PHYLOGENOMIC ANALYSIS OF THE TRANSCRIPTOME OF SPRUCE (GENUS PICEA)

by

Rokneddin M. Albouyeh

M.F.C. University of Toronto, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Forestry)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

June 2011

Abstract

Trees are sessile and exposed to a plethora of pests throughout their life-span and natural history. Diversification of the specialized (secondary) metabolism is known as a key factor in the co-evolution of trees with pests. This work focuses on tissue-related gene expression, and the expression of specialized pathways of phenolics and terpenoids in relation to the evolution of biochemical defenses in the coniferous species of spruce (genus Picea). Gene expression resources were assessed for the superimposition of tissue-related expression and cross-species expression profiling. Five species of spruce, *P. abies*, *P. glauca*, *P. jezoensis*, *P. omorika*, and *P. mariana*, were used to infer the evolution of gene expression among representative spruce species. As gene expression also depends upon tissue, I examined three sources of tissue: needles, outer stem (bark and the attached phloem) and xylem. The overall expression of phenolics was significantly diverged in the outer stem. At the gene family level, expression was predominantly stabile among species. Significant among-species divergence of gene expression, indicative of diversifying selection, was found for eight gene families. These families were: cinnamate 4- hydroxylases (C4H), dirigent-like proteins (DIR), glycosyl transferases (GLYTR), laccases (LAC), O-methyl transferases (OMT), phenylalanine ammonia lyases (PAL), putative caffeoyl CoA O-methyl transferases (pCCoAOMT) and putative phenylcoumaran benzylic ether reductases (pPCBER). Analysis of the expression of the terpenoid pathways in the outer stem revealed that for terpene synthase gene family (TPS), expression is significantly diverged among species. In a novel approach, heritability of gene expression using parent-offspring regression was inferred for Interior spruce (*P. glauca* x *engelmannii*), and average expression for TPS genes showed substantial heritability.

Preface

The present dissertation is built on a basic concept, differential expression of genes. On a larger scale, comparative genomics and evolutionary theories are used to explain the evolution of tissue-related gene expression and the expression of specialized biochemical pathways in the genus *Picea* (spruce). The outcome includes four data chapters:

Chapter 2 is published as Albouyeh, R., Farzaneh, N., Bohlmann, J., and Ritland K. 2010. Multivariate analysis of digital expression profiles identifies a xylem signature of the vascular tissue of white spruce (*Picea glauca*). Tree Genet. Genomes 6: 601-611. I proposed exploration of digital expression data for application in comparative and evolutionary genomics. I devised the statistical and bioinformatics strategies and prepared the manuscript. Nima Farzaneh built the contiguous sequences (contigs) and annotated the genes. Rick White helped performing two-way divisive cluster analysis. Drs. Joerg Bohlmann and Kermit Ritland offered constructive comments on the preparation of the manuscript. Dr. Kermit Ritland also provided feedback on data analysis and validation.

Chapter 3 is a prepared manuscript. I proposed the framework, cross-species comparisons. I devised the bioinformatics strategy; screened the genes; carried out the lab work; wrote data description and submitted the expression data to Gene Expression Omnibus (GEO); performed phylogenetic and statistical analyses; and wrote the manuscript. Rick White suggested the experimental design for microarray profiling, and pre-processed the expression data. Nima Farzaneh annotated the genes. Mack Yuen prepared the data files for GEO submission. Dr. Joerg Bohlmann shared data that helped compilation of the phenolics gene list. Dr. Carol Ritland supervised the lab work. Dr. Bjoern Hamberger offered expertise on the biosynthesis of phenolics; and revised the manuscript. Dr. Kermit Ritland identified a

suitable Picearetum; assigned a data base for contig assembly (Spruce V8); provided feedback on the overall design and data analysis; and revised the manuscript.

Chapter 4 is published as Albouyeh, R., and Ritland, K. 2009. Estimating heritability of gene expression using parent-offspring regression with 2-channel microarrays. J. Hered. 100: 114-118. I proposed the framework, heritability of gene expression using parent-offspring regression. I suggested publication of the experimental design as a stand-alone research; and wrote the manuscript. Dr. Kermit Ritland suggested alternative experimental designs; wrote the code for the simulations; and revised the manuscript.

Chapter 5 is a prepared manuscript. I carried out the lab work; wrote data description and submitted the expression data to GEO; devised the bioinformatics strategy; screened and compiled the list of terpenoid genes; analyzed the data; and wrote the manuscript. Rick White pre-processed the microarray data. Nima Farzaneh annotated the genes. Mack Yuen prepared the data files for GEO submission. Dr. Carol Ritland coordinated the lab work. Dr. Joerg Bohlmann advised regarding the biochemical aspects of the analysis. Dr. Kermit Ritland identified a suitable collection of Interior spruce (*Picea glauca* x *engelmannii*) for sampling; wrote the code to compute heritability values, provided feedback on data analysis, and revised the manuscript.

Table of Contents

List of Tables

List of Figures

Acknowledgements

My Ph.D. project along with my immense learning experience in this course would have not been possible without Kermit Ritland. I was inspired by his vision, knowledge, and his superior analytical skills. Thank you Kermit.

I would like to thank my committee, Joerg Bohlmann, and Keith Adams, for their advice and support of the project. I would also like to thank our project manager, Carol Ritland, for guidance in the lab; Yousry El-Kassaby for mentorship and encouragement; Barry Jaquish, and Bonny Hooge from British Columbia Forest Service, for the coordination of field works.

Members of Ritland Lab, Genetic Data Centre (G.D.C.), and Treenomix Conifer Health Program offered help or support in various ways throughout the project. Many thanks are due (in alphabetical order) to: Buschiazzo, E.; Chao, S.; Chen, C.; Cullis, C.; Dauwe, R., Farzaneh, N.; Gillan, T.; Hamberger, B.; Kolosova, N.; Leung, G.; Liew, C.; Lippert, D.; Mateiu, L.; Miscampbell, A.; Nguyen, A.; Porth, I.; Sun, M.; Tabanfar, L.; Tang, M.; Verne, S.; White, R.; Wytrykush, D.; Yueh, H.; Yuen, A.; Yuen, M.

I am grateful to my family; in particular my parents, Shahrokh and Shirin; to my brother Nouri; my sister Shadi, and my aunts, Sherry and Shahla, for their whole-hearted support. I would like to express my gratitude to Caroline Ames, for her friendship and support; and to Tali Conine, for her kindness and generosity.

I was financially assisted by UBC Faculty of Graduate Studies Tuition Fee Awards, and Faculty of Forestry Internal Award (Vandusen).This project was funded by Genome Canada and Genome British Columbia, and NSERC Discovery grants to Dr. Kermit Ritland.

1. Introduction

1.1 Prospects for the phylogenomic study of spruce transcriptome

Trees are long-lived, perennial organisms of a sessile nature. These features result in constant exposure to a plethora of stress agents throughout their life-span and natural history. Therefore, as biological systems, trees have unique challenges for adaptation and evolution.

Black cottonwood (*Populus trichocarpa*) is the primary model tree. Genomic studies of poplar provide understanding of the basic biotic interactions and adaptive traits unique to trees (Douglas and Jansson, 2007). In coniferous trees, genomic studies have been hindered by large genome sizes typical of conifers (10-40 Gbp), which are about two orders of magnitude larger than the genome of angiosperm tree poplar (450 Mbp). Furthermore, expansion of genes as families and paralogy within these gene families, hamper discovery and functional assignment of genes (Neal and Ingvarsson, 2008; Ralph et al. 2008).

For species such as conifers, an alternative for the discovery of genes and their functional assessment is the study of gene expression using transcriptome resources (Allona et al. 1998; Pavy et al. 2005; Ralph et al. 2008). For many lines of investigation in biology, transcriptome resources are of primary importance. In particular, research aimed at understanding differences of various tissue types (e.g. Whitehead and Crawford, 2005), or adaptation to the environments (e.g. Whitehead and Crawford, 2006), are rooted in the study of gene expression. Layered on top of this is the opportunity to use evolutionary comparisons in conifers.

Phylogenomics in the broad sense can be regarded as the intersection between the fields of evolution and genomics (Philippe and Blanchette, 2007). Within this paradigm, one can make use of evolutionary theory to assess the functionality and impact of genes in adaptive evolution.

The present dissertation is built on extensive transcription profiling. My basic questions involve the importance of tissue sources, and the expression of the genes in specialized (secondary) biochemical pathways; an engine that drives the sessile trees parallel in co-evolution with biotic stress agents. Furthermore, I demonstrate how two major concepts of the evolutionary theory, Darwinian selection and heritability, can be applied to transcriptome data to provide answers to some of these questions.

In this introductory chapter I provide an overview of the crucial factors that influenced my research objectives: (1) genus *Picea* as a study system in the family Pinaceae; (2) evolution of coniferous defense transcriptome; and (3) large-scale availability of transcriptome resources for spruce. I will then outline the research objectives of the data chapters for the analysis of spruce transcriptome.

1.2 Evolutionary properties of the genus Picea (spruce) as a system in Pinaceae

Spruce (*Picea* A. Dietr.) is a genus with 34 species in the coniferous family Pinaceae (Pine family) which comprises 11 genera and more than 200 species (Farjon, 1990). Molecular phylogenetic studies in Pinaceae mark a divergence time of 120-140 million years (MYR) between *Picea* and *Pinus*, and more recent divergence times of 13-20 MYR within *Picea* (Wang et al. 2000; Bouille and Bousqouet, 2005). A comparison of Expressed Sequence Tags (ESTs) contiguous sequences (contigs) between *Picea* and *Pinus* has demonstrated an average synonymous substitution rate of about 4×10^{-10} per year (Ritland et al. 2006). This is five to ten times less than what is observed for angiosperms such as poplar. Therefore, a relatively relaxed rate of evolution at both family and genus level can be inferred.

Coupled with the above, is the conservation of chromosome numbers across Pinaceae (except for Douglas fir, the base chromosome number across Pinaceae is generally 12).

Therefore, genetic maps, single nucleotide polymorphisms (SNPs) and expression profiles can be expected to be integrated in Pinaceae.

The phylogeny of spruce is not fully worked out at the level of the genus. However, for North American species, European species (i.e. *P. abies*, *P. omorika*) and some Asian species considerable works on phylogenies as well as biogeography exist in the literature (Bouille and Bousqouet, 2005; Campbell et al. 2005; Ran et al. 2006; Aizawa et al. 2007; Nasri et al. 2008; Tollefsrud et al. 2008). Moreover, monophyly of the species of *Picea* has never been debated (Ran et al. 2006).

In comparison, the closely related genus *Pinus* comprises about 111 species (Gernandt et al. 2005), with two major divisions (hard and soft pines; subgenera *Pinus* and *Strobus*), and further complexity within these two divisions. Therefore, comparative analysis of pines is more complicated. This is analogous to the problem that population structure poses to association genetics; variation of relatedness confounds inferences.

1.3 Perspectives on the study of the evolution of spruce defense transcriptome

Among various factors that contribute to the survival of conifers as ancient species is the evolution of their potent defenses. A comprehensive review by Franceschi et al. (2005) suggested a conceptual model for defense against insect pests. This model described an arms race involving co-evolution of conifer defenses and bark beetles, and focused on bark as a defense frontline against organisms trying to reach phloem tissues. It also highlighted diversification of the chemical compounds related to two biosynthetic pathways; phenolics and terpenoids.

In concert with the seminal review of Franceschi et al. (2005), conifer genomics has been increasingly focused on expression and structural level analyses of genes involved with these pathways, for purposes of identification in the processes of tree defense, as well as

functional assignment. These studies vary from case-control experiments (e.g. Ralph et al. 2006); and phylogenetic analyses (e.g. Hamberger and Bohlmann, 2006; Ralph et al. 2007) to targeted functional studies (e.g. Keeling et al. 2008). Nonetheless, overall structure of these pathways and the magnitude of evolutionary forces (various selection regimes, neutrality) on conservation or diversification of the different points of the pathways at the level of the transcriptome in conifers are largely unknown.

Finding the signature of natural selection on gene expression has significant practical implications in screening candidate genes for their adaptive importance. For instance, Whitehead and Crawford (2006) examined the expression of genes relative to an ecological parameter (temperature), and found a suit of metabolic genes to be under natural selection.

In coniferous species of spruce, Holliday et al. (2008) screened a suite of candidate genes for the adaptive trait of cold hardiness based on their pattern of expression. Ralph et al. (2008), summarizing the results of large-scale sequencing of spruce cDNAs, has concluded that an enormous capacity of coniferous genomes is invested in defense response. The study of the evolution of gene expression presents opportunities for understanding the adaptive role of genes in conifer defense.

1.4 Spruce transcriptome resources

A fundamental argument for the selection of a study system in comparative and evolutionary studies is "practicality" (Kellogg and Shaffer, 1993). Because of large conifer genomes (10-40 Giga base pair; Gbp) and consequent high content of repetitive DNA, conifers are not practical systems in forest genomics (Neale and Ingvarsson, 2008). Here, I would argue that transcriptome resources are "practical" regardless of the physical size of the genome, as the coding size of the transcribed genome are largely constant among higher plants, relative to total genome size.

1.4.1 Genomes

The completed genome sequencing projects lack representatives from any coniferous tree species in National Center for Biotechnology Information (NCBI) Plant Genome Central (http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList_hn.html, last accessed July 2010). Nevertheless, in anticipation of the completion of the sequencing project of Norway spruce (*Picea abies*), the first coniferous genome to be sequenced, (http://www.upsc.se/News/Latest/scientists-receive-sek-75-millions-to-map-the-genes-of-swedens-most-important-plant-norway-spruce.html), transcriptome resources have seen a considerable development. An additional project lead by Dr. David Neale, University of California at Davis, has recently been funded by the United States Department of Agriculture (USDA) to sequence not just one, but three conifer genomes using next generation sequencing (http://dendrome.ucdavis.edu/NealeLab/). These three species are Loblolly pine (*Pinus taeda*), Sugar pine (*P. lambertiana*), and Douglas fir (*Pseudotsuga menziesii*).

1.4.2 EST data bases

As of July, 2010 a total of 313,110 ESTs have been deposited for white spruce placing it behind only loblolly pine (*Pinus taeda*) on top of the list of all tree species at the National Center for Biotechnology Information EST database (ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). Gathered from a variety of tissue types, the dbEST division of NCBI comprises 186,637 EST from Sitka spruce (*P. sitchensis*), which ranks the fifth tree after horticultural species *Malus* and *Citrus*; 28,174 from interior spruce (*P. engelmannii* x *glauca*); 14,224 from Norway spruce and to a minor extent from black spruce. As a part of Genome Canada and Genome British Columbia projects (Treenomix and Treenomix II: Conifer Forest Health, http://www.treenomix.ca and Arborea

http://www.arborea.ulaval.ca) 23,589 and 6,464 full-length cDNAs (FL-cDNA) have been fully sequenced from white spruce and Sitka spruce respectively.

As an outcome of the availability of the EST resources, three generations of cDNA microarray platforms have been produced through the activities of the Treenomix projects. The first ever spruce cDNA microarray consisted of 9,700 unique clones in 2002. Followed a year later, a 16,700 (16.7 K) array and an 18.8 K array of unique clones of Interior spruce platforms released in 2005 (Ritland et al. 2006).

The use of the spruce cDNA platforms for the purpose of gene expression profiling at the level of species (cross-species) is justified following Bar-Or et al. (2007) who reviewed the literature of cross-species hybridizations. In order to obtain valid results from cross-species hybridization, these authors considered phylogeny and the length of the microarray probes as the major criteria. To achieve consistent reproducibility, they suggested longer length of the probes (i.e. typical of cDNA platforms), and the divergence between the target and reference species to be less than ~ 65 MYR. It is critical to recognize that not all genes diverge at similar rates among species, and that the generalization suggested by Bar-Or et al. (2007) has to be considered carefully. Nevertheless, because of the relaxed evolutionary rate of spruce species, it is expected that cross-species gene expression profiling among species of spruce would not be largely biased.

The applicability of the third generation Treenomix microarray platform for cross-strain (within-species) hybridizations, as well as limited cross-species hybridizations has been demonstrated by the hybridization of the RNA from Sitka spruce populations (Holliday et al. 2008).

1.5 Objectives of the dissertation

The principle objective of this dissertation is twofold: (1) Establishment of general phenomena surrounding the evolution of coniferous defense transcriptome; (2) Enhancement of the essential basic knowledge that is required for subsequent functional studies in spruce.

1.5.1 Multivariate analysis of digital expression profiles identifies a xylem signature of the vascular tissue of white spruce (*Picea glauca*)

Accomplishment of the above-mentioned objectives required exploring and assessment of the available resources for conifer comparative and evolutionary genomics. The second chapter of my dissertation was, therefore, dedicated to exploring digital expression profiles as an available transcriptome resource at the initiation of my research. This chapter started as a hypothesis-free research. The general objective was to assess the suitability of digital expression data as an independent data source versus microarray resources in respect to two prominent features: (1) cross-species comparisons; and (2) tissue related profiling. My research identified several limitations of the digital data for comparative analysis. However, exploratory analyses demonstrated that using this resource, tissue related gene expression can be differentiated within the vascular tissues of spruce. A practical implication of comparing tissues is screening variants of genes, such as those of secondary metabolism, that show differences in tissue-related expression. This research was also justified as an efficient way of utilizing digital expression data.

1.5.2 Evolution of the expression of phenolic gene families in the outer stem of spruce (genus Picea)

Felsenstein (1985) described the idea of sampling phylogenetic trees to infer biology (i.e. without the need for the representation of the full phylogeny). My third chapter is laid out within this context. I selected five species to create an evolutionary feature in the

expression data. The rationale was to maximize the genetic distance, while avoiding species with unresolved phylogenies.

Much of the chapter is based on the analysis of the divergence of phenolics gene expression among the species and across the tissues. The major hypothesis for this chapter questions the relevance of source of tissue in divergence of phenolics gene expression. The discussion expands on divergence of the expression of phenolic gene families coherent with their role in adaptive evolution. In addition, this chapter improves functional understanding of the uncharacterized coniferous phenolic gene families through evolutionary analysis.

1.5.3 Estimating heritability of gene expression using parent-offspring regression with 2-channel microarrays

Chapter four is inspired by the possibility of treating the abundances of gene transcripts as a phenotype. It introduces and discusses several alternative experimental designs for estimating heritability of gene expression using 2-channel microarrays. This chapter uses simulation to examine the power of various experimental designs. It is an original genetical genomics research, and the first of its kind to consider and compare such alternatives.

1.5.4 A novel experiment reveals heritability of terpenoid gene expression

This chapter deals with the expression of terpenoid pathways, which form a major class of natural products known or suggested to have roles in ecological functions. This chapter is inspired by the efficiency of transcript profiling to assess the impact of various pathway points in plant fitness. The hypothesis for this chapter is that there are differences for heritability of terpenoid pathway segments. This chapter also demonstrates how the connection between heritability and diversification can highlight functional relevance.

2. Multivariate analysis of digital expression profiles finds a xylem signature of the vascular tissue of white spruce (*Picea glauca*)

2.1 Introduction

Global analysis of gene expression is the simultaneous assessment of a large number of transcripts selected in an unbiased way. There are two main ways to achieve this: 1) the analogue approach, which uses either cDNA microarrays (Schena et al. 1995) or oligomer chips (Lockhart et al. 1996) and hybridization signal intensity to estimate transcript abundance, or 2) the digital approach, which uses counts of expressed sequence tags (ESTs) to estimate transcript abundance. Microarray-based methods are capable of assessing gene expression from numerous biological replicates, allowing for deep sampling of the tissue types. However, the assessment of gene expression using microarray technology is confined to the genes represented on the array. On the other hand, the tag-based methods, often referred to as an "open system", require no *a priori* knowledge of the genes, and hence are advantageous for gene discovery.

Conventional ESTs have been employed extensively for digital profiling of gene expression in the plant kingdom. This approach can be either "global" or "specific". An example of the "global" approach is the pioneering work of Ewing et al. (1999) who used clustering methods and raster displays (heatmaps) for a large-scale survey of digital gene expression in rice. Characterization of the grape transcriptome (da Silva et al. 2005) and the assembly of cotton ESTs (Udall et al. 2006) are other examples of the global approach. Examples of "specific" (or targeted) approaches to digital profiling with ESTs include the study of resistance genes in sugarcane (Wanderley-Nogueira, et al. 2007), cytochrome P450 genes in legumes (Li et al. 2007) and stress-responsive genes in rice (Gorantla et al. 2007).

In the past decade, differentiation and development of the vascular tissue has received attention in the conifer genomics literature (e.g. Allona et al. 1998; Zhang et al. 2000; Kirst et al. 2003; Zhang et al. 2003; Peter and Neale, 2004). Especially, as a subject of downstream research, development of wood forming xylem tissue has been under intensive study for identification of the genes that could improve wood quality (Paiva et al. 2008; Ukrainetz et al. 2008 a, b). Moreover, certain aspects of the development of the vascular system are important in the chemical ecology and evolution of the coniferous trees. For instance, the induction of phenolic and terpenoid conifer defenses is known to involve phloem (Ralph et al. 2007) and stem xylem (Martin et al. 2002; Franceschi et al. 2005; Miller et al. 2005; Keeling and Bohlmann, 2006a, b) tissues.

A number of recent studies have started to look at the tissue-specific patterns of gene expression within the vascular system with hybridization-based methodology. Using an Arabidopsis full-genome microarray, Ehlting et al. (2005) studied different stages of vascular differentiation and identified a group of genes associated with fiber development, secondary wall formation and lignifications. Foucart et al. (2006) used suppression subtractive hybridization (SSH; Diatchenko et al. 1996) method for transcript profiling of xylem versus phloem in Eucalyptus and found preferential expression of a set of gene involved in xylogenesis, hormone signaling, metabolism and proteolysis in xylem. In conifers, preferential expression of a set of defense genes, cell-wall modification and lignin biosynthesis, genes encoding protein kinases, and transcription factors is reported in co-ordination with the development of secondary xylem in spruce apical shoot growth (Friedmann et al. 2007).

While studies of digital expression have identified the expression signatures unique to phloem (in melon, Omid et al. 2007) and xylem (in poplar, Sterky et al. 1998; and in loblolly pine, Pavy et al. 2005a), studies of digital gene expression profiles in plant vascular tissue are few. Tissue types can be classified with respect to their global patterns of gene expression. Comparisons of transcript abundances across tissue types can identify tissue-specific markers, and in complement with sequence similarity, provide tentative annotations of newly discovered genes.

The ESTs of white spruce (*P. glauca*), and its hybrid, interior spruce (*P. glauca* x *engelmanii*) in the dbEST division of National Centre for Biotechnology Information (NCBI) comprise a collection of cDNA libraries that came from the efforts of two large-scale projects, Arborea I and II (http://www.arborea.ulaval.ca/) and Genome British Columbia Treenomix I and II (http://www.treenomix.ca; Pavy et al. 2005b; Ralph et al. 2006; Ralph et al. 2008). Representing a variety of tissues and organs, the collection of spruce cDNA libraries is a rich resource for the study of digital gene expression; in particular, libraries derived from vascular tissue are potentially suitable for profiling of xylem and phloem gene expression.

In this study, ESTs from the existing cDNA libraries of xylem and phloem tissue in white spruce and hybrid white spruce (Interior spruce) are analyzed with the specific objective of finding a set of transcripts that could serve as markers within the vascular tissue of conifers, representing signature of their source of tissue. To achieve this objective, a heuristic strategy is pursued that 1) finds a combination of ESTs and libraries that could maximize the phloem-xylem polarity; and 2) screens the elements (transcripts) with respect

to their proximity to the extremes of the poles. Results are corroborated with data derived from wet-lab experiment involving cDNA microarray spotted slides.

2.2 Methods

2.2.1 Derivation of the digital expression profiles

ESTs previously described by Pavy et al. (2005b) and Ralph et al. (2006, 2008) were assembled into contiguous segments (contigs) using CAP3 software (Huang and Madan, 1999) with 95% percent identity for match over a 40 base region. The contig assemblies were queried against the plant protein data base (Viridiplantae) using the NCBI Basic Local Alignment Search Tool (BLASTX). The best hit with an e-value less than 1^{e-25} was selected for the analyses. The results were then parsed using an in-house PERL script to generate numerical tables.

The cDNA libraries (Table 2.1) were scored for each constituent EST of each contig, producing a two-way contig versus library table of raw EST counts. Contigs with counts of less than five across the libraries were excluded following Ewing et al. (1998). Finally using a secondary script, contigs were blasted against The Arabidopsis Information Resource (TAIR) version 8 (using BLASTX with same e-value), and the Gene Ontology (GO) functional categories determined by the annotation of the corresponding Arabidopsis homologue.

2.2.2 Principal components analysis

Principal Components Analysis (PCA) was used for visualizing the data structure and extracting variance from the data. The options for PCA were selected following Pittelkow and Wilson (2005), who suggested double-centering and the use of projectory biplot "GE-biplot" for the analysis of microarray gene expression data. In the GE-biplot, one can represent treatments (specimens) as numerical vectors, simultaneously with the number of

associated elements, genes, as points projected in a lower dimensional space. This is an effective method for visualization of the trends in a multidimensional data set. It further allows the use of scores (coordinates) for genes on the first two or three principal components to identify genes specifically associated with a tissue.

To perform a double-centered PCA, and to generate its biplot, the Two-Way Weighted Summation (TWWS) algorithm was used as implemented in the statistical package CANOCO version 4 (ter Braak and Smilauer, 2003). This algorithm was introduced by ter Braak (1987) and is suited for the analysis of ecological data which are organized as tables containing counts of the different species, similar in structure to digital expression data. Moreover, when numerous combinations of variables and observations need to be tested, TWWS is extremely fast and efficient for outlier detection, visualization and assessment of data structure. In using TWWS to analyze the digital profiles, the cDNA libraries were treated as species and contigs were treated as samples.

2.2.3 Two-way divisive cluster analysis

As a complementary approach, the final combination of the libraries and contigs obtained from PCA was submitted to DIANA. The Cluster package of the open source software R 2.7.2 was used to perform DIANA with Euclidean Distance as the measure of similarity.

2.2.4 Validation with wet-lab analogue data

Microarray expression profiles from two sources of tissue, xylem and phloem with the attached bark, were used as an independent resource to validate screening of the genes. The expression profiles represent three biological replicates of white spruce comprising the expression intensities of 18,881 spotted ESTs which are part of the comparative microarray data generated by the Treenomix Conifer Health Project obtained from Treenomix 2 data

base (http://treenomix2.forestry.ubc.ca). The description of the comparative microarray experimental design and data processing is available at Gene Expression Omnibus (GEO) under series accession number GSE18374 (http://www.ncbi.nlm.nih.gov/geo/).

The same options used for the digital data were presumed for the PCA of the analogue data. Six replicates of white spruce tissues (three from xylem and three from phloem/bark) were treated as species, and spotted microarray ESTs were treated as samples.

2.3 Results

2.3.1 Variation of digital profiles

Table 2.1 summarizes the libraries that were derived from xylem or phloem tissue. Spruce cDNA libraries of mixed vascular tissue (i.e. containing both xylem and phloem), and those derived from other tissue types, were excluded from this list of white spruce libraries. In addition, a single library (GQ 022) that had been constructed solely from root vascular tissue was excluded in order to avoid the additional complexity of above vs. below ground polarity. This resulted in forming a 3,554 contig x 11 library table on which the preliminary analyses was based.

When all 11 libraries were subjected to PCA, nearly 50% of the variance was explained by the first two ordination axes (Table 2.2, Solution 1). However, the separation of the xylem and phloem libraries in the ordination space was not optimal; phloem libraries GQ006, GQ028 and WS005 showed affinity to the xylem libraries (Supplementary Fig. S1; http://www.springerlink.com/content/d0817125681203lm/supplementals/).

