
0 132aa, >Pglauca|GQ03913_M17.1_1... * PgDhn15
>Cluster 62

0 131aa, >Pglauca|PGTGY006705.1_1... * PgDhn40
>Cluster 63
0 125aa, >Ppinaster|AIN43960.1|... *
>Cluster 64
0 124aa, >Pglauca|PGTGY006706.1_1... * PgDhn41
>Cluster 65
0 120aa, >Psitka|ADE76347.1|... *
>Cluster 66
0 109aa, >Psitka|ADE76271.1|... *
>Cluster 67
0 109aa, >Pglauca|GQ03601_E22.2_1... * PgDhn24
>Cluster 68
0 109aa, >Pglauca|GQ03913_K08.F77.2_1... * PgDhn27
>Cluster 69
0 102aa, >Ppinaster|K7ZWF9|... *
>Cluster 70
0 102aa, >Ppinaster|K7ZV53|... *
>Cluster 71
0 100aa, >Pglauca|GQ03612_L10.F77.2_3... * PgDhn34
1 80aa, >Pglauca|GQ03903_F04.F77.2_1... at 100.00% PgDhn35
2 80aa, >Pglauca|GQ03914_P06.F77.2_1... at 100.00% PgDhn35
3 80aa, >Pglauca|GQ03918_F10.F77.2_1... at 100.00% PgDhn35
>Cluster 72
0 89aa, >Pglauca|GQ03122_M21.1_1... * PgDhn38
>Cluster 73
0 88aa, >Psitka|ABK21645.1|... *
1 88aa, >Psitka|ABK25374.1|... at 97.73%
2 88aa, >Psitka|ADE76499.1|... at 98.86%
>Cluster 74
0 87aa, >Pabies|ABU89751.1|... *
>Cluster 75
0 87aa, >Psitka|ABK21393.1|... *
1 87aa, >Pglauca|GQ03614_D04.F77.2_1... at 98.85% PgDhn31
>Cluster 76
0 84aa, >Pabies|AAX92687.1|... *
1 84aa, >Psitka|ADE76208.1|... at 97.62%
>Cluster 77
0 84aa, >Pglauca|GQ03603_F10.F77.2_1... * PgDhn33
1 84aa, >Pglauca|GQ03612_C16.F77.2_1... at 97.62% PgDhn32

2	84aa, >Pglauca|GQ03719_H02.F77.2_1... at 100.00%	PgDhn33
3	84aa, >Pglauca|GQ03911_M06.F77.2_1... at 97.62%	PgDhn32

Table S3.3- A total of 76 angiosperm dehydrins were clustered by sequence similarity (97%), 57 clusters were formed using CD-hit. Sequences indicated by an asterisk were used as representative sequence of the cluster.

Cluster	Sequence information
>Cluster 1	
0	326aa, >tr|B4FKQ0|B4FKQ0_MAIZE... *
1	326aa, >tr|B6UGH3|B6UGH3_MAIZE... at 98.47%
>Cluster 2	
0	326aa, >tr|Q53JR9|Q53JR9_ORYSJ... *
>Cluster 3	
0	325aa, >tr|K7V309|K7V309_MAIZE... *
>Cluster 4	
0	290aa, >tr|B4G1H1|B4G1H1_MAIZE... at 100.00%
1	291aa, >tr|B6TN77|B6TN77_MAIZE... *
2	291aa, >tr|B6SS21|B6SS21_MAIZE... at 99.66%
>Cluster 5	
0	137aa, >tr|O48672|O48672_ORYSA... at 97.81%
1	290aa, >tr|Q6ESR4|Q6ESR4_ORYSJ... *
>Cluster 6	
0	289aa, >tr|C4J477|C4J477_MAIZE... *
1	289aa, >tr|Q41824|Q41824_MAIZE... at 99.65%
2	281aa, >tr|B7U627|B7U627_MAIZE... at 97.86%
>Cluster 7	
0	284aa, >tr|J9PZL0|J9PZL0_MALDO... *
>Cluster 8	
0	277aa, >tr|J9PZB2|J9PZB2_MALDO... *
>Cluster 9	
0	272aa, >tr|Q4JNX4|Q4JNX4_PRUDU... *
>Cluster 10	
0	268aa, >tr|Q40955|Q40955_PRUPE... *
>Cluster 11	
0	268aa, >tr|M5VQQ9|M5VQQ9_PRUPE... *
>Cluster 12	
0	265aa, >sp|P31168|COR47_ARATH... *
>Cluster 13	
0	260aa, >sp|P42759|ERD10_ARATH... *
1	259aa, >tr|F4HST2|F4HST2_ARATH... at 98.84%
>Cluster 14	

