

Table des matières

Résumé	iii
Abstract.....	v
Table des matières	vii
Liste des tableaux	xi
Liste de figures.....	xiii
Remerciements.....	xv
Avant-propos.....	xvii
Liste des abréviations	xix

Chapitre 1.	1
Analyses transcriptomiques chez les conifères	1
1.1 Origine et importance économique et écologique des conifères	2
1.2 Biologie et caractéristique unique des arbres et des conifères.....	3
1.2.1 Différentiation tissulaire et formation du bois	3
1.2.2 Stratégies d'adaptation	6
1.3 Caractéristiques du génome des conifères	8
1.4 Méthodes de profilage transcriptionnel.....	9
1.5 Profilage transcriptionnel chez les conifères	10
1.5.1 Formation du bois	17
1.5.2 Stratégies d'adaptation	20
1.5.3 Autres études chez des conifères	25
1.6 Mise en contexte du projet	26
1.7 Objectifs.....	27
1.8 Références bibliographiques	29

Chapitre 2.	39
Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression	39
2.1 Résumé	41
2.1.1 Contexte	41
2.1.2 Résultats	41
2.1.3 Conclusion	41
2.2 Abstract	43
2.2.1 Background	43
2.2.2 Results	43
2.2.3 Conclusion	43
2.3 Background	45
2.4 Results and Discussion	47
2.4.1 Development of an oligonucleotide microarray for spruces (<i>Picea spp.</i>)	47
2.4.2 Differential expression in the vascular transcriptome is conserved among <i>Picea</i> species	48
2.4.3 Validation of the microarray and profiling results by RNA-Sequencing	51
2.4.4 PiceaGenExpress contains reference profiles that reveal patterns of tissue preferential expression	53
2.4.5 Non-annotated genes are expressed at low levels and in fewer tissues	54
2.4.6 Diversity of expression profiles varies within and among gene families with related functions	57
2.4.7 Expression of LTR retrotransposon sequences	58
2.5 Conclusions	61
2.6 Methods	63
2.6.1 Evaluation of array design and manufacture parameters with test oligonucleotide microarray	63
2.6.2 Microarray design and manufacture	64
2.6.3 Plant materials	65
2.6.4 RNA extraction, labelling and hybridization	66
2.6.5 Microarray data processing and analysis	67
2.6.6 The PiceaGenExpress database	68
2.6.7 RNA-Seq data processing and analysis	69
2.7 Acknowledgments	69
2.8 References	69
2.9 Additional files	74

Chapitre 3.	87
Modular organization of the white spruce (<i>Picea glauca</i> (Moench) Voss) transcriptome reveals functional organization and evolutionary signatures	87
3.1 Résumé	89
3.2 Abstract	91
3.3 Introduction	93
3.4 Materials and methods	95
3.4.1 Plant material	95
3.4.2 RNA isolation, labelling and hybridization	97
3.4.3 Microarray pre-processing and statistical analyses	97
3.4.4 Functional annotation analysis	98
3.4.5 Reverse transcription quantitative PCR (RT-qPCR)	99
3.5 Results	100
3.5.1 The expression of most white spruce genes vary significantly across vegetative tissues	100
3.5.2 The white spruce transcriptome is organized into coexpression groups classifying tissues based on their types or physiological functions	100
3.5.3 Gene ontology term enrichments show functional diversity among coexpression groups	102
3.5.4 Conifer-specific genes are more strongly associated with secondary meristematic tissues	104
3.5.5 Highly conserved invariant genes are associated with high expression and cell basic functions	104
3.5.6 Expressional and functional diversity among conserved secondary xylem and phelloderm preferential genes	107
3.5.7 Distinct expression patterns characterize wood formation at different stages during a growth season	107
3.5.8 Variation in secondary cell wall regulatory network hub genes is associated with xylem transcriptome reorganization during the growth season	109
3.6 Discussion	115
3.6.1 Evolutionary signatures of tissue differentiation	115
3.6.2 Modular transcriptome organization underpins tissue differentiation	116
3.6.3 Coexpression is associated with functional similarity	117
3.6.4 Temporal reorganization of gene expression networks in the vascular transcriptome	119
3.7 Concluding remarks	120
3.8 Acknowledgements	120
3.9 References	120
3.10 Supporting information	128

Chapitre 4.	139
Conclusions générales et perspectives	139
4.1 Discussion des résultats majeurs	139
4.1.1 La première puce à oligonucléotide	140
4.1.2 Étude sur plusieurs tissus: construction d'une base de données d'expression	140
4.1.3 Étude sur plusieurs tissus: organisation du transcriptome et réseaux transcriptionnels	142
4.1.4 Conservation de profil d'expression	144
4.2 Études complémentaires	145
4.2.1 L'approche RNA-seq pour les études complémentaires	145
4.2.2 Des questions pour les études complémentaires	146
4.3 Retombées pratiques et enjeux forestiers	147
4.4 Références bibliographiques	149

Liste des tableaux

Tableau 1.1	Études de profilage transcriptionnel sur la formation du bois chez les conifères	11
Tableau 1.2	Études de profilage transcriptionnel sur les stratégies d'adaptation chez les conifères	13
Table 2.1	Development a large-scale oligonucleotide array for spruces (<i>Picea</i> spp.)	47
Table 2.2	Analysis of probes: sequence similarity of probes aligned to <i>Picea glauca</i> and <i>P. sitchensis</i> sequences	48
Table 2.3	RNA-Seq data	51
Table 2.4	Validation of microarray results by RNA-Seq	52
Table 2.5	Tissue specificity in RNA-Seq	52
Table 2.6	The PiceaGenExpress database: sample characteristics, hybridizations and detected genes	54
Table S2.1	PiceaGenExpress transcript profiles.	85
Table S2.2	Summary statistics of PiceaGenExpress transcript profiles.	85
Table S2.3	Expression of cDNA selected for test microarray	85
Table 3.1	Numbers of genes detected and classification as variable and invariant based on transcript levels	99
Table 3.2	Function and cross-species conservation of invariant genes according to their expression levels	105
Table 3.3	Functional annotation of the top ten most connected genes (hubs) of the networks	113
Table S3.1	Compilation of gene expression results from experiments 1 and 2 and gene network analysis and annotation summary	134
Table S3.2	Functional annotation analysis of gene groups	134
Table S3.3	RT-qPCR primer sequences and microarray validation by RT-qPCR	134
Table S3.4	Number of genes per co-expression group in white spruce (<i>Picea glauca</i>)	134
Table S3.5	Under or overrepresentation of conifer specific genes across co-expression groups	135
Table S3.6	Enriched pathways in xylem (M2-7) preferential network	136
Table S3.7	Members of the secondary cell wall gene network shown to be transactivated by PgNAC-7 or MYB transcription factors in white spruce (<i>Picea glauca</i>)	137
Table S3.8	Distribution of detected genes across invariant gene and co-expression groups	138

Liste de figures

Figure 2.1	Interspecific comparison of hybridization intensities in secondary xylem	49
Figure 2.2	Preferential expression in secondary vascular tissues of three spruce species	50
Figure 2.3	The PiceaGenExpress database reveals tissue preferential and conserved expression patterns within three gene families	55
Figure 2.4	Expression classes and numbers of tissue of annotated and non-annotated sequences	56
Figure 2.5	Gene expression patterns in two osmotic regulation protein families based on the PiceaGenExpress database	59
Figure 2.6	Expression patterns of LTR retrotransposons based on the PiceaGenExpress database ...	60
Figure S2.1	Effect of SNPs on hybridization signal intensities and differential expression ratios	79
Figure S2.2	Comparison of differential expression results from a cDNA microarray and the test oligonucleotide microarray	80
Figure S2.3	Interspecific comparison of hybridization intensities in the phelloderm	81
Figure S2. 4	Hybridization signal intensities of genes in each of the 10 expression classes in each tissue in PiceaGenExpress transcript profiles	82
Figure S2. 5	Expression classes and numbers of tissue of annotated and non-annotated sequences	83
Figure S2.6	Hierarchical clustering dendrograms of gene expression within two osmotic regulation protein families.	84
Fig. 3.1	Classification of collected tissues based on gene expression	96
Fig. 3.2	Expression module and coexpression groups across tissues in white spruce (<i>Picea glauca</i>)	103
Fig. 3.3	Functional analysis of invariant genes (InvG) and coexpression groups (M1a-M11b) in white spruce (<i>Picea glauca</i>)	106
Fig. 3.4	Distribution of secondary xylem or phelloderm preferential genes conserved among Picea species across coexpression groups. Vascular tissue preferential genes in white (<i>Picea glauca</i>), Norway (<i>P. abies</i>) and black (<i>P. mariana</i>) spruces were identified in Raheison <i>et al.</i> (2012).	108
Fig. 3.5	Distribution of temporally variable xylem expressed genes across coexpression groups ..	110
Fig. 3.6	Phenylpropanoid pathway leading to lignin production	112
Fig. 3.7	Functional, network analysis and temporal expression patterns of PgNAC-7 connected genes	114
Fig. S3.1	Hierarchical clustering of white spruce (<i>Picea glauca</i>) tissues based on variable genes; .	129
Fig. S3.2	Representation of co-expression groups and their correlation to white spruce (<i>Picea glauca</i>) tissues	130
Fig. S3.3	Functional annotation of invariant genes and co-expression groups	132
Fig. S3.4	Gene degree distribution in a double log scale of the xylem (M2-7) preferential network .	133

Remerciements

Cette thèse ne se serait pas réalisée sans la contribution des personnes que je tiens à remercier ici.

J'exprime mes profonds remerciements à mon directeur de thèse, le Professeur John MacKay, qui m'a accueilli dans son laboratoire et m'a offert la chance de réaliser cette thèse. Merci pour ta disponibilité, tes conseils précieux, ton encadrement constant, et tes qualités humaines. Sois assuré de mon estime et de ma profonde gratitude.

Je tiens également à remercier Isabelle Giguère et Sébastien Caron, professionnels de recherche de l'équipe du laboratoire de John MacKay, pour leur rôle irremplaçable dans la réalisation de diverses expériences citées dans ce travail.

Ma vive reconnaissance va aussi à tous les membres de l'équipe du laboratoire de John MacKay pendant la période de 2010 à 2015. Vos commentaires et vos suggestions m'ont permis d'améliorer la qualité de ma recherche.

Ma gratitude va également à Jacqueline Grima-Pettenati, Armand Séguin et Louis Bernier d'avoir accepté d'évaluer ce travail en tant qu'examineur.

Enfin, je suis heureux de pouvoir exprimer ma profonde gratitude à ma chère famille. Sa patience, sa compréhension et son soutien indéfectible m'ont été extrêmement importants et font l'objet de mon admiration.

Avant-propos

Cette thèse comprend une introduction (chapitre 1), une conclusion générale (chapitre 4) ainsi que deux chapitres principaux (2 et 3) rédigés et présentés dans un format approprié pour une publication scientifique.

Le chapitre 2 est publié sous la référence : Raheison E, Rigault P, Caron S, Poulin P-L, Boyle B, Verta J-P, Giguère I, Bomal C, Bohlmann J, MacKay J. 2012. Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression. *BMC genomics* 13(1) : 434.

Le chapitre 3 est publié sous la référence : Raheison E, Giguère I, Caron S, Lamara M, MacKay J. 2015. Modular organization of the white spruce (*Picea glauca*) transcriptome reveals functional organization and evolutionary signatures. *New Phytologist*. doi : 10.1111/nph.13343

Les coauteurs ont contribué au développement des méthodes, à la réalisation de certaines parties des expériences, ou encore à la conception de l'étude. Moi, Élie Raheison, j'ai réalisé la majorité des analyses de résultats, leur interprétation et la rédaction de chaque chapitre sous la supervision de John MacKay, directeur de la thèse.

Liste des abréviations

4CL	<i>4-coumarate-CoA ligase</i>
ADN	Acide désoxyribonucléique
ADNc	Acide désoxyribonucléique complémentaire
AOC	Allène oxyde cyclase
AOS	Allène oxyde synthase
AP2	<i>Apatala2</i>
AP2/EREBP	<i>Apetala2/ethylene-responsive element-binding protein</i>
ARF2	<i>Auxin response factor 2</i>
ARNm	Acide ribonucléique messenger
bHLH	<i>Basic helix-loop-helix</i>
BSA	<i>Bovine serum albumin</i>
bZIP	<i>Basic leucine zipper</i>
C3H	<i>P-coumaroyl- shikimate/quinate 3'-hydroxylas</i>
CAD	<i>Cinnamyl alcohol dehydrogenase</i>
CCoAOMT	<i>Caffeoyl-CoA O-methyltransferase</i>
cDNA	<i>Complementary deoxyribose nucleic acid</i>
CesA	<i>Cellulose synthase</i>
EST	<i>Expressed sequence tag</i>
F5H	<i>Ferulate 5-hydroxylase</i>
FL-cDNA	<i>Full-length – complementary deoxyribose nucleic acid</i>
GGPP	<i>Geranylgeranyl diphosphate</i>
IAA	<i>Indole-3-acetic acid</i>
LAC	Laccase
LEA	<i>Late embryogenesis abundant</i>
LINE	<i>Long interspersed elements</i>
LOX	Lipoxygénase
LTR	<i>Long terminal repeat</i>
LTR	<i>Long terminal repeat</i>
NAC	NAM (<i>no apical meristem</i>) – ATAF (<i>Arabidopsis activation factor</i>) – CUC (<i>cup-shaped cotyledon</i>)
NBS-LRR	<i>Nucleotide binding site-leucine-rich repeat</i>
PAL	Phénylalanine ammonia-lyase

PB1	<i>Phox/Bem1p</i>
PCR	<i>Polymerase chain reaction</i>
Pfam	<i>Protein families</i>
PgCAD	<i>Cinnamyl alcohol dehydrogenase in Picea glauca</i>
PgDHS	<i>3-deoxy-d-arabino-heptulosonate 7-phosphate synthase in Picea glauca</i>
PgMYB8	<i>MYB8 in Picea glauca</i>
PgNAC-7	<i>NAC 7 in Picea glauca</i>
Pg β glu-1	<i>Picea glauca β-glucosidase-1</i>
RNA	<i>Ribonucleic acid</i>
RNA-seq	<i>Ribonucleic acid sequencing</i>
SAGE	<i>Serial analysis of gene expression</i>
SDS	<i>Sodium dodecyl sulfate</i>
SINE	<i>Short interspersed element</i>
SNP	<i>Single nucleotide polymorphism</i>
SSC	<i>Sodium chloride/sodium citrate</i>
SUS	<i>Suspensor</i>
TAIR	<i>The Arabidopsis Information Resource</i>
UBA	<i>Ubiquitin-like activating</i>

Chapitre 1.

Analyses transcriptomiques chez les conifères

Le transcriptome est l'ensemble des ARN messagers ou transcrits produits dans un tissu ou un organe. Le transcriptome se distingue du génome selon deux caractéristiques :

- Premièrement, la taille du génome peut être très différente selon les espèces (p.ex. 125 Mb pour l'*Arabidopsis thaliana*, GI [2000], 389-466 Mb pour le riz, Goff et coll. [2002], Yu et coll. [2002], IRGS [2005] et 20 000 Mb pour l'épinette blanche ou *Picea glauca* [Moench] Voss, Birol et coll. [2013]). Dans le cas du transcriptome, la taille varie beaucoup moins (p.ex. 40 Mb pour l'*A. thaliana*, 49 Mb pour le riz et à 41 Mb pour l'épinette blanche, Rigault et coll. [2011]).
- Deuxièmement, le génome est plus stable, car il change très peu à l'intérieur d'un individu. Le transcriptome est très dynamique, car il varie selon les tissus de l'individu (p.ex. Prasad et coll., 2013; Downs et coll., 2013), au sein d'un même tissu selon les stades de développement (p.ex. Li et coll., 2010; Mishima et coll., 2014) et au sein d'un même type de tissu selon les conditions biologiques (p.ex. Sato et coll., 2014; Mageroy et coll., 2015).

Le caractère dynamique du transcriptome fait qu'il est le reflet de l'activité de plusieurs gènes dans un tissu. C'est pourquoi l'analyse transcriptomique est utilisée depuis longtemps pour comprendre les mécanismes moléculaires reliés aux phénomènes biologiques chez les conifères (Allona et coll., 1998). Ces phénomènes biologiques qui incluent la formation du bois (p.ex. Porté et coll., 2002) et les stratégies d'adaptation face aux fluctuations environnementales (p.ex. King et coll., 2011; Alberto et coll., 2013), sont reconnus comme étant à l'origine de la grande taille et du mode de vie pérenne des conifères et autres arbres.

Dans ce chapitre, nous faisons un survol de l'origine, de l'importance écologique et économique des conifères, ainsi que des caractéristiques biologiques des conifères et autres arbres. Ensuite, nous passons en revue les analyses transcriptomiques chez les conifères. L'analyse de transcriptome peut être qualitative ou quantitative. L'analyse qualitative identifie les gènes exprimés et permet de découvrir des gènes qui pourraient être spécifiques aux conifères (p.ex. Mann et coll., 2013), de créer des banques d'ADNc (p.ex. Ralph et coll., 2008) et de construire des transcriptomes de référence sur des tissus à un stade de développement et à une condition biologique donnés (p.ex. Wanfeng et coll., 2014; Canales et coll., 2014; Pinosio et coll., 2014). L'analyse quantitative, aussi appelée profilage transcriptionnel, évalue le niveau d'expression des gènes à l'échelle du génome. Dans ce chapitre, nous portons notre attention sur les travaux de profilage transcriptionnel sur les conifères.

1.1 Origine et importance économique et écologique des conifères

Les conifères forment le plus grand groupe des gymnospermes. Avec 68 genres et 600 espèces, ils représentent à eux seuls plus de la moitié de toutes les espèces et plus des deux tiers de tous les genres de gymnospermes (Christenhusz et coll., 2011). Les conifères sont apparus il y a moins de 200 millions d'années (Lu et coll., 2014). On peut les regrouper en deux clades nord et sud selon leur histoire évolutive (Leslie et coll., 2012). Le clade du nord, qui est composé des Pinaceae et des Cupressoideae, est caractérisé par un temps de divergence plus récent et un taux de renouvellement d'espèces plus élevé que le clade du sud (Araucariaceae, Podocarpaceae, Callitroideae). La plus importante famille des conifères est celle des Pinaceae qui comprennent 225 espèces et 11 genres (Farjon, 2001). Bien que présentes dans des régions chaudes de la planète comme l'est de l'Asie, le nord de l'Afrique ou le sud des États-Unis, les membres de la famille des Pinaceae sont largement confinés aux régions tempérées et boréales de l'hémisphère nord (Farjon, 2001). Les conifères font l'objet de plusieurs études scientifiques à cause de leur importance écologique et économique (p.ex. Alberto et coll., 2013).

Les conifères sont écologiquement importants dans les zones boréales et tempérées de la terre et ils sont fortement représentés au Canada. Les genres *Pinus*, *Picea*, *Abies* et *Larix* des Pinaceae sont des composantes principales des forêts et des zones boisées qui couvrent 55 % de la région boréale canadienne (Brandt et coll., 2013). Cette dernière séquestre 208 milliards de tonnes de carbones; l'équivalent de plus de 25 années d'émissions des industries du monde (Carlson et coll., 2009). Grâce à l'importance de sa forêt, la région boréale canadienne offre des services écologiques non marchands estimés à 703 milliards de dollars par année (Anielski et Wilson, 2009). Ces services incluent le stockage du carbone, la purification de l'eau, la protection contre les inondations et le contrôle des ravageurs par les oiseaux. La forêt boréale canadienne est aussi un habitat essentiel dont dépend la survie de 325 espèces d'oiseaux, soit près de la moitié des espèces d'oiseaux de l'Amérique du Nord (Wells et Blancher, 2011). Elle abrite également des mammifères, dont certains (caribou des bois ou *Rangifer tarandus* L.; bison des bois ou *Bison bison athabasca* Rhoads; carcajou ou *Gulo gulo* L.) sont classés comme étant des espèces menacées d'extinction selon COSEWIC (2013).

Les conifères sont économiquement importants. Ils fournissent différents produits incluant le bois rond, le bois de chauffage, le bois de trituration, la grume de sciage et de placage (UNECE, 2013). En 2011, la production de bois ronds s'est élevée à 145 millions m³ au Canada, dont 82 % provenaient des conifères (UNECE, 2013). Peu de ce bois a été utilisé pour l'énergie; la plupart a servi à ravitailler les industries (UNECE, 2013). L'industrie forestière a rapporté environ 21 milliards de dollars; l'équivalent de 1,25 % du produit intérieur brut du Canada en 2013. L'importance

économique mondiale des conifères se manifeste également par le grand nombre de programmes d'amélioration de plusieurs pays repartis sur tous les continents (Mullin et coll., 2011). Au Canada, les programmes d'amélioration forestiers intègrent le sapin de Douglas (*Pseudotsuga menziesii*) et différentes espèces de pins (pin blanc ou *Pinus strobus* L., pin rouge ou *P. resinosa* Ait., pin gris ou *P. banksiana* Lamb., pin rigide ou *P. rigida* Mill., pin tordu ou *P. contorta* Douglas et pin argenté ou *P. monticola* Douglas), d'épinettes (épinette blanche, épinette noire ou *Picea mariana* [M] Britton, épinette rouge ou *P. rubens* Sarg., et épinette de Sitka ou *P. sitchensis* [Bong.] Carrière) et de cyprès (cyprès de Nootka ou *Chamaecyparis nootkatensis* [D. Don.] Spach et cyprès du Portugal ou *Cupressus lusitanica* Mill.) (Mullin et coll., 2011).

1.2 Biologie et caractéristique unique des arbres et des conifères

La grande taille et le mode de vie pérenne sont deux attributs qui distinguent les arbres des autres plantes. Ces attributs sont mis en place grâce à des phénomènes biologiques tels que la différenciation tissulaire liée à la grande production de bois, et les stratégies d'adaptation face aux agressions causées par des facteurs abiotiques et biotiques.

1.2.1 Différenciation tissulaire et formation du bois

La différenciation tissulaire a été cruciale pour l'évolution des premières plantes terrestres qui sont apparues vers la fin de l'Ordovicien, il y a 450 millions d'années ou Ma (Garbaye, 2007). Elle a eu lieu bien avant que les gymnospermes et les angiospermes ne divergent il y a environ 380 Ma (Kenrick et Crane, 1997) et avant que les conifères n'apparaissent il y a 200 Ma (Lu et coll., 2014). Les arbres, comme toutes les plantes vasculaires, ont pu réaliser la conquête du milieu terrestre grâce à des organes et des tissus bien différenciés et spécialisés (Garbaye, 2007). Ils possèdent trois organes végétatifs dont les rôles sont complémentaires : la feuille est un organe photosynthétique; la tige est un organe de soutien qui porte le feuillage et les bourgeons aériens; et la racine est un organe d'ancrage, d'absorption de l'eau et des sels minéraux, et d'entreposage de réserves. Chacun de ces organes possède des tissus avec un nom décrivant sa fonction. Il s'agit des tissus de revêtement (phelloderme et liège), des tissus vasculaires (xylème et phloème), des tissus fondamentaux (parenchyme, collenchyme, sclérenchyme) et des tissus méristématiques.

Les tissus méristématiques peuvent être primaires ou secondaires selon leur mode de croissance. Ainsi, les méristèmes primaires qui se trouvent à l'extrémité des tiges et des racines (méristèmes apicaux) et au niveau des entre-nœuds de la tige (méristèmes intercalaires) assurent la croissance en longueur. Les méristèmes secondaires se trouvent en position latérale dans les arbres et assurent la croissance en épaisseur. Les méristèmes secondaires sont au nombre de deux : le cambium et le

phellogène. Le phellogène est un anneau de méristème latéral qui produit le phelloderme vers l'intérieur et le liège vers l'extérieur. À l'intérieur, au-delà du phellogène, se trouve le cambium, un anneau latéral qui produit les tissus vasculaires secondaires tels que le xylème secondaire (vers le centre) et le phloème secondaire (vers l'écorce). Chez les arbres, l'effet de la croissance secondaire, plus particulièrement celle du xylème secondaire ou bois, est remarquable.

La grande taille des arbres est due à la grande production des matières ligneuses au niveau du xylème secondaire (bois). Le bois assure des rôles de soutien mécanique, de protection et de transport de l'eau et des sels minéraux. Il est particulièrement abondant dans les organes tels que les branches et le tronc. Chez le pin maritime (*Pinus pinaster* Ait), le taux en bois est estimé à près de 90 % la biomasse totale de ces organes ligneux peu importe l'âge de l'arbre (Porté et coll., 2002). Les troncs renferment près de 85 % de bois (le reste est constitué de l'écorce) selon une estimation basée sur des relations allométriques établies sur des conifères et des feuillus des forêts canadiennes (Lambert et coll., 2005; Ung et coll., 2008). L'importance en biomasse de ces organes ligneux augmente avec l'âge de l'arbre. Par rapport à la biomasse aérienne totale, l'importance en biomasse que représentent les branches et les troncs passe de 55 % dans des arbres âgés de 5 ans, à plus de 90 % dans des arbres âgés de plus de 25 ans (Porté et coll., 2002). Ces deux organes représentent environ 85 % de la biomasse aérienne chez les conifères et plus de 95 % chez les feuillus (Lambert et coll., 2005; Ung et coll., 2008). En raison de grande production des matières ligneuses, les arbres sont un énorme puits de carbone et doivent investir beaucoup d'énergie pour la synthèse des constituants du bois.

Le bois est essentiellement composé de cellulose (45 %), d'hémicellulose (25 %) et de lignine (25 %) (Sjostrom, 1993). Le bois des conifères est différent de celui des angiospermes. Chez les conifères, le bois est composé des trachéides (à 90 %) et non de vrais vaisseaux ou de fibres comme chez les angiospermes. De plus, chez les conifères, la lignine est composée principalement des sous-unités d'alcool coniférylique (G ou Guaïacyles) alors qu'elle est formée principalement du mélange des sous-unités d'alcool coniférylique et sinapylique (S/G ou Syringyles / Guaïacyles) chez les angiospermes (Donaldson, 2001). Aussi, l'hémicellulose des conifères est composée d'un mélange d'hétéromannanes (galactoglucomannane et arabinoglucuronoxylane) tandis que l'hémicellulose des angiospermes est composée des xylanes (glucuronoxylane).

La formation du bois implique cinq processus de différenciation : la division des cellules cambiales, l'expansion des cellules, le dépôt de la paroi secondaire, la lignification de la paroi et la mort cellulaire programmée (Hertzberg et coll., 2001). Le processus de la formation du bois qui est lié à l'activité cambiale est influencé par des facteurs reliés à l'environnement et aux stades de développement. Ces facteurs peuvent laisser des empreintes permanentes dans la structure du bois et

générer différents types de bois. Chez les conifères, il s'agit des bois initial et final, du bois de compression et du bois opposé, ou des bois juvénile et mature.

Dans les régions tempérées et boréales, le bois initial et le bois final se forment pendant la saison de croissance suivant le cycle de développement annuel des arbres. Leur formation est aussi influencée par la variation des facteurs abiotiques qui peut être très marquée entre les saisons (Hertzberg et coll., 2001). Ainsi, le bois initial se forme au début du printemps quand la température, la photopériode et la pluviométrie sont favorables à la période active de croissance. La formation du bois final a lieu vers la fin de la saison de croissance et quand les ressources hydriques s'amenuisent. Pendant cette période, la division et l'expansion des cellules cambiales diminuent au profit du dépôt et de la lignification de la paroi secondaire (Uggla et coll., 2001). C'est ainsi que le bois final est différent du bois initial sur le plan anatomique et chimique. Chez les conifères, le bois final est composé de trachéides généralement plus longues, à faible diamètre, à faible angle d'inclinaison des microfibrilles (donc plus dense), à paroi cellulaire plus épaisse et à teneur plus élevée en cellulose (Uggla et coll., 2001; Barnett et Jeronimidis, 2003).

Le bois de réaction et le bois opposé se forment sur des arbres soumis à des forces gravitationnelles et mécaniques (vent et pente) (Yamashita et coll., 2007). Les arbres qui se sont écartés de la position verticale (perpendiculaire au sol) présentent un tronc incliné. Chez les conifères, le bois opposé qui se forme à la face supérieure des tiges inclinées a des caractéristiques chimiques et anatomiques comparables à celles du bois normal. En réponse aux stimuli gravitationnels, le bois de réaction, qui est appelé bois de compression chez les conifères, se forme sur la face inférieure des tiges inclinées et exerce une pression longitudinale tendant à dresser l'axe. Le bois de compression possède des caractéristiques chimiques et anatomiques différentes du bois opposé ou normal. Il a des trachéides plus courtes, à angle d'inclinaison des microfibrilles plus important (par rapport à l'axe vertical), à paroi secondaire enrichie en lignine et à faible teneur en cellulose (Yamashita et coll., 2007; Donaldson et coll., 2004).

La formation du bois juvénile, du bois de transition et du bois mature est reliée à l'âge et à la phénologie de l'arbre. Le bois juvénile se forme pendant la période végétative où la croissance en hauteur de l'arbre est forte. La formation du bois de transition débute au moment où la croissance en diamètre commence; tandis que la formation du bois mature se met en place avec la fin de la croissance en hauteur (p.ex. Burdon et coll., 2004; Mansfield et coll., 2009). Les jeunes arbres sont constitués seulement du bois juvénile alors que les arbres adultes peuvent contenir les trois types de bois dans la partie inférieure dépourvue de branches vertes (Larson, 1969). Les cernes de croissance témoignent de l'âge du bois : au centre de l'arbre on trouve le bois juvénile (de la moelle) et le bois

mature vers l'extérieur de l'arbre (vers l'écorce). Ces bois sont reconnaissables par leurs caractéristiques différentes. Le bois juvénile des conifères par exemple, a une densité plus faible, des trachéides plus minces et plus courtes, et une teneur en cellulose plus faible que le bois mature (p.ex. Larson, 1969; Jozsa et Middleton 1994).

1.2.2 Stratégies d'adaptation

Les arbres ont un mode de vie pérenne. Ils peuvent être soumis à des agressions diverses et variables à cause du changement des facteurs abiotiques (paramètres climatiques) ou biotiques (insectes et agents pathogènes). Face à cette pression et devant l'impossibilité de fuir les conditions hostiles, ils ont développé des stratégies d'adaptation durables qui ne devraient pas être surmontées par des variations climatiques, ou par des insectes et des agents pathogènes qui ont un cycle de vie court. Cependant le cycle de vie court permet une adaptation rapides des agents pathogènes. En outre, ces derniers peuvent produire rapidement des populations de grande taille par reproduction sexuée et asexuée.

Les stratégies incluent les mécanismes de défense, la diversité génétique et la plasticité phénotypique. Les mécanismes de défense peuvent être préformés ou constitutifs et induits ou déclenchés en réaction aux blessures causées par des facteurs biotiques. Les conifères semblent notamment avoir investi dans le développement des défenses constitutives. Les mécanismes de défense constitutifs font référence aux barrières physiques et chimiques (métabolites secondaires). Par exemple, les cellules sclérifiées qui contiennent des dépôts de lignine et que l'on retrouve dans le phloème et l'écorce des conifères interfèrent physiquement avec l'alimentation du charançon du pin blanc (*Pissodes strobi* (Peck)) sur l'épinette de Sitka (King et coll., 2011). Les canaux résinifères remplis d'oléorésine et préformés au niveau des aiguilles, de l'écorce et du bois constituent aussi une barrière constitutive chimique contre les insectes et les champignons (Robert et coll., 2010; Hall et coll., 2011). L'oléorésine est composée de terpènes qui sont des métabolites secondaires répulsifs ou toxiques pour les insectes et les champignons (Zulak et Bohlmann, 2010; Hall et coll., 2011). Les mécanismes de défense induits peuvent être aussi physiques et chimiques. Le traitement au jasmonate de méthyle a augmenté la quantité de résines terpénoïdes et le nombre des canaux résinifères dans l'écorce d'épinette de Norvège (*Picea abies* [L.] H. Karst.; Zeneli et coll., 2006). Des résultats similaires ont été obtenus chez des épinettes de Sitka traitées avec des charançons du pin blanc (Byun-McKay et coll., 2006). Bien que distincts, ces deux mécanismes sont complémentaires. La défense induite se met en place quand la défense constitutive échoue.

Une grande diversité génétique caractérise de nombreuses espèces forestières et permet à une population ou à une espèce de mieux s'adapter aux changements des facteurs abiotiques et biotiques

dans l'environnement (St-Clair et Howe, 2011). La diversité génétique qui est nécessaire pour la survie d'une population ou d'une espèce est un paramètre de base des programmes de conservation de la biodiversité (Ledig, 2012). La diversité génétique au sein de la population est significativement plus grande chez les arbres forestiers que chez toute autre plante (Hamrick et coll., 1992). Chez les conifères, la diversité est plus grande chez les populations à large distribution de l'hémisphère nord que chez les populations fragmentées du sud (Alberto et coll., 2013). La grande diversité génétique des conifères du nord est en grande partie liée à leur vaste aire de répartition et au système d'accouplement allogame favorisé par l'anémophilie (Hamrick et coll., 1992; Beaulieu et coll., 2001; Alberto et coll., 2013).

La plasticité phénotypique est la capacité d'un organisme à exprimer des phénotypes différents à partir d'un génotype, en réponse à un changement des conditions climatiques (Franks et coll., 2014). À cause de leur mode de vie pérenne, les arbres ont une grande tolérance ou plasticité à différentes conditions environnementales (Alberto et coll., 2013). La plasticité phénotypique peut se manifester sous différentes formes. Chez les pins par exemple, en cas de sécheresse et d'accroissement de la température, la plasticité phénotypique se manifeste par la réduction du feuillage pour limiter la perte d'eau par évapotranspiration (Delucia et coll., 2000) ou par l'augmentation de la biomasse racinaire pour maximiser l'absorption d'eau (Richter et coll., 2012). La grande plasticité des conifères est démontrée par la réussite des tests de provenances effectués sur plusieurs espèces dans différentes régions du monde (Mullin et coll., 2011). Chacun de ces tests comprenait des populations implantées dans divers sites expérimentaux (Mullin et coll., 2011). Chaque population correspond aux individus qui proviennent d'une localité géographique et climatique. La régulation de l'expression des gènes et notamment la réorganisation du transcriptome en réponse aux changements de l'environnement biophysique sont à la base de la plasticité phénotypique (p.ex. Yeaman et coll., 2014).

Bref, la grande taille et le mode de vie pérenne des arbres sont liés aux phénomènes biologiques, dont la différenciation tissulaire, la formation du bois et les stratégies d'adaptation aux facteurs abiotiques et biotiques. Ces phénomènes biologiques sont en partie gouvernés par plusieurs gènes selon des études effectuées sur d'autres organismes (Roff et Bradford, 2000; Schiff et coll., 2001; Downs et coll., 2013; Porth et coll., 2013). Chez les conifères, des travaux considérant plusieurs gènes du génome ont été menés afin de comprendre les mécanismes moléculaires sous-jacents à ces phénomènes biologiques. La plupart de ces études se basent sur les gènes exprimés et le transcriptome et non directement sur des génomes à cause de la grande taille et de la complexité du génome de conifères.

1.3 Caractéristiques du génome des conifères

Les génomes nucléaires de trois espèces de conifères soit l'épinette blanche (Birol et coll., 2013) et l'épinette de Norvège (Nystedt et coll., 2013) et le pin à encens (*Pinus taeda* L.; Zimin et coll., 2014), ne sont séquencés que depuis 2013. Ces génomes sont énormes et complexes.

Les génomes dont la taille est d'environ 20 Gbp sont énormes comparés à ceux des autres plantes revues dans Michael et Jackson (2013). Ils sont 160 fois plus grands que le génome d'*A. thaliana* (GI, 2000), 43 fois plus grands que le génome du riz (*Oryza sativa* L.; Yu et coll., 2002) et environ 30 fois plus grands que les génomes du peuplier (*Populus trichocarpa* [Torr. & Gray]; Tuskan et coll., 2006) et de l'eucalyptus (*Eucalyptus grandis* Hill & Maiden; Myburg et coll., 2014) et 6 fois plus grands que celui de l'humain (IHGSC 2004). Les gènes des conifères ont des introns plus longs; les gènes de l'épinette blanche ont quatre fois plus de séquences introniques qu'*A. thaliana* et 2 fois plus que le peuplier et le maïs (Sena et coll., 2014). Neale et coll. (2014) rapportent que la longueur moyenne des introns du pin à encens, qui a été évaluée à 2,7 kpb, correspond à plus du double de celle de la vigne et à plus du septuple de celle du peuplier. Les introns ne contribuent pas de façon significative à expliquer la grande taille des génomes de conifères, car les gènes dans lesquels ils se trouvent ne représentent qu'une très faible portion du génome. Les séquences similaires à des séquences codantes représentent 3 % du génome (Morgante et De Paoli, 2011). Cette proportion tombera à moins de 1 % si l'on enlève les pseudogènes et les séquences codantes à l'intérieur des rétroéléments (séquences qui servent uniquement à la transposition) selon Warren et coll. (inédit).

Les génomes des conifères sont complexes, car ils sont riches en pseudogènes (Warren et coll., inédit). Leur grande taille s'explique par l'abondance des séquences répétées. Ces dernières sont en grande partie représentées par les éléments transposables à LTR des superfamilles des Ty3/Gypsy et des Ty1/Copia, qui formeraient plus de 80 % du génome chez le pin à encens et plus de 60 % chez l'épinette de Norvège (Nystedt et coll., 2013). La contribution des autres séquences est relativement faible avec moins de 3 % du génome pour les éléments transposables sans LTR (LINE et SINE) chez les pins (pin sylvestre ou *Pinus sylvestris* L. Wegrzyn et coll., 2013; pin à encens; Nystedt et coll., 2013) et chez les épinettes (épinette blanche, épinette de Norvège; Nystedt et coll., 2013). La même portion a été rapportée pour les séquences répétées en tandem (satellites, minisatellites et microsatellites) chez le pin à encens.

La fonction des gènes est moins bien connue chez les conifères que chez la plupart des plantes modèles. Des études ont montré que près de la moitié des séquences codantes identifiées dans les aiguilles de mélèze de Dahurie (*Larix gmelinii* Rupr.; Men et coll., 2013) et dans différents tissus et

organes d'épinette blanche (Rigault et coll., 2011) n'ont pas d'annotation fonctionnelle. Par la coexpression des gènes, le profilage transcriptionnel de plusieurs gènes et à l'échelle du génome peut renseigner sur la fonction des gènes (Allocco et coll., 2004; Prasad et coll., 2013).

1.4 Méthodes de profilage transcriptionnel

Le profilage transcriptionnel des gènes à l'échelle du génome peut être effectué avec des méthodes de séquençage ou la méthode d'hybridation, notamment grâce au développement des puces à ADN (aussi appelées biopuces).

Les méthodes de séquençage les plus connues sont l'analyse en série de l'expression des gènes (SAGE ou *serial analysis of gene expression*) et le séquençage à ARN (RNA-seq). La SAGE consiste à isoler les ARNm, à rétrotranscrire les ARNm en ADN complémentaires (ADNc), à identifier de courtes séquences spécifiques pour chaque ADNc (10 – 14 pb) appelées étiquettes, à concaténer les étiquettes en une longue molécule unique, à cloner et à séquencer la molécule. Le RNA-seq porte sur l'analyse de courts fragments d'ADNc (30 à 250 pb) appelés lectures, en utilisant une technologie de séquençage à haut débit. Les lectures sont positionnées sur un génome de référence ou assemblées (assemblage *de novo*) pour reconstituer la totalité ou une partie de la séquence des transcrits. Pour ces deux méthodes, le niveau d'expression d'un gène correspond au nombre d'occurrences des étiquettes ou de lectures qui lui sont associées.

La méthode d'hybridation utilise une puce à ADN qui est un support solide auquel sont greffées des sondes. Une sonde est une séquence d'ADNc ou d'oligonucléotides représentant un gène connu et qui est dans la mesure du possible spécifique à ce gène. Elle a pour rôle de détecter le niveau d'expression de ce gène dans un échantillon. La méthode d'hybridation comporte différentes étapes : les ARNm sont convertis en ADNc (simple brin); les ADNc sont marqués à l'aide de colorant(s) fluorescent(s) et hybridés sur la puce qui est ensuite scannée et l'image produite est analysée pour quantifier les signaux fluorescents d'hybridation. Le niveau d'expression d'un gène est proportionnel à l'intensité des signaux d'hybridation de la sonde correspondante. La qualité d'une puce peut être jugée selon la taille des sondes et le nombre de gènes qu'elle contient. Les puces qui renferment un grand nombre de gènes ont une plus grande couverture du génome. Aussi, les sondes d'ADNc plus longues (> 500 nucléotides) sont moins sensibles à l'hybridation croisée (Chou et coll., 2004), alors que les sondes oligonucléotidiques de courtes séquences (moins de 25 nucléotides) sont très sensibles aux polymorphismes (Pullat et coll., 2007).

Ces méthodes ont chacune leurs avantages et leurs inconvénients. La méthode de séquençage la plus utilisée de nos jours est le RNA-seq, car elle permet un séquençage direct sans passer par la lourde

étape du clonage bactérien comme c'est le cas pour la SAGE. Des études comparatives des méthodes d'hybridation et de RNA-seq ont mis en évidence une concordance des résultats issus de ces méthodes (Zhao et coll., 2014; Wang et coll., 2014). Comparé à la méthode d'hybridation, le RNA-seq présente des avantages reliés à sa capacité de détection. Le RNA-seq permet de détecter de nouveaux transcrits, car il n'utilise aucun *a priori* technique comme c'est le cas des sondes greffées sur la biopuce. Il offre aussi une plus grande sensibilité de détection pour une meilleure évaluation de niveau des transcrits (Zhao et coll., 2014; Wang et coll., 2014). Cette technique qui s'est développée avec les technologies de séquençage est plus récente que la technique d'hybridation. Elle est assez dispendieuse pour plusieurs échantillons et relativement nouvelle pour les chercheurs (Zhao et coll., 2014). Jusqu'à présent, le RNA-seq reste moins utilisé que la méthode d'hybridation chez les conifères. On peut cependant noter une tendance d'utilisation accrue du RNA-seq depuis quelques années.

1.5 Profilage transcriptionnel chez les conifères

Les travaux sur le profilage transcriptionnel des conifères dont nous discutons dans ce chapitre sont en grande partie répertoriés dans les tableaux 1.1 et 1.2. Dans ces études, les auteurs ont utilisé la méthode de RNA-seq et la méthode d'hybridation par puce à ADN pour évaluer l'expression de gènes. Ils se sont penchés sur l'étude de la dynamique du transcriptome portant sur les phénomènes biologiques tels que la formation du bois et les stratégies d'adaptation aux perturbations biotiques et abiotiques. Des études comparatives ont été menées afin d'identifier des gènes qui ont un niveau d'expression différentiel à travers les traitements considérés ou préférentiels à un traitement donné. Le nombre ou la proportion de gènes différentiels par rapport au nombre total des gènes analysés ont servi à mesurer l'ampleur de la dynamique du transcriptome (nombre ou proportion élevé des gènes différentiels = grand changement du transcriptome). L'annotation fonctionnelle des gènes différentiels a été effectuée dans le but d'identifier des gènes candidats impliqués dans un processus biologique donné.

Tableau 1.1 Études de profilage transcriptionnel sur la formation du bois chez les conifères

(a) Formation du bois liée aux stades de développement

Espèces	Méthodes	Comparaisons	N^{bre} de gènes différentiels (%)	Critères statistiques	N^{bre} de gènes testés	Références
<i>Pinus taeda</i>	Puce à ADNc	Bois sur cinq dates pendant une saison de croissance	667 (19)	$p \leq 0,0001$	3512	Paiva et coll. (2008)
<i>Cryptomeria japonica</i>	Puce à oligo	Bois initial (formation du bois initial) vs bois final (vers l'entrée en dormance)	10 380 (57)	$p \leq 0,05$; $p_a \leq 0,2$; $ \text{ratio}(\log_2) \geq 1$	18 082	Mishima et coll. (2014)
<i>Cunninghamia lanceolata</i>	Seq à ARN	Cambium pendant la période d'activation vs période de dormance	4415 (7,3)	$p_a \leq 0,001$, $ \text{ratio}(\log_2) \geq 2$	59 669	Qiu et coll. (2014)
-	-	Cambium pendant la période d'activation vs période de réactivation	883 (1,5)	-	-	-
-	-	Cambium pendant la période de réactivation vs période de dormance	4018 (6,7)	-	-	-
<i>P. radiata</i>	Puce à ADNc	Bois juvénile (5ans) : bois initial vs bois final	687 (21)	$p \leq 0,05$	3320	Li et coll. (2010)
-	-	Bois de transition (9ans) : bois initial vs bois final	995 (30)	-	-	-
-	-	Bois mature (30 ans) : bois initial vs bois final	381 (2,6)	-	-	-
<i>P. taeda</i>	Puce à ADNc	Bois à faible densité : bois initial vs bois final	87 (4)	$p \leq 0,01$	2171	Yang et Loopstra (2005)
-	-	Bois à forte densité : bois initial vs bois final	110 (5)	-	-	-
-	-	Bois initial : bois à faible densité vs bois à forte densité	51 (2,3)	-	-	-
-	-	Bois final : bois à faible densité vs bois à forte densité	131 (6)	-	-	-
<i>P. radiata</i>	Puce à ADNc	Bois initial : bois à forte rigidité vs bois à faible rigidité	112 (3,4)	$p \leq 0,05$	3320	Li et coll. (2011)
-	-	Bois final : bois à forte densité vs bois à faible rigidité	295 (8,9)	-	-	-

Tableau 1.1 (suite) Études de profilage transcriptionnel sur la formation du bois chez les conifères

(b) Formation du bois liée aux stress environnementaux

Espèces	Méthodes	Comparaisons	N ^{bre} de gènes différentiels (%)	Critères statistiques	N ^{bre} de gènes testés	Références
<i>Pinus pinaster</i>	Puces à ADNc	Bois de compression vs bois normal	496 (7,2)	$p_a \leq 0,001$; $ \text{ratio}(\log_2) \geq 1,5$	6841	Villalobos et coll. (2012)
<i>P. radiata</i>	Puce à ADNc	Bois de compression vs bois opposé	970 (29)	$p_a \leq 0,05$; $ \text{ratio}(\log_2) \geq 0,6$	3320	Li et coll. (2013)
<i>Chamaecyparis obtusa</i>	Seq à ARN	Bois de compression vs bois normal	2875 (7,1)	$p_a \leq 0,05$	40 602 t	Sato et coll. (2014)

Colonne des méthodes : Puce à ADNc ou Puce à oligo, méthode d'hybridation avec des puces à ADNc ou des puces à oligonucléotides; Seq à ARN, méthode de séquençage à ARN; **Colonne du N^{bre} de gènes différentiels (%)** : N^{bre}, nombre; valeur entre parenthèses, la proportion (%) de gènes différentiels par rapport au nombre total de gènes testés; **Colonne des Critères statistiques** : p, valeur de probabilité p; p_a , valeur de probabilité p corrigée; **Colonne du N^{bre} de gènes testés** : N^{bre}, nombre; t, transcrit unique.

