Title: A white spruce gene catalogue for conifer genome analyses.

Authors: Philippe Rigault1,5*, Brian Boyle2*, Pierre Lepage3, Janice E. K. Cooke4, Jean Bousquet2
and John J. MacKay2

1Gydle Inc., Québec, Canada
2Université Laval, Centre for Forest Research, Institute for Systems and Integrative Biology, and
Canada Research Chair in Forest and Environmental Genomics, Québec, Canada
3McGill University and Genome Québec Innovation Center, Québec, Canada
4University of Alberta, Department of Biological Sciences, Edmonton, Canada
5Corresponding author: Gydle Inc., 1363 avenue Maguire, Québec, Québec, Canada G1T 1Z2;
   prigault@gydle.com; Phone: +1 418 907 9377 Fax: +1 418 907 9372

*These authors contributed equally

Running title: A FL-cDNA conifer gene catalogue
Abstract

Several Angiosperms plant genomes including Arabidopsis, rice, poplar and grapevine have been sequenced but the lack of reference genomes in gymnosperm phyla reduces our understanding of plant evolution and restricts the potential impacts of genomics research. A gene catalogue was developed for the conifer tree Picea glauca through large-scale EST sequencing and FL-cDNA sequencing to facilitate genome characterizations, comparative genomics and gene mapping. The resource incorporates new and publicly available sequences into 27,720 cDNA clusters, 23,589 of which are represented by full length insert cDNAs. ESTs, mate-pair cDNA clones analysis, and custom sequencing were integrated through an iterative process to improve the accuracy of clustering outcomes. The entire catalogue spans 30 Mb of unique transcribed sequence. We estimated the P. glauca nuclear genome contains up to 32,520 transcribed genes owing to incomplete, partially sequenced and unsampled transcripts, and that its transcriptome could span up to 47 Mb. These estimates are in the same range as Arabidopsis and rice transcriptomes. Next-Gen methods confirmed and enhanced the catalogue by providing deeper coverage for rare transcripts, by extending many incomplete clusters and by augmenting the overall transcriptome coverage to 38 Mb of unique sequence. Genomic sample sequencing at 8.5% of the 19.8 Gb P. glauca genome identified 1,495 clusters representing highly repeated sequences among the cDNA clusters. With a conifer transcriptome in full view, functional and protein domain annotations clearly highlighted the divergences between conifers and angiosperms, likely reflecting their respective evolutionary paths.
Introduction

Angiosperms are the most diverse and widely studied among the five major phyla of seed plants, the Spermatophyta. They are also the only group of plants with sequenced genomes, which includes the model plant *Arabidopsis thaliana* (*Arabidopsis* Genome Initiative, 2000), and cultivated plants like rice (Yu et al., 2002; Goff et al., 2002), grapevine (Jaillon et al., 2007), and poplar (Tuskan et al., 2006), among others. The lack of reference genomes in other phyla representing the gymnosperms reduces the potential for insights into plant evolution and for impacts of genomics research.

Conifers are the largest and most ubiquitous group of gymnosperms. They are woody perennials that shape many Northern hemisphere ecosystems and support large industries through the provision of wood, fibre, and energy. Sequencing conifer genomes is relevant owing to their taxonomic position, their ecological significance and their economic importance, but it has been delayed by the very large size of their genomes (i.e. 15 to 30 Gigabases; Morgante and De Paoli 2011; Murray et al., 2010; Ohri and Khoshoo, 1986).

Full knowledge of genes, namely protein coding sequences, is the most fundamental outcome of genome sequencing projects. Identifying the gene complement of any genome opens up numerous opportunities to enhance our understanding of biological functions and their evolution. Genome sequencing and large-scale cDNA analysis are the major approaches which have been deployed to predict genes and obtain transcribed sequences, respectively. Because conifer genomes are extremely large, investigations of conifers including pines (*Pinus spp.*), spruces (*Picea spp.*), Douglas-fir (*Pseudotsuga menziesii*) and Japanese cedar (*Cryptomeria japonica*) have primarily focused on transcript sequencing and analysis. Clustering of expressed sequence tags (ESTs) obtained from capillary (Sanger) sequencing has been used to infer putative unigenes or transcript sets (Allona et al., 1998; Kirst et al., 2003; Cairney et al., 2006;
Futurama et al., 2008; Liang et al., 2007; Li et al., 2009; Lorenz et al., 2006; Pavy et al., 2005; Ralph et al., 2008). Recent high throughput 454 (GS-FLX) pyrosequencing has followed a similar strategy in *Pinus contorta* (Parchman et al., 2010). While EST clustering is a cost effective way to obtain large datasets of transcribed sequences, it is notorious for producing more sequences than there are transcribed genes (Vettore et al., 2003; Riken 2001) and the outcomes may be highly variable (MacKay and Dean, 2011). Futurama et al (2008) used a mate-paired analysis to improve clustering and identify a set of full length cDNA inserts. Regardless of the experimental approaches and the goals of genome projects, full length (FL)-cDNA sequencing remains a gold standard for assisting genome-wide characterization and annotation of eukaryotic genes. FL-cDNA analysis was only reported in one conifer, i.e. *Picea sitchensis* where 6,464 cDNA were sequence finished with the use of custom primers (Ralph et al., 2008).

In this report, we aimed to produce a resource that accurately and more completely describes the transcribed gene catalogue of a major conifer tree. To reach this goal, we clustered ESTs from numerous cDNA libraries, and used mate-pair analysis of cDNA clones together with FL sequencing to obtain 23,589 unique FL Insert cDNAs (FLICs) from *Picea glauca* (Moench) Voss (white spruce). Next-Generation (Next-Gen) sequencing technologies were used to extend our sampling of the transcriptome and obtain deeper coverage of the sequence clusters. Genomic sample sequencing of 1.69 Gb also proved useful to help characterize the set of cDNA clusters. This white spruce FLIC resource has already contributed to accelerating investigations of gene expression and genetic diversity *P. glauca* (Beaulieu et al., 2011; Pelgas et al., 2011) and in *P. marianna* (black spruce) (Prunier et al., 2011) and will likely enhance the outcomes of conifer genome sequencing initiatives.
**Results**

A clustering and cDNA sequence analysis process was developed with the aim of correctly grouping multiple transcripts of the same gene. An iterative approach was used to deal with several rounds of production of Sanger EST sequences (Figure 1A). Analyses featured the full exploitation of cDNA clone information. Each cDNA clone is unique, originates from a single transcript, exists as frozen stock, and can be used to generate multiple sequence reads. This clone-level analysis allows proper use of directional 5’ and 3’ sequencing, and multiple reads from a single clone can be assembled together to produce a higher quality sequence than those of individual reads. The final outcome of this process being a Gene CATalogue of *Picea glauca*, we refer to it as GCAT-Pgl.

The *Picea glauca* gene catalogue comprises 23,589 unique full length insert cDNAs.

The gene catalogue was developed by using EST data from 42 *P. glauca* cDNA libraries (Supplemental Table 1). A total of 146,616 white spruce high quality (HQ) ESTs were produced (from libraries GQ028 to GQ041) and analyzed together with 125,556 previously-described ESTs, from the Arborea ([www.arborea.ulaval.ca](http://www.arborea.ulaval.ca)) (Pavy et al., 2005) and Treenomix ([www.treenomix.ca](http://www.treenomix.ca)) (Ralph et al., 2008) research programs (Supplemental Table 1). In total, these 272,172 ESTs represent 201,405 distinct cDNA clones (Table 1).