0 258aa, >tr|G8FVC3|G8FVC3_EUCGL... *
>Cluster 15
0 256aa, >tr|J9Q177|J9Q177_MALDO... *
>Cluster 16
0 249aa, >tr|Q30E95|Q30E95_PRUPE... at 100.00%
1 254aa, >tr|M5XNV8|M5XNV8_PRUPE... *
>Cluster 17
0 236aa, >tr|B6UGS5|B6UGS5_MAIZE... *
>Cluster 18
0 232aa, >tr|J9PZB0|J9PZB0_MALDO... *
>Cluster 19
0 229aa, >tr|A1XSX2|A1XSX2_MALDO... at 98.69%
1 230aa, >tr|J9PZX2|J9PZX2_MALDO... *
>Cluster 20
0 228aa, >tr|M5VP39|M5VP39_PRUPE... *
>Cluster 21
0 228aa, >sp|P30287|DHR25_ORYSJ... *
>Cluster 22
0 225aa, >tr|A9PA80|A9PA80_POPTR... *
>Cluster 23
0 225aa, >tr|U3MKZ8|U3MKZ8_PRUAV... *
>Cluster 24
0 216aa, >tr|A0A0N6YJL4|A0A0N6YJL4_MALDO... *
>Cluster 25
0 212aa, >tr|W1NLK4|W1NLK4_AMBTC... *
>Cluster 26
0 210aa, >tr|A0A059BPM9|A0A059BPM9_EUCGR... *
>Cluster 27
0 207aa, >tr|A5BUW9|A5BUW9_VITVI... *
1 206aa, >tr|F6H0C4|F6H0C4_VITVI... at 99.51%
>Cluster 28
0 202aa, >tr|Q5QIC0|Q5QIC0_PRUPE... *
1 202aa, >tr|Q9SW89|Q9SW89_PRUDU... at 97.03%
>Cluster 29
0 200aa, >tr|A0A0N6W1E0|A0A0N6W1E0_MALDO... *
>Cluster 30
0 193aa, >sp|P42758|XERO2_ARATH... *
1 159aa, >tr|Q8H7A5|Q8H7A5_ARATH... at 100.00%
>Cluster 31

0 191aa, >tr|F6I0M9|F6I0M9_VITVI... *
>Cluster 32
0 190aa, >tr|J9Q0P7|J9Q0P7_MALDO... *
>Cluster 33
0 188aa, >tr|J9Q0P4|J9Q0P4_MALDO... *
>Cluster 34
0 188aa, >tr|A0A0P0V781|A0A0P0V781_ORYSJ... *
>Cluster 35
0 186aa, >sp|P30185|DHR18_ARATH... *
>Cluster 36
0 185aa, >sp|Q96261|DHLEA_ARATH... *
>Cluster 37
0 185aa, >sp|P42763|ERD14_ARATH... *
>Cluster 38
0 183aa, >tr|B9HRJ7|B9HRJ7_POPTR... *
>Cluster 39
0 182aa, >tr|M5VLL5|M5VLL5_PRUPE... *
>Cluster 40
0 177aa, >tr|C0Z2D8|C0Z2D8_ARATH... *
>Cluster 41
0 177aa, >tr|J9PZK8|J9PZK8_MALDO... *
>Cluster 42
0 172aa, >sp|A2ZDX9|DHR21_ORYSI... *
>Cluster 43
0 168aa, >sp|P12950|DHN1_MAIZE... *
1 168aa, >tr|A3KLI1|A3KLI1_MAIZE... at 99.40%
2 168aa, >tr|A3KLI0|A3KLI0_MAIZE... at 98.21%
3 168aa, >tr|A7RDP0|A7RDP0_MAIZE... at 97.62%
>Cluster 44
0 166aa, >tr|A5C8N2|A5C8N2_VITVI... *
>Cluster 45
0 165aa, >tr|Q40651|Q40651_ORYSA... *
>Cluster 46
0 164aa, >sp|A2ZDX6|DH16C_ORYSI... *
>Cluster 47
0 163aa, >tr|Q8GY17|Q8GY17_ARATH... *
>Cluster 48
0 160aa, >tr|Q6ESR3|Q6ESR3_ORYSJ... *

>Cluster 49
0 151aa, >sp|A2ZDX4|DH16D_ORYSI... *
>Cluster 50
0 148aa, >tr|W1NPP8|W1NPP8_AMBTC... *
>Cluster 51
0 147aa, >tr|H2EUU8|H2EUU8_EUCGL... *
>Cluster 52
0 137aa, >tr|A0A059AI61|A0A059AI61_EUCGR... *
>Cluster 53
0 130aa, >tr|A3REN2|A3REN2_VITVI... *
1 130aa, >tr|Q3ZNL4|Q3ZNL4_VITVI... at 99.23%
2 130aa, >tr|A5C8L5|A5C8L5_VITVI... at 97.69%
3 130aa, >tr|H9A0H3|H9A0H3_VITVI... at 98.46%
>Cluster 54
0 128aa, >sp|P25863|XERO1_ARATH... *
>Cluster 55
0 124aa, >tr|Q4VT48|Q4VT48_VITVI... *
1 124aa, >tr|A3REN1|A3REN1_VITVI... at 99.19%
>Cluster 56
0 91aa, >tr|A0A059BPV1|A0A059BPV1_EUCGR... *
>Cluster 57
0 83aa, >tr|A0A0M3R6L9|A0A0M3R6L9_ORYSI... *