Tableau 1.2 Études de profilage transcriptionnel sur les stratégies d'adaptation chez les conifères
(a) Plasticités phénotypiques liées aux facteurs abiotiques

Espèces	Méthodes	Comparaisons	N ^{bre} de gènes différentiels (%)	Critères statistiques	N ^{bre} de gènes testés	Références
<i>Picea sitchensis</i>	Puce à ADNc	Aiguilles au début (période de transition) vs à la fin de l'automne (dormance)	2224 (10,2)	$pa \leq 0,05$; $ \text{ratio}(\log_2) \geq 2$	21 840	Holliday et coll. (2008)
<i>Pinus contorta</i>	Seq à ARN	Aiguilles des arbres soumis à sept conditions qui varient selon des facteurs environnementaux et des stades de développement	11 658 (48,8)	$pa \leq 0.01$	23 889	Yeaman et coll. (2014)
<i>Picea glauca</i> x <i>P. engelmannii</i>	-	-	6413 (27,3)	-	23 519	-
<i>Pinus sylvestris</i>	Puce à ADNc de <i>Pinus taeda</i>	Hypocotyles soumis à la lumière rouge clair vs rouge sombre	644 (5,1)	$pa \leq 0,05$; $ \text{ratio}(\log_2) \geq 0,95$	12 523	Ranade et coll. (2013)
<i>Picea abies</i>	Seq à ARN	Cals embryonnaires soumis au froid (18 °C) vs chaud (30 °C)	1608 (1,1)	$ \text{ratio}(\log_2) \geq 1$	143 723 t	Yakovlev et coll. (2014)

Tableau 1.2 (suite) Études de profilage transcriptionnel sur les stratégies d'adaptation chez les conifères
(b) Mécanismes de défense liés aux facteurs biotiques

Espèces	Méthodes	Comparaisons	N^{bre} de gènes différentiels (%)	Critères statistiques	N^{bre} de gènes testés	Références
<i>Picea glauca</i> x <i>P. engelmannii</i> ¹	Puce à ADNc	Écorces des arbres non résistants vs résistants au charançon du pin blanc (<i>Pissodes strobi</i>)	191 (1)	$pa \leq 0.05$; $ \text{ratio}(\log_2) \geq 0,6$	17 825	Verne et coll. (2011)
<i>P. sitchensis</i> ¹	Puce à ADNc	Jeunes pousses terminales pourvues vs dépourvues de l'écorce	610 (0,4)	$pa \leq 0.01$; $ \text{ratio}(\log_2) \geq 1$	16 700	Friedmann et coll. (2007)
<i>P. glauca</i> ¹	Puce à oligo	Aiguilles des arbres non résistants vs résistants à la tordeuse des bourgeons de l'épinette	486 (2,1)	$pa \leq 0.05$	23 853	Mageroy et coll. (2015)
<i>Pinus monticola</i> ²	Seq à ARN	Aiguilles des arbres résistants et non traités (contrôles) vs traités au champignon responsable de la rouille vésiculeuse (<i>Cronartium ribicola</i>);	789 (3,4)	$pa \leq 0,05$; $ \text{ratio}(\log_2) \geq 0,6$	23 000	Liu et coll. (2013)
-	-	Aiguilles des arbres non résistants et non traités (contrôles) vs traités à <i>C. ribicola</i>	562 (2,4)	-	-	-
<i>Larix gmelinii</i> ²	Seq à ARN	Aiguilles des arbres non traités vs traités avec de l'acide jasmonique	2383 (4,7)	$pa \leq 0,001$; $ \text{ratio}(\log_2) \geq 1$	51 157 t	Men et coll. (2013)
-	-	Aiguilles des arbres non traités vs traités avec du jasmonate de méthyle	2767 (5,4)	-	-	-

Tableau 1.2 (suite) Études de profilage transcriptionnel sur les stratégies d'adaptation chez les conifères
(b) (suite) Mécanismes de défense liés aux facteurs biotiques

Espèces	Méthodes	Comparaisons	N ^{bre} de gènes différentiels (%)	Critères statistiques	N ^{bre} de gènes testés	Références
<i>Pinus sylvestris</i> ²	Puce à ADNc de <i>Pinus taeda</i>	Racines des arbres non traités vs traités avec un champignon saprophytique (<i>Trichoderma aureoviride</i>). Mesure après 15 jours d'inoculation	10 (0,5)	$pa \leq 0,01$; $ \text{ratio}(\log_2) \geq 0,3$	2109	Adomas et coll. (2008)
-		Racines des arbres non traités vs traités avec un champignon symbiotique (<i>Laccaria bicolor</i>). Mesure après 15 jours d'inoculation	16 (0,8)	-	-	-
-		Racines des arbres non traités vs traités avec un champignon pathogène (<i>Heterobasidion annosum</i>). Mesure après 15 jours d'inoculation	294 (13,9)	-	-	-
<i>Picea sitchensis</i> ²	Puce à ADNc	Écorces des arbres non traités (contrôles) vs traités au charançon du pin blanc	2382 (24,5)	$pa \leq 0,05$; $ \text{ratio}(\log_2) \geq 0,6$	9720 s (5500 gènes)	Ralph et coll. (2006)
-	-	Écorces des arbres non traités (contrôles) vs traités. Le traitement consiste à une blessure mécanique	3089 (31,8)	-	-	-
-	-	Jeunes pousses terminales des arbres non traités (contrôles) vs traités pendant 3 heures à la tordeuse de l'épinette (<i>Choristoneura occidentalis</i>)	358 (3,7)	-	-	-
-	-	Jeunes pousses terminales des arbres non traités (contrôles) vs traités pendant 52 heures avec <i>C. occidentalis</i>	3490 (35,9)	-	-	-

Tableau 1.2 (suite) Études de profilage transcriptionnel sur les stratégies d'adaptation chez les conifères
(b) (suite) Mécanismes de défense liés aux facteurs biotiques

Espèces	Méthodes	Comparaisons	N^{bre} de gènes différentiels (%)	Critères statistiques	N^{bre} de gènes testés	Références
<i>Pinus radiata</i> ²	Puce à oligo de pins	Xylème mucilagineux des arbres non traités (contrôles) vs traités avec l'éthéphon. Mesure après 8 semaines de traitement	23 084 (13)	$pa \leq 0,01$; $ \text{ratio}(\log_2) \geq 1$	175 614 s	Dubouzet et coll. (2014)
-	-	Xylème fibreux des arbres non traités (contrôles) vs traités avec l'éthéphon. Mesure après 8 semaines de traitement	12 718 (7,2)	-	-	-
-	-	Écorces des arbres non traités (contrôles) vs traités avec l'éthéphon. Mesure après 8 semaines de traitement	1761 (1)	-	-	-

Colonne des espèces : ¹ mécanisme de défense constitutive; ² mécanismes de défense induite; **Colonne des méthodes** : Puce à ADNc ou Puce à oligo, méthode d'hybridation avec des puces à ADNc ou des puces à oligonucléotides; Seq à ARN, méthode de séquençage à ARN; **Colonne des comparaisons** : acide jasmonique, jasmonate de méthyle et éthéphon sont des phytohormones impliquées dans la croissance et dans la signalisation de réponses de défense des plantes (Schnurr et coll., 1996; Guo et Ecker, 2004; Wasternack, 2007); **Colonne des N^{bre} de gènes différentiels (%)** : N^{bre}, nombre; valeur entre parenthèses, la proportion (%) de gènes différentiels par rapport au nombre total de gènes testés; **Colonne des Critères statistiques** : p, valeur de probabilité p; pa, valeur de probabilité p corrigée; **Colonne du N^{bre} de gènes testés** : N^{bre}, nombre; t, transcrit unique, s, sondes qui peuvent être des gènes ou d'autres types de séquences (p.ex. sondes contrôles, transcrits).

1.5.1 Formation du bois

Les caractéristiques du bois sont relativement stables car une fois mises en place, elles forment une sorte d'archive de la vie de l'arbre. Ainsi, on peut voir le bois initial et le bois final qui sont les résultats des changements développementaux physiologiques qui ont lieu au cours d'une saison de croissance. Ces changements physiologiques sont observés surtout en régions tempérées et boréales. Ils sont influencés par les changements des facteurs abiotiques (disponibilité hydrique, température et photopériode) dans l'environnement. Il a été également possible de voir chez les conifères la formation du bois de réaction, appelé bois de compression, résultant de contraintes mécaniques.

Des études transcriptomiques qui comparent le bois initial au bois final, et le bois de compression au bois opposé (ou normal) ont été menées sur du pin à encens (Yang et Loopstra, 2005; Paiva et coll., 2008), du pin de Monterey (*Pinus radiata* [D. Don.]; Li et coll., 2010; 2011; 2013), du pin maritime (Villalobos et coll., 2012), du sapin de Chine (Qiu et coll., 2013), du cèdre du Japon (*Cryptomeria japonica* D. Don.; Mishima et coll., 2014) et du cyprès du Japon (*Chamaecyparis obtusa* (Siebold & Zucc.); Sato et coll., 2014). Les gènes différentiels identifiés dans ces études montrent la dynamique du transcriptome lors de la formation du bois (tableaux 1.1).

1.5.1.1 Dynamique du transcriptome à différents stades de développement : cas du bois initial, bois initial et final, bois juvénile, bois de transition et bois mature.

Parmi les travaux où l'on a étudié le bois initial et le bois final, on observe une variation de la proportion des gènes différentiels entre ces deux stades de développement allant de 1,5 % à 57 % du nombre total de gènes analysés (Tableau 1.1a). Cette variation est attribuable aux variations techniques, dont le critère de signification, le nombre et la spécificité des gènes testés. La plus grande proportion a été enregistrée par Mishima et coll. (2014) qui ont considéré des seuils de signification plus permissifs (p-value corrigée 0,2; log ratio de 1) sur 18 082 gènes. La plus faible proportion a été obtenue par Qiu et coll. (2013) qui ont considéré un plus grand nombre de gènes et des seuils de signification plus contraignants (valeur p corrigée 0,001 et log ratio de 2) que Mishima et coll. (2014). Dans la plupart des cas, la proportion est à moins de 10 %, mais elle peut s'élever à 30 % dans le cas des études où l'on a testé des gènes identifiés à partir d'une banque d'ADNc de bois (Li et coll., 2010).

L'examen de chacune des études permet de constater une relation entre l'ampleur de la dynamique du transcriptome et des variations biologiques associées au stade de développement. Ainsi, plus il y a un nombre élevé de gènes différentiels entre deux conditions (ou de gènes préférentiels à une condition), plus grand est le changement transcriptomique associé à ces conditions. Par exemple :

- La formation du bois final entraîne un plus grand changement du transcriptome que la formation du bois initial. Une comparaison des bois à faible et à forte densité a identifié 2,5 fois plus de gènes à expression différentielle dans le bois final que dans le bois initial du pin à encens (Yang et Loopstra, 2005). Un résultat similaire a été rapporté par Li et coll. (2011) qui ont comparé le transcriptome des bois à faible *versus* bois à forte rigidité dans le bois final et dans le bois initial du pin de Monterey. Le bois final est un stade de développement qui est conditionné par des facteurs abiotiques moins favorables, à savoir la diminution de la température, de la photopériode et de la disponibilité hydrique.
- Des phases de transitions vers le début ou la fin de la dormance dans le cycle de vie annuel des arbres entraînent aussi un plus grand changement transcriptomique dans le bois. Mishima et coll. (2014) ont trouvé 1,6 fois plus de gènes préférentiels dans le bois au début de la dormance que pendant la formation du bois du cèdre du Japon. Après avoir étudié l'activité du cambium du sapin de Chine, Qiu et coll. (2013) ont trouvé un plus grand nombre de gènes différentiels vers le début (comparaison des cambiums en phase d'activation *versus* phase de dormance) et la fin (comparaison des cambiums en phase de dormance *versus* phase de réactivation) de la dormance. La période de transition entre la réactivation et l'activation est celle où l'on retrouve le plus faible nombre de gènes à expression différentielle (comparaison des cambiums en phase de réactivation *versus* phase d'activation).
- L'ampleur des changements du transcriptome dans le bois peut varier aussi selon l'âge de l'arbre. Ainsi, Li et coll. (2010), qui ont comparé le bois initial *versus* le bois final, ont identifié un plus grand nombre de gènes différentiels dans le bois de transition des arbres âgés de 9 ans (30 %) que dans les bois juvéniles (21 %; arbres âgés de 5 ans) ou matures (2,6 %; arbres âgés de 30 ans).

1.5.1.2 Dynamique du transcriptome en réponse aux stress environnementaux : cas du bois de compression

Les proportions de gènes différentiels identifiés dans des études où l'on a comparé le bois de compression *versus* le bois opposé ou bois normal sont présentées dans le tableau 1.1b. Toutefois, ces valeurs ne sont pas comparables entre elles à cause des variations techniques (critères de signification, nombre et nature des gènes testés) entre les études. La proportion des gènes différentiels a été évaluée à 7 % selon Villalobos et coll. (2012) qui ont analysé environ 6900 gènes. Une valeur semblable a été trouvée par Sato et coll. (2014) qui ont travaillé sur tous les gènes à l'échelle du génome (méthode de séquençage à ARN) avec un critère statistique moins contraignant que celui rétenu par Villalobos et coll. (2012). La proportion s'élève à 30 % d'après Li et coll. (2013) qui ont utilisé des puces avec des sondes (gènes) plus spécifiques au bois. Ces sondes ont en effet été développées à partir d'une banque d'ADNc du bois de pin de Monterey.

1.5.1.3 Fonctions des gènes surexprimés dans la formation du bois

Les analyses d'annotation fonctionnelle ont permis d'identifier la fonction et les processus biologiques auxquels sont associés les gènes surexprimés pendant la formation du bois;

- Les gènes impliqués dans la division cellulaire (p.ex. *cyclin*, *profiling-1* et *cyclin-like F-box*), l'expansion cellulaire (p.ex. *glucan endo-1,3-beta-D-glucosidase*, *expansins*, et *xyloglucan endotransglycolase*) et la transduction des messages cellulaires sont surexprimés pendant la formation du bois initial (Li et coll., 2010; Paiva et coll., 2008), pendant les périodes de réactivation et d'activation du cambium (Qiu et coll., 2013) et dans le bois de compression (Villalobos et coll. 2012; Li et coll., 2013). Des gènes responsables de la synthèse de la paroi primaire (p.ex. *pectinesterase*, *cellulase* et *pectate lyase*) peuvent être préférentiels aussi bien dans le bois initial (Li et coll., 2010) que dans le bois à faible rigidité (Li et coll., 2011).
- Les gènes impliqués dans la formation de la paroi secondaire des trachéides sont surexprimés dans le bois final (Yang et Loopstra, 2005; Li et coll., 2010), dans le bois à forte rigidité (Li et coll., 2011) et dans le bois de compression (Villalobos et coll. 2012; Li et coll., 2013). Parmi ces gènes, on peut citer des gènes responsables de la synthèse de la lignine (p.ex. 4CL, C3H, CAD, LAC et CCoAOMT), de la cellulose (p.ex. Cesa, SUS et *callose synthase-like*), de

l'hémicellulose (p.ex. protéines arabino-galactanes). Des gènes responsables de la formation de la paroi secondaire sont aussi surexprimés vers le début de la période de formation du bois d'après Mishima et coll. (2014) qui ont évalué l'expression des gènes dans le bois à quatre dates différentes d'une saison de croissance.

- Différents gènes impliqués dans la réponse au stress (p.ex. *dehydrin*, LEA et *desiccation-related protein*) sont surexprimés lors de l'entrée en dormance (Mishima et coll., 2014) et pendant la formation du bois final (Yang et Loopstra, 2005; Li et coll., 2010).
- Des gènes qui se rapportent à la formation du cytosquelette (p.ex. *actin related-protein*, *microtubule-associated protein* et *firbin profiling*) sont surexprimés pendant la formation du bois (Mishima et coll. 2014). Ils sont aussi préférentiels au bois initial (Li et coll. 2010), au bois final (Yang et Loopstra, 2005; Paiva et coll., 2008), au bois à forte rigidité (Li et coll., 2011) et au bois de compression (Villalobos et coll. 2012; Li et coll., 2013).
- Des facteurs de transcription qui sont reliés à la formation de la paroi secondaire (p.ex. *zinc finger*, R2R3-MYB, autres MYBs, NAC, AP2/EREBP, Lim domain, WRKY, bZIP) ont une expression différentielle entre le bois initial et le bois final (Paiva et coll., 2008; Li et coll., 2010; Qiu et coll., 2013, Mishima et coll., 2014). D'autres études ont également montré l'importance de certains MYBs dans la formation de la paroi secondaire chez les conifères. Ainsi, en se basant sur l'analyse des séquences de l'épinette blanche et du pin à encens, Bedon et coll. (2007) ont fait l'inventaire des MYBs dont certains, y compris MYB1 et MYB8, sont potentiellement impliqués dans la synthèse de la lignine et de la formation du bois. Bomal et coll. (2008) ont montré l'effet de la surexpression de MYB1 et de MYB8 qui augmente l'expression des gènes responsables de la formation de paroi secondaire. Des interactions entre ces deux MYBs et le promoteur de certains gènes de la paroi secondaire ont été validées grâce à un test fonctionnel (Duval et coll., 2014).

1.5.2 Stratégies d'adaptation

Les stratégies d'adaptation face aux perturbations abiotiques et biotiques ont aussi été l'objet de différentes analyses par profilage transcriptionnel chez les conifères tels que l'épinette de Sitka (Ralph et coll., 2006; Holliday et coll., 2008), l'épinette de Norvège

(Yakovlev et coll., 2014), l'épinette de l'intérieur (l'hybride *Picea glauca* x *P. engelmannii*; Verne et coll., 2011; Yeaman et coll., 2014), le pin tordu (Yeaman et coll., 2014), le pin argenté (Liu et coll., 2013), le pin sylvestre (Adomas et coll., 2008), le pin de Monterey (Dubouzet et coll., 2014) et le mélèze de Dahurie (Men et coll., 2013). Ces auteurs ont comparé des tissus et des organes d'arbres soumis à des facteurs climatiques, d'arbres non résistants *versus* résistants et d'arbres non traités *versus* traités aux divers facteurs biotiques. Les tissus et les organes analysés correspondent à ceux qui sont directement affectés par le stress causé par le facteur étudié. Il s'agit des aiguilles, des jeunes pousses terminales, des tissus embryonnaires, des écorces, des racines et des xylèmes. Des gènes différentiels identifiés dans ces travaux montrent la dynamique du transcriptome reliée aux mécanismes de défense et à la plasticité phénotypique des conifères (tableaux 1.2).

1.5.2.1 Dynamique du transcriptome en réponse aux facteurs biotiques et abiotiques

Les résultats des études sur la plasticité phénotypique des conifères aux facteurs abiotiques, dont la température, la photopériode, la disponibilité hydrique et la lumière, sont présentés dans le tableau 1.2a. La proportion des gènes dont l'expression change en réponse à l'un ou plusieurs de ces facteurs varie entre 1 et 49 % selon l'étude. Cette variation s'explique par la fluctuation des paramètres techniques, à savoir le nombre de gènes et de conditions analysés. La proportion maximale a été rapportée par Yeaman et coll. (2014) qui ont analysé l'expression de 23 889 gènes dans les aiguilles des pins tordus soumis à sept conditions différentes. Ces mêmes auteurs ont trouvé une proportion plus faible de gènes différentiels (27,3 %) chez l'épinette de l'intérieur à cause d'un nombre plus faible de répétitions de certaines conditions testées pour cette espèce. La plus faible proportion a été obtenue par Yakovlev et coll. (2014) qui ont travaillé sur des hypocotyles soumis seulement à deux conditions de lumière différentes, mais qui ont analysé un plus grand nombre de gènes que Yeaman et coll. (2014).

Les résultats obtenus par différentes études sur les mécanismes de défense face aux facteurs biotiques chez les conifères sont présentés dans le tableau 1.2b. Seulement trois de ces études ont analysé le mécanisme de défense constitutif. Les auteurs y ont rapporté une faible proportion (0,4 – 1 %) de gènes différentiels entre de jeunes pousses terminales pourvues d'écorce *versus* celles qui en étaient dépourvues chez l'épinette de Sitka (Friedmann et coll., 2007), entre les individus résistants *versus* ceux qui étaient non résistants au charançon du pin blanc chez l'épinette de l'intérieur

(Verne et coll. 2011) ou à la tordeuse des bourgeons de l'épinette (*Choristoneura fumiferana* [Clemens]) chez l'épinette blanche (Mageroy et coll., 2015). Les autres études ont porté sur le mécanisme de défense induit et ont rapporté une proportion de gènes différentiels variant de 0,5 à 36 % des gènes testés. La variation de la proportion des gènes différentiels entre ces études peut être associée aux variations techniques (seuil de significativité et spécificité des gènes analysés) et biologiques (nature du traitement). La proportion maximale a été rapportée par Ralph et coll. (2006). Ces derniers ont utilisé des critères de significativité moins restrictifs que les autres (Adomas et coll., 2008; Men et coll. 2013). Ils ont aussi analysé des gènes spécifiques à la défense induite en utilisant des puces renfermant des sondes (gènes) développées à partir d'une banque d'ADNc d'arbres soumis à des blessures d'origine mécanique et entomologique (Ralph et coll., 2008). Liu et coll. (2013), qui ont considéré les mêmes critères de significativité que Ralph et coll. (2006), ont rapporté des proportions plus faibles (1,2 à 4,3 %) sur un plus grand nombre de gènes testés (soit 2,4 fois plus élevé que dans Ralph et coll., 2006). Les plus faibles proportions (0,5 et 0,8 %) peuvent être attribuées à une variation biologique (la nature du traitement). Elles ont été obtenues à partir d'une comparaison de l'expression des gènes dans les racines d'arbres traités avec des champignons non pathogènes (saprophyte et symbiotique) après 15 jours d'inoculation (Adomas et coll. 2008).

L'examen de chacune des études a permis de constater l'existence d'un lien entre la dynamique du transcriptome et les variations biologiques reliées aux stratégies de réponse aux perturbations abiotiques et biotiques des conifères;

- Le stress causé par des facteurs abiotiques et biotiques augmente l'ampleur de la dynamique du transcriptome chez les conifères. Le nombre de gènes préférentiellement exprimés à la fin de l'automne est 1,3 fois plus élevé que le nombre de gènes fortement exprimés au début de l'automne dans les aiguilles de l'épinette de Sitka (Holliday et coll., 2008). On sait que les conditions climatiques sont plus contraignantes et causent un plus grand stress aux arbres à la fin qu'au début de l'automne. Adomas et coll. (2008) ont aussi obtenu un nombre de gènes différentiels de 18 à 29 fois plus élevé chez les arbres soumis au champignon pathogène que chez les arbres soumis aux champignons non pathogènes (saprophyte et symbiotique) après 15 jours d'inoculation. Des études sur les aiguilles du mélèze de Dahurie ont également mis en évidence

la surexpression de 63 à 75 % des gènes différentiels (contrôles *versus* non traités) dans les arbres traités avec du jasmonate de méthyle et de l'acide jasmonique (Men et coll., 2013). Chez les conifères, la réaction de défense induite par l'application exogène de jasmonate de méthyle est similaire à la réaction de défense induite par des blessures causées par des insectes (Moreira et coll., 2009).

- L'arbre a une réponse transcriptomique différente selon le type d'interaction avec les champignons auquel il est inoculé (Adomas et coll., 2008). La réponse transcriptomique se traduit par une plus grande proportion de gènes différentiels (contrôles *versus* inoculés) pour le champignon pathogène (13,9 %) que pour les champignons non pathogènes (0,5 – 0,8 %). En cas de stress, la réponse transcriptomique est différente selon les tissus d'un arbre. Dubouzet et coll. (2014) ont comparé des arbres traités et non traités avec l'éthéphon (une substance qui réduit la croissance en hauteur et en diamètre des conifères d'après Schnurr et coll., 1996). Ils ont constaté que la dynamique du transcriptome est plus importante dans le xylème mucilagineux proche du cambium (13,1 %) que dans le xylème fibreux (7,24 %) et que dans l'écorce (1 %).
- La dynamique du transcriptome devient plus importante avec la durée d'exposition au stress. La proportion de gènes différentiels passe de 1,6 % après une journée à 6,2 % après cinq jours et à 14 % après 15 jours d'inoculation du champignon pathogène (Adomas et coll., 2008). Dubouzet et coll. (2014), ont également découvert que la proportion de gènes différentiels augmentait avec le temps de traitement à l'éthéphon dans le xylème mucilagineux (de 0,4 % après une semaine à 13 % après huit semaines de traitement), dans le xylème fibreux (de 0,8 % après une semaine à 7 % après huit semaines de traitement) et dans l'écorce (de 0,07 % après une semaine à 1 % après huit semaines de traitement).
- Pour les arbres soumis au stress biotique, les arbres résistants ont une réaction transcriptomique plus importante que les arbres non résistants. Lui et coll. (2013) ont comparé le transcriptome des aiguilles d'arbres induits *versus* non induits au champignon responsable de la rouille vésiculeuse (*Cronartium ribicola* A. Dietr). Ils ont trouvé que les gènes différentiels sont 1,5 fois plus nombreux chez les arbres résistants que chez les arbres non résistants.

- L'étude comparative d'épinettes blanches résistantes *versus* non résistantes à la tordeuse des bourgeons de l'épinette a identifié un gène (*Pgβglu-1*) dont le différentiel d'expression était de 770 à 1000 fois supérieur chez les arbres résistants (Mageroy et coll., 2015). Cette différence est beaucoup plus grande que tous les autres gènes testés, soit près de 24 000 au total. Le gène identifié a été soumis à des tests fonctionnels et il a été montré qu'il était responsable de la synthèse de composés phénoliques présents uniquement chez les arbres résistants et qui diminue la survie des insectes (Delvas et coll., 2011).

1.5.2.2 Fonctions des gènes différentiels liés aux stratégies d'adaptation

Les analyses d'annotation fonctionnelle de gènes différentiels ont permis d'identifier des fonctions de gènes différentiels liés aux stratégies d'adaptation selon les conditions biologiques ou les facteurs environnementaux étudiés;

- Des gènes ont une expression préférentielle à certains facteurs qui causent le stress. Des gènes qui se rapportent à la régulation épigénétique (p.ex. *histone methyltransferase*, *histone acetyltransferase* et *bromodomain-containing proteins*) sont surexprimés dans des tissus embryonnaires soumis à une variation thermique (Yakovlev et coll., 2014). Bien que surexprimés dans l'écorce des arbres résistants au charançon du pin blanc (Verne et coll., 2011), certains gènes qui codent pour des protéines de chocs thermiques sont aussi induits en réponse aux variations des facteurs climatiques (Yakovlev et coll., 2014; Yeaman et coll., 2014). Un gène photorécepteur (*cryptochrome1*) est surexprimé dans des hypocotyles soumis à des lumières rouge clair et rouge sombre (Ranade et coll., 2013). Des gènes qui codent pour des enzymes LOX, AOS et AOC sont induits en réponse à une blessure mécanique, à l'attaque d'insectes et à l'application exogène de jasmonate de méthyle et d'acide jasmonique (Ralph et coll., 2006; Men et coll., 2013). LOX, AOS et AOC sont des enzymes impliqués dans la synthèse de l'acide jasmonique qui est une molécule de signal dans la défense.
- Les gènes impliqués dans la réponse au stress (p.ex. NBS-LRR, UBA, *chitinases* et PB1) et dans le métabolisme secondaire (exemples pour la synthèse des phenylpropanoïdes: PAL, C3H et F5H; exemples pour le métabolisme des terpénoïdes: *zeatin O-glucosyltransferase*, *isopentenyl diphosphate isomerase* et *GGPP synthase*) sont surexprimés dans les arbres résistants non soumis (Verne et coll. 2011) et soumis à des stress abiotiques (Holliday et coll., 2008; Ranade et

coll., 2013; Yakovlev et coll., 2014; Yeaman et coll., 2014) et biotiques (Adomas et coll., 2008; Ralph et coll., 2006; Men et coll., 2013; Liu et coll., 2013). Ils ont aussi une expression préférentielle dans l'écorce des arbres (Friedmann et coll., 2007).

- Certains facteurs de transcription sont préférentiellement exprimés dans les hypocotyles soumis à la lumière rouge sombre (Ranade et coll., 2013) et dans les arbres blessés (Ralph et coll., 2006; Liu et coll., 2013). Ces facteurs de transcription incluent ARF2, IAA, *zinc finger homeodomain*, NAC, *PLATZ-domain*, *dof-zinc finger*, *tubby-like F-box*, WRKY, *AP2-domain*, bHLH et bZIP.
- Le niveau d'expression des gènes impliqués dans le métabolisme primaire (p.ex. la ferrédoxine et les protéines de la glycolyse, de la synthèse du galactinol, du photosystème I et II, et de la liaison à la chlorophylle a et b) varie selon la nature de la relation que les arbres ont avec les facteurs biotiques. Ils sont sous-exprimés dans les arbres soumis à l'attaque d'insectes (Ralph et coll., 2006) et de champignon pathogène (Adomas et coll., 2008), mais ils sont surexprimés dans les arbres traités avec un champignon symbiotique. L'expression de ces gènes peut être différente selon le phénotype. Par exemple, des gènes du métabolisme primaire ont présenté un niveau d'expression relativement plus élevé chez les individus résistants que chez les non résistants, les deux ont été soumis au charançon responsable de la rouille vésiculeuse (Liu et coll., 2013).

1.5.3 Autres études chez des conifères

Aucune étude préalable ne s'est penchée sur l'étude du profil d'expression des gènes sur un grand nombre de tissus. Quelques-unes ont abordé la conservation du profil d'expression (Pavy et coll., 2008; Yeaman et coll., 2014). Pavy et coll. (2008) ont étudié l'expression des gènes dans les aiguilles, le phloème et le xylème secondaire afin d'identifier des gènes candidats impliqués dans la formation du bois. Sur les 10 400 séquences testées, 3,5 % sont préférentielles au xylème et 0,3 % ont des séquences et des profils d'expression préférentiels au xylème conservé entre des gymnospermes et des angiospermes. Entre les conifères, la proportion des gènes ayant une séquence et un profil d'expression conservés semble être plus élevée. Yeaman et coll. (2014) ont découvert que 74 % des gènes orthologues ont un profil d'expression conservé entre deux conifères: pin tordu et épinette de l'intérieur. Cette proportion est assez élevée si l'on se réfère au nombre de traitements (sept traitements *versus* trois



traitements ou tissus dans Pavy et coll., 2008) sur lesquels se sont basés les profils d'expression étudiés (Yeaman et coll., 2014).

1.6 Mise en contexte du projet

Cette thèse s'inscrit dans le cadre du projet SMarTForests qui exploite la génomique dans le but de développer des marqueurs génétiques diagnostiques pour des programmes d'amélioration des conifères. Les marqueurs serviront à identifier les arbres et les semis démontrant une meilleure croissance, de meilleures propriétés du bois et une forte résistance aux insectes. Un marqueur diagnostique est un gène ou une séquence d'ADN polymorphe qui est génétiquement lié à un phénotype d'intérêt. Pour être utiles en amélioration génétique, ces marqueurs doivent avoir une valeur prédictive du phénotype et de la valeur génétique des arbres. Étant donné la très grande taille des génomes des conifères (p.ex. Nystedt et coll., 2013; Warren et coll., inédit) et le faible déséquilibre de liaison dans leurs populations (Namroud et coll., 2010), plusieurs études réalisées chez les conifères s'appuient sur une approche par gènes candidats (p.ex. Dillon et coll., 2010; Beaulieu et coll., 2011; Eckert et coll., 2012; Prunier et coll., 2013). Cette approche cible les gènes pour lesquels il existe des indications préalables d'une implication potentielle dans la mise en place du phénotype. Selon Pflieger et coll. (2001), un bon gène candidat est testé selon au moins trois approches: étude d'expression, étude de polymorphisme et étude de fonction. Dans cette thèse, nous nous sommes intéressés à étudier l'expression de gènes à l'échelle du génome (profilage transcriptionnel), ce qui permet entre autres de développer des connaissances fondamentales nécessaires à l'identification des gènes candidats chez les conifères. Ces connaissances ont été jumelées aux informations disponibles sur les séquences géniques et contribuent au développement de marqueurs génétiques et à la découverte des bases moléculaires de plusieurs types de caractères (p.ex. Rutledge et coll., 2013; Verta et coll., 2013; Mageroy et coll., 2015).

Les recherches sur les conifères sont justifiées autant par leur importance économique que leurs rôles écologiques. À cause de leur grande taille et de leur complexité, les génomes de conifères n'ont été disponibles qu'à partir de 2013 (Birol et coll., 2013; Nystedt et coll., 2013; Neale et coll., 2014). Pour étudier les gènes à l'échelle du génome, les chercheurs ont depuis longtemps opté pour la méthode transcriptomique (Allona et coll., 1998). Ils ont fixé des objectifs spécifiques dans lesquels ils répondaient à des questions biologiques en lien avec la formation du bois (p.ex.

Mishima et coll., 2014; Sato et coll., 2014), avec les mécanismes de défense (Verne et coll., 2011; Men et coll., 2013) et avec l'adaptation aux conditions climatiques des conifères (Ranade et coll., 2013; Yeaman et coll., 2014). Dans l'ensemble, leurs travaux ne permettaient pas d'établir le profil d'expression des gènes regroupant les différents tissus et organes des conifères, car ils ne considéraient qu'un seul type de tissus dans la plupart des études, et au plus trois différents types de tissus. La variation du transcriptome reliée à la différenciation tissulaire n'a donc pas été abordée. Pourtant, des études chez des angiospermes ont mis en évidence la grande part de la différenciation tissulaire dans la dynamique du transcriptome par rapport à d'autres facteurs (Ma et coll., 2005; Downs et coll., 2013). Les études faites chez les conifères ont principalement identifié et analysé les gènes différentiels entre différents stades de développement ou conditions biologiques, ce qui a permis de mettre en évidence la dynamique du transcriptome et d'identifier des gènes candidats selon l'expression. Une autre déficience dans nos connaissances actuelles sur les conifères est représentée par les gènes invariables; malgré leur rôle potentiel dans le maintien des mécanismes cellulaires de base (Eisenberg et Levanon, 2013). Enfin, très peu d'études ont examiné l'évolution de la dynamique du transcriptome en comparant différentes espèces; les seules exceptions sont les travaux décrits par Pavy et coll. (2008) et Yeaman et coll. (2014). Des études comparant plusieurs tissus ou considérant plusieurs espèces permettraient de créer une assise sur laquelle appuyer les connaissances déjà acquises et les développements futurs.

1.7 Objectifs

Pour cette thèse, nous nous sommes fixés des objectifs à portée fondamentale qui ont pour objet l'organisation, l'évolution et la dynamique du transcriptome chez les conifères. Par le biais du profilage transcriptionnel sur plusieurs tissus, nous avons analysé l'ensemble des gènes exprimés, autant les gènes invariables que les gènes variables du transcriptome. Nous avons aussi étudié la variation spatiale (xylème secondaire et phelloderme) du transcriptome pour trois espèces d'épinette (épinette blanche, épinette de Sitka, épinette noire) et la variation temporelle pour deux espèces d'épinette (épinette blanche et épinette de Norvège). Nous avons répondu à des objectifs généraux et spécifiques dans les chapitres 2 et 3.

Le chapitre 2 présente d'abord le développement de la première puce à oligonucléotides pour l'étude des profils transcriptomiques chez les conifères, ainsi

qu'une analyse de la conservation et la diversité du profil d'expression chez les conifères. Il comporte trois objectifs.

- **Objectif 2.1** Montrer l'utilité de la puce à oligonucléotide pour obtenir des profils d'expression représentatifs du transcriptome. Nous présentons la première puce à oligonucléotide développée pour les conifères afin de permettre des études plus précises. Cette puce contient les séquences de près de 24 000 gènes distincts de l'épinette blanche et présente la plus large couverture du génome des conifères à ce jour. L'avantage des oligonucléotides est qu'ils sont fortement spécifiques et qu'ils permettent de limiter l'hybridation croisée. Toutefois, l'utilisation de sondes suffisamment longues permet de tolérer quelques polymorphismes et de faire des comparaisons interspécifiques.
- **Objectif 2.2** Mettre en évidence et caractériser la conservation des profils associés aux préférences tissulaires chez les conifères. Des gènes dont l'expression est différentielle entre le xylème secondaire et le phelloderme ont été identifiés et la conservation de leur profil d'expression entre les espèces d'épinette (épinette blanche, épinette de Sitka, épinette noire) a été étudiée.
- **Objectif 2.3** Dresser un portrait de la variabilité de l'expression entre les tissus et les gènes afin de dégager les tendances évolutives qui façonnent leur expression. Pour cet objectif, nous avons construit une base de données appelée PiceaGenExpress à partir des données d'expression provenant de diverses expériences. Pour permettre l'intégration de ces résultats, les données quantitatives d'expression ont été transformées en rangs ou en variables quantitatives discontinues.

Le chapitre 3 avait pour but de dresser un portrait de la dynamique du transcriptome et de construire des réseaux d'expression. Le chapitre 3 comporte deux objectifs.

- **Objectif 3.1** Étudier l'organisation du transcriptome des conifères par le biais des analyses des données d'expression du transcrit. Nous visions notamment à grouper les gènes selon différents profils de variations tissulaires (groupes de coexpression tissulaire) et à déterminer la classification biologique des gènes appartenant à chaque groupe dans le but de montrer les relations fonctionnelles qui existent à l'intérieur et entre les cohortes d'expression. Cet

objectif comporte aussi un regard évolutif au point de vue morphologique (acquisition des différents tissus) et phylogénétique (comparant les conifères et autres espèces végétales).

- **Objectif 3.2** Construire des réseaux transcriptionnels de la formation du bois (xylème secondaire) d'après une analyse de coexpression qui montre les relations entre les gènes.

Dans le cadre de cette thèse, des comparaisons de l'expression génique entre plusieurs types de tissus ont été réalisées pour l'objectif 2.3 et l'objectif 3.1. Ces comparaisons ont été faites avec des objectifs distincts, et ont suivi soit une approche semi-quantitative ou quantitative, respectivement. La conclusion générale de la thèse établira une comparaison des deux approches, mettra en lumière leurs buts respectifs et raisons d'être distinctes, ainsi que les retombées de chacune pour le développement des connaissances et leurs possibilités d'utilisation lors du développement de marqueurs pour la sélection génétique.

1.8 Références bibliographiques

- Adomas A, Heller G, Olson Å, Osborne J, Karlsson M, Nahalkova J, Van Zyl L, Sederoff R, Stenlid J, Finlay R. 2008.** Comparative analysis of transcript abundance in *Pinus sylvestris* after challenge with a saprotrophic, pathogenic or mutualistic fungus. *Tree Physiology* **28**(6): 885-897.
- Alberto FJ, Aitken SN, Alía R, González-Martínez SC, Hänninen H, Kremer A, Lefèvre F, Lenormand T, Yeaman S, Whetten R. 2013.** Potential for evolutionary responses to climate change—evidence from tree populations. *Global Change Biology* **19**(6): 1645-1661.
- Allocco DJ, Kohane IS, Butte AJ. 2004.** Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* **5**(1): 18.
- Allona I, Quinn M, Shoop E, Swope K, Cyr SS, Carlis J, Riedl J, Retzel E, Campbell MM, Sederoff R. 1998.** Analysis of xylem formation in pine by cDNA sequencing. *Proceedings Of The National Academy Of Sciences* **95**(16): 9693-9698.
- Anielski M, Wilson S. 2009.** *Counting Canada's Natural Capital : Assessing the Real Value of Canada's Boreal 26 Ecosystems*. Ottawa, Ontario: The Canadian Boreal Initiative, The Pembina Institute. 76p.
- Barnett J, Jeronimidis G. 2003.** *Wood quality and its biological basis*: Boca Raton, USA: CRC Press LLC.

- Beaulieu J, Doerksen T, Boyle B, Clément S, Deslauriers M, Beauseigle S, Blais S, Poulin P-L, Lenz P, Caron S. 2011.** Association genetics of wood physical traits in the conifer white spruce and relationships with gene expression. *Genetics* **188**(1): 197-214.
- Beaulieu J, Rainville A, Daoust G, Bousquet J. 2001.** La diversité génétique des espèces arborescentes de la forêt boréale. *Le Naturaliste Canadien* **125**: 193-202.
- Bedon F, Grima-Pettenati J, Mackay J. 2007.** Conifer R2R3-MYB transcription factors: sequence analyses and gene expression in wood-forming tissues of white spruce (*Picea glauca*). *BMC Plant Biology* **7**(1): 17.
- Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Saint Yuen MM, Keeling CI, Brand D, Vandervalk BP. 2013.** Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics* **29**(12): 1492-1497.
- Bomal C, Bedon F, Caron S, Mansfield SD, Levasseur C, Cooke JE, Blais S, Tremblay L, Morency M-J, Pavy N. 2008.** Involvement of *Pinus taeda* MYB1 and MYB8 in phenylpropanoid metabolism and secondary cell wall biogenesis: a comparative in planta analysis. *Journal Of Experimental Botany* **59**(14): 3925-3939.
- Brandt J, Flannigan M, Maynard D, Thompson I, Volney W. 2013.** An introduction to Canada's boreal zone: ecosystem processes, health, sustainability, and environmental issues 1. *Environmental Reviews* **21**(4): 207-226.
- Burdon RD, Kibblewhite RP, Walker JC, Megraw RA, Evans R, Cown DJ. 2004.** Juvenile versus mature wood: a new concept, orthogonal to corewood versus outerwood, with special reference to *Pinus radiata* and *P. taeda*. *Forest Science* **50**(4): 399-415.
- Byun-McKay A, Godard K-A, Toudefallah M, Martin DM, Alfaro R, King J, Bohlmann J, Plant AL. 2006.** Wound-induced terpene synthase gene expression in Sitka spruce that exhibit resistance or susceptibility to attack by the white pine weevil. *Plant Physiology* **140**(3): 1009-1021.
- Canales J, Bautista R, Label P, Gómez-Maldonado J, Lesur I, Fernández-Pozo N, Rueda-López M, Guerrero-Fernández D, Castro-Rodríguez V, Benzekri H. 2014.** *De novo* assembly of maritime pine transcriptome: implications for forest breeding and biotechnology. *Plant Biotechnology Journal* **12**(3): 286-299.
- Carlson M, Wells J, Roberts D. 2009.** *The carbon the world forgot: conserving the capacity of Canada's Boreal Forest region to mitigate and adapt to climate change*. Seattle, USA and Ottawa, Canada: Boreal Songbird Initiative and Canadian Boreal Initiative.
- Chou C-C, Chen C-H, Lee T-T, Peck K. 2004.** Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Research* **32**(12): e99-e99.
- Christenhusz M, Reveal J, Farjon A, Gardner MF, Mill RR, Chase MW. 2011.** A new classification and linear sequence of extant gymnosperms. *Phytotaxa* **19**:55-70

- COSEWIC. 2013.** *Canadian Wildlife Species at Risk. Committee on the Status of Endangered Wildlife in Canada.* Web site: http://www.cosewic.gc.ca/eng/sct0/rpt/rpt_csar_e.cfm [accessed 21 November 2013]
- Delucia EH, Maherali H, Carey EV. 2000.** Climate-driven changes in biomass allocation in pines. *Global Change Biology* **6**(5): 587-593.
- Delvas N, Bauce É, Labbé C, Ollevier T, Bélanger R. 2011.** Phenolic compounds that confer resistance to spruce budworm. *Entomologia Experimentalis Et Applicata* **141**(1): 35-44.
- Dillon SK, Nolan M, Li W, Bell C, Wu HX, Southerton SG. 2010.** Allelic variation in cell wall candidate genes affecting solid wood properties in natural populations and land races of *Pinus radiata*. *Genetics* **185**(4): 1477-1487.
- Donaldson LA. 2001.** Lignification and lignin topochemistry—an ultrastructural view. *Phytochemistry* **57**(6): 859-873.
- Donaldson LA, Grace J, Downes GM. 2004.** Within-tree variation in anatomical properties of compression wood in radiata pine. *IAWA journal* **25**(3): 253-271.
- Downs GS, Bi Y-M, Colasanti J, Wu W, Chen X, Zhu T, Rothstein SJ, Lukens L. 2013.** A developmental transcriptional network for *Zea mays* defines coexpression modules. *Plant Physiology* **161**(4): 1830-1843.
- Dubouzet JG, Donaldson L, Black MA, McNoe L, Liu V, Lloyd-Jones G. 2014.** Heterologous hybridisation to a *Pinus* microarray: profiling of gene expression in *Pinus radiata* saplings exposed to ethephon. *New Zealand Journal of Forestry Science* **44**(1): 21.
- Duval I, Lachance D, Giguère I, Bomal C, Morency M-J, Pelletier G, Boyle B, MacKay JJ, Séguin A. 2014.** Large-scale screening of transcription factor–promoter interactions in spruce reveals a transcriptional network involved in vascular development. *Journal Of Experimental Botany* **65**(9): 2319-2333.
- Eckert AJ, Wegrzyn JL, Cumbie WP, Goldfarb B, Huber DA, Tolstikov V, Fiehn O, Neale DB. 2012.** Association genetics of the loblolly pine (*Pinus taeda*, Pinaceae) metabolome. *New Phytologist* **193**(4): 890-902.
- Eisenberg E, Levanon EY. 2013.** Human housekeeping genes, revisited. *Trends In Genetics* **29**(10): 569-574.
- Farjon A. 2001.** *World checklist and bibliography of conifers*. Texas, USA: Forester-Artist.Com.
- Franks SJ, Weber JJ, Aitken SN. 2014.** Evolutionary and plastic responses to climate change in terrestrial plant populations. *Evolutionary Applications* **7**(1): 123-139.
- Friedmann M, Ralph SG, Aeschliman D, Zhuang J, Ritland K, Ellis BE, Bohlmann J, Douglas CJ. 2007.** Microarray gene expression profiling of developmental transitions in Sitka spruce (*Picea sitchensis*) apical shoots. *Journal Of Experimental Botany* **58**(3): 593-614.
- Garbaye J. 2007.** Evolution et adaptation des plantes aux conditions hydriques limitantes. *L'eau en horticulture: économiser maintenant*. 9e colloque du conseil scientifique de la SNHF,

http://www.jejardine.org/images/stories/2_La_plante_et_son_milieu/adaptation_plantes_garbaye.pdf.