A critical step in the GCAT process was the production of clone sequences by orienting 5’ and 3’ ESTs and assembling all ESTs from the same cDNA into into a higher quality sequence representing that transcript. Clone sequences were further analyzed for sequence composition, base quality and cloning context in order to eliminate chimeric constructs and only use oriented high-quality sequences during the clustering step. Clone sequences were also used to assess insert size distribution, gene discovery and sequence coverage in order to direct the strategy of cDNA sequencing.
In order to represent clusters by their most informative sequence, we identified the most 5’ cDNA in each cluster and sought to obtain the Full-Length Insert cDNA sequence (FLIC) of such clones through directed and internal sequencing steps. These steps were repeated to produce FLIC sequences for most of the cluster representative clones, merging and splitting some clusters at each iteration.

The resulting gene catalogue is a grouping of cDNA clone sequences into 27,720 unique cDNA clusters, each estimated to represent a distinct gene (Table 1). A reference clone was selected for each cluster for annotation and for assessing transcript completion. In total, 23,589 of the clusters (85%) are represented by a FLIC. The RNA transcripts encompass a wide range of lengths, with several hundred above 2000 nucleotides (Figure 1B). The entire catalogue currently spans 30.15 Mb of sequence.

The majority of clusters (61%) contain at least two distinct cDNA clones and the average is 7.3 clones; however, the median is only 2 clones, 39% of the clusters contain a single clone, and 18% contain 10 or more clones (Table 2). The low sequence redundancy indicated by these cluster composition data may be attributed to the wide diversity of the cDNA libraries, as well as the efficacy of the cDNA normalization (Supplemental Tables 1, 2).

Sequence similarity with plant genomes enables extensive functional annotation

The *P. glauca* cDNA clusters were annotated based on the detection of open reading frames (ORF) and based on sequence similarity searches (see methods). Sequence similarity to conifer sequences was found for 22,255 clusters (80%) at the protein level and 23,900 clusters (86%) at the nucleotide level, based on the large transcriptome datasets of *Picea sitchensis* and *Pinus taeda* (Supplemental Table 3). Overall protein similarity with angiosperm genomes was significantly lower: i.e. 16,453 (59%) for *A. thaliana*, 16,996 (61%) for *P. trichocarpa*, 16,093 (58%) for *O.*
sativa. A total of 49 clusters were identified as contaminating chloroplast encoded sequences based on the *P. sitchensis* chloroplast genome (Cronn et al., 2008).

Considering protein level similarity to the above plants, 23,262 (84%) of the GCAT-Pgl clusters had at least one hit and 54% had a putative function assigned (Table 3). Significant similarity hits were much more frequent for clusters with long representative sequences, as nearly all of the sequences above 1000 bp (96% of 13,311) matched a known sequence, and 73% of them were assigned a putative function (Table 3B). These data show that functional annotation in conifers can be significantly improved by obtaining more complete sequence information for transcribed sequences. The internal sequencing described in this report was effective at fully capturing many long cDNA clones, at generating a high proportion of FLICs, and ultimately in improving annotation. The longest 100 cDNAs clones were all estimated to be above 2800 nucleotides, and 75% of these were completely sequenced (Supplemental Table 4A). The presence of a gap in 25% of these long clones suggested that the FLICs are still somewhat biased toward shorter transcripts. It was also observed that the most sampled sequences had a high level of functional annotation. The 100 clusters with the deepest coverage (all >102 clones, all with several FLICs) gave strong matches with plant sequences of known function predominantly encoding basic housekeeping functions such as protein synthesis and turn-over or photosynthesis (Supplemental Table 4B), with only 10% that could not be assigned a putative function.

The distribution of protein domains points to divergent evolution with angiosperms

Similarity to known protein domains was searched using Pfam-A models (Finn et al., 2010) in the 26,217 sequences predicted to contain an ORF (Table 4). The protein families and clans were nearly identical in number and largely overlapping between spruce, *Arabidopsis*, and rice (Table 5), and poplar and grapevine (Supplemental Table 6); however, both the number and the
proportion of spruce sequences (59%) matching Pfam domains were lower than observed in *Arabidopsis* (78%) and rice (69%). The identification of fewer members per family on average may result from incomplete cDNA sequences, smaller gene families, rare transcripts not being sampled or genomic DNA contaminations in the EST dataset.

The occurrence of Pfam domains revealed classes of proteins that are likely to be overrepresented, underrepresented (Figure 2) or conserved in spruce (Supplemental Table 6). The FAD/NAD(P)-binding Rossmann fold superfamily (CL0063) appeared highly conserved as it comprised 1,924 spruce sequences representing 114 domains, compared to 1,920 sequences and 115 domains in *Arabidopsis*, and 1770 sequences and 108 domains in rice. In contrast, 28 protein domains were statistically overrepresented in spruce. Many of them could be linked to metabolic processes or to stress response (Figure 2A). Most of the overrepresented stress response proteins were related to osmotic stress, such as dehydrins, ABA/WDS (abscisic acid/water deficit stress) induced protein, LEA (late embryogenesis), AWPM-19-like family (involved in cold tolerance). The metabolic process proteins were very diverse. They included few secondary metabolism enzymes (e.g., O-methyltransferases) and few carbohydrate metabolism enzymes (e.g., glycosyl hydrolases family 16, found in xyloglucan endo-transglycosylase).

Proteins domains from type I transposable elements (retroelements) were overrepresented, whereas type II transposable elements were underrepresented, which is consistent with recent literature (Morgante and De Paoli, 2011). Specific domains of unknown functions (DUFs) also fell into both categories; while many other DUFs were variable or lineage specific among angiosperms.

The underrepresented protein domains (41 in total) included large protein families involved in cellular processes (e.g., F-box and associated FBA1 and FBA3 domains involved in ubiquitination), signalling (e.g., protein kinases) and transcription (Figure 2B). Underrepresented
domains were also part of different classes of cell wall related proteins (e.g. Extensin-like region; Glycosyl hydrolases family 28 Plant invertase/pectin methylesterase inhibitor), a few biotic interaction proteins (lectin binding), and the S-locus glycoprotein family (36 to 198 loci in angiosperms) which appears to be nearly absent from *Picea*.

The total number of *P. glauca* sequences containing one of the 39 Pfam domains found in plant transcription factors (TFs) and the representation of most of these domains was approximately half that of *Arabidopsis*, rice, poplar, and grapevine (Supplemental Table 7). The B3 DNA binding domain found ABI3/VP1 transcription factor was the most underrepresented in spruce, with 15-20% of the number of angiosperm sequences. The only overrepresented TF domain in spruce was that of histone-like TFs (CBF/NF-Y) and archaeal histones (PF00808) found in plant Hap proteins. In contrast, the relative representation of each TF domain among the total TF domains was largely conserved between spruce and angiosperm plants (Figure 3).

**The size of the spruce transcriptome may be estimated from FL inserts and mRNAs.**

We used the clone completion (Table 1) and gene annotation information to assess the mRNA status of *P. glauca* cDNA clusters (Table 4). Full-length mRNA status was achieved when a representative transcript sequence contained a complete protein coding sequence (CDS) as well as some 5'UTR and 3'UTR sequence. We assessed the CDS status of GCAT gene sequences as follows: first, sequences with BLASTX similarity to a reference protein in Arabidopsis, rice, poplar, vine or SwissProt were classified in regard to length and similarity of the match, into three classes: complete confirmed (class 1), complete predicted (class 2), and incomplete (class 4). Genes without similarity to a reference protein were classified, according to whether an ORF could be predicted or not, and the features of the predicted ORF (classes 3, 5 and 6).

Based on these criteria, 11,608 (49.2%) of the FLICs were estimated to contain a complete CDS (in classes 1-3). The average mRNA size was deduced from classes 1 and 2 to be 1396
nucleotides, with an average CDS of 871 bp (290 predicted amino acids), compared to a CDS of 1021 bp in Arabidopsis and 1096 in rice (Table 5), consistent with the sampling bias toward shorter FLICs and indicating that spruce has longer 3' and 5' UTRs.

