Genome Biology and Evolution Advance Access published March 5, 2015 doi:10.1093/gbe/evv044

Research Article

Genome-wide analysis reveals diverged patterns of codon bias, gene expression and rates of sequence evolution in Picea gene families

*Amanda R. De La Torre**[1], *Yao-Cheng Lin*[3], *Yves Van de Peer*[3,4], *Pär K. Ingvarsson*[1,2]

[1] Department of Ecology and Environmental Science, Umeå University, Linneaus väg 6, SE-901 87 Umeå, Sweden

[2] Umeå Plant Science Centre, Umeå, Sweden

[3] Department of Plant Systems Biology (VIB) and Department of Plant Biotechnology and Bioinformatics (Ghent university), Technologiepark 927, 9052 Ghent, Belgium

[4] Genomics Research Institute, University of Pretoria, Hatfield Campus, Pretoria 0028, South Africa.

*Author for correspondence: Amanda R. De La Torre, Umeå university, Sweden, 070-2262973, amanda.delatorre@emg.umu.se

Journal: Genome Biology and Evolution

Abstract: 238

Total word count for main body of the text (excluding references): 6323

Number of figures: 4 in color

Number of tables: 2

Supplemental figures: 9

Supplemental tables: 7

Abstract

The recent sequencing of several gymnosperm genomes has greatly facilitated studying the evolution of their genes and gene families. In this study, we examine the evidence for expression-mediated selection in the first two fully sequenced representatives of the gymnosperm plant clade (*Picea abies* and *Picea glauca*). We use genome-wide estimates of gene expression (>50,000 expressed genes) to study the relationship between gene expression, codon bias, rates of sequence divergence, protein length and gene duplication.

We found that gene expression is correlated with rates of sequence divergence and codon bias, suggesting that natural selection is acting on *Picea* protein-coding genes for translational efficiency. Gene expression, rates of sequence divergence and codon bias are correlated with the size of gene families, with large multi-copy gene families having, on average, a lower expression level and breadth, lower codon bias, and higher rates of sequence divergence than single-copy gene families. Tissue-specific patterns of gene expression were more common in large gene families with large gene expression divergence than in single copy families. Recent family expansions combined with large gene expression variation in paralogs and increased rates of sequence evolution suggest that some *Picea* gene families are rapidly evolving to cope with biotic and abiotic stress.

Our study highlights the importance of gene expression and natural selection in shaping the evolution of protein-coding genes in *Picea* species, and sets the ground for further studies investigating the evolution of individual gene families in gymnosperms.

Key words: Gene expression, gene duplication, single-copy genes, codon usage, sequence divergence.

Introduction

Understanding the molecular changes underlying phenotypic differences between species is of great biological interest (Hahn et al. 2007). Functional evolutionary innovations leading to new phenotypes often result from changes in gene expression (Gu et al. 2004; Gallego-Romero et al. 2012; Wang and Paterson, 2012; Jacquemin et al. 2014). Changes in gene expression are common between genes and reflect the differences in time and energy allocated to the expression of genes whose products are required by the cell in different concentrations (Williford and Demuth, 2012). The more "connected" (in a network of dependency) the gene products are, the more sensitive the phenotype is to changes in product concentration (altering an individual's fitness) (Freeling et al. 2009; Birchler and Veitia, 2012). Gene expression divergence have been linked to several gene features such as evolutionary rates, codon bias, intron size, coding sequence length and amino acid composition. The widespread occurrence of these associations across different phylogenetic taxa highlights the role of gene expression in the evolution of protein-coding genes (Williford and Demuth, 2012; Warnefors and Kaessmann, 2013).

Gene expression divergence has also been linked to gene duplication. (Ohno, 1970; Lynch and Katju, 2004; Conant and Wolfe, 2008; Flagel and Wendel, 2009). Gene copies resulting from duplication may evolve under fewer functional constraints and relaxed purifying selection and eventually acquire a new function (neo-functionalization). Alternatively, the functions originally performed by an ancestral gene may be partitioned between two descendant gene copies, resulting in genes having complementary roles (sub-functionalization). The maintenance of duplicated genes may also be explained by dosage balance, network connectivity and multiprotein complex issues (Birchler et al. 2001; Papp et al. 2003; Veitia et al. 2004). The dosage model states that any successful genome has evolved, by positive selection, an optimum balance of gene products that interact with each other to form protein complexes, involved in multiple steps of biological processes. This model has been used to explain the maintenance of duplicates in *Arabidopsis* (Blanc and Wolfe, 2004; Freeling and Thomas, 2006; Barker et al. 2008; Freeling, 2008) and yeast (Davis and Petrov, 2005) and to justify the presence of dosage-sensitive modifiers of the *white* eye color in *Drosophila* (Birchler et al. 2001). In contrast, duplication of some genes may be strongly deleterious because an increase in copy number may unbalance their interactions with other proteins within the cell, leading to dosage imbalance (Makino and McLysaght, 2010; De Smet et al. 2013). Because of these selective and functional constraints, single-copy genes are expected to evolve more slowly than genes in multigene families, a view widely supported by numerous studies (Han et al. 2009; Jaillon et al. 2009; De Smet et al. 2013) although a few studies have reported the opposite trend (Yang et al. 2003; Jordan et al. 2004). The relative proportions of neo-functionalization, sub-functionalization and pseudogenization may influence gene family sizes and the evolution of gene families (Chen et al. 2010).

Several properties of gene interaction networks (such as node connectivity and centrality) may also influence gene duplicability, and may reflect differential selective forces acting on various genes (Hahn et al. 2004; Ramsay et al. 2009). Duplicated genes are found more often at the periphery of networks in *E. coli*, yeast, and *Drosophila*; whereas in humans, they tend to occupy the most central positions (Doherty et al. 2012). Whole-genome studies have shown that highly connected network elements (e.g., nodes) tend to be more functionally constrained than nodes with fewer connections in both metabolic and protein-protein interaction networks (Hahn et al. 2004; Vitkup et al. 2006). Therefore the position of an element in a network certainly affects its evolutionary fate. Upstream genes are generally more selectively constrained than downstream genes in biochemical pathways because mutations in upstream genes would generate greater pleiotropic effects potentially leading to deleterious effects (Otto et al. 2004; Ramsay et al. 2009; Alvarez-Ponce et al. 2009). However, some studies have failed to detect a correlation between pathway position and sequence divergence (Yang et al. 2009; Jovelin and Phillips 2011).

Gymnosperms are a major plant clade that diverged from angiosperms about 300 MYA (Leslie et al. 2012). Despite initial attempts to understand their characteristic biology and unique genome architecture, studies in gymnosperms were hampered by the lack of reference genomes. The recent genome sequencing of three gymnosperms has revealed different features compared with other plant genomes (De La Torre et al. 2014). In brief, it is been shown that the enormous size of conifer genomes (20 to 40 Gb), by far the largest genomes sequenced to date, is not a consequence of whole-genome duplications (WGD) nor it is due to an increased number of protein-coding genes. Instead, conifer genomes have grown by a slow and steady accumulation of a diverse and large set of transposable elements (Hamberger et al. 2009; Nystedt et al. 2013; De La Torre et al., 2014). Although the number of protein-coding genes is not significantly higher in gymnosperms than in angiosperms, recent phylogenetic studies have shown that some gene families have evolved differently in these two plant clades (Hamberger and Bohlmann 2006; Porth et al. 2011; Nystedt et al. 2013; Neale et al. 2014). Previous studies investigating the selective forces and evolutionary rates in gymnosperms have been based on very few numbers of genes (Willyard et al. 2007; Palme et al. 2009; Chen et al. 2010). More recently, two studies included orthologous comparisons between *Picea* and *Pinus* species using a higher number of genes (3,000-5,000) obtained from transcriptome and EST data (Buschiazzo et al. 2012; Chen et al. 2012). The incipient state of knowledge on the evolution of gymnosperm gene families would surely be enhanced by genome-wide studies that include analyses of gene expression data and sequence divergence in gene families of several species.

In this study, we examine the evidence for expression-mediated selection in the first two fully sequenced representatives of the gymnosperm plant clade (*Picea abies* and *Picea glauca*). We use genome-wide estimates of gene expression to investigate the relationship between gene expression, codon bias, rates of sequence divergence and protein length. We also tested for the relationship between gene expression and gene duplication, using gene family size as a proxy; and pathway position and gene duplication (using the terpenoid pathway as an example). This study highlights the importance of gene expression and natural selection in shaping the evolution of protein-coding genes in *Picea* species.