- GI. 2000.** Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**(6814): 796.
- Goff SA, Rieke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H. 2002.** A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**(5565): 92-100.
- Guo H, Ecker JR. 2004.** The ethylene signaling pathway: new insights. *Current Opinion In Plant Biology* **7**(1): 40-49.
- Hall DE, Robert JA, Keeling CI, Domanski D, Quesada AL, Jancsik S, Kuzyk MA, Hamberger B, Borchers CH, Bohlmann J. 2011.** An integrated genomic, proteomic and biochemical analysis of (+)-3-carene biosynthesis in Sitka spruce (*Picea sitchensis*) genotypes that are resistant or susceptible to white pine weevil. *The Plant Journal* **65**(6): 936-948.
- Hamrick JL, Godt MJW, Sherman-Broyles SL. 1992.** Factors influencing levels of genetic diversity in woody plant species. *New Forests* **6**: 95-124
- Hertzberg M, Aspeborg H, Schrader J, Andersson A, Erlandsson R, Blomqvist K, Bhalerao R, Uhlén M, Teeri TT, Lundeberg J. 2001.** A transcriptional roadmap to wood formation. *Proceedings of the National Academy of Sciences* **98**(25): 14732-14737.
- Holliday JA, Ralph SG, White R, Bohlmann J, Aitken SN. 2008.** Global monitoring of autumn gene expression within and among phenotypically divergent populations of Sitka spruce (*Picea sitchensis*). *New Phytologist* **178**(1): 103-122.
- IHGSC 2004.** Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011): 931-945.
- IRGS. 2005.** The map-based sequence of the rice genome. *Nature* **436**(7052): 793-800.
- Jozsa LA, Middleton GR. 1994.** *A discussion of wood quality attributes and their practical implications*. Special Publication N° SP-34, Forintek Canada Corporation, Vancouver, BC, 42 p
- Kenrick P, Crane PR. 1997.** The origin and early evolution of plants on land. *Nature* **389**(6646): 33-39.
- King JN, Alfaro RI, Lopez MG, Van Akker L. 2011.** Resistance of Sitka spruce (*Picea sitchensis* (Bong.) Carr.) to white pine weevil (*Pissodes strobi* Peck): characterizing the bark defence mechanisms of resistant populations. *Forestry* **84**(1): 83-91.
- Lambert M, Ung C, Raulier F. 2005.** Canadian national tree aboveground biomass equations. *Canadian Journal Of Forest Research* **35**(8): 1996-2018.
- Larson PR. 1969.** Wood formation and the concept of wood quality. *Yale Univ. School Forestry, Bull*(74): 1-54.
- Ledig FT. 2012.** Climate Change and Conservation. *Acta Silvatica et Lignaria Hungarica* **8**(1): 57-74.

- Leslie AB, Beaulieu JM, Rai HS, Crane PR, Donoghue MJ, Mathews S. 2012.** Hemisphere-scale differences in conifer evolutionary dynamics. *Proceedings of the National Academy of Sciences* **109**(40): 16217-16221.
- Li X, Wu HX, Southerton SG. 2010.** Seasonal reorganization of the xylem transcriptome at different tree ages reveals novel insights into wood formation in *Pinus radiata*. *New Phytologist* **187**(3): 764-776.
- Li X, Wu HX, Southerton SG. 2011.** Transcriptome profiling of *Pinus radiata* juvenile wood with contrasting stiffness identifies putative candidate genes involved in microfibril orientation and cell wall mechanics. *BMC Genomics* **12**(1): 480.
- Li X, Yang X, Wu HX. 2013.** Transcriptome profiling of radiata pine branches reveals new insights into reaction wood formation with implications in plant gravitropism. *BMC Genomics* **14**(1): 768.
- Liu J-J, Sturrock RN, Benton R. 2013.** Transcriptome analysis of *Pinus monticola* primary needles by RNA-seq provides novel insight into host resistance to *Cronartium ribicola*. *BMC Genomics* **14**(1): 884.
- Lu Y, Ran J-H, Guo D-M, Yang Z-Y, Wang X-Q. 2014.** Phylogeny and Divergence Times of Gymnosperms Inferred from Single-Copy Nuclear Genes. *PloS One* **9**(9): e107679.
- Ma L, Sun N, Liu X, Jiao Y, Zhao H, Deng XW. 2005.** Organ-specific expression of *Arabidopsis* genome during development. *Plant Physiology* **138**(1): 80-91.
- Mageroy MH, Parent G, Germanos G, Giguère I, Delvas N, Maaroufi H, Baucé É, Bohlmann J, Mackay JJ. 2015.** Expression of the β -glucosidase gene Pg β glu-1 underpins natural resistance of white spruce against spruce budworm. *The Plant Journal* **81**(1): 68-80.
- Mann IK, Wegrzyn JL, Rajora OP. 2013.** Generation, functional annotation and comparative analysis of black spruce (*Picea mariana*) ESTs: an important conifer genomic resource. *BMC Genomics* **14**(1): 702.
- Mansfield SD, Parish R, Di Lucca CM, Goudie J, Kang K-Y, Ott P. 2009.** Revisiting the transition between juvenile and mature wood: a comparison of fibre length, microfibril angle and relative wood density in lodgepole pine. *Holzforschung* **63**(4): 449-456.
- Men L, Yan S, Liu G. 2013.** De novo characterization of *Larix gmelinii* (Rupr.) Rupr. transcriptome and analysis of its gene expression induced by jasmonates. *BMC Genomics* **14**(1): 548.
- Michael TP, Jackson S. 2013.** The first 50 plant genomes. *The Plant Genome* **6**(2): 1-7.
- Mishima K, Fujiwara T, Iki T, Kuroda K, Yamashita K, Tamura M, Fujisawa Y, Watanabe A. 2014.** Transcriptome sequencing and profiling of expressed genes in cambial zone and differentiating xylem of Japanese cedar (*Cryptomeria japonica*). *BMC Genomics* **15**(1): 219.
- Moreira X, Sampedro L, Zas R. 2009.** Defensive responses of *Pinus pinaster* seedlings to exogenous application of methyl jasmonate: concentration effect and systemic response. *Environmental And Experimental Botany* **67**(1): 94-100.

- Morgante M, De Paoli E. 2011. Toward the conifer genome sequence.** In *Genetics, Genomics and Breeding of Conifer Trees*. In: Plomion C, Bousquet J, Kole C, eds. *Genetics, Genomics and Breeding of Conifer Trees*. New Hampshire, UK: Science Publishers and CRC Press, 389-403.
- Mullin TJ, Andersson B, Bastien J-C, Beaulieu J., Burdon RD, Dvorak WS, King JN, Kondo T, Krakowski J, Lee SJ, McKeand SE, Pâques L, Raffin A, Russell JH, Sroppa T, Stoeck M, Yanchuk A. 2011.** Economic importance, breeding objectives and achievements. In: Plomion C, Bousquet J, Kole C, eds. *Genetics, Genomics and Breeding of Conifer Trees*. New Hampshire, UK: Science Publishers and CRC Press, 40-127.
- Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D. 2014.** The genome of *Eucalyptus grandis*. *Nature* **510**(7505): 356-362.
- Namroud M-C, Guillet-Claude C, Mackay J, Isabel N, Bousquet J. 2010.** Molecular evolution of regulatory genes in spruces from different species and continents: heterogeneous patterns of linkage disequilibrium and selection but correlated recent demographic changes. *Journal Of Molecular Evolution* **70**(4): 371-386.
- Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD. 2014.** Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology* **15**(3): R59.
- Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A. 2013.** The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**(7451): 579-584.
- Paiva J, Garnier-Géré P, Rodrigues J, Alves A, Santos S, Graça J, Le Provost G, Chaumeil P, Silva-Perez D, Bosc A. 2008.** Plasticity of maritime pine (*Pinus pinaster*) wood-forming tissues during a growing season. *New Phytologist* **179**(4): 1180-1194.
- Pavy N, Boyle B, Nelson C, Paule C, Giguère I, Caron S, Parsons LS, Dallaire N, Bedon F, Bérubé H. 2008.** Identification of conserved core xylem gene sets: conifer cDNA microarray development, transcript profiling and computational analyses. *New Phytologist* **180**(4): 766-786.
- Pflieger S, Lefebvre V, Causse M. 2001.** The candidate gene approach in plant genetics: a review. *Molecular Breeding* **7**(4): 275-291.
- Pinosio S, González-Martínez S, Bagnoli F, Cattonaro F, Grivet D, Marroni F, Lorenzo Z, Pausas J, Verdú M, Vendramin G. 2014.** First insights into the transcriptome and development of new genomic tools of a widespread circum-Mediterranean tree species, *Pinus halepensis* Mill. *Molecular Ecology Resources* **14**(4): 846-856.
- Porté A, Trichet P, Bert D, Loustau D. 2002.** Allometric relationships for branch and tree woody biomass of Maritime pine (*Pinus pinaster* Ait.). *Forest Ecology And Management* **158**(1): 71-83.
- Porth I, Klápště J, Skyba O, Friedmann MC, Hannemann J, Ehrling J, El-Kassaby YA, Mansfield SD, Douglas CJ. 2013.** Network analysis reveals

the relationship among wood properties, gene expression levels and genotypes of natural *Populus trichocarpa* accessions. *New Phytologist* **200**(3): 727-742.

- Prasad A, Kumar SS, Dessimoz C, Bleuler S, Laule O, Hruz T, Gruissem W, Zimmermann P. 2013.** Global regulatory architecture of human, mouse and rat tissue transcriptomes. *BMC Genomics* **14**(1): 716.
- Prunier J, Pelgas B, Gagnon F, Despons M, Isabel N, Beaulieu J, Bousquet J. 2013.** The genomic architecture and association genetics of adaptive characters using a candidate SNP approach in boreal black spruce. *BMC Genomics* **14**(1): 368.
- Pullat J, Fleischer R, Becker N, Beier M, Metspalu A, Hoheisel JD. 2007.** Optimization of candidate-gene SNP-genotyping by flexible oligonucleotide microarrays; analyzing variations in immune regulator genes of hay-fever samples. *BMC Genomics* **8**(1): 282.
- Qiu Z, Wan L, Chen T, Wan Y, He X, Lu S, Wang Y, Lin J. 2013.** The regulation of cambial activity in Chinese fir (*Cunninghamia lanceolata*) involves extensive transcriptome remodeling. *New Phytologist* **199**(3): 708-719.
- Ralph SG, Chun HJ, Kolosova N, Cooper D, Oddy C, Ritland CE, Kirkpatrick R, Moore R, Barber S, Holt RA. 2008.** A conifer genomics resource of 200,000 spruce (*Picea* spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (*Picea sitchensis*). *BMC Genomics* **9**(1): 484.
- Ralph SG, Yueh H, Friedmann M, Aeschliman D, Zeznik JA, Nelson CC, Butterfield YS, Kirkpatrick R, Liu J, Jones SJ. 2006.** Conifer defence against insects: microarray gene expression profiling of Sitka spruce (*Picea sitchensis*) induced by mechanical wounding or feeding by spruce budworms (*Choristoneura occidentalis*) or white pine weevils (*Pissodes strobi*) reveals large-scale changes of the host transcriptome. *Plant, Cell & Environment* **29**(8): 1545-1570.
- Ranade SS, Abrahamsson S, Niemi J, García-Gil MR. 2013.** *Pinus taeda* cDNA Microarray as a Tool for Candidate Gene Identification for Local Red/Far-Red Light Adaptive Response in *Pinus sylvestris*. *American Journal Of Plant Sciences* **4**: 479-493.
- Richter S, Kipfer T, Wohlgemuth T, Guerrero CC, Ghazoul J, Moser B. 2012.** Phenotypic plasticity facilitates resistance to climate change in a highly variable environment. *Oecologia* **169**(1): 269-279.
- Rigault P, Boyle B, Lepage P, Cooke JE, Bousquet J, MacKay JJ. 2011.** A white spruce gene catalog for conifer genome analyses. *Plant Physiology* **157**(1): 14-28.
- Robert JA, Madilao LL, White R, Yanchuk A, King J, Bohlmann J. 2010.** Terpenoid metabolite profiling in Sitka spruce identifies association of dehydroabietic acid, (+)-3-carene, and terpinolene with resistance against white pine weevil. *Botany* **88**(9): 810-820.
- Roff DA, Bradford MJ. 2000.** A quantitative genetic analysis of phenotypic plasticity of diapause induction in the cricket *Allonemobius socius*. *Heredity* **84**(2): 193-200.



- Rutledge RG, Stewart D, Caron S, Overton C, Boyle B, MacKay J, Klimaszcwska K. 2013.** Potential link between biotic defense activation and recalcitrance to induction of somatic embryogenesis in shoot primordia from adult trees of white spruce (*Picea glauca*). *BMC Plant Biology* **13**(1): 116.
- Sato S, Yoshida M, Hiraide H, Ihara K, Yamamoto H. 2014.** Transcriptome Analysis of Reaction Wood in Gymnosperms by Next-Generation Sequencing. *American Journal Of Plant Sciences* **5**, 2785-2798.
- Schiff C, Wilson I, Somerville S. 2001.** Polygenic powdery mildew disease resistance in *Arabidopsis thaliana*: quantitative trait analysis of the accession Warschau-1. *Plant Pathology* **50**(6): 690-701.
- Schnurr JP, Cheng Z, Boe AA. 1996.** Effects of plant growth regulators on sturdiness of Jack pine seedlings. *Journal Of Environmental Horticulture* **14**: 228-230.
- Sena JS, Giguère I, Boyle B, Rigault P, Birol I, Zuccolo A, Ritland K, Ritland C, Bohlmann J, Jones S. 2014.** Evolution of gene structure in the conifer *Picea glauca*: a comparative analysis of the impact of intron size. *BMC Plant Biology* **14**(1): 95.
- Sjostrom E, 1993.** *Wood Chemistry*. Academic Press, San Diego.
- St-Clair JB, Howe GT. 2011.** Strategies for conserving forest genetic resources in the face of climate change. *Turkish Journal Of Botany* **35**(4): 403-409.
- Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A. 2006.** The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**(5793): 1596-1604.
- Ugla C, Magel E, Moritz T, Sundberg B. 2001.** Function and dynamics of auxin and carbohydrates during earlywood/latewood transition in Scots pine. *Plant Physiology* **125**(4): 2029-2039.
- UNECE. 2013.** Forest product statistics 2007-2011. *Timber Bulletin ECE/TIM/BULL/65/2, Issue 2012, United Nations Economic Commission for Europe (UNECE)*, <http://www.unece.org/fileadmin/DAM/timber/statsdata/fps07-11.pdf>.
- Ung C-H, Bernier P, Guo X-J. 2008.** Canadian national biomass equations: new parameter estimates that include British Columbia data. *Canadian Journal Of Forest Research* **38**(5): 1123-1132.
- Verne S, Jaquish B, White R, Ritland C, Ritland K. 2011.** Global transcriptome analysis of constitutive resistance to the white pine weevil in spruce. *Genome Biology And Evolution* **3**: 851-867.
- Verta JP, Landry CR, MacKay JJ. 2013.** Are long-lived trees poised for evolutionary change? Single locus effects in the evolution of gene expression networks in spruce. *Molecular Ecology* **22**(9): 2369-2379.
- Villalobos DP, Díaz-Moreno SM, Said E-SS, Cañas RA, Osuna D, Van Kerckhoven SH, Bautista R, Claros MG, Cánovas FM, Cantón FR. 2012.** Reprogramming of gene expression during compression wood formation in pine: coordinated modulation of S-adenosylmethionine, lignin and lignan related genes. *BMC Plant Biology* **12**(1): 100.

- Wanfeng L, Suying H, Liwang Q. 2014.** Transcriptome resources and genome-wide marker development for Japanese larch (*Larix kaempferi*). *Frontiers of Agricultural Science And Engineering* **1**(1): 77-84.
- Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, Fang H, Hong H, Shen J, Su Z. 2014.** The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nature Biotechnology* **32**(9): 926-932.
- Warren R, Keeling C, Yuen M, Raymond A, Taylor G, Vandervalk B, Paulion D, Robertson G, Yang C, Hoffmann M, Weigel D, Ritland C, Isabel N, Jaquish B, Yanchuk A, Bousequet J, Jones S, Birol I, Bohlmann J. (inédit).** Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer defense metabolism. *The Plant Journal*.
- Wasternack C. 2007.** Jasmonates: an update on biosynthesis, signal transduction and action in plant stress response, growth and development. *Annals of Botany* **100**(4): 681-697.
- Wegrzyn JL, Lin BY, Zieve JJ, Dougherty WM, Martínez-García PJ, Koriabine M, Holtz-Morris A, Crepeau M, Langley CH, Puiu D. 2013.** Insights into the loblolly pine genome: characterization of BAC and fosmid sequences. *PloS One* **8**(9): e72439.
- Wells JV, Blancher P. 2011.** Global role for sustaining bird populations. In: Wells JV, ed. *Boreal Birds of North America*. Los Angeles, USA: University of California Press, 7-23.
- Yakovlev IA, Lee Y, Rotter B, Olsen JE, Skråppa T, Johnsen Ø, Fossdal CG. 2014.** Temperature-dependent differential transcriptomes during formation of an epigenetic memory in Norway spruce embryogenesis. *Tree Genetics & Genomes* **10**(2): 355-366.
- Yamashita S, Yoshida M, Takayama S, Okuyama T. 2007.** Stem-righting mechanism in gymnosperm trees deduced from limitations in compression wood development. *Annals Of Botany* **99**(3): 487-493.
- Yang S-H, Loopstra CA. 2005.** Seasonal variation in gene expression for loblolly pines (*Pinus taeda*) from different geographical regions. *Tree Physiology* **25**(8): 1063-1073.
- Yeaman S, Hodgins KA, Suren H, Nurkowski KA, Rieseberg LH, Holliday JA, Aitken SN. 2014.** 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* **203**(2): 578-591.
- Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X. 2002.** A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**(5565): 79-92.
- Zeneli G, Krokene P, Christiansen E, Krekling T, Gershenzon J. 2006.** Methyl jasmonate treatment of mature Norway spruce (*Picea abies*) trees increases the accumulation of terpenoid resin components and protects against infection by *Ceratocystis polonica*, a bark beetle-associated fungus. *Tree Physiology* **26**(8): 977-988.

- Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. 2014.** Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS One* **9**(1): e78644.
- Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, Puiu D, Roberts M, Wegrzyn JL, de Jong PJ. 2014.** Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics* **196**(3): 875-890.
- Zulak KG, Bohlmann J. 2010.** Terpenoid biosynthesis and specialized vascular cells of conifer defense. *Journal Of Integrative Plant Biology* **52**(1): 86-97.

Chapitre 2.

Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression

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

2.1 Résumé

2.1.1 Contexte

Les conifères ont de très grands génomes (13 à 30 Gigabases) qui sont pour la plupart non-caractérisés bien que de nombreuses banques d'ADNc sont disponibles depuis peu. Ce rapport présente un aperçu général de la variation transcriptomique dans un conifère d'une part, et la conservation et la diversité de profil d'expression des gènes selon les tissus végétatifs majeurs d'autre part.

2.1.2 Résultats

Une puce à oligonucléotide a été développée à partir des bases de données d'ADNc de *Picea glauca* et de *P. sitchensis*. Elle représente 23 853 gènes uniques et convient pour faire le profilage transcriptionnel de plusieurs espèces. Une comparaison des tissus xylèmes secondaires et phellodermes a démontré que l'expression préférentielle dans ces tissus vasculaires était hautement conservée entre les espèces de *Picea*. La séquence à ARN a confirmé la préférence tissulaire par expression et fourni une validation robuste de la conception de la puce. Une petite base de données des profils d'expression appelée PiceaGenExpress a été développée à partir de 150 hybridations couvrant huit types de tissus majeurs. Au total, les transcrits ont été détectés au moins dans un tissu chez 92% des gènes de la puce. Les gènes non annotés se sont généralement exprimés plus faiblement et dans peu de tissus que les gènes à fonction connue ou prédite. La diversité d'expression à l'intérieur des familles de gènes peut être rapidement estimée à partir de PiceaGenExpress. Les dehydrines et les protéines de régulation osmotique *late embryogenesis abundant* (LEA) sont de plus grandes familles de gènes chez les conifères que chez les angiospermes. Un grand contraste et une faible diversité ont été observés chez la famille des dehydrines tandis que divers profils suggèrent une plus grande diversification entre les gènes de LEAs.

2.1.3 Conclusion

La puce à oligonucléotide et la base de données PiceaGenExpress représentent les premières ressources de ce genre pour les gymnospermes. Ce rapport sur l'analyse transcriptomique de l'épinette est censé accélérer les études génétiques sur le vaste et important groupe des conifères.

2.2 Abstract

2.2.1 Background

Conifers have very large genomes (13 to 30 Gigabases) that are mostly uncharacterized although extensive cDNA resources have recently become available. This report presents a global overview of transcriptome variation in a conifer tree and documents conservation and diversity of gene expression patterns among major vegetative tissues.

2.2.2 Results

An oligonucleotide microarray was developed from *Picea glauca* and *P. sitchensis* cDNA datasets. It represents 23,853 unique genes and was shown to be suitable for transcriptome profiling in several species. A comparison of secondary xylem and phelloderm tissues showed that preferential expression in these vascular tissues was highly conserved among *Picea* spp. RNA-Sequencing strongly confirmed tissue preferential expression and provided a robust validation of the microarray design. A small database of transcription profiles called PiceaGenExpress was developed from over 150 hybridizations spanning eight major tissue types. In total, transcripts were detected for 92% of the genes on the microarray, in at least one tissue. Non-annotated genes were predominantly expressed at low levels in fewer tissues than genes of known or predicted function. Diversity of expression within gene families may be rapidly assessed from PiceaGenExpress. In conifer trees, dehydrins and late embryogenesis abundant (LEA) osmotic regulation proteins occur in large gene families compared to angiosperms. Strong contrasts and low diversity was observed in the dehydrin family, while diverse patterns suggested a greater degree of diversification among LEAs.

2.2.3 Conclusion

Together, the oligonucleotide microarray and the PiceaGenExpress database represent the first resource of this kind for gymnosperm plants. The spruce transcriptome analysis reported here is expected to accelerate genetic studies in the large and important group comprised of conifer trees.

2.3 Background

Microarray (MA) transcript profiling and RNA sequencing (RNA-Seq) represent powerful approaches to rapidly gain functional information on a genome-wide scale. Information on RNA transcript abundance is a key to assessing the biological role of gene products and cannot be directly deduced from a gene's sequence. This has lead researchers to develop databases of RNA abundance profiles, first and foremost for model organisms. For example, the AtGenExpress database was created for the model-plant *Arabidopsis* from a host of tissue preferential and stress response expression profiles [1]. Databases such as AtGenExpress are particularly useful for the identification of groups of co-expressed genes. Other plant oriented databases include the poplar PopGenIE made up of tissue, developmental and stress response profiles [2]. Reflecting the value of gene expression data, public organizations and institutes also maintain generic databases like the Gene Expression Omnibus or GEO (NCBI) and ArrayExpress (EBI), among others, which host datasets from a wide array of organisms.

Recent transcriptome-wide analyses underscore the importance of gene expression in the genetic architecture of complex traits. Studies in fruit flies, mice, humans and maize show that a proportion of the genetic variants underlying complex phenotypes exert their effects through gene expression [3,4]; so, discovering the genetic basis for the variation in transcript abundance is central to understanding phenotypic variation [5]. Gene expression studies also provide insights into the molecular impacts of natural selection. For example, expression profiling showed the differential action of selection pressure on different tissues and organs in humans [6]. A comparative analysis of mouse and human showed a high level of conservation in the expression of orthologous genes, showing the stability of house-keeping genes and the variability of tissue specific genes [7].

Transcriptome profiling is facilitated by the availability of a reference genome but many studies have also been based on large-scale cDNA sequence datasets. In plants, many angiosperm genomes have been sequenced, including the model plant *Arabidopsis thaliana* [8], rice [9], poplar [10] and grapevine [11]; however, reference genomes are still lacking for plant phyla belonging to the gymnosperms. The best studied gymnosperms are conifers, which as a group have extremely large genomes



(ranging from 18 to 30 Gb). In conifers including pines (*Pinus spp.*), spruces (*Picea spp.*), Douglas-fir (*Pseudotsuga menziesii*) and Japanese cedar (*Cryptomeria japonica*) over 1 million expressed sequence tags (ESTs) have been obtained from dideoxy sequencing and assembled to infer putative unigenes or transcript sets (reviewed in [12]). Large collections of cDNAs are available for white spruce (*Picea glauca*) [13] and Sitka spruce (*P. sitchensis*) [14]. From 30% to 40% of conifer sequences cannot be annotated because they lack sequence similarity to known genes ([13-16]).

This report describes a large-scale oligonucleotide microarray developed from the extensive cDNA datasets available for spruce trees (*P. glauca*, *P. sitchensis*) to achieve broad transcriptome coverage. Previously, microarrays were developed from PCR amplicons (cDNA microarrays) primarily in pines and spruces (reviewed in [12]). Many of the cDNA microarrays have ranged from a few hundred to several thousand cDNAs, and a few of them have included over 20,000 spots, i.e. in *Picea sitchensis* [17] and *Pinus taeda* [18]. They have essentially been used in comparative experiments (using two-dye designs) to investigate transcriptome remodeling during tissue differentiation, development, or in response to environmental cues [12], but a general characterization of conifer transcriptomes has been lacking.

A major goal of the present study was to assemble transcript profiles from spruce trees (*Picea spp.*) into a database called PiceaGenExpress, aiming to characterize the basic features of a conifer transcriptome such as the number of transcribed genes in a variety of tissues. The reference profiles in PiceaGenExpress enabled exploratory analyses of the diversity of expression patterns within and among gene families and the expression of retrotransposons. In addition, conservation of gene expression in secondary vascular tissue was studied based on interspecific comparisons of tissue preferential expression. The accuracy of microarray profiles and design was directly evaluated by RNA-Seq analysis of the same samples as those used for one of the microarray experiments included in PiceaGenExpress.

2.4 Results and Discussion

2.4.1 Development of an oligonucleotide microarray for spruces (*Picea spp.*)

A large-scale custom microarray containing 31,604 oligonucleotide probes was designed for broad representation of the spruce (*Picea spp.*) transcriptome. The 70 nucleotide probes were based on unique cDNA sequences from white spruce (*P. glauca*) [13] and Sitka spruces (*P. sitchensis*) [14] (Table 2.1). Both of these conifers have extensive ESTs and FL-cDNA sequence databases developed from dideoxy sequencing (Sanger method); 454 ESTs (GS-FLX) were also available for *P. glauca*. The probe design parameters and microarray manufacturing methods were experimentally determined through hybridization experiments with a microarray of 3,900 oligonucleotide probes specifically designed for optimization tests (for details see Supp. Mat. 2.1, Figure S2.1 and Figure S2.2).

Table 2.1 Development a large-scale oligonucleotide array for spruces (*Picea spp.*)

sequence information used to design oligonucleotide probes from *Picea glauca* and *P. sitchensis* sequences

Category	Number of probes	%	Probe designed ¹	Confirmation of <i>P. glauca</i> (Pgl) cDNA clone or other
1	11,214	35%	Pgl	Confirmed in Pgl and Psi cDNAs
2	12,251	39%	Pgl	Confirmed by Pgl 454 seqs or Psi cDNAs
3	4,840	15%	Psi	Confirmed by Pgl 454 seq
4	1,629	5%	Pgl	Unconfirmed but Pgl clones ≥ 2
5	1,670	5%	Psi	Psi full-length cDNA only (not found Pgl)
All	31,604			

¹Each probe was designed based on the sequence of *P. glauca* (Pgl) or *P. Sitchensis* (Psi)

The high level of sequence identity in the *Picea* datasets enabled the design of a single probe matching both *P. glauca* and *P. sitchensis* sequences, for most genes; however, the datasets did not overlap entirely so that some probes were unique to one of the species and could not be verified in the other. A level of 95.7% of sequence identity or higher (3 mismatches or less) was obtained for nearly all of the probes relative to *P. glauca* sequences, and for 59% of the probes relative to *P. sitchensis* (Table 2.2). Preliminary experiments indicated that 3 mismatches had a small impact on the hybridization signal intensity and the ability to detect differential expression (See Supp. Mat. 2.1, Figure S2.1 and Figure S2.2). Analyses presented in this report are

based on the set of 25,045 probes designed from *P. glauca* sequences, which match 23,853 unique cDNAs in the *P. glauca* gene catalogue of Rigault et al. [13].

Table 2.2 Analysis of probes: sequence similarity of probes aligned to *Picea glauca* and *P. sitchensis* sequences

<i>P. glauca</i>	<i>P. sitchensis</i>	Number of probes	%
67-70 (>95.7%)	67-70 (>95.7%)	17,279	54%
67-70 (>95.7%)	NA	11,999	38%
67-70 (>95.7%)	63-66 (>90%;<95%)	923	3%
63-66 (>90%;<95%)	67-70 (>95.7%)	529	2%
NA	67-70 (>95%)	869	3%

The potential for utilizing this oligonucleotide microarray in other species and genera of the Pinaceae family was evaluated by using comparative RNA hybridizations with four different spruces (*Picea spp.*), two pines (*Pinus spp.*) and a larch (*Larix laricina*) (Figure 2.1 and Figure S2.3). The hybridization outcomes were highly similar among the spruces (Figure 2.1A-C). A large majority of the probe signal intensities varied less than 2-fold among the spruces (Figure 2.1G). When compared to *P. glauca*, the pines and larch gave more divergent results as shown by the number of common positive probes (above background) and overall data correlations (Figure 2.1D-E). Surprisingly, the pines and larch gave signals of equal or greater intensity than *P. glauca* for most of the probes, and a decreased signal for 17% to 34% of the detected probes. Taken together, these observations indicated that the microarray is suitable for direct comparisons of transcript levels between spruces. For pines and larch, a subset of the probes may not be informative. In general, the MA appears appropriate for studies comparing data within the same genus.

2.4.2 Differential expression in the vascular transcriptome is conserved among *Picea* species.

Microarray transcript profiles compared two tissues that support secondary vascular growth, i.e. diameter stem growth, as it is a key feature of the life habit of trees. Secondary xylem is the wood forming tissue located on the internal side of the cambial meristem. It was compared to a composite sample of phloem and phelloderm tissues (referred to here as phelloderm) located on the outer side of the cambial meristem in three spruce species. Identical analyses were carried out in three spruces: *P. glauca*, *P. sitchensis* and *P. mariana*. The numbers of transcripts detected for these

two tissue types were highly conserved in the three species, ranging from 13,744 to 14,513 in xylem, and 14,990 to 15,697 in phelloderm.

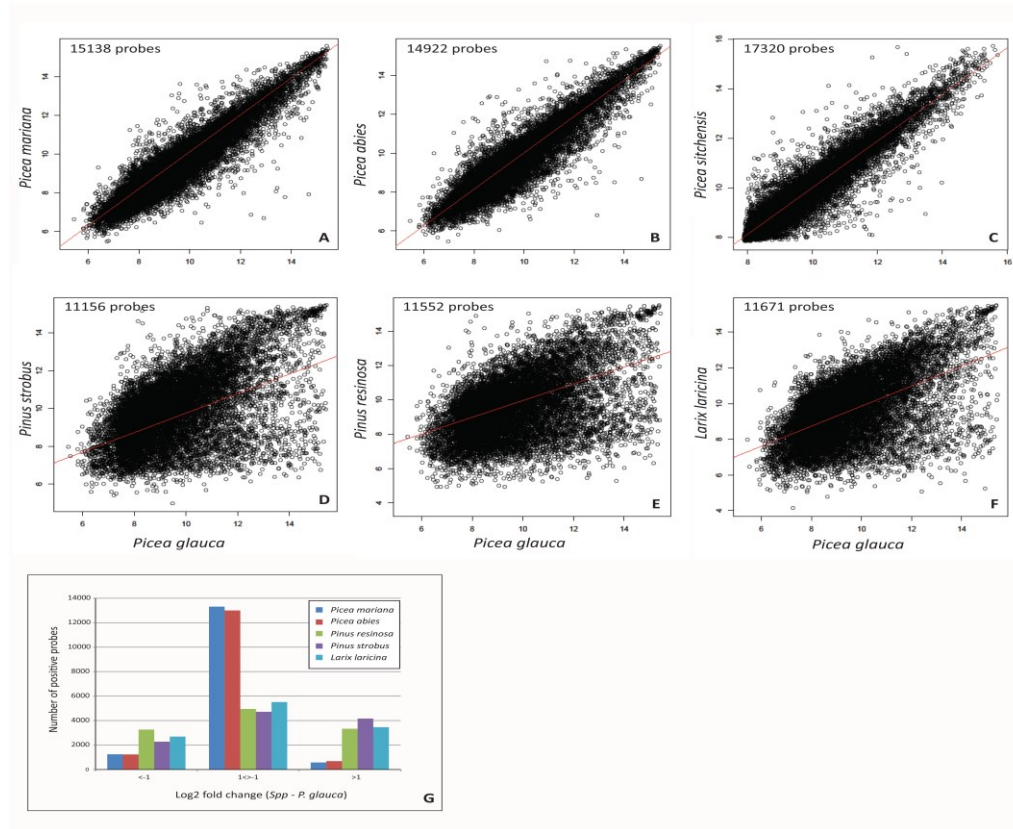


Figure 2.1 Interspecific comparison of hybridization intensities in secondary xylem

A-F: Pair-wise comparison of white spruce and six other species based on the number of shared positive probes indicated in the plots. The squared correlation coefficients (r^2) are as follows 0.86 (A), 0.85 (B), 0.89 (C), 0.29 (D), 0.22 (E) and 0.27 (F). G: Analysis of signal intensity variation between species; the fold change (FC) was determined from the average normalized signal intensities (log2 scale). An FC of 1 or -1 represents a two-fold signal increase or decrease, respectively. For phelloderm results, see Figure S2.3

A total of 5,407 genes were differentially expressed (DE) in all three species. Tissue preferential transcript accumulation and the fold difference between the tissues were very similar among the three species (Figure 2.2). The small number of genes that varied in their tissue specificity (60 genes or 1.1% of the DE genes) indicated that genes with small difference in expression between tissues were more prone to vary between species or be less accurately determined.

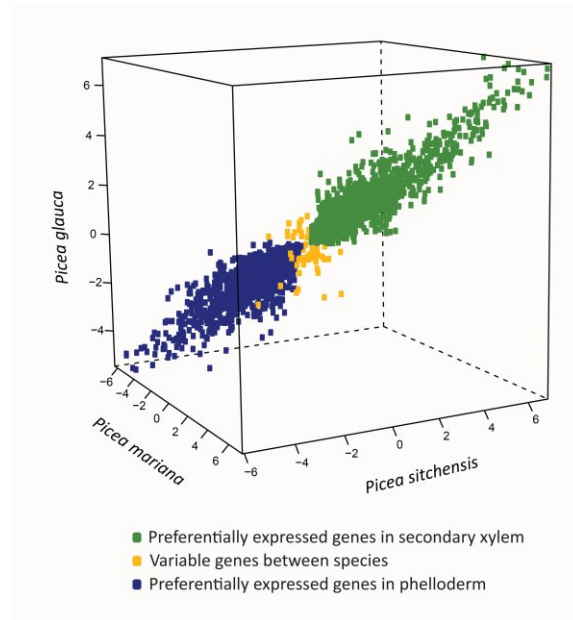


Figure 2.2 Preferential expression in secondary vascular tissues of three spruce species

The FC data in the plot represent genes with differential expression in all three spruces. The scale is the log2 fold change

Different MA experiments comparing xylem and phloem tissues in angiosperms including *Arabidopsis* [19,20] and poplar [21], and in *P. glauca* have helped to delineate groups of genes whose expression was of particular relevance to secondary vascular growth. A MA profiling study in *Arabidopsis* root-hypocotyl defined a set of 319 genes specifically regulated in secondary xylem compared to phloem or non-vascular tissues [19]. In young spruce trees, 360 sequences were shown to be xylem preferential compared to needles and phloem [22]. A core set of 52 xylem genes was identified by Ko et al. [21] based on transcriptome analyses of secondary xylem in *Arabidopsis thaliana* and poplar, and of cotton fibres. The expression patterns reported here for three spruces indicated that tissue preferential expression for xylem compared to phelloderm were conserved among spruces. These conserved patterns could be the basis for comparative genomics of conifers and angiosperms trees. This finding is also relevant for studying the genetic architecture of wood traits because it was shown that xylem preferential expression was a feature of genes associated with genetic variation in wood properties [23].

2.4.3 Validation of the microarray and profiling results by RNA-Sequencing.

An RNA-Seq study *P. glauca* secondary xylem and phelloderm used the same RNA samples as for the MA profiling, but the samples for each tissue type were pooled before sequencing. A total of 59.5 M high-quality sequences were obtained and mapped to the 27,720 cDNA clusters of the *P. glauca* gene catalogue as previously described [13] (Table 2.3). The sequence frequency data were normalized by transforming the data to reads per kilobase of exon model per million mapped reads (RPKM) as described by Mortazavi et al. [24]. These authors estimated that an RPKM > 1 represents one RNA molecule per cell; therefore, an RPKM > 1 was used as a threshold for detection.

Table 2.3 RNA-Seq data

	Xylem	Phelloderm	Total or Both
Total reads (HQ) (millions)	29.5	29.9	59.5
Reads mapped (millions)	17.3	20.4	37.7
Genes - RPKM ¹ > 1	19,604	21,366	22,012
Genes - RPKM > 3	16,168	18,297	19,108

¹ RPKM, Number of reads per kilobase of mapped sequences per megabase based on Mortazavi et al. [24]

Nearly all of the DE genes from the MA analysis (99%) were represented in at least one of the RNA-Seq samples, and a large majority (84% to 88%) of genes with a 2-fold difference on the MA, were also differentially represented in RNA-Seq (Table 2.4). These RNA-Seq data confirmed the tissue preferential expression (xylem vs. phelloderm) of 99.3% of the genes determined to be differentially expressed with both methods (Table 2.3). The cross-validations between the two methods were based on differential expression in 2,171 to 4,181 genes depending on the fold difference threshold; therefore, it represented a highly robust confirmation of the accuracy of the MA results. The overlap in the number of genes varied depending on the fold difference threshold, likely owing to differences in experimental design and technique; however, the preferential expression (xylem vs. phelloderm) was highly congruent between the analysis methods (Table 2.4).

Table 2.4 Validation of microarray results by RNA-Seq

Criteria		MA ¹	RNA-Seq and MA ²					RNA-Seq only ⁴		
P-value	FC (log2)		RPKM > 1		DE ³		Tissue pref.	DE	Add	
0.05	0.5	5,666	5,592	99%	4,181	75%	4,155	99%	776	19%
0.05	1.0	2,614	2,588	99%	2,265	88%	2,253	99%	677	30%
0.01	0.5	5,526	5,466	99%	3,608	66%	3,585	99%	542	15%
0.01	1.0	2,608	2,582	99%	2,171	84%	2,160	99%	491	23%

¹ DE positive genes determined by MA

² Number of DE genes detected by RNA-Seq among the DE genes found by MA hybridization; Tissue pref., indicates that the tissue preference for xylem or phelloderm was the same

³ DE, differential expression in RNA-Seq, determined by a Chi-squared test (df = 1) with Benjamini-Hochberg correction for multiple testing

⁴ RNA-Seq results for non-DE genes from MA (considering genes represented on the MA only)

The RNA-Seq detected more transcribed genes than the MA analysis, i.e. close to 6,000 genes with an RPKM > 1 (Table 2.3). The number of DE genes determined by RNA-Seq and not by MA ranged from 491 to 776 (Table 2.4), although these results are only suggestive given the experimental design used for the RNA-Seq. In addition, some of the genes shown to be tissue preferential by MA analysis were suggested by RNA-Seq to be specifically expressed only in one of the tissue types (Table 2.5). Interestingly, 50% to 65% of these putative tissue-specific sequences had no annotation based on similarity to TAIR sequences or the detection of Pfam domains, compared to less than 40% for the entire set of genes represented on the microarray [13]. This observation is consistent with findings from animal research showing that genes that have more specific expression patterns also tend to have less conserved sequences among species [4]. In evolutionary terms, genes that are expressed only in xylem or in phloem may either represent sequences that are unique or are more highly diverged in conifers compared to other plants.

Table 2.5 Tissue specificity in RNA-Seq

	Xylem Total	Specific Annotated	Phelloderm Total	Specific Annotated
Genes RPKM > 1	44	16 (36%)	186	79 (42%)
Genes RPKM > 3	17	6 (35%)	81	41 (50%)

2.4.4 PiceaGenExpress contains reference profiles that reveal patterns of tissue preferential expression.

The PiceaGenExpress database was developed from over 150 MA hybridizations of *Picea spp.* obtained for eight sample types representing different tissues and experiments (Table 2.6) (for procedures, see methods). Detailed analysis of each of the eight datasets will be presented elsewhere. In any given sample type, transcripts were detected for 10,067 to 17,070 genes (hybridization signal above background threshold); transcripts were detected for 21,939 different genes considering all of the tissues, i.e. 92% of the *P. glauca* genes represented on the microarray. The PiceaGenExpress is made available as a flat file (Table S2.2) for ease of upload and access. Less than 10% of the genes were unique to one sample type. A simple ranking procedure was applied, as a means to represent the relative expression and enable qualitative comparisons across samples. The genes were ranked within each dataset of PiceaGenExpress separately, based on average signal intensity in a sample type, and then equally divided into 10 intensity categories (expression classes): from the lowest 10% (class 1) to the highest 10% (class 10) (Table S2.1, Figure S2.4).

The PiceaGenExpress database was used as a tool to rapidly assess tissue preferential and invariant expression profiles (Figure 2.3). It allowed us to identify gene families with diverse patterns of tissue preferential expression. For example, transcripts for three cellulose synthase (CS) genes stood out as being strongly overrepresented in the secondary xylem (both juvenile and mature) compared to all other tissues, where as other CS transcripts were more ubiquitous (Figure 2.3A). This observation was consistent with studies in pine [25] and poplar [26] showing that CesA genes that are specialized in secondary cell wall formation occurred as triplets of genes. A different type of pattern was observed with transcripts for photosystem I and II proteins, which are generally strongly expressed in green tissues (Figure 2.3B). As might be expected, transcripts of all of the genes in this class were detected at high levels and were co-expressed in young needles, expanding buds, and phelloderm of young trees, and at low levels in non-photosynthetic tissues including roots, immature embryos and megagametophytes. For secondary xylem, the same transcripts were abundant in young trees but low in the mature trees, which could be accounted for by differences in exposure to light and bark thickness. Genes encoding house-keeping proteins like ubiquitin did not vary between the tissues (Figure 2.3C). These observations relating

to tissue preferential expression are rapid and simple within PiceaGenExpress. They are also consistent with known expression and physiology, and suggest that the methodology that was followed to develop the database is adequate for revealing key patterns of gene expression.

Table 2.6 The PiceaGenExpress database: sample characteristics, hybridizations and detected genes

	Plant Material			Microarrays			Genes ⁵		
	Tissue type	Sp ¹	Source sampling ²	Genotype ³	Slide	Imaging ⁴	Total	Unique	Non-annotated
1	Embryogenic cells	Pgl	a, a	1	6	SQ	10,066	100	2,456
2	Vegetative buds	Pgl	b, c	2	10	SQ	12,361	128	3,216
3	Xylem (Mature)	Pgl	d, d	60	60	SQ	14,686	176	4,232
4	Xylem (juvenile)	Pgl	e, f	30	20	SQ	13,807	56	3,701
5	Phelloderm	Pgl	e, f	30	20	SQ	15,803	214	4,391
6	Young needles	Pgl	e, f	30	10	SQ	12,819	167	3,025
7	Megagametophytes	Pgl	g, c	3	3	PA	17,056	1,111	5,205
8	Adventitious roots	Pab	h, c	8	20	PA	15,718	393	4,696
Total detected								21,24	2,345
Not detected								2,61	

¹ Pgl: *Picea glauca* (White spruce); Pab: *Picea abies* (Norway spruce);

² The source of materials and the sampling method are from the following studies: a, [43]; b, [37]; c, this paper, see methods; d, [23]; e, O-P seedlot, this paper see methods; f, [22]; g, C2856 parent from [38]; h, [44]

³ Total numbers of genotypes analyzed (either individually or in pools).

⁴ Microarray Scanner and Image Processing: SQ, ScanArray Express (Perkin Elmer, Waltham MA, USA) and QuantArray v3.0 (Packard BioChip Technologies, Billerica, MA, USA); PA, PowerScanner (Tecan Group Ltd., Männedorf, Switzerland) and ArrayPro Analyser v6.3 (Media Cybernetics, Bethesda, MD, USA)

⁵ Total: number of genes above background (see methods); Unique: genes detected only in one tissue.