The first PCA solution was then used to: (1) find a cluster within members of the same tissue that best represent shared expression of tissue-related genes; (2) maximize the separation between xylem and phloem tissues along the first principal component (PC1; which is the major axis gradient in the data set). Emphasis was placed on xylem libraries as

they had formed a tighter group than phloem. Performing PCA within the xylem libraries (Table 2.2, Solution 2) further confirmed uniformity of four (GQ004, GQ007, WS003, and WS008) of the seven xylem libraries (Supplementary Fig. S2; http://www.springerlink.com/content/d0817125681203lm/supplementals/), the members of which were selected to analyze against phloem libraries.

PCA was re-run several times with different combinations of the selected libraries, and outlier points were progressively removed from the contig list, until an optimal configuration of the libraries was obtained that could maximize the xylem-phloem polarity (Table 2.2; Solution 3). In this configuration, not only the three xylem libraries (GQ 007, WS 003, and WS 008) overlapped, but also their separation from phloem libraries resulted in PC1 explaining 68.6% of the total variance (Fig. 2.1).

As 100% of the variance was explained by the first 4 axes in Solution 3, the applicability of PCA to demonstrate the data structure in the lower dimensions was justified. The variance extracted by the second and higher axes was relatively low in this PCA solution, which suggested that the information contained in the higher axes is of less value for interpretation.

As shown in Fig 2.1, one of the libraries (GQ 006) had the most variance explained by PC 2 (as opposed to PC 1), and its projection on PC1 was close to the origin. Therefore, GQ 006 library was removed from the final analysis. This resulted in an increase in the variance explained by PC 1(Table 2.2; Solution 4), while the relative ranking of the contigs based on their scores (position of a contig along an ordination axis; eigenvector value of a contig) on this axis was largely unaffected (Supplementary Fig. S3; http://www.springerlink.com/content/d0817125681203lm/supplementals/).

2.3.2 Gene selection

To investigate tissue-related gene expression and gene selection, we focused on the xylem tissue, as represented by at least three libraries, with the WS 026 (which had the highest dispersion of the phloem libraries along PC1 in all PCA solutions) kept to retain polarity. The contig score was used here as a criterion to select contigs, with those on the negative end of PC1 regarded as highly associated with the xylem tissue. The top 10% with the highest PC1 scores was chosen, since selection of variably expressed genes is proven to be a useful step in exploratory analyses of gene expression data sets (Pittlekow and Wilson, 2003). Further comparison of the genes associated with xylem (128 contigs) using 10% criterion of PCA scores with DIANA output verified this criterion; the majority of the contigs (87) corresponded to three distinctive blocks on the heatmap which showed similar pattern of gene expression among all three xylem libraries (Fig. 2.2).

Functional assignment and representation of the selected genes

Of the 128 contigs selected, 33 (25%) had no similarity to any sequences (BLAST e-value greater than 1^{e-25}) in public data bases, and 38 (30%) comprised of ambiguous annotations including unknown, hypothetical, predicated, or putative proteins which were not confirmed by the combined BLAST searches (Supplementary data; http://www.springerlink.com/content/d0817125681203lm/supplementals/).

The total list of annotated contigs is compiled in Supplementary data. Overall, the annotations of the contigs were consistent with their assigned GO function and process categories. Genes associated with hydrolase activity, protein metabolism, and response to stress found to be the most abundant in the analysis of function and process categories (Fig. 2.3 a, b). The other abundant groups of genes were involved in transport, protein binding, DNA or RNA binding, developmental processes, signal transduction, transferase activity, and

transporter activity respectively (vague categories such as other enzyme activity, other binding, or other cellular processes are excluded).

Based on the analysis of GO component category, it was not possible to identify xylem, or even characterize vascular tissue. For instance, the occurrence of chloroplast, a cellular component associated with green tissue, or the lack of cell wall components for a woody tissue, was unexpected (Fig. 2.3, c). As well, the higher abundance of mitochondria, cytosol, or membrane components could be indicative of any active differentiating cell.

Assigning the function of the groups listed in Supplementary data, the meta-clusters (blocks) of DIANA did not resolve into uniform groups of related functional members, or biologically meaningful cliques.

Validation by the wet-lab analogue approach

PCA of the analogue data included a single attempt with no further optimization. The total variance in the data was extracted by four axes (Table 2.2, Solution 5); 75.1 % of which by the first axis alone, rendering an optimal separation of xylem from phloem/bark tissue replicates along this major gradient (Supplementary Fig. S4; http://www.springerlink.com/content/d0817125681203lm/supplementals/).

The majority (117) of the total 128 xylem-related contigs had at least one match in BLAST against xylem-related microarray ESTs with top 10% scores on PC1(n=1861). However, since the stringent e-value cut off of 1^{e-25} was adopted for this search, only one third of the contigs (n=44) were validated by the analogue data (Supplementary data; http://www.springerlink.com/content/d0817125681203lm/11295_2010_Article_275_ESM.html).

Validation by the analogue data highlighted the elements of phenolics biosynthesis laccase (for which 7 EST hits was found in the xylem microarray expression), and cinnamyl alcohol dehydrogenase enzymes which are instrumental in the formation of lignin. Together with these, the occurrence of the other fundamental enzyme of cell wall biosynthesis, cellulose synthase, underpin the characteristics of the wood forming tissue, xylem.

The overall pattern of the GO function and process categories as reflected by the screened genes were similar to the genes of the Supplementary data (http://www.springerlink.com/content/d0817125681203lm/11295_2010_Article_275_ESM.html) with response to stress being most prevalent (Fig. 2.3). In light of this, the presence of several contigs involved in response to stress including dehydrin , beta tubulin, and alcohol dehydrogenase (with respectively 6 , 4 and 2 microarray EST hits) is also emphasized (Table 2.3).

2.4 Discussion

2.4.1 Avoiding potential biases of digital analysis of EST libraries

To establish the transcriptome signatures specific to xylem or phloem, ideally, numerous vascular profiles (libraries) should be analyzed. However, the numbers of ESTs can often differ dramatically between libraries due to quality of cloning and the investment into sequencing. This can introduce biases into the multivariate analysis of gene expression. Ewing et al. (1999), in their analysis of rice ESTs, suggested that libraries should not differ by about 5 times for the numbers of ESTs collected. Table 2.1 indicates that in white spruce, we have a rich and diverse number of ESTs related to vascular tissue. However, the set of libraries used varied in number by about 14 times between the smallest and the largest library.

Another inherent problem that makes different EST collections non-uniform is that different cDNA libraries may capture different sets of transcripts from the same source of tissue, due to differences in their construction methods, biological factors including genotype, age and environmental factors such as season. Capturing different sets of ESTs during each library construction attempt can explain distant grouping of the phloem libraries in the present study.

Over-representation of housekeeping genes common to both sources of vascular tissue during the construction of the libraries, which might have caused grouping of several phloem and xylem libraries, may not mean that the identified genes are "xylem-specific". However, clustering and positioning of the selected xylem libraries (GQ007, WS003, WS008) in the ordination diagrams in two consecutive rounds of PCA (Solution 1 and 2) is an indication of sharing "xylem-related" expression profile among them. Since these libraries were constructed from the pooled xylem tissue, it is likely that the analyzed expression profile represent that of the xylem housekeeping genes. As well, using the pooled tissue could provide a means of buffering environmental and biological factors that could have strongly impacted analysis of the libraries from uniform tissues.

Detailed investigation into the complex relationships among transcripts, and their associations with the source of tissue, was partly hampered in this research due to the occurrence of about 30 % of genes with ambiguous annotations, and 25 % of genes with no sequence similarity in the data bases. While the validation of the results obtained here beyond the context of exploratory studies rests upon the improvement of plant functional data bases (which in turn, might change these proportions and reveal new insights into the relationships among genes and their association with their source of tissue), the latter finding

is of practical significance. An alternative explanation again indicates unique expression of these genes in the xylem tissue of conifer which fits the objective of tissue-related marker discovery, which in turn, reflects a proper method of gene selection.

The bioinformatics criteria adopted to ensure the objective of this research, screening tissue-related markers, were: first) the stringent cut-off of 1^{e-25}; second) the combinatory BLAST against TAIR and all plant data bases of NCBI in which the acceptance of an annotation was conditioned upon consistency in both searches. Relaxing these criteria would result in a higher proportion of the annotated genes. In particular, the addition of "no hit" and "ambiguous" contigs with significant BLAST hit on the xylem analogue profiles will increase the total number of the xylem related markers. However, the use of such conservative strategy not only highlighted the proportion of "no hit" as a potential conifer-specific group, but also avoided groupings of various genes under generalized groups. For instance, separation of cinnamyl-alcohol dehydrogenase from other group of more generally annotated alcohol dehydrogenase points to the divergence of this group from the latter in the conifers.

2.4.2 Relationship to earlier findings

Associations of a group of genes (as listed in Table 2.3) with the xylem tissue were consistent with literature of large-scale digital and analogue gene expression in trees, and particularly conifers; these include preferential expression of genes involved in lignin and cell wall biosynthesis such as cinnamyl alcohol dehydrogenase (Friedmann et al. 2007) and laccase (Sterky et al. 1998; Friedmann et al. 2007) and cellulose synthase (Friedmann et al. 2007; Pavy et al. 2005b), or universal stress proteins (Friedmann et al. 2007). These can serve a starting point for the targeted studies to determine their specific types (tissue-specific markers) in xylem (for instance done for cinnamyl alcohol in the floral stem of Arabidopsis

by Sibout et al. 2005) and differences of their expression patterns and sequences across various species and other sources of tissue.

Although, abundance of genes associated with protein metabolism including transporters and protein kinases as reported (Pavy et al. 2005b) can be a characteristic of any active developing tissue, over-representation of the genes involved in stress and defense response in this study warrants attention; especially since the analyzed libraries were constructed from normal tissue, presumably in the absence of biotic or abiotic stress (also absent in the analysis of the GO categories). With respect to the involvement of vascular tissue in stress and defense response (e.g. Martin et al. 2002; Miller et al. 2005), this feature can indicate tissue-specificity.

In addition, several previously recorded genes encoding zinc-fingers, late embryogenic abundant proteins (Friedmann et al. 2007; Foucart et al. 2006; Pavy et al. 2005b), ubiquitin-family proteins (Pavy et al. 2005b; Foucart et al. 2006) were selected in the analysis of the digital data (Supplementary data; http://www.springerlink.com/content/d0817125681203lm/11295_2010_Article_275_ESM.html). However, the final screening of these genes was not verified by analogue data which can be explained by the closed design of the microarray platforms, that might under-represent certain groups of genes and can be considered as another possible bias in the analyses.

2.4.3 Alternatives to digital analysis of ESTs

Pioneered by Adams et al. (1991) using conventional Expressed Sequence Tags (ESTs), digital analysis of gene expression could also be done by Serial Analysis of Gene Expression (SAGE) technique (Velculescu et al. 1995), Long SAGE (Saha et al. 2002), and Massively Parallel Signature Sequencing (MPSS) (Brenner et al. 2000). A key advantage of MPSS and SAGE techniques is efficiency in sampling large numbers of transcripts over

conventional EST or cDNA sequencing. Compared to one EST or a fraction of cDNA in the conventional method, between 12 and 20 transcripts are sampled per sequencing reaction in SAGE, and potentially all the tags in a library can be sequenced simultaneously in MPSS (Hene et al. 2007). Nevertheless, in publicly available databases, currently there are over 61 million ESTs archived (http://www.ncbi.nlm.nih.gov/sites/gquery, as of April 17, 2009) which can be utilized for digital expression analysis (congenitally known as digital Northern analysis).

The new generation of sequencing technology such as LCM-454, or Solexa's SBS (Sequence by Synthesis), are in principle capable to adapt to tag-based technology that will allow production of extremely large EST or SAGE libraries (Hene et al. 2007). Meanwhile, the current collection of vascular tissue is the only resource available for conducting a large-scale digital analysis in spruce. In spite of this, application of PCA as an exploratory method proved to be useful; most notably in finding a proper combination of the libraries within the existing collection. Using this combination, the differences between the two sources of tissue were illustrated with a reasonable level of assurance from variation in the total data structure, which further allowed elucidating the elements creating the greatest influence in tissue-related gene expression.

2.5 Concluding remarks

In the rapidly evolving genomics era, emergence of ultra high-throughput sequencing methods would allow rapid production of extremely large EST or SAGE libraries, and may overshadow conventional analysis of digital gene expression. While such deep transcriptomic data sets are yet to be generated and made available in the public databases, it was demonstrated here that with the application of suitable statistical methods, transcriptome signatures could be identified cost-effectively within an existing data.

The illustrative use of GO function and process categories (shown here as bar charts) could be a potentially useful approach for illustrating generalized tissue-related signatures; and can be used for comparisons with other studies using the similar tissue types (e.g. analogue data), and for comparisons of other sources of tissue, in approximation towards tissue-specific signatures of gene expression. However, until the future growth and improvement of the plant genomic data bases providing cellular information in a finer resolution, the use of this category should be cautioned.

While the validation of the results of this paper rests, in part, upon the improvement of plant genomic data bases, the presented research can be used in establishing priorities among genes for targeted functional studies. The subset of contigs validated with the analogue data could be introduced as xylem-related markers tissue with a reasonable level of assurance. Further analysis of the microarray data can provide insights into the lack of verification of some contigs, or the possibility of compilation of a more comprehensive list of tissue-related markers. However, detailed investigation of the microarray data was not within the scope of the present paper and can be treated as an independent report.

Table 2. 1 Description of White spruce (*Picea glauca*) and Interior spruce (*P. glauca* x *engelmanii* CV. PG29) cDNA libraries[1] that represent phloem and xylem sources of tissue.

cDNA library identifier	Tree species	Number of ESTs	Library description
GQ 004	White Spruce	2397	Non-lignified differentiating xylem from normal vertical trees
GQ 006	White Spruce	6136	Cambium and phloem region from normal vertical trees
GQ 007	White Spruce	999	Secondary xylem tissues of trees girdled by removing a ring of bark (tissue pooled from above and below the girdle)
GQ 028	White Spruce	15977	Cambium and phloem region from normal vertical trees
GQ 029	White Spruce	3736	Xylem scrapings from normal vertical trees
GQ 030	White Spruce	5712	Xylem scrapings from normal vertical trees
GQ 031	White Spruce	14501	Xylem scrapings 2 m above ground level
WS 003	Interior Spruce	2591	Xylem from early, mid and late developmental stages (pooled tissue)
WS 005	Interior Spruce	2827	Phloem from early, mid and late developmental stages(pooled tissue)
WS 008	Interior Spruce	14149	Xylem from early, mid and late developmental stages(pooled tissue)
WS 026	Interior Spruce	14343	Phloem from early, mid and late developmental stages(pooled tissue)

[1] Part of Pavy et al. 2005b, Ralph et al. 2006, and Ralph et al. 2008 publications, data based at Canada's Michael Smith Genome Sciences Centre at http://www.bcgsc.ca/).

Table 2. 2 Summarizing PCA attempts in respect to variance explained by four axes (PCs).

Solution	Number of libraries	Description	Cumulative % variance extracted			
			Axis 1	Axis 2	Axis 3	Axis 4
1	11	All libraries in Table 1	29.9	46.3	60.9	73.7
2	7	GQ004, GQ007, GQ029, GQ030, GQ031, WS003, WS008	39.6	71.6	84.9	93.0
3	5	GQ 006, GQ 007, WS 003, WS 008, WS 026	68.6	82.4	95.3	100
4	4	GQ 007, WS 003, WS 008, WS 026	79.6	94.6	100	-
5	N/A	Analogue data	75.1	90.1	99.7	99.9

Table 2. 3 List of the contigs screened as markers of the xylem tissue validated by the analogue data.

Annotation		GO category (Component)	GO category (Function)	GO category (Process)	Significant EST match (analogue)
Contig4454	Cinnamyl alcohol dehydrogenase	- Nucleus - Plasma membrane - Chloroplast	- Other enzyme activity - Other binding	- Response to stress	WS0261_G15 (6.3^{e-194})
Contig9576	Disease resistance gene	-	-	-	WS0048_M04 (2.6^{e-162}) WS0261_M07 (1.1^{e-28})
Contig2700	Leucine-rich repeat trans-membrane protein kinase	-	- Kinase activity - Nucleotide binding	- Protein metabolism - Signal transduction	WS0034_E15 (7.3^{e-152}) WS0079_A22 (1.5^{e-32}) WS0265_A16 (6.8^{e-31})
Contig213	Cellulose synthase	-	- Transferase activity	- Cell organization and biogenesis - Response to stress	WS00915_P16 (4.9^{e-26})

Table 2. 3 continued

	Annotation	GO category (Component)	GO category (Function)	GO category (Process)	Significant EST match (analogue)
Contig2	40S ribosomal protein S19 (RPS19A)	- Ribosome - Cytosol	- Structural molecule activity	- Protein metabolism	WS00910_P02 (4.1 $^{e-138}$)
Contig3201	Polygalacturonase	-	- Hydrolase activity	- Other metabolic processes	WS00924_M15 (7.8 $^{e-147}$)
Contig1806	Signal peptide, peptidase family protein	- Endoplasmic reticulum	-	-	WS0054_I20 (1.1 $^{e-154}$)
Contig6142	Cytochrome c oxidase	- Mitochondria	- Transporter activity	- Electron transport or energy pathways	WS0082_M03 (2 $^{e-135}$)
Contig1598	Auxin-induced protein/transcription factor	- Nucleus	- Transcription factor activity - Protein binding	- Developmental processes - Other biological processes	WS0261_M19 (3.6 $^{e-57}$)
Contig1862	Eearly nodulin	-	-	-	WS0078_F09 (6.2 $^{e-84}$)

Table 2. 3 continued

	Annotation	GO category (Component)	GO category (Function)	GO category (Process)	Significant EST match (analogue)
Contig2344	Oligopeptide transporter (OLP)	- Other membranes	- Transporter activity	- Transport	WS0039_K04 (2.1 $^{e-121}$) WS0093_F06 (2.9 $^{e-87}$)
Contig9364	14-3-3 protein, general regulatory protein	-	- Protein binding	- Signal transduction	WS0264_K11 (7.5 $^{e-163}$) WS00931_P23 (2.6 $^{e-46}$)
Contig8855	Response to abscisic acid (RAB), binding	- Other cellular components - Other intracellular components	- Nucleotide binding	- Transport	WS00924_H11 (5.8 $^{e-42}$)
Contig4685	Thioredoxin	-	-	-	WS0261_A20 (2.7 $^{e-153}$)
Contig12077	Aldolase (plastidic aldolase, fructose-bisphosphate aldolase)	- Mitochondria - Chloroplast - Plastid	- Other enzyme activity	- Response to stress	WS00925_A22 (3.7 $^{e-184}$)
Contig3668	Auxin-associated family protein	-	-	-	WS0012_N05 (1.3 $^{e-147}$)

Table 2. 3 continued

	Annotation	GO category (Component)	GO category (Function)	GO category (Process)	Significant EST match (analogue)
Contig185	Universal stress protein (USP) family protein	-	-	- Response to stress	WS01018_J20 (2.2^{e-118})
Contig13116	Laccase	- Other cellular components	- Other binding - Other enzyme activity	-	WS00813_M20 (1.7^{e-132}) WS0038_I15 (1.6^{e-46}) WS0086_J17 (1.6^{e-35}) WS00815_F23 (4.6^{e-34}) WS0038_B22 (1.8^{e-29}) WS0039_C04 (2.2^{e-28}) WS0056_P10 (1.6^{e-28})
Contig1782	Acid phosphatase	- Other cellular components	- Hydrolase activity	-	IS0014_D21 (9.3^{e-52})
Contig10633	Beta tubulin	- Other intracellular components	- Structural molecule activity	- Response to stress	WS0041_L16 (3.9^{e-153}) WS01018_N22 (1.8^{e-100}) WS0043_L20 (5.6^{e-49}) WS00913_J01 (3.6^{e-46})

Table 2. 3 continued

Contig ID	Annotation	GO category (Component)	GO category (Function)	GO category (Process)	Significant EST match (analogue)
Contig5946	Ribose 5-phosphate isomerase	- Plastid - Chloroplast	- Other enzyme activity	- Response to stress - Other cellular processes	WS0261_H01 (2.5^{e-153})
Contig12368	Alcohol dehydrogenase	- Other cellular components	-	- Response to stress - Developmental processes	WS00815_A15 (7.1^{e-150}) WS00721_A21 (2.9^{e-35})
Contig9056	Aspartic proteinase	- Other cytoplasmic components	-	- Protein metabolism	WS0042_E21 (3.2^{e-33})
Contig2864	Secretory carrier membrane protein (SCAMP) protein	- Other membranes	- Transporter activity	-	WS00924_H01 (1.4^{e-190})
Contig2214	Selenoprotein family protein	- Other cellular componens	- Other binding	-	WS0055_O18 (4.8^{e-145})

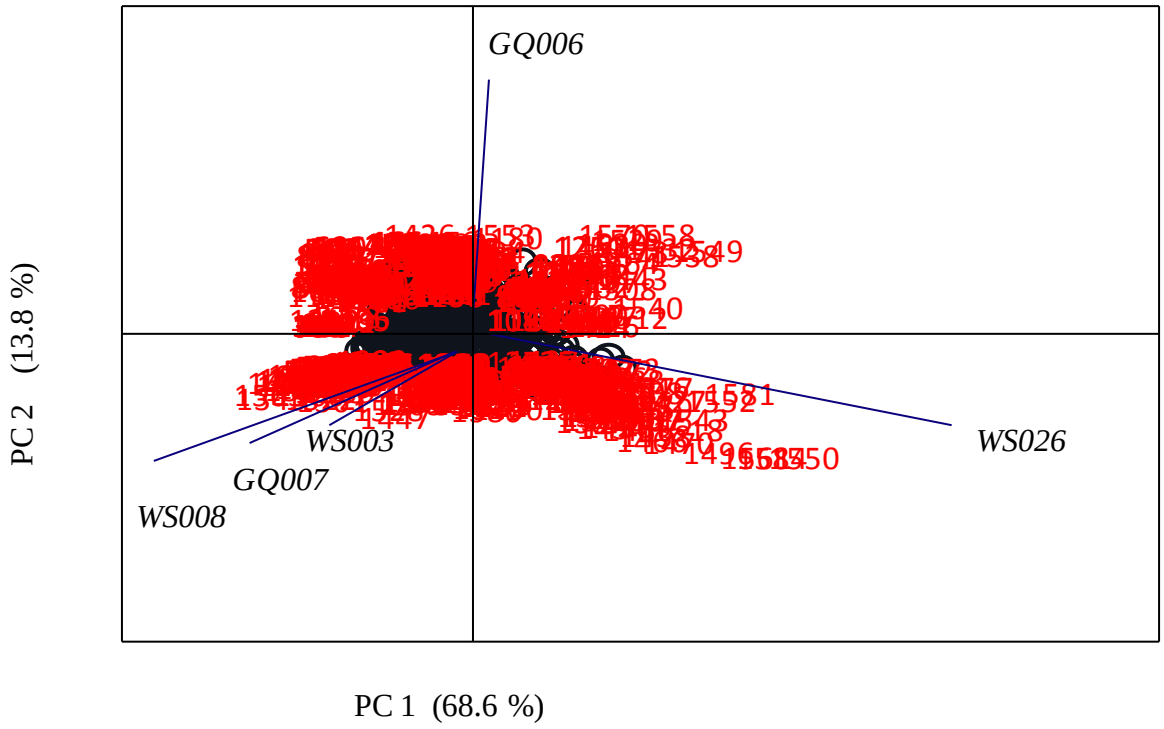


Figure 2. 1 Biplot projection of 5 libraries demonstrating simultaneous ordination of the contigs and libraries in a two dimensional space.

Xylem libraries (GQ007, WS003, and WS008) and WS026 phloem library are situated on the opposite sides of the first principal component (PC 1) which explains 68.6 % of the variance in the total data set. Each library is represented by a blue ray. Contigs (seen with red labels) are points spreading out from the origin, and are positively correlated if on the same side of the origin and negative if on the opposite sides.

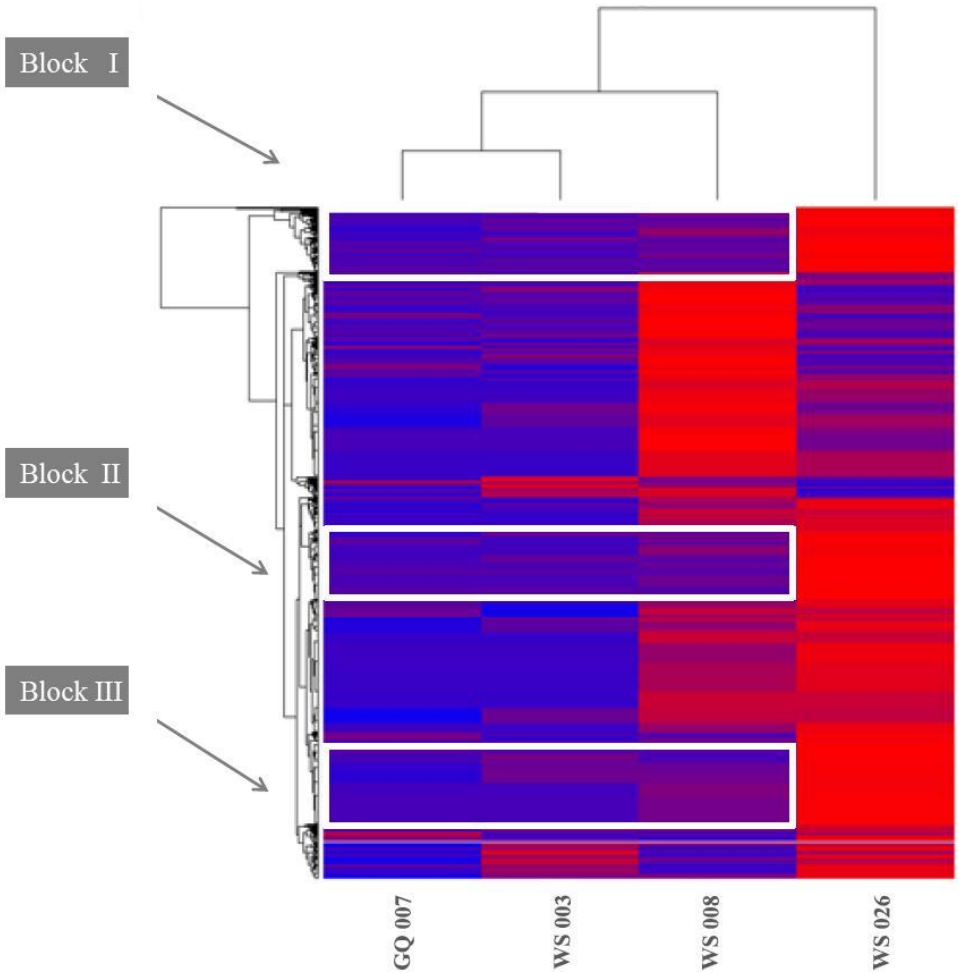
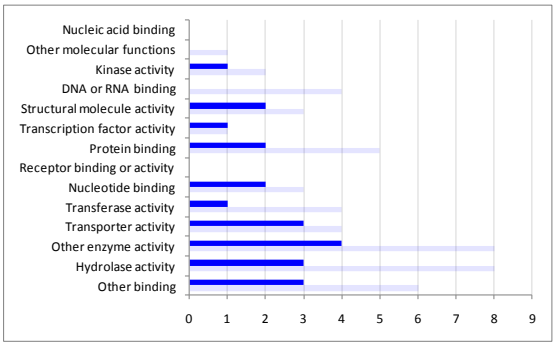


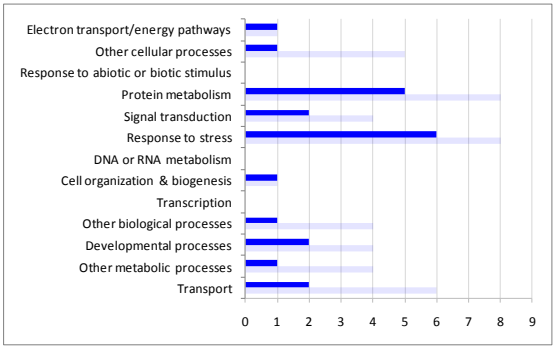
Figure 2. 2 Dendrogram and heatmap of DIANA showing simultaneous classification of the contigs and four cDNA libraries**.**

The blocks represent similar pattern of gene expression (in the range of dark blue color) among three xylem libraries (GQ 007, WS 003, and WS 008) versus WS 026 phloem library.

a)



b)



c)

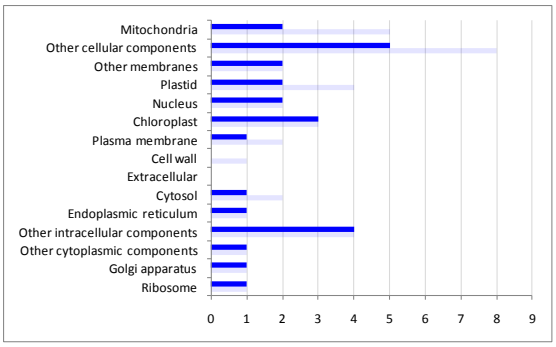


Figure 2. 3 Bar charts representing distribution of the contigs from Supplementary data (light colour bars) and verified genes (dark-colour bars) among gene ontology function (a), process (b), and component (c) categories. The horizontal axes represent the counts of the contigs.