The GC content of ORFs in class 1 sequences was determined as 46.6 %, which is very similar to Arabidopsis (45%) (Alexandrov et al., 2006) and significantly lower than maize (58%) (Alexandrov et al., 2009). The 47% GC content of the 3rd codon position was also similar to Arabidopsis, whereas maize (and other monocots) has a much higher and more variable GC content in the 3rd position (Alexandrov et al., 2009).

The P. glauca gene catalogue contains 27,720 cDNA clusters, representing 30.15 Mb of unique sequence. The majority of clusters (23,589) are represented by a FLIC but 12,736 (54%) of them lacked a complete CDS and 1509 of them completely lacked any ORF. We used the sequence completion status (Table 1) and CDS status (Table 4) to estimate the full size of the P. glauca transcriptome sampled to date (Table 4). The transcriptome represented in the P. glauca gene catalogue was predicted to span 40.75 Mb, i.e. 0.20% of the P. glauca genome reported to be 19.8 Gb by Ohri et al., (1986). We thus estimate that 11.60 Mb as missing (Table 4).

**Next-Gen sequencing enhances and extends the gene catalogue.**

The Sanger based gene catalogue was validated and augmented by cDNA analysis with 454 sequencers (GS-FLX, Roche) and by RNA sequencing (RNA-Seq) with GAII (Illumina) sequencers (See methods). For libraries and samples sequenced, see Supplemental Table 8. A total of 7.42 M GS-FLX and 59.5 M GA (RNA-Seq) high-quality sequences were obtained and mapped to the cDNA clusters.

The Next-Gen sequences provided independent validations of the Sanger-derived sequences (Table 2). They were obtained in part from separate libraries and tissue samples (Supplemental Table 8). Low coverage of cDNA sequences is a common feature of Sanger-based
EST catalogues due to factors such as limited sampling of rare transcripts and partial sequencing. A total 93% (25,853) of the *P. glauca* clusters were matched by at least one Next-Gen sequence, with 98% (16,496) for multiple-clone clusters and 86% (9,357) for the singletons (not shown). Figure 4 illustrates how GS-FLX sequences directly matching cDNA clones were useful to confirm sequence structure in areas of low coverage, such as 5' ends of transcripts (e.g. Figure 4A) or to help delineate unspliced introns (Figure 4B). The Next-Gen sequences also extended many incomplete clones and clusters. A total of 2.9 Mb was added through extensions at 5' and 3' ends, and by filling gaps in clusters with non overlapping 5' and 3' reads (Table 6).

The discovery of genes not represented in the Sanger-based gene catalogue was assessed in the Next-Gen sequencing data using several approaches. First, we determined the proportion of Next-Gen sequences that mapped to existing clusters based on direct sequence overlap. Depending on the library, 87% to 92% of the GS-FLX reads mapped to the GCAT-Pgl clusters, except for library GQ043-45, which gave 37% representation, indicating likely contamination: we were able to identify the contaminating sequences as predominantly derived from genomic *P. glauca* DNA (Supplemental Table 3). The GA-II reads were represented in 80% of the GCAT-Pgl clusters. The lower frequency of matches observed with GA reads is consistent with the lack of 3' bias in RNA-Seq compared to available *P. glauca* ESTs, and the shorter sequence lengths. Consistent with this expectation, the GA-II match rate increased to 87% after the Sanger-based cDNA clusters were extended with the directly matching GS-FLX reads.

Second, the orphan GS-FLX sequences (not mapped to a cDNA cluster) were assembled and produced 3,129 contigs longer than 400 bases and containing two or more reads. Approximately 35% (1,146) of these orphan contigs produced a strong match to a Pinaceae putative unique transcript (PUT) from PlantGDB (Table 7). The number of similar sequences (>85% identity) ranged from 88 to 567 depending on the species. Third, the contigs and HQ
singletons obtained from orphan GS-FLX data that were confirmed by RNA-Seq were also checked for similarity to plant proteins (TBLASTX). A total of 20,549 contigs ranging from 64 to 586 bp and spanning a total 5.0 Mb were thus identified as putative transcriptome sequence (not shown); however, the majority of the best matches were against proteins also matched by Sanger-based cDNA clusters, indicating that the set of orphan sequences likely represent missing sequence for many of the same genes as well as a smaller fraction of unsampled genes.

**Genomic sample sequencing identifies repeated sequences in the cDNA dataset.**

Sample sequencing of the *P. glauca* genome based on 4.91 million GS-FLX reads totalling 1.69 Gb and representing a genome coverage of 8.5% was used to further characterize the gene catalogue and specifically to detect putative repetitive sequences. Retroelements and non characterized repetitive elements were estimated to make up 75% of the 18 Gb *Picea abies* genome (Morgante and De Paoli, 2011). Over 20,000 complete copies of the gymny retroelement were estimated to be in the 22 Gb *Pinus taeda* genome (Morse et al., 2009). Therefore, identifying repetitive sequences is useful to assess genomic contamination. A total of 1,490 cDNA cluster sequences produced strong alignments with numerous genomics sequences, and were represented by only one or two ESTs that lacked similarity to plant proteins (Table 8). They included 1,195 clusters recognized by 40 or more genomic sequence reads (Table 8) and 223 clusters with more than 1,000 significant alignments to genomic sequence data (not shown). The majority of these sequences contained one or several short ORFs (68% in class 5) or had no ORF (15% in class 6) (Table 4). Only 7% had an ORF containing a Pfam domain, and most of the matches were weakly similar to retroelement proteins such as reverse transcriptases, RNase H, integrases. Despite these observations, a majority of the sequences were similar to putatively transcribed sequences from at least one other conifer tree, such as *Pinus taeda* or *Picea sitchensis*; therefore some of them may represent *bona fide* transcripts from retroelements or from unknown high copy genes.
Discussion

Development of a FL-cDNA transcriptome resource for conifer genomics

The development of a *P. glauca* gene catalogue creates a resource that will enhance our understanding of the conifer transcriptome and facilitate other genomic investigations including genome sequencing and comparative genomics. Large-scale FL-cDNA sequencing represents a proven approach to most effectively support these objectives. Previous cDNA analyses in conifers used large-scale clustering of end reads (5', 3', or both) to infer putative unigenes or transcript sets, primarily from *Pinus taeda* (Allona et al., 1998, Kirst et al., 2003, Cairney et al., 2006, Lorenz et al., 2006, Liang et al., 2007), *Pinus radiata* (Li et al., 2009), *Picea* spp. (Ralph et al., 2008), *Picea glauca* (Pavy et al., 2006), and *Cryptomeria japonica* (Futurama et al., 2008). The only exception to this approach was the FL analysis of 6,464 *Picea sitchensis* cDNA clones (Ralph et al., 2008). Futurama et al. (2008) used a mate-paired analysis to improve clustering and identify a set of full length cDNA inserts. Mate-pair analysis is routinely used in FL-cDNA sequencing; for example, it contributed in the recent characterization of the maize genome (Soderlund et al., 2009). Here, the *P. glauca* clone analyses integrated mate-pairs (non overlapping 5' and 3' reads) from the same cDNA insert and internal sequencing in an iterative process to improve the accuracy of clustering outcomes. The clone analysis helped to verify error models in our dataset (not shown), and full-length analysis helped to verify sequence orientation and to improve sequence annotations. Each cluster was represented by a single cDNA insert rather than a consensus sequence with the aim of minimizing potential informatics artefacts, and of reducing ambiguous or under clustering.