2.4.5 Non-annotated genes are expressed at low levels and in fewer tissues

From 30% to 40% of cDNAs from non-model organisms such as conifers could not be annotated by standard sequence similarity searches like BLAST [15, 16] or HMMER [13]. We investigated whether insights into the role of non-annotated sequences could be obtained by surveying their abundance and distribution among tissues in PiceaGenExpress (Figure 2.4; Figure S2.5). First, we observed that non-annotated sequences, i.e. sequences that lacked similarity to known plant genes, were more

represented among low abundance transcript classes. On average, the non-annotated sequences represented 37.1% of the expression class 1 genes and 22.4% in class 10 (not shown), and a same trend was observed in each of the tissue types (Figure 2.4A-C; Figure S2.5A-E). These non-annotated sequences could be either unique to conifers, owing to differential loss or acquisition of genes among taxa, or could be too highly diverged to permit functional annotation. Their general low level of expression may suggest that they were expressed in fewer cells or were tightly regulated in some manner, perhaps playing a more specialized role in metabolism or development.

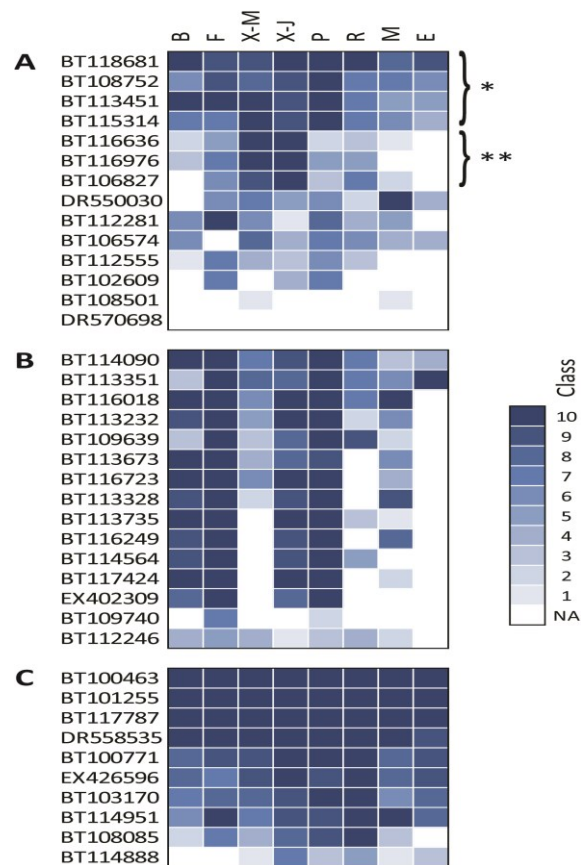


Figure 2.3 The PiceaGenExpress database reveals tissue preferential and conserved expression patterns within three gene families

A: Cellulose synthases. B: Photosystem I and II proteins. C: Ubiquitins. NA: Not detected. Tissues: B (Vegetative buds), F (Foliage), X-M (Xylem - mature), X-J (Xylem - juvenile), P (Phelloderm), R (Adventitious roots), M (Megagametophytes), E (Embryogenic cells)

Second, our data showed that annotated and non-annotated genes were strongly contrasted in regard to the number of tissues in which transcripts were detected (Figure 2.4D). The majority of the annotated sequences were detected in seven or eight of the tissues. In contrast, the non-annotated genes were much more likely to be found in few tissues. This observation is consistent with the idea that less conserved (including non-annotated) sequences may generally play more specialized roles. It is also consistent with findings from comparative expression studies of mice and humans showing that house-keeping genes were more highly expressed and were more conserved among species both in terms of their sequence and their expression [7].

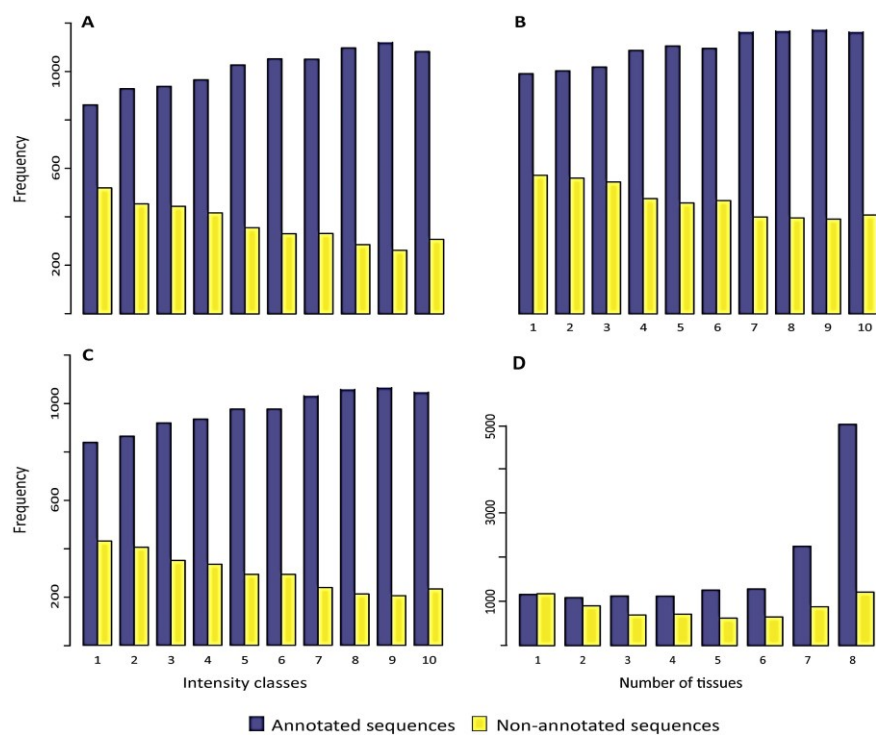


Figure 2.4 Expression classes and numbers of tissue of annotated and non-annotated sequences

A-C: Number of annotated and non-annotated sequences per expression class for xylem (A), roots (B) and young foliage (C). Other tissues are shown in Figure S2.5. D: Number of tissues in which each annotated and non-annotated sequence was detected. Frequency, number of genes in a given intensity class or detected in a given number of tissues types

2.4.6 Diversity of expression profiles varies within and among gene families with related functions

A total of 28 different gene families were reported to be statistically overrepresented in spruce compared to major angiosperms based on the occurrence of protein domains of known function [13]. Approximately one fifth of them were related to stress responses including osmotic regulation proteins like dehydrin and late embryogenesis abundant protein (LEA) gene families. The two families are of similar size in the white spruce gene catalogue, with 49 sequences each [13]. Expression profiles (distribution of transcript abundance classes) within these families were examined in PiceaGenExpress to gain insight into the extent of functional divergence or redundancy that might be associated with gene family expansion. The expression of 49 different sequences containing a dehydrin protein domain indicated striking differences between tissue types during normal development. Many more dehydrin sequences were detected in roots (high expression classes), megagametophytes and, to some extent, in the phelloderm than in the other tissues (Figure 2.5A). A large group of the sequences seemed to be co-expressed, i.e. all were expressed either very strongly (class 9–10) or more weakly (class < 5).

Expression profiles of the 49 gene sequences containing a LEA domain also varied between tissues and appeared more diversified than observed for dehydrins (Figure 2.5B). The extent of divergence among the members in each family was analyzed by clustering and determination of Euclidean distances among the sequences (Figure 2.5C). Overall most of the nodes (38 out of 49) were separated by a greater distance in the LEA family than in the dehydrin family, indicating that during normal development, the regulation of the dehydrin family members is less diversified. This observation may point at greater functional diversification among LEAs.

Both dehydrins and LEAs have been shown to be expressed during normal development and to be water stress responsive in conifers [18, 27, 28]. Our study only considered expression during normal development. Previous studies in conifers have monitored the expression of a small subset of these large gene families. For example, five and eight dehydrins transcript sequences were studied in foliage of maritime pine [27] and vegetative buds of Norway spruce, respectively [28]. We detected 22 distinct sequences by MA profiling in young foliage, strongly suggesting that a comprehensive

view of this protein family in response to stress remains to be developed. The expression profiles presented here indicated that osmotic-regulation during normal development may involve more genes (especially dehydrin genes) and potentially more varied functions in some of the tissues (megagametophytes, roots and phelloderm) than others (foliage, embryos). This observation indicates that osmotic response monitoring, whether it is related to drought conditions [27] or to normal developmental processes [28] may be sensitive to tissue preference. In the PiceaGenExpress dataset, several dehydrins were down regulated in vegetative buds that were sampled at the time of bud flush in the spring, as was observed through detailed time series analyses of Norway spruce dehydrins [28]. An interesting feature of the data is that the two seed derived tissues, i.e. immature somatic embryos and megagametophytes from germinating seeds, were highly contrasted in regard to the expression of specific dehydrin and LEA sequences.

2.4.7 Expression of LTR retrotransposon sequences

Recent reports indicated that LTR retrotransposons represented a large fraction of the conifer genome [29-31]. In addition, sequences traced to copia and gypsy-like retroelements were reported as being very frequent in *Pinus contorta* ESTs [32] and were overrepresented in a root xylem cDNA library [33], although some of these sequences may represent genomic contaminations [13]. More evidence is needed to show whether any of these elements are active or if activity may vary as a function of development or environmental stresses. The PiceaGenExpress database was scanned to obtain evidence of RNA transcript production, which is the first step for LTR retrotransposon mobilization. A total of 83 cDNA sequences represented on the MA coded only for protein domains expected for LTR retroelements, such as RNase H, integrase core domain, retrotransposon gag protein and reverse transcriptase. Many of the sequences were not detected at all; however, 46 sequences were detected in at least one tissue, most of them were found in two tissues or more, and only a few accumulated at high levels (expression class 10) (Figure 2.6A). Tissue preferential accumulation was suggested by the fact that few sequences were detected in immature somatic embryos and were present at low levels, where as many sequences were detected and in higher expression classes in mature xylem, roots and megagametophytes.

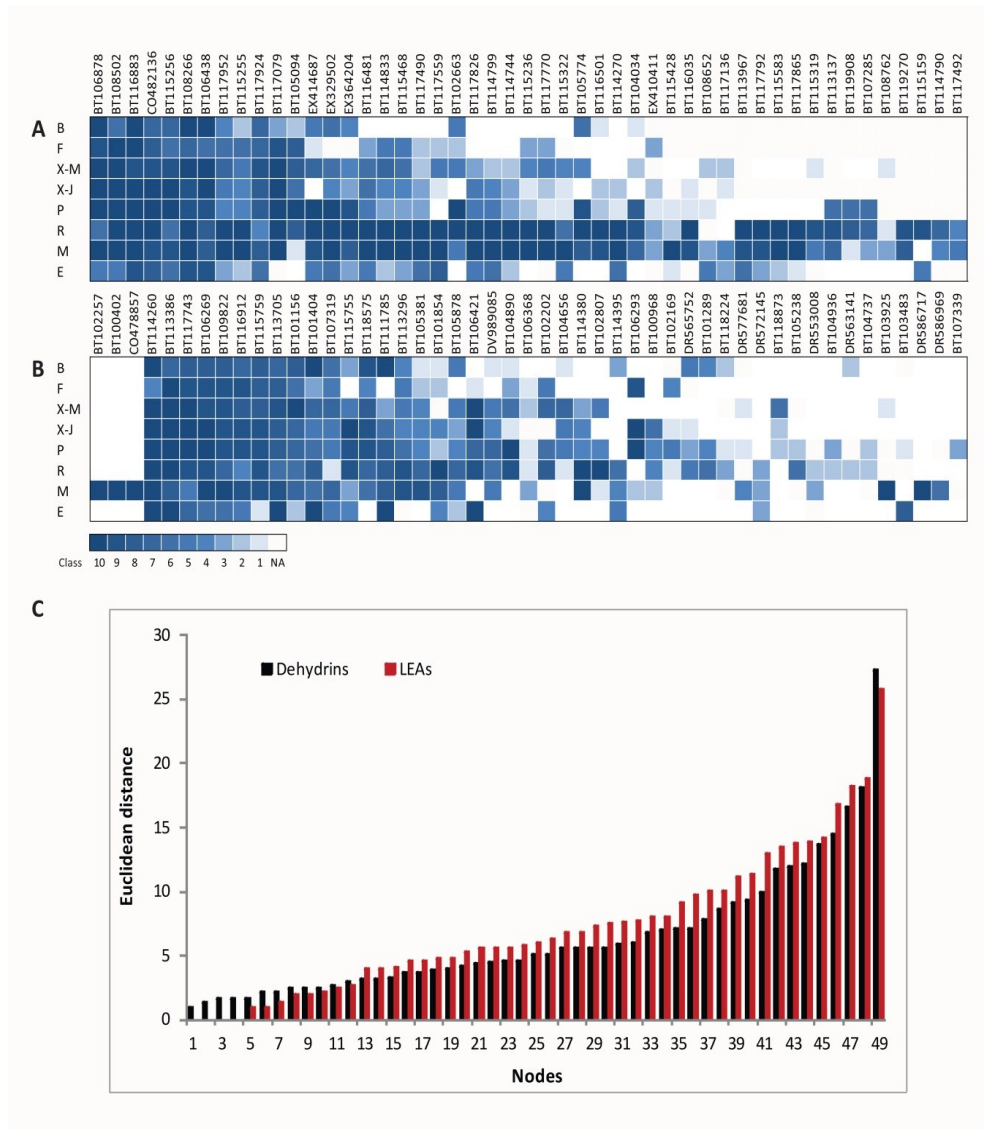


Figure 2.5 Gene expression patterns in two osmotic regulation protein families based on the PiceaGenExpress database

A: Dehydrins. B: Late Embryogenesis Abundant proteins. C: Distribution of Euclidean distances between the members of each protein family. The order of the bars is not representative of the order of the genes in panels A and B. A, B: Tissues: B (Vegetative buds), F (Foliage), X-M (Xylem - mature), X-J (Xylem - juvenile), P (Phelloderm), R (Adventitious roots), M (Megagametophytes), E (Embryogenic cells)

The MA sequences matching putative LTR were estimated to represent up to 78,520 unique copies in the *P. glauca* genome (Figure 2.6B) and a total of 1.18 Million sequences (not shown) based on their occurrence in shotgun sample sequencing data for *P. glauca* [13].

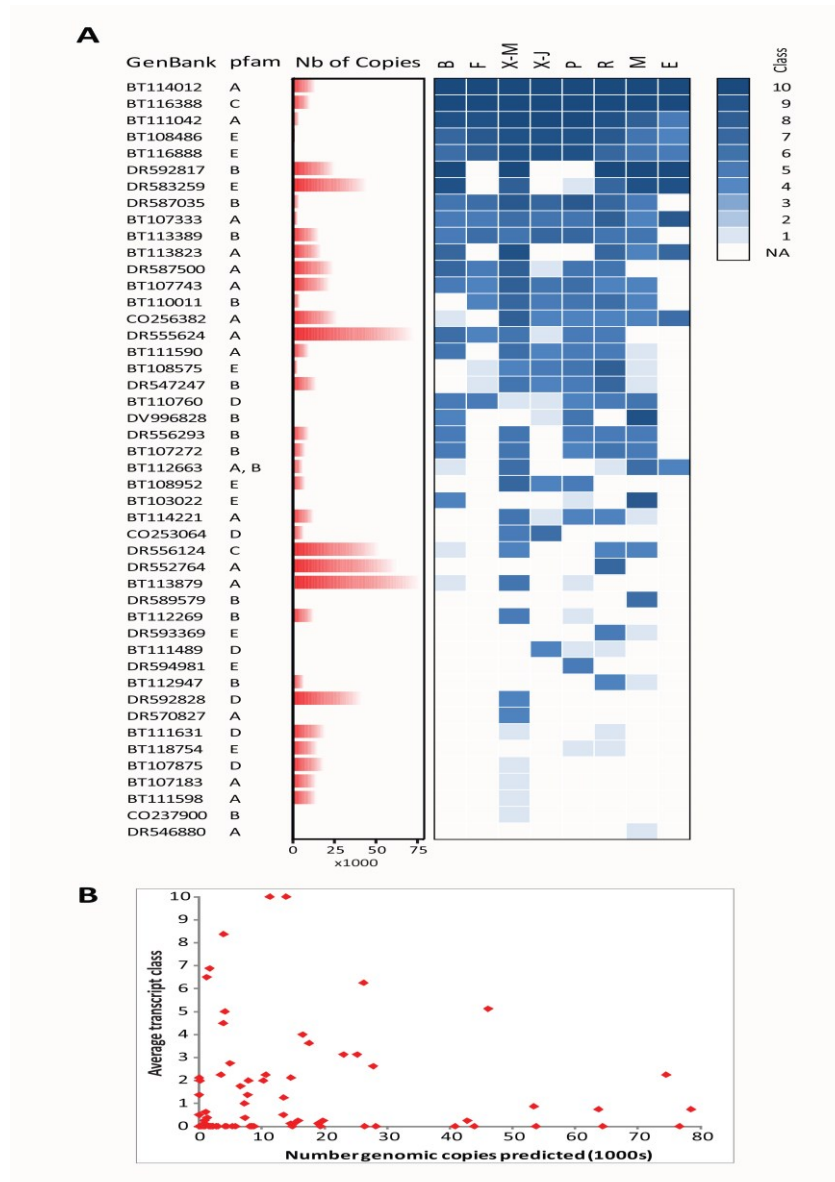


Figure 2.6 Expression patterns of LTR retrotransposons based on the PiceaGenExpress database

A: Sequences containing protein domains of LTR retroelements. Pfam annotations: a: Integrase core domain; b: Retrotransposon gag protein; c: Retroviral aspartyl protease; d: Reverse transcriptase (RNA-dependent DNA polymerase); e: RNase. Tissues: B (Vegetative buds), F (Foliage), X-M (Xylem - mature), X-J (Xylem - juvenile), P (Phelloderm), R (Adventitious roots), M (Megagametophytes), E (Embryogenic cells). B: Expression levels are not correlated with the number of genomic copies

Data are not currently available to estimate what proportion of the genome these sequences may represent; however, if the sequences were derived from an intact LTR of 4000 bp on average, they would represent nearly 5 Gbp or 20% of the *P. glauca* genome. In other words, such a large number of copies is expected to occupy a sizable fraction of the genome. The number of predicted copies did not appear to correlate with transcript accumulation, i.e. number of tissues in which they were detected or relative levels (Figure 2.6A, see right panel; 2.6B). This observation indicates that the data were more likely to result from transcription than genomic contaminations of the RNA samples. It also shows that the number of copies accumulated in the genome is not useful for predicting the level of transcript production. In fact, many of the high copy sequences (>30,000 copies) were not detected at all in any of the tissues. These results point to specific LTR sequences that have the highest potential for mobilization in *P. glauca*.

2.5 Conclusions

An oligonucleotide microarray was developed from *P. glauca* and *P. sitchensis* datasets. It represents 23,853 unique *P. glauca* genes or 85% of the recently reported gene catalogue [13]. Single dye analysis and a ranking procedure were used to develop PiceaGenExpress, a database of reference transcript profiles, based on 150 hybridizations in eight different tissue sample types. These data represent the first resource of this kind for a gymnosperm plant.

The pine family comprises over two hundred species belonging to eight genera. It is the largest and the most economically important of the conifers. Interspecific comparison experiments presented in this report indicated that the microarray may be applied to at least three of these genera. It could be a valuable tool for species where cDNA resources are lacking or underdeveloped. Our findings also indicate that expression profiles from *P. glauca* are likely to be representative of other conifers. RNA-sequencing has become a method of choice for transcriptome profiling but the analysis of RNA-Seq data can be complex owing to factors such as sequence polymorphisms, gene paralogs, and alternate splicing. Therefore, successful application of RNA-Seq depends on the availability or the development of a good quality reference genome or gene catalogue [24], and both are lacking in many non-model species. By mapping *P. glauca* RNA-Seq and microarray data back to the same

cDNAs sequences, we were able to show that the two methods were highly congruent. New sequencing technologies promise to generate high quality sequences in addition to very large volumes of data. Given the large size of conifer genomes, it may be advantageous to use these technologies to develop high quality gene catalogues rather than attempt to assemble entire genomes. Once a gene catalogue is produced, these high throughput sequences represent a powerful methodology for unrestricted gene expression studies [24].

Conifer genomes have a number of characteristics that make them unique, most prominently their enormous size which can reach or even exceed 30 Gbp [30] and their highly repetitive sequences [31]. They are also known to have high levels of heterozygosity and are believed to harbour many gene paralogs, at least in some gene families [12]. For the MA described here, the empirical tests of probe specificity (Figure S2.1) and the probe design parameters appeared sufficient to discriminate between paralogs that are known in the *P. glauca* catalogue of 27,720 genes [13]. The very high level of congruence observed between MA and RNA-Seq results, 99.3% of confirmation of tissue preference from over 2100 genes, also suggest that the MA results accurately reflect the expression of the target sequences. In contrast, issues of cross-hybridization between gene paralogs with different expression patterns would have likely resulted in a lower validation rate.

The tools and methods presented in this report may lead to diverse applications for fundamental discovery in forest genetics and evolutionary biology, such as understanding phenotypic variation in economic and adaptive traits. The genetic architecture of complex phenotypes in plants and trees is routinely probed by scanning the genome for DNA sequence polymorphisms through QTL mapping and association studies [34]. However, Huang et al. [5] summarized several recent studies by stating that discovering the genetic basis for the variation in transcript abundance was central to understanding phenotypic variation. Examining the genetic architecture of gene expression can provide functional insights into physiology and metabolism, for example by revealing the organization of gene networks [35,36].

2.6 Methods

2.6.1 Evaluation of array design and manufacture parameters with test oligonucleotide microarray.

A custom MA comprised of 3,900 oligonucleotide probes was developed to evaluate the impact of design parameters (for details see Supp. Mat. 2.1, Figure S2.1 and Figure S2.2). It contained multiple probes for 929 distinct genes as well as cDNA amplicons for 96 of those genes. The parameters tested include oligonucleotide lengths of 50, 60 and 70 nucleotides, the position of probes within the transcript, the impact of SNPs and indels. The presence of one or three SNP mismatches (distributed throughout the oligonucleotide probes) had a small effect on hybridization signal intensities and in many cases the expression ratio between tissues was largely conserved (See Figure S2.1). In contrast, the presence of seven SNP mismatches in the probe had a large effect on intensities and expression ratios. Based on these data, 70-mer probes that vary by up to three SNPs distributed throughout the probe (95.7% sequence identity) are expected to hybridize to the same RNA and on average produce a signal of similar strength, whereas probes with seven SNPs or more (less than 90 % sequence identity) give very little cross hybridization. These observations establish thresholds of sensitivity to sequence variation, i.e. up to three mismatches have little impact on sensitivity, and specificity, i.e. specificity is achieved with seven mismatches or more to other sequences. The cut-off is situated between four and six SNPs, i.e. between 95% and 91% sequence identity. The presence of short insertions and deletions (up to six nucleotides) located at the center of the 70-mer probes had a small impact on probe performance. The impacts of other parameters tested were generally small and less predictable.

The impact of different spotting buffers and surface chemistries used to manufacture the microarray were also assessed in regard to image quality and data reproducibility. For details on methods and findings, see Supp. Mat. 2.1. We found that the optimum conditions were obtained by using aminosilane coated slides and 3X SSC without betaine as a spotting buffer.

2.6.2 Microarray design and manufacture

The sequences included in the microarray were selected on the basis of reproducible sequence quality from all of the ESTs and FL-cDNA described for *P. glauca* in Rigault et al. [13] and *P. sitchensis* in Ralph et al. [14] (Table 2.1A). To obtain a robust probe set, we selected sequences that were either detected in the two species, verified with two technologies (Sanger and 454) or were derived from a FL-cDNA (Table 2.1A). The probes were 70 nucleotides in length, and were designed to minimize similarity with other sequences in the dataset. Sequence similarity between the probes and known sequences in *P. sitchensis* and *P. glauca* was determined from sequence alignments. The microarray contains 25,045 probes that match with known *P. glauca* sequences; however, they represent 23,853 unique genes based on the most recent clustering [13], such that 1,017 genes are represented by more than one probe.

The microarray consists of 33,984 spotted features including 33,024 sample spots (Invitrogen, Carlsbad, CA, USA), 240 negative buffer spots, and 480 Spot Report Alien Oligos (Agilent Technologies, Inc., Santa Clara, CA, USA). Oligonucleotides were consolidated into 384-well plates, lyophilized by speed-Vac, and resuspended in 3X SSC to a printing concentration of 30 μ M. Oligos were printed on aminosilane slides (Erie, Hudson, NH, USA) with a QArrayMax microarray printer (Genetix Limited, Hampshire, UK) using 946MP2 microarray pins (ArrayIt Corp, Sunnyvale, CA, USA) in a 48-pin tool depositing $\sim$ 0.5 nL per spot onto the slide. The resulting microarrays had a 4 x 12 subgrid layout with 708 spots per subgrid, each spot having approximate diameter and pitch of 90 μ m and 160 μ m, respectively. A 280-bp GFP (green fluorescent protein) oligonucleotide was printed in subgrid corners to assist in grid alignment during image processing. The slides were crosslinked in a UV Stratalinker 2400 (Agilent Technologies, Inc.) at 300 mJ. Array quality was assessed by visual inspection and hybridization of representative slides from a print run by dye-labeled random 9-mer oligonucleotides. The quality control images were acquired via the GenePix 4200AL scanner (Molecular Devices, CA, USA) at a 10 micron resolution and quantified with the Imogene 8.0 software suite. Oligonucleotide library management, printing of microarrays and quality control was performed by the Genome BC Microarray Platform (Vancouver, BC, Canada).

The array design details are available in the Gene Expression Omnibus (accession No. GPL15033). All of the transcript profiling data is also deposited in GEO (available upon acceptance of this manuscript for publication).

2.6.3 Plant materials

The origin, genotypes and method of collection of each type of material are described in Table 2.6, except for details described here. All of the tissue samples were frozen in liquid nitrogen immediately after removal from the trees, the seed or tissue culture vessels, and stored at -80°C until further use.

- Vegetative Buds: Buds were collected from branches of several clonal replicates of 9-year-old trees of *P. glauca* regenerated from two genetically distinct somatic embryogenesis lines as described [37] during the mid-Spring when the buds were just beginning to grow. For each genotype, five biological samples each consisting of 6 buds (approximately 80 mg fresh weight) were used for analysis.
- Secondary xylem, phelloderm (including phloem), young needles of juvenile trees: Nursery planting stock (from open-pollinated seed lots) were obtained as 3-year-old seedlings of *Picea glauca*, *Picea mariana*, *Picea abies*, *Picea Sitchensis*, *Pinus strobus*, *Pinus resinosa* and *Larix laricina*, were transferred to 8-inch pots and grown in a greenhouse under natural light conditions. Sampling of tissues was timed with the ending of primary shoot elongation, i.e. after 6 to 8 weeks of growth, and was as described [22]. For each tissue type five biological samples were prepared by pooling 6 independent trees within each species. These materials were used for interspecific comparisons (all 6 species) and for the development of PiceaGenExpress (*P. glauca* only).
- Megagametophytes: Control-pollinated seed (cross C962856) were obtained from *P. glauca* tree (80112) described in [38], were surface-sterilized for 1 minute in 70% EtOH and 10 minutes in 3% Na-hypochlorite, washing the seeds with sterile water three times between and after treatments. The seeds were then immersed in 4°C sterile water for 24 hours and stratified at 4°C for 28 days. Next, the seeds were moved to 26°C on petri dishes with a moist paper and kept in the dark to start the germination process. After 4 hours of incubation the seeds were opened and the embryo was separated from the



megagametophyte under a dissecting microscope. A total of 3 biological samples comprised of a single megagametophyte were used for analysis.

- Adventitious roots: Norway spruce (*P. abies*) seedlings were grown in the experimental nursery of Finnish Forest Research Institute for about one and a half years before sampling. They were grown in standard nursery growing media, light Sphagnum peat, and fertilized with mineral nutrients. Roots were washed under tap water to remove surrounding peat and approximately 1 cm of the root ends was collected for analysis. Between two and four biological samples comprised of several root ends were analyzed for each of eight different genotypes.

2.6.4 RNA extraction, labelling and hybridization

Total RNA was extracted following Chang et al. [39] as described in Pavy et al. [22] for all of the sample types, except for megagametophytes where poly A + RNA was extracted directly by using polyT coated magnetic beads (Dyna). One µg of total RNA was amplified for each sample replicates, except for the megagametophyte samples where 10 ng of poly A + RNA was used, with the Amino Allyl MessageAmp II aRNA Amplification Kit (Applied Biosystems by Life Technologies, Carlsbad, CA, USA) according to manufacturer's instructions. Five µg of amplified RNA (aRNA) was then labelled with Alexa Fluor 555 or 647 dyes (Invitrogen, Carlsbad, CA, USA) and purified as per the manufacturer's instructions. Dye incorporation efficiency was determined by using a Nanodrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) following the manufacturer's instructions. Depending on the experiment, each microarray was hybridized with one labeled aRNA sample or two samples labelled with different dyes. The sample(s) to be hybridized to a microarray were mixed and the volume was reduced to ~10 µl by evaporating excess water in a DNA 120 speedvac (Thermo Fisher Scientific). Labelled aRNAs were fragmented for 15 minutes at 70°C using Ambion's "RNA Fragmentation Reagents" (Applied Biosystems), placed on ice for 1 minute, denatured for 2 minutes at 95°C, put on ice for 2 min and resuspended in 120 µl hybridization buffer (50% formamide, 5X SSC, 0,1% SDS, 0,1 mg/mL Herring sperm DNA) preheated to 55°C. Samples were kept in a heating block at 50°C until hybridization.

Hybridizations were performed in HS400Pro hybridization stations (Tecan Group Ltd., Männedorf, Switzerland). The slides were heated at 80°C for 10 minutes, then washed once at 37°C with 0.5X SSC, 0.1% SDS for 20 seconds and once at 50°C with 2X SSC, 0.5% SDS for 20 seconds, and prehybridized for 1 hour at 65°C in 5X SSC, 0.1% SDS, 0.1 mg/ml BSA, 0.1 mg/ml Herring Sperm DNA. Next the slides were washed at 55°C with 2X SSC, 0.5% SDS for 1 minute with a 30 second soak and washed again at 45°C for 1 minute with the same solution. The resuspended labeled targets were injected into the chambers and hybridized for 16 hours at 45°C with sample agitation. The slides were then washed as follows: 2 times 1 minute 30 seconds at 45°C with 30 seconds soaking in 2X SSC, 0.5% SDS, 1 time 1 minute at 45°C in 2X SSC, 0.5% SDS, 2 times 1 minute 30 seconds at 45°C with 30 seconds soaking in 0.5X SSC, 0.1% SDS, 1 time 1 minute at 37°C with 20 seconds soaking in 0.5X SSC, 0.1% SDS, 1 time 1 minute 30 seconds at 23°C with 30 seconds soaking in 0.5X SSC, 0.1% SDS, 1 time 30 seconds at 23°C in 0.1X SSC and 2 times 30 seconds at 23°C in milliQ filtered water. Finally slides were dried for 2 minutes 30 seconds with nitrogen gas. Slide scanning and image processing were performed as described in table 2.6.

2.6.5 Microarray data processing and analysis

A procedure was developed to process and analyze microarray intensity data from single dyes, as opposed to fold change methods routinely used for spotted arrays. Data analyses were performed using customized scripts for R and Bioconductor (<http://www.r-project.org> and <http://www.bioconductor.org>). Spots that were flagged as presenting abnormal morphology during the image processing were replaced by mean value of the remaining spots of the same probe from the other slides from the same sample type. Background intensities were subtracted from the foreground intensities. Background-subtracted data were log2-transformed and normalized using quantile correction approach.

A filtering step was applied to select positive genes to be used for further analysis. The mean intensity of spots containing buffer only was calculated for each row of subgrids, and was taken as the minimum intensity of probes for that subgrid. A probe was called positive (detected above background) when its signal intensity was above the buffer intensity on at least 50% of slides within a given sample type. When determining differential expression, positive probes were probes that were detected

according to this criterion in at least one of the tissues (e.g. phelloderm and xylem juvenile). Mean probe intensity was determined for genes represented by more than one positive probe. All microarray experiment data has been submitted to the Gene Expression Omnibus (GEO) under accession numbers GSE35624, GSE35847 and GSE35922.

Statistical testing for differential gene expression used the linear modeling approach and the empirical Bayes statistics [40]; the p-values were adjusted for multiple testing according to Benjamini and Hochberg [41]. Differential genes were those meeting an adjusted p-value ≤ 0.01 , unless stated otherwise.

2.6.6 The PiceaGenExpress database

The PiceaGenExpress database was developed from transcript profiles obtained for eight different tissue types coming from five independent experiments (see Table 2.6). For this, the average signal intensity was determined from all of the slides available for a sample type. The genes were then ranked based on their average signal intensities within a tissue type and equally divided into 10 separate classes according to their signal intensity. Genes from class 1 or class 10 were the 10% with lowest and highest signal intensities, respectively (for the number of genes per class in each type, see Table S2.1 and Figure S2.5). Functional annotations were based on the matches with Arabidopsis proteins (TAIR 9 release) with E-value $< 1e-10$ and on the detection of Pfam domains as described [13]. PiceaGenExpress is made available as a flat file (Table S2.2), which may be uploaded to any type of data processing or spreadsheet platform.

Agglomerative hierarchical clustering with the complete linkage method was performed using hclust function in R [42] on the expression levels for two gene families, i.e. dehydrins and LEA proteins. This approach used a similarity matrix based on Euclidean distance. A smaller distance means that two genes or clusters have more similar expression levels in the tissues analyzed. First, the clustering analysis placed each gene into its own singleton group or cluster. Second, the closest clusters were iteratively joined together until all genes were merged into a single cluster based upon similarity/distance measures between clusters. Dendrograms showing clusters of genes were drawn (Figure S2.6).

2.6.7 RNA-Seq data processing and analysis

Two composite samples were analysed by RNA- Sequencing for validation and comparison to MA results. Samples were prepared by combining equal molar amounts of the *P. glauca* RNAs isolated from secondary xylem (juvenile trees) and phelloderm. These samples were also used in the MA profiling of each of the tissues. RNA-Sequencing, filtering of quality reads and mapping of reads onto the cDNA clusters were described [13]. The number of sequences matching each cDNA cluster was normalized by transforming the data to the number of reads per kilobase of exon model per million mapped reads (RPKM) following the method of Mortazavi et al. [24]. Unless specified otherwise, an RPKM >1 was used as a minimum threshold of detection for RNA-Seq. Differential expression of genes was determined by using the Chi-squared test corrected for multiple testing according to Benjamini and Hochberg [41].

2.7 Acknowledgments

The authors thank A. Séguin, I. Duval, K. Klimaszewska, S. Velmala, and T. Pennanen for sharing data before publication, for incorporation into PiceaGenExpress reference transcript profiles. The authors would like to acknowledge Stephane LeBihan of the Genome BC Microarray Platform for advice and dedicated support of microarray printing. Funding for this work was obtained from Genome Canada, Genome Quebec, Genome BC for the Arborea, Treenomix , and SMarTForests Projects, and the Discovery grant program of the Natural Sciences and Engineering Council (JM).

2.8 References

1. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU: **A gene expression map of *Arabidopsis thaliana* development.** *Nat Genet* 2005, **37(5)**:501–506.
2. Sjödin A, Street NR, Sandberg G, Gustafsson P, Jansson S: **The Populus Genome Integrative Explorer (PopGenIE): a new resource for exploring the *Populus* genome.** *New Phytol* 2009, **182(4)**:1013–1025.
3. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G: **Genetics of gene expression surveyed in maize, mouse and man.** *Nature* 2003, **422(6929)**:297–302.

4. Lemos B, Meiklejohn CD, Cáceres M, Hartl DL: **Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories.** *Evolution* 2005, **59(1)**:126–137.
5. Huang GJ, Shifman S, Valdar W, Johannesson M, Yalcin B, Taylor MS, Taylor JM, Mott R, Flint J: **High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues.** *Genome Res* 2009, **19(6)**:1133–1140.
6. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M: **The evolution of gene expression levels in mammalian organs.** *Nature* 2011, **478(7369)**:343–348.
7. Zheng-Bradley X, Rung J, Parkinson H, Brazma A: **Large scale comparison of global gene expression patterns in human and mouse.** *Genome Biol* 2010, **11(12)**:R124.
8. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408(6814)**:796–815.
9. Yu J, Hu S, Wang J, Wong GKS, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*).** *Science* 2002, **296(5565)**:79–92.
10. Tuskan G, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, *et al*: **The genome of black cottonwood, *Populus trichocarpa* (Torr & Gray ex Brayshaw).** *Science* 2006, **313**:1596–1604.
11. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature* 2007, **449(7161)**:463–467.
12. MacKay J, Dean J: **Transcriptomics.** In *Genomics and Breeding of Conifers*. Edited by Plomion C, Bousquet J, Kole C. New York: Edenbridge Science Publishers and CRC Press; 2011:333–357.
13. 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(1)**:14–28.
14. Ralph S, Chun H, Kolosova N, Cooper D, Oddy C, Ritland C, Kirkpatrick R, Moore R, Barber S, Holt R: **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.
15. Kirst M, Johnson AF, Baucom C, Ulrich E, Hubbard K, Staggs R, Paule C, Retzel E, Whetten R, Sederoff R: **Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*.** *Proc Nat Acad Sc USA* 2003, **100(12)**:7383–7388.

16. Cairney J, Pullman GS: **The cellular and molecular biology of conifer embryogenesis.** *New Phytol* 2007, **176(3)**:511–536.
17. Holliday JA, Ralph SG, White R, Bohlmann J, Aitken SN: **Global monitoring of autumn gene expression within and among phenotypically divergent populations of Sitka spruce (*Picea sitchensis*).** *New Phytol* 2008, **178(1)**:103–122.
18. Lorenz WW, Sun F, Liang C, Kolychew D, Wang H, Zhao X, Cordonnier-Pratt MM, Pratt LH, Dean JFD: **Water stress-responsive genes in loblolly pine (*Pinus taeda*) roots identified by analyses of expressed sequence tag libraries.** *Tree Physiol* 2006, **26**:1–16.
19. Zhao C, Craig JC, Petzold HE, Dickerman AW, Beers EP: **The xylem and phloem transcriptomes from secondary tissues of the *Arabidopsis* root-hypocotyl.** *Plant Physiol* 2005, **138(2)**:803–818.
20. Oh S, Park S, Han KH: **Transcriptional regulation of secondary growth in *Arabidopsis thaliana*.** *J Exper Bot* 2003, **54(393)**:2709–2722.
21. Ko JH, Beers EP, Han KH: **Global comparative transcriptome analysis identifies gene network regulating secondary xylem development in *Arabidopsis thaliana*.** *Mol. Gen. Gen.* 2006, **276(6)**:517–531.
22. Pavy N, Boyle B, Nelson C, Paule C, Giguère I, Caron S, Parsons LS, Dallaire N, Bedon F, Bérubé H: **Identification of conserved core xylem gene sets: conifer cDNA microarray development, transcript profiling and computational analyses.** *New Phytol* 2008, **180(4)**:766–786.
23. Beaulieu J, Doerksen T, Boyle B, Clément S, Deslauriers M, Beauseigle S, Blais S, Poulin PL, Lenz P, Caron S: **Association genetics of wood physical traits in the conifer white spruce and relationships with gene expression.** *Genetics* 2011, **188(1)**:197–214.
24. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5(7)**:621–628.
25. Nairn CJ, Haselkorn T: **Three loblolly pine CesA genes expressed in developing xylem are orthologous to secondary cell wall CesA genes of angiosperms.** *New Phytol* 2005, **166(3)**:907–915.
26. Suzuki S, Li L, Sun YH, Chiang VL: **The cellulose synthase gene superfamily and biochemical functions of xylem-specific cellulose synthase-like genes in *Populus trichocarpa*.** *Plant Physiol* 2006, **142(3)**:1233–1245.
27. Velasco-Conde T, Yakovlev IP, Majada JP, Aranda I, Johnsen O: **Dehydrins in maritime pine (*Pinus pinaster*) and their expression related to drought stress response.** *Tree Genetics & Genomes* 2012.

28. Yakovlev IA, Asante DK, Fossdal CG, Partanen J, Junttila O, Johnsen O: **Dehydrins expression related to timing of bud burst in Norway spruce.** *Planta* 2008, **228(3)**:459–472.
29. Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, Garcia SA, Kubisiak TL, Amerson HV, Carlson JE, Nelson CD: **Evolution of genome size and complexity in *Pinus*.** *PLoS One* 2009, **4(2)**:e4332.
30. Morgante M, De Paoli E: **Toward the conifer genome sequence.** In *Genetics, Genomics and Breeding of Conifers Trees*. Edited by Plomion C, Bousquet J, Kole C. New York: Edenbridge Science Publishers and CRC Press; 2011:389–407.
31. Kovach AS, 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.
32. Parchman T, Geist K, Grahnen J, Benkman C, Buerkle CA: **Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery.** *BMC Genomics* 2010, **11**:180.
33. Pavy N, Laroche J, Bousquet J, Mackay J: **Large-scale statistical analysis of secondary xylem ESTs in pine.** *Plant Mol Biol* 2005, **57(2)**:203–224.
34. Grattapaglia D, Plomion C, Kirst M, Sederoff RR: **Genomics of growth traits in forest trees.** *Curr Opin Plant Biol* 2009, **12(2)**:148–156.
35. Keurentjes JJB, Fu J, Terpstra IR, Garcia JM, Van Den Ackerveken G, Snoek LB, Peeters AJM, Vreugdenhil D, Koornneef M, Jansen RC: **Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci.** *Proc Nat Acad Sci USA* 2007, **104(5)**:1708–1713.
36. Drost DR, Benedict CI, Berg A, Novaes E, Novaes CRDB, Yu Q, Dervinis C, Maia JM, Yap J, Miles B: **Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of *Populus*.** *Proc Nat Acad Sci USA* 2010, **107(18)**:8492–8497.
37. Klimaszewska K, Overton C, Stewart D, Rutledge RG: **Initiation of somatic embryos and regeneration of plants from primordial shoots of 10-year-old somatic white spruce and expression profiles of 11 genes followed during the tissue culture process.** *Planta* 2011, **233(3)**:635–647.
38. 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.
39. Chang S, Puryear J, Cairney J: **A simple and efficient method for isolating RNA from pine trees.** *Plant Mol Biol Rep* 1993, **11(2)**:113–116.

40. Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.** *Stat Applic Gen Mol Biol* 2004, **3**:1.
41. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Royal Stat Society Series B (Methodological)* 1995, **57(1)**:289–300.
42. Kaufman L, Rousseeuw P: *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, Inc; 1990.
43. Zhao N, Boyle B, Duval I, Ferrer JL, Lin H, Seguin A, Mackay J, Chen F: **SABATH methyltransferases from white spruce (*Picea glauca*): gene cloning, functional characterization and structural analysis.** *Tree Physiol* 2009, **29(7)**:947–957.
44. Korkama T, Pakkanen A, Pennanen T: **Ectomycorrhizal community structure varies among Norway spruce (*Picea abies*) clones.** *New Phytol* 2006, **171(4)**:815–824.

2.9 Additional files

Supplemental Materials 1 Experimental tests of optimal oligonucleotide design and manufacture methods

Introduction

A test microarray was developed and utilized to assess the impacts of oligonucleotide design parameters and manufacturing methods. The key parameters that were tested are oligonucleotide probe length (affecting T_m), the impact of SNPs at different positions on the oligonucleotide, impact of insertions or deletions of up to 6 nucleotides. The manufacturing characteristics were those expected to affect spot morphology and signal quality such as slide surface chemistry and spotting buffer. The results reported here are from a tissue comparison experiment was carried out with RNAs isolated from needles (foliage) and secondary xylem of young white spruce trees (*P. glauca*) that previously studied by using an 11k cDNA microarray [1].

Materials and methods

Genes used for test chip development

A set of 1,000 cDNAs were selected based upon differential expression between needles and xylem according to Pavy et al. [1], on their expression level secondary xylem in mature trees (this report). See Table S2.1. The sequences were mapped to 929 cDNA clusters or full-length cDNA described in *P. glauca* [2]. The full-length cDNAs or the longest representative cDNA clone from each cluster were used to design and analyse the oligonucleotide probes. In addition, 96 of the cDNAs were also amplified by PCR as described in Pavy et al. [1] and amplicons were spotted along side with the oligonucleotides (Table S2.3).

Oligonucleotide design parameters

Variations of probe designs and control probes were as follows:

- Perfect match probes 70-mer oligonucleotides probes (1093 probes, representing 929 cDNAs) with one or two probe per gene. These probes were used as reference probes.
- Truncated probes: one 60-mer oligonucleotide and one 50-mer oligonucleotide was designed to overlap with the 70-mer reference probe for comparison purposes, for 50 genes.
- Sets of probes with equidistant SNP mismatches distributed throughout the oligonucleotide (N=1, 3, 7, 10, 14, 21) (%id=98, 95, 90, 85, 80, 70), for 50 genes.
- Probes with mismatching ends: 70-mers designed to match with the corresponding reference probe over 60 or and 50 nucleotides and dangling ends, for 50 genes.

- Sets of probes with insertions and deletions of (1, 2, 3, 6) nucleotides in the middle of a probe, for 50 genes.
- Position controls (50 genes): probes located in the 3', middle and 5' parts of the same gene, as defined by 500-base intervals
- Synthesis controls (50 probes): same oligo synthesized twice
- Negative controls (50 alien control probes)

Microarray manufacture

The oligonucleotides were obtained from Invitrogen (Carlsbad, CA, USA). The microarray printing was carried out at the Genome BC Microarray Platform (Vancouver, BC, Canada) using 4 x 12 spot sub-arrays. Glass slides with two types of surface chemistry were tested and compared: aminosilane coated slides (Erie, Hudson, NH, USA) and epoxysilane coated slides (Erie, Hudson, NH, USA). In addition, variations of the spotting buffer were used to resuspend the oligonucleotides: 3x SSC or 3x SSC with 1M betaine spotted side-by-side on the microarray.