3. Evolution of the expression of phenolic gene families in the outer stem of spruce (genus Picea)

3.1 Introduction

Diversification and specialization of chemical compounds can be driven by the evolution of plant defense against a diverse array of pests (Benderoth et al. 2006; O'Reilly-Wapstra et al. 2004). Phenolics are a broad class of compounds produced as metabolites of the phenylpropanoid and associated pathways, and include hydroxycinnamates, flavonoids, tannins, anthocynins and lignins (Fig. 3.1) These compounds are involved in constitutive and induced defenses, and provide plants with physical barriers as well as repelling protection against invasive organisms including pests and their associated pathogens (Franceschi et al. 2005; Bernards and Bastrup-Spohr, 2008; Lev-Yadun and Gould, 2008). Researchers have detected the signature of natural selection in the nucleotide sequences of several genes of phenylpropanoid and flavonoid pathways (Ramos-Onsins et al. 2008; Kuelheim et al. 2009). Diversification and expansion of several phenylpropanoid-related gene families has also been documented (Hamberger et al. 2007).

In addition to sequence evolution, changes of gene expression have long been recognized as a fundamental component of adaptive evolution (Britten and Davidson, 1969; King and Wilson, 1975). The evolution of gene expression has been investigated across species (e.g. human-mice-rat by Jordan et al. 2005; various *Drosophila* species by Lemos et al. 2005). Among populations of the same species, alteration of gene expression in response to environmental factors has been documented (e.g. *Fundulus heteroclitus* by Whitehead and Crawford, 2006; *Picea sitchensis* by Holliday et al. 2008). The divergence of gene expression among spruce species has not yet been studied, and the evolution of gene expression for genes in phenolic pathways has not been studied in any plant species. Detecting the

signature of natural selection on gene expression allows assessment of the relative contribution of gene expression vs. structural changes in evolution (Nuzhdin et al. 2004). This can also facilitate the functional assignment of genes in species that lack a reference genome.

The evolution and expression of phenolic genes among coniferous species is, to a large extent, unknown. Functional characterization of phenolic genes is limited to the genes upstream of the pathways (e.g. 4CL, -4-coumarate-coenzyme A ligase, Wagner et al. 2009). For the majority of genes, homology based comparisons with angiosperm reference pathways are used for classification. For the phenylpropanoid and associated pathways, reference pathway maps are available for angiosperms (e.g. www.kegg.com/kegg/pathway/map/map00940.html). However, homology based annotation is often hampered by the expansion of genes as families in gymnosperms, and subsequent paralogy (i.e. duplication, neo- and -subfunctionalization events) within these expanded gene families (Ralph et al. 2008). As well, large evolutionary distance (385 million years since a common ancestor, Zimmer et al. 2007), makes the linkage between conifer function and angiosperm function at best tenuous, and places importance on conifer specific studies.

In this study, we examine patterns of gene expression divergence for phenolics among five species of spruce, spanning much of the phylogeny of this genus. Our objectives are (1) to examine the relevance of the source of tissue in the diversification of phenolic gene families, (2) to infer the modes of evolution (diversifying selection, balancing selection, stabilizing selection, neutrality) underlying the expression of the various phenolic genes and their corresponding families, and (3) to further understand the structure of the coniferous

phenolic biosynthesis pathway through such investigation of the divergence of gene expression.

3.2 Methods

3.2.1 Sampling

The arboretum of British Columbia Forest Service's Kalamalka Research Station, Vernon, BC, Canada maintains 23 species of spruce at a similar environment since planted in 1977. The phylogenetic tree of the microsatellite genetic distances of the 23 species (Rungis et al. 2007) served as a guide to sample five mutually unrelated spruce species; black spruce (*Picea mariana*), Jezo spruce (*P. jezoensis*), Norway spruce (*P. abies*), Serbian spruce (*P. omorika*), and white spruce (*P. glauca*). Three biological replicates per species (confirmed by Amplified Fragment Length Polymorphism markers, see below) gave a total of 15 samples. The samples of bark with the attached phloem (pooled from various aerial branches) were separated from the inner layers (xylem) in the field during mid-afternoon hours, flash-frozen in liquid nitrogen, and transferred into separate containers on April 24, 2007.

3.2.2 Screening and classification of phenolic genes

Sequences of the third generation Treenomix microarray cDNA platform were used for an exhaustive screen specific for phenolic biosynthesis gene families. This platform, submitted under accession GPL5423 to Gene Expression Omnibus (GEO), comprises 18,881 spotted EST elements from 12 cDNA libraries of white spruce, Interior spruce (*P. glauca* x *engelmannii*) and Sitka spruce (*P. sitchensis*). Contiguous sequences (contigs) were formed from the array cDNA sequences and the sequences of an in-house database of spruce ESTs (Spruce V8; http://treenomix2.forestry.ubc.ca/). The contigs were used for a translated

nucleotide query (BLASTX) against protein database of the Arabidopsis Information Resource (TAIR version 8), and a translated nucleotide query (TBLASTX) against a curated data base of translated nucleotides of all plants (Viridiplantae) at National Center for Biotechnology Information (NCBI, downloaded March 9, 2009) in which unknown and hypothetical proteins had been removed. Because of the large evolutionary distance between the gymnosperms and angiosperms, an E-value cutoff was not adopted as a search criterion for the annotation. Genes were accordingly categorized into the corresponding families based on annotation identity.

BLAST searches effectively assigned the genes to families. However, the disagreement of annotation results between the narrow search of TAIR and the broad search of Viridiplantae suggested the possibility of functional divergence of several categories. Hence, the designation "putative" was decided for pCCoAOMT, to incorporate changes of the annotation from caffeoyl CoA O-methyl transferase in angiosperms to catechol O-methyltransferase in gymnosperms for several genes; pDFR in which the annotations shifted from dihydro flavonol reductase and BANYULS in angiosperms to anthocyanidin reductase in gymnosperms; and pPCBER encompassing phenylcoumaran benzylic ether reductase, pinoresinol-laricinol reductase, isoflavone reducatase, and leucoanthocyanidin reductase, comprising members with relatively high sequence similarity, but divergent functions. Similarly, the annotations of cytochromes p450 (CYP) categories CYP84 (ferulate 5-hydroxylase, F5H), CYP71, CYP 736 (angiosperms), and CYP750 (gymnosperms) were exchanged. Therefore, these categories were also regarded as putative, and termed "F5H-like" after their distant relatives with known function only in angiosperm phenolic biosynthesis.

The majority of annotations for the best hits of 2-Oxoglutarate-Ferrous-dependent Oxygenases (2OGFeII) and glycosyltransferases (GLYTR) were obtained only at the level of superfamily, which rendered consistent separation of flavonol synthase, anothcyanidin synthase (2OGFeII), or anthocyanidin 3-O-glucosyltransferase (GLYTR). The separation of other categories with similar hierarchical relationship, O-methyl transferases (OMT, superfamily) and putative caffeoyl-CoA-O-methyl transferases (pCCoAOMT, family), showed high consistency, which suggested divergence of a discreet functional group. Members of pCCoAOMT were distinct from OMT which encompassed other putative O-methyltransferases, beta-alanine n-methyltransferase, and caffeic acid O-methyltrasnferases (COMT).

Because of the paralogy of genes in conifers, and the challenges in annotation of individual genes we focused upon gene families. On the basis of the above, 18 categories (Appendix A1) were recognized, hereafter termed "gene families".

3.2.3 Microarray profiling and analysis

A list of 332 phenolic cDNA sequences was compiled for expression profiling (Appendix A1). Gene expression profiling was performed using 15 cross-species hybridizations in a loop design (Fig. 3.2a). The array 350 kit (Genisphere, Hatfield, PA, USA) was used for microarray hybridizations. The microarrays were scanned with a ProScanArray scanner (Perkin Elmer, Downers Grove, IL, USA), and the scanned TIF images were processed by ImaGene ver. 6.0.1 (BioDiscovery Inc. El Segundo, CA, USA) to quantify spot intensities. The microarray data was submitted to GEO under the series accession GSE18374 in compliance with Minimum Information About a Microarray Experiment (MIAME). The spot signals were normalized within, as well as between arrays

using the variance stabilizing (VSN) method (Huber et al. 2002) with the Cy3 and Cy5 expression intensities being averaged to arrive at a matrix of gene x individual expression intensities.

A linear mixed model of the form $y_{gki}=\mu+\alpha_g+\lambda_k+\gamma_{ki}+\varepsilon_{gki}$ was used to test the significance of variation among species (i.e. divergence) of the overall phenolics expression in bark/phloem. In this model, y_{gki} is the normalized intensities of gene expression for genes (1,…,332), μ is the intercept, α_g is a fixed effect term that indexes g=(1,…, 18) gene family categories, λ_k is a fixed effect term that indexes k=(1,…, 5) species, γ_{ki} is the random effect of individual i of the species k, (biological effect) and ε_{gki} is the error. Restricted maximum likelihood (REML) was used for estimation. This model was also used to analyze microarray data of Albouyeh et al. (2010, series accession GSE18374) from xylem and needle cross-species comparisons, profiled according to the same experimental design.

To examine variation within each gene family category in bark/phloem, the normalized signal intensities were again analyzed by fitting mixed effect ANOVAs of the same form but with α_g here being a fixed effect term that indexes number of genes per family category. Q values (Storey and Tibshirani, 2003) were calculated to adjust for multiple testing over 18 categories.

3.2.4 Analysis of the neutral divergence of five spruce species

Amplified Fragment Length Polymorphism (AFLP) markers (Vos et al. 1995) were assayed to estimate neutral genetic distances between species. A total of 6 *Eco*RI and *Mse*I primer combinations were selected for scoring following the method of Goodwillie et al. (2006). The binary data of 388 AFLP loci (200 polymorphic) was used to calculate the pair-wise genetic distances in accordance with Nei and Li (1979) model. Restdist program of

PHYLIP ver. 3.66 (Felsenstein, 2006) was used to compute a distance matrix from DNA fragments. A neighbor-joining phylogenetic tree was constructed using Consense and Neighbor packages of PHYLIP 3.66. TreeView ver. 1.6.6. (Roderic, 2001) was used to view the tree (Fig. 3.2b).

3.2.5 Analysis of the neutral divergence of family categories

The spot signal intensity values for the genes of divergent families on array were converted to the pair-wise differences of expression on the individual (biological replicate) basis. Spearman rank correlations were calculated to quantify the strength of the association between expression difference and AFLP (neutral) genetic distance over 105 pair-wise comparisons for each gene, and averaged over gene members for each family category (Appendix A2).

3.2.6 Analysis of the divergence of expression versus DNA sequences

For each gene family, the corresponding sequences of array spots were aligned with CLUSTALW application of BioEdit ver. 7.0.0. (Hall, 1999). Among species of spruce, evolution of genomic loci is denoted by a relatively slow mutation rate (Wang et al. 2000; Bouille and Bousquoute, 2005). Therefore, divergence among the sequences of the array from white spruce, Sitka spruce, and Interior spruce was assumed to reflect the divergence of the sequences among the five species studied. The DNADist application of BioEdit was used to construct a distance matrix for the aligned sequences of each family. The average DNA distance per family was used to plot against F values of species expression. Ordinary least squares (OLS) was used to assess the strength of the relationship between *Z*-scores of average DNA distance and *F* value of species expression. The significance of the relationship between the divergence of the expression of gene families (*F* value species) and the

divergence of their corresponding sequences (average DNA distance) among species was computed using Wilcoxon signed rank test.

3.3 Results and discussion

3.3.1 Expression of phenolic biosynthesis genes in the bark and phloem compared with xylem and needles

The results of the statistical tests for the overall variation of the phenolics gene expression are presented in Table 3.1. The magnitude of variation of gene expression among species, as well as the driving factors for this variation, was compared in different tissues. Defining families was a significant factor in the variation of gene expression, for all three tissue sources. However, only three families in needle, and no families in xylem were found to have significantly high variation among species at the gene family level (after adjustment of multiple testing; Appendix A2).

In the bark and attached phloem, the variation of gene expression among species was significantly high for eight families of genes (after adjustment of multiple testing; Table 3.2). Coupled with this, high overall variation of gene expression (based on the F value of mixed ANOVA test; Table 1) in bark and the attached phloem was observed. Among the tissue sources, bark and the attached phloem was the only tissue for which among species divergence was a significant ($P < 0.05$) factor in the variation of gene expression.

The finding of the statistical analysis is in line with the well established role of coniferous bark as a defense barrier against the diverse array of pests aiming for the nutrient rich phloem (reviewed by Franceschi et al. 2005). This places emphasis on the possibility of defense-driven variation in the expression of the phenolic genes in the outer stem of conifers. In addition, genome-wide perspective of the function of phenolic pathways in plant defense

suggests a model of polygenic response in the induction of phenolic gene families (Dixon et al. 2002). This agrees with finding higher number of gene families of different functions having diverged expression in the outer stem, compared to the other tissues (Table 3.2).

If such diversification is the result of the co-evolution with pests, natural selection is the explanation for the variation in the expression of genes to increase the survival of the trees. Our inferences about the modes of selection underlying the divergence of the expression of gene families in the outer stem are based on statistical analysis (ANOVA). Applying this framework to the analysis of gene expression (e.g. Whitehead and Crawford, 2006), three modes of selection is possible to infer: high variation in gene expression among vs. within species indicates diversifying selection whereas the opposite indicates balancing selection , and low within and among species variation in gene expression is an indication of stabilizing selection. Using this framework, it is also possible to use *F* statistic as a measure of expression divergence (e.g. Nuzhdin et al. 2004) to assess the extent to which divergence of the expression of genes is related to divergence of their corresponding sequences.

3.3.2 Modes of the evolution of phenolic gene families

The majority of gene families in the outer stem (11 out of 18), did not show significant variation in expression among species (Table 3.2). Gilad et al. (2006), using a similar statistical framework (linear mixed model), suggest that a set of genes for which expression levels have remained constant across a phylogeny is likely under stabilizing selection. We delineated a space that characterized these gene families as well by having both low between and within species variation (Fig. 3.3).The predominance of this mode of selection over the majority of families was not unexpected. Previous studies of the evolution of gene expression in other biological systems report similar findings (e.g. Gilad et al. 2006;

Rifkin et al. 2003). Within the context of plant defense, it is the steady-state expression among species for these gene families that is likely of adaptive importance.

Significant ($Q < 0.05$) variation among species (divergence from the space of stabilizing selection) distinguished eight families, phenylalanine ammonia lyases (PAL), cinnamate 4- hydroxylases (C4H), O-methyl transferases (OMT) putative caffeoyl CoA O-methyl transferases (pCCoAOMT), dirigent-like proteins (DIR) , laccases (LAC), and putative phenylcoumaran benzylic ether reductases (pPCBER), and glycosyl transferases (GLYTR).

The analysis of the expression of the above gene families against neutral divergence did not find substantial associations between the divergence of gene expression and neutral evolution at the family level (Table 3.3). The expression of few individual genes, especially in DIR and GLYTR families, did show strong positive correlations with neutral genetic distance (0.5 or higher; Appendix A3). However, considering the pattern of expression at the family level cancels strong associations with neutrality (Table 3.3). Therefore, at the level of gene families, diversifying (positive Darwinian) must be the driving force underlying the divergence of expression among species. Integration of the divergence of these gene families in the bark of spruce could be an indication for an inherent polygenic diversification response against the co-evolving pests aiming to pass this barrier.

The advantage of the statistical framework used here as opposed to neutrality models (e.g. Rifkin et al. 2003; Khaitovich et al. 2004) is the possibility of inferences about a third mode, balancing selection, in addition to diversifying and stabilizing selection. Our interest to infer this mode of selection, in particular, was to explain the divergence of the expression of gene families upstream of the pathways. Under this mode of selection, a copy of the

upstream gene can expected to be kept conserved while the other copies of the gene diverge in function.

Conforming to the pattern of high within versus low among species variation in expression (Fig. 3.3), balancing selection has likely impacted two gene families, chalcone synthases (CHS), and coumarate 3-hydroxylases (C3H), immediately following the core phenylpropanoid pathway and at the entry points of the flavonoid and general phenylpropanoid pathway, respectively. However, balancing selection was also expected to derive the evolution of PAL and C4H, two of the gene families defining, with 4CL, the core of the phenylpropanoid pathway. These enzymes are positioned at the entry point of the phenolics biosynthesis pathways where divergence could affect the functionality of the general phenylpropanoid and the associated downstream pathways. The lack of species by genes interaction (Table 3.3, also seen for pCCoAOMT), however, stresses uniform responses among all the species to the divergence of PAL and C4H at the family level, strong indication for diversifying selection.

3.3.3 Evolutionary trends in the expression of coniferous phenolic pathways

We used the divergence of the expression of phenolic gene families in bark as a surrogate to understand the trends in diversification of phenolic biosynthetic pathways. The divergence of the expression of gene families in bark was proportional to the divergence of their corresponding sequences (Table 3.4; Fig. 3.4). Working with the array ESTs, it was not possible to estimate the rates of DNA substitution with precision, and further correlate with the divergence of expression. However, significant relationship ($P > 0.01$) between divergence at the two levels of DNA sequences and expression (Table 3.4) could be taken that much of the observed divergence is likely to be present at the structural level. Therefore,

divergence of the expression of gene families must be of adaptive importance (less likely to be explained by drift). This pattern was shown to be stronger for the downstream gene families, GLYTR, and pPCBER. However, DIR was the gene family exhibiting the strongest relation between the divergence of expression and DNA sequences (Fig. 3.4). With respect to the magnitude (> 3 absolute values in *Z*-scores) of the divergence in the expression, this family could be treated as an outlier. Removing DIR as an outlier does not largely affect the overall relation between the divergence of expression and sequences (Table 3.4, Solution II). Nevertheless, this result could emphasis the different nature of the members of this gene family compared to the other families. The DIR family proteins dictate the stereochemistry of compounds synthesized by other enzymes. Therefore, the need for the modification of the products from a diverse array of evolving phenolic genes is a potentially strong driving force for the high rate of divergence in the expression of their gene family.

Overall, the pattern of the divergence of bark gene families was supported by patterns of nucleotide diversity reported in angiosperms by Ramos-Onsins et al. (2008) and Kuelheim et al. (2009) in Arabidopsis and Eucalyptus, where strong diversity trends along the sequential positioning of the phenylpropanoid and flavonoid pathway genes had not been recovered. Although the divergence of families such as GLYTR, or LAC indicates higher diversification downstream of the pathways (leading to the biosynthesis of lignin), our findings of the divergence of pCCoAOMT, or pDFR suggest diversification at upstream, and entry points of the associated pathways respectively (Fig. 3.1). Notably, divergence of PAL positioned at the upstream most point of the general phenylpropanoid pathway, suggests specific diversification of single steps of the phenolics pathway for which participation in general or specific metabolism is not fully understood. The divergence of upstream gene

families demonstrates conifer-specific pathway variations, and can be an indication for the yet unknown auxiliary pathways shunting from the upstream of the phenylpropanoid pathway.

Our findings were in contrast with the findings of the terpenoids pathway (Ritland and Ramsay, 2009) where the evolution of the downstream is marked by a higher degree of diversification compared to the upstream genes. A possible explanation for this trend might be the more distant separation in the functionality of the downstream metabolites of the terpenoid pathways compared to phenolics. An example of such separation could be seen downstream in the parallel biosynthesis of gibberellins and defense-related diterpens, contrasting features of the primary and specialized metabolism.

3.4 Concluding remarks

We used divergence of gene expression as a proxy to understand the relevance of the phenolics biosynthesis in adaptive evolution. Our independent treatment of the group of genes studied here among the other components of plant defense can indicate such relevance. In our inferences of the evolution of pheolics expression, we have used statistical analyses. Taken the results of these analyses to discuss the evolution of conifer defense implies that: 1) outer stem is the instrumental source of tissue in the diversification; 2) inferences about natural selection for gene expression holds at the level of gene families; 3) divergence of the various groups of genes should be considered integrated to fit in the concept of polygenic defense response.

The practical implication of the results of this paper (in the absence of a sequenced genome in conifers and limited understanding of the function of the phenolic genes) is in targeted functional approaches and understanding the coniferous variations of phenolic

genes. For instance, pCCoAOMT, and pPCBER families could be suitable candidates in which the divergence of the expression was proportional by sequence divergence. The functional divergence of these families in angiosperms places emphasis on the importance of these genes for conifer specific studies. In that, we found our combinatory bioinformatics approach a powerful strategy for the study of the evolution of phenolic genes.

Table 3. 1 Mixed model analysis of cross-species microarray data from three different sources of tissue testing for the divergence of overall phenolic gene expression among five species. DF, degrees of freedom; numDF, numerator of DF; denDF, denominator of DF; Species x Category, species by category (family) interaction.

Tissue source	Response	numDF	denDF	*F*	*P*
Bark/phloem	Intercept	1	4875	776429700	<0.01**
	Species	4	10	5.41	0.01 *
	Category	18	4875	27.61	<0.01**
	Species x Category	72	4875	1.02	0.42
Needle	Intercept	1	4875	1481091	<0.01**
	Species	4	10	1.16	0.39
	Category	18	4875	41.09	<0.01**
	Species x Category	72	4875	0.49	0.99
Xylem	Intercept	1	4875	3676270	<0.01**
	Species	4	10	0.75	0.58
	Category	18	4875	44.79	<0.01**
	Species x Category	72	4875	0.5	0.99

* Significant at $\alpha < 0.05$

**Significant at $\alpha < 0.01$

Table 3. 2 Summary of mixed effects analysis of variance for the expression (response) of each gene category in bark and phloem

DF, degrees of freedom; nmDF, numerator of DF; dnDF, denominator of DF; Species x Gene, species by gene interaction; Var. comp. between, between-species variance component (*Z*-scores); Var. comp. within, within-species variance component (*Z*-scores).

Category	Factor	nmDF	dnDF	*F*	*P*	*Q*	Var. comp. between	Var. comp. within
2-Oxoglutarate Ferrous dependent oxygenase (2OGFE)	Intercept	1	560	727433200	< 0.01		0.05	-0.62
	Species	4	10	1.19	0.37	0.33		
	Gene	56	560	88.81	< 0.01			
	Species x Gene	224	560	2.05	< 0.01			
4-Coumaryl CoA ligase (4CL)	Intercept	1	140	200902400	< 0.01		-0.13	-0.41
	Species	4	10	0.9	0.50	0.37		
	Gene	14	140	50.6	< 0.01			
	Species x Gene	56	140	1.16	0.24			
Coumarate 3-hydroxylase (C3H)	Intercept	1	20	28483820	< 0.01		-0.45	1.81
	Species	4	10	1.32	0.33	0.33		
	Gene	2	20	65	< 0.01			
	Species x Gene	8	20	3.95	0.01			
Cinnamate 4-hydroxylase (C4H)	Intercept	1	10	52931640	< 0.01		-1.48	-0.82
	Species	4	10	5.23	0.02	0.03*		
	Gene	1	10	< 0.01	0.95			
	Species x Gene	4	10	0.53	0.72			

Table 3. 2 continued

Category	Factor	nmDF	dnDF	*F*	*P*	*Q*	Var. comp. between	Var. comp. within
Cinnamyl alcohol dehydrogenase (CAD)	Intercept	1	120	203295700	< 0.01		-0.3	-0.82
	Species	4	10	0.68	0.62	0.39		
	Gene	12	120	130.53	< 0.01			
	Species x Gene	48	120	1.26	0.16			
Cinnamoyl CoA reductase (CCR)	Intercept	1	160	255401300	< 0.01		-0.2	-0.82
	Species	4	10	0.87	0.51	0.37		
	Gene	16	160	83.49	< 0.01			
	Species x Gene	64	160	1.44	0.04			
Chalcone isomerase (CHI)	Intercept	1	30	83572510	< 0.01		-0.99	-0.82
	Species	4	10	0.72	0.6	0.39		
	Gene	3	30	97.23	< 0.01			
	Species x Gene	12	30	1.59	0.15			
Chalcone synthase (CHS)	Intercept	1	90	26725600	< 0.01		0.45	2.59
	Species	4	10	2.26	0.14	0.17		
	Gene	9	90	62.78	< 0.01			
	Species x Gene	36	90	1.66	0.28			

Table 3. 2 continued

Category	Factor	nmDF	dnDF	*F*	*P*	*Q*	Var. comp. between	Var. comp. within
Dirigent-like proteins (DIR)	Intercept	1	290	232824300	< 0.01		0.28	0.08
	Species	4	10	19.88	< 0.01	< 0.01*		
	Gene	29	290	65.68	< 0.01			
	Species x Gene	116	290	1.68	< 0.01			
Flavonoid 3-hydroxylase (F3H)	Intercept	1	120	99306480	< 0.01		-0.21	0.45
	Species	4	10	1.20	0.37	0.33		
	Gene	12	120	118.14	< 0.01			
	Species x Gene	48	120	1.69	0.01			
Ferulate 5-hydroxylase-like (F5H-like)	Intercept	1	80	108296033	< 0.01		0.12	-0.84
	Species	4	10	1.81	0.2	0.23		
	Gene	8	80	44.41	< 0.01			
	Species x Gene	32	80	1.02	0.45			
Glycosyl transferase (GLYTR)	Intercept	1	520	530103600	< 0.01		0.44	-0.23
	Species	4	10	8.4	< 0.01	0.01*		
	Gene	52	520	52.15	< 0.01			
	Species x Gene	208	520	2.76	< 0.01			

Table 3. 2 continued

Category	Factor	nmDF	dnDF	*F*	*P*	*Q*	Var. comp. between	Var. comp. within
Laccase (LAC)	Intercept	1	240	122899700	< 0.01		-0.6	-0.72
	Species	4	10	9.93	< 0.01	< 0.01*		
	Gene	24	240	93.58	< 0.01			
	Species x Gene	96	240	3.57	< 0.01			
O-methyl transferase (OMT)	Intercept	1	200	172087100	< 0.01		0.58	0.07
	Species	4	10	6.22	< 0.01	0.02*		
	Gene	20	200	39.17	< 0.01			
	Species x Gene	80	200	3.17	< 0.01			
Phenylalanine ammonia lyase (PAL)	Intercept	1	40	60786880	< 0.01		-1.5	1.09
	Species	4	10	6.92	< 0.01	0.02*		
	Gene	4	40	113.13	< 0.01			
	Species x Gene	16	40	1.89	0.05			

Table 3. 2 continued

Category	Factor	nmDF	dnDF	*F*	*P*	*Q*	Var. comp. between	Var. comp. within
Caffeoyl CoA O-methyl transferase, putative (pCCoAOMT)	Intercept	1	80	47740450	< 0.01		2.17	0.66
	Species	4	10	5.49	0.01	0.03*		
	Gene	8	80	29.26	< 0.01			
	Species x Gene	32	80	0.75	0.82			
Dihydro flavonol reductase, putative (pDFR)	Intercept	1	180	324458600	< 0.01		-0.52	-0.82
	Species	4	10	0.89	0.50	0.37		
	Gene	18	180	105.07	< 0.01			
	Species x Gene	72	180	3.89	< 0.01			
Phenylcoumaran benzylic ether reductase, putative (pPCBER)	Intercept	1	265	206905200	< 0.01		1.29	-0.26
	Species	4	10	4.26	0.03	0.04*		
	Gene	25	265	42.15	< 0.01			
	Species x Gene	100	265	2.08	< 0.01			

* Significant *Q* value (false discovery rate level: 0.05)

Table 3. 3 Summarizing correlations of the expression differences with neutral genetic distances averaged over each diverged gene families.