Table S3.4- Classification of angiosperm and conifer dehydrins based on their conserved amino-acid segments (segment-K, A, E, S, Y and N1). The graphical representation of all possible classifications (models) is in Fig.3. Sequences indicated by //, *, ** presented one degenerate A, K or Y segment, respectively.

Dehydrin sequence	Classification	Model
Ppinaster-K7ZWF9 Ppinaster-K7ZV53	N1 K2	1
Pglauca-PgDhn3 Pglauca-PgDhn13 Pabies-ABS58629.1	N1 ESK2	2
Pabies-ABU89748.1 Pelliottii-ACA51879.1 Pelliottii-AHY28772.1	N1 AESK	3
Ptabuliformis-AJA33584.1 Ppinaster-CCG34063.1 Pglauca-PgDhn10 Psitka-ABK23574.1 Pglauca-PgDhn4 Pabies-ABS58627.1 Pglauca-PgDhn12 * Psitka-ABK21332.1 *	N1 AESK2	4
Pdensata-AFP89984.1 Psylvestris-ACJ37795.1 Pdensata-AFP89971.1 Pmugo-ADV18574.1 Psylvestris-ADV31783.1 Pnigra-ACL27838.1 Pmugo-AFN44791.1 Psylvestris-ACJ37783.1 Pabies-ABU89747.1 * Ppinaster-K7ZUX7 Ppinaster-K7ZVB9	N1 AESK3	5

Ppinaster-CCG34067.1 Psitka-ABK26857.1 Pglauca-AAD28175.1-PgDhn1 Pglauca-PgDhn2 Pglauca-PgDhn5 Pglauca-PgDhn6 Pglauca-PgDhn11 Psitka-ADE76524.1 Psitka-ADE77666.1 * Pmassoniana-AHJ86266.1	N1 AESK4	6
Psitka-ABK26840.1 Pglauca-PgDhn7	N1 A2E2SK4	7
Ppinaster-K7ZV46	N1 A2E2SK5	8
Pglauca-PgDhn8 *	N1 A4SK	9
Ppinaster-AIN43960.1 //	N1 AS	10
Pglauca-PgDhn9 * Psitka-ADE77238.1 *	N1 ASK3	11
Ppinaster-AIN43961.1 * Ppinaster-AIN43962.1 *	N1 SK	12
C0Z2D8-ARATH * P42763-ERD14_ARATH G8FVC3-EUCGL A0A059BPM9_EUCGR A5BUW9_VITVI A9PA80_POPTR Q8GY17_ARATH	N1 SK2	13
C4J477-MAIZE B6TN77-MAIZE Q6ESR4-ORYSJ P31168-COR47_ARATH P42759-ERD10_ARATH	N1 SK3	14

J9PZL0-MALDO J9PZB2-MALDO M5VLL5_PRUPE		
M5XNV8-PRUPE	N1 SK4	15
W1NLK4-AMBTC A5C8N2_VITVI	SK3	16
Pabies-ABU89751.1 Pglauca-PgDhn31 Psitka-ABK21393.1	K3	17
Pglauca-PgDhn14 * Pglauca-PgDhn15 * Pglauca-PgDhn16	K4	18
Pobovata-AFM97413.1 * Pglauca-PgDhn17 * Pglauca-PgDhn18 * Pglauca-PgDhn19 * Pabies-ABS58630.1 * Pabies-ABS58631.1 * Pglauca-PgDhn20 *	K6	19
Pglauca-PgDhn21 *	K8	20
Psitka-ADE76451.1 * Pglauca-PgDhn22 *	K10	21
Pglauca-PgDhn23 *	K11	22
Pglauca-PgDhn24 Pglauca-PgDhn25 Psitka-ADE76347.1 Pglauca-PgDhn26 Pglauca-PgDhn27	N1 K3	23
Pglauca-PgDhn28	N1 K4	24