Material and Methods

Sequence retrieval and expression profiles

Coding sequences were obtained for 26,597 genes from the high-confidence gene set in the *Picea abies* genome (Nystedt et al. 2013, http://congenie.org). Following the same procedure, coding sequences from 27,721 genes (derived from full-length cDNA) were retrieved from the *Picea glauca* gene catalogue (Rigault et al. 2011, http://www.arborea.ulaval.ca).

Expression profiles for 23,854 *P. glauca* genes were obtained for eight different tissue types, including vegetative buds, needles, xylem (mature), xylem (juvenile), phelloderm, adventitious roots, megagametophytes and embryonic cells, from the PiceaGenExpress database (Raherison et al. 2012). Samples were collected from clonal replicates of young *P. glauca* trees in Canada. RNA was extracted, labeled and hybridized using microarrays, as fully described in Raherison et al (2012). Using customized Perl scripts, these genes were matched with those in the *P. glauca* gene catalogue. Functional annotations were based on the detection of Pfam domains and on matches with *Arabidopsis* (TAIR 9 release) with e-value <1e-10. Blast2GO v.2.7.0 was used to perform a *Blastx* search (e-value<1e-10) and Gene Ontology mapping with the plant GO-Slim terms (Conesa et al. 2005).

Expression profiles from *P. abies* were obtained from 22 samples for eight different tissues that included needles, male and female cones, shoots, buds, pineapple galls, stems, and early and late wood. Samples were collected from multiple, pooled biological samples obtained from clonal copies growing in northern Sweden. After RNA extraction, Paired-end RNA sequencing (RNASeq) was performed using the Illumina HiSeq 2000 platform. All details of the sequencing, de novo transcriptome assembly and generation of transcripts' expression values are fully described in Nystedt et al. (2013). Genes were ranked based on their average signal intensities within a tissue type and divided in ten equally large classes (10% quantiles) to allow comparisons with the *P. glauca* gene expression data. Expression breadth, defined as the number of tissues where a gene scored at least one hit, was also calculated for both *P.glauca* and *P.abies*. Total coding sequence length was calculated using the fastalength program from the exonerate package (version 2.2.0, Slater and Birney, 2005).

Identification of orthologous groups and alignments

Open reading frames (ORF) were predicted from the cDNAs in the *P.glauca* gene catalogue using the program FrameDP (Gouzy et al. 2009). Some redundancy was found when estimating the ORFs as 2,197 cDNAs matched to more than one ORF. In these cases, the ORF with the longest sequences were kept. After ORF prediction and UTR removal, 19,057 coding sequences (from an initial set of 27,721) were kept and used for posterior gene family analysis. In *P. abies*, we did not detect any ORF redundancy since we only used the "high-confidence" genes identified in the *P. abies* genome paper (Nystedt et al. 2013). A few genes (273 coding sequences) however, had partial initial codons, likely caused by partial gene lengths. This reading frame shift was corrected with an in-house BioPerl script. After ORF detection, UTR removal, and frame correction 26,164 sequences were kept from an original data set of 26,597 sequences.

We used an all-against-all BLASTP followed by a Markov Cluster algorithm to group (putative) orthologous protein sequences between the genomes of *P.glauca* and *P. abies* with the program Ortho-MCL (Li et al. 2003, http://orthomcl.org). These Ortho-MCL orthologous groups (OG) or "gene families" were composed by orthologs (between species) and recent paralogs (within species). The Ortho-MCL clustering was used to estimate gene family size. We grouped families of similar sizes according to the number of genes in each OG as follows: single-copy (one gene in each species); 2-5 genes; 5-10 genes; 10-20 genes; 20-100 genes; and more than 100 genes. Annotation for orthologous gene families was based on Pfam domain information. Because the Ortho-MCL results may be susceptible to the choice of the inflation parameter, which controls the orthologous groups' size and on the accuracy of the alignments, we manually revised the alignments of all gene families and constructed phylogenetic trees to assess whether the genes in each orthologous group could be considered true orthologs.

In addition to Ortho-MCL, we used MUSCLE (Edgar et al. 2004) to generate multiple alignments for each gene family. Gene families having more than 500 orthologous genes (for both *Picea* species) did not align well in MUSCLE because the similarity in the conserved domain of genes was very high whereas outside this region the similarity decreased significantly. Under these circumstances, the global multiple alignments inferred by

MUSCLE were not reliable, and we therefore decided to exclude all alignments and codeml results for four very large families due to this problem. Alignments containing a majority of gaps and missing data were not considered for subsequent analyses. Non-unambiguously aligned regions in the alignment were removed based on BLOSUM62 (Henikoff and Henikoff, 1992) scoring matrix values allowing each aligned amino acid position with 10% of divergence, and converted to Phylip format.

Estimation of codon bias

Codon bias, measured as the frequency of optimal codons (Fop), was obtained for 26,052 genes in *P. abies* and for 19,056 genes in *P. glauca* using the program CodonW (version 1.4.2, http://codonw.sourceforge.net). Only trimmed coding sequences (excluding UTRs) were analyzed. Differences in codon usage between highly expressed and lowly expressed genes in both *P. glauca* and *P.abies* were based on comparing the positions of each codon on the first and second axis of a correspondence analysis of synonymous codon usage. The identification of codons that are preferentially used in highly expressed genes (optimal codons) was done by observing the clustering of codons along the first axis (correlated with gene expression) in the correspondence analysis (Supporting Information Figs. S2 and S3). This set of optimal codons was then used as input in CodonW to estimate measures of codon adaptation index (CAI) and codon bias index (CBI). GC content and GC content at third position (GC3s) were also calculated with CodonW.

Estimation of substitution rates

Transcript sequences for *P. glauca* and *P. abies* were trimmed and only the coding part was kept during the pairwise alignment. Based on the Ortho-MCL results we built a list of pairwise sequences for each gene family. Then we aligned the mRNA sequences per codon using Needle (http://www.ebi.ac.uk/Tools/psa/emboss_needle) to obtain the input files for the codeml analysis. Needle uses the Needleman-Wunsch dynamic programming algorithm to globally align two protein or nucleotide sequences along their length. Alignments containing a majority of gaps and missing data were not considered for subsequent analyses.

Synonymous (dS) and nonsynonymous (dN) nucleotide substitution rates per site were calculated using the maximum likelihood method of Goldman and Yang (1994) in the Codeml program from the PAML package (version 4.6, Yang, 2007). For each sequence pair, only the results with the highest lnL (log likelihood) were retained. Average scores of retained values were taken after ten repeats. The synonymous/nonsynonymous ratio (dN/dS or ω) is a measure of natural selection acting on the protein, in which values of ω<1 mean negative purifying selection; ω=1, neutral evolution; and ω>1, positive selection. We discarded genes with dS values lower than 0.01, as these values may result in inaccurate estimates of ω, and genes with dS or dN>2 which suggest saturation of substitutions. Abnormally high ω ratios (ω>10) were also discarded (Villanueva-Cañas et al. 2013).

Functional enrichment analyses

We tested for over-representation of functional categories in two of the different gene family sets (single-copy gene families, and large gene families with more than 100 genes) using the BINGO 2.44 Cytoscape plugin (Maere et al. 2005). Corrections for multiple testing were done using the Benjamini and Hochberg method (1995) with a false discovery rate threshold of 0.05.

Statistical analyses

Expression level and expression breadth were tested for correlations with all variables including codon bias (Fop, CBI, CAI), GC content (GC and GC3s), protein length, number of synonymous, non-synonymous substitutions and their ratio (dN, dS, ω) and total number of introns. Kruskal-Wallis one-way analysis of variance by ranks, which is a non-parametric test suitable for comparing groups of unequal size, was used to test associations between family size and gene expression, codon bias and rates of sequence divergence. All analyses were made using the R statistical package (version 3.0.3, R Core Team, 2014).

Correlations of substitution rates, codon bias, and gene expression with pathway position

Protein sequences involved in the conversion to lutein, abscisic acid, gibberellic acid and brassinosteroids in the terpenoid pathway were analyzed. We used an all-against-all BLASTP followed by a Markov Cluster algorithm in Ortho-MCl, to identify orthologous protein sequences of *Arabidopsis* (as reported in Ramsay et al. 2009) in the genome of *Picea abies*. Pathway position was measured following Ramsay's (2009) pathway pleiotropic index, in which groups of enzymes are numbered relative to pathway branch points from most upstream to most downstream. Correlations with codon bias (Fop), substitution rates (ω), gene expression level, and gene expression breadth were tested using the R statistical package (version 3.0.3, R Core Team, 2014).