Spearman rank correlation coefficient (Appendix A3) of 0.5or higher is considered substantial. CC, correlation coefficient; St. error, standard error for the arithmetic mean of the correlation coefficients per category; N, number of genes in each category.

Category	N	Minimum	Maximum	CC	
				Mean	St. error
C4H	2	-0.08	0.2	0.06	-
DIR	30	-0.12	0.49	0.08	0.03
GLYTR	53	-0.19	0.51	0.09	0.02
LAC	25	-0.19	0.42	0.1	0.04
OMT	21	-0.18	0.33	0.07	0.03
PAL	5	-0.03	0.25	0.13	0.05
pCCoAOMT	9	-0.2	0.23	0.01	0.05
pPCBER	27	-0.12	0.4	0.14	0.03

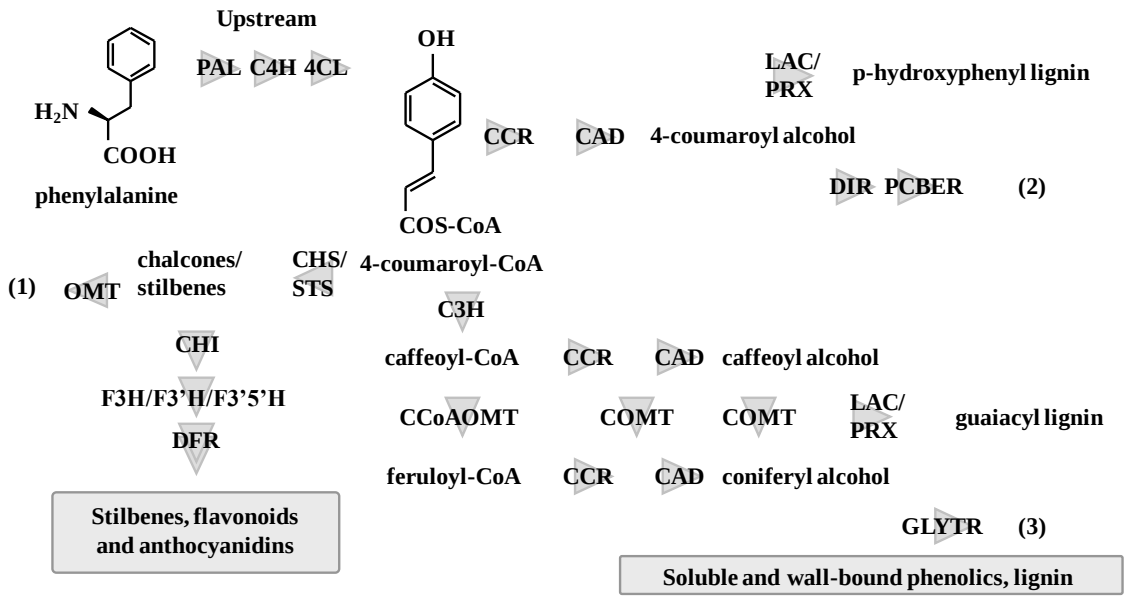
Table 3. 4 Relationship between the divergence of the expression gene families and the divergence of their corresponding sequences.

On the left side of the table ordinary least squares (OLS) results are summarized. Least squares is fitted to the *Z*-scores calculated from *F* species expression of the mixed model analysis averaged within families (dependent), and average DNA distance within families *Z*-scores(predictor). Right side of the table presents Wilcoxon signed rank test. The analyses are repeated with the total number of gene families (I) and without DIR gene family (II). N, frequencies; Asymp. sig., asymptotic significance (two-tailed).

Solution	OLS				Signed Rank test				
	Source	Sum of squares	Mean Square	R^2	Ranks	N	Mean rank	Sum of ranks	Asymp. sig.
I	Model†	3.87	3.87	0.23	Negative ranks [a]	2	4	8	< 0.01
	Error	13.14	0.77		Positive ranks [b]	16	10.19	163	
	Uncorrected total	17			Ties [c]	0	-	-	
II	Model†	1.04	1.04	0.15	Negative ranks [a]	2	4	8	< 0.01
	Error	5.79	0.36		Positive ranks [b]	15	9.67	145	
	Uncorrected total	6.83			Ties [c]	0	-	-	

† no intercept included; [a] *F* species expression < average DNA distance; [b] *F* species expression > average DNA distance; [c] *F* species expression = average DNA dist

a)



b)

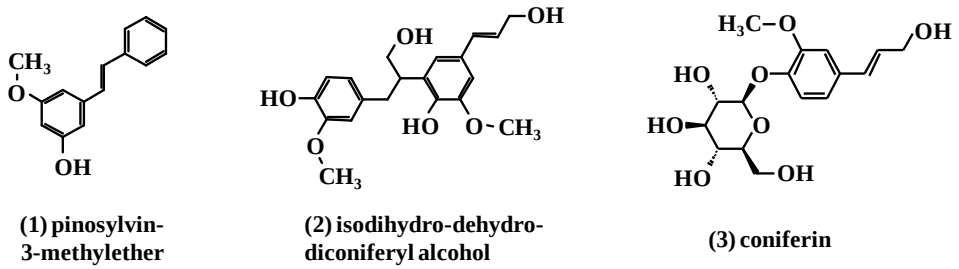
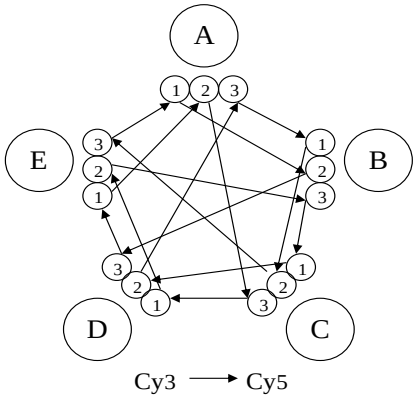


Figure 3. 1 A simplified schematic representation of the core phenylpropanoid pathway to the downstream of the phenolic pathways.

Downstream of the pathways leads to the biosynthesis of (a) stilbenes, flavonoids and anthocyanidins, as well as soluble and wall-bound phenolics and lignin, culminating to (b) selected conifer phenylpropanoid-derived specialized metabolites shown with chemical structure. Enzymatic steps are indicated by triangles. PAL, Phenylalanine ammonia lyase [EC:4.3.1.24]; C4H, Trans-cinnamate 4-monooxygenase [EC:1.14.13.11]; 4CL, 4-coumarate:CoA ligase [EC:6.2.1.12]; CHS, Chalcone synthase [EC:2.3.1.74]; STS, Stilbene

synthase [EC: 2.3.1.95]; OMT/COMT, O-Methyl transferase/Caffeic acid 3-O-methyltransferase [EC: 2.1.1.68]; CHI, Chalcone isomerase [EC:5.5.1.6]; DFR, Dihydroxyflavonol 4-reductase [EC:1.1.1.219]; F3H/ F3’H/; F3’5’H, Flavanone 3β-hydroxylase/Flavonoid 3’-hydroxylase/Flavonoid 3'5'-hydroxylase[EC:1.14.-.-]; CCR, Cinnamoyl CoA reductase [EC:1.2.1.44]; CAD, Coniferaldehyde dehydrogenase [EC:1.1.1.195]; CCoAOMT, Caffeoyl CoA O-methyltransferase [EC:2.1.1.104]; DIR, Dirigent-like protein; PCBER, Phenylcoumaran benzylic ether reductases [EC: 1.3.1.-]; PRX/LAC, Laccase/peroxidases [EC:1.11.1.-]; GLYTR, Glycosyl transferase [EC 2.4.1.-].

a)



b)

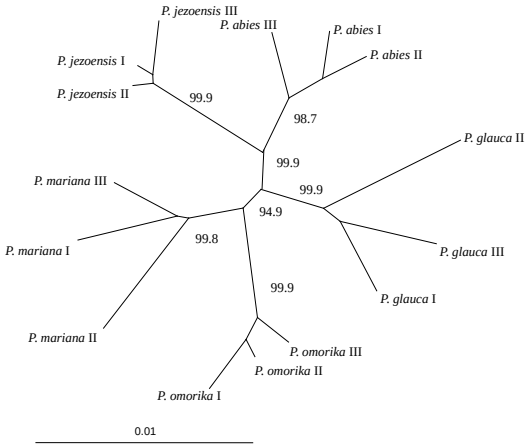


Figure 3. 2 Cross-species comparisons scheme.

(a) Experimental design for cross-species hybridizations on a two channel microarray (green Cy3 and red Cy5). The design consists of 3 biological replicates (individuals) of 5 species, *Picea abies* (A), *P.mariana* (B), *P. omorika* (C), *P. jezoensis* (D), and *P. glauca* (E), including 15 hybridizations; (b) Unrooted neighbor-joining tree of the species on the individuals basis constructed from AFLP genetic distances with bootstrap scores (1000 replicates). The biological replicates are denoted by Roman numbers.

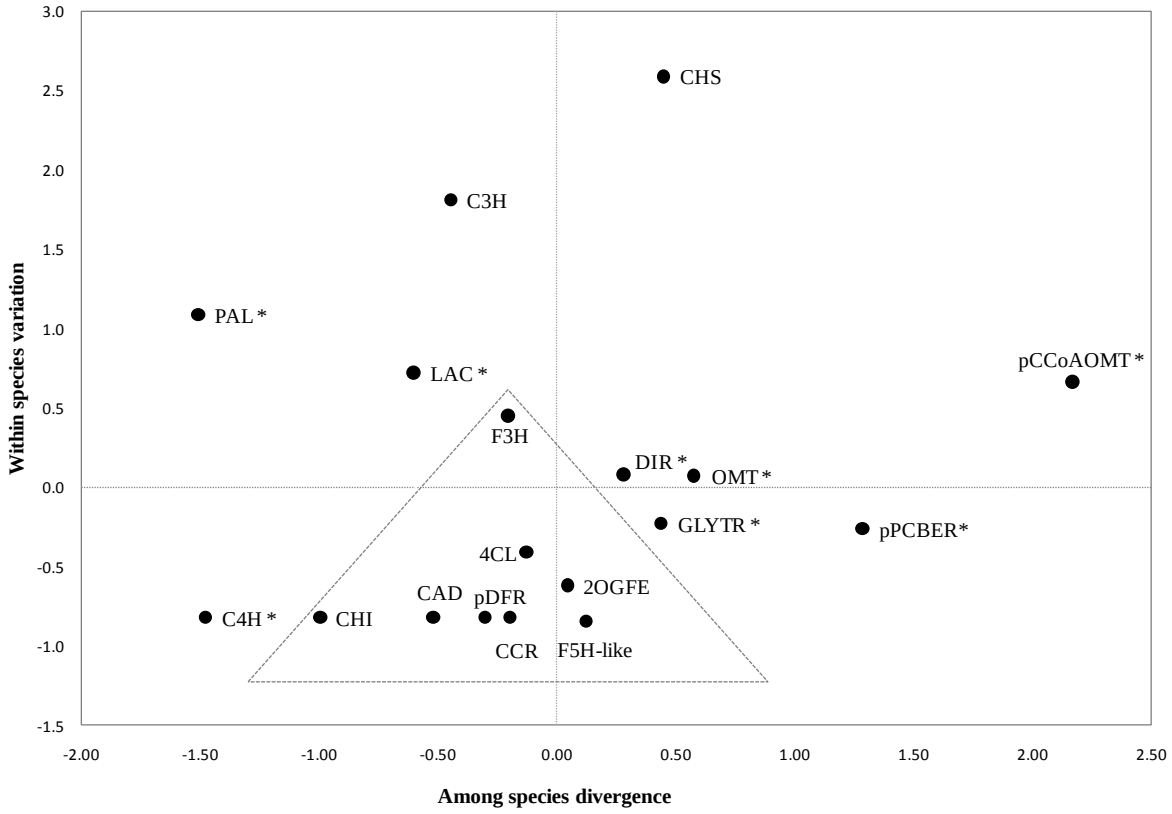


Figure 3. 3 Plot of the variance components (centered and standardized) from Table 2.2. The vertical and horizontal axes represent within and between species variance components respectively. The categories significantly diverged among species ($Q < 0.05$) are emphasized with an asterisk. A triangle is used to proximate the hypothetical space of stabilizing selection. Family labels follow Table 3.2.

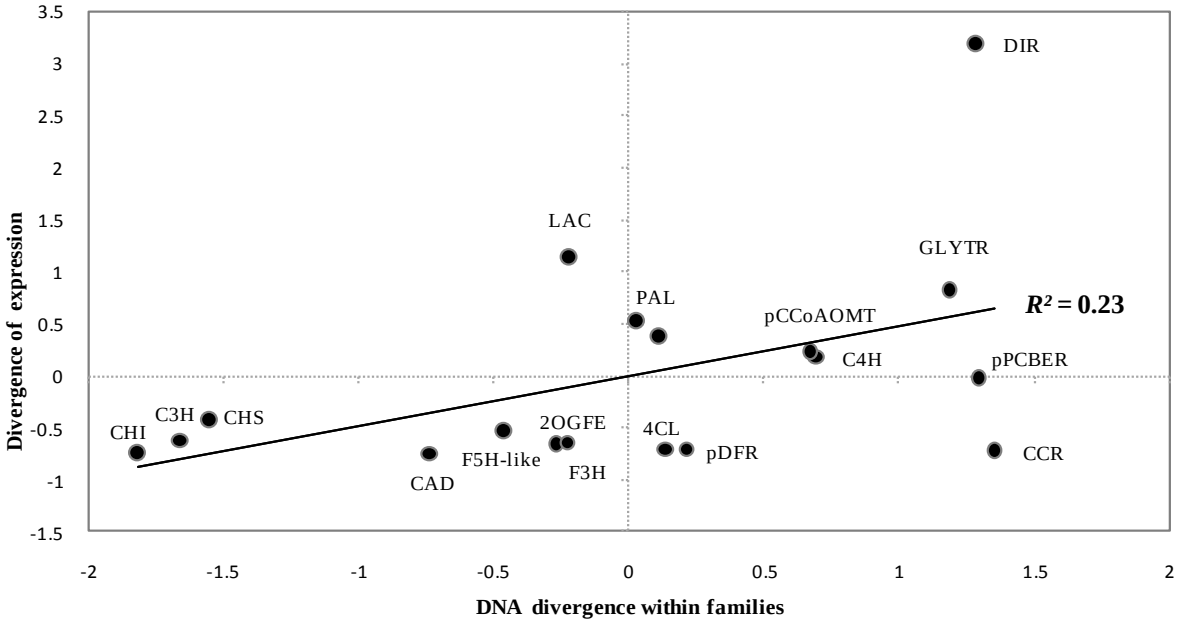


Figure 3. 4 Relationship between the divergence of the expression of gene families and their corresponding DNA sequences.

Horizontal axis represent average DNA distance within family, and the vertical axis *F* value of species per family from the linear mixed model analysis (values are centered and standardized). The trend line shows the least squares fit (with R^2) to the Z-scores. Family labels follow Table 3.2.

4. Estimating heritability of gene expression using parent-offspring regression with 2 channel microarrays

4.1 Introduction

The newly developed field of genetical genomics makes use of microarray technology to (a) infer regulatory networks controlling gene expression (Chesler et al. 2004; Bystrykh et al. 2005); (b) map expression quantitative trait loci (Jansen and Nap, 2001; Brem et al. 2002; Darvasi, 2003; Schadt et al. 2003); and (c) infer the heritability of transcript abundances (Monks et al. 2004; Vuylsteke et al. 2005). The importance of understanding the genetic basis of gene expression is predicated on the widely held view that phenotypic diversity is generated not only by changes of DNA sequence, but also by changes in the levels of gene expression (Li and Burmeister, 2005).

The focus of this article is on the estimation of heritability of transcript abundances using parent-progeny regression, specifically with the single parent-offspring design. This is one of four major designs for inferring the heritability of a quantitative trait (the others being midparent-offspring, half sib family, and full sib family designs; Falconer, 1989). Parent-offspring regression is the most straightforward method for estimating heritability for three reason. First, because it is possible to base the essential computations on least-squares regression; the statistical properties are well known. Second, neither dominance nor linkage influences the covariance between parents and offspring. Third, it is not biased when parents are selected on the basis of their phenotype (see Lynch and Walsh, 1998 for an in-depth discussion).

Traditionally, to estimate the heritability for a quantitative trait of interest, measurements are taken directly on parents and offspring. This is followed by regression of offspring measurements upon parent measurements; the slope of the regression is

proportional to the heritability of the trait. For many species, parents and their offspring are easily identified in the field, and in plants, progeny can be sampled as seed. However, in many species including most plant species, only one parent can be identified: the mother, as the male parent is an unknown pollen donor (Lynch and Walsh, 1998). In tree breeding programs, the genetic value of a candidate "plus" tree is often evaluated by growing open-pollinated progeny of a tree, where the progeny are produced by cross-pollination to many male parents of unknown location. Such material is also used to evaluate heritability of traits important in breeding programs.

With the advent of microarrays, it is possible to view the transcriptome of an organism as a suite of quantitative traits (Rockman and Kruglyak, 2006). An obvious first question in genetical genomics is regarding the extent to which levels of transcription are genetically determined; more specifically, what is the heritability of transcript levels? If gene expression is heritable, then natural selection can act on differences of transcript abundances to increase fitness, and adaptive evolution of gene expression will take place.

However, in two color microarrays, transcript abundance is not directly observable; rather, the difference of transcript abundance between two labelled mRNA populations (one for each color) is observed. This is of no major issue if one is interested, say, in the response of the transcriptome to some experimental treatment, as compared to a control treatment. But with genetical genomic inferences, a more complex experimental design than a common reference or circular structure is needed (Bueno-Filho et al. 2006). The literature on comparison of the alternative designs specific to the estimation of heritability of gene expression is limited. In particular, the case of parent-offspring regression with two channel microarrays has not been examined. This is of immense importance for the research

programs that need to utilize two-color microarrays as an alternative to oligomere platforms due to the budget restrictions, or the nature of the biological systems under study. Here, three alternative designs for single parent-offspring regression are introduced and examined with respect to bias and statistical power for inferring heritability of gene expression with two-channel microarrays.

4.1.1 Extension of the concept on two-channel microarrays

Two-channel microarrays do not measure absolute values of gene expression levels (unlike real-time PCR), but rather the differences between the green (Cy3) channel and red (Cy5) channel hybridization intensities are obtained. The objective is thus to estimate the covariances between parent and offspring for gene expression, and the variance of gene expression among parents, using differences of gene expression. For a single parent and an offspring where, say the parent RNA is labeled with Cy3 and the progeny RNA labeled with Cy5, let the i-th parent have a level X_i, and its progeny have a level Y_i. In the population, we define V_X as the variance of all X_i, V_Y the variance of all Y_i, and C_{XY} the covariance between X_i and Y_i. Because parent and progeny may be of different age, or differ for other factors, we allow for the possibility that parents have different means than their progeny: let X_i have mean U_x and Y_i have mean U_y. Between a parent and an offspring, the difference of gene expression is (X_i - Y_i). To solve for the heritability using paired differences, we also require the observation of the difference of gene expression between parent i and unrelated progeny j, (X_i - Y_j), and the difference of gene expression between unrelated parents i and j, (X_i - X_j). The expected squared differences of these quantities are:

$$\mathrm{E}\left((X_i - Y_i)^2\right) = U_X^2 + V_X + U_Y^2 + V_Y - 2U_X U_Y - 2C_{XY} = a$$

$$\mathrm{E}\left((X_i - Y_j)^2\right) = U_X^2 + V_X + U_Y^2 + V_Y - 2U_X U_Y = b$$

$$\mathrm{E}\left((X_i - X_j)^2\right) = 2V_X = c$$

From these three quantities, we can solve for the heritability as twice the regression of offspring on parent values:

$$h^2 = 2b_{op} = 2\left(\frac{C_{XY}}{V_X}\right) = 2\left(\frac{b-a}{c}\right) \qquad \text{(Equation 4.1)}$$

In any actual experiment, *a*, *b* and *c* are found by equating the above expectations to the corresponding sample moments. For example, $a = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2$ for n parent-offspring pairs.

If the mean values are the same in both generations, and the variances also the same, then

$$\mathrm{E}\left((X_i - Y_i)^2\right) = 2V_X - 2C_{XY} = a$$

and

$$h^2 = 2b_{op} = 2\left(\frac{C_{XY}}{V_X}\right) = 2\left(\frac{c-a}{c}\right) \qquad \text{(Equation 4.2)}$$

Hence, the hybridization of parent *i* to offspring *j* is not needed (Fig. 4.1).

4.2 Evaluation of alternative hybridization designs

Here, three alternative designs of paired samples, "chain design", "independent quartets design" and "completely independent design" that allow estimation of heritability using a two-dye microarray experiment are considered (Fig. 4.2). To balance these designs, equal numbers of Cy3 and Cy5 hybridizations need to be done. For this, each maternal parent and each offspring has to be hybridized four and two times respectively as illustrated by the directions of the arrows.

To evaluate the statistical properties of these three designs, we generated simulated data with pre-specified heritabilities. For each parent-progeny pair, two independent, normally-distributed random numbers with variances of one were generated (denoted Z_1 and Z_2). Then we transformed these as:

$$X_i = Z_1$$

$$Y_i = \sqrt{1-(h^2)^2}\,Z_2 + h^2 Z_1$$

where h^2 is the pre-specified, true heritability. This results in an expected covariance between X_i and Y_i equal to h^2, and a variance equal to unity for both parents and progeny. We also considered a fourth design, "common reference design", used by Monks et al. (2004) in their study of the heritability of human gene expression. In this design, one of the two RNAs are simply pooled RNA from all samples, and used in one of the dyes. The difference of the parent from the reference is the direct observation of the level of gene expression in the parent; likewise for the progeny. Note that two, instead of three, microarrays are required for each parent-progeny pair. Parent-offspring regression is then performed on these directly inferred levels of expression.

We considered sample sizes ranging from 3 to 30 parent-progeny pairs, involving each of the three designs in Fig. 4.2 plus the reference design, and evaluated a range of true heritabilities. 100,000 replicates were run to evaluate the mean and variance of the estimated heritabilities. The FORTRAN code for the simulations is available upon request.

To assess the effect of single-gene inheritance on the above estimators of heritability, we considered a simple scenario of a single diallelic locus with equal gene frequencies and with additive effects of $-\sqrt{2}$, 0 and $+\sqrt{2}$ for the homozygous, heterozygous and alternative homozygous genotypes. This results in a variance of 1 as used above. Progeny data were generated by drawing one parent allele at random, and choosing a second parent allele at random. Only the chain design was considered.

4.3 Results and discussion

The results of simulations are illustrated in Fig. 4.3 as the average variance of estimates per parent-progeny pair (the observed variance divided by the number of parent progeny pairs, for a range of sample size) vs. the number of parent progeny pairs used. Each design was evaluated over the range of possible heritabilities; Fig. 4.3 shows the case for h^2=0.5; other heritabilities showed the same pattern. In all designs, the standard error (SE) becomes roughly asymptotic for 15 pairs. Thus, it is necessary to include at least 15 parent-progeny pairs for a reasonable degree of precision. This is in line with previous findings related to the estimation of heritability as Lynch and Walsh (1998) have also established number of families N >15 for a reasonable degree of precision.

Figure 4.3 (and other simulations not shown involving different heritabilities) also clearly illustrates that the chain design is much more efficient than the other two paired designs, and is also more efficient than a design where one dye consists of pooled RNA from

all samples. Therefore, the chain design is most efficient and quite fortuitous for the practical reason as it requires the least number of RNA extractions.

It is also significant that the chain design is more efficient than the reference design, even taking into account that three instead of two arrays per parent-progeny pair are required. On a per-array basis, the asymptotic information between the two designs can be compared by computing information values (the inverse of the variance) per array. For the chain design, the information per parent-progeny pair is about (1/0.24) = 4.16 for the three arrays required, or about 1.40 per array. For the reference design, the information is about (1/0.83) =1.20 for the two arrays, or 0.60 per array. Hence, the chain design has over twice the power of the reference design.

For the chain design, we also considered the bias and variance of heritability estimates as a function of the true heritability. Table 4.1 shows that there is no bias (to 0.01 precision) and that the estimation variance decreases as the heritability increases down to zero when h^2=1 (as parents and progeny are perfectly correlated).

The nearly asymptotic sample size of 15 is used here as a minimum sample size requirement. For an experiment involving 15 parent -progeny pairs (45 hybridization), the average SE is found to be half of the true heritability at about h^2=0.5. Hence, at h^2=0.5 and higher, the estimated heritability is expected to be statistically significant.

Monks et al. (2004) found that when a common reference design is used, a sample size of 15 provides 28% power to detect $h^2 = 0.1$, 63% power to detect $h^2 = 0.2$, 85% power to detect $h^2 = 0.3$, 94% power to detect $h^2 = 0.4$, and 100% power to detect $h^2 \geq 0.5$. Our results here show that a paired design, as opposed to using a common reference design, is

considerably more statistically efficient, which should improve the power values they presented.

Because individual elements of an array are the products of single gene loci, we also considered the effect of simple (one-locus) inheritance on the estimators we presented. Genetic variation at the actual gene underlying the mRNA transcript, at a *cis*-acting regulatory element, or at a trans-acting regulatory element, has the potential to cause mono-genetic inheritance for a given array element, if no other genes are involved in regulation. We found, based upon the above described single-gene inheritance simulation, that unbiased estimates of heritability were still obtained. With no environmental effect added, the estimated heritability equaled unity (the true value), and the estimate of heritability declined in proportion to the relative amount of environmental variance added (results not shown). However, the variance of the estimate was several times greater; under the chain design, the variance of h^2 per parent-progeny pair was asymptotically about 5.5 when true h^2=1, and about 3.5 when true h^2=0.5. Most interestingly, this trend is the opposite of what is found for continuous traits (Table 4.1); more estimation variance with greater genetic determination. This is due to the greater stochasticity caused by the presence of discrete phenotypes.