The advantages associated with FL-cDNA have been abundantly illustrated for aiding genome assembly and annotation (e.g. Haas et al., 2002) and for characterizing transcriptomes (e.g. Alexandrov et al., 2006). Numerous EST assembly methods and integrated platforms have been developed and optimized (e.g. Forment et al., 2008; Lee et al., 2007). However, by their
nature, EST assemblies typically overestimate the number of genes in transcriptome samples. The ESTs from the same locus may fail to assemble owing to several factors including alternate splicing, multiple polyadenylation sites, sequence polymorphism, and sequencing errors. Levels of redundancy after EST assembly have been estimated to range from 20% to 22% even under optimal conditions (Riken 2001; Vettore et al., 2003). The outcomes of EST clustering also vary depending on the objectives and the analytical approach, especially in uncharacterized genomes. For example, conifers with large EST datasets (P. taeda, P. glauca, P. sitchensis) produced Unigene builds ranging from 18,709 to 19,825 (NCBI, Unigene assemblies accessed in Dec 2010) but gave assemblies of putative unique transcripts (PUTs) of 31,054 to 72,829 (PlantGdb; Dong et al., 2004). Sequence data from multiple species also inflates the number of unique entities resulting from sequence assemblies, as observed with the 75,758 tentative consensus sequences (TCs) for spruces (DFCI Spruce Gene Index, release 4.0; Quackenbush et al., 2001) and the 69,968 pine genus TCs (DFCI Pine Gene Index, release 8.0).

Until quality drafts of conifer genomes become available, FL-cDNAs provide us with models that most accurately represent the gene coding portion of the genome. As such, they are essential to facilitate downstream applications of genomics. For example, tracking sequence polymorphisms such as SNPs has become central to understanding the relationship between genotype and phenotype targeted by association genetics or other approaches (Namroud et al., 2008; Eckert et al., 2009; Beaulieu et al., 2011; Neale and Kremer 2011). Without adequate reference sequences, allelic and paralogous variants may be confounded and lead to erroneous genomic inferences and representations (Ritland et al 2011), which is a particularly acute problem in undomesticated and highly heterozygous organisms such as conifer trees. Applications such as RNA-Seq and short read clustering will also be facilitated by the availability of accurate cDNA models to be used in reference assembly methods.
Next-Gen sequences from two types of technologies (GS-FLX and GAII) were used to confirm and annotate Sanger clustering results, to obtain deeper coverage, and to extend the transcriptome sampling. Recent studies involving uncharacterized genomes such as lodgepole pine (*Pinus contorta*) (Parchman et al., 2010) and European oaks (*Quercus petraea* and *Q. robur*) (Ueno et al., 2010) have shown that GS-FLX is efficient for transcriptome sequencing but clustering methods are prone to generate exceedingly large numbers of unique sequences upon assembly, particularly when multiple individuals are pooled for sequencing. Even when integrating Sanger and GS-FLX reads in a hybrid analysis, the assembly of oak sequences yielded a minimum of 116,000 unique sequence elements (Ueno et al., 2010). Assembly of GS-FLX data alone generated short contigs averaging at 353 bp in *Eucalyptus* (Noveas et al., 2008; de novo assembly) and 500 bp in pine (Parchman et al., 2010; combined de novo and reference assembly).

Our results indicate that sequences under 1,000 bp are suboptimal for annotation of conifer genes. Kumar and Blaxter (2010) tested several transcriptome assembly tools and approaches, and reported that the best results were obtained in a stepwise use of two different assembly programs. Similarly, for *P. glauca*, we mapped Next-Gen sequences to existing cDNA clusters, and only the unmapped sequences were used for new assemblies. With this approach, assembly of orphan spruce sequences produced over 20,000 unique but relatively short sequences that did not directly overlap with the cDNA clusters.

Genomic sample sequencing of *P. glauca* (8.5% of the genome) enabled the annotation of 1% of the clones as putative repetitive sequences. These sequences accounted 5.4% of the clusters owing to their uniqueness in the dataset. These sequences represent potential genomic DNA contaminants although some of them could represent *bona fide* transcripts. A recent report of GS-FLX *Pinus contorta* transcriptome sequencing identified approximately 6% of the sequences as putative retrotransposons (Parchman et al., 2010). The authors indicated that they
were transcribed sequences found at a much higher frequency than in previous reports. Our observations lead us to suggest they could also represent genomic DNA contaminants, which could be more difficult to account for when using Next-Gen methods, especially when poly-A tail homopolymers are avoided by using random hexamers for reverse transcription.

**A conifer transcriptome in full view**

Estimating the number of transcribed genes in a conifer genome is relevant for many downstream investigations and has been the object of interest and debate (reviewed in MacKay and Dean 2011). The GCAT-Pgl catalogue is based on the sampling of over 200,000 cDNA clones from 42 diverse cDNA libraries. These sequences were grouped into 27,720 cDNA clusters spanning 30.15 Mb. The removal of 1,545 putative contaminating sequences (genomic and chloroplastic) gives 26,185 clusters spanning 29.35 Mb of the nuclear transcriptome. We estimated that 11.60 Mb were missing from the representative sequences containing a CDS or predicted ORF, owing to incomplete clones or partial sequencing. The transcriptome sampled by Sanger sequencing was thus predicted to be up to 40.75 Mb (Table 4), which is 2% more and 17% less than the gene space in *Arabidopsis* and in rice, respectively (Table 5). Taken together, these predictions suggest that several genes and transcripts were most likely not represented in the *P. glauca* catalogue based on Sanger sequences.

The data reported here support different approaches to assess the overall size of the *P. glauca* transcriptome and thus estimate the number of missing sequences. First, the number of transcribed genes may be estimated based on the number of matching sequences in two independent sets of transcripts. This approach was used for humans (Ewing and Green 2000) and maize (Alexandrov et al., 2009), and may be applied to *P. glauca* because the ESTs represented two independent sets of cDNA libraries (GQ001-041 and WS001-034; Supplemental Table 1). The number of transcribed genes (m) was estimated as 30,214 excluding putative
genomic contaminants, and 32,720 when they were included (see methods). These estimates indicate that 83-85% of the expressed P. glauca genes have been sampled by Sanger sequencing. The 4,029 to 4,800 transcribed genes not represented could span 5.50 to 6.55 Mb (based on 1,365 bp of class 1 sequences), so the total transcriptome could be as large as 47.30 Mb or 0.24% of the Picea genome.

Second, 7.78 Mb of new sequence data were discovered through analyses of Next-Gen sequence data that matched with P. glauca cDNA clusters or with known plant protein sequences. When combined, Sanger and Next-Gen sequencing data produced a total of up to 37.93 Mb of unique transcribed sequence. The data are consistent with our estimates of the overall transcriptome size. Future assemblies of transcript sequences integrating these different types of sequence data will help to improve overall transcriptome coverage and bring an even higher resolution to the number of transcribed genes.

**Insights into conifer genome structure and evolution**

Our highest estimate of 32,720 transcribed genes in the P. glauca genome is larger than Arabidopsis (Arabidopsis Genome Initiative, 2000) and is nearly identical to maize at 32,450 genes (Schnable et al., 2009). However, it remains at the lower end of the spectrum observed among angiosperms. From an evolutionary perspective this finding is not surprising and, given the high degree of genome and sequence conservation among the Pinaceae (Ritland et al., 2011), it is likely to apply to other members of the family. Early studies hypothesized that the Pinaceae had many more genes (Kinlaw and Neale 1997) ranging as high as 225,000 genes (Rabinowicz et al., 2005); however, evidence has not yet been obtained for genome-wide gene amplifications or polyploidization events that could result in such large numbers of transcribed genes (Morgante and De Paoli 2011). In fact, it appears that gene contents may have been overestimated for reasons that have been uncovered in recent reports. On the one hand, conifer genomes contain a
substantial fraction of single copy sequences (over 15%, i.e. 3 Gb) including more or less degenerate retroelements which may appear as unique sequences in the genome (Morgante and De Paoli 2011). On the other hand, the fraction of the genomic sequences that share significant similarity with genes appears exceedingly large (Magbanua et al., 2011). It was reported to roughly 3% in *P. abies* (Morgante and De Paoli, 2011), which is close to 600 Mb (our estimate), a size that is larger than most plant genomes sequenced to date. Such a finding appears to be at odds with genome evolution theory presented by authors such as Linch (2007) but may be explained in part by the rather large number of pseudogenes; for example, Garcia-Gil (2008) found that pseudogenes were much more numerous than intact coding genes in the case of the phytochrome gene family in Scots pine. Large numbers of pseudogenes could have contributed to inflating indirect estimations of gene content (Rabinowicz et al., 2005) and to producing complex genomic hybridization patterns (Kinlaw and Neale 1997).