Pglauca-PgDhn29 Pglauca-PgDhn30 Psitka-ADE76271.1 Psitka-ADE76163.1		
Pabies-AAX92687.1 Pglauca-PgDhn32 Pglauca-PgDhn33 Pglauca-PgDhn34 Pglauca-PgDhn35 Psitka-ABK21645.1	N1 K2	25
Psitka-ABK23338.1 Pglauca-PgDhn36 Psitka-ABK24891.1 Psitka-ACN41062.1 Pglauca-PgDhn37 Pglauca-PgDhn38	K1	26
Pglauca-PgDhn39 * Pglauca-PgDhn40 Pglauca-PgDhn41	K2	27
Q40651_ORYSA	YSK	28
A2ZDX4-DH16D_ORYSI A2ZDX6-DH16C_ORYSI P12950-DHN1_MAIZE A2ZDX9-DHR21_ORYSI W1NPP8_AMBTC **	YSK2	29
K7V309_MAIZE B4FKQ0_MAIZE	YSK3	30
P30287-DHR25_ORYSJ B6UGS5_MAIZE A0A059AI61_EUCGR	Y2SK2	31
H2EUU8_EUCGL	Y3SK2	32

M5VP39_PRUPE A0A0N6W1E0_MALDO F6I0M9_VITVI B9HRJ7_POPTR Q96261-DHLEA_ARATH		
J9PZK8_MALDO	K4	18
P42758-XERO2_ARATH	K6	18
J9PZX2_MALDO	YSK3	30
J9PZB0_MALDO Q5QIC0_PRUPE U3MKZ8_PRUAV J9Q0P4_MALDO	Y2SK3	33
J9Q0P7_MALDO	YK3	34
J9Q177_MALDO	YK4	35
M5VQQ9_PRUPE	Y3K3	36
A0A0N6YJL4_MALDO	Y3SK3	37
P30185-DHR18_ARATH	Y2SK2	31
P25863-XERO1_ARATH	YSK2	29
Q4VT48_VITVI A3REN2_VITVI	YSK2	29

Table S3.5- The one-way ANOVA tested if the expression levels between the three tissues (phelloderm, xylem and young foliage) were different.

	Source (Tissues)	
Genes	**F value**	**Pr(>F)**
PgDh10	9.21	0.01
PgDh19	36.38	4.42E-04
PgDh38	1980.00	3.46E-09
PgDh23	11.98	0.01
PgDh2	2.75	0.14
PgDh9	10.78	0.02
PgDh37	0.61	0.58
PgDh12	4.51	0.06
PgDh7	6.77	0.03
PgDh8	6.18	0.03
PgDh32	118.80	6.10E-05
PgDh15	14.51	0.01
PgDh36	2.92	0.13

Table S3.6- A three-way ANOVA with water potential as a function of type of treatment (watering regimes), genotype, sampling dates and their interaction.

Source	**F value**	**Pr(>F)**
genotype	1.665	0.195
date	135.094	<2e-16
trait	634.769	<2e-16
genotype:date	0.291	0.967
genotype:trait	0.56	0.573
date:trait	141.09	<2e-16
genotype:date:trait	0.414	0.91

Table S3.7- A three-way ANOVA with expression as a function of type of treatment (watering regimes), genotype, sampling dates and their interaction.

Genes	Source	F value	Pr(>F)
PgDhn10	genotype	48.533	6.11E-16
	date	34.677	< 2e-16
	trait	299.92	< 2e-16
	genotype:date	2.409	0.02
	genotype:trait	4.931	0.01
	date:trait	61.287	< 2e-16
	genotype:date:trait	3.209	0.00
PgDhn36	genotype	1.044	0.36
	date	8.092	8.96E-06
	trait	1.567	0.21
	genotype:date	2.526	0.01
	genotype:trait	0.925	0.40
	date:trait	7.545	2.01E-05
	genotype:date:trait	1.03	0.42
PgDhn17	genotype	0.48	0.62
	date	3.794	6.42E-03
	trait	2.644	0.11
	genotype:date	0.654	0.73
	genotype:trait	1.944	0.15
	date:trait	1.535	0.20
	genotype:date:trait	1.151	0.34
PgDhn33	genotype	0.588	0.56
	date	87.04	< 2e-16
	trait	452.94	< 2e-16
	genotype:date	1.887	0.07
	genotype:trait	1.097	0.34
	date:trait	52.215	< 2e-16
	genotype:date:trait	1.628	0.13
PgDhn9	genotype	33.149	6.26E-12
	date	25.384	8.47E-15
	trait	9.418	2.72E-03
	genotype:date	1.434	0.19