Recent studies imply pervasive non-additivity of gene expression and that for some transcripts, regulatory polymorphism in *cis* and in *trans* could affect expression (Gibson and Weir, 2005). In this research, the objective was rather to evaluate a number of alternative designs using a straightforward quantitative genetic methodology, than to dissect the complicated genetic basis of transcription. Estimation of non-additive effects would involve sibship designs, and could be a subject of further research.

In searching for an optimal design for estimating the heritability of gene expression, general strategies have been discussed by Rosa et al. (2005) and Bueno-Filho et al. (2006) based on statistical precision or power, and more specifically with two-channel microarrays applications by (Wit et al. 2005). Here, we compared a set of balanced designs (i.e. each parent and each offspring is hybridized once with the green and once with the red dye). Therefore, we limited the analysis of our hypothetical parent-offspring scenarios to least-squares methodology.

However, extension of the concept of single parent–offspring design to gene expression requires that the parent and offspring were sampled at the same developmental stage, since gene expression levels are specific for each developmental stage. In the laboratory, this could be achieved by using tissue of parents and offspring cultured under similar conditions. However, circumstances might necessitate that the parents and offspring of different ages be sampled under different environmental conditions which, in turn, complicates the design and analysis of the data. Using mixed model analyses that account for environmental differences may be more suitable in these cases.

Table 4. 1 Average estimated heritabilities (h^2) and their standard errors (SE). Estimates computed under the chain design with 15 parents, for various values of true heritabilities.

true h^2	est. h^2	Ave. SE
0.00	0.00	0.417
0.10	0.10	0.373
0.20	0.20	0.329
0.30	0.30	0.285
0.40	0.40	0.244
0.50	0.50	0.203
0.60	0.60	0.162
0.70	0.70	0.122
0.80	0.80	0.082
0.90	0.90	0.041
1.00	1.00	0.000

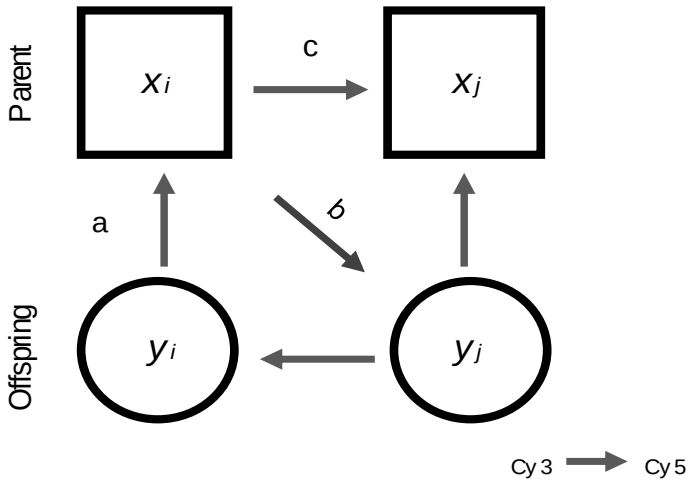


Figure 4. 1 Parent-offspring alternatives.

Some alternatives of the many possible maternal parent (*x*) and off spring (*y*) combinations are demonstrated, including the hybridizations (a, b, and c) necessary to perform in order to solve for h^2, the heritability of gene expression. With respect to Equations 4.1 and 4.2, the required pair wise combinations include hybridization of (a) maternal parent with her own offspring; (b)with another maternal parent; and (c) offspring of an adjacent pair.

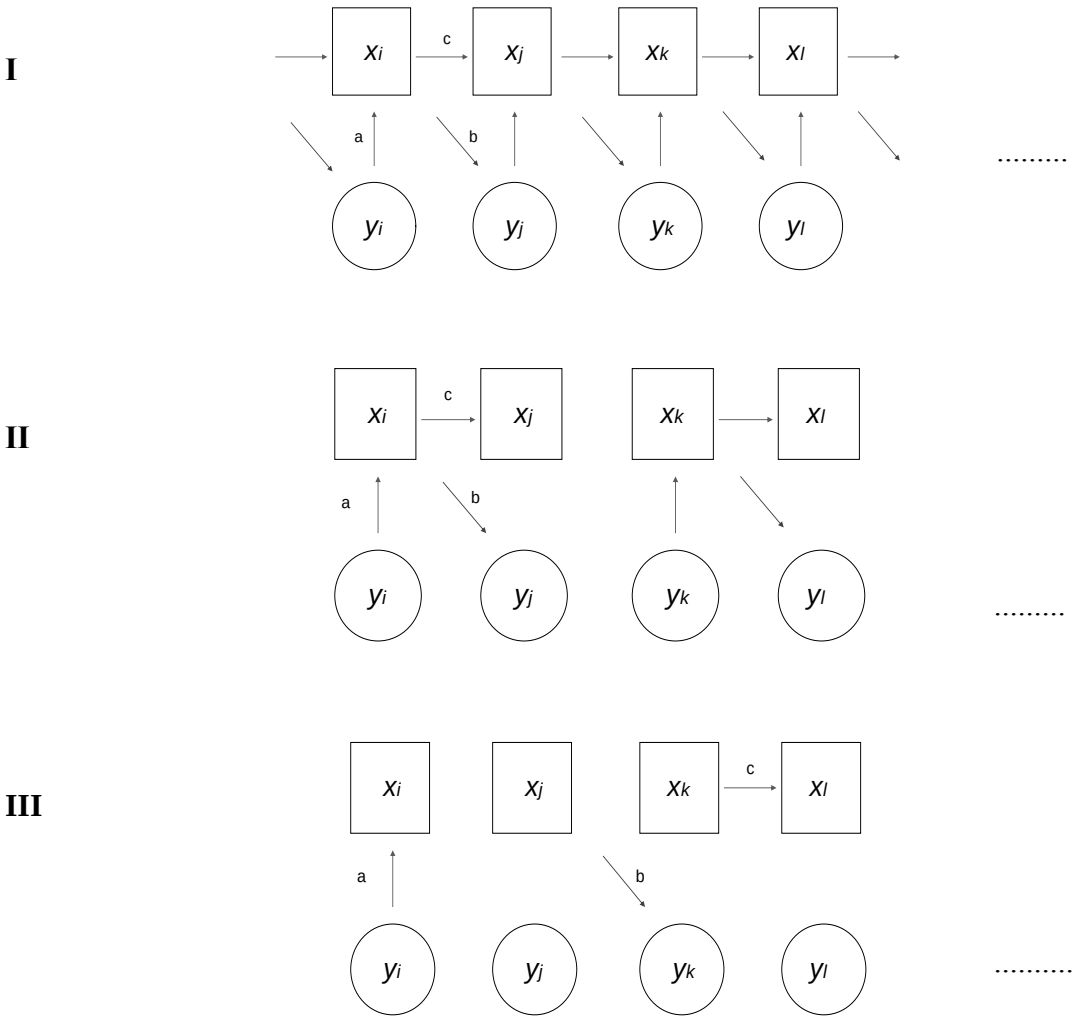


Figure 4. 2 Alternative experimental designs.

Three alternative designs involving parent-offspring pairs when (I) parent-progeny pairs are connected as a chain, (II) form independent quartets, and (III) all pair wise hybridizations are independent. Dotted lines indicate the continuation of the pattern until satisfaction of the sample size requirement.

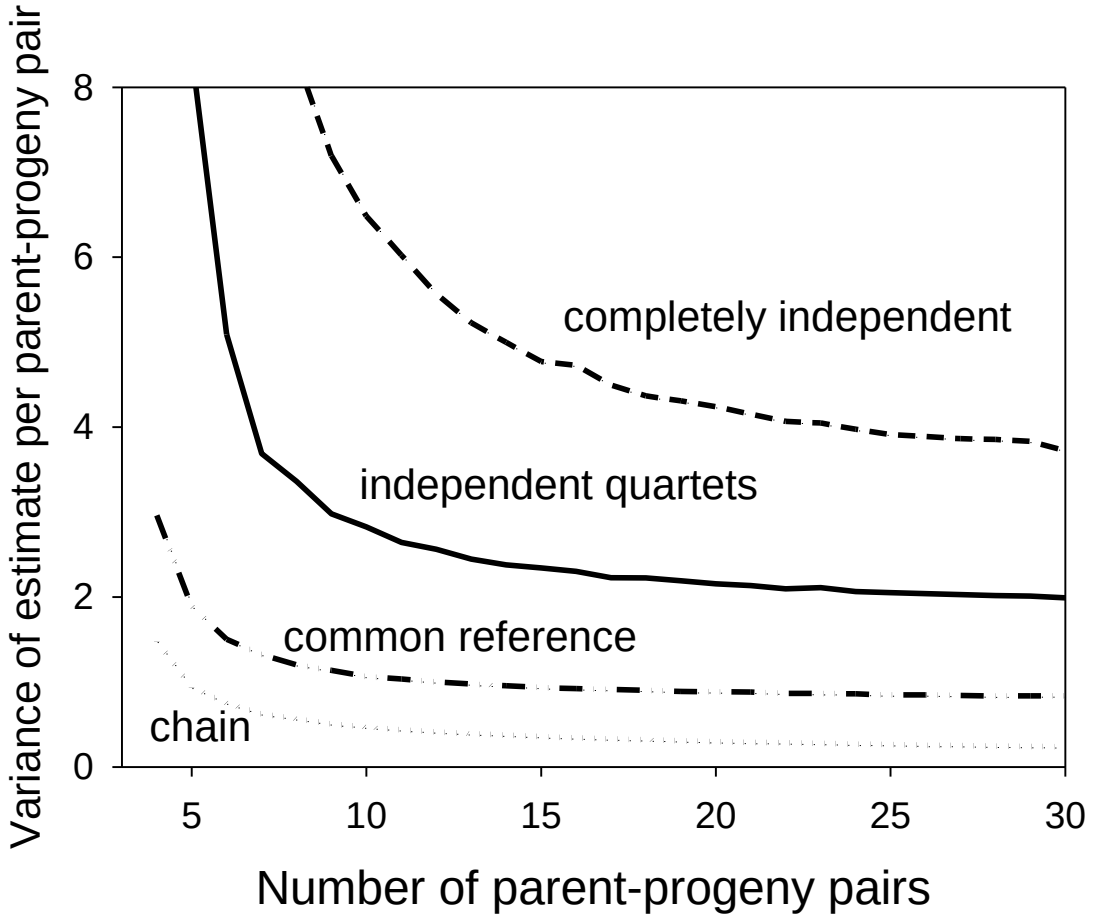


Figure 4. 3 Results of simulations under three alternative designs.

Results of simulations under three alternative designs for estimating heritability using paired samples of parents and offspring, plus the design of a common reference.

Plotted are variance of estimate per parent-progeny pair, vs. the number of parent-progeny pairs used. True heritability was 0.5; bias was evident only for the "completely independent" design.

5. A novel experiment reveals heritability of terpenoid gene expression

5.1 Introduction

Evolution by natural selection requires heritable variation in traits. Heritability of a trait is a measure that expresses the genetic proportion of its phenotypic variance. The magnitude of the genetic variance is, therefore, a key determinant in the relation between phenotypic variation and adaptive evolution. The ratio of the additive genetic variance to phenotypic variance is narrow-sense heritability, which is a practical key measure in breeding methods for improving plants or animals (Falconer and Mackay, 1996).

In plants, a diverse array of terpenoid compounds (isoprenoid derivatives) are known or assumed to have specialized for ecological functions. Interacting within the context of reproduction, defense or symbiosis, terpenoids act as attractants, repellents, anti-feedants, toxins, or antibiotics (Bohlmann and Keeling, 2008). Arguably, such specializations in the function of terpenoids result in increased fitness.

Earlier works on the heritability of terpenoids have been based on the measurements of their end products, metabolites. The pioneering studies that have explicitly estimated the parameters of heritability using the terpenoid metabolite profiles come from the works on the coniferous family Pinaceae. For instance, for monoterpenes heritabilities have been estimated in Slash pine (*Pinus elliottii*; Squillace, 1971) and Loblolly pine (*P. Taeda*; Rockwood, 1973) which have generally been substantial (50 % or higher).

The advent of microarrays allows treating the abundances of gene transcripts as a phenotype. Accordingly, inferences about heritability have used gene transcripts as opposed to gene products (e.g. Monks et al. 2004; Vuylsteke et al. 2005). An advantage of this approach is that heritability can be dissected to segments of a metabolic pathway. Using a case-control design, transcription profiling has shown the impact of differential expression of

pathway points on plant defense mechanism (Ralph et al. 2006). In a similar manner, heritability estimates related to the expression of pathway points can be used to assess their relative contribution in plant fitness.

Outer stem tissues of coniferous species comprise a defense barrier against organisms trying to reach nutrient-rich phloem. Diversification of terpenoids has been recognized as a defense mechanisms in conifers (Franceschi et al. 2005; Keeling and Bohlmann, 2006b). At the level of gene pathways, there is evidence for increased diversification from upstream to the downstream, as shown for terpenoids pathways of angiosperms (Ramsay et al. 2009).

The most straightforward method for the estimation of heritability is parent-offspring regression (Lynch and Walsh, 1998). We have previously introduced a chain design to estimate the heritability of gene expression using single parent-offspring regression with microarrays (Albouyeh and Ritland, 2009). Here we utilize this experimental design to compare gene expression of maternal parents and offspring of Interior spruce (*Picea glauca* x *engelmannii*).

In this study, our aims were to: (a) document trends in the heritability of the expression from upstream to the downstream of terpenoid pathways; (b) test the difference in the heritability of the expression when partitioning pathways into segments; (c) associate observed patterns of heritability with patterns of adaptive evolution in spruce.

5.2 Methods

5.2.1 Sampling

A total of 30 maternal parents and offsprings (15 pairs) of Interior spruce were sampled. The open pollinated progenies were planted in 1972 at the Tree Improvement station of British Columbia Forest Service, Prince George, BC, Canada. The parents were located nearly 1 Km west (122 ° 42’ 43’’ W, 53° 45’ 41’’ N) of the progenies site,

established as grafts in1970. The outer stem tissues (bark and the attached phloem pooled from various aerial branches) were separated from the inner layers in the mid-morning hours of June 25, 2008, flash frozen in liquid Nitrogen, and transferred to separated containers.

5.2.2 Expression profiling and genetic analysis

Parents and offsprings were hybridized using the chain design of Albouyeh and Ritland (2009) for a total of 45 hybridizations (based on the minimum sample size requirement of 15 pairs). Array 350 kit (Genisphere, Hatfield, USA) was used for the hybridization of the samples cDNA to the third generation Treenomix platform (accession number GPL5423 of Gene Expression Omnibus; GEO). The slides were scanned with a ProscanArray scanner (PerkinElmer, Downers Grove, IL, USA) and the scanned TIF images were processed by ImaGene software version 6.0.1 (BioDiscovery Inc., El Segundo, CA, USA). The spot signal values were normalized between arrays using the variance stabilizing method (VSN, Huber et al. 2002). All expression profiles were processed in compliance with the Minimum Information for A Microarray Experiment (MIAME) and submitted to GEO under Series GSE22921.

The normalized Cy3/Cy5 channel ratios were used to compute the heritability values for the genes on the array. Given different mean heritability values and variances in different generations (parents and offsprings), narrow-sense heritability on the individual gene basis was obtained as:

$$h^2 = 2\left(\frac{C_{XY}}{V_X}\right) = 2\left(\frac{b-a}{c}\right)$$

Here h^2 is the narrow-sense heritability; V_X is the variance of the expression levels of all parents; C_{XY} is the covariance of the expression levels of all parents and offsprings; a, b, and

c denote the expression levels when offsprings and own parents, parents and unrelated offsprings, and parents and parents are hybridized respectively.

5.2.3 Annotation and classification of genes

Contiguous sequences (contigs) were formed from the array cDNA sequences and the sequences of an in-house database of spruce ESTs (Spruce V8; http://treenomix2.forestry.ubc.ca/). This strategy enhances the power of the Basic Local Alignment Search Tool (BLAST) because of the increased length of the queries. To assess the consistency of annotations, a combinatory BLAST strategy was devised; a translated nucleotide query (BLASTX) against protein database of the Arabidopsis Information Resource (TAIR version 9), and a translated nucleotide query (TBLASTX) against non-redundant (NR) translated nucleotides at National Center for Biotechnology Information (NCBI, downloaded August, 2010).

We defined a “segment” as a sampling entity that encompasses genes representative of a distinctive division in the terpenoid pathways. Six segments were considered from upstream to the downstream of the terpenoids pathways using KEGG PATHWAY (2010) as a template (Fig. 5.1). The assignment of genes to each segment was done based on the annotations of the array elements after an exhaustive screen for the enzymes specific to each segment (Appendix B). The differences in the average heritability of gene expression of the segments were then tested on the pair-wise basis using Wilcoxon signed rank test.

Gene Ontology (GO) terms were compared for a total of 94 terpenoid genes. Due to the limited number of GO terms (Appendix B) obtained for the “Cellular component” and “Function” categories, these categories were excluded from the comparisons. Among various term of the “Biological process” category, “Response to stress” was considered the most meaningful to assess the impact of terpenoid segments in plant defense.

5.2.4 Analysis of the divergence of the expression of the segments

We previously profiled gene expression from bark and the attached phloem tissues across several spruce species (series accession GSE18374; Albouyeh et al. 2010). This microarray data includes cross-species hybridizations of three biological replicates from five species of spruce; *P. abies*, *P. glauca*, *P. jezoensis*, *P. mariana*, *P. omorika* (a total of 15 hybridizations).

To test the divergence of the expression for each terpenoid pathway segment among species, the array signal intensities were normalized within as well as between arrays using the VSN method. The Cy3 and Cy5 expression intensities were averaged to arrive at a matrix of gene x individual expression intensities. The expression intensities were then analyzed by fitting a mixed model of the form:

$$y_{gki} = \mu + \alpha_g + \lambda_k + \gamma_{ki} + \varepsilon_{gki}$$

In this model, y_{gki} is the normalized intensities of gene expression for n genes in each segment (1,…,n), μ is the intercept, α_g is a fixed effect term that indexes g=(1,…, n) genes in the segment, λ_k is a fixed effect term that indexes k=(1,…, 5) species, γ_{ki} is the random effect of individual i of the species k, (biological effect) and ε_{gki} is the error. Restricted maximum likelihood (REML) was used for estimation.

5.3 Results

5.3.1 Overall patterns

Overall, there was not a sequential trend of increase in the heritabilities from upstream to all downstream segments of the pathways. The heritability of gene expression was relatively low for the Mevalonate pathway (MEV) in the upstream, as well as two downstream biosynthetic pathways of steroids (STRL) and carotenoids (CAR). However, there was an increasing trend from the Mevalonate (MEV) and 2-*C*-methyl-D-erythritol 4-

phosphate/1-deoxy-D-xylulose 5-phosphate (MEP) pathways (upstream) to the downstream, peaking at terpene synthases (TPS). This finding suggested the likelihood of capturing a sequential biological process with relevance in fitness; especially highlighting the peak of the trend. To further investigate the adaptive relevance of these pathways points where an increasing trend in heritability was observed, we set the difference of heritability upstream versus downstream as an hypothesis and used the divergence of gene expression and impact on response to stress as complementary analyses.

5.3.2 Segmental differences

The results of the statistical analysis showed that average heritability of the expression was different across the segments; in particular for the TPS (Fig. 5.2b; Table 5.1). On this basis, TPS had significantly higher average heritability than other segments, except for MEP and BKBNE. The BKBNE segment had significantly higher average heritability than MEV and CAR which had the lowest average heritability. The average heritability for TPS (0.77 ± 0.16 S.E.) and BKBNE (0.62 ± 0.33 S.E.) were substantial (> 50 %).

Analysis of the divergence of the expression of segments again highlighted TPS (Table 5.2). The results of the tests for the divergence of expression showed that gene expression of five terpenoid segments was conserved among species of spruce. For TPS, gene expression had significantly ($P > 0.05$) diverged among species.

The analysis of the association of Response to stress GO term with the segments showed that TPS had the highest impact on response to stress. This biological process was to a lesser extent impacted by MEV, MEP, BKBNE, and STRL, and had not been impacted by CAR (Fig. 5.3).

5.4 Discussion

The presence of extensive heritable variation for secondary metabolism indicates important physiological and ecological functions that impact fitness. Heritability studies in plants have shown that characters closely related to fitness tend to have high heritabilities (Geber and Griffen, 2003). However, how heritability, gene expression and secondary metabolism interact is not addressed extensively in the previous studies.

Here, the trend in the heritability of the expression of terpenoids was not monotonously increasing from upstream to all downstream points of the pathways. We found an increasing trend from upstream to the downstream of the pathways, peaking at a segment encompassing terpene synthase (TPS) family of genes. This finding suggested capturing a sequential process likely to be explained by an underlying adaptive functional property; especially at the peak of the trend.

Our partitioning of pathways into "segments" in particular enhanced explanation of the underlying functional property for terpene synthase gene family**:**

First, it provided units for testing the divergence of expression on different pathway points. Divergence of gene expression among species can be used as a surrogate to reflect the diversification of the expression of pathway points that is likely due to the act of diversification selection (Albouyeh et al. in prep; Chapter 3). Therefore, our functional inferences were comprehended by much of the evolutionary theory to explain the adaptive relevance of the trait under study (gene expression). Second, working with the average heritabilities of the segments allowed comparing stronger patterns which, despite the presence of large variation within each segment, revealed functional differences. Third, analysis of the "Biological process" terms on the segment-by-segment basis portrayed a meaningful sketch of the pathways function.

5.4.1 Relation to adaptive evolution

We considered high degree of heritability, diversification of gene expression and response to stress, as indicators of adaptation in plant defense. An example of such combination was seen here for TPS gene family. Although response to stress might occur as a result of many different environmental conditions, it can be seen as a substantial investment in the processes of defense response of conifers. This is in line with the large body of literature regarding the involvement of terpene synthase genes in the mechanism of plant defense, as well as the inherent diversification response of the bark of conifers (reviewed in Franceschi et al. 2005; Keeling and Bohlmann, 2006b).

On the other hand, CAR had a low impact on response to stress which matched by its relatively low heritability. Although, a few studies in the literature (e.g. Adams and Demmigadams, 1994; Bauer et al. 2000) have discussed the role of carotenoids as a secondary photosynthetic pigments in conifers, the literature on the involvement of this class of terpenoids in adaptive processes of conifers is largely silent.

The biosynthesis of campestrol and resulting brassinosteroids have roles in plant defenses (Dangl and Jones, 2001).However, it was not possible to elaborate on relatively lower impact of these on plant fitness (based on their lower average heritability and low impact on response to stress).

In spite of their interpretive value, there are important considerations in using GO terms. Often, there is high inconsistency in the number of GO terms obtained for different genes that can vary from none to several. To circumvent this, we converted the number of terms obtained for each gene to a binary state for the subsequent analyses (Appendix B). The number of GO terms for each gene is partly a function of the number of studies done for that

gene, but can also be an indication for pleiotropy, or the extent to which different genes can impact biological processes. Analysis of binary data is likely to mask such impact.

5.4.2 Heritability of the expression of TPS

The heritability of the expression of terpene synthase gene family is substantial (i.e. above 50). Typically one might be interested to see how the heritability estimates for the expression of these genes are comparable to the other organization levels such as proteome or metabolome. The heritability estimates are specific to the populations in space and time, and for the population under study, data from other organizational levels have not been collected in parallel. However, based on the existing heritability estimates from metabolite profiles of monoterpene in Pinaceae (Squillace, 1971; Rockman, 1973), and the evidence presented here from the level of gene expression, this family of genes is of high adaptive relevance in the population of study.

In our report here we dealt with the heritability of terpene synthase as a family of genes which encompasses various classes (mono-, sesqui-, and di-terpene synthases). In contrast, existing heritability estimates from terpene metabolits of the closest species, pines, belong to single products such as Pinene, Limonene, or Myrcine.which might not provide grounds for direct comparisons. However, previous molecular and biochemical characterization of selected conifer defense systems support a model of multigenic defense (Ralph et al. 2006). Therefore, from the plant defense perspective, integrating the co-expression of these genes better explains outcome of the aggregate phenotype of plant defense. Therefore, considering different members of this family of genes as a single segment and using their average heritability to explain tree fitness deemed appropriate.

5.5 Conclusion

We provided the first estimates for the heritability of the expression of terpenoids in conifers. Our results highlighted terpene synthase (TPS) family of genes for which the heritability of expression was substantial.

In respect to the diversification of expression and impact on response to stress, we concluded that the expression of TPS gene family is closely connected with the adaptation of spruce trees as a possible defense trait. This finding as well supports the importance of bark as a defense barrier against the co-evolving pests.

Table 5. 1 Results of Wilcoxon tests of average heritability.

Wilcoxon's test for the differences of average heritability of the pathway segments on a pair-wise basis. Values include test statistic *Z* , and the asymptotic significance in parenthesis (two-tailed). Segment abbreviations follow Fig.1.

Segment	MEV	MEP	BKBNE	TPS	STRL	CAR
MEV	-	- 1.33 [a]	- 2.16 [a]	- 3.01 [a]	- 1.21 [a]	- 0.73 [a]
		(0.18)	(<0.05) *	(< 0.01)**	(0.23)	(0.46)
MEP		-	- 0.98 [a]	- 1.88 [a]	- 1.07 [b]	- 2.2 [b]
			(0.33)	(0.06)	(0.29)	(<0.05)*
BKBNE			-	- 1.7 [a]	- 1.14 [b]	- 2.2 [b]
				(0.09)	(0.26)	(<0.05)*
TPS				-	- 2.7 [b]	- 2.2 [b]
					(< 0.01)**	(0.03)*
STRL					-	- 1.36 [b]
						(0.17)

a Based on negative ranks

b Based on positive ranks

* Significant at $\alpha < 0.05$

** Significant at $\alpha < 0.01$

Table 5. 2 Divergence of terpenoids gene expression.

Results of the tests for the divergence of the expression of various pathway segments among five species of spruce. Segment abbreviations follow Fig. 5.1.

Segment	*F* value	*P* value
	(Num DF: 4, Den DF: 10)	
MEV	0.4	0.81
MEP	1.36	0.32
BKBNE	2.23	0.14
TPS	4.83	< 0.05*
STRL	0.62	0.66
CAR	0.72	0.6

NumDF, Degrees of freedom of numerator

DenDF, Degrees of freedom of denominator

* Significant at $\alpha < 0.05$

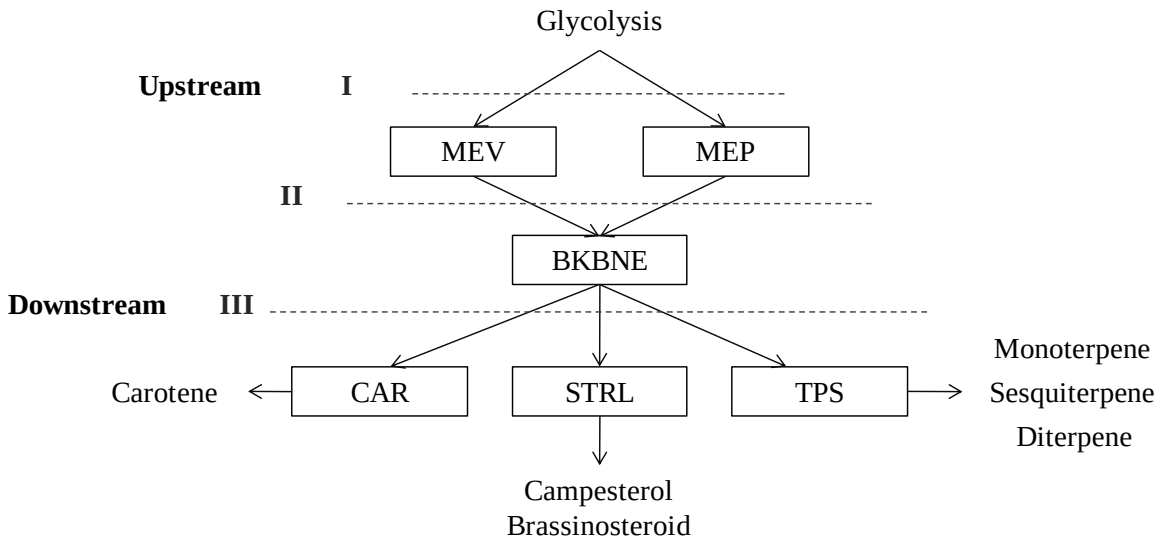


Figure 5. 1 Simplified representation of the terpenoid segments.