Angiosperms genome sequences have been used to study the evolution of genome structure and of gene families. For example, the evolution of most transcription factors (TFs) was shown to involve two major bursts of gain/expansion, coinciding with the water-to-land transition and the radiation of flowering plants (angiosperms) (Lang et al., 2010). Given that their radiation largely predates that of flowering plants and their apparent slow rate of morphological evolution (Gernandt et al., 2011) and their conserved genome macro-structure (Ritland et al., 2011), it stands to reason that gymnosperms may have fewer TFs in their genomes than angiosperms. Conifers were previously shown to have fewer members overall and to lack members in specific subfamilies for TF families such as Knox I and HD-zip-III, and to have more numerous recent gene duplications in closely related subfamilies (Guillet-Claude et al., 2004; Coté et al., 2010) and in R2R3-MYBs (Bedon et al., 2010), among others. Consistent with the model presented by Lang and colleagues (2010), the occurrence and distribution of TF protein
domains in the cDNA resource described here strongly suggests that lower overall representation applies to the majority of TF families (Figure 3). Our data also suggest that the distribution of sequences among TF families is largely conserved between angiosperms and gymnosperms (conifers). These hypotheses may be fully tested through phylogenetic analyses and verified once a quality draft genome is obtained.

Conclusions

The cDNA resource described here will enable a host of downstream applications ranging from functional and evolutionary investigations to population and association genetic studies. Approximately 85% of the cDNA clusters are represented by fully sequenced inserts, and we estimated the gene catalogue represents up to 85% of the transcribed genes in the P. glauca genome. This coverage and the FL-cDNA approach have allowed us to begin describing the extent of conservation and divergence in transcribed genes between a representative gymnosperm (conifer) and angiosperms, more systematically than previously possible. More detailed investigations of protein families may help to uncover metabolic and regulatory mechanisms relevant to the unique biology of conifers and gymnosperms. The extent of the sequence data available for P. glauca will also be instrumental to leverage the throughput of Next-Gen sequencers for whole genome, exome analysis or RNA-Seq.

Methods

Sanger and Next-Gen sequencing of cDNA, RNA and genomic DNA

Nucleic acid isolation methods and cDNA libraries

For cDNA libraries, tissue sampling was as described (Bedon et al., 2007)(for tissue descriptions see Supplemental Table 1). RNA extractions were performed according to Chang et al. (1993). Total RNA was ethanol precipitated and shipped at room temperature to Evrogen (Moscow,
Russia) for library synthesis. For libraries GQ028 to GQ041, complementary DNA (cat. # CS-040) and normalized cDNA (cat. # CS011-2B) libraries were constructed and directionally cloned into pAL17.3 by Evrogen using the SMART procedure (Zhu et al., 2001). Normalization of amplified cDNA was performed using the DSN normalization method (Zhulidov et al., 2004). The normalization efficiency was monitored by qPCR based on four reference genes (Supplemental Table 2). For 454 sequencing, four uncloned normalized cDNA libraries for Next-Gen sequencing (cat. # CS010) were constructed by Evrogen (Supplemental Table 8). For RNA-Seq, total RNA was extracted from secondary xylem and secondary phloem of *P. glauca* as described (Bedon et al., 2007). The samples were created by pooling the RNA from 20 three year-old open-pollinated trees (Pavy et al., 2008a). For genomic sequencing, DNA was isolated using a the NucleoSpin 96 Plant II extraction system (Macherey-Nagel, Duren, Germany) following the manufacturer’s instructions from the same *P. glauca* genotype as library F7701; quality analysis was as described (Pavy et al., 2008a).

**Sequencing methods for cDNA, RNA and genomic DNA**

The Sanger cDNA sequences reported here [i.e. not reported previously by Pavy et al., (2005) or Ralph et al., (2008)] were generated at the McGill University and Génome Québec Innovation Centre (Montreal, Canada). Briefly, the cDNA libraries were plated on LB-agar containing ampicillin and Xgal. White colonies were picked into 384 well plates containing 2X LB and 50% glycerol using a QPix2XT (Genetix, Beaverton, USA) and grown overnight. Plasmid DNA was isolated using an alkaline lysis procedure and PALL filtration plates (PALL, Port Washington, USA). Sanger sequencing was produced on 3730xl DNA Analyzers with the Big Dye Terminator sequencing kit (Applied Biosystems, Foster City, CA, USA). The full length sequencing of clones was performed as described above using a combination of 5′ (vector), 3′ (vector or Poly[A]), and internal (clone specific) oligonucleotides. The 454 reads were obtained on GS-FLX sequencers
(Roche, San Diego, CA, USA) from different service providers using standard or Titanium protocols (Supplemental Table 8). Normalized amplified cDNAs obtained from Evrogen or genomic DNA were fragmented and used as starting material for shotgun library synthesis, library amplification onto beads by emulsion PCR and sequencing following manufacturer recommended procedures. For RNA-Seq, total RNA was quantified by Bioanalyzer (Agilent, Santa Clara, CA, USA), and then submitted to the standard protocol for RNA-seq library construction and sequencing on the Genome Analyzer (GAIIx) (Illumina, San Diego, USA).

Analysis of Sanger cDNA sequencing data

Production of high quality ESTs and cDNA clone sequences

Base calling and quality scoring was performed by Phred (Ewing and Green, 1998). Vector identification was performed by cross-match and followed by vector-specific detection of cloning features, adapter sequences, and Poly(A) tails, which were clipped to retain six adenines. The resulting insert sequences were submitted to a quality filter requiring 100 consecutive high-complexity bases above Phred score 20, and then to an identification step for bacterial and fungal contamination removal.

Clone sequences were produced by orienting ESTs and assembling all ESTs from the same clone using CAP3 (Huang and Madan 1999), followed by an analysis of the cloning context (gathered from the EST production step) to eliminate chimeras and determine the sequence completion status of the clone (5'-end, 3'-end, 5'-3' with a gap, full-length insert).

Iterative Clustering of cDNA clone sequences

Clones were grouped into gene clusters using an iterative clustering scheme. First, pairwise sequence comparisons using Nuclear (Gydlé) were used to group clones having 97% or more sequence identity over 100 bases of their clusterable regions and no large region of dissimilarity (defined by 150 bases with identity between 85% and 94%). Second, we proceeded with several
iterations of clone addition and clone sequence modification as a result of adding sequence reads to existing clones. These modifications resulted in the merging and splitting of clusters based on human and automated curation steps.

**Cluster representative sequences and FL sequencing**

The selection of the most 5’ cDNA was based on pair-wise sequence alignments using Nuclear (Gydle) between clone sequences. Many clusters had a selected clone that was already complete, owing to the length of Sanger reads (800 bases) and the large number of clones with 5’ and 3’ reads (enough to span over 1.5kb). The clones with incomplete sequence were re-racked and submitted to end-sequencing for verification and extension, then to internal sequencing steps with the use of custom internal primers if needed. Each round of sequence completion was followed by clone reassemblies and a cluster analysis, merging or splitting clusters and selecting an alternative clone whenever necessary. In total, 11,421 clones were thus selected, 18,350 additional sequences were produced, resulting in a FLIC for 9,812 of these clones, bringing the number of clusters to 27,720 and the number of clusters with a FLIC to 23,589.