	genotype:trait	1.38	0.26
	date:trait	18.704	1.11E-11
	genotype:date:trait	1.906	0.07
	genotype	16.103	1.15E-06
	date	7.507	3.10E-05
	trait	1.295	0.26
PgDhn37	genotype:date	1.262	0.27
	genotype:trait	0.177	0.84
	date:trait	3.022	2.20E-02
	genotype:date:trait	1.688	0.11
	genotype	1.091	0.34
	date	0.381	0.82
	trait	10.625	1.53E-03
PgDhn12	genotype:date	1.928	0.06
	genotype:trait	3.266	4.23E-02
	date:trait	4.663	1.72E-03
	genotype:date:trait	2.182	0.04
	genotype	10.433	7.80E-05
	date	3.609	8.67E-03
	trait	10.288	1.81E-03
PgDhn7	genotype:date	1.287	0.26
	genotype:trait	3.108	0.05
	date:trait	5.054	9.57E-04
	genotype:date:trait	3.004	4.67E-03
	genotype	13.883	4.47E-06
	date	79.138	< 2e-16
	trait	432.11	< 2e-16
PgDhn35	genotype:date	3.368	1.78E-03
	genotype:trait	9.32	1.88E-04
	date:trait	88.359	< 2e-16
	genotype:date:trait	1.189	0.31
	genotype	3.227	0.04
PgDhn16	date	31.176	<2e-16
	trait	117.62	<2e-16

genotype:date	1.414	0.20
genotype:trait	0.328	0.72
date:trait	44.554	<2e-16
genotype:date:trait	2.037	0.05

Chapter 4: Conclusions

This thesis is devoted to developing a better understanding of the evolution of genes in conifers. The focal species was *P. glauca* and the work included several data and comparisons to other conifers in addition to flowering plants to enable comparative analyses relevant for inferring evolutionary differences. This investigation into gene evolution covered two complementary aspects: the structure of individual genes and the factors that impact on structural differences (Chapter 2) and the organization of a large gene family which has diverged between conifers and flowering plants (Chapter 3). We begin by reviewing the major findings and conclusions from the work (section 4.1); this is followed by a critical overview of the contributions to the field (section 4.2) and perspectives for future developments and application (section 4.3).

4.1 Major findings and conclusions

In this section, we present the major results and conclusions based on the main thesis objectives described in Chapter 1.

4.1.1 Evolution of gene structure

In order to understand some of the forces that could influence the evolution of gene structure in conifers, we evaluated whether genome size, composition and gene expression profile could have an impact on gene structure and intron sizes (Chapter 2).

First we reported a detailed analysis of the gene structure of 35 genes from *Picea glauca* and their closest homologous from *Arabidopsis thaliana*, *Zea mays* and *Populus trichocarpa*. We observed that species with larger genomes have longer introns, but not proportionally to their genome size. For example, the *Picea glauca* genome is 158 times larger than *Arabidopsis* genome; however it presented four times more intron sequence per gene on average than *Arabidopsis*. Often in the reports of conifer genomes the attention has been directed to the elevated average intron lengths. In our detailed analysis of gene structure we observed that conifers have long introns on average because of the presence of

few long introns per gene, however the median intron length remains similar to other plant genomes.

To delineate the level of gene structure conservation between conifers we carry out a detailed pairwise comparison of introns and exons between *Picea glauca* and *Pinus taeda* for 23 genes. We found high exon sequence similarity and conserved number of exons and introns, as expected because conifers present a conserved genome macrostructure and low rates of genome evolution (Pavy et al., 2012). Surprisingly short intron sequences were also conserved between the two species while only a few long introns were conserved; suggesting that short introns may be under stronger selection than longer ones.

When we looked at the gene expression profile of the 35 genes from *P. glauca*, we observed that highly transcribed genes presented more intronic sequence on average than genes with more specialized expression; however there was a large variation of total intronic sequence among genes from each expression group. In our observations there was no clear correlation between intron size and expression profile.

We were also curious to know if the repetitive elements that are responsible for the large genomes in conifers have impacted the evolution of gene structure. We developed a *P. glauca* repeat library and screened the sequenced BAC clones containing genes. We observed that the amount of repetitive elements is variable and lower in the intergenic region when compared to estimations of the whole genome. We also searched for repetitive elements in almost 2 000 genes. We showed that repetitive elements had an impact in the evolution of gene structure, not only contributing to the size of long introns as described by Nystedt et al., 2014, but were also found in smaller introns. We detected in majority small fragments of 114 bp in median in introns, probably because part of the original inserted element have been lost over time.

These observations altogether indicated that the evolution of gene structure seems to be ruled by many factors in different proportions according to the characteristics of each species. Our results suggests that genome size and composition have impacted gene

structure evolution, probably in combination with other factors not studied in the present work such as recombination rate and effective population size.

4.1.2 Gene family evolution: a case study of dehydrins

Evolutionary history and genome properties also shape the evolution of gene families. A few large-scale gene discovery and genome sequencing projects in conifers have carried out broad analyses that give an overview of the numbers and sizes of gene families (Rigault et al. 2011; Nystedt et al. 2013; Wegrzyn et al. 2014). We aimed to develop a more detailed understanding of evolutionary paths by examining a large gene family that diverged between conifers and flowering plants. We decided to follow up on the identification of a large number of dehydrin genes in *P. glauca* (Rigault et al. 2011; Raherison et al. 2012). Our main objective was to characterize this gene family and trace its evolutionary history with an emphasis on conifers and, study their expression responsiveness during dehydration stress.