The segments are in three progressive steps (I, II, III) from upstream to the downstream of the terpenoid pathways. The segments shown in boxes encompass genes encoding enzymes of Mevalonate (MEV) and 2-*C*-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate (MEP)pathways to the backbone (BKBNE) including prenyl synthases/transferases; Terpene synthases (TPS), entry points to the biosynthesis of various terpenes; Sterol (STRL) including several enzymes of the steroids pathways leading to Campesterol and Brassinosteroid; and several enzymes from the carotenoids (CAR) pathway that lead to the biosynthesis of carotenes.

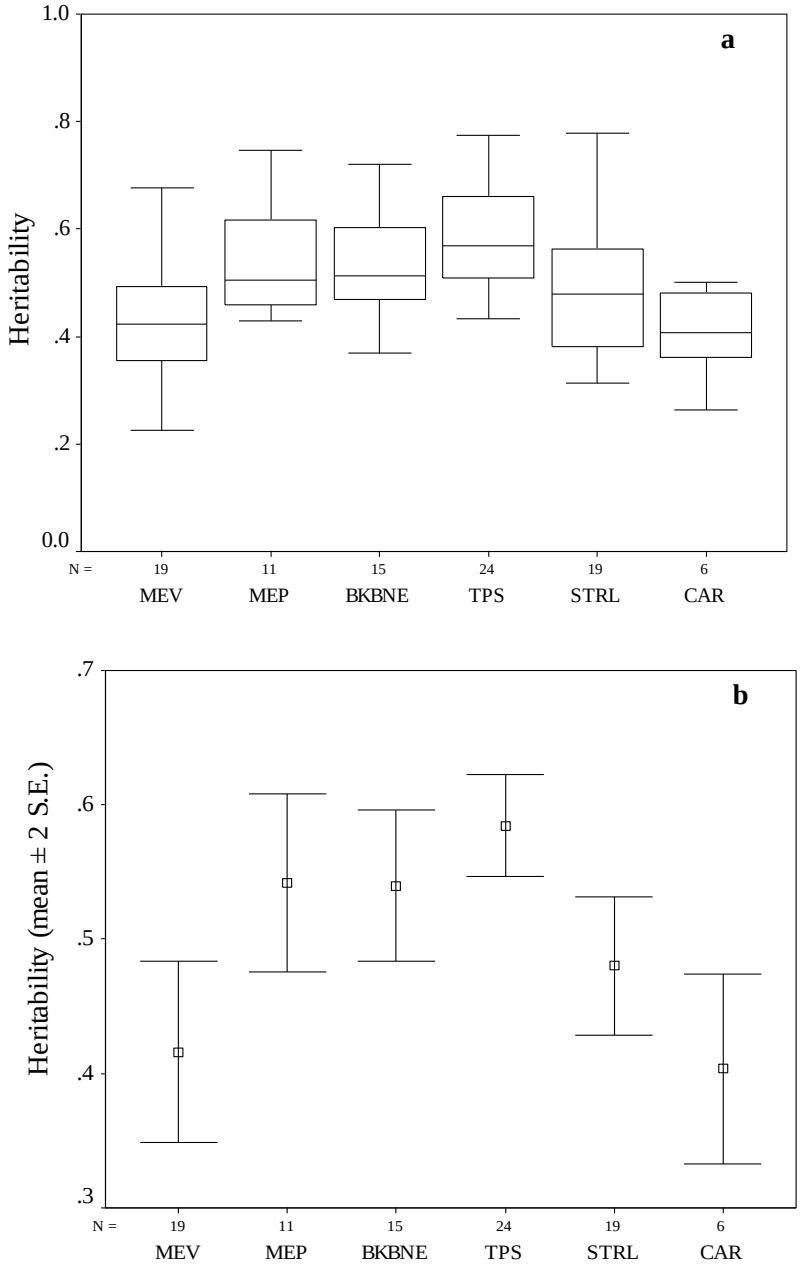


Figure 5. 2 Mean and median of the heritability of the expression of segments**.**

Box plots showing median and range of the heritability values in each segment (a); and differences in the average heritability of the segments (b). Segment abbreviations follow Fig.5.1 (re-scaled heritability values). N, number of ESTs; S.E. Standard error of the mean.

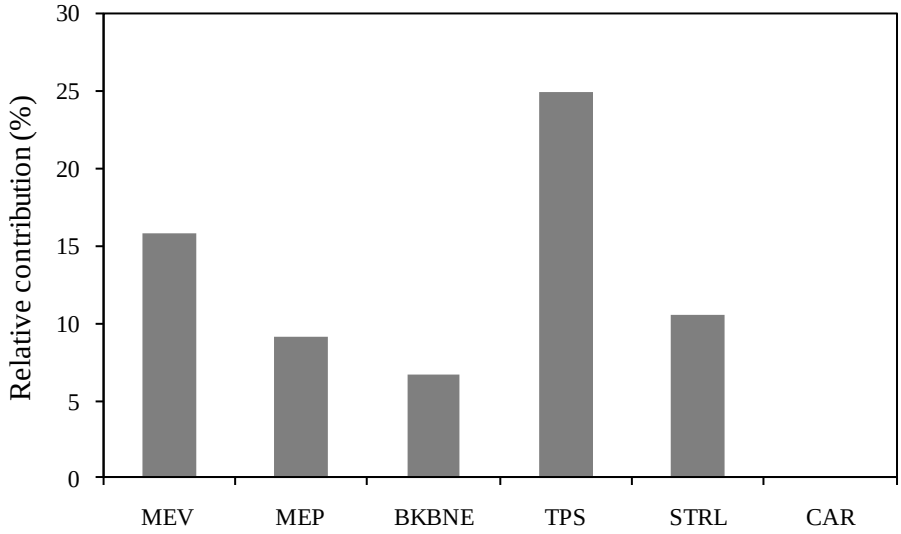


Figure 5. 3 Impact of various pathway segments in response to stress.

The relative contribution is calculated as (P_i/P_s) x100 where P_i is the abundance of genes with response to stress GO term in a given segment, and P_s is the total number of genes in that segment.

6. Conclusion

The work presented here can be identified, at least, by five prominent features. These include: (1) cross-species expression profiling; (2) tissue-related gene expression; (3) expression of the specialized biochemical pathways; (4) heritability of gene expression; and (5) the evolution of gene expression. The overlap of these features in conifers defined areas for which previous research was lacking.

To this end, my dissertation has provided the first generalized snapshot of the evolutionary structure of phenolic and terpenoid biochemical pathways at the level of the transcriptome in conifers. I employed a "phylogenomic" method, which involves several species ideally of nearly equal relationship, which the spruce genus offers. My phylogenomic analyses demonstrated that transcriptome does further the hypothesis of (Franceschi et al. 2005), *visa vi* that: outer stem (bark) tissue sources form a defense barrier against the co-evolution of pests in conifers. It also showed that gene expression data have the resolution to differentiate between inner- and outer -stem layers in spruce.

To infer the interaction between gene expression and its role in plant defense and fitness, I combined evolutionary theory and bioinformatics strategies. Using this framework, I found that the functional relevance of genes in specialized metabolism is best explained at the level of gene families. This implies a cumulative effect of gene expression in the outcome of defense traits. It is in line with the polygenic view of defense response in conifers (Ralph et al. 2006).

Chapter 5 demonstrated a clear example of how the combination of bioinformatics and the components of the evolutionary theory can specify a quantitative trait in the defense response. In this chapter, the average expression for the members of terpene synthase family achieved a substantial heritability.

Below I remark on some aspects of my overall findings related to the analysis of the expression of phenolics and terpenoids pathways.

6.1 Evolutionary structure of phenolics and terpenoids pathways at the level of transcriptome

The study of the evolution of gene expression for the biochemical pathways of phenolics and terpenoids demonstrated both unsystematic diversification and heritability trends. Contrasting to these, studies in theory (e.g., Waxman and Peck, 1998; Otto, 2004) suggest pleiotropic effects where there is higher connections between the genes (i.e. upstream) and gradients of selective pressure from upstream to the downstream of the biochemical pathways. Holliday (2009) pointed to the limitations of empirical examples in terms of the number of genes studied. I emphasize on the need to examine empirical data from the pathways gene expression (level of transcriptome). My results, however, might be adjusted by the discovery of the yet unknown parts of the pathways.

In the outer stem, phenolic families appeared to be more diversified than terpenoids. My reasons for this are based on differential expression as well as bioinformatics. Between the bioinformatics attempts, less inconsistency observed in the annotations of terpenoids. For a higher proportion of phenolics, the annotation shifted between the attempts.

In the set of genes studied, phenolics pathways were largely represented by the genes involved in the specialized metabolism. Whereas, in the set of terpenoids studied primary and specialized metabolism were intertwined to the downstream of the pathways.

6.2 Specialized metabolism: the engine of tree defense

Population-level studies use measurable environmental gradients to test for the adaptive relevance of gene expression (e.g., Whitehead and Crawford, 2006; Holliday et al. 2008). However, pressure from the pests in the natural history is a latent factor. I focused on the diversification of two major pathways of phenolics and terpenoids because of the considerable literature (e.g. Franceschi et al. 2005; Keeling and Bohlmann, 2006b; Hamberger et al. 2007) relevant to their role in tree defenses. The diversification of the specialized metabolites of these pathways had been identified as a key in the co-evolution of trees with bark borer pests. However, the extent of diversification and the relative importance of the various categories of genes in the evolution coniferous defense response; especially at the level of gene expression was largely unknown.

I used the divergence of gene expression among species as a surrogate to the diversification of the pathways. Cross-species microarray profiling demonstrated the overall divergence of gene expression in the outer stem. It further identified those families of genes for which the expression was diversified.

The unsystematic trends in the heritability and diversification of the pathways, does not guarantee conjunction of the concepts for every diverged gene family. However, an example of such conjunction was provided for terpene synthase gene family, interpreted as a strong implication for relevance in plant fitness.

Overall, these results support the evolution of a diversification response in the outer stem, likely as the result of pressures posed by bark borer pests. The families of genes that were identified with diverged gene expression could serve as a starting point for functional analyses, as well to model networks of gene expression to reflect inherent diversification response of trees to the bark borer pests.

6.3 Limitations

The sets of genes used here to represent phenolics and terpenoids biochemical pathways were compiled using the annotation of the ESTs spotted on the microarray. However, these sets under-represent the total number of genes involved in phenolics and terpenoids pathways. First, microarray is a closed system (i.e. it is a subset of screened ESTs from a certain number of cDNA libraries). Second, as also shown in the analysis of digital expression data (chapter 2), a considerable proportion of the total genes were unknown, which might have functions in the terpenoids or phenolics pathways.

A likely consequence of this data limitation is incomplete representation of the structure of the pathways. For example, downstream phenolic pathways shunting from the flavonoid pathway (anthocynine, flavone and flavonol, and isoflavonoid biosynthesis) were truncated in my study. It is unclear that having fewer representatives in these pathways is due to the bias of the platform, or their lower occurrences in conifers.

Using ESTs of the array also limits evolutionary analyses because of the incomplete representation of the total length of the genes. As a result, the precision of annotations and classification of genes to their corresponding families, and transcript profiling (due to the chance of cross hybridization) are reduced. In addition, it lacks the resolution to differentiate between orthologous and paralogous gene copies. Therefore, genetic distances between the constructed families are not possible to be calculated precisely.

6.4 Future work

With the completion of the sequencing project of Norway spruce, conifer genomics continues to be an interesting field of research. A direct implication of having a sequenced reference genome to this work is higher accuracy in the annotation and classification of genes into their corresponding families. Moreover, with the definition of the coding sequences for

the ESTs analyzed here, the precision of the tests of selection at level of genomic DNA sequences (e.g. analysis of DNA substitution rates) will be increased. These will lead to the improvement of the functional assessment of genes and the joint evolutionary analyses of genome and transcriptome. An intriguing related aspect is the identification of pseudogenes and their related evolutionary analyses, such that is presented by Khaitovich et al. (2006) for expressed pseudogenes.

Future work could incorporate the controls of the expression on the pathways; especially on identification of the parts and the connections involved in the induced defenses. Relevant to this, network thinking has become incorporated in system biology. Integrating metabolomic and proteomics can shed light on the nature of the connections and the evolutionary properties of the networks.

References

Adams MD, Kelley JM, Gocayne JD, et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651-1656

Adams WW, Demmigadams B (1994) Carotenoid composition and down-regulation of photosystem-II in 3 conifer species during winter. Physiol Plantarum 92: 451-458

Aizawa M, Yoshimaruth H, Saito H, Katsuki T, Kawahara T, Kitamura K, Shi F, Kaji M (2007) Phylogeography of a northeast Asian spruce, *Picea jezoensis*, inferred from genetic variation observed in organelle DNA markers. Mol Ecol 16: 3393-3405

Albouyeh R, Farzaneh N, Bohlmann J, Ritland K (2010) Multivariate analysis of digital gene expression profiles identifies a xylem signature of the vascular tissue of white spruce (*Picea glauca*). Tree Genet Genomes 6: 601-611

Albouyeh R, Ritland K (2009) Estimating Heritability of Gene Expression Using Parent-Offspring Regression with 2-Channel Microarrays. J Hered 100: 114-118

Allona I, Quinn M, Shoop E, et al (1998) Analysis of xylem formation in pine by cDNA sequencing. Proc Nat Acad Sci USA 95: 9693-9698

Bar-Or C, Czosnek H, Koltai H (2007) Cross-species microarray hybridizations: a developing tool for studying diversity. Trend Genet 23:200-207

Bauer H, Plattner K, Volgger W (2000) Photosynthesis in Norway spruce seedlings infected by needle rust *Chrysomyxa rhodedendri*. Tree Physiol 20: 211-216

Benderoth M, Textor S, Windsor AJ, Mitchell-Olds T, Gershenzon J, Kroymann J (2006) Positive selection driving diversification in plant secondary metabolism. Proc Natl Acad Sci USA 103: 9118-9123

Bernards MA, Bastrup-Spohr L (2008) Phenylpropanoid metabolism induced by wounding and insect herbivory. In: Schaller A, editor. Induced Plant Resistance to Herbivory. Springer, Berlin

Bohlmann J, Keeling CI (2008) Terpenoid biomaterials. Plant J 54: 656-669

Bouille M, Bousquet J (2005) Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer Picea (Pinaceae): Implications for the long-term maintenance of genetic diversity in trees. Am J Bot 92: 63-73

Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296: 752-755

Brenner S, Johnson M, Bridgham J, et al (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotech 18:630-634

Britten RJ, Davidson EH (1969) Gene regulation for higher cells - a theory. Science 165:349-357

Bueno Filho JSS, Gilmour SG, Rosa MJG (2006) Design of microarray experiments for genetical genomics studies. Genetics 174: 945-957

Bystrykh L, Weersing E, Dontje B, et al (2005) Uncovering regulatory pathways affecting hematopoietic stem cell function using “genetical genomics”. Nat Genet 37: 225-232

Campbell CS, Wright WA, Cox M, Vining TF, Smoot Major C, Arsenault MP (2005) Nuclear ribosomal DNA internal transcribed spacer 1 (ITS1) in Picea (Pinaceae): sequence divergence and structure. Mol Phylogenet Evol 35: 165-185

Chesler EJ, Lu L, Wang J, Williams RW, Manly KF (2004) WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. Nat Neurosci 7: 485-486

da Silva FG, Iandolino A, Al-Kayal F, et al (2005) Characterizing the grape transcriptome. Analysis of expressed sequence tags from multiple Vitis species and development of a compendium of gene expression during berry development. Plant Physiol 139: 574-597

Dangl JL, Jones JDG (2001) Plant pathogens and integrated defense responses to infection. Nature 411: 826-831

Darvasi A, (2003) Gene expression meets genetics. Nature 422: 269-271

Diatchenko L, Lau Y-FC, Campbell AP, et al (1996) Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc Nat Acad Sci, USA 93: 6025-6030

Dixon RA, Achnine L, Kota P, Liu CJ, Reddy MSS, Wang LJ (2002) The phenylpropanoid pathway and plant defense - a genomics perspective. Mol Plant Pathol 3:371-390

Ehlting J, Mattheus N, Aeschilman DS, et al (2005) Global transcript profiling of primary stems from Arabidopsis thaliana identifies candidate genes for missing links in lignin biosynthesis and transcriptional regulators of fiber differentiation. Plant J 42: 618-640

Ewing RM**,** Ben Kahla A, Poirot O, et al (1999) Large-Scale Statistical Analyses of Rice ESTs Reveal Correlated Patterns of Gene Expression. Genome Res 9: 950-959

Falconer DS (1989) Introduction to quantitative genetics, 3rd edn. Longman, Scientific & Technical, New York

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Pearson Education Ltd, Harlow

Farjon A (1990) Pinaceae: drawings and descriptions of the genera *Abies*, *Cedrus*, *Pseudolarix*, *Keteleeria*, *Nothotsuga*, *Tsuga*, *Cathaya*, *Pseudotsuga*, *Larix* and *Picea*.

Koeltz Scientific Books, Königstein

Felsenstein J (2006) PHYLIP: the PHYlogeny Inference Package version 3.66 [updated 2008 September 8]. Available from: http://evolution.genetics.washington.edu/phylip.html

Felsenstein J (1985) Phylogenies and the comparative method. Am Naturalist 125: 1-15

Foucart C, Paux E, Ladouce N et al (2006) Transcript profiling of a xylem vs phloem cDNA subtractive library identifies new genes expressed during xylogenesis in Eucalyptus. New Phytol 170: 739-752

Franceschi VR, Krokene P, Christiansen E, Krekling T (2005) Anatomical and chemical defenses of conifer bark against bark beetles and other pests. New Phytol 167: 353-375

Friedmann M, Ralph SG, Aeschliman D, et al (2007) Microarray gene expression profiling of developmental transitions in Sitka spruce (*Picea sitchensis*) apical shoot. J Exp Bot 58: 593-614

Geber MA, Griffen LR (2003) Inheritance and natural selection of functional traits. Int J Plant Sci 164: S21-42

Gernandt DS, Lopez GG, Garcia, SO, Liston A (2005) Phylogeny and classification of Pinus. Taxon 54: 29-42

Gibson G, Weir B (2005) The quantitative genetics of transcription. Trends Genet 21: 616-623

Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP (2006) Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature 440:242-245

Goodwillie C, Ritland C, Ritland K (2006) The genetic basis of floral traits associated with mating system evolution in Leptosiphon (Polemoniaceae): An analysis of quantitative trait loci. Evolution 60:491-504

Gorantla M, Babu PR, Lachagari VBR, et al (2007) Identification of stress-responsive genes in an indica rice (*Oryza sativa* L.) using ESTs generated from drought-stressed seedlings. J Exp Bot 58: 253-265

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95-98

Hamberger B, Ellis M, Friedmann M, Souza CDA, Barbazuk B, Douglas CJ (2007) Genome-wide analyses of phenylpropanoid-related genes in *Populus trichocarpa*, *Arabidopsis thaliana*, and *Oryza sativa*: Populus lignin toolbox and conservation and diversification of angiosperm gene families. Can J Bot 85:1182-1201

Hamberger B, Bohlmann J (2006) Cytochrome P450 mono-oxygenases in conifer genomes: discovery of members of the terpenoid oxygenase superfamily in spruce and pine. Biochem Soc Transact 34: 1209-1214

Hene L, Sreenu VB, Voung MT et al (2007) Deep analysis of cellular transcriptomes LongSAGE versus classic MPSS. BMC Genomics 8:333

Holliday JA (2009) Genomics of adaptation to local climates in Sitka Spruce (*Picea sitchensis*). Ph.D. thesis, University of British Columbia

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 Phytol 178:103-122

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9: 868-877

Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18:S96-104

Jansen RC, Nap J (2001) Genetical genomics: the added value from segregation. Trends Genet 17: 388-391

Jansson S, Douglas KJ (2007) Populus: a model system for plant biology. Annu Rev Plant Biol 58: 435-458

Jordan IK, Marino-Ramirez L, Koonin EV (2005) Evolutionary significance of gene expression divergence. Gene 345:119-126

Keeling C, Bohlmann J (2006a) Diterpene resin acids in conifers. Phytochemistry 67: 2415-2423

Keeling C, Bohlmann J (2006b) Genes, enzymes and chemicals of terpenoid diversity in the constitutive and induced defense of conifers against insects and pathogens. New Phytol 170: 657-675

Keeling CI, Weisshaar S, Lin RPC, Bohlmann, J (2008) Functional plasticity of paralogous diterpene synthases involved in conifer defense. Proc Natl Acad USA 105: 1085-1090

KEGG PATHWAY Database [Internet] (2010) Kanehisa Laboratories. [accessed 2010 Nov 9]. Available from: http://www.genome.jp/kegg/pathway.html

Kellogg EA, Shaffer HB (1993) Model organisms in evolutionary studies. Syst Biol 42: 409-414

Khaitovich P, Weiss G, Lachmann M, et al (2004) A neutral model of transcriptome evolution. Plos Biology 2:682-689

Khaitovich P, Enard, W, Lachmann M, Paabo S (2006) Evolution of primate gene expression. Nat Rev Genet 7: 693-702

King MC, Wilson AC (1975) Evolution at 2 levels in humans and chimpanzees. Science 188:107-116

Kirst M, Johnson AF, Baucom C, Ulrich E, et al (2003) Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana.* Proc Nat Acad Sci USA 100:7383-7388

Kuelheim C, Yeoh SH, Maintz J, Foley WJ, Moran GF (2009) Comparative SNP diversity among four eucalyptus species for genes from secondary metabolite biosynthetic pathways. BMC Genomics 10:452

Lemos B, Meiklejohn CD, Caceres M, Hartl DL (2005) Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution 59:126-137

Lev-Yadun S, Gould KS (2008) Role of anthocyanins in plant defense. In: Gould KS, Davies KM, Winefield C (ed) Life's colorful solutions: the biosynthesis, functions, and applications of anthocyanins. Springer, Berlin

Li LY, Cheng H, Gai JY, et al. (2007) Genome-wide identification and characterization of putative cytochrome P450 genes in the model legume *Medicago truncatula*. Planta 226: 109-123

Li J, Burmeister M (2005) Genetical genomics: combining genetics with gene expression analysis. Hum Mol Genet 14: R163-R169

Lockhart DJ, Dong HL, Byrne MT et al (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotech 14: 1675-1680

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Assoc. Sunderland

Martin D, Tholl D, Gershenzon J, Bohlmann J (2002) Methyl jasmonate induces traumatic resin ducts, terpenoid resin biosynthesis, and terpenoid accumulation in developing xylem of Norway spruce stems. Plant Physiol 129: 1003-1018

Miller B, Madilao LL, Ralph S, Bohlmann J (2005) Insect-induced conifer defense: white pine weevil and methyl jasmonate induce traumatic resinosis, de novo formed volatile emissions, and accumulation of terpenoid synthase and putative octadecanoid pathway transcripts in Sitka spruce. Plant Physiol 137: 369-382

Monks SA, Leonardson A, Zhu H et al (2004) Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 75, 1094-1105

Nasri N, Bojovic S, Vendramin GG, Fady B (2008) Population genetic structure of the relict Serbian spruce, *Picea omorika*, inferred from plastid DNA. Plant Syst Evol 271: 1-7

Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. Curr Opin Plant Biol 11: 149-155

Nei M, Li WH (1979) Mathematical-model for studying genetic-variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269-5273

Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM (2004) Common pattern of evolution of gene expression level and protein sequence in drosophila. Mol Biol Evol 21:1308-1317

Omid A, Keilin T, Glass A, et al (2007) Characterization of phloem-sap transcription profile in melon plants. J Exp Bot 58: 3645-3656

O'Reilly-Wapstra JM, McArthur C, Potts BM (2004) Linking plant genotype, plant defensive chemistry and mammal browsing in a eucalyptus species. Funct Ecol 18:677-684

Otto, SP (2004) Two steps forward, one step back: the pleiotropic effects of favoured alleles. Proc Biol Sci 271: 705-714

Paiva JAP, Garces M, Alves A, et al (2008) Molecular and phenotypic profiling from the base to the crown in maritime pine wood-forming tissue. New Phytol 178: 283-301

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

Pavy N, Paule C, Parsons L (2005b) Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters. BMC Genomic 6:144

Peter G, Neale D (2004) Molecular basis for the evolution of xylem lignifications. Curr Opin Plant Biol 7: 737-742

Philippe H, Blanchette M (2007) Overview of the first phylogenomics conference. BMC Evol Biol 7 (Suppl 1): S1

Pittelkow Y, Wilson S (2005) Use of Principal Component Analysis and the GE-Biplot for the graphical exploration of gene expression data. Biometrics 61: 630-634

Pittelkow YR, Wilson S (2003) Visualisation of Gene Expression Data – the GE-biplot, the Chip-plot and the Gene-plot. Stat Appl Genet and Mol Biol 2: 1-17

Ralph SG, Chun HJE, Kolosova N et al (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 9:484

Ralph SG, Jancsik S, Bohlmann J (2007) Dirigent proteins in conifer defense II: Extended gene discovery, phylogeny, and constitutive and stress-induced gene expression in spruce (Picea spp.) Phytochemistry 68: 1975-1991

Ralph SG, Yueh H, Friedmann M, et al (2006) Conifer defense against insects: microarray gene expression profiling of Sitka spruce (*Picea sitchensis*) induced by mechanical wounding or feeding by spruce budworms (*Choristoneura occidentalis*) or white pine weevils (*Pissodes strobi*) reveals large-scale changes of the host transcriptome. Plant Cell Environ 29: 1545-1570

Ramos-Onsins SE, Puerma E, Balana-Alcaide D, Salguero D, Aguade M (2008) Multilocus analysis of variation using a large empirical data set: phenylpropanoid pathway genes in *Arabidopsis thaliana*. Mol Ecol 17:1211-1223

Ramsay H, Rieseberg LH, Ritland K (2009) The Correlation of Evolutionary Rate with Pathway Position in Plant Terpenoid Biosynthesis. Mol Biol Evol 26: 1045-1053

Ran JH, Wei, XX, Wang XQ (2006) Molecular phylogeny and biogeography of Picea (Pinaceae): implications for phylogeographical studies using cytoplasmic haplotypes. Mol Phylogenet Evol 41: 405-419

Rifkin SA, Kim J, White KP (2003) Evolution of gene expression in the *Drosophila melanogaster* subgroups. Nat Genet 33: 138-144

Ritland K, Ralph S, Lippert D, Rungis D, Bohlmann J (2005) New directions for conifer genomics. In Landscapes, Genomics and Transgenic Conifer Forests. Claire G. Williams(ed), Kluwer-Springer Press, Netherlands

Rockman MV, Kruglyak L (2006) Genetics of global gene expression. Nat Rev Genet 7: 862-872