**Functional annotations**

Gene sequences (i.e cluster representative clone sequences) were characterized as follows. Protein-level similarity searches were performed with BLASTX (E-value <= 1e-10) against proteins of Arabidopsis thaliana (TAIR version 9), Oryza Sativa (RefSeq version 43), Populus trichocarpa (version 43) and Vitis vinifera (RefSeq version 43) and with TBLASTX (E-value <= 1e-10) against representative transcript sequences of *Pinus taeda* (UniGene build 11) and *Picea sitchensis* (UniGene build 15). Nucleotide level searches were also performed using Nuclear (Gydle) against EST sequences of *P. taeda* (identity >= 85%, 186k ESTs) and *P. sitchensis* (identity >= 95%, 293k ESTs), the different identity criteria reflecting the average transcript similarity with *P. glauca*.
For ORF identification in sequences not found to contain a complete CDS and possibly containing frame shifts, gene sequences were translated in all 6 frames and the resulting ORFs were screened as follows. The first ORF on the plus strand was clipped to the first methionine if it appeared in the first 300 bases of the sequence. The ORFs on the plus strand were retained if the longest ORF was larger than N amino acids (N=40), in which case all other ORFs of length N/2 or longer were also kept. On the minus strand, the longest ORF was retained only if it was longer than all ORFs on the plus strand.

Clusters were assigned protein family domains in two steps. First, ORFs from gene sequences were assigned directly against Pfam (Pfam-A, version 24.0) using HMMER (v3.0, domain-specific thresholds). Second, indirect Pfam assignations were made for those genes that gave no direct Pfam match but had a TAIR homolog which both qualified the gene sequence as an incomplete CDS and itself contained a Pfam domain.

_Estimation of the number of transcribed genes in the _P. glauca_ genome_

The number of transcribed genes was estimated based on the method developed by Ewing and Green (2000) to estimate the number of human genes. This method was also applied to estimate the number of transcripts in the maize genome (Alexandrov et al., 2009), such that:

\[ m = \frac{n_1}{N} \left( \frac{n_2}{N} \right) N = \frac{n_1n_2}{N} \quad \text{(equation 1)} \]

where,

- \( m \) is the total number of genes
- \( n_1 \) is the number of genes in set 1
- \( n_2 \) is the number of genes in set 2
- \( N \) is the number of genes shared in both sets

The sequence datasets (set 1 GQ001 to GQ041; set 2 WS001 to WS034) represent two independent samplings of the _P. glauca_ transcriptome derived from separate biological materials.
and cDNA libraries (Supplemental Table 1). The estimate of m (total number of transcribed genes) was based on 27,720 clusters, or 26,185 after removal of putative genomic and chloroplastic contaminating sequences (see results). The two sets represent 20,849 \( (n_1) \) and 17,215 clusters \( (n_2) \) respectively, with 11,879 clusters in common \( (N) \).

**Analysis of Next-Gen cDNA and RNA sequence data**

**Sequence processing and mapping onto cDNA clusters**

Next-Gen reads from 454 and RNA-Seq were base called with manufacturer software to produce FASTQ files. Sequence filtering was performed using Gydle software, by detection of cloning features, adapter sequences, and Poly(A) tails (clipped to retain six adenines) followed by composition filtering. Sequences containing 50 or more high-complexity bases of high quality (90% scores over Q20) were kept. After identification and removal of bacterial, chloroplast and ribosomal sequences, the resulting ESTs were aligned to cDNA cluster sequences using Nuclear (Gydle Inc.), requiring a minimum of 95% identity over 50 bases for assignment to their best hit.

**Clustering and analysis of orphan GS-FLX sequences**

The 454 sequences that could not be mapped to a GCAT cluster (371,372 reads; 89 Mb in total) were assembled using the genomic mode of Newbler (version 2.3, Roche); 183,369 reads were assembled into 11,032 contigs and 179,437 singletons. First, the contigs were used as queries to search for transcribed conifer homologs using BLAST (E value < 1e-10) in PlantGDB, i.e. assemblies 157a for *P.taeda* and *Picea engelmannii X P.glauc*, 175a for *P.glauc*, *P.sitchensis Picea abies*, and *Pinus contorta*; and 177a for *Pinus pinaster*. Second, both the contigs and singletons were used for sequence similarity matches against *A. thaliana, O. sativa, P. trichocarpa, and V. vinifera* using TBLASTX, as described above.

**Analysis of genomic sample sequencing data**
Reads were trimmed using the default settings and then assembled using the genomic mode assembly of Newbler. The assembler eliminated 1.19 million reads based on quality criteria, placed 2.44 million reads into 264,420 contigs, leaving 2.47 million singletons that were longer than 100 bases of quality sequence of high complexity. These contigs and singletons were mapped to the GCAT clusters using Nuclear (Gydle Inc.) as described for the cDNA sequences.

Data Release

The Sanger produced ESTs are in dbEST (accessions CK434215-CK445168, CO472624-CO490610, DV970116-DV998428, EX305736-EX445576, GE471531-GE483395, GO360018-GO372194, GO653068-GO653136, GR222328-GR223664, GR435961-GR435991, GR953522-GR954913, GT051873-GT052669, GT738463-GT740210); the full length insert cDNA (FLICs) are in Genbank (accessions BT10331-BT120001) and the Next-Gen sequences are in the NCBI short read archive (accession SRA0293921). The cluster compositions (ESTs and clones) and are made available as Supplemental files. The sequence compiled in a single fasta file and a quality file, and cluster compositions are available for download at www.Aborea.ulaval.ca

List of Supplemental Materials

Supplemental Table S1. Descriptions, ESTs, clones and clusters of cDNA libraries.

Supplemental Table S2. Assessment of normalization efficiency in cDNA libraries.

Supplemental Table S3. Functional annotations of P. glauca cDNA clusters based on sequence similarity with major angiosperm genes and proteins, and conifer cDNAs.

Supplemental Table S4. Clusters with the longest sequence (100) and deepest coverage (100).

Supplemental Table S5. Protein domains in Pgl cDNA clusters based on Pfam-A models.

Supplemental Table S6. Distribution of Pfam domains in P. glauca and major angiosperm plants.
Supplemental Table S7. Distribution of transcription factor Pfam domains in *P. glauca* and major angiosperm plants.

Supplemental Table S8. Next-Generation libraries and sequences

**Acknowledgments.**

This study is an achievement of the Arborea research project in forest genomics. The authors are grateful for the provision of trace sequence files from the Treenomix project by J Bohlmann and collaborators (Univ of British Columbia, Canada). The authors thank V Jacobi, I Giguère, S Caron, F Gagnon, S. Blais (Laval University) and CG Allen (University of Alberta) for RNA and DNA isolations, and A Villeneuve and D Sinnett for Sanger sequencing work (McGill University and Genome Québec Innovation Center). Biological materials were provided by N Isabel (Canadian Forest Service) and J Beaulieu (Canadian Wood Fibre Center). Funding for the research was obtained from Genome Canada and Genome Québec for the Arborea project (PR, JB, JM) and from Genome Québec for the genomic sampling sequencing (JB, JM).