The phylogenetic analyses suggest that a lineage-specific duplication contributed to the expansion of dehydrins in the genus *Picea*. The diversity of conifer dehydrin sequences were reflected in a wide range of structural types, represented by the modular variations in the amino acid motif composition. The variation was also observed within the K-segments, which were found to be less conserved than in angiosperms. The gene family expansion and the structural diversity suggest that subfonctionalization would be involved in the increase of dehydrin diversity in conifers. We did not observe a direct relationship between amino acid structural classification and expression profile under normal conditions and under dehydration stress. The *Picea glauca* dehydrins showed differential expression profiles across vegetative tissues under normal conditions and in leaves under dehydration stress. The N1 K2 and N1 AESK2 dehydrins were very responsive to water stress, as shown in other conifer studies.

Many conifers are long lived species that experience variable and sometimes extreme environmental conditions during their life span. The elevated number of dehydrins and their high level of diversification at the structural and functional levels may reflect the adaptive

plasticity required for living in environments with highly variable conditions such as the boreal forest. We showed that some of the dehydrins are responsive to dehydration, supplementary studies will be needed to elucidate the complete picture of their functional role during abiotic stresses.

4.2 Critical overview and contributions

Evolution of gene structure

In the study of gene structure evolution in conifers, the data from *Picea glauca* BAC sequencing and gene space obtained from sequence capture were presented for the first time. The 21 BAC clones, each containing a single gene, represented a significant advance for conifers. A few studies (Hamberger et al., 2009; Magbanua et al., 2011; Kovach et al., 2010; Bautista et al., 2007) have isolated and analyzed BACs from pines and spruces; they reported the analysis of 2 to 10 BACs. Many of the BACs they analyzed only contained incomplete gene sequences (many were pseudogenes and others were presumably genes split between two BACs) or no recognizable gene sequence at all. This problem was associated with BAC library screening using probe hybridization and may be explained in part by the abundance of pseudogenes in conifer genomes. We overcame this problem by using a PCR screen and validations during the screening stages. Therefore, the set of 21 BACs may represent a small set relative to reports in other organisms but it was unprecedented for conifers.

At the time we started this project, the conifer genomes had not yet been published. We planned the analyses of gene structure in genes isolated from the *Picea glauca* BAC clones, because with this strategy we were able to assembly long inserts of intergenic region (average of 125 Kbp), which included complete genomic sequence of targeted genes. When the first version of the *Picea glauca* genome sequence became available we expanded our analyses to 35 genes (18 genes from *Picea glauca* genome assemble and 17 from BAC clones). This number is relatively small but, at the time, we could not increase our analysis because *Picea glauca* assembly was still highly fragmented and gene annotation not sufficiently advanced for our analyses. Although our study has focused on 35 genes our

findings were in agreement with the whole-genome reports in regard to the average intron size and median size in both *Picea abies* and *Pinus taeda.* The significant benefit of our strategy is that we were able to deliver a carefully curated analysis of individual genes. Certainly, additional analyses including more genes are needed especially in the section related to relationship of gene expression profile and intron size.

In collaboration with a Swedish research group we developed a library of *Picea glauca* repeat elements. The advantage of using this library was that we were able to identify repeat elements that were absent in Repbase (database of eukaryotic repetitive and transposable elements) (Kapitonov and Jurka, 2008). Combining the library of repeat elements developed for *Picea glauca* and the gene space obtained from sequence capture we were capable to analyze the impact of repeat elements in the gene structure of *Picea glauca* at a large scale. We showed that 32% of the genes sequenced (total of 1836 genes) contained repeat elements in their introns, despite the fact that the introns were for the most part under in 1 Kbp in the sequence capture dataset. This has allowed us to show the ubiquitous distribution of repetitive sequences in *Picea glauca* genes and intergenic regions.

Dehydrin gene family

This is the first study in conifers to report and describe such a high number of dehydrins. Until now, less than ten dehydrins had been reported in other conifers such as *Picea abies*, *Picea obovata*, *Pinus pinaster* and *Pinus pinea* (Yakovlev et al., 2008; Perdiguero et al., 2012; Kjellsen et al., 2013; Perdiguero, Soto and Collada, 2015). Instead of remaining restricted to *Picea glauca* dehydrin gene family characterization, we carried out an exhaustive analysis in which we utilized the available sequence resources in conifers and included the sequences from several flowering plants to show an evolutionary and structural scenario involving conifer and angiosperm dehydrins. This strategy has allowed us to confirm the deeply rooted differences in the modular structure of the amino acid motifs among angiosperms and conifer dehydrins.

