RESOURCE

Improved white spruce (Picea glauca) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism

René L. Warren\(^1\), Christopher I. Keeling\(^2,1\), Macaire Man Saint Yuen\(^2\), Anthony Raymond\(^1\), Greg A. Taylor\(^1\), Benjamin P. Vandervalk\(^1\), Hamid Mohamadi\(^1\), Daniel Paulino\(^1\), Readman Chiu\(^1\), Shaun D. Jackman\(^1\), Gordon Robertson\(^1\), Chen Yang\(^1\), Brian Boyle\(^8\), Margarete Hoffmann\(^3\), Detlef Weigel\(^3\), David R. Nelson\(^4\), Carol Ritland\(^5\), Nathalie Isabel\(^6\), Barry Jaquish\(^7\), Alvin Yanchuk\(^7\), Jean Bousquet\(^8\), Steven J. M. Jones\(^1,9,10\), John MacKay\(^8,11\), Inanc Birol\(^1,9,10,\)* and Joerg Bohlmann\(^2,5,12,\)*

\(^1\)Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC V5Z 4S6, Canada,
\(^2\)Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada,
\(^3\)Max Planck Institute for Developmental Biology, Spemannstrasse 35, 72076 Tübingen, Germany,
\(^4\)Department of Microbiology, Immunology and Biochemistry, University of Tennessee Health Science Center, Memphis, TN 38163, USA,
\(^5\)Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada,
\(^6\)Natural Resources Canada, Laurentian Forestry Centre, Québec, QC G1V 4C7, Canada,
\(^7\)British Columbia Ministry of Forests, Lands, and Natural Resource Operations, Victoria, BC V8W 9C2, Canada,
\(^8\)Department of Wood and Forest Sciences, Université Laval, Québec, QC G1V 0A6, Canada,
\(^9\)Department of Medical Genetics, University of British Columbia, Vancouver, BC V6H 3N1, Canada,
\(^10\)School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada,
\(^11\)Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK, and
\(^12\)Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

Received 24 February 2015; accepted 15 May 2015; published online 28 May 2015.
*For correspondence (e-mails ibirol@bcgsc.ca or bohlmann@msl.ubc.ca).
\(^1\)These authors contributed equally to this work.

SUMMARY

White spruce (Picea glauca), a gymnosperm tree, has been established as one of the models for conifer genomics. We describe the draft genome assemblies of two white spruce genotypes, PG29 and WS77111, innovative tools for the assembly of very large genomes, and the conifer genomics resources developed in this process. The two white spruce genotypes originate from distant geographic regions of western (PG29) and eastern (WS77111) North America, and represent elite trees in two Canadian tree-breeding programs. We present an update (V3 and V4) for a previously reported PG29 V2 draft genome assembly and introduce a second white spruce genome assembly for genotype WS77111. Assemblies of the PG29 and WS77111 genomes confirm the reconstructed white spruce genome size in the 20 Gbp range, and show broad synteny. Using the PG29 V3 assembly and additional white spruce genomics and transcriptomics resources, we performed MAKER-P annotation and meticulous expert annotation of very large gene families of conifer defense metabolism, the terpene synthases and cytochrome P450s. We also comprehensively annotated the white spruce mevalonate, mehtleythritol phosphate and phenylpropanoid pathways. These analyses highlighted the large extent of gene and pseudogene duplications in a conifer genome, in particular for genes of secondary (i.e. specialized) metabolism, and the potential for gain and loss of function for defense and adaptation.

Keywords: conifer genomes, whole-genome shotgun assembly, ABYSS, Bloom filter, genome scaffolding, genome finishing.
INTRODUCTION

Conifers are among the longest living plant species on the planet, with the natural life span of individual trees often exceeding several hundred years, and ranging up to more than 5000 years – an order of magnitude older than the confirmed oldest living angiosperm (Brown, 2014). With over 600 extant species (Farjon, 2014), conifers dominate the landscape across large areas of the northern hemisphere. Conifers have existed for over 300 Myr (Gernandt et al., 2011), and thus have survived through periods of extreme climatic sways (Jaramillo-Correa et al., 2004; Anderson et al., 2006, 2011; Tollefsrud et al., 2008), pest infestations (Food and Agriculture Organization of the United Nations 2009) and natural disasters (Kelly et al., 2013; Arbello et al., 2014). Much of the resistance of conifers against biotic stresses, such as microbial disease, insect pests, and browsing animals, originate from the tree’s massive metabolic investment in anatomical and chemical defense systems. Chemical defenses of different biochemical pathways are present in all organs of a conifer tree. Some specialized cell types and tissues such as resin ducts and polyphenolic parenchyma cells produce and accumulate large quantities of oleoresin terpenoids and various phenolic metabolites (Franceschi et al., 2005; Hammerbacher et al., 2011; Huggins et al., 2004; Keeling and Bohlmann 2006b, Mageroy et al., 2015). Despite their long generation times, conifers’ diverse and dynamic chemical defense systems and physical strength of lignified tissues allows these trees, under most conditions, to survive against faster evolving pests and pathogens (Kolosova et al., 2014).