Results

Gene expression

Because the number of different tissues a gene is expressed in (expression breadth) influences gene expression, we ranked the genes based on their average signal intensities within a tissue type, to avoid possible biased correlations between codon bias and gene expression. Despite this, our analysis of 26,597 genes in *P. abies* and 27,721 genes in *P. glauca* suggests both expression level and expression breadth were strongly correlated in *P. glauca* (r= 0.79, P<0.001) and *P. abies* (r= 0.56, P<0.001), as it is been reported in other species such as *Populus tremula* (Ingvarsson 2007). Expression breadth was positively correlated with protein length, and total intron length in *P. abies* (r= 0.17, P<0.001; and r= 0.21, P<0.001). Expression level was only weakly correlated with GC content in both *P. abies* and *P. glauca* (r= 0.08, P<0.001; and r= 0.07, P<0.001), but not correlated with GC content at third position in neither *P. abies* nor *P. glauca* (r= 0.01, P= 0.3705; and r= -0.01, P= 0.3596). Results of the correlations among all variables can be found in Supporting Information Tables S1 and S2).

Codon bias

Codon bias, defined as the preferential use of a subset of synonymous codons, for optimal translational efficiency, is most pronounced in highly expressed genes in species whose effective population size is large, such as many tree species (Ingvarsson, 2008). In our study, we found high levels of codon bias, measured as the frequency of optimal codons (Fop); Fop averaged 0.58 across 19,057 genes in *P. glauca*; and 0.60 across 26,164 genes in *P. abies*. Codon bias (Fop) was highly and negatively correlated with GC content and GC content at third position in both *P. glauca* (r= -0.78, P<0.001 and r= -0.98, P<0.001, respectively) and *P.abies* (r= -0.81, P<0.001 and r= -0.98, P<0.001). This suggests that optimal codons in *Picea* do not usually have G or C at the third codon position, which differs from other plant species (Serres-Giardi et al. 2012). GC content and GC at third position were also highly correlated in both species (r= 0.76, P<0.001 in *P. glauca* and r= 0.79, P<0.001 in *P. abies*).

There was a significant correlation between expression breadth and the frequency of optimal codons (Fop) in *P. abies* (r= 0.11, P<0.001) and *P. glauca* (r= 0.14, P<0.001). When genes were grouped based on their total expression breadth and average Fop was calculated for each class, the correlation between Fop and expression breadth increased significantly in *P. glauca* (r= 0.93, P<0.001; Figure 1). In contrast, Fop was not correlated with gene expression, when gene expression was based on average signal intensities within a tissue type and divided in classes; but it was weakly correlated to gene expression when maximum gene expression across tissues was used (r= 0.03, P<0.001).

Significant clustering of codons along the first axis of the corresponding analysis was observed in both *P. glauca* and *P. abies*. Codons in highly expressed genes were located in one extreme and, codons of lowly expressed genes in the other extreme of the first axis (Supporting Information Figs. S1 and S2). This is usually taken as evidence that the major trend driving codon usage is correlated with gene expression, supporting the choice for optimal codons. In addition, the fact that optimal codons are the same in both spruce species (with the exception of optimal codons coding for Arginine (Arg) amino acid) further supports the selection for optimal codons. The similarity in codon usage for both spruce species is consistent with the relatively shallow phylogenetic distance between species (species divergence was ca. 14 MYA). Patterns of codon usage are shown in Supporting Information Tables S3 (*P. abies*) and S4 (*P. glauca*).

Rates of sequence divergence

Synonymous (dS) and nonsynonymous (dN) nucleotide substitution rates per site were calculated for all *P. glauca* and *P. abies* genes. The median number of synonymous substitutions (dS) across all gene pair combinations was 0.086 and 0.056 for nonsynonymous substitutions (dN). Even though the synonymous/nonsynonymous ratio (ω) varied widely among gene pair combinations, the median ratio was equal to 0.404, suggesting very strong functional constraint at most of the genes. We found 9126 genes with a ratio (ω) lower than 1, suggesting negative purifying selection; and 1054 genes with a ratio higher than 1, suggesting positive selection. Over-represented functional categories of genes under diversifying selection (ω>1) included genes involved in biotic and abiotic stress responses (Leucine rich repeat, Late embryogenesis abundant, Actin, Histone, Pollen allergen, Salt stress response, Stress responsive, Heat repeats, Auxins, Dehydrins and Heavy-metal associated domain, Protein kinases and Cytochrome P450); and genes involved in carbohydrate metabolism and transport (several families of Glycosyl hydrolases including Cellulase and Chitinase, and several families of Glycosyl transferases) (Table S7). Similar over-represented functional categories of genes involved in biotic and abiotic stress response were found in a smaller scale study comparing EST data between *Picea sitchensis* and *Pinus taeda* (Buschiazzo et al. 2012).

In relation to putative pseudogenes, we found 31 genes with ω>1 that were not expressed in any tissue, and 80 genes for which gene expression data was not available in *P.glauca*. In *P. abies*, we found 30 genes with ω>1 that lack expression data. These genes, however, had start and end codons in their coding sequences, and matched known genes in other species (http://congenie.org). Considering that we tested over 50,000 genes in this study, the number of putative pseudogenes in the data set is likely negligible, and is therefore unlikely to bias any of our results.

Gene family analysis

A total of 38,662 genes (22,972 from *P. abies* and 15,690 from *P. glauca*) from both species were assigned to 5,151 orthologous groups (gene families). The frequency distribution of gene family sizes follows power-law distributions that tend to become flatter as the number of genes in the genomes increased (Huynen and van Nimwegen, 1998). Most of the genes were clustered in small families, containing 2 to 10 genes; or belonging to single-copy gene families (Figure S3). We define single-copy genes as those that are present in both *P. glauca* and *P. abies*, and that possess a one-to-one orthologous relationship in these genomes, meaning that they have remained single-copy since their last common ancestor (approximately 14 MYA), or that have been restored to single-copy status following gene duplication during that time.

We identified a few families that contained more than 100 genes. *P. abies* had eight families with 100 to 200 genes, six families with 200 to 300 genes and two families with 463 and 523 genes respectively. *P. glauca* had five families with 100 to 200 genes, and one with 269 genes. Based on their Pfam annotations, these very large gene families were mainly composed of protein kinases, LRR (Leucine-Rich Repeats) and PPR (Pentatricopeptide repeats). We also identified some gene families that lack orthologous gene families in the other species (740 in *P. abies* and 331 in *P. glauca*). These lineage-specific families, also called “orphans”, may arise from duplication of previously existing genes followed by rapid divergence or by de novo evolution of new genes (Heinen et al. 2009, Carvunis et al. 2012; Neme and Tautz, 2014). Alternatively they may just be artifacts of the threshold used for clustering or a result of missed annotations of genes in incompletely sequenced genomes (Hahn et al. 2007; Tautz and Domazet-Lošo, 2011). The smaller number of *P. glauca* specific gene families may be due to the incomplete nature of FL-cDNA data. In addition, 3,625 genes in *P. abies* and 3,368 genes in *P. glauca* could not be grouped using Ortho-MCL.

Gene family size is correlated with gene expression, rates of sequence divergence and codon bias

Significant correlations between family size and all variables studied, including gene expression, rates of sequence divergence (dN, dS, ω), codon bias (Fop, CAI, CBI), GC content and protein length were found in both *P. abies* and *P.glauca* (Table 1). Expression level and expression breadth tend to decay as the size of the families increased, being the highest in single-copy gene families, and the lowest in families with over 100 genes ($P < 2.2e^{-16}$, one-sided Mann-Whitney U test) (Figure 2 and Supporting Information Figure S4). Interestingly, genes in the single-copy family group had the highest levels of expression breadth, meaning that they were widely expressed in most or all tissues ($P < 2.2e^{-16}$, one-sided Mann-Whitney U test).

Codon bias followed a similar pattern as gene expression, where families having a high expression level also had a high codon bias, due to the positive correlation between gene expression and codon bias ($P < 2.2e^{-16}$, one-sided Mann-Whitney U test). With regard to protein length, single-copy gene families seemed to encode longer proteins than gene families of bigger sizes. GC content at third position increased as the size of the gene families increased ($P < 2.2e^{-16}$, one-sided Mann-Whitney U test), being the lowest in single-copy gene families and the highest in families with 20-100 genes (Figure 2 and Supporting Information Figure S4). Lower levels of synonymous and non-synonymous substitutions were found in single-copy gene families than in multigene families of different sizes, with the largest families (>100 genes) having the highest substitution rates ($P < 2.2e^{-16}$ (*P.abies*) and $P = 4.34 e^{-7}$ (*P.glauca*), one-sided Mann-Whitney U test) (Figure S5).