The phylogenetic analysis permitted us to infer possible evolutionary paths for the origin of the different dehydrin classes and to explain to some extent the expansion of the dehydrin gene family in *Picea*. However, the topology of the dehydrin phylogenetic tree lacked resolution near the origin and we were unable to provide a complete and clear evolutionary history of the early stages of the dehydrin gene family. More sampling of dehydrins in other conifer genera, other gymnosperms and more primitive plants would help to fill this gap and thus improve our understanding of the evolutionary history of the dehydrin gene family.

Our initial plan was to analyze stress responsiveness of dehydrins in roots in addition to foliage. However, we found that the *Picea glauca* plants were very sensitive to the dehydration stress treatment and after 14 days without watering the roots were dehydrated and yielded low quality RNA extracts. We decided to only analyze the foliage and eliminate the root samples from our expression analyses. It would be interesting to rethink the method of analyzing roots in stressful conditions and sample other tissues in future analyzes such as the phelloderm which accumulated several dehydrin transcripts under normal conditions.

We measured the accumulation of *Picea glauca* dehydrin transcripts by using quantitative RT-PCR. We faced significant challenges in amplifying many of the dehydrins because of high levels of sequence similarity between genes. This was a major limiting factor for the design of specific primers. Despite this fact we successfully amplified 17 dehydrins. It revealed a partial picture of the expression profile of *Picea glauca* dehydrins. This problem could be circumvented by using an RNA-seq approach which we expect would be more successful at revealing the expression profile of a large gene family such as the dehydrins.

4.3 Perspectives

4.3.1 Large scale gene structure evolution and comparative analyses of intergenic regions

The initial idea of the project was to isolate and sequence BAC clones containing targeted genes in *Picea glauca* and other conifer species in order to conduct comparative analyses of both the gene structure and the intergenic region. Other research groups were involved in the BAC isolation studies in other conifers; unfortunately due to technical problems they isolated just a few of the targeted genes, which limited our ability to conduct the analyses.

The sequencing of more BAC clones containing target genes in *Picea glauca* and other conifer species would be interesting to expand the comparative analyses to the intergenic level. As the genome assemblies still fragmented, the sequencing of targeted BAC clones in conifers could provide intergenic sequences of 130-150 Kbp. A comparative analysis of equivalent intergenic regions among conifers could reveal the degree of conservation of these regions as well its composition, besides the fact that these long sequences could also be useful to improve the genome assemblies. When attempted to analyze the few BAC sequences that were available in more *Picea glauca* and another conifer by sequence alignment, we observed that the sequences of intergenic regions were highly variable. It may be that a more fruitful approach would be to set up analyses on a larger scale spanning many genes along a chromosome. For example, this could involve identifying landmarks such as LTR retrotransposons and investigating their distribution on a chromosomal scale or examining distances and sequence between genes.

When a large portion of *P. glauca* genes will be available in contiguous genomic sequences we will be able to do a detailed large scale comparative analysis of gene structure among homologous sequences from other plant species and be able to draw wider conclusions on the impact of large genomes in the gene structure as well the relationships between expression profile and intron sizes. At the moment the gene models and annotations in the conifer genome assemblies have not yet reached this level (Prunier et al. 2015; Warren et al., 2015).

Another aspect that could enrich the *P. glauca* genome characterization would be the classification of *P. glauca* repetitive elements that do no match characterized repetitive sequences i.e. that have no significant hits in Repbase and nr genbank. Our analyses showed that many of the repetitive elements in the intergenic regions and in introns were unclassified. Future investigations could deepen our knowledge of these unique elements present in conifers and show the differences between these sequences and already known elements.

4.3.2 Dehydrins: multi-function proteins

A more complete understanding of the genes and pathways involved in the process of plant adaptation to limiting water conditions would improve our ability to preserve the relevant genetic diversity. The dehydrins *PgDhn10, 16, 33* and *35* were highly responsive to dehydration in *Picea glauca* foliage. The characterization of genotypes with different levels of tolerance to drought could reveal the potential of these genes as molecular markers for tolerance to drought which has been identified as a developing threat to the health of boreal forests (Gauthier et al., 2015)

A continuation of this work would aim to develop a better understanding of the response of dehydrins to dehydration conditions by increasing the types of sampled tissues and testing other water-stress conditions such as salinity and cold. An RNA seq analysis would give us this complete picture of dehydrin expression responses and would have the potential to reveal many other genes that respond to drought stress and thus help to characterize the genetic architecture of drought stress.

Another way of increasing our understanding of dehydrin functional roles would be to realize *in vitro* experiments to test their capability to prevent water loss, cryoprotection activity, capability to prevent protein aggregation at high temperatures and capability to bind ions and nucleic acid. These *in vitro* functions have been investigated in a few angiosperm dehydrins (Alsheikh, Heyan and Randall, 2003; Goyal, Walton and Tunnacliffe, 2005; Momma et al., 2005; Tompa et al., 2006) and their protective capabilities have been proven. Even the importance of the modularity nature of the

dehydrins has been tested, for example the removal of K-segments has affected their protective capability (Reyes et al., 2008). As *Picea glauca* dehydrins presented an elevated structural diversity, testing their functional roles by *in vitro* essays would likely expand our understanding of the relationship between structure and function in this family.