Biological material

For this experiment, 2 year-old white spruce seedlings produced from open-pollinated seed lots were obtained from a commercial nursery (CPPFQ nursery, Ste-Anne de Beaupré, QC, Canada). Seedlings were transferred to 8" diameter pots, and allowed to grow under greenhouse conditions with standard watering and fertilizing as applied for optimal growth, but no artificial lighting or heating. After eight weeks, most of the plants had completed their primary growth for the season and formed a small terminal bud. At that time, young needles (formed during the same season) and secondary xylem tissue were collected from the main stem of 30 trees. To minimize uncontrolled variability between samples, individual samples were randomly mixed to form pools of five trees. Each pool was considered as a biological replicate, and five replicates were analysed for each tissue. Total RNA was extracted as described in Pavy et al. [1].

RNA amplification and labelling

One µg of total RNA was amplified for each replicate with the Amino Allyl MessageAmp II aRNA Amplification Kit (Applied Biosystems by Life Technologies, Carlsbad, CA, USA) following manufacturer's instructions. Five µg of amplified RNA (aRNA) was then labelled with Alexa Fluor 555 or 647 dyes (Invitrogen, Carlsbad, CA, USA) and purified as described in the manual. Incorporation efficiency of the dyes was evaluated using a Nanodrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). The two samples to be hybridized together were mixed and volume was reduced to ~10µl in a DNA 120 speedvac (Thermo Fisher Scientific, Waltham, MA, USA). Labelled aRNAs were fragmented for 15

minutes at 70°C using Ambion's «RNA Fragmentation Reagents» (Applied Biosystems by Life Technologies, Carlsbad, CA, USA), placed on ice for 1 minute, denatured for 2 minutes at 95°C, put on ice for 2 min and resuspended in 120 µl hybridization buffer preheated to 55°C. Samples were kept in a heating block at 46°C until hybridization. Hybridizations were performed in HS400Pro hybridization stations (Tecan Group Ltd., Männedorf, Switzerland).

Results

Assessment of microarray manufacture methods

When processed with our standard hybridization protocol (see materials & methods), aminosilane coated slides were clean, with low background, high spot signal intensity and good spot morphology. On the other hand, epoxysilane coated slides presented high background, low spot intensity and significant spots bleeding all over. We also tested the addition of betaine to the 3X SSC spotting buffer to reduce the occurrence of donut-shaped spots. We observed that the addition of betaine had an unexpected effect: spots which contained betaine showed a systematical effect of “bleeding” causing the spots appear as long streaks. We concluded that the optimum conditions were obtained by using aminosilane coated slides and 3X SSC without betaine as a spotting buffer.

Impacts of probe design.

The impact of probe design parameters was assessed in terms of probe specificity and assay sensitivity based on hybridization signal intensities (Fig. S2.1A, C, E), differential transcript accumulation ratios when comparing different tissues (secondary xylem and phloem, see methods) (Fig. S2.1B, D, F), and detection of differential transcript accumulation with oligonucleotide probes compared to cDNA amplicons (Fig. S2.2). The data presented here are primarily for 70-mer oligonucleotides unless specified as being different.

The presence of single nucleotide polymorphisms (SNPs) had a most strong and consistent effect on the microarray hybridization outcomes. These findings were based on signal intensities and ratios and by comparing probes with SNPs to reference probes with no mismatches. Figure S2.1 shows that a single SNP mismatch in a 70-mer probe had a small effect on probe intensities (Fig. S2.1A). Similarly, the ratio of expression when comparing two tissues, secondary xylem and young needles in the present study, was largely unaffected (Fig. S2.1B). Because the SNP was positioned at the center of the oligonucleotide probe, its impact on hybridization intensity was assumed to be maximal, as SNPs that are closer to one end of the probe would result in longer stretches of perfect match sequences. We were able to infer that SNP near the end of the probes would have very little impact, because we observed that 70-mer probes with degenerate ends gave results that were comparable to the reference (not shown).

Probes were also designed with 3, 7, 10, 14, or 21 SNP mismatches, representing 95% to 70% sequence identity. In each case the SNPs were equidistant from one another and distributed throughout the probe. The presence of three SNP mismatches (distributed throughout the oligonucleotide probes) had a small to moderate effect on hybridization signal intensities (Fig. S2.1C) and in many cases the expression ratio was not largely altered (Fig. S2.1D). In contrast, the presence of seven SNP mismatches in the probe had a large effect on intensities and expression ratios (Fig. S2.1E, F). Based on these data, 70-mer probes that vary by up to three SNPs distributed throughout the probe (95.7% sequence identity) are expected to hybridize to the same RNA and on average produce a signal of similar strength. Probes with seven SNPs or more, on the other hand, give very little cross hybridization. These observations establish a threshold of sensitivity to sequence variation and for specificity with a cut-off between four and six SNPs, i.e. between 95% and 91%.

These findings on the impacts of SNPs were compared to independently obtained results for the same genes based on analyses using an amplicons cDNA microarray and comparing the same tissue samples as those analysed here from secondary xylem and young needles [1]. A total of 93% probes with no mismatches gave the same differential expression response (significant DE at $p < 0.01$, with correction for multiple testing) as found for the same gene by using cDNA microarray (Fig. S2.2). The presences of 1 or 3 SNP mismatches decreased the agreement of results slightly to 87% and 78% respectively. In contrast, almost no agreement was found with 7 SNPs or more.

The presence of single insertion or deletion of 1, 2, 3, or 6 nucleotides positioned at the center of the probe had little effect on the microarray hybridization outcomes. Fig. S2.2 show that the despite these insertions and deletions, tissue preferential expression was highly congruent with the cDNA array results (85%-90% of genes) found in Pavy et al. [1].

Both SNPs and indels are frequent in many conifer coding sequences and represented potential sources of variation that were expected to influence the outcomes of microarray in spruce coding sequences. The data obtained from these probe comparisons are sufficient to model their effects and minimize their impacts in global probe design and microarray experiments.

The other factors we examined had minor and less predictable effects on hybridization outcomes. There was little variation across different high quality 70-mer probes designed at different position within the same gene. The variation appeared random and had a minor impact on the detection of differential expression; the data did not suggest a significant bias of probes located near the 3' or the 5' of the cDNA sequence (not shown). The different probe lengths (50-mer, 60-mer and 70-mer) also had an effect on hybridization signal intensities. As might be expected, the 60-mer and 70-mer oligonucleotides gave slightly more robust results and generally produced stronger hybridization signals than the 50-mers.

References

1. Pavy, N, Boyle, B, Nelson, et al. (2008) Identification of conserved core xylem gene sets: conifer cDNA microarray development, transcript profiling and computational analyses. *New Phytol.* 180: 766-786
2. Rigault P, Boyle B, Lepage P, Cooke J, Bousquet J, MacKay J (2011) A white spruce gene catalogue for conifer genome analyses. *Plant Physiol.* 157: 14-28

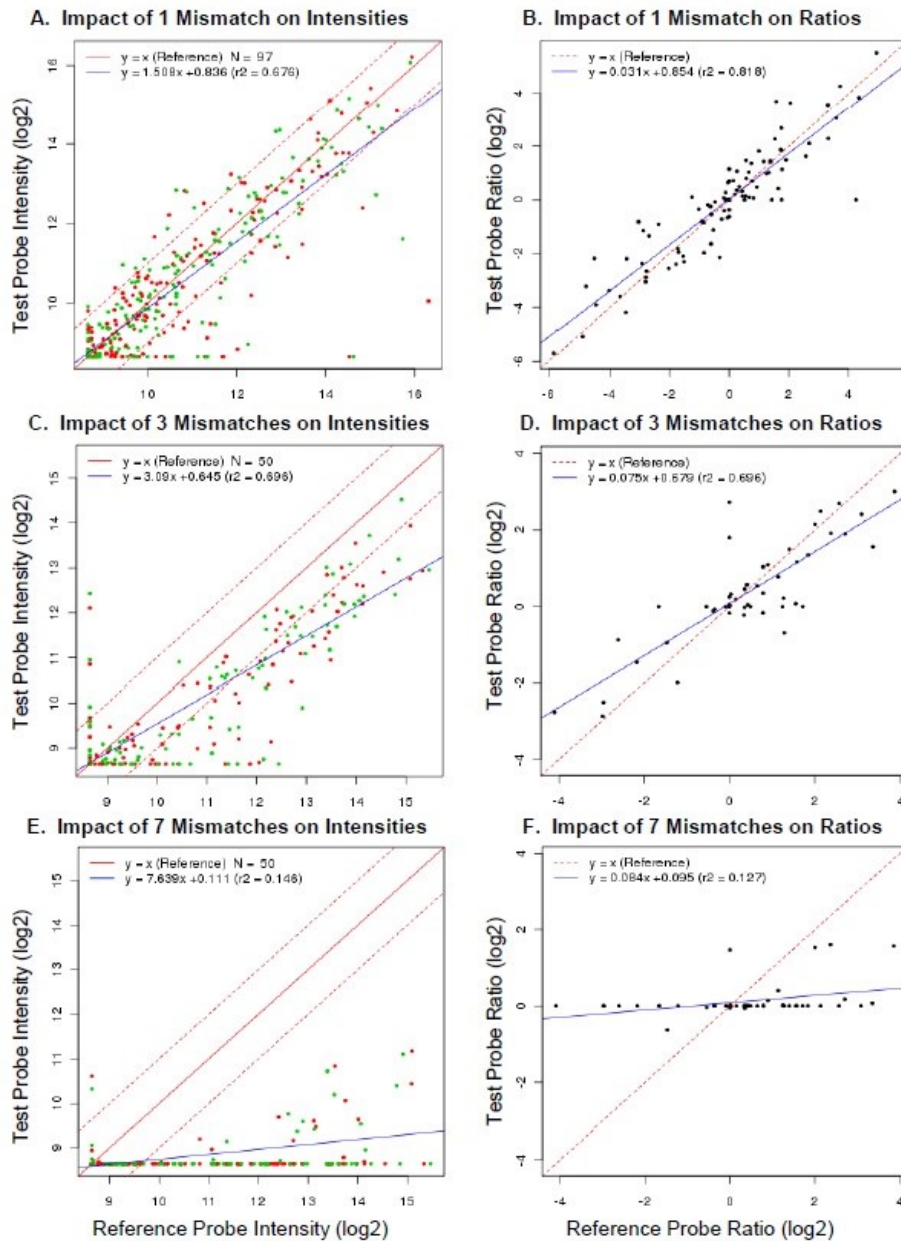


Figure S2.1 Effect of SNPs on hybridization signal intensities and differential expression ratios

Hybridization data were based on five biological replications of each white spruce tissue tested, and two technical replicates (dye swaps) were used for each sample. Each data point represents the mean value for the five biological replicates. For probe intensities (A, C, E), the data are based on hybridizations with total RNA from secondary xylem; each point represents the mean value data for Alexa Fluor 555 (green) or Alexa Fluor 647 (red). The ratios (B, D, F) were obtained from pair-wise comparisons of secondary xylem and young needles; each dot represents the mean ratio obtained from the dye-swaps of all five biological replicates.

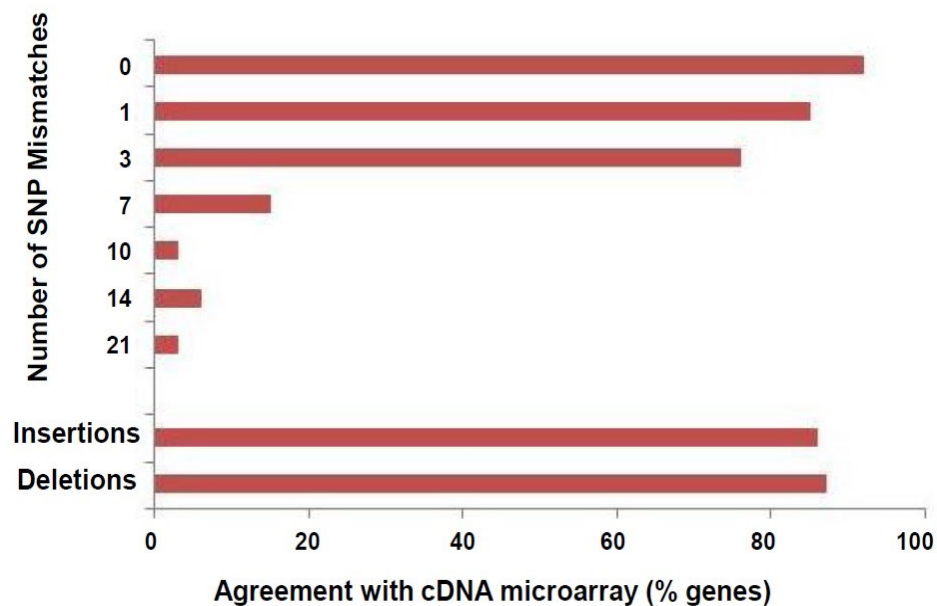


Figure S2.2 Comparison of differential expression results from a cDNA microarray and the test oligonucleotide microarray

Hybridization data were based on five biological replications of each white spruce tissue tested, and two technical replicates (dye swaps) were used for each sample. Tissue preferential expression was determined as described (Pavy et al. 2008) for secondary xylem and young needles. The outcomes of the two types of arrays were compared by assessing the presence or absence of statistically significant tissue preference.

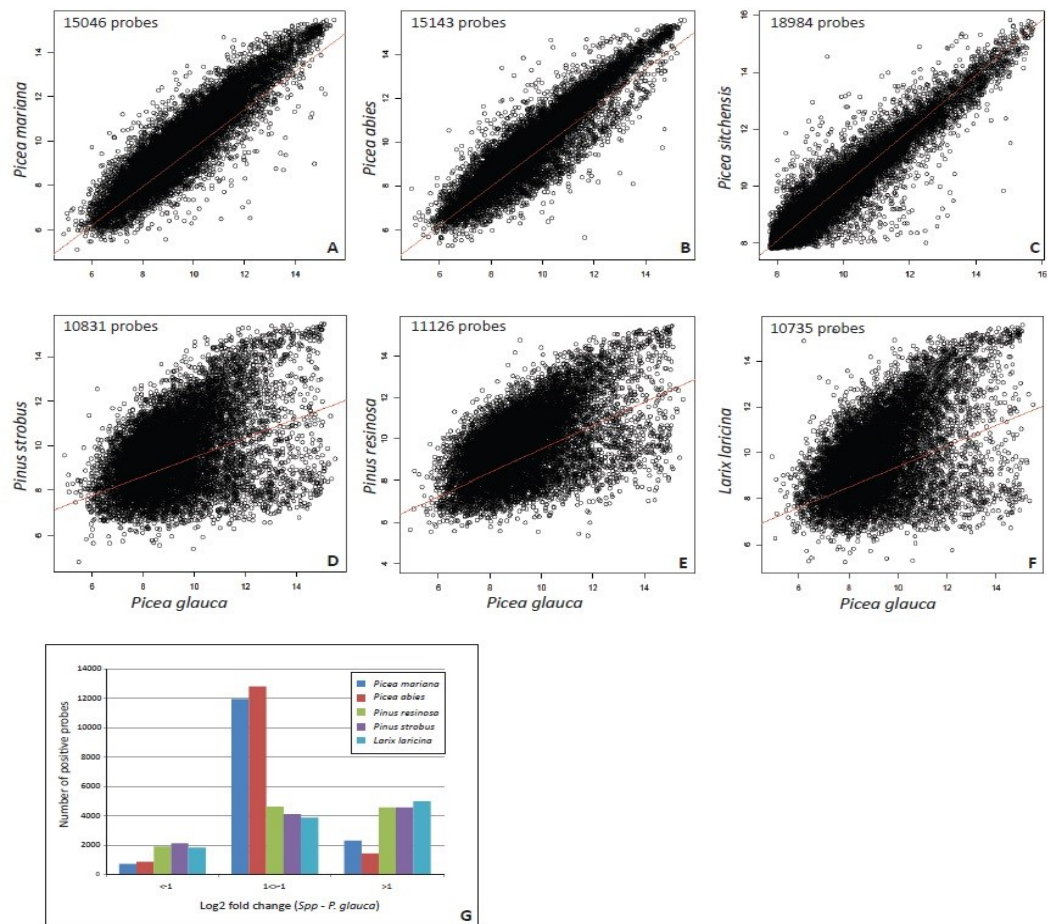


Figure S2.3 Interspecific comparison of hybridization intensities in the phelloderm

A-F: Pair-wise comparison of white spruce and six other species based on the number of shared positive probes indicated in the plots. The squared correlation coefficients (r^2) are as follows 0.83 (A), 0.84 (B), 0.90 (C), 0.18 (D), 0.30 (E) and 0.18 (F). G: Analysis of signal intensity variation between species; the fold change (FC) was determined from the average normalized signal intensities (log2 scale). An FC of 1 or -1 represents a two-fold signal increase or decrease, respectively.

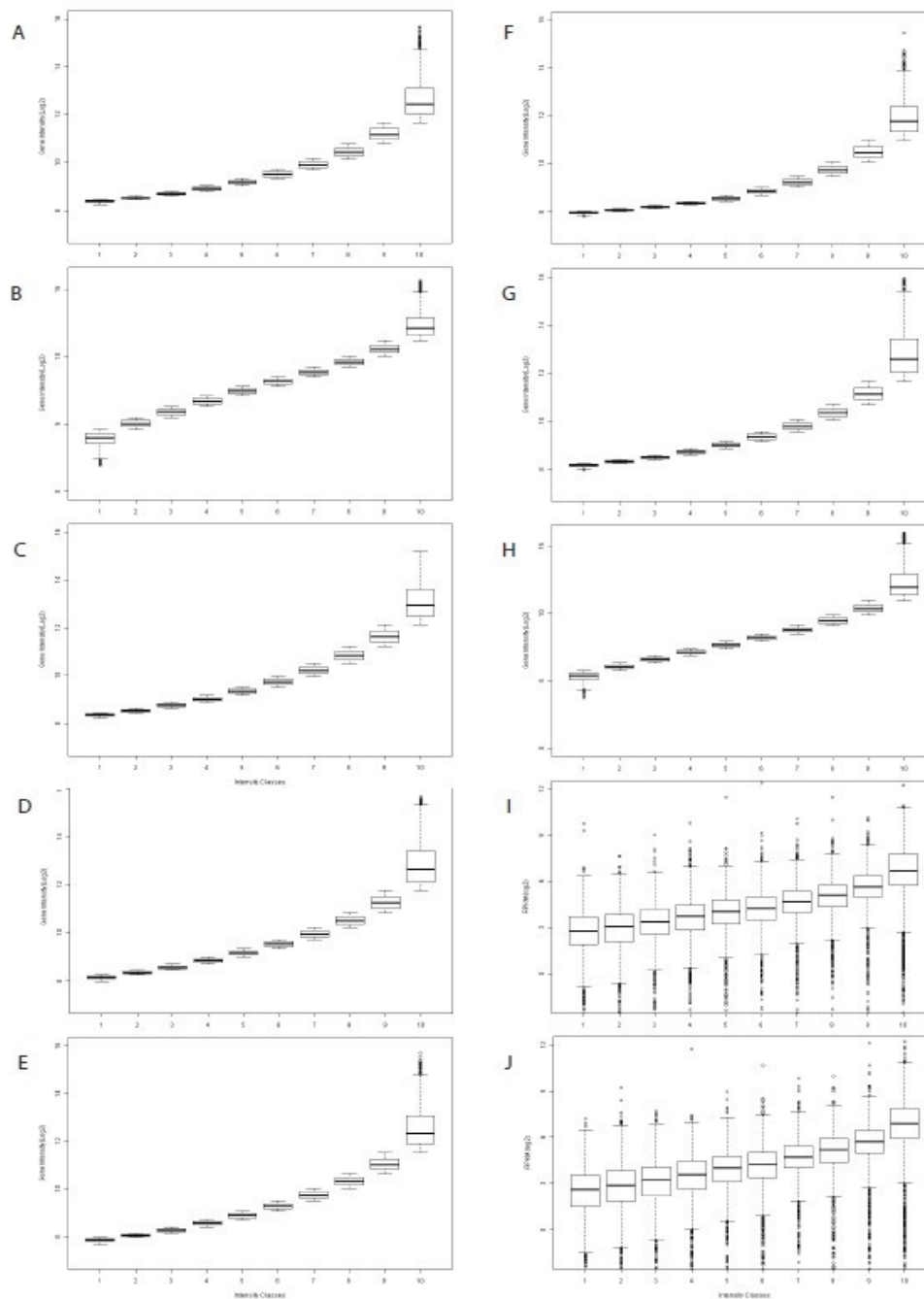


Figure S2.4 Hybridization signal intensities of genes in each of the 10 expression classes in each tissue in *Picea*GenExpress transcript profiles

Vegetative buds (A), megagametophytes (B), xylem from mature trees (C), phelloderm from juvenile trees (D), xylem from juvenile trees (E), embryogenic cells (F), needles (G) and roots (H). RPKM values from RNA-sequencing of phelloderm (I) and xylem (J) of juvenile trees are also presented

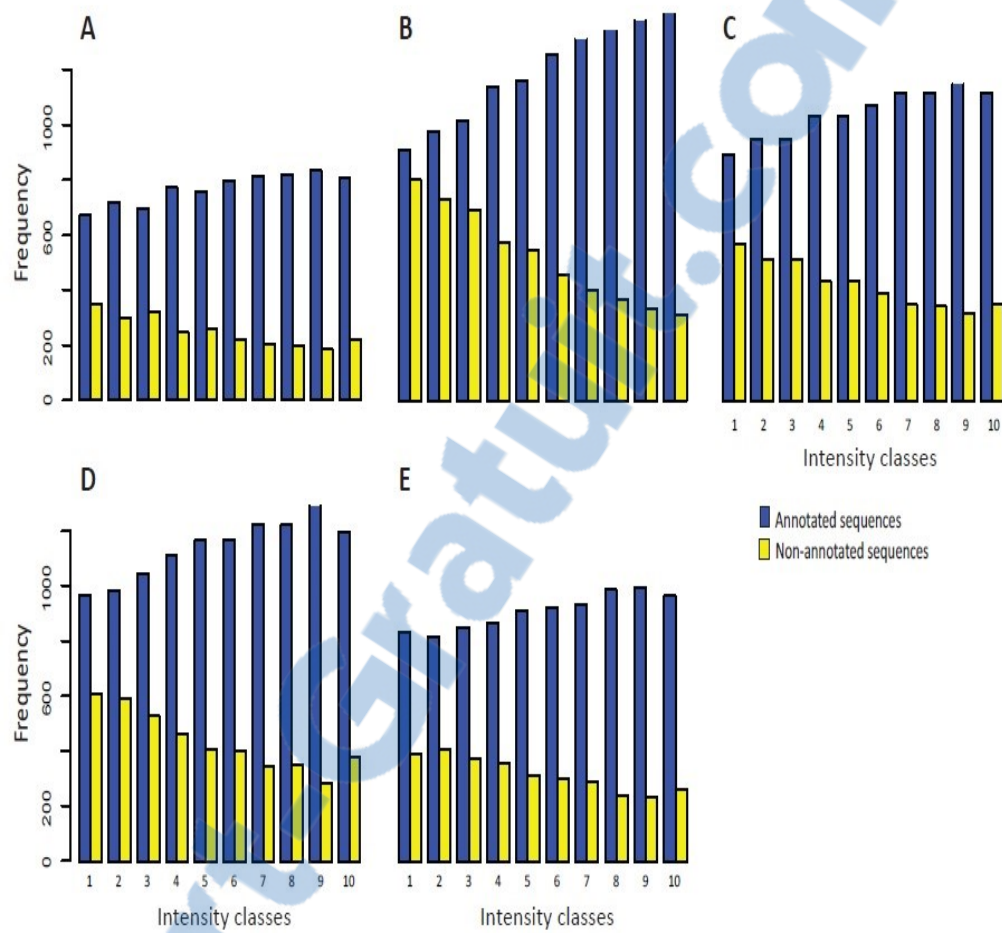


Figure S2.5 Expression classes and numbers of tissue of annotated and non-annotated sequences

A-E: Number of annotated and non-annotated sequences per expression class for embryogenic cells (A), megagametophytes (B), xylem from mature trees (C), phelloderm (D) and vegetative buds (E).

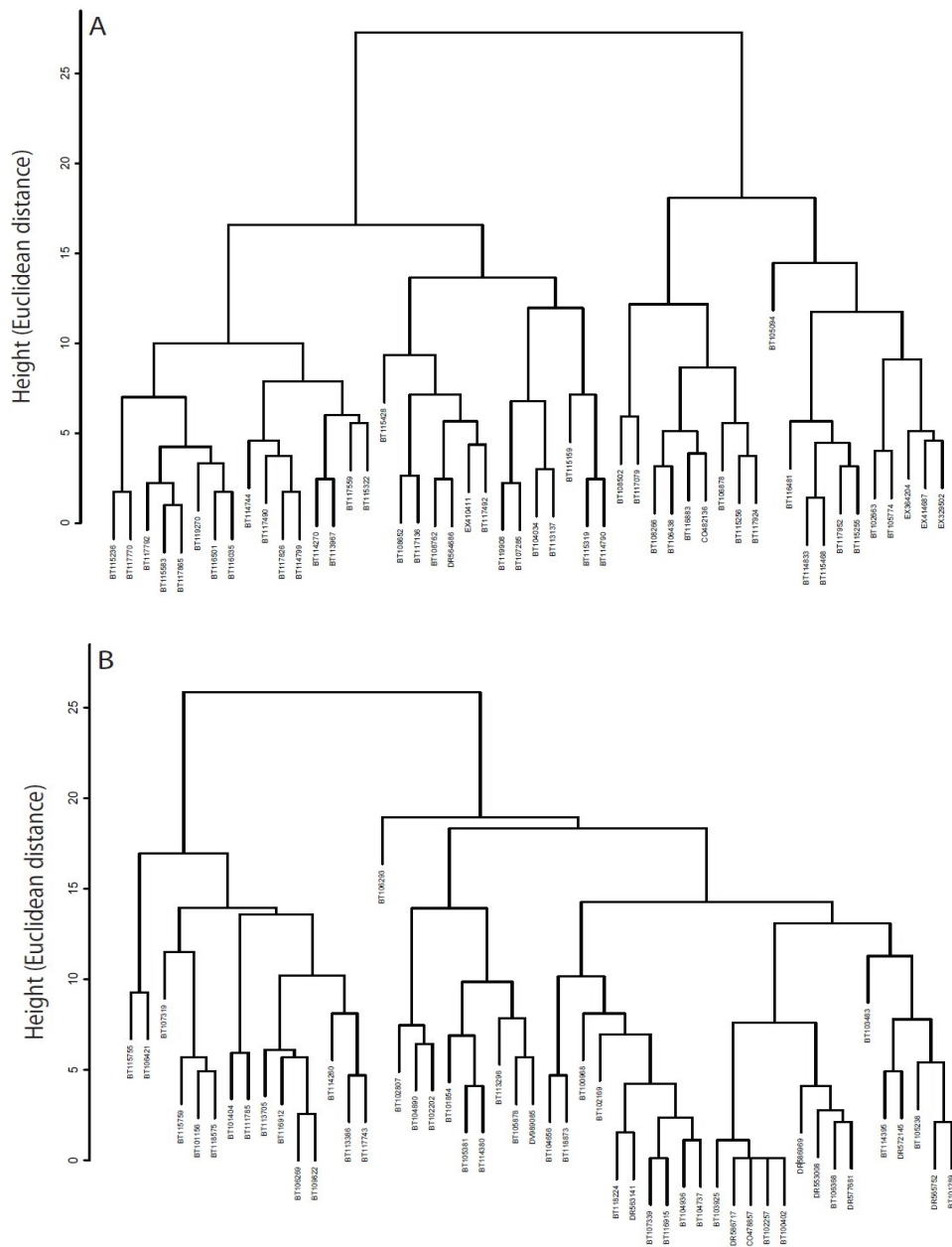


Figure S2.6 Hierarchical clustering dendrograms of gene expression within two osmotic regulation protein families.

A) dehydrins, and B) late embryogenesis abundant (LEA) proteins. Each leaf node of the dendrograms corresponds to an individual gene, and each node (horizontal line) represents a gene cluster. A gene cluster is composed of individual genes or existing gene cluster with the fusion point. Each gene cluster was placed at a height level as shown on the vertical axis. Height values refer to the similarity/distance measures between genes and gene clusters.

Table S2.1 and Table S2.2 are not shown here but available in xls format at <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3534630/>

Table S2.1 PiceaGenExpress transcript profiles.

Table S2.2 Summary statistics of PiceaGenExpress transcript profiles.

Table S2.3 Expression of cDNA selected for test microarray

Sequences selected from cDNA microarray studies	Tissue preferential transcript accumulation			Intensity of transcripts detected in secondary xylem of mature trees			
	Needle	Xylem	None	Total	Strong	Moderate	Low
Total number of cDNAs selected	295	295	410	512	109	126	277
%	30%	30%	41%	51%	21%	25%	54%
cDNA represented by an amplicon on test microarray	29	29	38	50	11	12	27

Chapitre 3.

Modular organization of the white spruce (*Picea glauca* (Moench) Voss) transcriptome reveals functional organization and evolutionary signatures

[Raheison E, Rigault P, Giguère I, Caron S, Mebarek Lamara & MacKay J (2015). Modular organization of the white spruce (*Picea glauca* (Moench) Voss) transcriptome reveals functional organization and evolutionary signatures. *New Phytologist*] doi : 10.1111/nph.13343]

3.1 Résumé

- Le profilage transcriptionnel a montré les bases moléculaires de plusieurs processus biologiques dans les plantes mais rares soient les études qui ont développé une compréhension sur la variation transcriptomique globale. Nous avons étudié la structure du transcriptome de l'épinette blanche afin de délimiter l'organisation modulaire, la fonction et les caractères évolutifs du transcriptome.
- Des analyses utilisant des puces à ADN ont été effectuées afin de i) faire l'identification et la caractérisation fonctionnelle des groupes de gènes co-exprimés; ii) faire l'étude de la diversité de profils et de fonctions des gènes dont la préférence d'expression au tissu vasculaire est conservée entre les espèces des *Picea*; iii) faire l'identification des réseaux d'expression reliés à la formation du bois.
- Nous avons classé par profilage sur plusieurs tissus végétatifs 22 857 gènes comme variables (79%, 22 groupes de coexpression) ou invariables (21%). L'organisation modulaire et la restructuration complexe du transcriptome entre les gènes préférentiels des tissus vasculaires ont été révélées par l'attribution à des groupes de coexpression qui ont des profils qui se chevauchent partiellement et des fonctions partiellement distinctes. Des analyses intégrées sur des tissus et des profils spatialement et temporellement variables ont identifié des réseaux de gènes du xylème secondaire, montré leur changement pendant une saison de croissance et identifié le facteur de transcription PgNAC-7 (facteur de transcription *no apical meristem* (NAM), *Arabidopsis transcription activation factor* (ATAF) et *cup-shaped cotyledon* (CUC) 007 dans *P. glauca*) comme un gène central majeur spécifique à la formation du bois initial.
- Le profilage référentiel a identifié des groupes de coexpression complets et statistiquement robustes qui révèlent que l'organisation modulaire est à la base de la conservation évolutive de la structure de transcriptome.

3.2 Abstract

- Transcript profiling has shown the molecular bases of several biological processes in plants but few studies have developed an understanding of overall transcriptome variation. We investigated transcriptome structure in white spruce (*Picea glauca* (Moench) Voss) aiming to delineate its modular organization and associated functional and evolutionary attributes.
- Microarray analyses were used to: i) identify and functionally characterize groups of co-expressed genes; ii) investigate expressional and functional diversity of vascular tissue preferential genes which were conserved among *Picea* species; iii) identify expression networks underlying wood formation.
- We classified 22857 genes as variable (79%, 22 coexpression groups) or invariant (21%) by profiling across several vegetative tissues. Modular organization and complex transcriptome restructuring among vascular tissue preferential genes was revealed by their assignment to coexpression groups with partially overlapping profiles and partially distinct functions. Integrated analyses of tissue-based and temporally variable profiles identified secondary xylem gene networks, showed their remodelling over a growing season and identified the transcription factor PgNAC-7 (no apical meristem (NAM), *Arabidopsis* transcription activation factor (ATAF) and cup-shaped cotyledon (CUC) transcription factor 007 in *P. glauca*) as a major hub gene specific to earlywood formation.
- Reference profiling identified comprehensive, statistically robust co-expressed groups revealing that modular organization underpins the evolutionary conservation of the transcriptome structure.

3.3 Introduction

Transcriptome analysis is among the most widely used large-scale experimental approaches to investigate genome function and to help annotate genome sequences (e.g. Pickrell *et al.*, 2010; Prasad *et al.*, 2013; Ellison *et al.*, 2014). The transcriptome as a whole is highly dynamic owing to quantitative variations over spatial and temporal scales. The observed changes in transcript levels result from the combinatorial action of genetically encoded regulatory pathways that orchestrate development and responses to environmental stimuli. The accumulation of individual genes transcripts may be viewed as a bridge between the genotype and the phenotype and as such, may be acted upon through natural selection (Romero *et al.*, 2012; Prasad *et al.*, 2013).

The fundamental interest in and application of transcriptome analysis are reflected by its broad application in non-model organisms including many taxa such as conifer trees where large genome size or complexity represents a technical barrier to genomic investigation. Conifer tree genomes are highly repetitive and with sizes of 18 to 30 Gb, they are several folds larger than most angiosperm plant genomes sequenced to date including *Arabidopsis thaliana* (GI, 2000), black cottonwood (*Populus trichocarpa*; Tuskan *et al.*, 2006), maize (*Zea mays*; Schnable *et al.*, 2009), and eucalyptus (*Eucalyptus grandis*; Myburg *et al.*, 2014) as well as animal genomes such as Human (IHGSC, 2004). Three conifer genomes were recently sequenced and include white spruce (*Picea glauca*; Birol *et al.*, 2013), Norway spruce (*P. abies*; Nystedt *et al.*, 2013), and loblolly pine (*Pinus taeda*; Neale *et al.*, 2014). In contrast, cDNA and transcriptome sequencing in conifers has been ongoing for over 15 years (e.g. Allona *et al.*, 1998), has led to many functional investigations, and produced outcomes that enable genetic mapping research, association studies and investigations of evolutionary adaptation (MacKay *et al.*, 2012). Despite this extensive body of work, a general framework describing transcriptome organization in conifers is lacking and would help to develop a more unified understanding of genome function.

Conifer transcriptome studies have been conducted to gain insight into the molecular underpinnings of growth, development and stress responses. Conifers are long lived and need to cope with diverse biotic and abiotic stress factors over time, which has led to a number of studies on the theme of resistance and acclimation. For examples, Sun *et al.* (2011) identified genes responsive to fungal elicitation in scots pine (*Pinus*

sylvestris) while Verne *et al.* (2011) identified up and down regulated genes in white pine weevil resistant individuals in interior spruce (*P. glauca* x *Picea engelmannii*). In the context of global environmental change, Yeaman *et al.* (2014) investigated patterns of gene expression related to acclimation across environmental conditions in lodgepole pine (*Pinus contorta*) and interior spruce; and Baldi *et al.* (2011) identified candidate genes for cold tolerance in cypress (*Cupressus sempervirens*). Considering the ecological and economic importance of conifer woods, another area of interest is transcriptome variation and reorganization during secondary xylem growth. Genes with significant changes in expression between earlywood and latewood formation were identified in loblolly pine trees from two seed sources (Yang & Loopstra, 2005), in maritime pine (*Pinus pinaster*) during a growing season (Paiva *et al.*, 2008), and in radiata pine (*P. radiata*) at different stages of developmental aging (Li *et al.*, 2010). A few studies have addressed related evolutionary aspects by identifying genes that are conserved in expression patterns and sequence between species such as white spruce and *A. thaliana* (Pavy *et al.*, 2008), among different spruce species (Raheison *et al.*, 2012) and between pine and spruce (Yeaman *et al.*, 2014).

A majority of transcriptome profiling experiments in conifers and plants in general have aimed to identify and functionally describe expression variations in subsets of genes associated with biological processes or genotypic differentiation (*e.g.* Lorenz *et al.*, 2011; Palle *et al.*, 2011). Genes whose transcript levels do not vary across samples have been less investigated, despite their potentially key housekeeping roles (Aceituno *et al.*, 2008; Chang *et al.*, 2011; Eisenberg & Levanon, 2013). Furthermore, many studies have focused on a single or a few tissues or organs; however, recent studies in flowering plants show that tissue differentiation is the major driver of transcriptome reorganization in species such as *A. thaliana* (Ma *et al.*, 2005), maize (Downs *et al.*, 2013), barley (*Hordeum vulgare*; Downs *et al.*, 2013), and tobacco (*Nicotiana tabacum*; Druka *et al.*, 2006). The conservation of tissue-specific gene expression patterns is also well-established in vertebrates including mammals (Prasad *et al.*, 2013) and non-mammals (Chan *et al.*, 2009). Taken together, these observations indicate that studies considering several tissues are a powerful approach to delineate transcriptome organization across a whole organism. Furthermore, the rate of gene expression evolution can differ depending on tissue or organ sources in angiosperms (Yang & Wang, 2013) and in mammals (Brawand *et al.*, 2011).

We investigated transcriptome structure by interrogating 23853 genes on a custom designed microarray profiling platform (Raheison *et al.*, 2012) in two different controlled experiments with biological replications. We have previously shown a high level of concordance between results obtained with this microarray and RNA-Seq data (Raheison *et al.*, 2012). Other recent studies used RNA-Seq profiling (*e.g.* Yeaman *et al.*, 2014) and were able to interrogate nearly identical numbers of genes (23889 in spruce and 23519 in pine).

Our objectives were threefold:

- Characterize the organization of the white spruce transcriptome based on multiple tissue data generated from a single experiment enabling statistical model analyses and the identification of high-confidence coexpression groups. This led us to define and functionally analyze expression groups including invariant and variable genes across vegetative tissues of white spruce;
- Investigate expressional and functional diversity of vascular tissue preferential genes which were conserved among *Picea* species
- Delineate gene coexpression networks and identify genes involved in secondary cell wall development during wood formation, and study network reorganization over the course of a growth season.

3.4 Materials and methods

We used microarray datasets from two independent experiments (experiments 1 and 2) using the same microarray platform described in Raheison *et al.* (2012). These experiments are described in the plant material section. We also used gene expression results in shoot secondary xylem and phelloderm tissues of white (*Picea glauca* (Moench) Voss), black (*P. mariana* (M) Britton) and Sitka (*P. sitchensis* (Bong.) Carrière) spruces previously described in Raheison *et al.* (2012). Unless specified otherwise, all statistical analyses were performed using the base R packages; the Benjamini and Hochberg method was employed where appropriate to correct sets of *P*-values for multiple comparisons (Benjamini & Hochberg, 1995) using the *multtest* R package; and hypergeometric test for under or overrepresentation analysis.

3.4.1 Plant material

We used young spruce trees from open-pollinated seed lots that were grown for 6 – 16 wk in a greenhouse under natural light conditions. Tissue samples were collected



either at one (experiment 1) or at four different time points (experiment 2) in a growing season. In experiment 1, in July we collected from three-year-old seedlings of white spruce seven vegetative tissues: shoot apex, needles of young branches (young foliage), shoot or root secondary xylems, shoot or root phelloderms, and root tips (Fig. 3.1a).

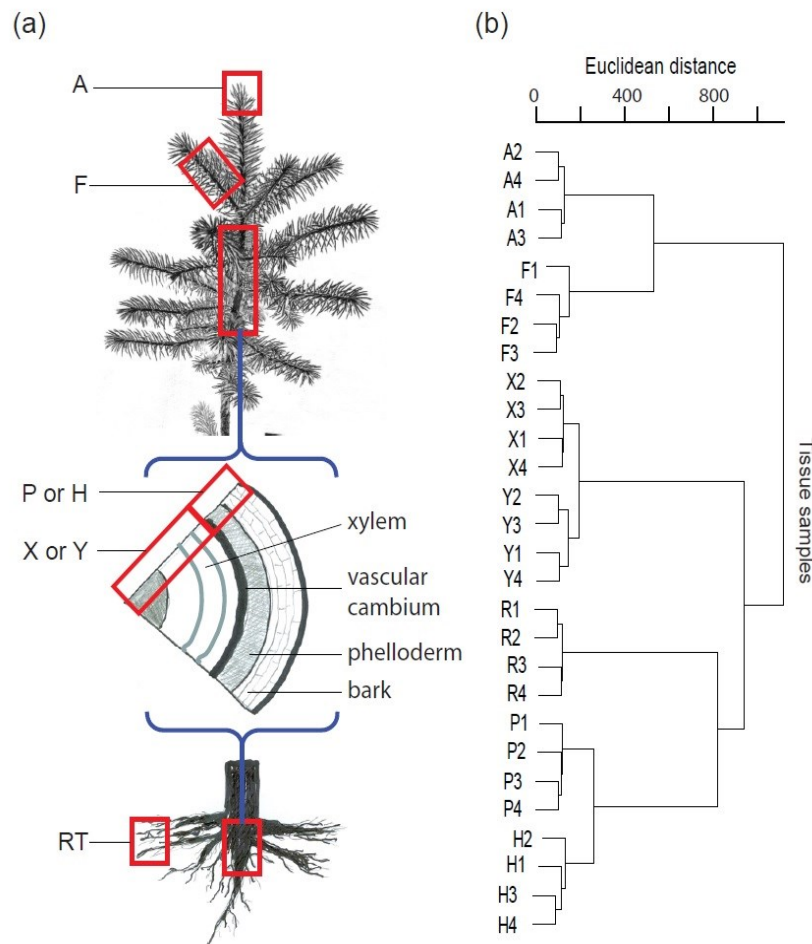


Fig. 3.1 Classification of collected tissues based on gene expression

(a) Tissue sampling (image provided by Dr A Séguin); (b) hierarchical clustering of white spruce (*Picea glauca*) tissues based on microarray data ($\log_2$ scale). Analysis used the *hclust* function of R with Ward's method (Ward, 1963) and based on expression data of detected genes. A, shoot apex; X, shoot secondary xylem; Y, root secondary xylem; P, shoot phelloderm; H, root phelloderm; F, young foliage; and R, root tips; each tissue was replicated four times (labelled 1-4)

The shoot apex sample was about 1 cm in length and contained the shoot tip with the apical bud as well as primary needles. Secondary xylem and phelloderm tissues lied inside and outside of vascular cambium, respectively (Raherison *et al.*, 2012). Four biological replicates were used for each tissue sample, and each replicate was a pool of samples from five independent trees. In experiment 2, shoot secondary xylem tissues of four-year-old seedlings of Norway and white spruces were sampled on June 22, July 13, August 17 and September 21. Samples were collected from six individual trees for each time point. All of the samples were immediately frozen in liquid nitrogen and stored at -80°C until RNA isolation.

3.4.2 RNA isolation, labelling and hybridization

Total RNA was isolated from tissue samples by using the CTAB extraction method of Chang *et al.* (1993) as described in Pavy *et al.* (2008). Total RNA of each sample replicate was amplified using the Amino Allyl MessageAmp II aRNA amplification kit (Ambiom by Life Technologies, Carlsbad, CA, USA) following the manufacturer's instructions. In experiment 1, dye-swap technical replicates for each biological replicates were performed. Each sample of amplified RNA (aRNA) were paired and labelled with Alexa Fluor 555 and 647 dyes (Invitrogen, Carlsbad, CA, USA). In experiment 2, aRNAs were labelled with Alexa Fluor 555 dye. RNA amplification, aRNA labelling, hybridization, array scanning and image processing were performed as described in Raherison *et al.* (2012).

3.4.3 Microarray pre-processing and statistical analyses

In Experiment 1, intensity data from single dyes of the two-dye experiment were analyzed separately. Replacement of abnormal or flagged spots, data normalization, and identification of positive probes and detected genes were performed as described in Raherison *et al.* (2012). A total of 24 850 probes corresponding to 23 623 unique white spruce genes were detected at least in one tissue (Table 3.1, Table S3.1). Using the *nlme* R package, we applied to detected genes a mixed analysis of variance (ANOVA) model with tissue, dye and interaction tissue x dye as fixed effects; and interaction tree x tissue as a nested random effect. A set of 846 genes were not included in the ANOVA because their expression levels were non-normally distributed (Anderson-Darling's test; *nortest* R package), had non-homogenous variances (Levene's test; *car* R package), or were affected by the factor dye or interaction tissue x dye. We identified genes that had changes (variable genes,

adjusted P -value < 0.05) or were unchanged (invariant genes, adjusted P -value > 0.05) across expression tissues from the remaining detected genes. These remaining detected genes were also used for further analyses. In experiment 2, flagged spots of expression data were removed from analysis. Data normalization and identification of positive probes were carried out as described in Raheison *et al.* (2012). A two-way ANOVA with species and time as factors was performed to identify the 1366 temporally variable genes or genes differentially expressed in secondary xylems across time points in a growing season. Temporally variable genes were those with one or more statistically significant probes (adjusted P -value < 0.05). The average expression values in two dye technical replicates for each tissue samples (experiment 1) or six replicates in two species of each time point (experiment 2) were used for Weighted Correlation Network Analysis (WGCNA R package; Langfelder & Horvath, 2008), template matching (Harris *et al.*, 2009), correlational and hierarchical clustering analyses. In WGCNA, we used soft-thresholding power values of 14 and 18 for the experiments 1 and 2, respectively to identify expression modules. An expression module is a cluster of highly (positively and negatively) correlated genes. It reveals two negatively correlated expression profiles. Each expression profile is associated with a group of genes that we referred to as a coexpression group in experiment 1 or temporally variable cluster in experiment 2 (Table S3.1). The WGCNA script used is included in the Method S1. The raw and preprocessed data have been deposited in the National Center for Biotechnology Information - Gene Expression Omnibus (NCBI - GEO) with accession numbers GSE60277 and GSE51884 for experiments 1 and 2, respectively.