The recently published draft genome sequences of white spruce (Picea glauca) (Birol et al., 2013), Norway spruce (P. abies) (Nystedt et al., 2013), and loblolly pine (Pinus taeda) (Neale et al., 2014) represent unprecedented technical achievements, possible in part due to advances in bioinformatics technology development (Simpson et al., 2009; Zimin et al., 2014; Mohamadi et al., 2015), enabling genome assemblies and analyses on the 10–100 Gbp order, a scale unimaginable only 5 years ago. These first conifer draft genome assemblies also provide first opportunities for identification of complete sets of large gene families and metabolic systems that contribute to resilience and resistance of conifers. Previous work on conifer defense gene families, such as those of oleoresin defense, was based on a subset of high-quality sequences from full-length cDNA sequences, expressed sequence tag (EST) sequences, and shotgun transcriptome sequencing (Ralph et al., 2008; Hamberger et al., 2011; Keeling et al., 2011b; Hall et al., 2013a).

White spruce (P. glauca) is widespread in North America and is the most widely planted conifer species in Canada. We reported on the shotgun genome sequence assembly of the P. glauca PG29 genotype (V2, Birol et al., 2013), which is an important genotype used in tree-breeding programs in British Columbia, where insect resistance is a major focus. Here, we present an updated draft PG29 genome assembly (V3) that is 70% more contiguous than the version previously reported (scaffold NG50 length V3 = 71.5 kbp versus V2 = 41.9 kbp; see Earl et al. (2011) for an explanation of the use of NG50 instead of N50 to compare assemblies). In brief, we re-scaffolded the previously published V2 PG29 draft genome sequence using a reference PG29 RNA-seq transcriptome assembly, a set of 27 720 white spruce cDNA clone sequences (Rigault et al., 2011), and large-fragment (3, 8, 12 kbp) mate pair sequences. We assessed the quality of this assembly using sequence capture data and a PG29-only subset of an updated and more comprehensive cDNA clone resource (42 440 cDNA downloaded from GCAT; https://web.gy-dle.com/smartforests/gcat) (Rigault et al., 2011). These resources enabled a genome-wide P. glauca gene annotation, along with more detailed analysis of specific gene families and pathways that are hallmarks of conifer defense, including the terpene synthase (TPS) and cytochrome P450 (P450) gene families, the mevalonate pathway, methylenetetrol phosphate pathway, and phenylpropanoid pathway. These gene families and pathways are responsible for biosynthesizing much of the specialized metabolome that forms the chemical defense of conifers (Franceschi et al., 2005, Keeling and Bohlmann 2006d).

In addition, we introduce the draft genome assembly of a second white spruce genotype, WS77111, a representative genotype from eastern Canada used in breeding programs in Quebec. This genotype has been used to develop a pedigree for constructing genetic linkage maps (Pelgas et al., 2011; Pavy et al., 2012). We note a high level of shared synteny between WS77111 and PG29, even though PG29 has recently been found to have features of a complex genetic admix of white spruce with Engelmann spruce (P. engelmannii) and Sitka spruce (P. sitchensis) (De La Torre et al., 2014b; Hamilton et al., 2014). Hence, we used the WS77111 V1 assembly, a genotype not known to be a genetic admix, to further re-scaffold the PG29 V3 genome draft, providing the most contiguous spruce genome yet (V4, scaffold NG50 length = 83.0 kbp), and a valuable reference for conifer genomics in general. In the context of this work, we discuss enabling, high-performance bioinformatics technologies for large genome assembly (e.g. ABYSS and DIDA) and comparative genomics (ABYSS-Bloom) that are expected to find broad applications, especially for large-scale genomics in which their impact can be more readily realized.
RESULTS