Functional enrichment of gene families

Results of the functional enrichment analyses based on GO categories suggest very different over-representation of functional categories in large multi-copy gene families compared to single-copy gene families (Table 2). Functional enrichment in large gene families included broad molecular functions such: protein kinase activity, transferase and phosphotransferase activity, binding, catalytic activity, and signal transduction; response to different stimulus such as response to acids and bacteria; and regulation of different processes such as regulation of cellular processes, meristem growth, immune response. It also included genes involved in reproductive development processes and post-embryonic development (Supporting Information Table S5). Functional enrichment in single-copy gene families included genes involved in the process of gene expression such as: translation; metabolism and biosynthesis of nucleotides, DNA, ncRNA, mRNA, RNA and proteins; RNA (binding, methylation, modification, splicing) and ncRNA processing; and nitrogen compound metabolic and biosynthetic processes (Supporting Information Table S6).

Correlations of substitution rates, codon bias, and gene expression with pathway position

Using a set of 473 orthologous protein sequences from *P. abies*, we found significant negative correlations between codon bias (Fop) and pathway position in the groups of enzymes involved in the conversion from glucose to abscisic acid (r=-0.61, P=0), and glucose to gibberelic acid (r=-0.45, P=0). Expression breadth was also negatively correlated with pathway position in the groups of enzymes involved in the conversion from glucose to brassinosteroid (r=-0.26, P=0.0064), and to gibberelic acid (r=-0.21, P=0.0053) (Figure 4). All other correlations with codon bias and expression breadth were not significant. Correlations with substitution rates (ω) were also not significant. Pathway position of all studied branches are detailed in Supporting Information Figure S7.

Discussion

Gene expression and rates of sequence divergence

Although changes in gene expression may play an important role in phenotypic divergence and adaptation, the role of natural selection on the evolution of gene expression levels remains largely unknown in non-model species (Gilad et al. 2006; Bedford and Hartl, 2009). Our study in *Picea* shows that selection is likely acting on gene expression to increase the efficiency and accuracy of transcription and protein synthesis and processing (Figure 1). Our results are further sustained by the fact that highly expressed genes may be under greater selective constraints than lowly expressed genes, based on the negative correlation between dN, ω, and expression level and expression breadth and in both *P. glauca* and *P. abies* (Supporting Information Tables S1 and S2). This relationship between dN and gene expression has also been observed in other plant species such as *Populus* and *Arabidopsis* (Ganko et al. 2007; Ingvarsson et al. 2007). However, protein length and intron sizes increased with gene expression, which is contrary to the expectation for selection to reduce transcriptional cost, and suggests that other factors may be influencing protein length in *Picea*.

The relationship between dN and gene expression divergence may also be interpreted as a correlation between protein divergence and expression divergence, suggesting an important role of gene expression in the evolution of protein coding genes in *Picea* species. Our results showing that highly expressed genes are associated with slow-evolving protein sequences and less divergent gene expression patterns may suggest that the correlation between protein divergence and expression divergence is a result of between-gene variation in expression levels. However, our study also suggests that other gene characteristics such as tissue specificity, mutation rate (dS), gene family size, and connectivity may also contribute to explain the correlation. This suggests that the correlation between expression divergence and protein divergence is not linked to a specific gene characteristic, but instead reflects more general selective constraints, supporting recent studies in mammals (Warnefors and Kaessmann 2013).

Gene expression and codon bias

Changes in gene sequence and structure that lead to a reduction in time or energy spent in the complex processes of transcription and translation may be particularly favored in highly expressed genes (Akashi, 1994; Stoletzki and Eyre-Walker, 2007; Zhou et al. 2009; Williford and Demuth, 2012). Codon bias, defined as the preferential use of a subset of synonymous codons, is determined by a balance between drift, mutation and natural selection for optimal translational efficiency and/or accuracy (Akashi 2001). In species whose effective population size is large, such as prokaryotes, unicellular eukaryotes and some tree species (e.g *Populus*), natural selection may be the main force shaping codon usage (Ingvarsson, 2008, 2009). In our study in *Picea*, we found that a subset of synonymous codons is preferentially used in highly and widely expressed genes (Supporting Information Tables S3 and S4, and Figures S1 and S2), supporting the role of selection for translational efficiency. This significant positive correlation between gene expression and codon bias has been previously observed in several organisms including *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Drosophila melanogaster* (reviewed in Plotkin and Kudla, 2011) and more recently in *Populus tremula* (Ingvarsson et al. 2007), *Silene latifolia* (Qiu et al. 2011), *Cardamine spp* (Ometto et al. 2012), and *Tribolium castaneum* (Williford and Demuth, 2012). Codon bias may also have a role in protein export. High-frequency of non-optimal codons have been found in the signal sequences of the N-terminal regions of proteins exported via secretory pathways, which seems to be important for the correct folding of pre-exported proteins (Humphreys et al. 2000; Power et al. 2004; Palazzo et al. 2007; Zalucki et al. 2009).

Translational efficiency may also be influenced by modifications in the non-coding portion of the genes. Changes in the promoter region and alternative promoter usage during the process of transcription and post-transcriptional regulation may lead to transcripts exhibiting reduced or enhanced translational efficiency in plants and animals (Larsen et al. 2002; Hong et al. 2012; Huang et al. 2013). Although studying promoter regions was out of the scope of this work, our analysis of nine *Picea* families of different sizes, showed a greater variation of motifs (located 1kb upstream UTR sequences) in large gene families (showing a lower codon bias) than in single copy and small gene families (showing a higher codon bias). More studies are needed to understand how the variation in promoter regions affects translational efficiency in gymnosperm species. An example of the most significant motifs for three of the studied gene families can be found in Figures S8 and S9.

Gene expression and gene duplication

By using gene family size as a proxy for gene duplication, we found a strong correlation between gene expression and gene duplication in the two *Picea* species, with large gene families having, on average, a lower expression level and breadth, lower codon bias, and higher rates of sequence divergence than single-copy gene families (Table 1, Figures 2, S4 and S5). A correlation between gene duplication and gene expression variation has also been previously observed in a small-scale study of gene expression networks in *P. glauca,* and also in *Arabidopsis* and rice (Hanada et al*,* 2008; Verta et al. 2013). Our results suggest the evolution of gene family size in *Picea* is under strong functional and selective constraints.

Gene duplication may also be influenced by the position of genes in the pathway. As an example, our study of the terpenoid metabolic pathway in *Picea* suggests that duplicated genes in large gene families are more often found in the downstream branches of the pathway, supporting previous studies in *E.coli*, yeast and *Drosophila*, but differing from those in humans (Doherty et al. 2012). These results may suggest more selective constraints in upstream genes than in downstream genes, however we did not find a significant correlation between rates of sequence divergence (ω) and pathway position for any of the branches of the pathway studied.

In contrast, we found a strong negative correlation between codon bias, and gene expression breadth with pathway position in several of the branches (Figure 4), with upstream genes having a higher codon bias and being more widely expressed than genes in downstream positions. These findings further supports our previous results showing a correlation between gene expression and gene duplication; and suggest pathway position is influencing the patterns of gene duplication in *Picea* species.

Gene expression in single-copy gene families

Single copy genes may be under strong selective constraints, because an increase in copy number may unbalance their interactions with other proteins resulting in deleterious effects (Makino and McLysaght, 2010; De Smet et al. 2013). Our study in *Picea* showed that there is selective pressure to maintain genes encoded ancient conserved biological functions such as translation, DNA/RNA metabolism and nuclease activity as singletons; supporting recent studies in angiosperm species (Armisen et al. 2008; Duarte et al. 2010; Zhang et al. 2012; De Smet et al. 2013).

In addition, our results suggest that single-copy genes are widely expressed in all or most tissues and have higher expression levels, on average, than genes in multigene families (Figures 2 and 3 and Supporting Information Figure S4). Gene expression of single copy genes showed more divergence among tissues in *P.glauca* than in *P.abies*, with paralogs that are expressed in megagametophytes showing the highest expression and least diverged patterns in *P. glauca* (Figure S6). Single-copy genes also had the lowest levels of tissue-specific expression in both *P. glauca* and *P. abies*, when compared when duplicated gene families (Figure 3). The comparison between single copy orthologs, however, was limited because of the different experimental procedures used to generate the expression data and the different tissues analyzed in each of the studied species. Despite the differences, overall expression levels of single copy orthologs were significantly correlated between *P.abies* and *P.glauca* (r=0.54, P<0.0001).