3.4.4 Functional annotation analysis

The cDNA sequences from which the oligonucleotide probes were designed (Raheison *et al.*, 2012) were annotated against the *A. thaliana* protein dataset (The Arabidopsis Information Resource 10) using the BLASTX (E -value $< 10^{-10}$). Functional classes presented on Table S3.2 were developed as follows: firstly, we identified and clustered enriched biological process gene ontology (GO) terms using DAVID (Database for Annotation, Visualization and Integrated Discovery, enrichment at P -value < 0.05 , Huang *et al.*, 2009); secondly, enriched GO terms were classified into functional categories of plant GO slim using cateGORizer (Hu *et al.*, 2008); and finally based on the GO term itself, its cluster and its functional category,

each GO term was manually assigned to a consensus functional class. Six sequences that we analyzed represented *P. glauca* MYB proteins. These sequences were named in this report: BT112255: PgMYB29; BT117714: PgMYB30; BT119291: PgMYB31; DR571012: PgMYB32; BT108182: PgMYB33; and BT106711: PgMYB34.

Table 3.1 Numbers of genes detected and classification as variable and invariant based on transcript levels

Category	Total genes	Annotated genes (%)³
Interrogated genes	23 853	15 724 (66)
Detected genes ¹	23 623	15 586 (66)
Statistically tested genes (ANOVA) ²	22 857	15 096 (66)
Invariant genes	4805	2410 (50)
Variable genes	18 052	12 726 (70)
High-confidence genes (adjusted <i>P</i> -value < 0.0001)	10 548	7651 (73)
Low-confidence genes (0.05 < adjusted <i>P</i> -value > 0.0001)	7504	5075 (68)

¹Transcripts were detected above background levels in at least one tissue of white spruce (*Picea glauca*); ²Genes that met assumptions for ANOVA testing (for details, see methods); ³Genes that matched an *Arabidopsis* sequence (BLASTX, *E*-value < 10⁻¹⁰).

3.4.5 Reverse transcription quantitative PCR (RT-qPCR)

We performed RT-qPCR on 28 genes to verify results from microarray analyses (Table S3.3). Gene specific primers were designed with Primer3Plus (Untergasser *et al.*, 2007), evaluated with OligoCalc (Kibbe, 2007) and checked for uniqueness in the white spruce gene catalog (Rigault *et al.*, 2011). The same RNA samples that were used for microarray experiments were used for reverse transcription and RT-qPCR as described in Boyle *et al.* (2009) by using the QuantiFast SYBR Green PCR Kit (Qiagen, Carlsbad, CA, USA) with 5 ng of cDNA in a final volume of 15 µL. The number of molecules was determined following the method of Rutledge & Stewart (2008) with modifications described in Boyle *et al.* (2009). A one-way ANOVA (for the factor tissue) and the template matching method were used to determine if results from microarray and RT-qPCR analyses were statistically matched with regard to gene category and expression profile, respectively.

3.5 Results

3.5.1 The expression of most white spruce genes vary significantly across vegetative tissues

We analysed transcripts levels of 23 853 unique genes in seven vegetative tissues during normal development of white spruce (experiment 1). These tissues represent above-ground (apex, young foliage or needles, shoot secondary xylem, and shoot phelloderm) and below-ground (root secondary xylem, root phelloderm and root tips) organs. Microarray analysis of each tissue sample comprised four biological replicates that were analysed separately. Dye-swap technical replicates were also used. We identified 18 052 genes whose expression level varied significantly across tissues (variable genes) and 4805 whose expression level did not vary across tissues (invariant genes) (Table 3.1, Table S3.1, S 3.4). A total of 10 548 variable genes were identified as high confidence (adjusted P -value < 0.0001) and 7504 genes ($0.0001 < \text{adjusted } P\text{-value} < 0.05$) as low confidence (Table 3.1, Table S3.1, S3.4).

3.5.2 The white spruce transcriptome is organized into coexpression groups classifying tissues based on their types or physiological functions

As a first step to explore expression data and transcriptome organization, we used hierarchical clustering. Biological samples of each tissue were clustered within a single unique group, thus indicating high repeatability of transcript level determinations as well as uniformity among samples of each type (Fig. 3.1b). Identical results were obtained when analysing either the whole set of detected genes (Fig. 3.1b) or only the set of variable genes (Fig. S3.1). The clustering clearly formed groups according to tissue type. Root xylem clustered closely with shoot xylem, and root phelloderm with shoot phelloderm, thus indicating a strong similarity between tissues from above- and below-ground organs. The cluster mostly similar to foliage (*i.e.* needles) was the shoot apex, an organ that contains primordial leaves.

To further characterize the organization of the white spruce transcriptome, we clustered variable genes and identified 22 coexpression groups or sets of genes sharing a similar expression profile ranging in size from 48 to 1929 genes (Fig. 3.2a, Fig. S3.2, Table S3.1, Table S3.4). Gene clustering to delineate the coexpression groups proceeded in two steps: i) we clustered high-confidence genes into 11 modules each containing two coexpression groups with the *WGCNA* package of R and defined

the expression profile associated to each of the groups; ii) we used the resulting average expression level within each of the 22 groups as templates to determine membership of low-confidence genes within the coexpression groups (adjusted P -value < 0.05). In this study, expression modules (M1-M11) are comprised of genes whose expression was positively or negatively correlated, whereas members of a coexpression group had only positively correlated profiles (M1a-M11b, 22 groups in total), as shown in heatmaps (Fig. 3.2a, Fig. S3.2a-j). A principal component analysis (eigengene modules) was used to show the differentiation of tissues within each expression module (Fig. 3.2, Fig. S3.2). Eigengene groups were used to study correlation between groups and tissues (Fig. 3.2b, Fig. S3.2k) and between tissues (Fig. 3.2c).

Our results indicate a modular organization of the transcriptome that is linked to differentiation between tissue types or physiological functions (Fig. 3.2a, b, Fig. S3.2, Table S3.4). For example, four coexpression groups containing 42% of the variable genes were positively correlated with a single tissue (average Pearson correlation coefficient $r > 0.7$, shoot apex in M4a, M8b, M9a and M11b; and needles in M5b) or tissue type ($r > 0.6$; secondary xylem from shoots and roots in M2a, M7b and M9b; and phelloderm from shoots and roots in M6a and M10a). The profile of M1a group comprising 11% of the variable genes was positively correlated with photosynthetic tissues ($r = 0.8$).

The identification of coexpression groups also supports a strong similarity between tissues from above- and below-ground organs. For example, fourteen profiles comprising 70% of variable genes closely clustered xylem samples from the shoot and the root, and similar phelloderm samples from the shoot and the root (M2-M7 and M10). In contrast, only 21% of variable genes were separately clustered for vascular tissues from the root and the shoot (M1 and M11). Analysis based on eigengene groups revealed highly positive correlation between vascular tissues from the shoot and the root (Fig. 3.2c).

To confirm microarray analysis results, we determined and statistically analyzed transcript levels of selected genes using RT-qPCR, including 24 variable genes and four invariant genes (Table S3.3). Results were strongly consistent between the two methods, with 92% of tested genes having similar expression profile and similar category (invariant and variable genes) (Table S3.3).

3.5.3 Gene ontology term enrichments show functional diversity among coexpression groups

Functional analysis was performed for the 22 coexpression groups and invariant genes based on annotations and enrichment for biological process gene ontology (GO) terms for genes. Most of the detected genes (66%) were similar to an *A. thaliana* gene sequence (Table 3.1) and were annotated accordingly. The proportion of white spruce sequences that matched an *A. thaliana* sequence was somewhat higher among variable genes (70% for the high-confidence group) than for the invariant genes (50%; Table 3.1). We found 4324 genes associated with enriched GO terms, representing 50% of the genes with assigned GO term annotations (Fig 3.3a, Fig. 3.3, Table 3.2).

The GO term enrichment results showed that the dominant enriched biological processes varied across the coexpression groups (Fig. 3.3b). In addition, strong or preferential expression in a tissue type may be associated with several coexpression groups from different modules, and thus with enrichment of different processes. For example, group M6a with genes strongly expressed in phelloderm tissues compared to all other tissues was enriched for stress response. Group M10a was also highly expressed in phelloderm tissues but contrasted with M6a through stronger expression in needles; it was enriched for terms related to lipid/carbohydrate metabolic process. Observations such as these show how assignment to a coexpression group is a useful first step for developing hypothesis for different groups of phelloderm preferential genes. Other similar examples were observed for groups of genes with strong expression in apex tissue (M4a, M8b, M9a and M11b), or in xylem tissues (M2a, M7a and M9b). These findings revealed the diversity of functions among genes preferentially expressed in a tissue type.

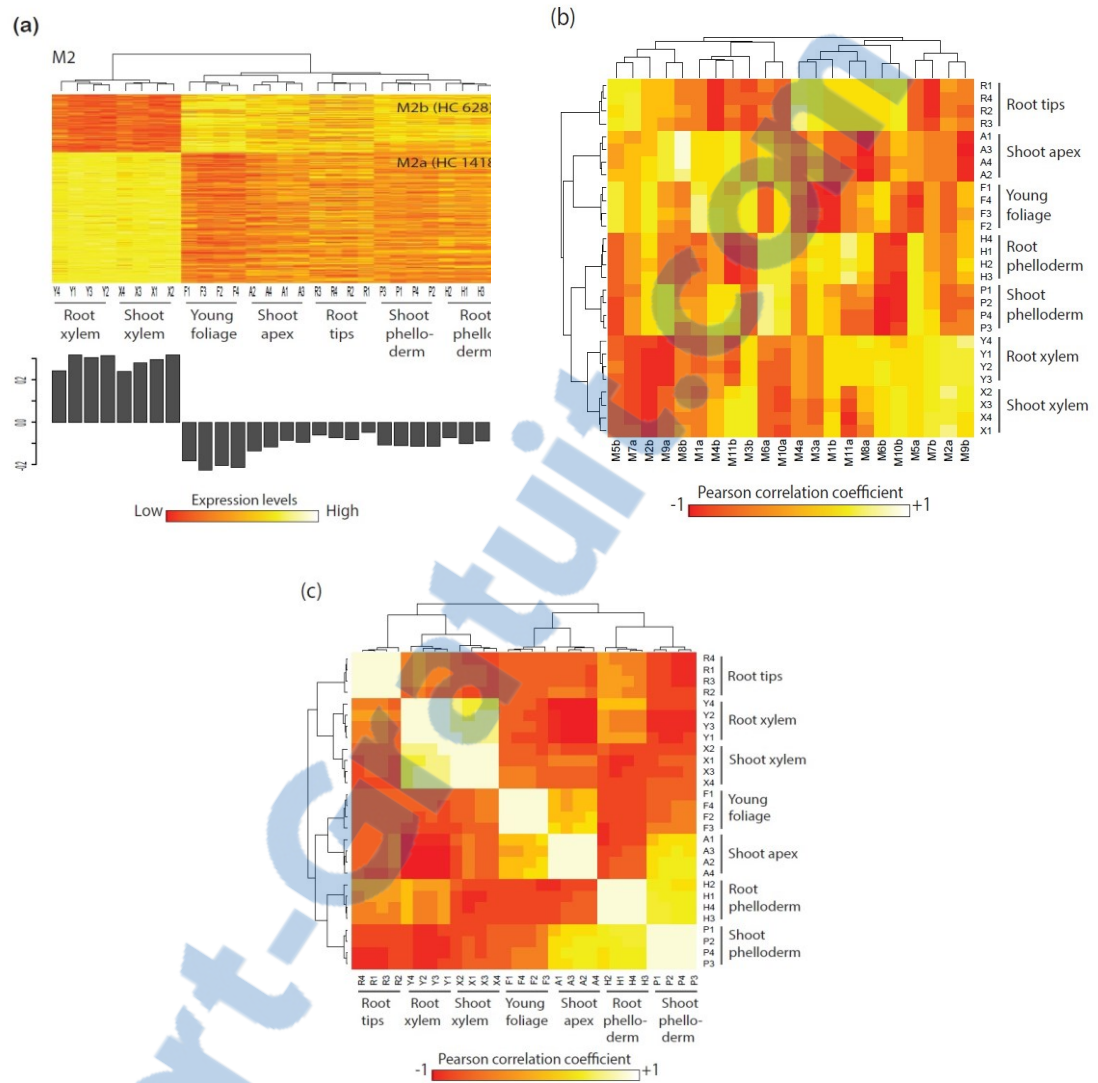


Fig. 3.2 Expression module and coexpression groups across tissues in white spruce (*Picea glauca*)

(a), Heatmaps show transcript abundance (log₂ scale) of high-confidence variable genes in the M2 expression module and M2a-M2b coexpression groups determined with the WGCNA package of R; Number in parentheses, number of high-confidence (HC) variable genes in the coexpression group; Total and low-confidence variable gene numbers are listed in Table S3.4; Rows (Y axis), genes which are listed in Table S3.1; Columns (X axis), each of the four replicates of tissues (A1-4, shoot apices; X1-4, shoot secondary xylems; Y1-4, root secondary xylems; P1-4, shoot secondary phelloderms; H1-4, shoot secondary phelloderms; F1-4, young foliages; and R1-4, root tips); Bar plots display the eigengene modules (or the first principal components); See Fig. S3.2 for graphically representation of other expression modules and coexpression groups; (b), Correlation between coexpression groups (M1a-M11b) and tissues based on eigengene groups; (c) Correlation between tissues based on eigengene groups.

3.5.4 Conifer-specific genes are more strongly associated with secondary meristematic tissues

Functional and structural differences have been extensively documented in secondary vascular tissues when comparing conifer and angiosperms trees. For example, functions of water transport and mechanical support are attributed to a single cell type in conifers (tracheids) and to two distinct cell types in angiosperms (vessels and fibers); in addition, different types of hemicelluloses and lignins make up their secondary cell walls. We thus explored to what extent genes that are identified as unique or specific to conifers may be more represented in coexpression groups with strong expression in these tissue types. We used results of searches for sequence similarity to *A. thaliana* sequences (this work) and to Sitka spruce and loblolly pine sequences (Rigault *et al.*, 2011) to identify putative conifer specific genes and analyzed their distribution across coexpression groups. We found that conifer-specific genes were overrepresented in groups with a preference for expression in secondary vascular tissues (M2a, M6a), and underrepresented in groups without a clear preference in any tissue (M1b, M3a) or with high expression in tissues more closely derived from shoot meristem (M1a, M8b) (Table S3.5).

3.5.5 Highly conserved invariant genes are associated with high expression and cell basic functions

We observed that invariant genes spanned a broad range of expression levels (Table 3.2) and that fewer of them had functional annotations compared to the variable genes (Fig. 3.3a, Fig. S3.3).

Overall, the 596 invariant genes within enriched GO terms were predominantly associated to basic cellular activities and to a lesser extent to transport, cell communication/transduction, cell/organ development, and stress response (Fig. 3.3b), suggesting an overrepresentation of housekeeping genes. We analysed the group further with the aim of uncovering relationships between expression levels and functional annotations or sequence similarities.

Results showed that genes that were conserved between conifers and *A. thaliana* were more strongly expressed as they were overrepresented among the high expression classes. Inversely, the putative conifer- specific genes were overrepresented in low expression classes (Table 3.2). Genes associated with the major enriched functions were distributed in similar proportions in all expression level classes except for stress response category which was statistically overrepresented in the lowest expression class. This latter group likely represents genes whose expression may vary in response to a stress factor and had a low basal level in tissues grown in permissive conditions.

Table 3.2 Function and cross-species conservation of invariant genes according to their expression levels

Expression classes ¹	Total	Cross-species sequence conservation ²			Functional classes ³			
		Conifers	Conifers - arabidopsis	White spruce- arabidopsis	A (%)	B (%)	C (%)	D (%)
Low (<Q1)	1202	142*	225*	110*	79	7	16*	11
Low to moderate (Q1-Q2)	1201	147*	227*	89*	94	6	8	10
Moderate to high (Q2-Q3)	1201	109	419*	134*	94	10	7	13
High (>Q3)	1201	82*	686*	397*	90	9	10	20
Total in number	4805	480	1557	730				

¹Based on their mean expression from all tissue samples of white spruce (*Picea glauca*), invariant genes were divided into four equal classes delimited by the first, second and third quartiles (Q1, Q2, or Q3); ²Conifers: number of genes with conserved sequences between white spruce (*P. glauca*), sitka spruce (*P. sitchensis*) and loblolly pine (*Pinus taeda*); Conifers–*Arabidopsis thaliana*: number of genes with conserved sequences between one of the three conifer species and *A. thaliana*; and White spruce–*A. thaliana*: number of genes with conserved sequences between white spruce and *A. thaliana* based on sequence similarity approach (this work, Rigault *et al.*, 2011; BLASTX, *E*-value < 10⁻¹⁰); ³Frequency (number in %) of genes associated with each of the functional classes; A, basic cellular activities; B, cell/organ development; C, response to abiotic and biotic factors; D, transport, cell communication/transduction; the sum of percentages per row exceeds 100% because a gene may be involved in one or more functional classes; ^{2, 3}Hypergeometric tests performed for expressional classes by column: *statistically significant (adjusted *P*-value < 0.05), for over (bold numbers) or underrepresentation (italic numbers).



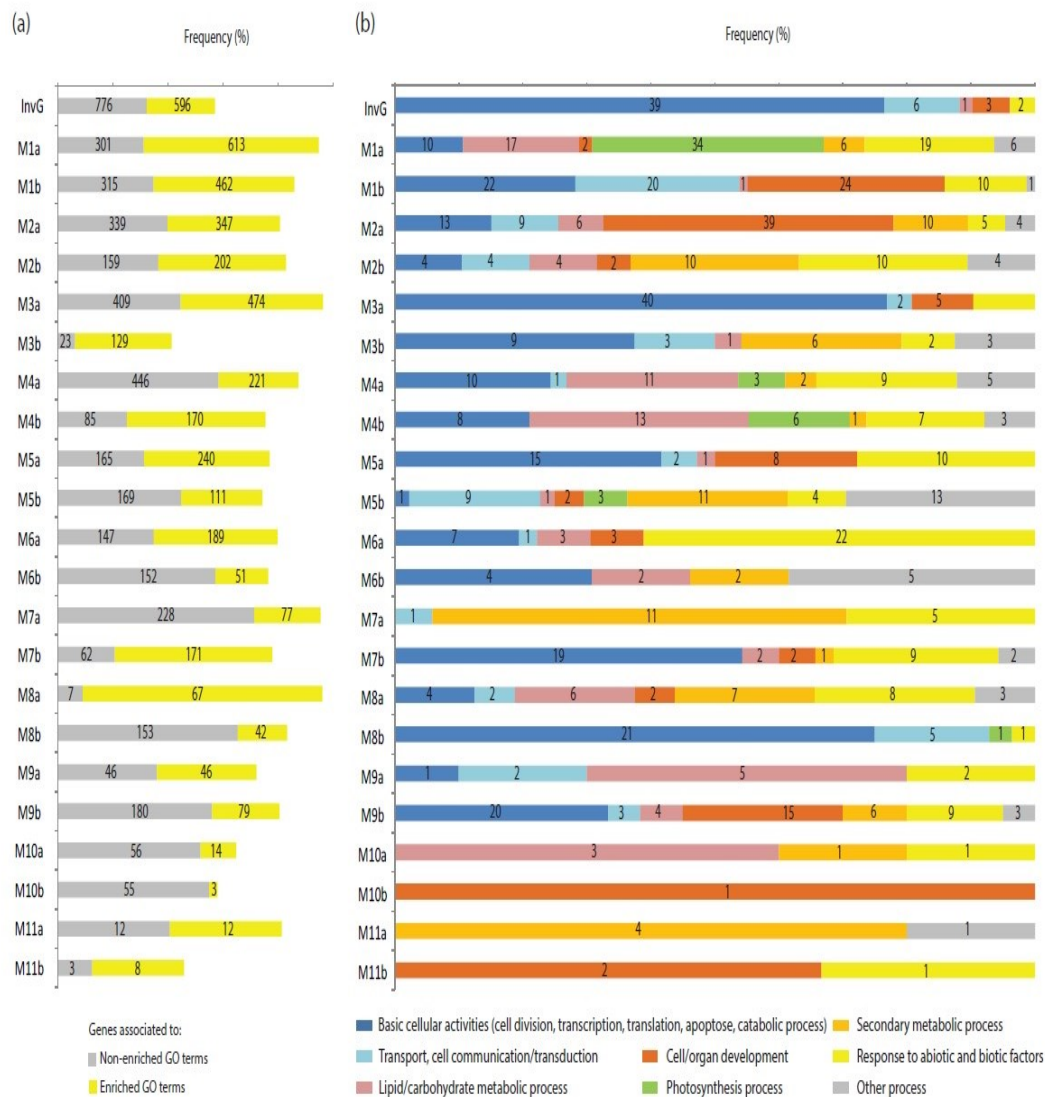


Fig. 3.3 Functional analysis of invariant genes (InvG) and coexpression groups (M1a-M11b) in white spruce (*Picea glauca*)

(a) X axis, frequency (number in %) of genes associated with each of the two annotation categories: enriched and non-enriched biological process gene ontology (GO) terms; Number inside bars, number of genes per category; (b) X axis, Frequency (number in %) of enriched GO terms associated with consensus functional classes; Number inside bar, number of enriched GO terms per consensus functional class.

3.5.6 Expressional and functional diversity among conserved secondary xylem and phelloderm preferential genes

We used the defined coexpression groups to characterize sets of genes that we previously identified as preferentially expressed in shoot secondary xylem (2537 genes) and phelloderm (2800 genes) tissues and were conserved in white, black and sitka spruces (Raheison *et al.*, 2012; Table S3.1). We found that 95% of the conserved xylem and phelloderm preferential genes were distributed among all coexpression groups, indicating that their modular organization is represented by diverse expression profiles (Fig. 3.4). Their respective distributions among coexpression groups were highly contrasted and matched closely with the profiles of the coexpression groups with regard to expression in xylem and phelloderm tissues; except for group M1a (with similar expression levels in both xylem and phelloderm tissues). The two tissue types are derived from different but neighbouring secondary stem meristems; the xylem is derived from the cambial zone (located more internally) and the phelloderm is derived from the phellogen (located more externally). Our results show the extent of transcriptome reorganization resulting from differentiation pathways from these two meristems.

3.5.7 Distinct expression patterns characterize wood formation at different stages during a growth season

Secondary vascular development and growth in woody plants follow an annual cycle that is characterized by an initial rapid growth phase (earlywood formation), followed by a slower phase that transitions to dormancy (latewood formation). We carried out a time-course microarray profiling experiment (experiment 2) and identified 1366 genes with significant temporal variation in white and Norway spruces (adjusted P -value < 0.05). The genes were clustered with WGCNA into six temporal coexpression groups of 41 to 507 genes (Fig. 3.5a, Table S3.1). Overall, transcript levels did not change between June and July, remained stable (Fig. 3.5a; cluster T1, 726 genes) or varied progressively (Fig. 3.5a; cluster T2, 547 genes) in August and changed dramatically from August to September. Transcripts accumulating preferentially during June-July and September represented 51% and 43%, respectively of the genes (Fig. 3.5a). In contrast, only 3% of the genes were preferentially expressed in August.

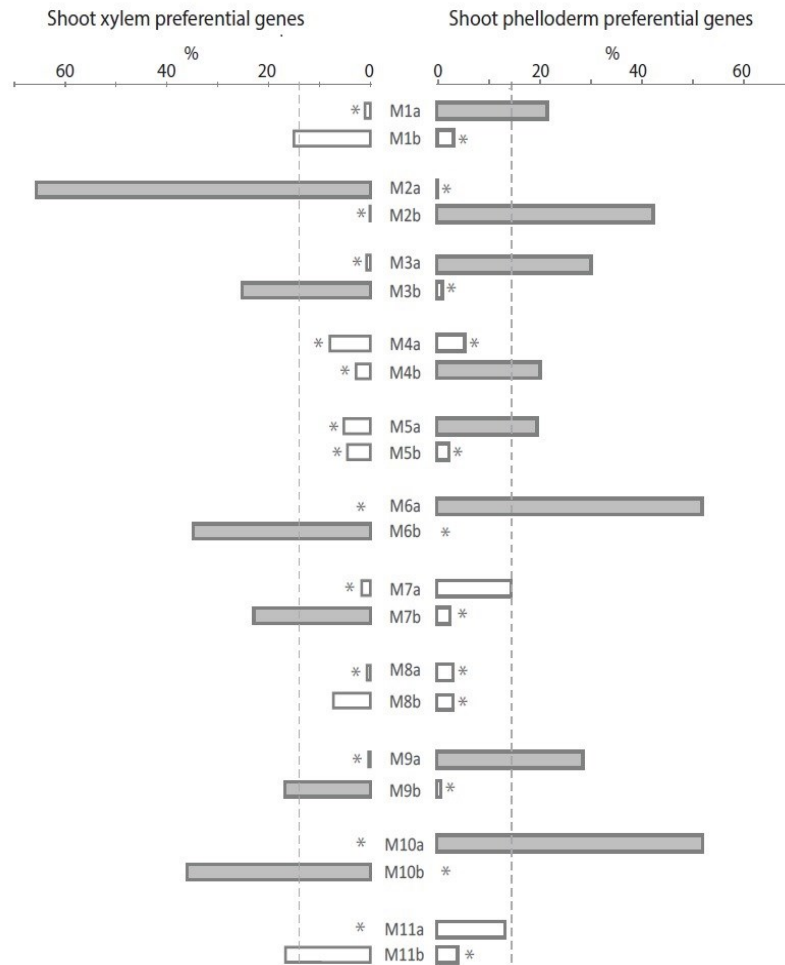


Fig. 3.4 Distribution of secondary xylem or phelloderm preferential genes conserved among *Picea* species across coexpression groups. Vascular tissue preferential genes in white (*Picea glauca*), Norway (*P. abies*) and black (*P. mariana*) spruces were identified in Raheison *et al.* (2012).

Dashed grey vertical lines represent the expected frequency based on chance alone (%) of preferential genes across coexpression groups. Overrepresentation of xylem or phelloderm preferential genes compared to chance alone (hypergeometric test; adjusted P -value < 0.05) is shown by grey-colored bars and underrepresentation of xylem or phelloderm preferential genes is shown by open bars with an asterix (hypergeometric test; adjusted P -value < 0.05). Each of the over and underrepresentations of xylem and phelloderm preferential genes closely matched with the expression profiles of the corresponding coexpression groups as shown in Fig. 3.2a and S 3.2a-j

Next, we analyzed the temporally variable gene clusters for membership in the 22 coexpression groups derived from our comparison of vegetative tissues in white spruce (Fig. 3.5b). We found that June-July preferential genes (average Pearson correlation coefficient $r = 0.6$) were overrepresented in groups exhibiting high expression in xylem tissues (M2a, M10b); these genes are called earlywood genes. This observation is consistent with the fact that the secondary xylem tissues used to define the tissue-based coexpression groups were sampled in July during earlywood formation (Fig. 3.5b). Inversely, September preferential genes ($r = 0.8$) were overrepresented in groups with a strong preference in phelloderm tissues or in other tissues than xylem (M2b, M6a, and M10a); these genes are referred as latewood genes (Fig. 3.5b). These observations indicate that many of the earlywood genes were specific to xylem tissues at the start of the growing season while latewood genes were more active in other parts of the trees during early phases of the growth season. Analysis of functional annotations indicated that basic cellular activities, lipid/carbohydrate reserves and secondary metabolic process were specific to earlywood genes and that stress-responsive genes were specific to latewood genes. Cell/organ development is a common function with a higher representation in latewood genes (Fig. 3.5c, Table S3.2).

3.5.8 Variation in secondary cell wall regulatory network hub genes is associated with xylem transcriptome reorganization during the growth season

We developed three unweighted and signed networks targeting two xylem preferential modules (M2 and M7), June-July and September preferential genes. In these analyses, two genes whose expression profiles have an absolute Pearson correlation coefficient above 0.9 were connected. We computed the number of genes connected to each gene (*i.e.* degree) to identify hub(s), which represent(s) the most connected gene(s). We identified the enriched pathways for M2-7 network in Table S3.6. Among the enriched pathways, we investigated the branch of phenylpropanoid biosynthesis pathway leading to lignin biosynthesis (Fig. 3.6). Based on gene expression and sequence similarity, twelve of the 23 presented genes were preferentially expressed in xylem tissues and candidate genes for lignin biosynthesis. The remaining genes may contribute to other branches of the phenylpropanoid metabolic pathway.

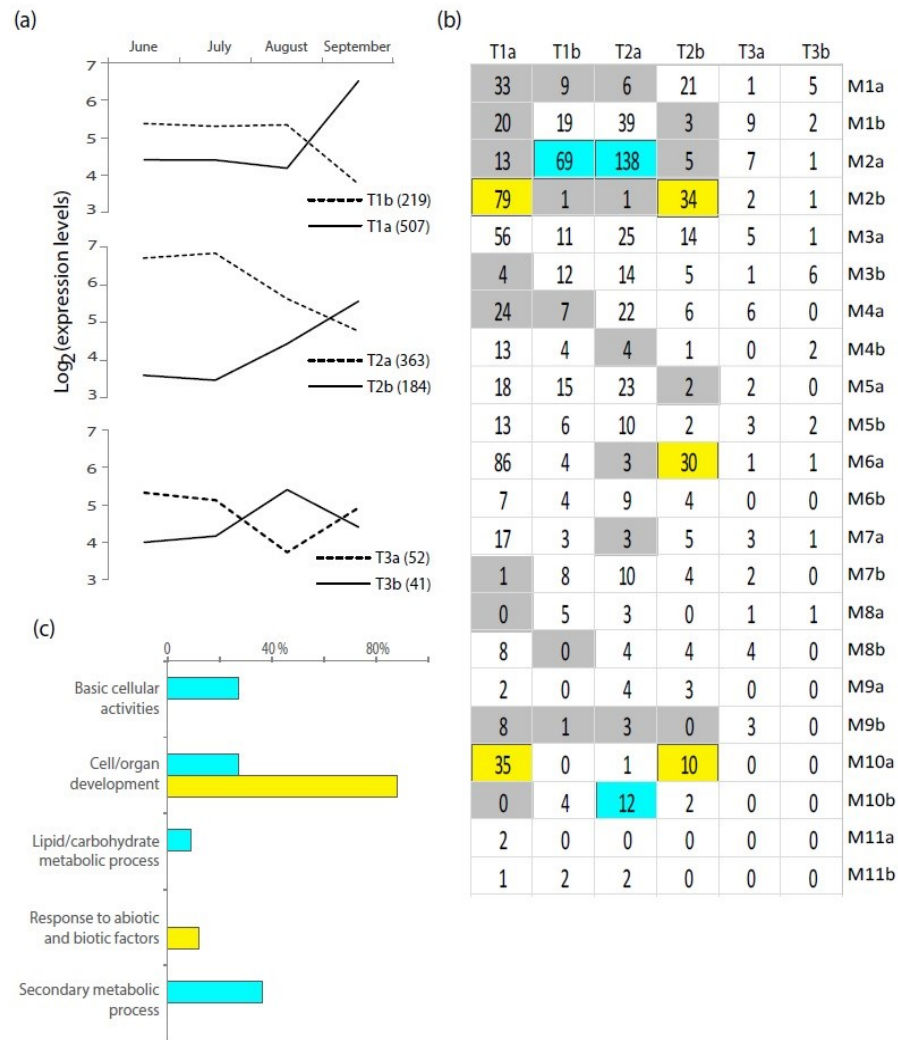


Fig. 3.5 Distribution of temporally variable xylem expressed genes across coexpression groups

(a) Expression profiles of temporally variable genes in secondary xylem of white (*Picea glauca*) and Norway (*P. abies*) spruces; Number in brackets, number of genes within temporally variable clusters (T1a-T3b); (b) Number of temporally variable genes across tissue-based coexpression groups (M1a-M11b); Grey and blue/yellow background, under and overrepresentation (hypergeometric test, adjusted P -value < 0.05), respectively; Blue and yellow background, earlywood and latewood genes, respectively *i.e.* June-July (T1b and T2a) and September (T1a and T2b) preferential genes that were overrepresented in M2a and M10b; and in M2b, M6a and M10a, respectively; (c) Functional annotation analysis based on GO term enrichment of earlywood (blue-colored histogram) and latewood (yellow-colored histogram) genes

Xylem preferential (M2-7) network was scale free (Fig. S3.4) and had PgNAC-7 (no apical meristem (NAM), *Arabidopsis* transcription activation factor (ATAF) and cup-shaped cotyledon (CUC) transcription factor 007 in *Picea glauca*) as the most connected hub gene (Table 3.3a). A sub-network containing common interactions (connections) and genes between xylem (M2-7) and June-July preferential networks also had PgNAC-7 as the most connected hub gene (Table 3.3a; Table S3.1). PgNAC-7 was recently identified as a key regulator of secondary cell wall development in white spruce, which is functionally analogous to VND6/7 (vascular-related NAC-domain 6/7; Duval *et al.*, 2014). Functional annotation analysis was carried among genes positively and negatively connected to PgNAC-7. Genes positively connected to PgNAC-7 were specifically enriched for basic cellular, development and reserve metabolisms, whereas stress response and secondary metabolism terms were enriched in both negatively and positively connected genes (Fig. 3.7a,b).

We present a network showing PgNAC-7 and genes involved in secondary cell wall formation as reviewed in Mizrachi *et al.* (2012) or shown to be transactivated in conifers by regulators of secondary cell wall formation including MYB transcription factors (Bomal *et al.*, 2008, Duval *et al.*, 2014, Fig. 3.7c, Table S3.7). The network includes temporally variable and non-variable genes (Fig. 3.7c). Temporally variable genes included PgNAC-7, five of the 10 MYBs and seven of the 22 enzymes, indicating that the network may vary over the course of a growth season. We observed that all of the earlywood genes were positively connected to PgNAC-7 and the latewood genes were negatively connected to PgNAC-7.

A network was constructed for September preferential genes. The most connected hub genes were three stress-responsive genes, *i.e.* a nucleotide-binding-site-leucine-rich-repeat- or NBS-LRR-encoding gene (activated disease resistance 1 (ADR1)), a dehydrin-encoding gene (DHR), and a tubby-protein-encoding gene (TUB) (Table 3.3b, Table S3.1). Several other dehydrins were also co-regulated in this network (Table S3.1). These findings are consistent with changes in environmental factors including water deficit and temperature decrease that occur at the end (September) of a growth season and influence latewood formation in conifers.

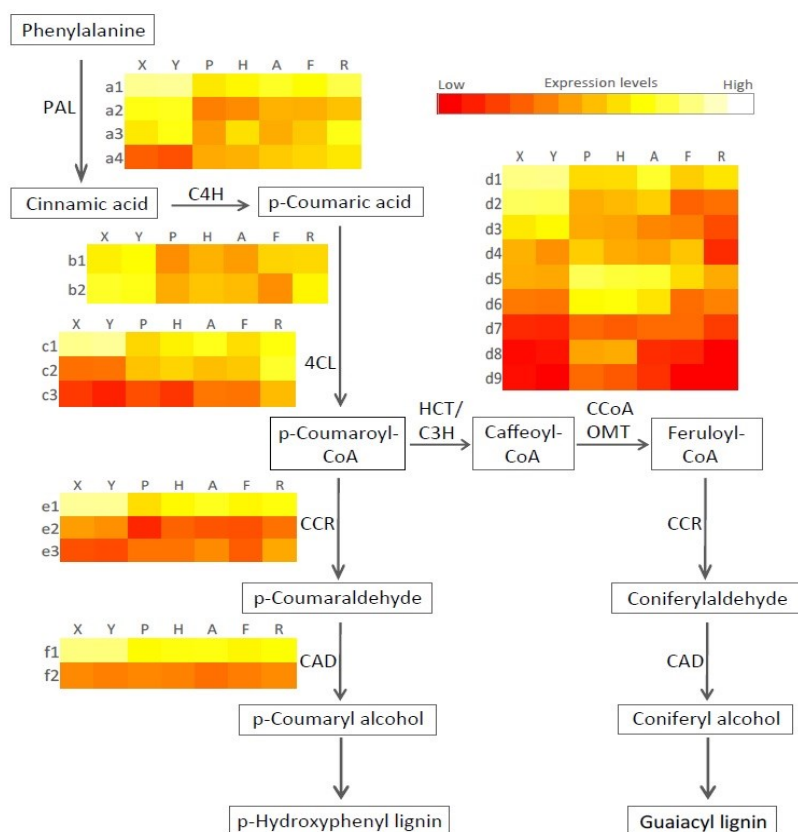


Fig. 3.6 Phenylpropanoid pathway leading to lignin production

Boxes, metabolite; Arrows, enzyme reactions; enzymes are reported with three letter code: PAL, phenylalanine ammonia lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate-CoA ligase; HCT/C3H, shikimate/quinate hydroxycinnamoyltransferase / p-coumarate 3-hydroxylase; CCoAOMT, caffeoyl-CoA O-methyltransferase; CCR, cinnamoyl-CoA reductase; CAD, cinnamoyl alcohol dehydrogenase; Heatmaps next to enzyme name indicate expression value of genes (a1-4, b1-2, c1-3, d1-9, e1-3, f1-2) across tissues (X, Y, P, H, A, F, R); Each expression value is the average of four the biological replicates per tissue; a1-4, BT112211, BT106538, BT119163, BT114680; b1-2, BT116118, BT117086; c1-3, BT106671, BT116171, DR551141; d1-9, BT106698, BT106398, BT101243, BT117977, BT110631, BT102905, BT112333, BT104804, BT114515; e1-3, DR573886, BT111802, BT112289; f1-2, BT112280, BT116920; X, shoot secondary xylem; Y, root secondary xylem; P, shoot phelloderm; H, root phelloderm; A, shoot apex; F, young foliage; R, root tips.

Table 3.3 Functional annotation of the top ten most connected genes (hubs) of the networks

GenBank Accession Number	Cluster ID ¹	Degree ²	Rank ³	PFAM description ⁴	Functional annotation ⁵
(a)					
BT102049	GQ0165_B14	53	1	No apical meristem (NAM) protein	PgNAC-7, NAC 007
CO235762	WS0021_I12	43	2		
BT117400	GQ03818_L16	40	3	Subtilase family	Subtilase family protein
DR591085	WS00830_F12	40	3		
BT108414	GQ03121_H22	39	5	Myb domain protein	PgMYB2*
BT106274	GQ02906_F17	36	6		
BT107150	GQ03103_E24	35	7	Xyloglucan endo-transglycosylase (XET) C-terminus; Glycosyl hydrolases family	xyloglucan endotransglucosylase/hydrolase
BT107836	GQ03113_F02	34	8	Aldehyde dehydrogenase family	aldehyde dehydrogenase 3F1
DR594044	WS00839_J22	33	9	Subtilase family; protease-associated domain; Peptidase inhibitor I9	Subtilisin-like serine endopeptidase family protein
BT106749	GQ03009_M04	32	10		Ring/U-box superfamily protein
(b)					
BT105806	GQ02828_L20	496	1	Tubby C 2 (TUB)	Protein of unknown function (DUF567)
BT112527	GQ03316_N22	496	1	Leucine Rich Repeat	ADR1, Disease resistance protein (CC-NBS-LRR class) family
BT117952	GQ03918_F10	496	1	Dehydrin (DHR)	
BT103744	GQ02801_H14	495	4	ThiF family	ubiquitin-activating enzyme
BT109195	GQ03205_L14	494	5	Glycosyl hydrolase family, N-terminal domain	glucuronidase
BT113137	GQ03326_D07	494	5	Dehydrin	
BT114680	GQ03519_E06	494	5	Aromatic amino acid lyase	phenylalanine ammonia-lyase 4
BT104565	GQ02811_O09	493	8	Embryo-specific protein 3, (ATS3)	Embryo-specific protein 3, (ATS3)
BT115870	GQ03707_L13	493	8	Glycosyl hydrolase family	beta glucosidase
BT101047	GQ0065_L03	492	10		
BT104876	GQ02816_C10	492	10	Hsp20/alpha crystallin family	heat shock protein 17.4
BT104881	GQ02816_D12	492	10		sucrose synthase 3
EX329848	GQ02830_C21	492	10	Dehydrin (DHR)	
BT114833	GQ03603_F10	492	10	Dehydrin (DHR)	
BT116286	GQ03715_F15	492	10	Pyruvate phosphate dikinase, PEP/pyruvate binding domain	Pyruvate phosphate dikinase, PEP/pyruvate binding domain
DR574449	WS00737_M23	492	10		Cox19 family protein (CHCH motif)

(a), Sub-network (Xylem (M2-7) and June-July preferential network) which contains 185 genes (see Table S3.1 for details); (b), September preferential network which contains 691 genes (see Table S3.1 for details); ¹From by Rigault *et al.* (2011); ²The number of genes connected to each gene within the network; ³The rank of genes based on their degree within the network, genes with similar degree have the same rank; ⁴PFAM, the protein families database; ATS3, *Arabidopsis thaliana* seed gene 3; PEP, *phosphoenolpyruvate*; ⁵Functional annotation based on TAIR10 (The *Arabidopsis* Information Resource version 10); NAC 007, no apical meristem (NAM), *Arabidopsis* transcription activation factor (ATAF) and cup-shaped cotyledon (CUC) transcription factor 007; PgNAC-7, name of the NAC 007 gene in white spruce (*Picea glauca*) according to Duval *et al.* (2014); DUF, Domain of unknown function; ADR1, activated disease resistance 1; CC-NBS-LRR, coiled coil-nucleotide-binding-site-leucine-rich-repeat.

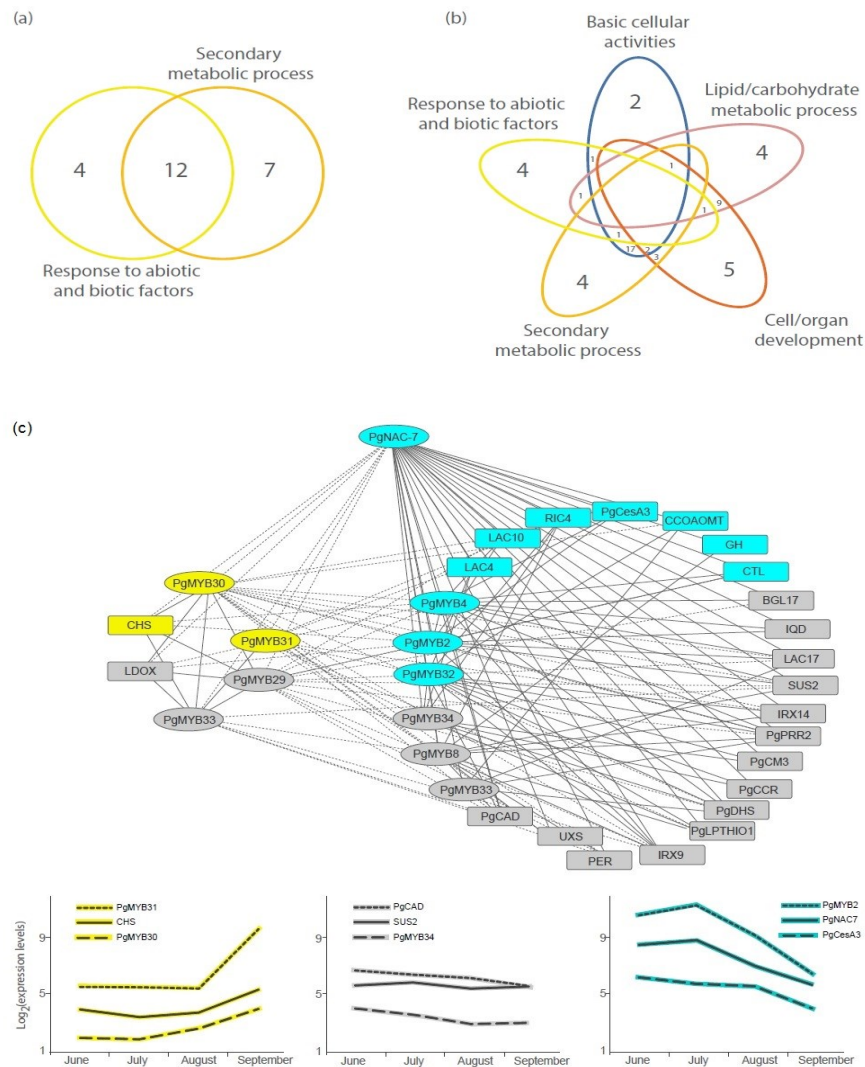


Fig. 3.7 Functional, network analysis and temporal expression patterns of *PgNAC-7* connected genes

PgNAC-7, no apical meristem (NAM), *Arabidopsis* transcription activation factor (ATAF) and cup-shaped cotyledon (CUC) transcription factor 007 in *Picea glauca*; (a,b) Venn diagrams show distribution of *PgNAC-7* negatively (a) and positively (b) connected genes among functional classes; (c) Upper panel, an unweighted and signed network shows connections of *PgNAC-7* with MYB transcription factors (oval-shaped nodes) and putative genes (rectangle-shaped nodes) associated with secondary cell wall formation (Duval *et al.*, 2014; Mizrahi *et al.*, 2012); *PgNAC-7* / MYB transcription factors and genes are listed in Table S3.1. Edges or connections are coexpression between genes with a Pearson correlation coefficient < -0.9 (dashed line) or $\geq +0.9$ (solid line). Lower panel, temporal variation of genes; node and graph colors, temporal expression of genes in secondary xylem tissue of white (*P. glauca*) and Norway (*P. abies*) spruces during a growing season: blue and yellow, highly expressed genes in June, July or August (earlywood genes) and in September (latewood genes), respectively; grey, temporally invariable genes.

3.6 Discussion

Our findings described the modular organization of white spruce transcriptome which we divided into invariant genes and coexpression groups of variable genes across tissues. We assigned 18 052 genes to 22 coexpression groups that are statistically robust and 4805 genes as being invariant across tissues. We also showed the complexity, conservation and reorganization of the transcriptome associated with secondary vascular tissue formation in spruces. These findings relating to transcriptome organization are discussed in the context of plant evolution, gene functions and expression networks.