**References**


Prunier J, Laroche J, Beaulieu J, Bousquet J (2011) Scanning the genome for gene SNPs related


**Table 1. Sequence completion of cDNA clones and cluster representative clones.**

<table>
<thead>
<tr>
<th>cDNA clones (all)</th>
<th>201405</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete (FLIC)</td>
<td>131,148</td>
</tr>
<tr>
<td>5' and 3' reads</td>
<td>4,270</td>
</tr>
<tr>
<td>5' reads only</td>
<td>35,150</td>
</tr>
<tr>
<td>3' reads only</td>
<td>30,837</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genes (Clusters)</th>
<th>27,720</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete (FLIC)</td>
<td>23,589</td>
</tr>
<tr>
<td>5' and 3' reads</td>
<td>577</td>
</tr>
<tr>
<td>5' reads only</td>
<td>861</td>
</tr>
<tr>
<td>3' reads only</td>
<td>2,693</td>
</tr>
</tbody>
</table>

**Table 2. Number of clones per cluster and validations by Next-Gen sequencing.**

<table>
<thead>
<tr>
<th>Sanger sequencing: cluster depth</th>
<th>Clusters matched by 454 sequences (GS-FLX)</th>
<th>Clusters matched by Illumina sequences</th>
<th>Total Next-Gen matches</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1-5</td>
<td>&gt;5</td>
</tr>
<tr>
<td>Clones per cluster</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of clusters (% of total)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10,941 (39%)</td>
<td>2,522</td>
<td>2,389</td>
</tr>
<tr>
<td>2-9</td>
<td>11,275 (41%)</td>
<td>479</td>
<td>832</td>
</tr>
<tr>
<td>10-49</td>
<td>4,979 (18%)</td>
<td>40</td>
<td>42</td>
</tr>
<tr>
<td>50-99</td>
<td>416 (1.5%)</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>≥100</td>
<td>109 (0.4%)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total (per number of reads)</td>
<td>3,043</td>
<td>3,264</td>
<td>21,413</td>
</tr>
<tr>
<td>Representation reads)</td>
<td>11%</td>
<td>12%</td>
<td>77%</td>
</tr>
</tbody>
</table>

1 Number of quality sequences matching a given cDNA cluster (classes of 0 reads, 1-5, or >5 reads) for GS-FLX, Illumina, or GS-FLX and Illumina data combined (0 or >1 reads). Library details are in Supplemental Table 8.
Table 3. Sequence similarity and functional annotations according to cluster length\(^1\).

<table>
<thead>
<tr>
<th>cDNA clusters (bp)</th>
<th>All clusters</th>
<th>&lt;501</th>
<th>501-1,000</th>
<th>1,001-1,999</th>
<th>&gt;2,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of clusters</td>
<td>27,720</td>
<td>2,429</td>
<td>1,1903</td>
<td>1,1903</td>
<td>1,408</td>
</tr>
<tr>
<td>Total sequence (unique) (Mb)</td>
<td>30.15</td>
<td>0.92</td>
<td>9.27</td>
<td>16.7</td>
<td>3.26</td>
</tr>
<tr>
<td>Similar to a known plant sequence</td>
<td>84%</td>
<td>52%</td>
<td>78%</td>
<td>95%</td>
<td>98%</td>
</tr>
<tr>
<td>Putative function assigned</td>
<td>54%</td>
<td>16%</td>
<td>41%</td>
<td>71%</td>
<td>83%</td>
</tr>
<tr>
<td>Transcribed locus, unknown or predicted protein(^2)</td>
<td>30%</td>
<td>36%</td>
<td>37%</td>
<td>23%</td>
<td>15%</td>
</tr>
<tr>
<td>Unassigned (no significant hit)</td>
<td>16%</td>
<td>48%</td>
<td>22%</td>
<td>5%</td>
<td>2%</td>
</tr>
</tbody>
</table>

\(^1\) Based on matches to *A. thaliana, P. trichocarpa, or O. Sativa* annotated genome sequences.

\(^2\) The only significant hits had no assigned function (unknown proteins, predicted proteins, transcribed loci (undefined), unknown mRNA, and transcribed loci weakly or moderately similar to a known or predicted protein).
Table 4. RNA transcript sequence completion, CDS status and transcriptome coverage.

<table>
<thead>
<tr>
<th>Class</th>
<th>CDS coverage</th>
<th>Genes</th>
<th>Mb</th>
<th>FLIC cDNAs</th>
<th>FLIC avg length</th>
<th>Estimated missing Mb$^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Complete confirmed (protein)</td>
<td>5,837</td>
<td>7.97</td>
<td>96%</td>
<td>1,365</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Complete predicted (protein)</td>
<td>3,941</td>
<td>5.68</td>
<td>90%</td>
<td>1,419</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Complete Predicted (ORF)</td>
<td>1,830</td>
<td>1.49</td>
<td>94%</td>
<td>812</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Incomplete (protein)</td>
<td>7,949</td>
<td>8.67</td>
<td>78%</td>
<td>1,130</td>
<td>8.07</td>
</tr>
<tr>
<td>5</td>
<td>Incomplete (ORF)</td>
<td>6,654</td>
<td>5.63</td>
<td>77%</td>
<td>839</td>
<td>3.53</td>
</tr>
<tr>
<td>6</td>
<td>No CDS or ORF detected</td>
<td>1,509</td>
<td>0.71</td>
<td>91%</td>
<td>462</td>
<td>0</td>
</tr>
</tbody>
</table>

1 Class 1-3, similar to a reference protein (Arabidopsis, Rice, Poplar, Vine, SwissProt BLASTX Evalue = 1e-10): 1, complete confirmed if the CDS is similar over the entire protein; 2, complete predicted if the CDS is similar over part of the protein, but the transcript extends long enough on each side to cover the protein length; 3, incomplete if the ORF does not extend enough beyond the similarity region to cover the protein length. Class 4-6, no similarity to Arabidopsis proteins: 4, complete predicted if the ORF starts with a methionine upstream stop codons in transcript); 5, incomplete if ORF starts at the beginning of a transcript sequence, but first Methionine was well downstream or missing; 6, no ORF detected.

2Missing transcriptome sequence was estimated based on sequence alignments with the closest reference homolog and statistics of class 1 clusters used as a reference. Class 2 and 3 clusters with a predicted complete CDS were estimated not to lack any sequence in 5’ or 3’ UTRs. For incomplete clones of class 4, reference protein matches were used to estimate 6.29 Mb of coding sequence, 1.2 Mb of 5’ UTR and 0.58 Mb of 3’ UTR. The estimate for class 5 was based on the average transcript statistics in class 1.
Table 5. Pfam-A assignations in spruce, Arabidopsis, and rice.

<table>
<thead>
<tr>
<th></th>
<th>P. glauca</th>
<th>Arabidopsis</th>
<th>Rice $^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes or cDNA clusters</td>
<td>27,720$^2$</td>
<td>27,379</td>
<td>28,292</td>
</tr>
<tr>
<td>Transcriptome sequence (Mb)</td>
<td>41</td>
<td>40</td>
<td>49</td>
</tr>
<tr>
<td>Pfam Families</td>
<td>3,498$^3$</td>
<td>3,594</td>
<td>3,440</td>
</tr>
<tr>
<td>Pfam Clans</td>
<td>2,309</td>
<td>2,417</td>
<td>2,312</td>
</tr>
<tr>
<td>Genes with Pfam-A - Number</td>
<td>15,508$^3$</td>
<td>21,233</td>
<td>19,629</td>
</tr>
<tr>
<td>-Proportion</td>
<td>59%$^4$</td>
<td>78%</td>
<td>69%</td>
</tr>
<tr>
<td>Transcript size</td>
<td>1,365$^5$</td>
<td>1,326</td>
<td>1,507</td>
</tr>
<tr>
<td>CDS</td>
<td>867</td>
<td>1021</td>
<td>1096</td>
</tr>
<tr>
<td>5' UTR size</td>
<td>202$^3$</td>
<td>101</td>
<td>121</td>
</tr>
<tr>
<td>3' UTR size</td>
<td>322$^3$</td>
<td>204</td>
<td>290</td>
</tr>
</tbody>
</table>

$^1$The analysis only considered the 26,217 sequences containing one or more ORFs.

$^2$The assignation to Pfam-A families (version 24.0) was done in two steps: 1) Direct assignation: 13,971 gene clusters with direct assignation to 3,162 Pfam-A families; 2) Indirect assignation of an additional 1037 clusters by association with a TAIR homolog added another 186 Pfam_A families (see methods).

$^3$The estimation is conducted using class 1, 2 and 3 cDNA clusters (Table 4), which are biased toward short transcripts.

$^4$The number of genes in Arabidopsis is based on TAIR (version 9), and in O. sativa based on RefSeq (Version 43).