We found that single-copy genes were longer, on average, than genes in multigene families, which agrees with the view that housekeeping and highly expressed genes are less compact than tissue-specific genes in humans and plants (Ren et al. 2006; Zhu et al. 2008). Based on the low number of synonymous and non-synonymous substitutions, we believe these housekeeping genes may have evolved more slowly and are therefore more conserved than genes in multigene families. This is further corroborated by a recent study based on nucleotide polymorphisms that showed that genes with high expression level and breadth were more conserved than tissue-specific genes in *P. glauca* (Pavy et al. 2013). Other examples of genes under highly functional constraint on gene duplication are the genes encoding the LEAFY transcription factor, one of the few transcription factors found as single-copy in all plant species, with the exception of gymnosperms (Baum et al. 2005). The LEAFY gene constitutes a special case of single-copy gene that has evolved new DNA binding specificities through a promiscuous intermediate, without losing its initial function (Kovach and Lamb, 2014; Sayou et al. 2014).

Gene expression divergence in large gene families

We found that large gene families in *Picea* had higher gene expression variation and higher rates of sequence divergence than genes in smaller gene families including single-copy genes. Because different copies of duplicate genes may become specialized at different times, duplicate genes may have more diversified expression profiles than single-copy genes, so that the expression patterns of duplicate genes are expected to diverge between species faster than those of single-copy genes (Gu et al. 2004). Gene expression also varied among tissues, with *Picea* paralogs having diversifying expression from tissue-specific to broadly expressed in most or all tissues. In a recent study in *Arabidopsis*, 97% of paralogous pairs showed evidence of functional diversification as a result of both neo- and sub-functionalization (Guo et al. 2013). Similarly, expression divergence of ancient paralogs leading to tissue specialization was common in *Gossypium* (Renny-Byfield et al. 2014). Successive rounds of sub- or neo-functionalization may lead to high expression variation among paralogous copies that in turn result in fitness advantages (Schmid et al. 2005). An example of this are the genes in the terpene synthase family in *Picea* which have undergone repeated rounds of neo-functionalization resulting in a broad diversity of secondary metabolites that are crucial in warding off pathogens and herbivores (Keeling et al. 2008).

Our study showed that Leucine rich repeat (LRR) and Protein kinases, which are two of the largest gene families found in conifers, have gone through recent gene family expansions in conifers and also have some of the most diversified gene expression patterns among *Picea* paralogous genes (Figure 3). Interestingly, these families also showed over-representation among genes with ω>1, suggesting they are evolving as a response to natural selection in *Picea*. Studies in *Arabidopsis* and rice suggest the kinase family has largely evolved as a response to biotic stress and has greatly expanded as a consequence of adaptation to fast-evolving pathogens (Hanada et al. 2008; Jacquemin et al. 2013). Similarly, LRR families, which have a role in disease resistance and protein-protein interactions mediated by specific amino acids, seem to be under diversifying selection in *Arabidopsis* and *Solanum* (Flagel and Wendel 2009; Slotte et al. 2011). Recent studies have suggested that different types of LRR (e.g. toll-interleukin receptor/nucleotide binding/leucine-rich repeat gene) have expanded in conifers and angiosperms (Neale et al. 2014).

By using recently available genomic resources of an understudied yet important major plant clade, our study sheds light on the role of gene expression and natural selection on the evolution of protein-coding genes in gymnosperms. This work advances our current understanding of plant science by showing the complex relationships between gene expression, codon bias, rates of sequence divergence and gene duplication in gymnosperms. Our ongoing research includes phylogenomic and molecular evolution analyses of duplicated and single copy genes in gymnosperm species.

Supplementary Material

Supplementary tables S1-S7 and figures S1-S9 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the European 7th Framework Programme under the ProCoGen (Promoting Conifer Genomic Resources) project. We thank the Norway spruce project

(UPSC, Sweden), UPPMAX computer cluster (Uppsala University, Sweden) and Ghent University Multidisciplinary Research Partnerships N2N “Bioinformatics: from nucleotides to networks”.

Literature Cited

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136: 927-935.

Alvarez-Ponce D, Aguade M, Rozas J. 2009. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes. Genome Research 19:234-242.

Armisen D, Lecharny A, Aubourg S. 2008. Unique genes in plants: specificities and conserved features throughout evolution. BMC Evol Biol 8:820.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acid Research 37: 202-208.

Barker MS. et al. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal patterns of duplicate gene retention after millions of years. Mol Biol Evol 25: 2445-2455.

Baum DA, Yoon HS, Oldham RL.2005. Molecular evolution of the transcription factor LEAFY in Brassicaceae. Mol Phylogenet Evol 37: 1-14.

Bedford T, Hartl DL. 2009. Optimization of gene expression by natural selection. Proc Natl Acad Sci USA 106(4): 1133-1138.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate- a practical and powerful approach to multiple testing. J R Stat Soc Series B 57(1): 289-300.

Birchler JA, Bhadra U, Bhadra MP, Auger DL. 2001. Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploidy syndromes and quantitative traits. Dev Biol. 234:275-88.

Birchler JA, Veitia RA. 2012. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proc Natl Acad Sci USA 109(37): 14746-14753.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16: 1679-91.

Buschiazzo E, Ritland C, Bohlmann J, Ritland K. 2012. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. BMC Evolutionary Biology 12:8.

Carvunis AR et al. 2012. Proto-genes and de novo gene birth. Nature 487: 370-374.

Chen FC, Chen CJ, Li WH, Chuang TJ. 2010. Gene family size conservation is a good indicator of evolutionary rates. Mol Biol Evol 27(8): 1750-1758.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nature 9: 938-950.

Conesa A et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.

Davis JC, Petrov DA. 2005. Do disparate mechanisms of duplication add similar genes to the genome? Trends Genet 21:548-551.

De La Torre AR et al. 2014. Insights into Conifer Giga-genomes. Plant Physiology 166:1-9.

De Smet R et al. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci USA 110(8): 2898-2903.

Doherty A, Alvarez-Ponce D, McInerney JO. 2012. Increased genome sampling reveals a dynamic relationship between gene duplicability and the structure of the primate protein-protein interaction network. Mol. Biol. Evol 29(11):3563-3573.

Duarte JM et al. 2010. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. BMC Evol Biol 10: 61.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5): 1792-1797.

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. New Phytol 183(3): 557-564.

Francino MP. 2005. An adaptive radiation model for the origin of new gene functions. Nature Gen 37(6): 573.

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res 16:805-814.

Freeling M. 2008. The evolutionary position of subfunctionalization, downgraded. Genome Dyn 4: 25-40.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol 60: 433-453.

Gallego-Romero I, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews 13: 505-516.

Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in *Arabidopsis*. Mol Biol Evol 24: 2298-2309.

Gilad Y, Oshlack A, Rifkin SA. 2006. Natural selection on gene expression. Trends in Genetics 22(8): 456-461.

Goldman N, Yang ZH. 1994. Codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725-736.

Gouzy J, Carrere S, Schiex T. 2009. FrameDP: sensitive peptide detection on noisy matured sequences. Bioinformatics 25: 670-671.

Gu Z, Rifkin SA, White KP, Li WH. 2004. Duplicate genes increase expression diversity within and between species. Nature Genet 36(6): 577-579.

Guo H, Lee T-H, Wang X, Paterson AH. 2013. Function relaxation followed by diversifying selection after whole-genome duplication in flowering plants. Plant Physiol 162:769-778.

Hahn MW, Conant GC, Wagner A. 2004. Molecular evolution in large genetic networks: does connectivity equal constraint?. J Mol Evol. 58:203–211.

Hahn MW, Han MV, Han SG. 2007. Gene family evolution across 12 *Drosophila* genomes. PLoS Genet 3(11): 2135-2146.

Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW. 2009. Adaptive evolution of young gene duplicates in mammals. Genome Res. 19:859-867.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol 148: 993-1003.

Hamberger B, Bohlmann J. 2006. Cytochrome P450 mono-oxygenases in conifer genomes: discovery of members of the terpenoid oxygenase superfamily in spruce and pine. Bioch Soc Trans 34(6): 1209-1214.

Hamberger B et al.,2009. Targeted isolation, sequence assembly and characterization of two white spruce (*Picea glauca*) BAC clones for terpenoid synthase and cytochrome P450 genes involved in conifer defence reveal insights into a conifer genome. BMC Plant Biol. 9: 106.

Heinen TJAJ, Staubach F, Häming D, Tautz D. 2009. Emergence of a new gene from an intergenic region. Curr Biol 19: 1527-1531.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915-10919.

Hong GJ, Xue XY, Mao YB, Wang LJ, Chen XY. 2012. Arabidopsis MYC2 interacts with DELLA proteins in regulating sesquiterpene synthase gen expression. The Plant Cell 24: 2635-2648.

Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. 2013. Highly recurrent TERT promoter mutations in human melanoma. Science 339: 957.

Humphreys DP. et al. 2000. High-level periplasmic expression in Escherichia coli using a eukaryotic signal peptide: Importance of codon usage at the 5' end of the coding sequence. Protein expression and purification 20: 252-264.