Assembly of the white spruce genotype WS77111 genome

Since the original white spruce PG29 draft genome was released (Birol et al., 2013), reduced sequencing costs and improved throughput made it feasible to generate the draft genome of another white spruce genotype of commercial and ecological interest in eastern North America. Doing so, we established two references for this species, and provided genomic resources for spruce breeding programs in both eastern and western North America. Following the sequencing strategy of the PG29 genome assembly, we designed multiple fragment libraries from genotype WS77111 to maximize representation of genomic content. We size selected some of these libraries to permit merging of paired-end reads into longer pseudo-reads in preparation for de novo assembly. We assembled over 3.3 billion paired-end Illumina reads totaling 0.97 Tbp, with at most 264, 12-core, compute nodes and a wall clock run time of slightly less than 5 days, producing a 22.4 Gbp WS77111 white spruce genome draft (Tables S1–S5). When comparing contiguity statistics between the assemblies of the two genotypes (Tables S6 and S7), we noted the WS77111 genome assembly contiguity to be slightly higher (at approximately 20 kbp, the scaffold N50 length of WS77111 was 2.3 kbp longer compared with that of the same-stage PG29 assembly when we followed the same assembly protocol). This finding was not surprising given that WS77111 whole-genome shotgun (WGS) reads were 287 bp long on average, approximately 130 bp longer than those generated for PG29 a year earlier. The strategy of combined sequencing platforms and fragment lengths was targeted to maximize paired read overlap and merging (e.g. PE250 on 400 bp fragments and PE300 on 600 bp fragments generated on Illumina HiSeq2500 and MiSeq, respectively). Accordingly, we were able to merge 71.5 and 60.5% of the paired reads derived from those corresponding combined fragment libraries, increase the value of k to 116, and improve assembly contiguity at every ABySS assembly stage (Table S6). The assembly figures for both PG29 and WS77111 corroborate a genome size in the 20 Gbp range, and highlight the value of using longer, low base error reads for assembly.

Updated assembly of the white spruce genotype PG29 genome

Since the first publication on the white spruce PG29 genome (Birol et al., 2013), we have improved upon the assembly by several means (Figure 1). We re-scaffolded the PG29 V2 genome using a newly obtained Trans-ABySS reference transcriptome assembly derived from eight PG29 RNA-seq libraries representing different organs and tissues (Table S7, NCBI BioProject PRJNA210511) to improve the contiguity in the genic space, reconstructing 13.7% more Core Eukaryotic Genes (CEGs), as assessed with CEGMA (Parra et al., 2007) (Table S8). Even though transcriptome re-scaffolding only had a modest impact on contiguity (increasing scaffold N50 length to 20.9 kbp) this was expected since the white spruce coding gene space represents a mere 0.11–0.37% of the genome. Further iterative re-scaffolding of this assembly using large-fragment mate pair sequences (mean length ± SD, 122 ± 22.6 kbp; 8.0 ± 3.3 kbp; and 3.3 ± 1.7 kbp) increased the NG50 contiguity of the genome scaffolds to 71.5 kbp, a 70.3% increase compared to the previously published assembly (V2). The contiguity of the genomic space of the resulting assembly (V3) is further improved, and now stands at 17.9% more complete since the original publication (Table S8), as assessed by CEGMA.

In the absence of a reference sequence, the assembly quality of large genomes is challenging to assess, especially for genomes riddled with repeat sequences such as those of conifers (e.g. approaching 86% in loblolly pine Wegrzyn et al., 2013). We have used orthogonal datasets to address the critically important point of assembly quality, namely 32 795 PG29 sequence capture contigs and a
set of sequences from 42,440 cDNA clones (Rigault et al., 2011; downloaded from GCAT; https://web.gydle.com/smartforests/gcat). The sequence capture was aimed at retrieving a portion of the gene space of PG29. In contrast to most exon capture experiments, it used long DNA fragments (mean insert size of 1.2 kb) and reads (629 bp on average) so as to span shorter introns. We obtained 3.7 million reads that were assembled into 126,508 contigs, which mapped to 23,184 (97%) of the target cDNAs used to design the probes and could be reduced to a set of 32,796 non-redundant contigs. In separate experiments, sequence contigs from sequence capture data and cDNA clone sequences were mapped onto the V3 PG29 assembly. We found over 84.5% of the capture contigs and 69.9% of the cDNAs aligned to the assembly with a sequence identity of 90% or higher, and covered over 80% of their length (Figure S1). When considering complete sequences aligning to a single assembly scaffold with 80% or more of their length, the majority (78.96 and 81.47%, respectively) of sequence capture contigs and cDNAs fell in this category (Table S9). Partial sequences are those covered over 20–80% of their length and represented 17.04 and 11.01% of the sequence data. It is unclear why certain sequences were not found in the PG29 V3 assembly (4.00%). Although as over 90% of the cDNA clones were mapped completely to either the PG29 V3 or WS77111 V1 assemblies, it points to minor genotype differences as well as assembly contiguity and gap differences in one relative to the other (Table S9).