3.6.1 Evolutionary signatures of tissue differentiation

We measured transcript levels in seven vegetative tissues of above- and below-ground organs of white spruce using a microarray platform and found that a large proportion of genes were differentially expressed across tissues with strong (10 548 high-confidence genes) or moderate statistical support (7504 low-confidence genes). Tissue differentiation has generally been linked to deeper transcriptome reorganization than developmental stages or environmental conditions in plants and animals. Our results from the comparison of white spruce tissues identified a much larger proportion of variable genes than comparisons of biological conditions or developmental stages in conifer trees including interior hybrid spruce (Verne *et al.*, 2011), maritime pine (Paiva *et al.*, 2008), radiata pine (Li *et al.*, 2010) and white spruce (this work). Similar findings have been reported in other plants including *A thaliana* (Ma *et al.*, 2005), tobacco (Edwards *et al.*, 2010) and maize (Downs *et al.*, 2013). In *A thaliana*, Aceituno *et al.* (2008) indicated that tissue types were a more influential variable for transcriptome variation than any of the tested experimental conditions. In mammals, tissue affiliation explained generally at least 10 times more expression variance than perturbations or diseases (Prasad *et al.*, 2013). In mouse, Su *et al.* (2002) identified 78% of genes with differential expression across tissues *versus* 9-10% across time points (Kwon *et al.*, 2012) or 1-7% between influenza-infected lungs and control (Pommerenke *et al.*, 2012). Tissue differentiation is evolutionarily ancient and has been linked to transcriptome signatures that are conserved across species (Prasad *et al.*, 2013). In our previous work, we found that only 60 of the 5407 vascular tissue differential genes have different expression patterns when comparing three spruce species (Raherison *et al.*, 2012). These findings showed how transcriptome signatures



highlighted the ancient evolutionary origins of tissue differentiation (~380 Mya; Kenrick & Crane 1997).

In light of these observations and the high level of genomic conservation among spruces (Ritland *et al.*, 2011), expression profiles may be regarded as a fundamental property of genes and their orthologs across the genus. It is also to be expected that profiles extend to other genera within the Pinaceae by extrapolation from comparison of white spruce and loblolly pine showing a high level of conservation in gene coding sequences (Rigault *et al.*, 2011) and in genome structure (Pavy *et al.*, 2013). The response to abiotic stress factors is also well conserved between pine and spruce as shown by Yeaman *et al.* (2014) who reported that the expression pattern was conserved for 74% of genes identified as sequence orthologs.

3.6.2 Modular transcriptome organization underpins tissue differentiation

The features of sequences that make up the different tissue transcriptomes largely overlap but their quantitative variations in transcript amounts reveal the underlying transcriptome reorganization. These quantitative changes are underpinned by a complex modular organization that was represented here by coexpression groups (Fig. 3.2, Fig. S3.2, Table S3.3, S3.4). The modular organization may be illustrated by considering secondary xylem and phelloderm from the stem. Transcripts were detected for 22 970 and 23 296 genes in xylem and phelloderm, respectively, and were distributed among invariant genes and all of the 22 coexpression groups (Table S3.8). This modular organization showed the expressional diversity of gene profiles in a given tissue.

Tissue preferential expression has been used to classify genes and to carry out functional studies and evolutionary investigations of tissue differentiation (Ma *et al.*, 2005; Pavy *et al.*, 2008; Yang & Wang, 2013; Duval *et al.*, 2014). We analyzed xylem and phelloderm preferential genes conserved in three spruce species and found that they were distributed among all coexpression groups as well as being differentially represented in many of them (Fig. 3.4). These observations show the diversity of expression profiles among genes identified as preferentially expressed in a given tissue.

3.6.3 Coexpression is associated with functional similarity

We report a high rate of GO term enrichment among coexpression groups. Almost 50% of the genes with assigned GO terms were associated with GO term enrichment in the different groups (Fig. 3.3a, Fig. S3.3). Coexpression groups were for the most part very distinct in regard to predominant enriched biological functions. Taken together, these observations indicated that several members within coexpression groups shared functions in addition to expression profiles, which support the notion that members of coexpression groups participate in similar or common processes. For example, enriched terms for photosynthesis, stress response and lipid metabolic processes were largely observed in the group of genes highly expressed in foliage and shoot apex tissues (M1a in Fig. 3.3b, Table S3.2). The shoot apex requires lipids to support intense cell division activity and rapid growth rate (Jordy, 2004). Both shoot apex and needles are photosynthetic tissues and regularly exposed to stress conditions. In turn, cell division, cell/organ development and cell transduction were the predominant enriched functions in profiles with higher expression in root vascular tissues (M1b in Fig. 3.3b, Table S3.2). Root tips contain a meristematic zone in active state of cell division (Reddy & Chary, 2003; Verbelen *et al.*, 2006); root growth and development are controlled by cell signal transduction pathways (Perrin *et al.*, 2005; Grieneisen *et al.*, 2007; Laskowski *et al.*, 2008).

Coexpression groups with opposite profiles within an expression module may have contrasted GO term enrichments (Fig. 3.2a, Fig. S3.2). For example, group with strong expression in secondary xylem tissues (M2a) had higher representation of terms related to cell/organ development than group with low expression in secondary xylem (M2b). The opposite was observed for the terms related to carbohydrate/lipid reserves and defense mechanisms. Coexpression groups with high (M7a) and low (M7b) expression in root tips and xylem were enriched for secondary metabolic processes and for basic cellular mechanisms, respectively; and both were enriched for stress responsive genes.

Considering these observations, enriched GO terms may represent indirect information on the biological function of co-expressed genes which lack GO terms or functional annotations (*i.e.* guilt by association). Genes highly co-expressed have a greater chance of sharing a common regulatory mechanism and therefore have similar functions (Allocco *et al.*, 2004). In this regard, a few of the coexpression profiles had

a relatively low proportion of genes with associated GO terms (*e.g.* M3b and M11b) where less than 25% of genes with assigned GO terms were associated with enriched GO terms. Genes without predicted functions are numerous in conifers (Rigault *et al.*, 2011) and represent significant numbers even in model organisms. Coexpression groups may help not only to functionally annotate these genes but also to improve inference of gene function across species. According to Prasad *et al.* (2013), an analysis combining tissue expression profiles with sequence similarity provides a better prediction of functions of homologous genes across species than using sequence information alone.

We identified 4805 genes that did not vary across tissues (invariant genes) and thus could be designated as housekeeping genes according to Chang *et al.* (2011) and Eisenberg & Levanon (2013). The number of putative housekeeping genes decreases to 1201 when using expression level criteria (moderate to high expression genes, Table 3.2) defined by Aceituno *et al.* (2008) in *A. thaliana*. We found about the same number of housekeeping genes in white spruce as reported for humans (Chang *et al.*, 2011) and three-fold more than reported for *A. thaliana* (Aceituno *et al.*, 2008), both of which studies considered many more tissue types and biological conditions. As relevant for housekeeping functions, the majority of invariant genes had a biological function related to basic cellular activities, in agreement with previous reports (Chang *et al.*, 2011; Eisenberg & Levanon 2013). Based on sequence similarity analysis, we also found that most of the highly expressed invariant genes were conserved between *A. thaliana* and white spruce (Table 3.2). This finding is consistent with previous reports indicating high expression of the majority of genes conserved between closely (mouse and Human; Liao & Zhang, 2006) or distantly related species (mouse and fruit fly; Tamura *et al.*, 2004). Highly expressed genes were shown to have lower evolutionary rates of change in protein sequences (Tamura *et al.*, 2004; Liao & Zhang, 2006) and to be under stronger selection for mRNA folding (Park *et al.*, 2013).

3.6.4 Temporal reorganization of gene expression networks in the vascular transcriptome

Network analysis identified PgNAC-7 as a network hub that is positively connected to several genes involved in secondary cell wall formation. These findings are consistent with the role of PgNAC-7 and downstream MYB genes (PgMYB1, 2, 4 and 8) shown in white spruce to act as secondary cell wall regulators in transient assays (Duval *et al.*, 2014) and in transgenic trees (Bomal *et al.*, 2008). This network is functionally similar to that regulated by class IIB NAC proteins VND4, VND5 and VND6 of *A. thaliana* (Duval *et al.*, 2014) and several downstream genes of MYBs (Zhong & Ye, 2009; Zhong *et al.*, 2010), which is also conserved in woody angiosperms including poplar (*Populus trichocarpa*) and eucalyptus (*Eucalyptus gunnii*) (Zhong & Ye, 2009; Zhong *et al.*, 2010; Hussey *et al.*, 2013).

Temporal analysis in white spruce indicated that PgNAC-7 and three of the spruce MYBs had expression profile strongly related to earlywood, while no temporal variation was observed for other MYBs and most of the downstream genes encoding enzymes involved in secondary cell wall assembly. These observations indicate that the regulatory network may be reconfigured over the course of a growth season. A scan of the latewood preferential cluster (Table S3.1) identified one other NAC sequence and four other MYBs but these sequences have not been tested functionally. In turn, they represent candidate genes to be investigated for involvement in the regulation of members of this network.

Remodeling of the regulatory network in latewood was supported by the finding that the most connected hubs are stress responsive genes, *i.e.* ADR1, a tubby gene and a dehydrin gene. ADRs encode NBS-LRR proteins involved in receptor-mediated signaling (Bonardi *et al.*, 2011). The network also contained many other co-regulated dehydrin genes. ADR1, tubby protein and dehydrin proteins enhance drought tolerance in plants (Chini *et al.*, 2004; Puhakainen *et al.*, 2004; Wardhan *et al.*, 2012). Our findings are consistent with the transition to dormancy that is initiated during latewood formation and involves several cellular changes including shifts in osmotic processes (Paiva *et al.*, 2008). Differences between earlywood and latewood formation in terms of biological function and transcriptional regulatory network indicate transcriptional changes associated with seasonal network reorganization during a growing season in conifers.

3.7 Concluding remarks

We classified expressed white spruce genes into coexpression groups that are indicative of the modular organization of the transcriptome. Our results showed that deeper transcriptome reorganization is associated with tissue differentiation than with developmental stages or environmental conditions, and that patterns are conserved between spruce species as might be expected given the ancient evolutionary origins of tissue differentiation. Time-course analyses of wood formation indicated network reorganization and a transition from cell wall formation toward defense related genes. This report extends previous work on the PiceaGeneExpress database (Raheison *et al.*, 2012) by statistical modeling of transcriptome organization leading to network analysis. We have developed a framework of information for spruce genes and an approach that could be readily applied to any number of tree and other plant genomes. Combined with other types of information, coexpression groups may enable basic discoveries relating to genome function, regulatory networks, biomarkers and candidate genes in addition to supporting the development of applications for breeding and the conservation of genetic diversity.

3.8 Acknowledgements

We thank Dr M. Mazerolle (Centre for Forest Research) for his statistical advice on the mixed ANOVA, Dr A. Séguin (Centre de Foresterie des Laurentides) for the tissue illustration in Fig. 3.1a, Dr K. Plourde and G. Germanos (Centre for Forest Research) for sample preparations. Funding to J.M. was from Genome Canada and Genome Québec as a part of the SMarTForests project and from an NSERC Discovery grant.

3.9 References

- Aceituno FF, Moseyko N, Rhee SY, Gutiérrez RA. 2008.** The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in *Arabidopsis thaliana*. *BMC Genomics* **9**: 438.
- Allocco DJ, Kohane IS, Butte AJ. 2004.** Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* **5**: 18.
- Allona I, Quinn M, Shoop E, Swope K, Cyr SS, Carlis J, Riedl J, Retzel E, Campbell MM, Sederoff R. 1998.** Analysis of xylem formation in pine by cDNA sequencing. *Proceedings of the National Academy Of Sciences* **95**: 9693-9698.

- Baldi P, Pedron L, Hietala AM, La Porta N. 2011.** Cold tolerance in cypress (*Cupressus sempervirens* L.): a physiological and molecular study. *Tree Genetics & Genomes* **7**: 79-90.
- Benjamini Y, Hochberg Y. 1995.** Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**: 289-300.
- Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Saint Yuen MM, Keeling CI, Brand D, Vandervalk BP. 2013.** Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics*: **29**: 1492-1497.
- Bomal C, Bedon F, Caron S, Mansfield SD, Levasseur C, Cooke JE, Blais S, Tremblay L, Morency M-J, Pavy N. 2008.** Involvement of *Pinus taeda* MYB1 and MYB8 in phenylpropanoid metabolism and secondary cell wall biogenesis: a comparative in planta analysis. *Journal Of Experimental Botany* **59**: 3925-3939.
- Bonardi V, Tang S, Stallmann A, Roberts M, Cherkis K, Dangl JL. 2011.** Expanded functions for a family of plant intracellular immune receptors beyond specific recognition of pathogen effectors. *Proceedings of the National Academy of Sciences* **108**: 16463-16468.
- Boyle B, Dallaire N, MacKay J. 2009.** Evaluation of the impact of single nucleotide polymorphisms and primer mismatches on quantitative PCR. *BMC Biotechnology* **9**: 75.
- Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M. 2011.** The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343-348.
- Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe M, Wilde A, Brudno M. 2009.** Conservation of core gene expression in vertebrate tissues. *J Biol* **8**: 33.
- Chang C-W, Cheng W-C, Chen C-R, Shu W-Y, Tsai M-L, Huang C-L, Hsu IC. 2011.** Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PloS One* **6**: e22859.
- Chang S, Puryear J, Cairney J. 1993.** A simple and efficient method for isolating RNA from pine trees. *Plant Molecular Biology Reporter* **11**: 113-116.

- Chini A, Grant JJ, Seki M, Shinozaki K, Loake GJ. 2004.** Drought tolerance established by enhanced expression of the CC–NBS–LRR gene, ADR1, requires salicylic acid, EDS1 and ABI1. *The Plant Journal* **38**: 810-822.
- Downs GS, Bi Y-M, Colasanti J, Wu W, Chen X, Zhu T, Rothstein SJ, Lukens L. 2013.** A developmental transcriptional network for *Zea mays* defines coexpression modules. *Plant Physiology* **161**: 1830-1843.
- Druka A, Muehlbauer G, Druka I, Caldo R, Baumann U, Rostoks N, Schreiber A, Wise R, Close T, Kleinhofs A. 2006.** An atlas of gene expression from seed to seed through barley development. *Functional & Integrative Genomics* **6**: 202-211.
- Duval I, Lachance D, Giguère I, Bomal C, Morency M-J, Pelletier G, Boyle B, MacKay JJ, Séguin A. 2014.** Large-scale screening of transcription factor–promoter interactions in spruce reveals a transcriptional network involved in vascular development. *Journal Of Experimental Botany* **65**: 2319-2333.
- Edwards KD, Bombarely A, Story GW, Allen F, Mueller LA, Coates SA, Jones L. 2010.** TobEA: an atlas of tobacco gene expression from seed to senescence. *BMC Genomics* **11**: 142.
- Eisenberg E, Levanon EY. 2013.** Human housekeeping genes, revisited. *Trends In Genetics* **29**: 569-574.
- Ellison CE, Kowbel D, Glass NL, Taylor JW, Brem RB. 2014.** Discovering Functions of unannotated genes from a transcriptome survey of wild fungal isolates. *mBio* **5**: e01046-01013.
- GI 2000.** Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796.
- Grieneisen VA, Xu J, Marée AF, Hogeweg P, Scheres B. 2007.** Auxin transport is sufficient to generate a maximum and gradient guiding root growth. *Nature* **449**: 1008-1013.
- Harris LW, Lockstone HE, Khaitovich P, Weickert CS, Webster MJ, Bahn S. 2009.** Gene expression in the prefrontal cortex during adolescence: implications for the onset of schizophrenia. *BMC Medical Genomics* **2**: 28.
- Hu Z-L, Bao J, Reecy J. 2008.** CateGORizer: a web-based program to batch analyze gene ontology classification categories. *Online J Bioinformatics* **9**: 108-112.
- Huang DW, Sherman BT, Lempicki RA. 2009.** Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**: 1-13.

- Hussey SG, Mizrahi E, Creux NM, Myburg AA. 2013.** Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. *Frontiers In Plant Science* **4**: 325.
- IHGSC 2004.** Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931-945.
- Jordy MN. 2004.** Seasonal variation of organogenetic activity and reserves allocation in the shoot apex of *Pinus pinaster* Ait. *Annals Of Botany* **93**: 25-37.
- Kenrick P, Crane PR. 1997.** *The origin and early diversification of land plants. A cladistic study.* Washington DC, USA: Smithsonian Institution Press.
- Kibbe WA. 2007.** OligoCalc: an online oligonucleotide properties calculator. *Nucleic Acids Research* **35**: W43-W46.
- Kwon E-Y, Shin S-K, Cho Y-Y, Jung UJ, Kim E, Park T, Park JHY, Yun JW, McGregor RA, Park YB. 2012.** Time-course microarrays reveal early activation of the immune transcriptome and adipokine dysregulation leads to fibrosis in visceral adipose depots during diet-induced obesity. *BMC Genomics* **13**: 450.
- Langfelder P, Horvath S. 2008.** WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559.
- Laskowski M, Grieneisen VA, Hofhuis H, Colette A, Hogeweg P, Marée AF, Scheres B. 2008.** Root system architecture from coupling cell shape to auxin transport. *PLoS Biology* **6**: e307.
- Li X, Wu HX, Southerton SG. 2010.** Seasonal reorganization of the xylem transcriptome at different tree ages reveals novel insights into wood formation in *Pinus radiata*. *New Phytologist* **187**: 764-776.
- Liao B-Y, Zhang J. 2006.** Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Molecular Biology And Evolution* **23**: 530-540.
- Lorenz WW, Alba R, Yu Y-S, Bordeaux JM, Simões M, Dean JF. 2011.** Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (*P. taeda* L.). *BMC Genomics* **12**: 264.
- Ma L, Sun N, Liu X, Jiao Y, Zhao H, Deng XW. 2005.** Organ-specific expression of *Arabidopsis* genome during development. *Plant Physiology* **138**: 80-91.
- MacKay J, Dean JF, Plomion C, Peterson DG, Cánovas FM, Pavy N, Ingvarsson PK, Savolainen O, Guevara MA, Fluch S. 2012.** Towards decoding the conifer gigagenome. *Plant Molecular Biology* **80**: 555-569.

- Mizrachi E, Mansfield SD, Myburg AA. 2012.** Cellulose factories: advancing bioenergy production from forest trees. *New Phytologist* **194**(1): 54-62.
- Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D. 2014.** The genome of *Eucalyptus grandis*. *Nature* **510**: 356-362.
- Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD. 2014.** Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology* **15**: R59.
- Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A. 2013.** The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**: 579-584.
- Paiva J, Garnier-Géré P, Rodrigues J, Alves A, Santos S, Graça J, Le Provost G, Chaumeil P, Silva-Perez D, Bosc A. 2008.** Plasticity of maritime pine (*Pinus pinaster*) wood-forming tissues during a growing season. *New Phytologist* **179**: 1180-1194.
- Palle SR, Seeve CM, Eckert AJ, Cumbie WP, Goldfarb B, Loopstra CA. 2011.** Natural variation in expression of genes involved in xylem development in loblolly pine (*Pinus taeda* L.). *Tree Genetics & Genomes* **7**: 193-206.
- Park C, Chen X, Yang J-R, Zhang J. 2013.** Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences* **110**: E678-E686.
- Pavy N, Boyle B, Nelson C, Paule C, Giguère I, Caron S, Parsons LS, Dallaire N, Bedon F, Bérubé H. 2008.** Identification of conserved core xylem gene sets: conifer cDNA microarray development, transcript profiling and computational analyses. *New Phytologist* **180**: 766-786.
- Pavy N, Gagnon F, Rigault P, Blais S, Deschênes A, Boyle B, Pelgas B, Deslauriers M, Clément S, Lavigne P. 2013.** Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. *Molecular Ecology Resources* **13**: 324-336.
- Perrin RM, Young L-S, Murthy UN, Harrison BR, Wang Y, Will JL, Masson PH. 2005.** Gravity signal transduction in primary roots. *Annals Of Botany* **96**: 737-743.

- Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. 2010.** Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**: 768-772.
- Pommerenke C, Wilk E, Srivastava B, Schulze A, Novoselova N, Geffers R, Schughart K. 2012.** Global transcriptome analysis in influenza-infected mouse lungs reveals the kinetics of innate and adaptive host immune responses. *PloS One* **7**: e41169.
- Prasad A, Kumar SS, Dessimoz C, Jaquet V, Bleuler S, Laule O, Hruz T, Gruissem W, Zimmermann P. 2013.** Global regulatory architecture of human, mouse and rat tissue transcriptomes. *BMC Genomics* **14**: 716.
- Puhakainen T, Hess MW, Mäkelä P, Svensson J, Heino P, Palva ET. 2004.** Overexpression of multiple dehydrin genes enhances tolerance to freezing stress in *Arabidopsis*. *Plant Molecular Biology* **54**: 743-753.
- Raherison E, Rigault P, Caron S, Poulin P-L, Boyle B, Verta J-P, Giguère I, Bomal C, Bohlmann J, MacKay J. 2012.** Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression. *BMC Genomics* **13**: 434.
- Reddy S, Chary S. 2003.** *University Botany II: (Gymnosperms, Plant Anatomy, Genetics, Ecology)*. New Delhi, India: New Age International (P) Ltd.
- Rigault P, Boyle B, Lepage P, Cooke JE, Bousquet J, MacKay JJ. 2011.** A white spruce gene catalog for conifer genome analyses. *Plant Physiology* **157**: 14-28.
- Ritland K, Krutovsky KV, Tsumura Y, Pelgas B, Isabel N, Bousquet J. 2011.** Genetic mapping in conifers. In: Plomion C, Bousquet J, Kole C, eds. *Genetics, genomics and breeding of conifers*. New York, USA: Edenbridge Science Publishers & CRC Press, 196-238.
- Romero IG, Ruvinsky I, Gilad Y. 2012.** Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics* **13**: 505-516.
- Rutledge RG, Stewart D. 2008.** A kinetic-based sigmoidal model for the polymerase chain reaction and its application to high-capacity absolute quantitative real-time PCR. *BMC Biotechnology* **8**(1): 47.
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA. 2009.** The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112-1115.

- Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A. 2002.** Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy Of Sciences* **99**: 4465-4470.
- Sun H, Paulin L, Alatalo E, Asiegbu FO. 2011.** Response of living tissues of *Pinus sylvestris* to the saprotrophic biocontrol fungus *Phlebiopsis gigantea*. *Tree Physiology* **31**: 438-451.
- Tamura K, Subramanian S, Kumar S. 2004.** Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Molecular Biology And Evolution* **21**: 36-44.
- Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A. 2006.** The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596-1604.
- Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA. 2007.** Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research* **35**: W71-W74.
- Verbelen J-P, De Cnodder T, Le J, Vissenberg K, Baluška F. 2006.** The root apex of *Arabidopsis thaliana* consists of four distinct zones of growth activities: meristematic zone, transition zone, fast elongation zone and growth terminating zone. *Plant Signaling & Behavior* **1**: 296.
- Verne S, Jaquish B, White R, Ritland C, Ritland K. 2011.** Global transcriptome analysis of constitutive resistance to the white pine weevil in spruce. *Genome Biology And Evolution* **3**: 851-867.
- Ward JH. 1963.** Hierarchical grouping to optimize an objective function. *Journal Of The American Statistical Association* **58**: 236-244.
- Wardhan V, Jahan K, Gupta S, Chennareddy S, Datta A, Chakraborty S, Chakraborty N. 2012.** Overexpression of CaTLP1, a putative transcription factor in chickpea (*Cicer arietinum* L.), promotes stress tolerance. *Plant Molecular Biology* **79**: 479-493.
- Yang R, Wang X. 2013.** Organ evolution in angiosperms driven by correlated divergences of gene sequences and expression patterns. *The Plant Cell Online* **25**: 71-82.
- Yang S-H, Loopstra CA. 2005.** Seasonal variation in gene expression for loblolly pines (*Pinus taeda*) from different geographical regions. *Tree Physiology* **25**: 1063-1073.

- Yeaman S, Hodgins KA, Suren H, Nurkowski KA, Rieseberg LH, Holliday JA, Aitken SN. 2014.** 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* **203**: 578-591
- Zhong R, Lee C, Ye ZH. 2010.** Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. *Trends Plant Science* **15**: 625-632.
- Zhong R, Ye ZH. 2009.** Transcriptional regulation of lignin biosynthesis. *Plant Signaling & Behavior* **4**: 1028-1034.

3.10 Supporting information

METHOD S 3.1 Weighted Correlation Network Analysis (WGCNA) script

WGCNA is an R package that can be briefly described as follow. A pair-wise Pearson correlation matrix was computed and then transformed into an adjacency matrix using a power function: adjacency value of two genes = $| \text{correlation between two genes} |^\beta$, where β is the weight. The adjacency matrix was transformed into a distance matrix using the topological overlap measure (Li & Horvath, 2007). A hierarchical clustering analysis was applied to the dissimilarity matrix (1 - distance matrix). Dendrogram leaves were cut and separated genes into distinct expression modules using the dynamic tree cut algorithm (Langfelder et al., 2008). An expression module can be defined as a cluster of highly (positively and negatively) correlated genes (Langfelder & Horvath, 2008). It reveals two negatively correlated expression profiles. Each expression profile is associated with a group of genes that we referred to a co-expression group in experiment 1 or temporally variable cluster in experiment 2 (Table S3.1). Below is the WGCNA script we used.

Loading expression data

Expression data is a matrix with genes in rows and tissues in columns

```
getwd();
workingDirectory="C:/Users/...";
setwd(workingDirectory);
library(WGCNA);
options(stringsAsFactors = FALSE);
datExp=read.table("ExpressionData.txt")
```

Transposing data and removing unnecessary data

```
datExpr0 = as.data.frame(t(datExp[,]));
names(datExpr0)=datExpr0[1,]
datExpr0=datExpr0[-1,]
```

Detecting outlier samples with hierarchical clustering analysis

no sample of experiments 1 and 3 was removed after the outlier verification

```
sampleTree = flashClust(dist(datExpr0), method = "average");
plot(sampleTree)
```

Choosing the soft-thresholding power: analysis of network topology

```
powers = c(c(1:10), seq(from = 12, to=20, by=2));
sft = pickSoftThreshold(datExpr0, powerVector = powers, verbose = 5);
plot(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fitIndices[,2],
xlab="Soft Threshold (power)", ylab="Scale Free Topology Model Fit,
signed R^2",type="n", main = paste("Scale independence"));
text(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fitIndices[,2],
labels=powers,cex=cex1,col="red");
abline(h=0.86,col="red")
plot(sft$fitIndices[,1], sft$fitIndices[,5],
xlab="Soft Threshold (power)",ylab="Mean Connectivity", type="n",
main = paste("Mean connectivity"))
text(sft$fitIndices[,1], sft$fitIndices[,5], labels=powers, cex=cex1,col="red")
```

Identifying expression modules

We used power = 14 and 18 to identify modules across tissues (Experiment 1)

and times (Experiment 3), respectively

```
net = blockwiseModules(datExpr0, power = 14, minModuleSize = 30,
maxBlockSize=12000, reassignThreshold = 0, mergeCutHeight = 0.25,
numericLabels = TRUE, pamRespectsDendro = FALSE, saveTOMs = TRUE,
saveTOMFileBase = "TOM1", verbose = 3)
```

Calculating eigengene of modules

```
MEs = moduleEigengenes(datExpr0, moduleColors)$eigengenes
MET =orderMEs(MEs)
```

References

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* 9: 559.

Langfelder P, Zhang B, Horvath S. 2008. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24: 719-720.

Li A, Horvath S. 2007. Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics* 23: 222-231

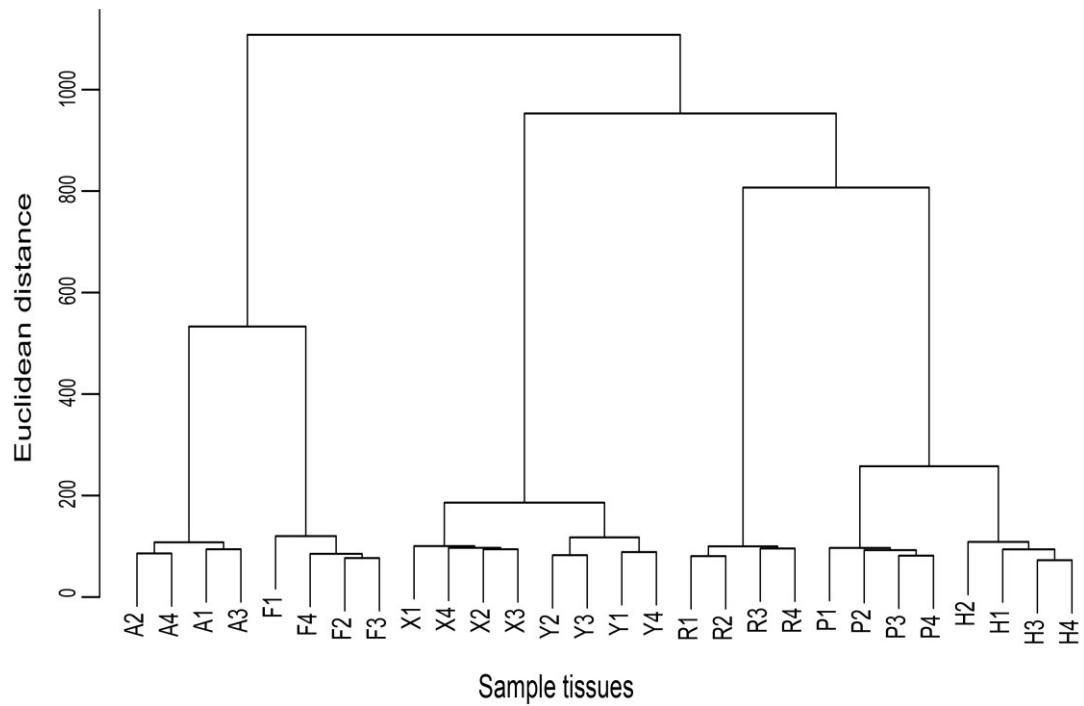


Fig. S3.1 Hierarchical clustering of white spruce (*Picea glauca*) tissues based on variable genes;

Analysis was performed using the *hclust* function of R with Ward's method (Ward, 1963) and microarray data (log₂ scale) of variable genes; A, shoot apex; X, shoot secondary xylem; Y, root secondary xylem; P, shoot phelloderm; H, root phelloderm; F, young foliage; and R, root tips; each tissue was replicated four times (labelled 1-4).

Reference

Ward JH. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58**: 236-244.

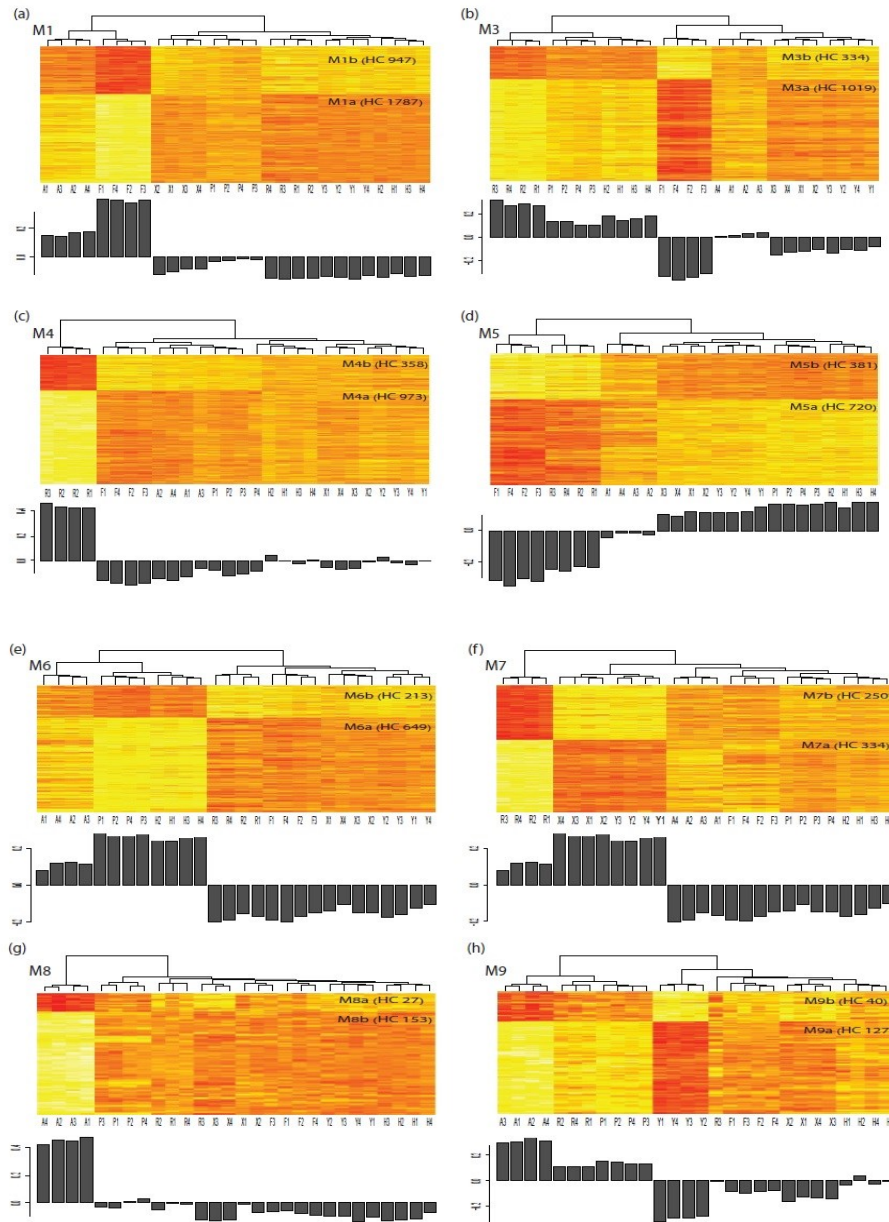


Fig. S3.2 Representation of co-expression groups and their correlation to white spruce (*Picea glauca*) tissues

(M1, M3-M9), expression modules; (M1a, M1b, M3a-M9b), co-expression groups; (a-h), Heatmaps show transcript abundance (log₂ scale) of high-confidence (HC) variable genes in expression modules and co-expression groups determined with the *WGCNA* package of R; Number in parentheses, number of high-confidence variable genes in the co-expression group; Total and low-confidence variable gene numbers are listed in Table S3.2; rows (Y axis), genes which are listed in Table S3.1; columns (X axis), tissues (A, shoot apex; X, shoot secondary xylem; Y, root secondary xylem; P, shoot phelloderm; H, root phelloderm; F, young foliage; and R, root tips); Each tissue was represented by four biological replicates (labelled 1-4); bar plots display the module eigengenes (or the first principal components).); (k), Heatmap shows the Pearson correlation coefficient between co-expression groups and tissue samples.

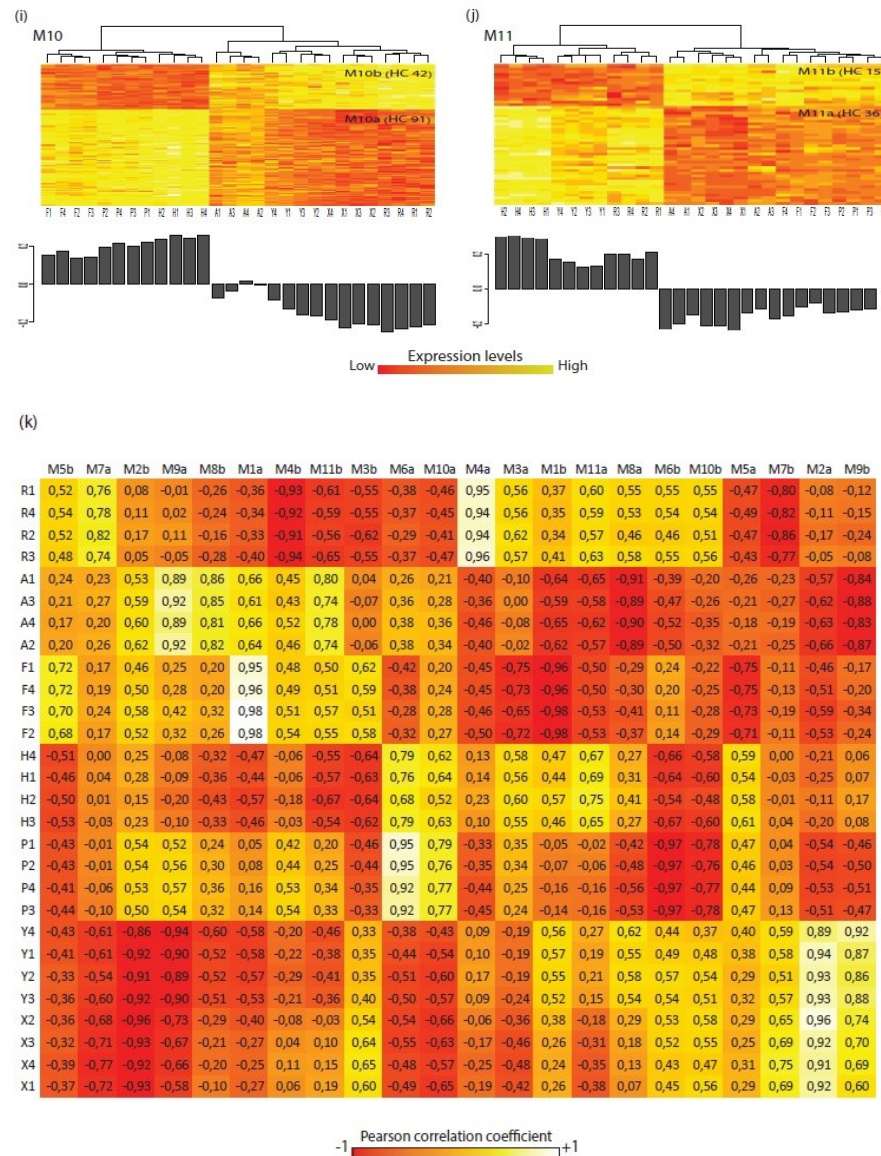


Fig. S3.2 (suite) Representation of co-expression groups and their correlation to white spruce (*Picea glauca*) tissues

(M10-M11), expression modules; (M10a-M11b), co-expression groups; (i-j), Heatmaps show transcript abundance (log2 scale) of high-confidence (HC) variable genes in expression modules and co-expression groups determined with the *WGCNA* package of R; Number in parentheses, number of high-confidence variable genes in the co-expression group; Total and low-confidence variable gene numbers are listed in Table S3.2; rows (Y axis), genes which are listed in Table S3.1; columns (X axis), tissues (A, shoot apex; X, shoot secondary xylem; Y, root secondary xylem; P, shoot phelloderm; H, root phelloderm; F, young foliage; and R, root tips); Each tissue was represented by four biological replicates (labelled 1-4); bar plots display the module eigengenes (or the first principal components); (k), Heatmap shows the Pearson correlation coefficient between co-expression groups and tissue samples.

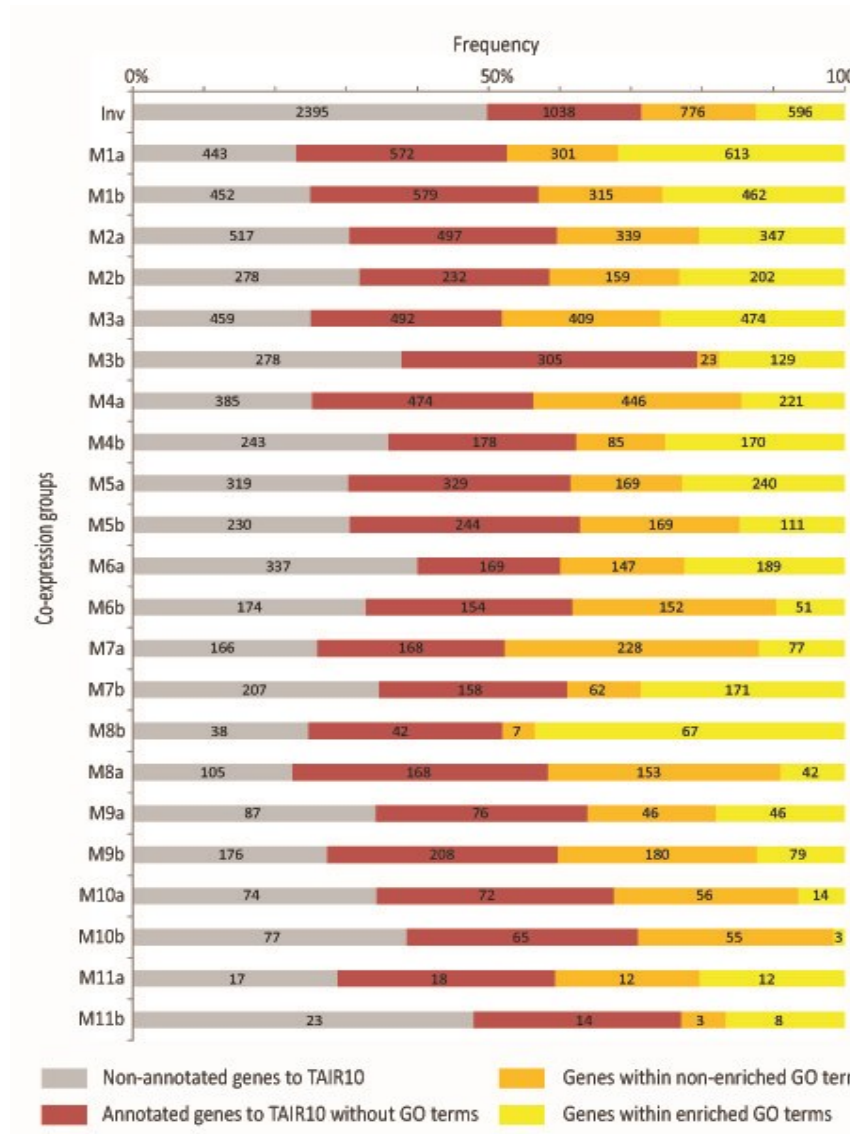


Fig. S3.3 Functional annotation of invariant genes and co-expression groups

Inv, invariant genes; M1a-M11b, co-expression groups; X axis, frequency (number in %) of genes that were associated to each of the four annotation categories: i) non-annotated genes to The *Arabidopsis* Information Resource version 10 (TAIR10, www.arabidopsis.org), ii) annotated genes to TAIR10 but without gene biological process gene ontology terms (GO terms), genes within iii) enriched and iv) non-enriched GO terms; Numbers inside each bar, numbers of genes in each category; Enrichment test was performed using DAVID tool (Database for Annotation, Visualization and Integrated Discovery; <http://david.abcc.ncifcrf.gov>, Huang *et al.*, 2009)

Reference

Huang DW, Sherman BT, Lempicki RA. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* 37: 1-13.

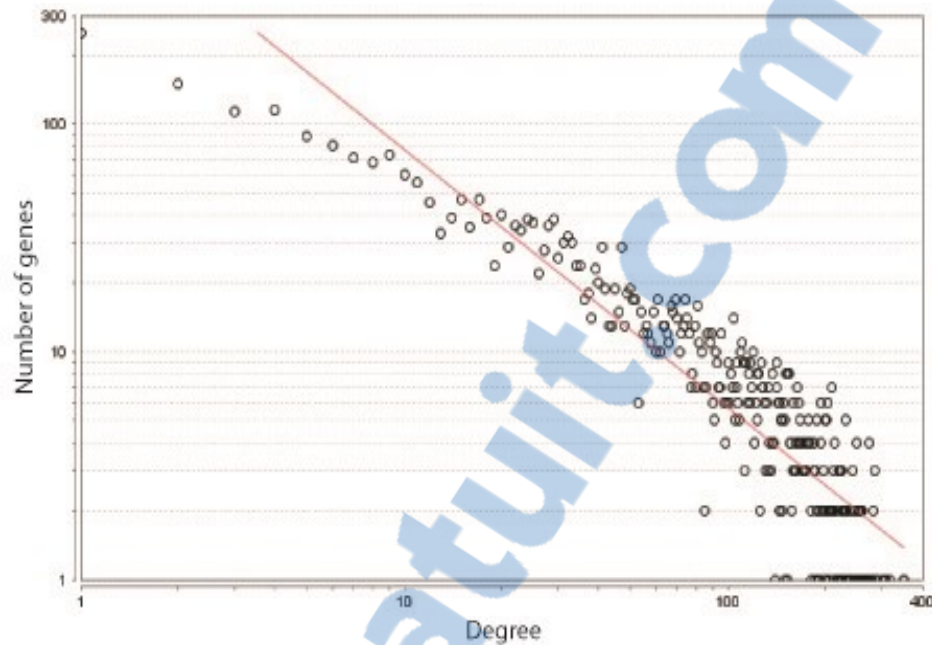


Fig. S3.4 Gene degree distribution in a double log scale of the xylem (M2-7) preferential network

The M2-7 network had a scale-free topology since its gene degree distribution fitted to a power-law distribution (*i.e.* few hubs hold together numerous small-degree nodes; Barabási *et al.*, 2001). Figure as output directly from the Cytoscape software (Assenov *et al.*, 2008). June-July and September preferential networks were not scale free.

References

Assenov Y, Ramirez F, Schelhorn S, Lengauer T, Albrecht M. 2008. Computing topological parameters of biological networks. *Bioinformatics* 24: 282-284.

Barabasi AL, Erzsébet R, Vicsek T. 2001. Deterministic scale-free networks. *Physica A* 299: 559-564.