Huynen MA, van Nimwegen E. 1998. The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 15(5): 583-589.

Ingvarsson PK. 2007. Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula*. Mol Biol Evol 24(3): 836-844.

Ingvarsson PK. 2008. Molecular evolution of synonymous codon usage in *Populus*. BMC Evol Biol 8:307.

Ingvarsson PK. 2009. Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in *Populus tremula*. Mol Biol Evol 27(3): 650-660.

Jacquemin J et al. 2014. Fifteen million years of evolution in the *Oryza* genus shows extensive gene family expansion. Mol Plant 7(4): 642-656.

Jaillon O, Aury JM, Wincker P. 2009. "Changing by doubling", the impact of whole genome duplications in the evolution of eukaryotes. C R Biol. 332:241-253.

Jordan IK, Wolf YI, Koonin EV. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol Biol 4:22.

Jovelin R , Phillips PC. 2011. Expression level drives the pattern of selective constraints along the Insulin/Tor signal transduction pathway in *Caenorhabditis*. Genome Biology and Evolution 3:715-722.

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

Kovach JD, Lamb RS. 2014. There can be only one. Science 343: 623-624.

Larsen LK, Amri EZ, Mandrup S, Pacot C, Kristiansen K. 2002. Genomic organization of the mouse peroxisome proliferator-activated receptor β/δ gene: alternative promoter usage and splicing yield transcripts exhibiting differential translational efficiency. Biochem.J 366:767-775.

Leslie AB, Beaulieu JM, Rai HS, Crane PR, Donoghue MJ. 2012. Hemisphere-scale differences in conifer evolutionary dynamics. Proc Natl Acad Sci USA 109(40): 16217-16221.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of Ortholog groups for Eukaryotic genomes. Genome Res 13: 2178-2189.

Lynch M, Katju V. 2004. The altered evolutionary trajectories of gene duplicates. Trends Genet 20(11): 544-549.

Maere S, Heymans K, Kuiper M. 2005. BINGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. Bioinformatics 21: 3448-3449.

Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci USA 107(20): 9270-9274.

Neale DB et al. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol 15:R59.

Neme R, Tautz D. 2014. Evolution: Dynamics of de novo gene emergence. Curr Biol 24(6): R238-R240.

Nystedt B et al. 2013. The Norway spruce genome sequence and conifer genome evolution. Nature: 1-6.

Ohno S. 1970. Evolution by gene Duplication. Springer, New York, p 160.

Ometto L, Li M, Bresadola L, Varotto C. 2012. Rates of evolution in stress-related genes are associated with habitat preference in two *Cardamine* lineages. BMC Evolutionary Biology 12:7.

Otto SP. 2004. Two steps forward one step back: the pleiotropic effects of favoured alleles. P Roy Soc B-Biol Sci. 271:705–714.

Palazzo AF et al. 2007. The signal sequence coding region promotes nuclear export of mRNA. PLOS Biology 5(12): e322.

Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194-197.

Pavy N et al. 2013. The landscape of nucleotide polymorphism among 13,500 genes of the conifer *Picea glauca*, relationships with functions and comparison with *Medicago truncatula*. Genome Biol Evol 5(10): 1910-1925.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nature 12: 32.

Porth I, Hamberger B, White R, Ritland K. 2011. Defense mechanisms against herbivory in *Picea*: sequence evolution and expression regulation of gene family members in the phenylpropanoid pathway. BMC Genomics 12: 608.

Power PM, Jones RA, Beacham IR, Bucholtz C, Jennings MP. 2004. Whole genome analysis reveals a high incidence of non-optimal codons in secretory signal sequences of *Escherichia coli*. Biochemical and Biophysical Research Communications 322: 1038-1044.

Qiu S, Bergero R, Zeng K, Charlesworth D. 2011. Patterns of codon usage bias in *Silene latifolia*. Mol Biol Evol 28(1): 771-780.

R Core Team. 2014. R: A language and environment for statistical computing. http://www.R-project.org.

Raherison E et al. 2012. Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression. BMC Genomics 13:434.

Ramsay H, Rieseberg LH, Ritland K. 2009. The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. Mol. Biol. Evol 26(5): 1045-1053.

Ren XY. 2006. In plants, highly expressed genes are the least compact. Trends Genet 22: 528-532.

Renny-Byfield S et al. 2014. Ancient gene duplicates in *Gossypium* (cotton) exhibit near-complete expression divergence. Genome Biol Evol 6 (3): 559-571.

Rigault P, Boyle B, Lepage P, Cooke JEK, Bousquet J, MacKay JJ. 2011. A white spruce gene catalog for conifer genome analyses. Plant Physiol 157(1): 14-28.

Sayou C et al. 2014. A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. Science 343: 645-648.

Schmid M et al. 2005. A gene expression map of *Arabidopsis thaliana* development. Nature Genetics 37 (5).

Serres-Giardi L, Belkhir K, David J, Glemin S. 2012. Patterns and evolution of nucleotide landscapes in seed plants. The Plant Cell 24:1379-1397.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.

Slotte T et al. 2011. Genomic determinants of protein evolution and polymorphism in *Arabidopsis*. Genome Biol Evol 3:1210-1219.

Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Mol Biol Evol 24: 374-381.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution 30: 2725-2729.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. Nature 12: 692-702.

Veitia RA. 2004. Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. Genetics 104:569-574.

Verta JP, Landry CR, Mackay JJ. 2013. Are long-lived trees poised for evolutionary change? Single locus effects in the evolution of gene expression networks in spruce. Mol Ecol 22: 2369-2379.

Villanueva- Cañas JL, Laurie S, Alba MM. 2013. Improving genome-wide scans of positive selection using protein isoforms of similar length. Genome Biol Evol 52(2): 457-467.

Vitkup D, Kharchenko P, Wagner A. 2006. Influence of metabolic network structure and function on enzyme evolution. Genome Biol. 7:R39.

Wang Y, Wang X, Paterson AH. 2012. Genome and gene duplications and gene expression divergence: a view from plants. Annals of the New York Academy of Sciences 1256: 1-14.

Warnefors M, Kaessmann H. 2013. Evolution of the correlation between expression divergence and protein divergence in mammals. Genome Biol Evol 5(7):1324-1335.

Williford A, Demuth JP. 2012. Gene expression levels are correlated with synonymous codon usage, amino acid composition, and gene architecture in the red flour beetle, *Tribolium castaneum*. Mol Biol Evol 29 (12): 3577-3766.

Yang J, Gu Z, Li WH. 2003. Rate of protein evolution versus fitness effect of gene deletion. Mol Biol Evol 20: 772-774.

Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591.

Yang YH, Zhang FM, Ge S. 2009. Evolutionary rate patterns of the Gibberellin pathway genes. BMC Evolutionary Biology 9:206.

Zalucki YM, Beacham IR, Jennings MP. 2009. Biased codon usage in signal peptides: a role in protein export. Trends in Microbiology 17(4): 146-150.

Zhang N, Zeng L, Shan H, Ma H. 2012. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol 195: 923-937.

Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. Mol Biol Evol 26: 1571-1580.

Zhu J, He F, Hu S, Yu J. 2008. On the nature of human housekeeping genes. Trends Genet 24(10): 481-484.

Supplementary Material

Table S1. Relationships among studied variables based on 27,597 genes in *Picea abies.* Significant correlations are in bold numbers.

Table S2. Relationships among studied variables based on 25,384 genes in *Picea glauca.* Significant correlations are in bold numbers.

Table S3. Codon usage in *Picea abies.*

Table S4. Codon usage in *Picea glauca.*

Table S5. Over-representation of functional categories among *Picea* genes in large size gene families (>100 genes).

Table S6. Over-representation of functional categories among *Picea* genes in single-copy gene families.

Table S7. Genes with rates of sequence divergence higher than 1 (w>1), potentially under diversifying selection (on-line version).

Figure S1. Plot of axis 1 to 4 of a correspondence analysis of codon usage in 26,052 genes from *P. abies*. Black circles represent codons from highly expressed genes, and red circles, codons from lowly expressed genes.

Figure S2. Plot of axis 1 to 4 of a correspondence analysis of codon usage in 19,056 genes from *P. glauca*. Black circles represent codons from highly expressed genes, and red circles, codons from lowly expressed genes.

Figure S3. Frequency distribution of gene family size in *P.glauca* and *P.abies*. Each dot represents one orthologous gene family.

Figure S4. Gene family size showed significant associations with expression level, expression breadth, GC content at third position and codon bias (Fop) in *P. glauca*.

Figure S5. Number of synonymous nucleotide substitutions (dS) and non-synonymous nucleotide substitutions (dN) vs. gene family size.