Table 6. Contribution of Next-Gen sequencing to extending the P. glauca gene catalogue

A. Extension of GCAT clusters

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Average length (bp)</th>
<th>Total (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3' clone extensions</td>
<td>9,012</td>
<td>176</td>
<td>1.58</td>
</tr>
<tr>
<td>5' clone extensions</td>
<td>7,844</td>
<td>151</td>
<td>1.19</td>
</tr>
<tr>
<td>Gaps filled</td>
<td>69</td>
<td>184</td>
<td>0.012</td>
</tr>
</tbody>
</table>

B. Unattached sequences

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Average length (bp)</th>
<th>Total (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20,945</td>
<td>243</td>
<td>5.0</td>
</tr>
</tbody>
</table>
Table 7. Similarity of orphan Next-Gen sequences with conifer transcripts in public databases.\(^1\)

<table>
<thead>
<tr>
<th>PUTs in PGDB(^2)</th>
<th>5,184</th>
<th>13,880</th>
<th>29,178</th>
<th>12,901</th>
<th>13,570</th>
<th>15,648</th>
<th>72,829</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identity</strong></td>
<td><strong>Number of hits</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;95%</td>
<td>68</td>
<td>188</td>
<td>67</td>
<td>490</td>
<td>15</td>
<td>22</td>
<td>67</td>
</tr>
<tr>
<td>&gt;90%</td>
<td>82</td>
<td>206</td>
<td>88</td>
<td>541</td>
<td>126</td>
<td>110</td>
<td>368</td>
</tr>
<tr>
<td>&gt;85%</td>
<td>88</td>
<td>216</td>
<td>92</td>
<td>567</td>
<td>194</td>
<td>140</td>
<td>547</td>
</tr>
<tr>
<td>&gt;70% (Total)</td>
<td>91</td>
<td>223</td>
<td>96</td>
<td>593</td>
<td>593</td>
<td>214</td>
<td>578</td>
</tr>
</tbody>
</table>

**Average**

| Identity (%)       | 96    | 98     | 96     | 97     | 90     | 91     | 91     |
| Length (bp)        | 307   | 446    | 424    | 470    | 539    | 376    | 478    |
| Score              | 504   | 768    | 706    | 799    | 685    | 503    | 625    |

\(^1\) A total of 3129 *P. glauca* contigs (>400 bp, >2 reads) that were obtained by cDNA assembly from the orphan GS-FLX reads (that could not be mapped to any *Pg*GCAT cluster, see methods) were compared to other *Pinaceae* transcriptomes, producing 1146 unique hits.

\(^2\) Putative Unique Transcript (PUT) from the Plant Genome Database project (PGDB) of the versions described in methods, [http://www.plantgdb.org/prj/ESTCluster/progress.php](http://www.plantgdb.org/prj/ESTCluster/progress.php).
Table 8. Identification of putative repetitive genomic in sequences the cDNA dataset.

<table>
<thead>
<tr>
<th>Type of sequence</th>
<th>Number of cDNA clusters</th>
<th>Sample sequencing$^1$</th>
<th>Poisson probability$^2$</th>
<th>Genomic copy number estimate$^3$</th>
<th>Occurrence in the cDNA dataset</th>
<th>Similarity to Ath proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly repetitive</td>
<td>1,195</td>
<td>40+</td>
<td>4.84 e-68</td>
<td>400+</td>
<td>1 or 2 ESTs</td>
<td>None</td>
</tr>
<tr>
<td>Moderately repetitive</td>
<td>295</td>
<td>30+</td>
<td>9.38 e-48</td>
<td>300-399</td>
<td>Only in GQ043-GQ045$^4$</td>
<td>None</td>
</tr>
</tbody>
</table>

$^1$Number of sequences that matched the cluster sequence (98% or more).

$^2$The Poisson probabilities are for the number of matches observed (30 or 40) if the sequence is a single copy gene. The probability of a single match is 0.333, assuming random sampling and gene distribution, and given a 0.085X genome coverage, the average length of the clusters (1348) and the average length of the genomic sequences (344 bp) but without accounting for introns.

$^3$Based on sample sequencing data assuming a random sampling and a 0.085X coverage of the P. glauca genome.

$^4$These libraries were determined to contain a large fraction of contaminating genomic DNA
Figure 1. The GCAT process applied to *P. glauca*. **A.** Overview of the EST, clone and FL-cDNA analyses steps. The clone analysis and the FL-cDNA sequencing enable iterative clustering, ultimately helping to optimize gene models, annotations and downstream applications. **B.** Size distribution and sequence completion of the representative cDNA clones for the 27,720 cDNA clusters. For incomplete clones, a minimum length was estimated based on available sequence.

Figure 2. Occurrence and classification of protein family domains in *P. glauca* relative to angiosperms. The total number of *P. glauca* cDNA clusters containing each of the Pfam-A domains was compared to *Arabidopsis*, rice and poplar genes (for entire list, see Supplemental Table 6) by using normalized angiosperm data to account for the number of overall genes in each species. The number of over- (**A**) and under- (**B**) represented protein domains in *P. glauca* was determined by Chi-squared testing (p-value < 0.05, with Bonferroni correction), for Pfam domains found six times or more in at least one of the species compared, and with a 50% difference between the species. The Pfam domains that were statistically different in at least two out of three comparisons were classified into major biological based upon TAIR annotations of *Arabidopsis* homologs (GO process), and Interpro and Pfam descriptions.

Figure 3. Relative frequencies of major Pfam domains found in transcription factor. Frequencies were determined for Pfam-A domains with hits in three genes or more in *P. glauca*, and calculated relative to the total number of genes containing TF Pfam domains within each of the species (Table 5). Stars indicate frequencies that are significantly different from Ath and Osa.

Figure 4. Coverage and validation EST clusters with GS-FLX (454) ESTs. **A.** Illustration of the validation of cDNA clusters with low clone coverage in the 5’ proximal region (GQ04008_F14, CC3HC4-type RING finger protein sequence). **B.** Confirmation of an unspliced intron in a unique
cDNA clone that was suggested by sequence similarity analyses (GQ0011_p18, RING 1A sequence, containing a putative unspliced intron).
Figure 1. The GCAT process and its application to *P. glauca*. A. Overview of the EST, clone and FL-cDNA analysis process. The clone analysis step and the FL-cDNA sequencing enable iterative sequence clustering and analyses to optimize gene models, annotations and downstream applications. B. Size distribution and cDNA sequence completion of the representative cDNA clones of the 27,720 GCAT-Pgl clusters. For incomplete clones, a minimum length was estimated based on available sequence.
Figure 2. Occurrence and classification of protein family domains in *P. glauca* relative to angiosperms. The total number of *P. glauca* cDNA clusters containing each of the Pfam-A domains was compared to *Arabidopsis*, rice and poplar genes (for entire list, see Supplemental Table 5) by using normalized angiosperm data to account for the number of overall genes in each species. The number of over-(A) and under-(B) represented in *P. glauca* was determined by Chi-squared testing (p-value < 0.05, with Bonferroni correction), for Pfam domains found six times or more in at least one of the species compared, and with a 50% difference between the species. The Pfam domains that were statistically different in at least two out of three comparisons were classified into major biological based upon TAIR annotations of *Arabidopsis* homologs (GO process), and Interpro and Pfam descriptions.
Figure 3. Relative frequencies of major Pfam domains found in transcription factor. Frequencies were determined for Pfam-A domains with hits in three genes or more in P. glauca, and calculated relative to the total number of genes containing TF Pfam domains within each of the species (Table 5). Stars indicate frequencies that are significantly different from Ath and Osa.
Figure 4. Coverage and validation EST clusters with GS-FLX (454) ESTs. A. Illustration of the validation of a cDNA cluster with low clone coverage in the 5' proximal region (GQ04008_F14, CC3HC4-type RING finger protein sequence). B. Confirmation of an unspliced intron in a unique cDNA clone that was suggested by sequence similarity analyses (GQ0011_P18, RING 1A sequence, containing a putative unspliced intron).