Validation of the PG29 genome assembly size relative to other spruce genomes

The previous publications of draft genome assemblies for two species of spruce revealed an approximately 8 Gbp genome reconstruction size discrepancy reported for Norway spruce ((Nystedt et al., 2013) relative to white spruce ((Birol et al., 2013). Although genome size determination in gymnosperms is not without some uncertainty (Murray, 1998), DNA C-value flow cytometry measurements place the genome sizes at approximately 19.6 Gbp for Norway spruce (Nystedt et al., 2013) and 15.8 Gbp for white spruce (Bai et al., 2012). This suggests there is an approximately 8 Gbp under-assembly for Norway spruce (Nystedt et al., 2013) and an approximately 5 Gbp over-assembly for white spruce (this study and Birol et al., 2013). We investigated whether the PG29 genome was indeed over-assembled by aligning the assembly to itself and tallying secondary alignments at given size thresholds (Table S10). The PG29 assembly had at most 4.36% exact 1 kb and larger sequence duplicates, or approximately 0.9 Gb, which is approximately one-fifth the observed discrepancy. Repeated sequences are common in conifers, and likely accounted for the bulk of these duplicate sequences (Wegrzyń et al., 2014), instead of being exclusively caused by de novo assembly artifacts. The WS77111 genotype assembly reconstruction (22.4 Gbp) was within the same genome size range of PG29.

Bioinformatics tools for assembly assessment of very large genomes

Continued advances in sequencing and bioinformatics technologies have made sequencing and assembly of very large (>10 Gbp) genomes a reality, opening the field of genomics to organisms with previously prohibitively sized genomes. For instance, with at most 264 compute nodes (12 cores each; 3168 total cores) and a run time of slightly less than 5 days (Table S2) we assembled de novo a 22.4 Gbp WS77111 white spruce genome, once a bioinformatics feat that is now becoming a customary task, thanks to the continued development of technologies such as ABYSS for parallel assembly (Simpson et al., 2009) and DIDA that permits read alignments with big data where compute memory is limiting (Mohamadi et al., 2015). There is, however, much work to do to facilitate and streamline the analysis of these large assemblies on a genome-wide scale. With genomes of three different conifer species (white spruce, Norway spruce, loblolly pine) now sequenced, comparing the sequence content and organization on that order is an attractive, but challenging proposition. We developed a scalable bioinformatics solution (ABYSS-Bloom, ABYSS release v1.5.2) within the ABYSS toolbox, which makes use of memory-efficient Bloom filters (BF, Bloom, 1970) for analyzing and comparing the sequence content of large genomes.

In brief, two genome sequences are decomposed into their respective k-mer content, loaded in separate BF data structures and their set k-mer bit intersection measured. To calculate the sequence identity on a genome scale, we initially used a calibration of two synthetic genomes (human chromosome 21) with known sequence divergence (not shown). The calibration was used to test the system, since the relation between the intersecting set and sequence identity approximately follows the geometric distribution. In practice, the potentially high false positive rates (FPR) associated with Bloom filters as they approach saturation impacts the k-mer content calculation, and was factored into an FPR-corrected metric. We tested ABYSS-Bloom with k = 24 on published genomes with known sequence divergence such as human–chimp and human–macaque and produced values of divergence comparable with the published figures of 1.3% (±5.77 × 10−6) and 5.7% (±2.22 × 10−5), respectively (Chimpanzee Sequencing and Analysis Consortium 2005, Rhesus Macaque Genome Sequencing and Analysis Consortium et al. 2007, Scally et al., 2012). In each comparison, the intersecting set was corrected for the Bloom filter FPR and reported as a fraction of the smallest denominator, which is the genome with the least number of k-mers. Using this system with
**Table 1** Genome sequence divergence and estimated divergence between white spruce, Norway spruce and loblolly pine; and for comparison, genome sequence divergence and estimated divergence between human, chimpanzee and macaque