Figure S6. Heatmap showing gene expression profiles of orthologous single copy genes in *P. glauca*. Each row is a single copy gene, and each column is a tissue. Expression levels go from 0 (lowly expressed) to 10 (highly expressed). Tissues studied included buds (b), needles (n), wood-early (we), wood-late (wl), megagametophytes (meg), phelloderm (phe), adventitious roots (ro) and embryonic cells (ec).

Figure S7. Plant terpenoid simplified pathway, modified from Ramsay et al. 2009. Numbers indicate pathway positions in each of the studied branches.

Figure S8. Sequence logo showing motif enrichment in the three most statistically significant motifs (from top to bottom) in the Promoter region of genes pertaining to two *Picea abies* gene families: A) Gene family “Picea 1012”, containing 118 genes, and annotated as MYb-like DNA-binding domain (PF00249.26, PF1392.1); and B) Gene family “Picea 1028”, containing 62 genes, annotated as Pectinesterase (PF01095.14; PF04043.10). The height of

the motif block is proportional to –log (p-value), truncated as the height for a motif with a p-value of 1e-10. Sequence logos were obtained using MEME (Bailey et al. 2009).

Figure S9. Figure shows A) sequence logo showing motif enrichment in the three most statistically significant motifs (from top to bottom) in the Promoter region of genes pertaining to *Picea abies* gene family "Picea 1236", containing 23 genes, annotated as Fasciclin (PF02469.17). Sequence logos were obtained using MEME (Bailey et al. 2009). B) Maximum-Likelihood phylogenetic tree with 100 bootstrap replicates showing all orthologous and paralogous genes for *P. abies* and *P. glauca*, for the same gene family "Picea 1236". Phylogenetic tree was constructed using MEGA 6.06 (Tamura et al. 2013).

Table 1. Kruskal-Wallis one-way analysis of variance showed significant differences between gene families of different sizes for the variables under study.

		Picea abies		*Picea glauca*	
variable	df	K	P-value	K	P-value
expression breadth	5	1411.763	<2.2e-16	791.343	<2.2e-16
expression level	5	1363.149	<2.2e-16	175.164	<2.2e-16
protein length	5	249.836	<2.2e-16	--	--
Fop	5	1595.21	<2.2e-16	931.690	<2.2e-16
CBI	5	1674.378	<2.2e-16	986.938	<2.2e-16
CAI	5	1919.456	<2.2e-16	1195.27	<2.2e-16
GC	5	675.907	<2.2e-16	411.547	<2.2e-16
GC3s	5	1498.092	<2.2e-16	950.950	<2.2e-16
dN	5	137.538	<2.2e-16	489.402	<2.2e-16
dS	5	166.499	<2.2e-16	75.580	7.03e-15
dNdS	5	32.926	3.89e-06	140.641	<2.2e-16

Table 2. Functional categories showing over-representation in large (>100 genes) and single-copy gene families in *Picea*.

GO term	GO-ID	Adjusted P-value (FDR<0.05)	No. of *Picea* genes in big families	Total no. of *Picea* genes
Large gene families				
protein kinase activity	4672	0.0000E-100	224	914
phosphotransferase activity, alcohol group as acceptor	16773	0.0000E-100	224	999
phosphorylation	16310	0.0000E-100	200	859
kinase activity	16301	0.0000E-100	229	1286
phosphate metabolic process	6796	3.0000E-100	200	941
phosphorus metabolic process	6793	4.0000E-100	200	943
transferase activity, transferring phosphorus-containing groups	16772	2.0915E-86	230	1566
protein serine/threonine kinase activity	4674	3.1894E-68	122	449
binding	5488	2.1444E-49	344	5263
nucleotide binding	166	1.4177E-46	190	1776
Single-copy gene families				
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	6139	3.6236E-59	264	1649
cellular nitrogen compound metabolic process	34641	1.1583E-57	300	2128
intracellular	5622	2.3325E-57	570	6668
nitrogen compound metabolic process	6807	1.7355E-56	302	2187
nucleic acid metabolic process	90304	1.4867E-53	213	1192
cellular macromolecule metabolic process	44260	1.3065E-51	355	3052
cellular metabolic process	44237	6.7892E-50	493	5409
gene expression	10467	1.7948E-49	181	931
macromolecule metabolic process	43170	1.9708E-48	368	3346
primary metabolic process	44238	1.2592E-47	469	5056

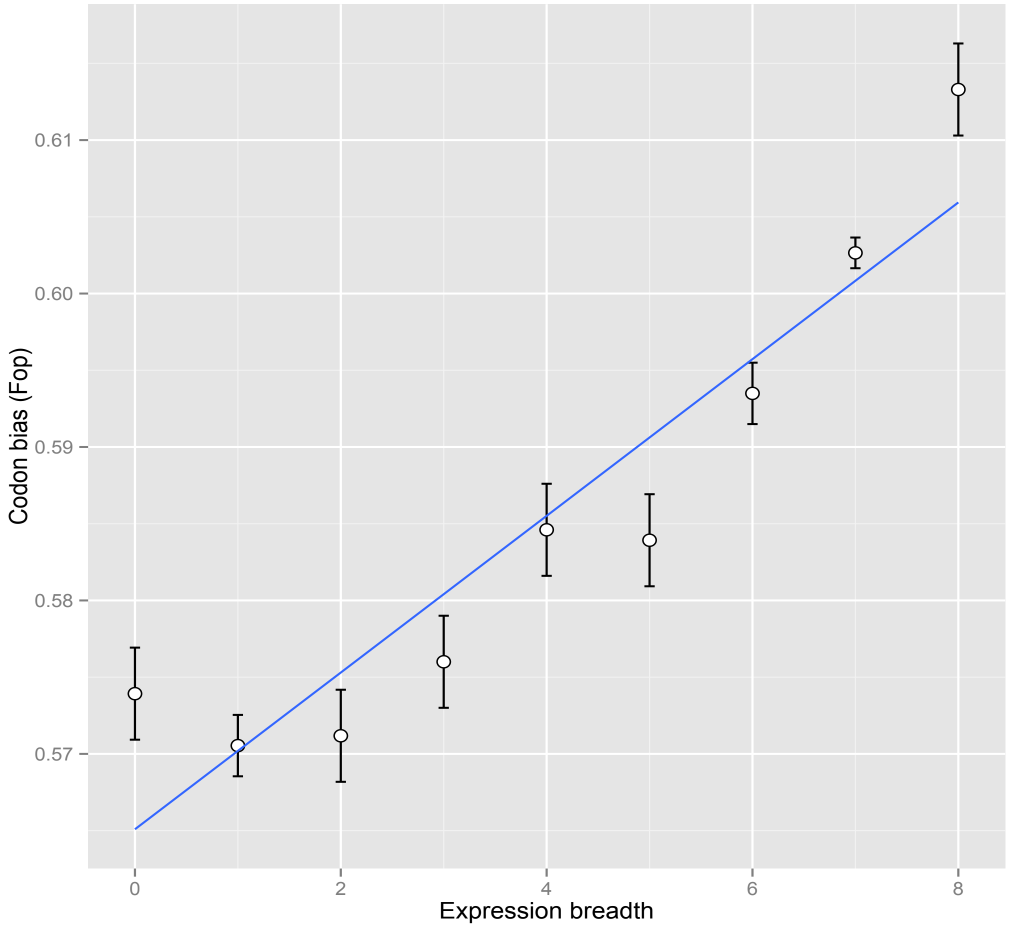

Fig. 1. Correlation between codon bias (Fop) and expression breadth in *P. glauca*. Genes were divided in groups based on total expression breadth and average Fop was calculated for each class. Standard errors of the means were plotted using vertical lines.