Rockwood DL (1973) Variation in Monoterpene Composition of 2 Oleoresin Systems of Loblolly-Pine. For Sci 19: 147-153

Roderic DM (2001) TreeView: Tree drawing software for Apple Macintosh and Windows version 1.66 [updated 2001 July 31]. Available from: http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

Rungis D, Berube Y, Zhang J et al (2004) Robust simple sequence repeat markers for spruce (Picea spp.) from expressed sequence tags. Theor Appl Genet 109:1283-1294

Saha S, Sparks AB, Rago C, et al (2002) Using the transcriptome to annotate the genome. Nat Biotechnol 20:508-512

Schadt EE, Monks SA, Drake TA et al (2003) Genetics of gene expression surveyed in maize, mouse, and man. Nature 422: 297-302

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

Sibout R, Eudes A, Mouille G, et al (2005) Cinnamyl alcohol dehydrogenase-C and -D are the primary genes involved in lignin biosynthesis in the floral stem of Arabidopsis. Plant Cell 17: 2059-2076

Squillac AE (1971) Inheritance of monoterpene composition in cortical oleoresin of Slash pine. For Sci 17: 381-387

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

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100:9440-9445

ter Braak C, Smilauer, P (2003) CANOCO 4.5 Reference manual. User's guide to CANOCO for Windows. Centre for Biometry, Wageningen

ter Braak CJF (1987) CANOCO- A FORTRAN program for canonical community ordination by Correspondence Analysis, Principal Components Analysis and Redundancy Analysis (version 2.1) Agricultural Mathematics Group, Wageningen

Tollefsrud MM, Kissling R, Gugerli F et al (2008) Genetic consequences of glacial survival and postglacial colonization in Norway spruce: combined analysis of mitochondrial DNA and fossil pollen. Mol Ecol 17: 4134-4150

Udall JA, Swanson JM, Haller K et al (2006) A global assembly of cotton ESTs. Genome Res 16:441-450

Ukrainetz NK, Kang K-Y, Aitken SN, Stoehr M, Mansfield SD (2008a) Heritability, phenotypic and genetic correlations of coastal Douglas-fir (*Pseudotsuga menziesii*) wood quality traits. Can J For Res 38:1536-1546

Ukrainetz NK, Ritland K, Mansfield SD (2008b) Identification of quantitative trait loci for wood quality and growth across eight full-sib coastal Douglas-fir families. Tree Genet Genomics 4:159-170

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

Vos P, Hogers R, Bleeker M et al (1995) AFLP - a new technique for DNA-fingerprinting. Nucl Acids Res. 23:4407-4414

Vuylsteke M, van Eeuwijk F, Van Hummelen P, Kuiper M, Zabeau M (2005) Genetic analysis of variation in gene expression in Arabidopsis thaliana. Genetics 171: 1267-1275

Wagner A, Donaldson L, Kim H et al (2009) Suppression of 4-coumarate-CoA ligase in the coniferous gymnosperm *Pinus radiata*. Plant Physiol 149:370-383

Wanderley-Nogueira AC, Soares-Cavalcanti NM, Morais DAL, et al (2007) Abundance and diversity of resistance genes in the sugarcane transcriptome revealed by in silico analysis. Genet Mol Res 6: 866-889

Wang XQ, Tank DC, Sang T (2000) Phylogeny and divergence times in Pinaceae: evidence from three genomes. Mol Biol Evol 17:773-781

Waxman D, Peck JR (1998) Pleiotropy and the preservation of perfection. Science 279: 1210-1213

Whitehead A, Crawford DL (2005) Variation in tissue-specific gene expression among natural populations. Genome Biol. 6: R13

Whitehead A, Crawford DL (2006) Neutral and adaptive variation in gene expression. Proc Nat Acad Sci USA 103: 5425-5430

Wit E, Nobile A, Khanin R (2005) Near-optimal designs for dual channel microarray studies. Appl Statis 54: 817-830

Zhang Y, Brown G, Whetten R, et al (2003) An arabinogalactan protein associated with secondary cell wall formation in differentiating xylem of loblolly pine. Plant Mol Biol 52: 91-102

Zhang Y, Sederoff RR, Allona I (2000) Differential expression of genes encoding cell wall proteins in vascular tissues from vertical and bent loblolly pine trees. Tree Physiol 20:457-466

Zimmer A, Lang D, Richardt S, Frank W, Reski R, Rensing SA (2007) Dating the early evolution of plants: Detection and molecular clock analyses of orthologs. Mol Genet Genomics 278:393-402

Appendices

Appendix A Chapter 3 supplementary data

A.1 Screened array elements related to phenolics biosynthesis.

Table S. 1 List of the 332 array elements corresponding to 18 phenolic gene families. Expressed Sequence Tags (ESTs) are listed with the identifiers of their annotated best hits in blast searches against all plant data bases of National Center for Biotechnology Information (NCBI Viridiplantae) and The Arabidopsis Information Resource version 8 (TAIR 8). Family abbreviations follow Table 3.2.

Family	Array EST ID	Annotation (NCBI Viridiplantae)	Annotation (TAIR 8)
2OGFE	WS0037_D03	gb|AAA85365.1	AT3G21420.1
	WS0044_J09	gb|AAS21058.1	AT5G05600.1
	WS0048_J11	gb|ACB42271.1	AT4G22880.1
	WS0264_J18	gb|ACC66093.1	AT4G22880.1
	WS0045_K01	gb|ACC66093.1	AT4G22880.1
	WS0094_C12	ref|NP_196179.1	AT5G05600.1
	WS0048_M19	ref|NP_192787.1	AT4G10490.1
	WS00716_N13	ref|XP_002330269.1	AT3G19000.1
	WS0092_M15	emb|CAH67943.1	AT4G10490.1
	WS00926_P02	gb|AAA85365.1	AT3G21420.1
	WS01016_H12	ref|XP_002330269.1	AT3G19000.1
	WS0062_D15	ref|XP_002330269.1	AT3G19000.1
	WS00716_D02	ref|XP_002330269.1	AT3G19000.1
	WS0076_K21	ref|NP_566623.1	AT3G19000.1
	WS00918_N23	ref|NP_192788.1	AT4G10500.1
	WS00931_I06	ref|NP_192787.1	AT4G10490.1
	WS01021_D05	ref|NP_192787.1	AT4G10490.1
	WS01027_A20	ref|XP_002330269.1	AT3G19000.1
	WS01028_G09	ref|NP_181207.2	AT2G36690.1
	WS02613_J05	ref|NP_566685.1	AT3G21420.1
	WS0262_M22	ref|XP_002330269.1	AT3G19000.1
	WS0051_P20	gb|ACC66093.1	AT4G22880.1
	WS01029_N08	gb|AAA85365.1	AT2G38240.1
	WS0072_C18	ref|NP_001045319.1	AT3G21420.1

Table S. 1 continued

Family	Array EST ID	Annotation (NCBI Viridiplantae)	Annotation (TAIR 8)
2OGFE	WS01010_O03	ref|XP_002304499.1	AT4G23340.1
	WS00825_J22	ref|NP_192787.1	AT4G10490.1
	WS0076_O18	gb|ACG44904.1	AT3G21420.1
	WS00928_O22	ref|NP_192787.1	AT4G10490.1
	WS01017_M20	ref|XP_002304499.1	AT1G52820.1
	WS01013_G01	ref|XP_002324836.1	AT1G52820.1
	WS00929_H17	gb|ACC66093.1	AT5G05600.1
	WS00821_E12	ref|XP_002324836.1	AT4G23340.1
	WS00935_G13	ref|NP_001054534.1	AT3G11180.1
	WS00812_H04	dbj|BAG68574.1	AT4G10490.1
	WS00916_M15	gb|ACG44904.1	AT2G38240.1
	WS0063_B12	ref|NP_196179.1	AT5G05600.1
	WS00927_L21	ref|NP_187728.1	AT3G11180.1
	WS00912_H11	ref|NP_187728.1	AT3G11180.1
	WS0262_L21	gb|AAA85365.1	AT2G38240.1
	WS00946_D04	ref|NP_196179.1	AT5G05600.1
	WS00937_K24	ref|NP_181359.1	AT2G38240.1
	WS0108_P03	gb|AAA85365.1	AT3G21420.1
	WS01021_L05	gb|ACG44904.1	AT2G38240.1
	WS00918_P23	ref|NP_192787.1	AT4G10490.1
	WS0104_N20	emb|CAC14568.1	AT4G10490.1
	WS0042_I07	ref|NP_192787.1	AT4G10490.1
	WS00931_G15	ref|NP_192787.1	AT4G10490.1
	WS0082_L07	ref|NP_192788.1	AT4G10500.1
	WS01014_B09	gb|AAA85365.1	AT3G21420.1
	WS01024_K17	gb|ACG44904.1	AT3G21420.1
	WS00827_N24	ref|NP_192788.1	AT4G10500.1
	WS00930_E20	ref|XP_002324836.1	AT1G52800.1
	WS00940_O12	ref|NP_194065.3	AT4G23340.1
	WS01018_O04	ref|NP_192788.1	AT4G10500.1
	WS00923_F22	gb|ACC66093.1	AT4G22880.1
	WS00820_G05	gb|AAA85365.1	AT3G21420.1
4CL	WS00931_E22	sp|Q10S72.1	AT1G20510.1
	WS00930_N07	gb|AAB42383.1	AT3G21240.1
	WS00816_O23	sp|Q7F1X5.1	AT5G38120.1
	WS00928_I04	AT3G16910.1	AT3G16910.1
	WS01027_O10	AT3G16910.1	AT3G16910.1
	WS00926_I20	sp|Q9M0X9.1	AT4G05160.1
	WS00112_J15	gb|AAC97389.1	AT1G65060.1
	WS0097_M19	AT1G20510.1	AT1G20510.1
	WS01010_M10	ref|XP_002325815.1	AT1G65060.1

Table S. 1 continued

Family	Array EST ID	Annotation (NCBI Viridiplantae)	Annotation (TAIR 8)
4CL	WS0269_E03	sp|Q10S72.1	AT1G20510.1
	WS0043_P03	AT1G20510.1	AT1G20510.1
	WS00913_G23	ref|XP_002305282.1	AT4G05160.1
	WS00923_H15	ref|XP_002300662.1	AT1G20510.1
	WS00824_G20	ref|NP_190468.1	AT3G48990.1
C3H	WS00823_C02	gb|AAL47685.1	AT2G40890.1
	WS01017_C19	FJ483941.1	AT2G40890.1
	WS0099_J20	gb|AAU00415.1	AT2G40890.1
C4H	WS00931_C11	gb|AAW70021.1	AT2G30490.1
	WS00824_E05	gb|AAD23378.1	AT2G30490.1
CAD	WS01020_G05	gb|AAP68279.1	AT1G72680.1
	WS00823_O18	dbj|BAE48658.1	AT5G19440.1
	WS0078_K01	ref|XP_002309270.1	AT4G39330.1
	WS00821_H07	sp|Q9ZRF1.1	AT4G39330.1
	WS00712_E24	ref|XP_002322761.1	AT4G39330.1
	WS0261_G15	dbj|BAE48658.1	AT5G19440.1
	WS0074_B14	ref|XP_002309270.1	AT4G39330.1
	WS00938_N21	gb|AAQ55962.1	AT4G39330.1
	WS01017_J05	sp|Q9ZRF1.1	AT4G39330.1
	WS00932_K15	emb|CAD29291.1	AT4G37970.1
	WS01030_G23	ref|NP_001064283.1	AT4G37970.1
	WS00911_M22	gb|AAQ20892.1	AT4G34230.1
	WS00943_D20	gb|AAP68279.1	AT1G72680.1
CCR	WS0094_C01	emb|CAK18610.1	AT1G15950.1
	WS00921_J24	AT5G58490.1	AT5G58490.1
	WS00920_P12	ref|NP_194776.1	AT4G30470.1
	WS00732_L12	AT4G37680.1	AT4G37680.1
	WS0104_E12	ref|NP_001051706.1	AT5G58490.1
	WS0046_F13	AT5G58490.1	AT1G15950.1
	WS00111_I15	gb|ACE95172.1	AT1G15950.1
	WS0013_L15	ref|NP_001058104.1	AT2G33590.1
	WS00933_O12	gb|ACJ38670.1	AT1G15950.1
	WS01013_A02	gb|AAR23420.1	AT5G58490.1
	WS01016_F01	gb|AAR23420.1	AT5G58490.1
	WS0106_D11	ref|NP_001051706.1	AT5G58490.1
	WS0109_N01	gb|ACB45309.2	AT1G15950.1
	WS0264_B20	ref|NP_178345.1	AT2G02400.1
	WS0061_J17	ref|NP_565557.1	AT2G23910.1
	WS0078_N05	ref|XP_002303845.1	AT1G80820.1

Table S. 1 continued

Family	Array EST ID	Annotation (NCBI Viridiplantae)	Annotation (TAIR 8)
CCR	WS00924_B17	ref|NP_177021.1	AT1G15950.1
CHI	WS00927_F05	sp|A5ANT9.1	AT3G55120.1
	WS00720_J05	sp|A5ANT9.1	AT3G55120.1
	WS0047_E22	gb|ACG35950.1	AT3G63170.1
	WS00912_C22	ref|NP_001048296.1	AT2G26310.1
CHS	WS00925_G22	sp|P48408.1	AT5G13930.1
	WS0024_K18	gb|ABD24236.1	AT5G13930.1
	WS00915_A19	dbj|BAA94594.1	AT5G13930.1
	WS0016_K10	gb|ABD38613.1	AT5G13930.1
	WS0064_P09	gb|ABD24253.1	AT5G13930.1
	WS0062_C03	sp|Q9M5M0.1	AT5G13930.1
	WS0044_E23	dbj|BAA89680.1	AT5G13930.1
	WS00731_E22	gb|ABD24228.1	AT5G13930.1
	WS0043_C06	gb|ABD24236.1	AT5G13930.1
	WS0014_M21	dbj|BAA89680.1	AT5G13930.1
DIR	WS00914_H24	gb|ABD52124.1	AT1G64160.1
	WS01018_J08	gb|AAF25368.1	AT1G64160.1
	WS00911_I09	gb|ABR27728.1	AT1G64160.1
	WS0262_G24	gb|ABR27718.1	AT5G42510.1
	WS0078_C23	gb|ABR27723.1	AT5G42510.1
	WS01012_J06	gb|ABD52128.1	AT5G42510.1
	WS0262_J09	gb|ABR27723.1	AT1G22900.1
	WS0043_N14	gb|ABR27722.1	AT1G22900.1
	WS0094_C05	gb|ABR27721.1	AT5G42510.1
	WS0058_H22	gb|ABR27722.1	AT1G22900.1
	WS01012_K18	gb|ABD52122.1	AT2G21100.1
	WS0104_A04	gb|ABR27724.1	AT1G22900.1
	WS0262_G08	gb|ABD52118.1	AT1G65870.1
	WS0265_H11	gb|ABR27722.1	AT1G22900.1
	WS01011_J07	gb|ABD52130.1	AT1G64160.1
	WS0107_K16	gb|ABR27726.1	AT5G42510.1
	WS0064_L20	gb|ABD52119.1	AT1G64160.1
	WS00815_A07	gb|ABD52123.1	AT5G42500.1
	WS01012_K12	gb|ABR27726.1	AT5G42510.1
	WS00923_D20	gb|ABD52116.1	AT1G64160.1
	WS00924_E04	gb|ABD52120.1	AT1G65870.1
	WS00927_L10	gb|ABD52115.1	AT1G65870.1
	WS01031_M14	gb|ABD52123.1	AT2G21100.1
	WS01032_M02	gb|ABD52127.1	AT1G64160.1

Table S. 1 continued

Family	Array EST ID	Annotation (NCBI Viridiplantae)	Annotation (TAIR 8)
DIR	WS0063_M13	gb|ABR27725.1	AT1G22900.1
	WS0261_J16	gb|ABD52121.1	AT2G21100.1
	WS0262_I24	gb|ABR27719.1	AT5G42510.1
	WS0064_J21	gb|ABD52117.1	AT1G64160.1
	WS00825_K05	gb|ABD52129.1	AT1G64160.1
	WS0086_I04	gb|ABR27724.1	AT1G22900.1
F3H	WS0071_I18	dbj|BAE47004.1	AT5G07990.1
	WS00824_C19	gb|AAS91654.1	AT5G07990.1
	WS0092_D21	gb|AAU93347.1	AT3G51240.1
	WS0022_J09	sp|Q50EK4.1	AT5G07990.1
	WS0091_H05	ref|XP_002303241.1	AT5G07990.1
	WS0015_K19	ref|XP_002303241.1	AT5G07990.1
	WS01021_N08	sp|Q50EK4.1	AT5G07990.1
	WS0022_G09	sp|Q96418.1	AT5G07990.1
	WS00922_H05	gb|ACM89788.1	AT5G07990.1
	WS00931_D17	dbj|BAH22519.1	AT5G07990.1
	WS0262_B20	gb|AAU93347.1	AT3G51240.1
	WS00824_O01	gb|AAP88702.1	AT5G07990.1
	WS0045_J16	gb|ACM89789.1	AT5G07990.1
F5H-like	WS00924_I12	gb|ACM89789.1	AT4G36220.1
	WS01035_B16	gb|ACM89789.1	AT4G36220.1
	WS00934_G23	gb|ACM89788.1	AT4G36220.1
	WS00923_E07	sp|Q50EK4.1	AT4G36220.1
	WS00935_B19	gb|ACM89789.1	AT4G36220.1
	WS00937_F04	gb|ACM89789.1	AT4G36220.1
	WS0075_C23	gb|ACM89788.1	AT3G48270.1
	WS0047_J11	sp|Q50EK4.1	AT2G30770.1
	WS00715_F07	sp|Q50EK4.1	AT3G26230.1
GLYTR	WS00110_P02	dbj|BAD77944.1	AT5G65550.1
	WS00111_D08	dbj|BAD77944.1	AT5G65550.1
	WS0071_C13	sp|Q40287.1	AT3G50740.1
	WS0094_H09	sp|P0C7P7.1	AT1G05680.1
	WS0097_K18	sp|Q40287.1	AT2G36790.1
	WS0044_D10	sp|Q40287.1	AT1G01390.1
	WS0086_B18	dbj|BAF75890.1	AT4G34131.1
	WS0021_D24	dbj|BAA83484.1	AT2G36780.1
	WS00912_N05	sp|P0C7P7.1	AT1G05680.1
	WS00912_M21	gb|AAR06917.1	AT2G15490.1

Table S. 1 continued

Family	Array EST ID	Annotation (NCBI Viridiplantae)	Annotation (TAIR 8)
GLYTR	WS00931_K21	dbj|BAG31950.1	AT2G36800.1
	WS0104_G16	sp|Q9M156.1	AT4G01070.1
	WS0072_B21	gb|AAR06921.1	AT5G03490.1
	WS00929_B02	gb|ACM09899.1	AT2G31750.1
	WS0061_M07	ref|NP_001044898.1	AT5G65550.1
	WS00939_J04	dbj|BAF75890.1	AT3G53160.1
	WS00813_G09	gb|ABR15470.1	AT4G02280.1
	WS00715_J21	emb|CAI62049.1	AT1G05680.1
	WS0079_E22	gb|ABV68925.1	AT1G22370.1
	WS0094_I12	dbj|BAG80543.1	AT3G21560.1
	WS0042_H15	gb|ABR15470.1	AT4G02280.1
	WS00933_I11	gb|ABR15470.1	AT4G02280.1
	WS0076_D08	emb|CAA47264.1	AT4G02280.1
	WS0075_C05	ref|NP_171646.1	AT1G01390.1
	WS01020_M12	dbj|BAG80549.1	AT1G73880.1
	WS01021_G12	gb|ABY73540.1	AT4G01070.1
	WS0022_A06	ref|NP_199780.1	AT5G49690.1
	WS00720_P10	dbj|BAF75879.1	AT2G36970.1
	WS0082_E13	gb|ACB56925.1	AT3G50740.1
	WS0071_K06	ref|NP_001051422.1	AT5G42765.1
	WS0078_G07	gb|ACG39738.1	AT4G34131.1
	WS0081_H13	dbj|BAG80542.1	AT1G22380.1
	WS01017_J15	gb|ABR15470.1	AT4G02280.1
	WS0097_O04	dbj|BAG80544.1	AT2G31750.1
	WS0094_A14	ref|NP_001051320.1	AT2G36760.1
	WS0015_N03	gb|ABL85472.1	AT4G01070.1
	WS0046_G19	gb|AAM13356.1	AT1G22360.1
	WS00713_P23	sp|Q4R1I9.1	AT1G01390.1
	WS00935_E02	dbj|BAG80551.1	AT2G36970.1
	WS00937_L20	gb|ABN08258.1	AT1G22340.1
	WS0106_G08	sp|Q40287.1	AT4G01070.1
	WS0106_M04	gb|AAF61647.1	AT3G21560.1
	WS0263_O04	gb|ACB56927.1	AT2G15490.1
	WS0076_C19	ref|NP_181234.1	AT2G36970.1
	WS00712_I24	gb|AAR06921.1	AT5G03490.1
	WS00820_C05	emb|CAM31954.1	AT4G01070.1
	WS00928_N23	ref|NP_181234.1	AT2G36970.1
	WS00921_I15	ref|NP_001044898.1	AT5G65550.1
	WS01024_C22	ref|NP_196793.1	AT5G12890.1

Table S. 1 continued

Family	Array EST ID	Annotation (NCBI Viridiplantae)	Annotation (TAIR 8)
GLYTR	WS01028_M19	dbj|BAD91803.1	AT5G12890.1
	WS00916_J06	ref|NP_181216.1	AT2G36780.1
	WS00830_A18	sp|Q9M156.1	AT4G01070.1
	WS0063_M21	ref|XP_002518724.1	
LAC	WS0105_M20	gb|AAK37826.1	AT2G30210.1
	WS01037_G16	gb|AAK37826.1	AT5G05390.1
	WS0039_C04	gb|AAK37824.1	AT5G05390.1
	WS00815_F23	gb|AAK37824.1	AT5G05390.1
	WS0035_M11	gb|AAK37826.1	AT5G05390.1
	WS0062_E09	gb|AAK37824.1	AT2G30210.1
	WS0033_E16	gb|AAK37824.1	AT5G05390.1
	WS0038_B22	gb|AAK37826.1	AT5G05390.1
	WS00923_A19	gb|AAK37826.1	AT5G05390.1
	WS0056_P10	gb|AAK37824.1	AT5G05390.1
	WS0016_N04	gb|AAK37826.1	AT5G05390.1
	WS00730_B15	gb|AAK37826.1	AT5G05390.1
	WS00825_I05	gb|AAK37827.1	AT5G01190.1
	WS00923_N05	gb|AAK37824.1	AT2G40370.1
	WS01037_E20	gb|AAK37826.1	AT5G05390.1
	WS00813_M20	gb|AAK37830.1	AT5G03260.1
	WS0086_J17	gb|AAK37823.1	AT2G38080.1
	WS00914_J21	gb|AAK37826.1	AT5G05390.1
	WS00920_E03	emb|CAK29863.1	AT2G38080.1
	WS01010_I07	AT2G38080.1	AT2G38080.1
	WS01014_L08	gb|AAK37829.1	AT5G03260.1
	WS01021_E17	gb|AAK37826.1	AT2G30210.1
	WS0038_D12	gb|AAK37827.1	AT5G60020.1
	WS0038_I15	gb|AAK37823.1	AT5G60020.1
	WS0055_F20	gb|AAK37826.1	AT5G05390.1
OMT	WS0104_J07	prf||2119166A	AT5G54160.1
	WS0046_C03	prf||2119166A	AT5G54160.1
	WS00925_J22	gb|AAQ01668.1	AT1G51990.1
	WS00927_O09	emb|CAC21601.1	AT5G54160.1
	WS0261_A24	gb|AAD24001.1	AT5G54160.1
	WS00715_G04	emb|CAI30878.1	AT5G54160.1
	WS0074_N24	gb|AAW80883.1	AT5G42760.1
	WS0078_K09	ref|XP_002319364.1	AT1G51990.1
	WS00915_B09	gb|AAC49708.1	AT1G51990.1
	WS01013_K11	gb|AAD24001.1	AT1G63140.2

Table S. 1 continued

Family	Array EST ID	Annotation (NCBI Viridiplantae)	Annotation (TAIR 8)
OMT	WS0099_A18	ref|XP_002326388.1	AT3G53140.1
	WS0263_L06	emb|CAC21601.1	AT5G54160.1
	WS0039_D14	gb|AAD24001.1	AT5G54160.1
	WS0023_M11	emb|CAI30878.1	AT5G54160.1
	WS0086_P13	emb|CAI30878.1	AT5G54160.1
	WS00936_K15	emb|CAC21601.1	AT5G54160.1
	WS02611_F21	emb|CAI30878.1	AT5G54160.1
	WS0264_B22	emb|CAC21601.1	AT5G54160.1
	WS0071_H13	gb|AAC49708.1	AT5G54160.1
	WS00723_D18	gb|AAD24001.1	AT4G35160.1
	WS00727_F10	ref|XP_002302692.1	AT3G62000.1
PAL	WS0044_O08	gb|AAW80645.1	AT3G53260.1
	WS0047_L15	gb|AAP85250.1	AT2G37040.1
	WS00821_C05	sp|P52777.1	AT3G53260.1
	WS0071_D22	sp|P52777.1	AT3G53260.1
	WS01030_B11	sp|P45735.1	AT2G37040.1
pCCoAOMT	WS0031_B01	emb|CAK18782.1	AT4G34050.1
	WS0064_O09	emb|CAK18782.1	AT4G34050.1
	WS0099_L04	gb|ABZ69058.1	AT4G34050.1
	WS01013_K15	emb|CAL55505.1	AT4G34050.1
	WS0107_O19	gb|ABZ69057.1	AT4G34050.1
	WS0057_P01	gb|ABZ69058.1	AT4G34050.1
	IS0014_O19	emb|CAJ43712.1	AT4G34050.1
	WS0261_E15	emb|CAK18782.1	AT4G34050.1
	WS0022_P17	emb|CAJ43712.1	AT4G34050.1
pDFR	WS0016_L17	gb|AAU95082.1	AT1G61720.1
	WS00926_B24	gb|AAU95082.1	AT5G42800.1
	WS0262_E14	gb|AAU95082.1	AT1G61720.1
	WS0075_F05	gb|AAU95082.1	AT5G42800.1
	WS01035_I24	gb|AAU95082.1	AT1G61720.1
	WS0109_E11	emb|CAE47010.1	AT5G42800.1
	WS01024_C11	gb|AAU95082.1	AT5G42800.1
	WS00928_F20	gb|AAU95082.1	AT1G61720.1
	WS01012_M21	gb|AAU95082.1	AT1G61720.1
	WS00929_C24	gb|AAU95082.1	AT5G42800.1
	WS00723_B22	gb|AAU95082.1	AT1G61720.1
	WS00725_B17	gb|AAU95082.1	AT5G42800.1
	WS01035_P13	gb|ABM64800.1	AT5G42800.1
	WS00724_C03	gb|AAU95082.1	AT1G61720.1

Table S. 1 continued

Family	Array EST ID	Annotation (NCBI Viridiplantae)	Annotation (TAIR 8)
pDFR	WS00923_J21	gb|AAU95082.1	AT1G61720.1
	WS0041_A05	gb|AAU95082.1	AT1G61720.1
	WS00113_A17	gb|AAU95082.1	AT4G27250.1
	WS0064_A19	dbj|BAD67186.1	AT5G42800.1
	WS0091_C23	gb|AAU95082.1	AT5G42800.1
pPCBER	WS01011_J14	gb|AAF64182.1	AT4G39230.1
	WS0104_P18	pdb|1QYC|B	AT4G39230.1
	WS00727_J11	tpe|CAI56321.1	AT4G39230.1
	WS00920_D17	gb|AAF64181.1	AT4G39230.1
	WS00723_J06	gb|AAF64181.1	AT4G39230.1
	WS0031_J08	gb|AAF64178.1	AT4G39230.1
	WS0044_K23	tpe|CAI56321.1	AT1G75290.1
	WS00914_A23	ref|XP_002310455.1	AT4G39230.1
	WS00922_A24	tpe|CAI56321.1	AT1G75290.1
	WS0048_K01	gb|AAF64178.1	AT1G75280.1
	WS00916_E05	gb|AAF64185.1	AT1G32100.1
	WS0058_F16	ref|XP_002310455.1	AT1G75290.1
	WS00910_P15	tpe|CAI56321.1	AT1G75290.1
	WS00928_M02	pdb|1QYC|B	AT4G39230.1
	WS0086_D12	gb|AAF64176.1	AT1G75280.1
	WS0089_H07	pdb|1QYD|D	AT1G75280.1
	WS0081_P02	gb|AAF64181.1	AT4G39230.1
	WS00812_I14	gb|AAF64185.1	AT4G13660.1
	WS0031_E17	gb|AAF64185.1	AT1G32100.1
	WS0074_H16	pdb|1QYC|B	AT4G39230.1
	WS00911_H03	gb|AAF63508.1	AT1G32100.1
	WS00820_P24	gb|AAF64178.1	AT4G39230.1
	WS00912_J13	sp|P52580.1	AT4G39230.1
	WS01030_I02	pdb|1QYC|B	AT1G75280.1
	WS01031_B17	pdb|1QYC|B	AT4G39230.1
	WS00712_E21	pdb|1QYC|B	AT4G39230.1
	WS0062_E02	ref|XP_002326577.1	AT4G13660.1

A.2 Expression of phenolic gene families in other tissue sources

Table S. 2 Summary of mixed effects ANOVAs for the expression of gene families in needle.