Tables S3.1, S3.2 and S3.3 are not shown here but available in xls format at
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3534630/>

Table S3.1 Compilation of gene expression results from experiments 1 and 2 and gene network analysis and annotation summary

Table S3.2 Functional annotation analysis of gene groups

Table S3.3 RT-qPCR primer sequences and microarray validation by RT-qPCR

Table S3.4 Number of genes per co-expression group in white spruce (*Picea glauca*)

Co-expression group	High-confidence gene	Low-confidence gene	Total
M1a	1787	142	1929
M1b	947	861	1808
M2a	1418	282	1700
M2b	628	243	871
M3a	1019	815	1834
M3b	334	401	735
M4a	973	553	1526
M4b	358	318	676
M5a	720	333	1053
M5b	381	373	754
M6a	649	193	842
M6b	213	318	531
M7a	334	305	639
M7b	250	348	598
M8a	153	1	154
M8b	27	441	468
M9a	127	128	255
M9b	40	603	643
M10a	91	125	216
M10b	42	158	200
M11a	36	23	59
M11b	15	33	48

High- and low-confidence genes are genes differentially expressed across tissues at $\text{adj}P\text{-value} < 0.0001$ and $0.05 < \text{adj}P\text{-value} > 0.0001$, respectively; (M1a-M11b) are tissue-based co-expression groups that are graphically presented in Fig. 3.2a and Fig. S3.2a-j.

Table S3.5 Under or overrepresentation of conifer specific genes across co-expression groups

Co-expression group	Number of conifer specific genes	Underrepresentation	Overrepresentation
M1a	154	x	
M1b	171	x	
M2a	203		x
M2b	107		
M3a	162	x	
M3b	81		
M4a	148		
M4b	86		
M5a	138		
M5b	101		
M6a	158		x
M6b	58		
M7a	60		
M7b	72		
M8a	8		
M8b	36	x	
M9a	32		
M9b	49		
M10a	24		
M10b	30		
M11a	6		
M11b	7		

Conifer specific genes are *Picea glauca* (Moench) Voss genes whose sequence has significant match with a *P. sitchensis* (Bong.) or *Pinus teada* L. sequence (BLASTX, $E\text{-value} < 10^{-10}$) but has no significant match with any *Arabidopsis* sequence; x indicates co-expression groups where conifer specific genes were over or underrepresented at $\text{adj}P\text{-value} < 0.05$

Table S3.6 Enriched pathways in xylem (M2-7) preferential network

Enriched pathways ¹	Gene number	Genes ²	adjP-val ³ <0.05
ath00940:Phenylpropanoid biosynthesis	23	AT5G05340, AT3G18080, AT4G36220, AT5G51890, AT1G15950, AT4G36430, AT1G65060, AT2G41480, AT4G16270, AT3G21240, AT1G26560, AT4G34050, AT1G44970, AT2G44480, AT2G37040, AT1G05260, AT2G30490, AT3G53260, AT4G34230, AT2G18150, AT5G06720, AT1G71695, AT5G66390	6.881E-6
ath01061:Biosynthesis of phenylpropanoids	40	AT3G06350, AT5G05340, AT1G61720, AT1G48030, AT5G66120, AT4G33510, AT2G45300, AT4G36430, AT5G13420, AT2G37040, AT1G05260, AT2G18150, AT4G34230, AT5G06720, AT1G43670, AT3G29200, AT4G22870, AT3G51240, AT5G51890, AT4G36220, AT1G15950, AT5G49460, AT5G07990, AT2G42790, AT2G41480, AT1G65060, AT1G48850, AT4G16270, AT1G44970, AT4G34050, AT3G21240, AT2G30490, AT3G53260, AT3G55120, AT5G13930, AT2G16370, AT4G22880, AT4G27070, AT1G71695, AT5G66390	1.07E-4
ath00941:Flavonoid biosynthesis	10	AT5G07990, AT4G34050, AT5G42800, AT1G61720, AT3G51240, AT2G30490, AT3G55120, AT5G13930, AT4G22880, AT4G22870	3.41E-4
ath00460:Cyanoamino acid metabolism	9	AT4G39640, AT1G26560, AT2G44480, AT3G18080, AT5G22300, AT3G61440, AT1G73050, AT4G13930, AT3G07990	0.00587
ath00360:Phenylalanine metabolism	15	AT5G05340, AT5G51890, AT4G36430, AT2G41480, AT4G16270, AT1G44970, AT2G37040, AT1G05260, AT2G30490, AT3G53260, AT2G18150, AT2G42490, AT5G06720, AT1G71695, AT5G66390	0.0092
ath00680:Methane metabolism	14	AT2G44160, AT5G05340, AT5G43940, AT5G51890, AT4G13930, AT4G36430, AT2G41480, AT4G16270, AT1G44970, AT1G05260, AT2G18150, AT5G06720, AT1G71695, AT5G66390	0.02720

¹identified using DAVID tool (Database for Annotation, Visualization and Integrated Discovery; <http://david.abcc.ncifcrf.gov>; Huang *et al.*, 2009); ² Gene identifiers according to TAIR (The Arabidopsis Information Resource); ³ Benjamini-Hochberg correction (Benjamini & Hochberg, 1995)

References

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**: 289-300.

Huang DW, Sherman BT, Lempicki RA. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**: 1-13.

Table S3.7 Members of the secondary cell wall gene network shown to be transactivated by PgNAC-7 or MYB transcription factors in white spruce (*Picea glauca*)

GenBank Acc. Nbr	Cluster ID ¹	Gene name ²	Function	Trans activation ³	PgNAC-7 (R>0.9) ⁴	PgMyb8 (R>0.9) ⁴	Co-expression group	Temporal cluster ⁵	Preferential expression in spruces ⁶
BT102049	GQ0165_B14	PgNAC-7	NAC-domain transcription factor			x	M2a	T2a	Xylem
BT108136	GQ03117_E18	PgMYB8	myb domain protein	a, b	x		M2a	None	Xylem
BT108631	GQ03124_J09	PgMYB1	myb domain protein	a, b			M3b	T2a	Xylem
BT108414	GQ03121_H22	PgMYB2	myb domain protein	b	x		M2a	T2a	Xylem
BT107883	GQ03113_N22	PgMYB4	myb domain protein	b	x		M2a	T2a	Xylem
BT116706	GQ03805_C07	PgDHS-2	3-deoxy-d-arabino-heptulosonate 7-phosphate synthase	a, b	x		M2a	None	Xylem
BT115945	GQ03709_D11	PgCM-3	Chorismate mutase-1	b	x		M2a	None	Xylem
BT112280	GQ03312_O11	PgCAD	cinnamyl alcohol dehydrogenase	a, b	x	x	M2a	None	Xylem
BT106671	GQ03007_K12	Pg4CL	4-coumarate:CoA ligase	a, b			M2a	None	Xylem
BT111802	GQ03303_N02	PgCCR-1	cinnamoyl coa reductase	b	x		M2a	None	Xylem
BT111350	GQ03236_G10	PgPRR-2	pinoresinol reductase	b	x	x	M2a	None	Xylem
BT116976	GQ03810_K09	PgCesA-3	Cellulose synthase family protein	a, b	x		M2a	T1b	Xylem
BT106155	GQ02904_A20	PgTUA-1	tubulin alpha-4 chain	a, b			M6a	None	None
BT116258	GQ03714_M12	PgXTH8-1	xyloglucan endotransglucosylase / hydrolase	a, b			M4a	T1b	Phelloderm

¹ From by Rigault *et al.* (2011); ² From Duval *et al.* (2014); ³ Functional evidence of transcriptional control was obtained from a, transient co-transfection with PgNAC-7, PgMYB-8 or both (Duval *et al.*, 2014) or b, stable over-expression of PgMYB8 (Bomal *et al.*, 2008); ⁴ Correlations with PgNAC-7 and PgMYB-8 from network analyses; ⁵ Temporal cluster T2a and T1b were highly and moderately June-July preferential genes, respectively; None indicates no temporal variation detected; ⁶ From Raherison *et al.* (2012); conserved preferential expression in three spruces (*Picea glauca*, *P. abies*, *P. sitchensis*) comparing stem secondary xylem and phelloderm tissues determined by microarray

References

Bomal C, Bedon F, Caron S, Mansfield SD, Levasseur C, Cooke JE, Blais S, Tremblay L, Morency M-J, Pavy N. 2008. Involvement of *Pinus taeda* MYB1 and MYB8 in phenylpropanoid metabolism and secondary cell wall biogenesis: a comparative in planta analysis. *Journal of experimental botany* **59**: 3925-3939.

Duval I, Lachance D, Giguère I, Bomal C, Morency M-J, Pelletier G, Boyle B, MacKay JJ, Séguin A. 2014. Large-scale screening of transcription factor–promoter interactions in spruce reveals a transcriptional network involved in vascular development. *Journal of experimental botany* **65**: 2319-2333.

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

Table S3.8 Distribution of detected genes across invariant gene and co-expression groups

	Foliage	Shoot apex	Shoot phelloderm	Shoot xylem	Root phelloderm	Root xylem	Root tips
Inv	4554	4414	4638	4511	4606	4375	4492
M1a	1928	1925	1908	1882	1892	1818	1840
M1b	1797	1791	1807	1808	1807	1806	1806
M2a	1675	1671	1686	1700	1685	1696	1661
M2b	871	869	868	849	868	838	860
M3a	1807	1824	1834	1816	1833	1810	1831
M3b	734	733	730	732	726	728	715
M4a	1469	1472	1492	1472	1512	1471	1524
M4b	675	675	672	669	672	661	654
M5a	1042	1049	1052	1051	1053	1048	1032
M5b	754	748	743	733	738	725	749
M6a	831	839	842	830	842	819	822
M6b	524	519	526	521	522	523	530
M7a	623	632	631	609	633	594	636
M7b	591	591	596	598	597	597	584
M8a	139	154	145	136	134	105	124
M8b	467	465	467	467	467	467	468
M9a	249	255	254	247	254	237	252
M9b	637	633	641	641	640	642	634
M10a	215	212	216	210	216	210	210
M10b	196	196	196	198	196	191	199
M11a	57	54	58	53	59	57	59
M11b	48	48	48	48	47	46	46
Unclassified	500	502	506	496	497	487	492
Tested_genes	22383	22271	22556	22277	22496	21951	22220
Detected_genes	23104	22951	23296	22970	23225	22608	22913

Inv, invariant genes; M1a-M11b, co-expression groups; Unclassified, variable genes that could not be classified to any co-expression groups using WGCNA or template matching methods; Tested genes, detected gene that met assumptions for ANOVA testing (for details, see methods); Detected genes, genes whose transcript levels were detected above background levels in the tissue

Chapitre 4.

Conclusions générales et perspectives

Dans ce chapitre, nous revenons sur les résultats majeurs que nous avons obtenus et nous en discutons. Nous cernons ensuite les limites de nos résultats et suggérons des études complémentaires qui pourraient être envisagées pour combler des lacunes. Enfin, nous citons des retombées pratiques potentielles découlant de nos résultats et expliquons brièvement les enjeux forestiers que devraient saisir les recherches dans le futur.

4.1 Discussion des résultats majeurs

Motivées par l'importance économique et écologique des conifères, des études génomiques ont été menées afin de comprendre les mécanismes moléculaires pouvant expliquer des phénomènes biologiques reliés à la formation du bois et aux stratégies d'adaptation face aux facteurs environnementaux chez ceux-ci. Plusieurs chercheurs se sont mis à étudier l'expression de gènes ou à analyser les profils transcriptionnels. Ils se sont fixés des objectifs spécifiques pour lesquels ont été identifiés des gènes à expression différentielle entre les stades de développement et les conditions biologiques (arbres stressés *versus* non stressés; arbres résistants à une maladie *versus* non-résistants). Ils ont étudié un petit nombre de tissus ou d'organes, mais n'ont que d'effleurer la variation de l'expression de gènes dans différents tissus de l'arbre. En outre, peu d'auteurs ont traité de l'aspect évolutif de l'expression de gènes en comparant différentes espèces.

Pour nos travaux, nous nous sommes fixés des objectifs à portée fondamentale se rapportant à l'organisation, à l'évolution et à la dynamique du transcriptome chez les conifères. Nous avons construit la première puce à oligonucléotide et nous avons utilisé la méthode d'hybridation dans nos expériences. Nous avons réalisé principalement deux analyses pour étudier la variation du niveau d'expression des gènes sur plusieurs tissus. Ces deux analyses sur plusieurs tissus diffèrent par leurs objectifs respectifs, la nature des données et les analyses statistiques effectuées. Nous avons également réalisé une expérience sur des tissus vasculaires de trois espèces d'épinette pour étudier la conservation des profils d'expression de gènes entre les espèces.

4.1.1 La première puce à oligonucléotide

Nous avons développé la première puce à oligonucléotide pour le profilage transcriptionnel chez les conifères. Quand on la compare à la plupart des puces utilisées dans les travaux antérieurs, cette puce présente deux avantages :

- Premièrement, elle est parmi celles qui présentent une plus grande couverture du génome. Elle comporte près de 24 000 gènes; ce qui représente environ 80 % du nombre total de gènes de l'épinette blanche selon l'estimation de Rigault et coll. (2011). Le nombre de gènes que contient la puce est aussi comparable au nombre de gènes analysés à l'échelle du génome selon la méthode de séquençage à ARN chez le pin tordu et l'épinette de l'intérieur (Yeaman et coll., 2014).
- Deuxièmement, cette puce à oligonucléotide de 70 nucléotides de longueur permet également d'obtenir une plus grande spécificité en limitant les hybridations croisées, contrairement aux puces constituées d'amplicons (ADNc), qui sont généralement plus longs et moins spécifiques en hybridation (Chou et coll., 2004).

De plus, par rapport aux puces à oligonucléotides plus courts (moins de 30 nucléotides de longueur [Pullat et coll., 2007]), notre puce est tolérante. Bien que composée des séquences d'épinette blanche, elle est compatible avec d'autres épinettes. Compte tenu de la conservation des séquences codantes chez les conifères (p.ex. 80 à 86 % selon Rigault et coll., 2011; 83,5 % selon Chen et coll., 2012; 82 % selon Mann et coll., 2013; 99 % selon Niu et coll., 2013; 85 % selon Liu et coll., 2013; 87,5 à 99 % selon Canales et coll., 2014; 94 % selon Wanfeng et coll., 2014; et 84 % selon Li et coll., 2014), la puce pourrait être utilisable pour d'autres genres de conifères en éliminant par exemple les gènes dont la séquence diverge relativement peu de celle de l'épinette blanche.

4.1.2 Étude sur plusieurs tissus: construction d'une base de données d'expression

La première étude sur plusieurs tissus présentée dans cette thèse (chapitre 2) a compilé et intégré des données issues de plusieurs expériences pour créer une base de données d'expression chez l'épinette. Les données utilisées dans cette étude sont moins contrôlées. Les tissus sont représentés par un nombre de répétitions différent allant de 3 à 60. Ils ont été aussi récoltés sur des arbres d'âges différents, introduisant un

échantillonnage de plusieurs stades de développement. Toutefois, pour pallier les variations techniques entre les expériences, nous avons appliqué la transformation en rang aux données et avons établi des profils basés sur des classes d'expression semi-quantitative (données quantitatives discontinues). Ces résultats d'expression tissulaire de près de 24 000 gènes forment la première base de données d'expression construite chez les conifères.

Cette base de données d'expression tissulaire que nous avons appelée PiceaGenExpress a servi de référence à d'autres études. Par exemple, Mageroy et coll. (2015) ont trouvé que *Pgβglu-1* a une expression préférentielle dans les aiguilles et chez les arbres d'épinette blanche résistants contre la tordeuse des bourgeons de l'épinette. En étudiant la structure de gènes, Sena et coll. (2014) ont pu vérifier que les gènes d'expression élevée et ubiquitaire ont plus de séquences introniques que les gènes préférentiels à des tissus. Enfin, Pavy et coll. (2013) ont rapporté que les gènes avec des déficits de substitutions non synonymes présentent une expression élevée et ubiquitaire alors que les gènes avec excès de substitutions non synonymes sont préférentiellement exprimés à un seul ou deux tissus.

PiceaGenExpress représente moins de tissus et de conditions quand on le compare aux bases de données d'expression établies sur des espèces modèles. À titre de comparaison, AtGenExpress est une base de données d'expression d'*Arabidopsis* qui contient, entre autre, des profils selon des tissus et aussi des profils selon des stades de développement et des conditions biologiques. De plus, AtGenExpress a été le fruit d'une collaboration entre des équipes de recherche de divers pays (Japon, Suisse, Angleterre, Allemagne et États-Unis). (<https://www.arabidopsis.org/portals/expression/microarray/ATGenExpress.jsp>).

BioGPS est une des bases de données développées sur l'humain et la souris (<http://biogps.org/#goto=welcome>) et contient l'expression de gènes sur 17 fois plus de tissus que PiceaGenExpress.

4.1.3 Étude sur plusieurs tissus: organisation du transcriptome et réseaux transcriptionnels

La deuxième étude sur plusieurs tissus avait pour objectif d'étudier l'organisation du transcriptome et de construire des réseaux (chapitre 3). Les tissus proviennent d'un seul groupe d'arbres du même âge, qui ont été cultivés ensemble et échantillonnés en même temps. Les données sont équilibrées (même nombre de répétitions pour tous les tissus analysés) et générées à partir d'une même expérience. Ce design expérimental a permis de faire des analyses statistiques quantitatives, dont un test d'ANOVA et un test de corrélation de Pearson, à l'instar de la base de données PiceaGenExpress qui permettait des analyses semi-quantitatives. L'ANOVA a servi à identifier les gènes dont le niveau d'expression varie significativement (gènes variables) ou non (gènes invariables) selon les tissus. Le test de corrélation de Pearson a été utilisé pour identifier l'interaction dans les réseaux transcriptionnels des gènes.

Cette étude a permis de mettre en évidence plusieurs phénomènes /conclusions;

- La différenciation tissulaire est évolutivement ancienne (environ 380 Ma selon Kenrick et Crane, 1997) et donc profonde chez les conifères. Environ 80 % des gènes analysés ont varié significativement suivant les tissus. Cette proportion de gènes variables entre les tissus est plus grande que ce qui a été trouvé par rapport à d'autres facteurs (stade de développement et condition biologique) selon les travaux de Paiva et coll. (2008), de Li et coll. (2010) et de Verne et coll. (2011). Une plus grande dynamique du transcriptome associée aux tissus a également été rapportée chez l'*Arabidopsis* (Aceituno et coll., 2008) et la souris (Su et coll., 2002; Pommerenke et coll., 2012).
- Le transcriptome de l'épinette blanche possède une organisation modulaire. Cette organisation se manifeste par la répartition des gènes en groupes de coexpression. Un groupe de coexpression est l'ensemble des gènes qui partagent le même profil d'expression tissulaire. La subdivision des gènes en groupes de coexpression selon des tissus a été aussi rapportée chez d'autres plantes dont le tabac (*Nicotiana tabacum* L., Edwards et coll., 2010) et le maïs (*Zea mays* L., Downs et coll., 2013) et chez des mammifères tels que l'humain et la souris (Su et coll., 2002; Onorati et coll., 2014).
- L'analyse des annotations fonctionnelles, qui se base sur les gènes avec des fonctions connues, a été effectuée pour chacun des groupes de coexpression. Cette analyse a permis d'identifier les fonctions moléculaires ou les processus

biologiques surreprésentés au sein de chacun des groupes. Par inférence, les résultats de coexpression permettent de reconnaître les processus biologiques dans lesquels seraient impliqués plus de 7500 gènes non annotés selon l'approche de similarité des séquences. Ce chiffre représente environ les 60 % de gènes dont les fonctions prédites restent inconnues d'après Rigault et coll. (2011). Cette méthode d'annotation fonctionnelle basée sur la coexpression de gènes a été utilisée par d'autres auteurs (p.ex. Summers et coll., 2010; Mabbott et coll., 2011). La méthode d'inférence selon l'approche de coupable par association se base sur les propos d'auteurs incluant Allocco et coll. (2004) et Mabbott et coll. (2011). Ainsi, selon Allocco et coll. (2004), les gènes coexprimés sont potentiellement corégulés et interviennent dans les mêmes processus biologiques. Et d'après Mabbott et coll. (2011), les gènes qui codent pour des protéines impliquées dans une voie métabolique donnée sont souvent coexprimés.

- La coexpression des gènes a été utilisée depuis plusieurs années pour construire des réseaux transcriptionnels (p.ex. Stuart et coll., 2003; Zheng et Zhao, 2013; Zhao et coll., 2015). Dans ces réseaux, deux gènes présentent une interaction quand leurs profils sont corrélés selon un certain seuil du coefficient de Pearson. Dans le cadre de cette étude, nous avons utilisé la coexpression tissulaire et nous avons identifié un réseau transcriptionnel dans le xylème. Plusieurs des gènes qui composent ce réseau interviendraient dans des voies métaboliques secondaires d'après leurs annotations fonctionnelles (p.ex. voie de phénylpropanoïde, voie de flavonoïde). Comme plusieurs réseaux biologiques (p.ex. Barabási et coll., 2001; Lorenz et coll., 2011; Netotea et coll., 2014), notre réseau est invariant d'échelle (*scale free*), c'est-à-dire qu'il contient peu de gènes fortement connectés (hub) et plusieurs gènes faiblement connectés. Le facteur de transcription NAC-7 de l'épinette blanche (PgNAC-7) a été identifié comme étant le plus important hub ou le gène le plus fortement connecté du réseau. Des études fonctionnelles ont montré l'implication des orthologues de PgNAC-7 chez l'*Arabidopsis* (Kubo et coll., 2005) et le peuplier (Zhong et coll., 2010) dans la synthèse de la paroi secondaire cellulaire. De récents travaux ont rapporté que PgNAC-7 agirait aussi comme un régulateur transcriptionnel dans la synthèse de la paroi secondaire (Duval et coll., 2014). L'interaction de PgNAC-7 avec le

promoteur de certains gènes, dont PgMYB8, PgCAD et PgDHS a été indiquée par une analyse de transactivation (Duval et coll., 2014). Dans le cadre de cette thèse, nous avons également réalisé une expérience où nous analysons la dynamique temporelle du transcriptome. Cette expérience a permis de préciser que PgNAC-7 est actif dans le xylème pendant la formation du bois initial au cours de la saison de croissance. La même expérience a également permis la construction d'un réseau impliqué dans la formation du bois final. Les gènes hubs de ce deuxième réseau sont des gènes de réponse au stress. Cette découverte indique que le réseau de coexpression change lors de la formation du bois final. Les changements observés reflètent l'apparition de stress associés aux conditions environnementales prévalant vers la fin de la saison de croissance, soit une disponibilité réduite en eau et des températures ambiantes plus froides

4.1.4 Conservation de profil d'expression

Nous avons analysé la conservation des profils d'expression de gènes entre le xylème secondaire et le phelloderme chez trois espèces d'épinettes, soit l'épinette blanche, l'épinette de Sitka, et l'épinette noire. Ces espèces ont un temps de divergence de 7 à 27 millions d'années ou Ma (Lockwood et coll., 2013). Nous avons découvert que 99 % des gènes différentiels ont un profil d'expression conservé chez les épinettes étudiées. Cette proportion est plus grande que celle (74 %) rapportée par Yeaman et coll. (2014), qui ont comparé des profils de gènes orthologues en réponse à des changements de conditions environnementales entre deux espèces plus éloignées (pin tordu et épinette de l'intérieur; 140 Ma). Le nombre de gènes préférentiels au xylème secondaire conservés entre les épinettes est environ 80 fois plus grand que celui rapporté par Pavy et coll. (2008). Cette différence n'est pas surprenante étant donné que ces derniers ont analysé moins de gènes et comparé l'épinette blanche et l'*Arabidopsis*; deux espèces qui sont phylogénétiquement beaucoup plus éloignées (gymnosperme *versus* angiosperme; 300 Ma) et morphologiquement différentes (arbre *versus* plante herbacée). Des travaux effectués chez des vertébrés ont montré la conservation des profils d'expression tissulaire de plus d'un tiers des gènes orthologues analysés seulement (Chan et coll., 2009). Cette faible proportion est attribuable au fait que les auteurs ont analysé un plus grand nombre de tissus (provenant du cerveau, des yeux, du cœur, des muscles, de l'intestin, de l'estomac, du

foie, du rein, de la rate et des testicules) et d'espèces (quatre espèces éloignées dont l'humain, la souris, la volaille *Gallus gallus* L. [310 Ma], la grenouille *Xenopus tropicalis* G. [360 Ma] et le poisson *Tetraodon nigroviridis* Marion de Procé [450 Ma]).

En bref, nous avons développé la première puce à oligonucléotides. Cette puce a été utilisée pour mesurer l'expression génique lors d'expériences réalisées dans le cadre de cette étude. Des analyses des données d'expression nous ont donné la possibilité de créer une base de données permettant d'explorer le transcriptome des conifères. D'autres ont permis d'étudier l'organisation du transcriptome et sa conservation entre des espèces et de construire des réseaux de gènes. Par rapport à ces réalisations, d'autres questions méritent d'être creusées à l'aide d'études complémentaires. Pour ces dernières, l'utilisation de l'approche de séquençage (RNA-seq) représente une méthodologie de recherche particulièrement prometteuse.

4.2 Études complémentaires

4.2.1 L'approche RNA-seq pour les études complémentaires

Dans le cadre de cette étude, nous avons utilisé la méthode d'hybridation dont les résultats ont été validés à l'aide de méthodes indépendantes telles que le qPCR et le RNA-seq. Toutefois, les résultats issus de cette approche présentent certaines limites quant au nombre de gènes ou de transcrits analysés. Les puces à ADNs sont reconnues pour avoir une plage de dynamique d'expression moins élevée que le RNA-seq (p.ex. Zhao et coll., 2014). Elles sont sujettes à une saturation d'hybridation pour détecter les gènes très fortement exprimés. À cause du bruit de fond, elles ont une faible sensibilité pour détecter les transcrits rares.

Comparée au RNA-seq, la méthode d'hybridation possède une capacité de détection plus faible. On sait qu'un gène peut produire différents transcrits selon l'épissage alternatif. La méthode d'hybridation ne permet pas de détecter le niveau d'expression des gènes et des transcrits dont la séquence n'est pas imprimée sur la puce. Notre puce comporte plus de 25 000 séquences de transcrits réparties dans près de 24 000 gènes d'épinette blanche. Certains auteurs ayant utilisé le RNA-seq ont pu découvrir un nombre de gènes plus élevé que ce que nous avons sur la puce. Bien que plus élevés, les chiffres peuvent être très différents à cause des variations d'origine technique entre les études, notamment la méthodologie bio-informatique qui est mise en œuvre pour



traiter les données. Par exemple, Liu et coll. (2013) ont utilisé l'assemblage de *novo* et ont identifié près de 40 000 gènes d'une longueur moyenne d'environ 1300 pb dans les aiguilles du pin argenté. Avec la même technique d'assemblage, Wang et coll. (2013) ont rapporté un nombre de gènes 1,5 fois plus élevé que Qiu et coll. (2013), mais avec une longueur moyenne de séquences de seulement 497 pb dans les tissus cambiaux de *Cunninghamia lanceolata* (L.) Hook. Li et coll. (2014) sont les rares auteurs ayant utilisé un génome de référence et qui ont trouvé un nombre de gènes comparable au nôtre dans *Picea balfouriana* (Rehd. Et Wils.) Hillier. Leurs résultats semblent raisonnables grâce à une grande proportion des séquences (85 %) alignées aux séquences d'épinette de Norvège.

Le RNA-seq est relativement nouveau et ne dispose pas nécessairement de techniques d'analyse aussi standardisées que la méthode d'hybridation. Mais avec l'amélioration de l'assemblage et l'annotation de génomes de références des conifères, ce problème d'analyses diminuera sans doute. Par ailleurs, grâce au progrès de la technologie de séquençage à haut débit qui réduira son coût, et compte tenu des avantages qu'il offre, le RNA-seq remplacerait à terme la méthode d'hybridation. L'utilisation de cette technique pour l'étude des conifères requiert l'amélioration des techniques d'analyse et l'utilisation d'un génome de référence pour réduire les erreurs au niveau de l'assemblage des séquences.

4.2.2 Des questions pour les études complémentaires

Au terme de notre étude, des questions biologiques restent sans réponses et méritent d'être creusées. Ces questions touchent d'une part l'aspect évolutif de la différenciation tissulaire et des réseaux de gènes et d'autre part l'aspect fonctionnel des gènes.

- En utilisant l'expression génique, nous avons démontré la divergence profonde des tissus chez les conifères. Pour arriver à cette conclusion, nous avons déterminé le nombre de gènes à expression différentielle entre les tissus. Ce nombre a été comparé à des chiffres provenant d'autres études. Il serait intéressant de confirmer cette profonde divergence tissulaire à l'aide d'analyses plus robustes. Ces analyses devraient être effectuées sur des données issues d'une expérience dans laquelle seraient considérées différentes sources de variation biologique dont le tissu, le stade de développement, la condition biologique et l'espèce.

- Dans le cadre de cette étude, un grand réseau transcriptionnel du bois de l'épinette blanche a été construit. Des groupes de coexpression qui permettent de construire des réseaux préférentiels à d'autres types de tissus (phellodermes, tissus photosynthétiques) ont été également identifiés. L'aspect évolutif qui consiste à étudier la conservation et la divergence des réseaux entre les espèces de conifères et d'autres plantes n'a pas été abordé. La conservation des réseaux serait intéressante dans la mesure où elle montre la régulation des gènes potentiellement conservés entre les espèces et permettrait d'élucider la fonction des gènes (p.ex. Stuart et coll., 2003). L'étude de la divergence des réseaux est justifiée notamment par le fait que selon certains auteurs, la divergence au niveau de la régulation des gènes contribue plus au phénomène de spéciation que la divergence des séquences codantes (p.ex. Romero et coll., 2012).
- L'analyse d'annotation fonctionnelle a permis d'identifier les processus biologiques dans lesquels seraient impliqués les groupes de coexpression. En se basant sur la coexpression, des fonctions peuvent être hypothétiquement attribuées à des gènes qui n'ont pas été annotés selon la méthode basée sur la similarité de séquences. Ces fonctions nouvellement attribuées devraient être vérifiées à l'aide des études de perte ou de gain de fonction (p.ex. suppression, surexpression; Bomal et coll., 2008) ou d'interaction moléculaire. À titre d'exemple, les résultats sur les réseaux suggéraient des interactions potentielles entre des facteurs de transcription et des gènes. Ces interactions pourraient être testées grâce à la méthode de retard sur gel afin de montrer la liaison aux séquences régulatrices et grâce à l'expression transitoire afin de montrer l'activation de la transcription tel montré dans Duval et coll. (2014).

4.3 Retombées pratiques et enjeux forestiers

Notre étude porte sur l'organisation, l'évolution et la dynamique du transcriptome chez les conifères et peut servir de référence dans l'interprétation des résultats d'autres études (p.ex. Pavy et coll., 2013; Sena et coll., 2014). Elle relève plus de la recherche fondamentale, mais peut néanmoins permettre d'identifier des gènes qui sont de bons candidats pour servir de marqueurs diagnostiques pour des programmes

d'amélioration. Par exemple, le gène *Pgβglu-1* a été découvert par Mageroy et coll. (2015) en utilisant PiceaGenExpress (chapitre 2) et peut être considéré comme un marqueur d'expression pour identifier des arbres résistants contre la tordeuse des bourgeons de l'épinette dans les forêts. Les variations phénotypiques liées à la résistance contre la tordeuse ont pu être seulement reliées avec des variations au niveau de l'expression, mais non au niveau des séquences de ce gène (Mageroy et coll., 2015). De la même façon, des marqueurs d'expression ont été identifiés et suggérés pour diagnostiquer la maladie d'Alzheimer à un stade précoce chez l'humain (Lunnon et coll., 2013). Un autre exemple est l'utilisation des gènes candidats par expression (chapitre 3) dans l'étude de polymorphismes pour identifier des marqueurs diagnostiques. Les gènes candidats par expression peuvent être utilisés comme données d'entrée dans les études d'association qui identifient des SNPs, ou marqueurs génétiques, reliés à un phénotype d'intérêt (Lamara et coll., inédit). Les marqueurs génétiques peuvent être utilisés dans le processus de sélection assistée par marqueurs (SAM). Cette technique permet d'identifier dès leur bas âge les arbres qui possèdent des gènes d'intérêt sans devoir passer par des cultures expérimentales. Par rapport aux méthodes conventionnelles d'amélioration, la SAM permet entre autres une réduction considérable de la longueur des cycles d'amélioration (Neale et Kremer, 2011).

Parmi les grands enjeux forestiers que les chercheurs doivent aborder, on retrouve les changements climatiques ainsi que les insectes ravageurs et maladies. Face aux changements climatiques, les arbres peuvent s'acclimater, s'adapter et migrer. Le rythme des changements est beaucoup plus rapide que la capacité d'adaptation naturelle des arbres (à cause de leur longévité; Savolainen et coll., 2007). C'est pourquoi les forêts sont particulièrement sensibles aux changements rapides des conditions climatiques. Des simulations ont par exemple montré que plus de 60 % des essences forestières des régions boréales et tempérées s'ajusteraient difficilement au réchauffement climatique prédit en 2085 (Hamann et Wang, 2006). Le réchauffement climatique entraîne des perturbations qui incluent la sécheresse et la pullulation d'insectes ravageurs ainsi que de certains agents pathogènes. Une sécheresse sévère a causé la perte d'un million d'ha de peuplier faux-tremble (*Populus tremuloides* Michx.) en Saskatchewan et en Alberta (Hogg et coll., 2008). La prolifération de petits scolytes a entraîné la mortalité de plus d'un million d'ha d'essences d'épinettes en Alaska (*Dendroctonus rufipennis* [Kirby]; Berg et coll., 2006) et plus de 10 millions d'ha de pin tordu en Colombie-Britannique (*Dendroctonus ponderosae* Hopkins; Kurz

et coll., 2008). À cause de la mondialisation des échanges commerciaux, un oomycète (*Phytophthora ramorum* W., De Cock & Man in't Veld) qui infectait initialement les feuillus (chêne vert ou *Quercus ilex* L., chêne à tan ou *Lithocarpus densiflora* [Hook & Arn.] R.) aux États-Unis se retrouve en Angleterre depuis 2009 et s'adapte à un conifère (mélèze du Japon; Brasier et Webber, 2010).

Face à ces enjeux, les études transcriptomiques devraient contribuer dans le futur à la compréhension des mécanismes moléculaires de base d'adaptation aux contraintes abiotiques et biotiques des arbres. Ces informations s'avéreront utiles lors du processus décisionnel dans la gestion des volets santé et production des forêts existantes.

4.4 Références bibliographiques

- Aceituno FF, Moseyko N, Rhee SY, Gutiérrez RA. 2008.** The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in *Arabidopsis thaliana*. *BMC Genomics* **9**(1): 438.
- Allocco DJ, Kohane IS, Butte AJ. 2004.** Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* **5**(1): 18.
- Barabási A-L, Ravasz E, Vicsek T. 2001.** Deterministic scale-free networks. *Physica A: Statistical Mechanics And Its Applications* **299**(3): 559-564.
- Berg EE, Henry JD, Fastie CL, De Volder AD, Matsuoka SM. 2006.** Spruce beetle outbreaks on the Kenai Peninsula, Alaska, and Kluane National Park and Reserve, Yukon Territory: relationship to summer temperatures and regional differences in disturbance regimes. *Forest Ecology And Management* **227**(3): 219-232.
- Bomal C, Bedon F, Caron S, Mansfield SD, Levasseur C, Cooke JE, Blais S, Tremblay L, Morency M-J, Pavy N. 2008.** Involvement of *Pinus taeda* MYB1 and MYB8 in phenylpropanoid metabolism and secondary cell wall biogenesis: a comparative in planta analysis. *Journal Of Experimental Botany* **59**(14): 3925-3939.
- Brasier C, Webber J. 2010.** Plant pathology: sudden larch death. *Nature* **466**(7308): 824-825.
- Canales J, Bautista R, Label P, Gómez-Maldonado J, Lesur I, Fernández-Pozo N, Rueda-López M, Guerrero-Fernández D, Castro-Rodríguez V, Benzekri H. 2014.** De novo assembly of maritime pine transcriptome: implications for forest breeding and biotechnology. *Plant Biotechnology Journal* **12**(3): 286-299.
- Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe M, Wilde A, Brudno M. 2009.** Conservation of core gene expression in vertebrate tissues. *J Biol* **8**(3): 33.

- Chen J, Källman T, Ma X, Gyllenstrand N, Zaina G, Morgante M, Bousquet J, Eckert A, Wegrzyn J, Neale D. 2012.** Disentangling the roles of history and local selection in shaping clinal variation of allele frequencies and gene expression in Norway spruce (*Picea abies*). *Genetics* **191**(3): 865-881.
- Chou C-C, Chen C-H, Lee T-T, Peck K. 2004.** Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Research* **32**(12): e99-e99.
- Downs GS, Bi Y-M, Colasanti J, Wu W, Chen X, Zhu T, Rothstein SJ, Lukens L. 2013.** A developmental transcriptional network for *Zea mays* defines coexpression modules. *Plant Physiology* **161**(4): 1830-1843.
- Duval I, Lachance D, Giguère I, Bomal C, Morency M-J, Pelletier G, Boyle B, MacKay JJ, Séguin A. 2014.** Large-scale screening of transcription factor–promoter interactions in spruce reveals a transcriptional network involved in vascular development. *Journal Of Experimental Botany* **65**(9): 2319-2333.
- Edwards KD, Bombarely A, Story GW, Allen F, Mueller LA, Coates SA, Jones L. 2010.** TobEA: an atlas of tobacco gene expression from seed to senescence. *BMC Genomics* **11**(1): 142.
- Hamann A, Wang T. 2006.** Potential effects of climate change on ecosystem and tree species distribution in British Columbia. *Ecology* **87**(11): 2773-2786.
- Hogg E, Brandt J, Michaelian M. 2008.** Impacts of a regional drought on the productivity, dieback, and biomass of western Canadian aspen forests. *Canadian Journal Of Forest Research* **38**(6): 1373-1384.
- Kenrick P, Crane PR. 1997.** The origin and early evolution of plants on land. *Nature* **389**(6646): 33-39.
- Kubo M, Udagawa M, Nishikubo N, Horiguchi G, Yamaguchi M, Ito J, Mimura T, Fukuda H, Demura T. 2005.** Transcription switches for protoxylem and metaxylem vessel formation. *Genes & Development* **19**(16): 1855-1860.
- Kurz WA, Dymond C, Stinson G, Rampley G, Neilson E, Carroll A, Ebata T, Safranyik L. 2008.** Mountain pine beetle and forest carbon feedback to climate change. *Nature* **452**(7190): 987-990.
- Lamara M, Raherison E, Lenz P, Beaulieu J, Bousquet J, MacKay J. (inédit).** Genetic architecture of wood properties based on association analysis and co-expression networks in white spruce.
- Li Q, Zhang S, Wang J. 2014.** Transcriptome analysis of callus from *Picea balfouriana*. *BMC Genomics* **15**(1): 553.
- Li X, Wu HX, Southerton SG. 2010.** Seasonal reorganization of the xylem transcriptome at different tree ages reveals novel insights into wood formation in *Pinus radiata*. *New Phytologist* **187**(3): 764-776.
- Liu J-J, Sturrock RN, Benton R. 2013.** Transcriptome analysis of *Pinus monticola* primary needles by RNA-seq provides novel insight into host resistance to *Cronartium ribicola*. *BMC Genomics* **14**(1): 884.
- Lockwood JD, Aleksić JM, Zou J, Wang J, Liu J, Renner SS. 2013.** A new phylogeny for the genus *Picea* from plastid, mitochondrial, and nuclear sequences. *Molecular Phylogenetics And Evolution* **69**(3): 717-727.

- Lorenz WW, Alba R, Yu Y-S, Bordeaux JM, Simões M, Dean JF. 2011.** Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (*P. taeda* L.). *BMC Genomics* **12**(1): 264.
- Lunnon K, Sattlecker M, Furney SJ, Coppola G, Simmons A, Proitsi P, Lupton MK, Lourdusamy A, Johnston C, Soininen H. 2013.** A blood gene expression marker of early Alzheimer's disease. *Journal Of Alzheimer's Disease* **33**(3): 737-753.
- Mabbott NA, Kenneth Baillie J, Kobayashi A, Donaldson DS, Ohmori H, Yoon SO, Freedman AS, Freeman TC, Summers KM. 2011.** Expression of mesenchyme-specific gene signatures by follicular dendritic cells: insights from the meta-analysis of microarray data from multiple mouse cell populations. *Immunology* **133**(4): 482-498.
- Mageroy MH, Parent G, Germanos G, Giguère I, Delvas N, Maaroufi H, Bauce É, Bohlmann J, Mackay JJ. 2015.** Expression of the β -glucosidase gene Pg β glu-1 underpins natural resistance of white spruce against spruce budworm. *The Plant Journal* **81**(1): 68-80.
- Mann IK, Wegrzyn JL, Rajora OP. 2013.** Generation, functional annotation and comparative analysis of black spruce (*Picea mariana*) ESTs: an important conifer genomic resource. *BMC Genomics* **14**(1): 702.
- Neale DB, Kremer A. 2011.** Forest tree genomics: growing resources and applications. *Nature Reviews Genetics* **12**(2): 111-122.
- Netotea S, Sundell D, Street NR, Hvidsten TR. 2014.** ComPIEx: conservation and divergence of co-expression networks in *A. thaliana*, *Populus* and *O. sativa*. *BMC Genomics* **15**(1): 106.
- Niu S-H, Li Z-X, Yuan H-W, Chen X-Y, Li Y, Li W. 2013.** Transcriptome characterisation of *Pinus tabuliformis* and evolution of genes in the *Pinus* phylogeny. *BMC Genomics* **14**(1): 263.
- Onorati M, Castiglioni V, Biasci D, Cesana E, Menon R, Vuono R, Talpo F, Goya RL, Lyons PA, Bulfamante GP. 2014.** Molecular and functional definition of the developing human striatum. *Nature Neuroscience* **17**: 1804-1815.
- Paiva J, Garnier-Géré P, Rodrigues J, Alves A, Santos S, Graça J, Le Provost G, Chaumeil P, Silva-Perez D, Bosc A. 2008.** Plasticity of maritime pine (*Pinus pinaster*) wood-forming tissues during a growing season. *New Phytologist* **179**(4): 1180-1194.
- Pavy N, Boyle B, Nelson C, Paule C, Giguère I, Caron S, Parsons LS, Dallaire N, Bedon F, Bérubé H. 2008.** Identification of conserved core xylem gene sets: conifer cDNA microarray development, transcript profiling and computational analyses. *New Phytologist* **180**(4): 766-786.
- Pavy N, Gagnon F, Rigault P, Blais S, Deschênes A, Boyle B, Pelgas B, Deslauriers M, Clément S, Lavigne P. 2013.** Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. *Molecular Ecology Resources* **13**(2): 324-336.

- Pommerenke C, Wilk E, Srivastava B, Schulze A, Novoselova N, Geffers R, Schughart K. 2012.** Global transcriptome analysis in influenza-infected mouse lungs reveals the kinetics of innate and adaptive host immune responses. *PloS One* 7(7): e41169.
- Pullat J, Fleischer R, Becker N, Beier M, Metspalu A, Hoheisel JD. 2007.** Optimization of candidate-gene SNP-genotyping by flexible oligonucleotide microarrays; analyzing variations in immune regulator genes of hay-fever samples. *BMC Genomics* 8(1): 282.
- Qiu Z, Wan L, Chen T, Wan Y, He X, Lu S, Wang Y, Lin J. 2013.** The regulation of cambial activity in Chinese fir (*Cunninghamia lanceolata*) involves extensive transcriptome remodeling. *New Phytologist* 199(3): 708-719.
- Rigault P, Boyle B, Lepage P, Cooke JE, Bousquet J, MacKay JJ. 2011.** A white spruce gene catalog for conifer genome analyses. *Plant Physiology* 157(1): 14-28.
- Romero IG, Ruvinsky I, Gilad Y. 2012.** Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics* 13(7): 505-516.
- Savolainen O, Pyhäjärvi T, Knürr T. 2007.** Gene flow and local adaptation in trees. *Annual Review of Ecology, Evolution, And Systematics*: 595-619.
- Sena JS, Giguère I, Boyle B, Rigault P, Birol I, Zuccolo A, Ritland K, Ritland C, Bohlmann J, Jones S. 2014.** Evolution of gene structure in the conifer *Picea glauca*: a comparative analysis of the impact of intron size. *BMC Plant Biology* 14(1): 95.
- Stuart JM, Segal E, Koller D, Kim SK. 2003.** A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302(5643): 249-255.
- Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A. 2002.** Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy Of Sciences* 99(7): 4465-4470.
- Summers KM, Raza S, Van Nimwegen E, Freeman TC, Hume DA. 2010.** Co-expression of FBN1 with mesenchyme-specific genes in mouse cell lines: implications for phenotypic variability in Marfan syndrome. *European Journal Of Human Genetics* 18(11): 1209-1215.
- Verne S, Jaquish B, White R, Ritland C, Ritland K. 2011.** Global transcriptome analysis of constitutive resistance to the white pine weevil in spruce. *Genome Biology And Evolution* 3: 851-867.
- Wanfeng L, Suying H, Liwang Q. 2014.** Transcriptome resources and genome-wide marker development for Japanese larch (*Larix kaempferi*). *Frontiers of Agricultural Science And Engineering* 1(1): 77-84.
- Wang Z, Chen J, Liu W, Luo Z, Wang P, Zhang Y, Zheng R, Shi J. 2013.** Transcriptome characteristics and six alternative expressed genes positively correlated with the phase transition of annual cambial activities in Chinese Fir (*Cunninghamia lanceolata* (Lamb.) Hook). *PloS One* 8(8): e71562.
- Yeaman S, Hodgins KA, Suren H, Nurkowski KA, Rieseberg LH, Holliday JA, Aitken SN. 2014.** 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* **203**(2): 578-591.

Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. 2014. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS One* **9**(1): e78644.

Zhao X, Liu Z-Y, Liu Q-X. 2015. Gene coexpression networks reveal key drivers of phenotypic divergence in porcine muscle. *BMC Genomics* **16**(1): 50.

Zheng Z-L, Zhao Y. 2013. Transcriptome comparison and gene coexpression network analysis provide a systems view of citrus response to ‘*Candidatus Liberibacter asiaticus*’ infection. *BMC Genomics* **14**(1): 27.

Zhong R, Lee C, Ye Z-H. 2010. Functional characterization of poplar wood-associated NAC domain transcription factors. *Plant Physiology* **152**(2): 1044-1055.