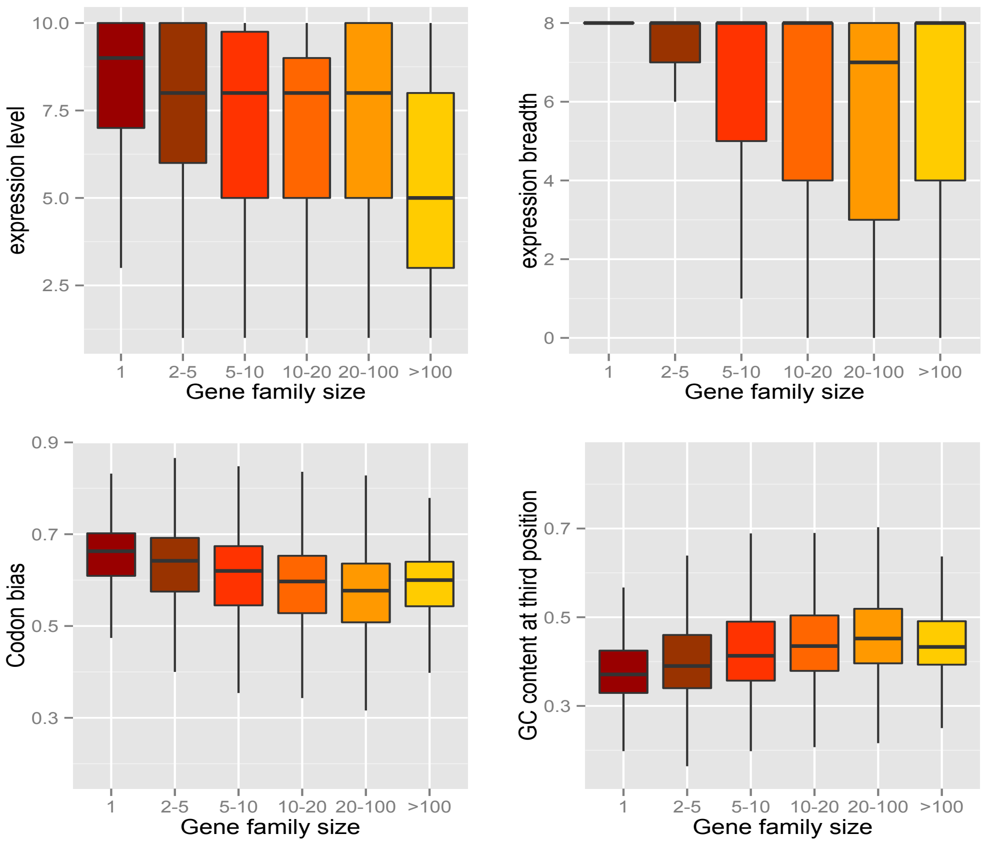

Fig. 2. Gene family size showed significant associations with expression level, expression breadth, GC content at third position and codon bias (frequency of optimal codons, Fop) in *P. abies*.

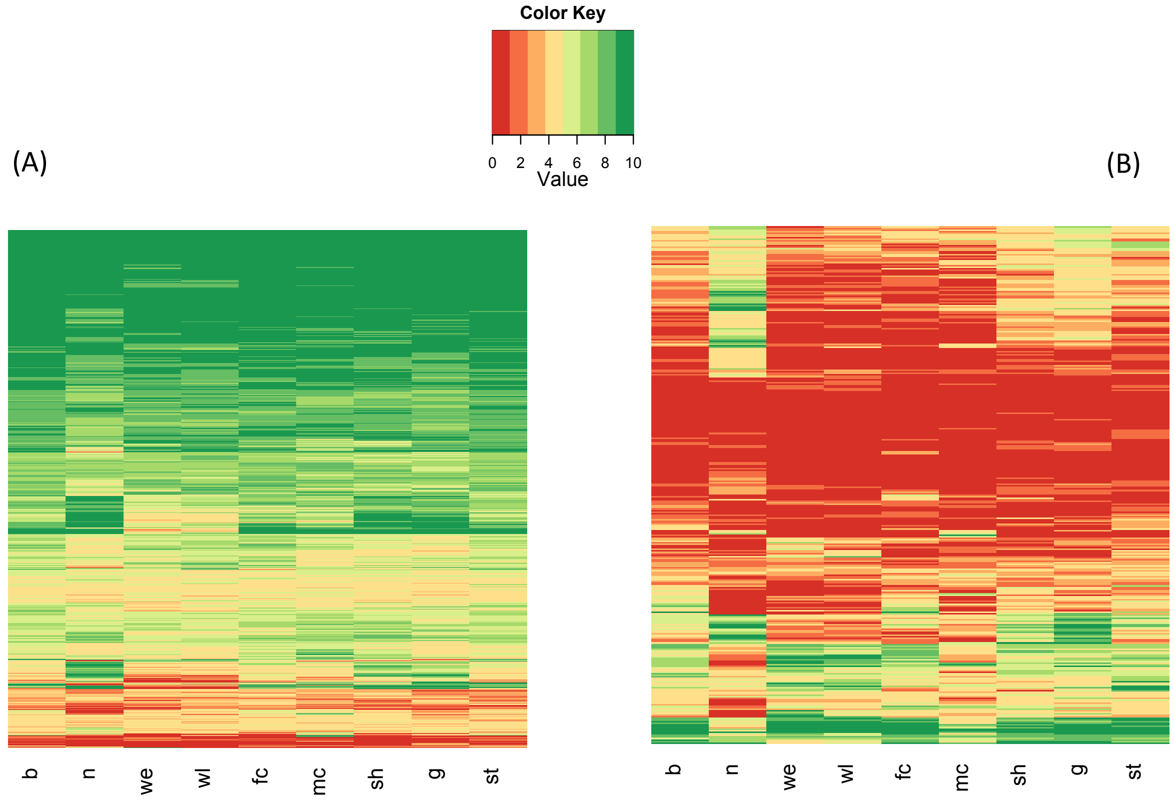

Fig. 3. Heatmaps showing gene expression profiles in *P.abies*. (A) Expression profiles of orthologous single copy genes in *P.abies*; (B) Heatmap showing diversified expression profiles among paralogous genes of the Leucine Rich Repeat (LRR) gene family. Each row is a gene and each column is a tissue. Expression levels go from 0 (lowly expressed) to 10 (highly expressed). Tissues studied included buds (b), needles (n), wood-early (we), wood-late (wl), female cone (fm), male cone (mc), shoots (s), galls (g), stems (st).

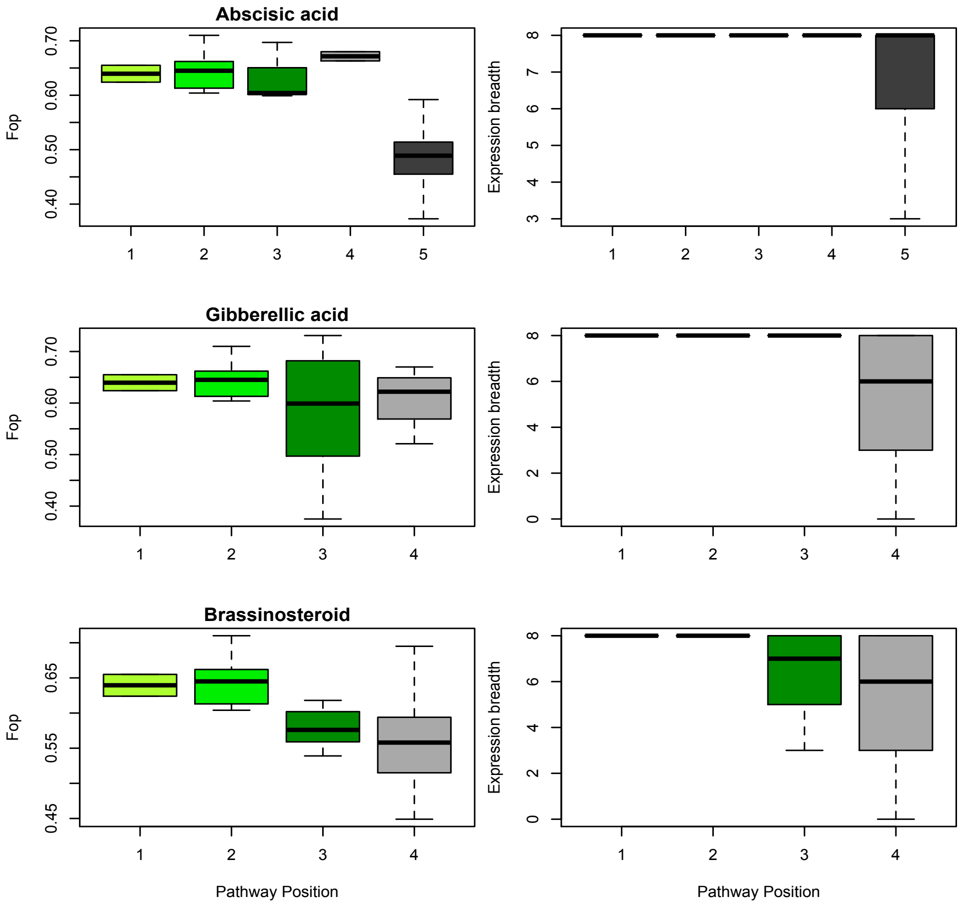

Fig. 4. Relationship between codon bias (Fop), and expression breadth with pathway position for *P.abies* in three of the four branches of the terpenoid synthesis. Plant terpenoid simplified pathway showing the pathway positions of each of the branches can be found in Figure S7.