Family	Factor	nmDF	dnDF	*F*	*P*	*Q*
2-Oxoglutarate Ferrous dependent oxygenase (2OGFE)	Intercept	1	560	1235955.4	<0.01	
	Species	4	10	0.92	0.49	0.28
	Gene	56	560	163.14	<0.01	
	Species x Gene	224	560	1.78	<0.01	
4-Coumaryl CoA ligase (4CL)	Intercept	1	140	508929.6	<0.01	
	Species	4	10	0.9	0.69	0.35
	Gene	14	140	70.82	<0.01	
	Species x Gene	56	140	0.58	0.48	
Coumarate3-hydroxylase (C3H)	Intercept	1	20	33653.94	<0.01	
	Species	4	10	0.99	0.33	0.22
	Gene	2	20	52.61	<0.01	
	Species x Gene	8	20	3.95	<0.01	
Cinnamate 4-hydroxylase (C4H)	Intercept	1	10	108310.54	<0.01	
	Species	4	10	14.65	<0.01	< 0.01*
	Gene	1	10	29.62	<0.01	
	Species x Gene	4	10	4.76	0.02	
Cinnamyl alcohol dehydrogenase (CAD)	Intercept	1	120	348835.5	<0.01	
	Species	4	10	1.82	0.2	0.19
	Gene	12	120	218.17	<0.01	
	Species x Gene	48	120	2.44	<0.01	
Cinnamoyl CoA reductase (CCR)	Intercept	1	160	317597.9	<0.01	
	Species	4	10	0.5	0 .74	0.19
	Gene	16	160	72.52	<0.01	
	Species x Gene	64	160	1.1	0.39	
Chalcone isomerase (CHI)	Intercept	1	30	97791.7	<0.01	
	Species	4	10	1.58	0.25	0.29
	Gene	3	30	150.48	<0.01	
	Species x Gene	12	30	1.01	0.46	

Table S. 2 continued

Family	Factor	nmDF	dnDF	*F*	*P*	*Q*
Chalcone synthase (CHS)	Intercept	1	90	105503.61	<0.01	
	Gene	4	10	1.53	0.27	0.17
	Species x Gene	9	90	86.7	<0.01	
		36	90	2.17	<0.01	
Dirigent proteins (DIR)	Intercept	1	290	423406.6	<0.01	
	Species	4	10	0.95	0.49	0.28
	Gene	29	290	116.4	<0.01	
	Species x Gene	116	290	1.12	0.13	
Flavonoid 3-hydroxylase (F3H)	Intercept	1	120	358173	<0.01	
	Species	4	10	5.4	0.01	0.17
	Gene	12	120	214.18	<0.01	
	Species x Gene	48	120	2.52	0.01	
Ferulate 5-hydroxylase-like (F5H)	Intercept	1	80	283024.16	<0.01	
	Species	4	10	6.19	<0.01	< 0.01*
	Gene	8	80	69.19	<0.01	
	Species x Gene	32	80	2.45	<0.01	
Glycosyl transferase (GLYTR)	Intercept	1	520	594945.3	<0.01	
	Species	4	10	1.64	0.61	0.32
	Gene	52	520	46.1	<0.01	
	Species x Gene	208	520	1.87	<0.01	
Laccase (LAC)	Intercept	1	240	878015.6	<0.01	
	Species	4	10	1.83	0.2	0.19
	Gene	24	240	98.3	<0.01	
	Species x Gene	96	240	1.47	<0.01	
O-methyl transferase (OMT)	Intercept	1	200	371366.4	<0.01	
	Species	4	10	2.64	0.1	0.14
	Gene	20	200	28.22	<0.01	
	Species x Gene	80	200	1.99	<0.01	
Phenylalanine ammonia lyase (PAL)	Intercept	1	40	104497.41	<0.01	
	Species	4	10	5.37	0.01	< 0.05*
	Gene	4	40	333.09	<0.01	
	Species x Gene	16	40	2.97	<0.01	

Table S. 2 continued

Family	Factor	nmDF	dnDF	*F*	*P*	*Q*
Caffeoyl CoA O-methyl trasnferase, putative (pCCoAOMT)	Intercept	1	80	156315.37	<0.01	
	Species	4	10	2.94	0.08	0.13
	Gene	8	80	68.61	<0.01	
	Species x Gene	32	80	0.62	0.93	
Dihydro flavonol reductase, putative (pDFR)	Intercept	1	180	288191.59	<0.01	
	Species	4	10	1.69	0.23	0.19
	Gene	18	180	94.19	<0.01	
	Species x Gene	72	180	2.29	<0.01	
Phenylcoumaran benzylic ether reductase, putative (pPCBER)	Intercept	1	265	420349.5	<0.01	
	Species	4	10	4.26	0.03	0.07
	Gene	25	265	68.57	<0.01	
	Species x Gene	100	265	2.74	<0.01	

DF, degrees of freedom; nmDF, numerator of DF; dnDF, denominator of DF; Species x Gene, species by gene interaction. * Significant Q value at FDR rate: 0.05.

Table S. 3 Summary of mixed effects ANOVAs for the expression of gene families in xylem.

No significant *Q* value is reported

Family	Factor	nmDF	dnDF	*F*	*P*	*Q*
2-Oxoglutarate Ferrous dependent oxygenase (2OGFE)	Intercept	1	560	2518373.7	<0.01	
	Species	4	10	0.26	0.9	0.57
	Gene	56	560	62.52	<0.01	
	Species x Gene	224	560	1.4	<0.01	
4-Coumaryl CoA ligase (4CL)	Intercept	1	140	614564.3	<0.01	
	Species	4	10	1.07	0.42	0.39
	Gene	14	140	58.45	<0.01	
	Species x Gene	56	140	1.48	0.04	
Coumarate3-hydroxylase (C3H)	Intercept	1	20	379996.9	<0.01	
	Species	4	10	2.75	0.09	0.11
	Gene	2	20	491.92	<0.01	
	Species x Gene	8	20	6.61	<0.01	
Cinnamate 4-hydroxylase (C4H)	Intercept	1	10	113142.95	<0.01	
	Species	4	10	1.42	0.3	0.34
	Gene	1	10	19.34	<0.01	
	Species x Gene	4	10	0.69	0.62	
Cinnamyl alcohol dehydrogenase (CAD)	Intercept	1	120	997241.4	<0.01	
	Species	4	10	4.21	0.03	0.08
	Gene	12	120	77.57	<0.01	
	Species x Gene	48	120	2.73	<0.01	
Cinnamoyl CoA reductase (CCR)	Intercept	1	160	588785.8	<0.01	
	Species	4	10	0.38	0.82	0.55
	Gene	16	160	59.22	<0.01	
	Species x Gene	64	160	2.37	<0.01	
Chalcone isomerase (CHI)	Intercept	1	30	232750.63	<0.01	
	Species	4	10	2.95	0.08	0.11
	Gene	3	30	33.74	<0.01	
	Species x Gene	12	30	3.43	<0.01	

Table S. 3 continued

Family	Factor	nmDF	dnDF	*F*	*P*	*Q*
Chalcone synthase (CHS)	Intercept	1	90	514966.6	<0.01	
	Species	4	10	0.62	0.66	0.47
	Gene	9	90	37.71	<0.01	
	Species x Gene	36	90	2.14	<0.01	
Dirigent proteins (DIR)	Intercept	1	290	652388.8	<0.01	
	Species	4	10	0.6	0.66	0.47
	Gene	29	290	83.4	<0.01	
	Species x Gene	116	290	2.3	<0.01	
Flavonoid 3-hydroxylase (F3H)	Intercept	1	120	895581.4	<0.01	
	Species	4	10	0.92	0.49	0.4
	Gene	12	120	170.98	<0.01	
	Species x Gene	48	120	2.39	<0.01	
Ferulate 5-hydroxylase-like (F5H-like)	Intercept	1	80	645499	<0.01	
	Species	4	10	4.79	0.02	0.08
	Gene	8	80	30.01	<0.01	
	Species x Gene	32	80	1.22	0.24	
Glycosyl transferase (GLYTR)	Intercept	1	520	1561081.5	<0.01	
	Species	4	10	1.4	0.45	0.39
	Gene	52	520	77.6	<0.01	
	Species x Gene	208	520	1.8	<0.01	
Laccase (LAC)	Intercept	1	240	520604.2	<0.01	
	Species	4	10	4.2	0.03	0.09
	Gene	24	240	141.6	<0.01	
	Species x Gene	96	240	2.4	<0.01	
O-methyl transferase (OMT)	Intercept	1	200	1352110.2	<0.01	
	Species	4	10	3.1	0.07	0.11
	Gene	20	200	113.5	<0.01	
	Species x Gene	80	200	1.5	0.02	
Phenylalanine ammonia lyase (PAL)	Intercept	1	40	202277.91	<0.01	
	Species	4	10	1.15	0.39	0.39
	Gene	4	40	97.32	<0.01	
	Species x Gene	16	40	0.52	0.09	

Table S. 3 continued

Family	Factor	nmDF	dnDF	*F*	*P*	*Q*
Caffeoyl CoA O-methyl trasnferase, putative (pCCoAOMT)	Intercept	1	80	596036.6	<0.01	
	Species	4	10	3.5	0.05	0.1
	Gene	8	80	98.51	<0.01	
	Species x Gene	32	80	2.1	<0.01	
Dihydro flavonol reductase, putative (pDFR)	Intercept	1	180	1452408	<0.01	
	Species	4	10	7.29	<0.01	0.06
	Gene	18	180	64.99	<0.01	
	Species x Gene	72	180	2.82	<0.01	
Phenylcoumaran benzylic ether reductase, putative (pPCBER)	Intercept	1	265	1187960.7	<0.01	
	Species	4	10	3.4	0.06	0.1
	Gene	25	265	123.8	<0.01	
	Species x Gene	100	265	1.6	<0.01	

DF, degrees of freedom; nmDF, numerator of DF; dnDF, denominator of DF; Species x Gene, species by gene interaction. No significant Q value is reported at 0.05 FDR level.

A.3 Correlation of the divergence of gene expression with neutral divergence

Table S. 4 Bivariate correlations between Amplified Fragment Length Polymorphism (AFLP) distances and differences of gene expression among five species.

Spearman rank correlation coefficient (*r*) is computed per gene (EST; Expressed Sequence Tag) basis for the families having diverged expression.

Family	Transcript ID	Spearman *r*
C4H	WS00931_C11	0.198
	WS00824_E05	-0.078
DIR	WS00914_H24	0.485
	WS01018_J08	0.424
	WS00911_I09	0.108
	WS0262_G24	0.102
	WS0078_C23	-0.095
	WS01012_J06	0.113
	WS0262_J09	0.124
	WS0043_N14	0.046
	WS0094_C05	-0.051
	WS0058_H22	0.156
	WS01012_K18	0.073
	WS0104_A04	0.228
	WS0262_G08	0.071
	WS0265_H11	0.103
	WS01011_J07	0.076
	WS0107_K16	0.058
	WS0064_L20	-0.103
	WS00815_A07	0.174
	WS01012_K12	-0.115
	WS00923_D20	-0.044
	WS00924_E04	0.077
	WS00927_L10	-0.062
	WS01031_M14	0.056
	WS01032_M02	0.117
	WS0063_M13	-0.010
	WS0261_J16	0.052
	WS0262_I24	0.135

Table S. 4 continued

Family	Transcript ID	Spearman *r*
DIR	WS0064_J21	0.241
	WS00825_K05	-0.058
	WS0086_I04	-0.087
GLYTR	WS00110_P02	0.251
	WS00111_D08	-0.007
	WS0071_C13	0.164
	WS0094_H09	0.137
	WS0097_K18	0.139
	WS0044_D10	0.068
	WS0086_B18	0.302
	WS0021_D24	-0.017
	WS00912_N05	0.383
	WS00912_M21	0.007
	WS00931_K21	-0.188
	WS0104_G16	0.249
	WS0072_B21	-0.006
	WS00929_B02	0.279
	WS0061_M07	0.078
	WS00939_J04	0.512
	WS00813_G09	0.122
	WS00715_J21	0.015
	WS0079_E22	0.124
	WS0094_I12	0.208
	WS0042_H15	0.020
	WS00933_I11	0.298
	WS0076_D08	0.222
	WS0075_C05	0.121
	WS01020_M12	0.147
	WS01021_G12	0.103
	WS0022_A06	-0.020
	WS00720_P10	-0.117
	WS0082_E13	0.180
	WS0071_K06	0.155
	WS0078_G07	0.168
	WS0081_H13	0.394
	WS01017_J15	-0.135
	WS0097_O04	0.280
	WS0094_A14	-0.094
	WS0015_N03	-0.050
	WS0046_G19	0.222
	WS00713_P23	-0.044
	WS00935_E02	0.193
	WS00937_L20	-0.100

Table S. 4 continued

Family	Transcript ID	Spearman *r*
GLYTR	WS0106_G08	-0.073
	WS0106_M04	-0.170
	WS0263_O04	0.024
	WS0076_C19	-0.016
	WS00712_I24	0.084
	WS00820_C05	-0.033
	WS00928_N23	0.065
	WS00921_I15	-0.048
	WS01024_C22	-0.145
	WS01028_M19	0.288
	WS00916_J06	0.009
	WS00830_A18	-0.046
	WS0063_M21	-0.063
LAC	WS0105_M20	0.301
	WS01037_G16	0.065
	WS0039_C04	0.373
	WS00815_F23	0.235
	WS0035_M11	0.058
	WS0062_E09	0.036
	WS0033_E16	0.342
	WS0038_B22	0.415
	WS00923_A19	0.243
	WS0056_P10	0.278
	WS0016_N04	-0.117
	WS00730_B15	0.131
	WS00825_I05	0.135
	WS00923_N05	0.129
	WS01037_E20	-0.101
	WS00813_M20	0.301
	WS0086_J17	0.035
	WS00914_J21	-0.099
	WS00920_E03	-0.004
	WS01010_I07	-0.173
	WS01014_L08	-0.072
	WS01021_E17	-0.188
	WS0038_D12	0.048
	WS0038_I15	0.047
	WS0055_F20	0.057
OMT	WS0104_J07	0.130
	WS0046_C03	0.083
	WS00925_J22	0.124
	WS00927_O09	0.330
	WS0261_A24	0.050

Table S. 4 continued

Family	Transcript ID	Spearman *r*
OMT	WS00715_G04	-0.183
	WS0074_N24	0.069
	WS0078_K09	0.129
	WS00915_B09	0.073
	WS01013_K11	0.005
	WS0099_A18	0.046
	WS0263_L06	0.098
	WS0039_D14	0.015
	WS0023_M11	0.032
	WS0086_P13	0.137
	WS00936_K15	0.011
	WS02611_F21	-0.008
	WS0264_B22	0.056
	WS0071_H13	0.017
	WS00723_D18	0.289
	WS00727_F10	0.060
PAL	WS0044_O08	0.095
	WS0047_L15	0.205
	WS00821_C05	0.114
	WS0071_D22	0.250
	WS01030_B11	-0.027
pCCoAOMT	WS0031_B01	0.116
	WS0064_O09	-0.068
	WS0099_L04	-0.026
	WS01013_K15	-0.056
	WS0107_O19	-0.198
	WS0057_P01	-0.117
	IS0014_O19	0.138
	WS0261_E15	0.226
	WS0022_P17	0.081
pPCBER	WS01011_J14	0.384
	WS0104_P18	-0.022
	WS00727_J11	0.093
	WS00920_D17	0.402
	WS00723_J06	0.147
	WS0031_J08	0.270
	WS0044_K23	0.176
	WS00914_A23	0.362
	WS00922_A24	0.148
	WS0048_K01	0.128
	WS00916_E05	0.323
	WS0058_F16	0.140
	WS00910_P15	0.305

Table S. 4 continued

Family	Transcript ID	Spearman *r*
pPCBER	WS00928_M02	-0.021
	WS0086_D12	0.118
	WS0089_H07	0.103
	WS0081_P02	-0.019
	WS00812_I14	-0.049
	WS0031_E17	-0.013
	WS0074_H16	0.049
	WS00911_H03	-0.055
	WS00820_P24	-0.123
	WS00912_J13	0.204
	WS01030_I02	0.178
	WS01031_B17	0.211
	WS00712_E21	0.234
	WS0062_E02	0.078

Appendix B Chapter 5 supplementary data

Table S. 5 List of array elements corresponding to six terpenoid biosynthetic segments.

For each Expressed sequence Tag (EST) the identifiers of their annotated best hits in blast searches against National Center for Biotechnology Information non-redundant data base (NCBI NR) and The Arabidopsis Information Resource version 9 (TAIR 9), and presence or absence of Response to stress Gene Ontology (GO) term is included. Narrow-sense heritability (h^2) is calculated for the expression of each EST and re-scaled within an interval of ±4. MEV, Mevalonate pathway; MEP, non-mevalonate pathway; BKBNE, backbone; TPS; terpene synthases; STRL, steroids segments leading to the biosynthesis of campesterol and brassinosteroid; CAR, carotenoids biosynthetic segment.

Segment	EST ID	Annotation (NCBI NR)	Annotation (TAIR 9)	Response to stress	h^2
MEV	WS00933_A20	gb|EDU44946.1	AT2G33150.1	1	0.343
	WS01021_K16	dbj|BAA11117.1	AT2G33150.1	1	0.397
	WS0099_E21	dbj|BAA11117.1	AT2G33150.1	1	0.368
	WS00923_B12	emb|CAA65250.1	AT4G00960.1	0	0.226
	WS00820_M23	emb|CAA65250.1	AT4G11820.2	0	0.442
	WS00813_F13	gb|AAU89123.1	AT1G76490.1	0	0.481
	WS00930_B21	gb|AAQ82685.1	AT1G76490.1	0	0.549
	WS0072_I23	gb|ABY20976.1	AT2G17370.1	0	0.677
	WS01016_E19	gb|AAQ82685.1	AT2G17370.1	0	0.448
	WS00722_H11	gb|AAQ82685.1	AT1G76490.1	0	0.664

Table S. 5 continued

Segment	EST ID	Annotation (NCBI NR)	Annotation (TAIR 9)	Response to stress	h^2
MEV	WS0022_N13	gb|AAQ82685.1	AT1G76490.1	0	0.423
	WS01032_B02	gb|ACG45252.1	AT2G26800.2	0	0.385
	WS00925_M22	gb|ACG45252.1	AT2G26800.2	0	0.077
	WS0052_A23	gb|AAL18925.1	AT5G27450.1	0	0.506
	WS01021_M09	gb|AAL18926.1	AT1G31910.1	0	0.261
	WS00825_J03	gb|EEY22280.1	AT1G31910.1	0	0.407
	WS0104_E21	gb|AAV32433.1	AT3G54250.1	0	0.463
	WS01018_H09	gb|AAV32433.1	AT3G54250.1	0	0.529
	WS01029_I02	gb|EDU46934.1	no hit	0	0.261
MEP	WS0097_H02	gb|ABS50518.1	AT4G15560.1	0	0.491
	WS00930_F08	gb|ABS50520.1	AT4G15560.1	0	0.607
	WS01028_M14	gb|ABS50519.1	AT4G15560.1	0	0.746
	WS00930_P05	gb|ACJ67022.1	AT5G62790.1	0	0.454
	WS0078_L16	gb|AAZ80386.1	AT2G02500.1	0	0.700
	WS0031_C22	gb|AAY40863.1	AT1G63970.1	0	0.628
	WS00822_P08	gb|ABD73009.1	AT1G63970.1	0	0.428
	WS01033_I18	gb|ABB78087.1	AT5G60600.1	1	0.507
	WS01030_N10	gb|ABO26588.1	AT4G34350.1	0	0.504
	WS00815_M18	gb|ABO26588.1	AT4G34350.1	0	0.433
	WS0099_C01	gb|ABO26587.1	AT4G34350.1	0	0.465
BKBNE	WS0074_I12	gb|ACU56978.1	AT5G16440.1	0	0.617
	WS0013_N24	gb|ACA21460.1	AT5G47770.1	1	0.453
	WS00111_A06	emb|CAB91608.1	AT3G59380.1	0	0.701
	WS0038_A18	gb|AAL17614.2	AT4G36810.1	0	0.837
	WS00911_G14	gb|ACA21462.1	AT4G36810.1	0	0.720
	WS0104_P09	gb|AAN01134.1	AT4G36810.1	0	0.513

Table S. 5 continued

Segment	EST ID	Annotation (NCBI NR)	Annotation (TAIR 9)	Response to stress	h^2
BKBNE	WS01030_E02	dbj|BAF98303.1	AT4G38460.1	0	0.589
	WS00911_G14	gb|AAN01134.1	AT4G36810.1	0	0.513
	WS01039_D11	gb|ACA21459.1	AT2G34630.1	0	0.505
	WS0062_C04	gb|AAL17614.2	AT1G04550.2	0	0.434
	WS0106_K12	gb|EAA30713.1	AT1G78510.1	0	0.508
	WS00939_L12	gb|EDU41081.1	AT2G30920.1	0	0.486
	WS0045_M09	gb|EDU41081.1	AT2G30920.1	0	0.369
	WS00819_P07	dbj|BAF29571.1	AT1G78510.1	0	0.433
	WS0261_K09	dbj|BAH10639.1	AT1G74470.1	0	0.529
TPS	WS0022_E04	gb|AAP72020.1	AT4G16730.1	0	0.660
	WS00924_B02	gb|AAP72020.1	AT2G24210.1	1	0.508
	WS00929_M11	gb|AAK83564.1	AT4G16740.2	1	0.774
	WS0105_B05	gb|AAS47692.1	AT4G16730.1	0	0.717
	WS0063_I21	gb|AAP72020.1	AT2G24210.1	1	0.580
	WS0092_L05	gb|AAO73863.1	AT2G24210.1	1	0.651
	WS0094_F18	gb|AAS47693.1	AT2G24210.1	1	0.662
	WS00923_A21	gb|AAK39127.2	AT2G24210.1	1	0.555
	WS00926_E08	gb|AAS47697.1	AT1G61680.1	0	0.621
	WS00723_E14	gb|AAS47693.1	AT4G02780.1	0	0.469
	WS0092_I21	gb|AAS47693.1	AT4G02780.1	0	0.588
	WS00819_E12	gb|AAS47693.1	AT1G61680.1	0	0.434
	WS00724_C19	gb|ABA86247.1	AT1G61680.1	0	0.511
	WS00712_A10	gb|ABV44452.1	AT1G61680.1	0	0.706
	WS0019_A03	gb|AAO73863.1	AT1G61680.1	0	0.496
	WS00112_B15	gb|ABA86247.1	AT4G02780.1	0	0.459
	WS0106_I22	gb|AAS47695.1	AT1G61680.1	0	0.540

Table S. 5 continued

Segment	EST ID	Annotation (NCBI NR)	Annotation (TAIR 9)	Response to stress	h^2
	WS0078_K20	gb|AAK83565.1	AT1G48800.1 |	0	0.592
	WS0063_F08	gb|ACM04452.2	AT4G16730.1	0	0.715
	WS00929_B22	gb|ABA86249.1	AT1G70080.1	0	0.670
	WS00927_M20	gb|AAS47695.1	AT4G16730.1	0	0.528
	WS0072_A14	gb|AAK39129.2	AT4G16730.1	0	0.548
	WS01033_J22	gb|AAS47689.1	AT4G02780.1 |	0	0.504
	WS01031_F19	gb|AAS47690.2	AT4G16730.1	0	0.540
STRL	WS01035_P02	gb|ABI14439.1	AT4G34640.1	0	0.778
	WS0045_C22	gb|ABX64425.2	AT3G45020.1	0	0.571
	WS00922_A05	gb|ACG45015.1	AT1G58440.1	1	0.528
	WS0021_N23	gb|ACJ05633.1|	AT1G58440.1	1	0.486
	WS00930_J18	gb|AAZ83345.1	AT5G13710.1	0	0.369
	WS00112_C09	gb|AAM91592.1	AT1G20330.1	0	0.379
	WS0097_B02	gb|AAG44096.1	AT2G07050.1	0	0.314
	WS0024_M18	gb|AAG44096.1	AT2G07050.1	0	0.380
	WS0074_F24	gb|ACG37002.1	AT1G20050.1	0	0.584
	WS00111_M22	AT3G02580.1	AT3G02580.1	0	0.409
	WS01037_P04	gb|AAH63347.1	no hit	0	0.560
	WS00712_J02	gb|ABA01479.1	AT3G19820.1	0	0.405
	WS0056_K01	gb|ABA01479.1	AT3G19820.1	0	0.585
	WS00823_B03	emb|CAD41584.3	AT3G50660.1	0	0.463
	WS0264_A16	gb|ACG33439.1	AT3G50660.1	0	0.480
	WS00712_B06	gb|EFH47957.1	AT5G16010.1	0	0.383
	WS00822_H12	gb|EFH47957.1	AT5G16010.1	0	0.565
	WS01027_C02	gb|EFH47957.1	AT5G16010.1	0	0.516
	WS00730_D04	gb|EFH47957.1	AT5G16010.1	0	0.366

Table S. 5 continued

Segment	EST ID	Annotation (NCBI NR)	Annotation (TAIR 9)	Response to stress	h^2
CAR	WS00917_H06	gb|ACO53104.1	AT5G17230.1	0	0.263
	WS0075_O23	gb|ABD91578.1	AT3G10230.1	0	0.397
	WS00112_N12	gb|ABA43903.1	AT4G25700.1	0	0.502
	WS0073_E12	gb|AAG10793.1	AT4G25700.1	0	0.418
	WS0013_L22	gb|ACO53105.1	AT3G53130.1	0	0.361
	WS01012_C16	gb|AAQ04224.1	AT3G04870.1	0	0.481