4.3.3 Linking gene structure and gene family evolution.

In the present thesis, the evolution of gene structure and gene families were investigated separately but a more integrated understanding these two complementarity aspects of gene evolution could shed new light into conifer genomics. For example, the analysis of factors that impacted the evolution of gene structure could be applied to furthering our understanding of the dehydrin gene family. Gene structure evolution in large gene families such as the dehydrins could underpin variations in the modular organization of amino acid motifs. Phenomena such as insertions and deletions associated with repetitive element activity or sequence variations such as exon doubling or shuffling could be at the origin of the structural variation we have described.

4.4 References

Alsheikh MK, Heyen BJ, Randall SK. Ion Binding Properties of the Dehydrin *ERD14* Are Dependent upon Phosphorylation. Journal of Biological Chemistry. 2003;278:40882–9.

Bautista R, Villalobos DP, Díaz-Moreno S, Cantón FR, Cánovas FM, Claros MG. Toward a *Pinus pinaster* bacterial artificial chromosome library. Annals of Forest Science. 64:855–64.

Gauthier S, Bernier P, Kuuluvainen T, Shvidenko AZ, Schepaschenko DG. Boreal forest health and global change. Science. 2015;349:819–22.

Goyal K, Walton LJ, Tunnacliffe A. LEA proteins prevent protein aggregation due to water stress. Biochemical Journal. 2005;388:151–7.

Hamberger B, Hall D, Yuen M, Oddy C, Hamberger B, Keeling CI, et al. Targeted isolation, sequence assembly and characterization of two white spruce (*Picea glauca*) BAC clones for terpenoid synthase and cytochrome P450 genes involved in conifer defense reveal insights into a conifer genome. BMC Plant Biology. 2009;9:106.

Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nature Reviews Genetics. 2008;9:411–2.

Kjellsen TD, Yakovlev IA, Fossdal CG, Strimbeck GR. Dehydrin accumulation and extreme low-temperature tolerance in Siberian spruce (*Picea obovata*). Tree Physiology. 2013;33:1354–66.

Kovach A, Wegrzyn JL, Parra G, Holt C, Bruening GE, Loopstra CA, et al. The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics. 2010;11:420.

Magbanua ZV, Ozkan S, Bartlett BD, Chouvarine P, Saski CA, Liston A, et al. Adventures in the enormous: a 1.8 million clone BAC library for the 21.7 Gb genome of Loblolly pine. PLoS ONE. 2011;6:e16214.

Momma M, Kanego S, Haraguchi K, Matsukura U. Peptide mapping and assessment of cryoprotective activity of 26/27-kDa dehydrin from soybean seeds. Bioscience, Biotechnology, and Biochemistry. 2003;67:1832–5.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, et al. The Norway spruce genome sequence and conifer genome evolution. Nature. 2013;497:579–584.

Pavy N, Pelgas B, Laroche J, Rigault P, Isabel N, Bousquet J. A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. BMC Biology. 2012;10:84.

Perdiguero P, Barbero MC, Cervera MT, Soto Á, Collada C. Novel conserved segments are associated with differential expression patterns for *Pinaceae* dehydrins. Planta. 2012;236:1863–74.

Perdiguero P, Soto Á, Collada C. Comparative analysis of *Pinus pinea* and *Pinus pinaster* dehydrins under drought stress. Tree Genetics & Genomes. 2015;11:70.

Prunier J, Verta J-P, MacKay JJ. Conifer genomics and adaptation: at the crossroads of genetic diversity and genome function. New Phytolologist. 2016;209:44–62.

Reyes JL, Campos F, Wei H, Arora R, Yang Y, Karlson DT, et al. Functional dissection of hydrophilins during in vitro freeze protection. Plant Cell Environ. 2008;31:1781–90.

Tompa P, Bánki P, Bokor M, Kamasa P, Kovács D, Lasanda G, et al. Protein-water and protein-buffer interactions in the aqueous solution of aniIntrinsically unstructured plant dehydrin: NMR intensity and DSC aspects. Biophysical Journal. 2006;91:2243–9.

Warren RL, Keeling CI, Yuen MMS, Raymond A, Taylor GA, Vandervalk BP, et al. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. Plant J. 2015;83:189–212.

Yakovlev IA, Asante DKA, Fossdal CG, Partanen J, Junttila O, Johnsen O. Dehydrins expression related to timing of bud burst in Norway spruce. Planta. 2008;228:459–72.