<table>
<thead>
<tr>
<th>Species</th>
<th>Est. million years lineages diverged</th>
<th>Published sequence divergence %</th>
<th>ABySS-bloom sequence divergence %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-chimpanzee</td>
<td>9–5a</td>
<td>1.1–1.4b</td>
<td>1.3</td>
</tr>
<tr>
<td>Human-macaque</td>
<td>28–25</td>
<td>6.5c</td>
<td>5.7</td>
</tr>
<tr>
<td>White spruce PG29 V3-WS77111</td>
<td>Not known</td>
<td>NA</td>
<td>2.2</td>
</tr>
<tr>
<td>White spruce PG29 V3-Norway spruce</td>
<td>20–15d</td>
<td>NA</td>
<td>3.0</td>
</tr>
<tr>
<td>White spruce PG29 V3-Loblolly pine v1.01</td>
<td>140–120d</td>
<td>NA</td>
<td>17.8</td>
</tr>
</tbody>
</table>

*Based on the k-mer content of the smaller genome. Divergence calculations are more accurate for more similar genomes.

Chimpanzee Sequencing and Analysis Consortium 2005.

Resesus Macaque Genome Sequencing and Analysis Consortium et al. 2007.

Bouillé and Bousquet, 2005.

Savard et al., 1994.

$k = 26$, we estimated the genome sequence content of white spruce PG29 to be 97.8% identical to that of white spruce WS77111, 97.0% identical to Norway spruce, and 82.2% identical to loblolly pine (Tables 1 and S11).

We assessed the structural divergence between the PG29 V3 and WS77111 V1 draft genome assemblies using PAVfinder, a pipeline initially developed for detecting chromosomal aberrations in human cancer genome assemblies (https://github.com/bcgsc/pavfinder). Although PAVfinder was designed to compare draft genome assemblies to a complete reference genome such as human, we used it here to compare the white spruce genotype WS77111 V1 assembly to the PG29 V3 assembly, analyzing whole-scaffold sequence alignment to gain insights on genome dissimilarity and thus evaluate the possibility to re-scaffold PG29 V3 using the white spruce WS77111 genotype sequence. Our analysis revealed a possible single translocation and no other major structural rearrangement between the two genotypes (Table S12). When considering WS77111 scaffolds 500 bp or larger for alignments $\geq 100$ bp, we profiled additional structural variations in the WS77111 genome in relation to PG29, including approximately 1.1 M deleted bases, approximately 201 $k$ duplicated bases, 21 $k$ inserted bases and 7.2 $k$ inverted bases in total. On a 22.4 Gbp scale, these represent small fractions of all assembled bases (0.0049, 0.009, 0.0001 and <0.0001%, respectively). This result, and the fact that scaffolds from the two assemblies did not necessarily reconstruct the same way, provided solid grounds for re-scaffolding PG29 using the WS77111 assembly.

The scaffolding process does not necessarily increase the resolved sequence content of an assembly. This situation is especially true when the read-to-read overlap parameter $k$ is globally optimized to give the best overall assembly. Yet, local coverage fluctuations due to the random sampling nature of the shotgun sequencing process may require higher (in higher coverage regions, to disambiguate sequence similarities) or lower (in lower coverage regions, to detect shorter overlap) $k$-mer lengths. As such, sequencing data that was used to assemble a genome would in principle contain information to close some of the scaffold gaps, without the need for additional sequencing experiments. To use this latent information, we developed Sealer (https://github.com/bcgsc/abyss/tree/sealer-prelease), a high-throughput sequence-finishing tool, using the computational engine of the paired-end read connecting utility, Konnector (Vandervalk et al., 2014). Sealer is currently the only high-throughput sequence finishing tool that scales up to large (>3 Gbp) genome assemblies (https://github.com/bcgsc/abyss/tree/sealer-prelease). We ran Sealer iteratively at different $k$-values for both PG29 V3 and WS77111 V1 assemblies, and also performed two rounds of gap closing on the PG29 V4 draft genome. Separate Sealer runs on the WS77111 V1 ($k = 64, 80, 96$) and PG29 V3 ($k = 84, 96$) draft assemblies closed a total of 461 196 (25.52%) and 399 476 (13.79%) of the gaps, respectively, in the first round of automated genome finishing, finishing the gap-filling PG29 V3 assembly was re-scaffolded using information derived from PG29-to-WS77111 scaffold alignments (Table S13, PG29 V4, GCA_000411955.4). The process merged over 1.3 M scaffolds, and yielded a scaffold NG50 length of 83.0 kbp, a 1.2-fold contiguity improvement compared with the V3 assembly. Further gap-filling with Sealer closed 498 251 gaps (13.08% of 3 807 770 gaps) on PG29 V4, which attested to the validity of the method that used a closely related genome for whole-genome re-scaffolding, since subsequent gap-filling was only permitted when the consensus sequence linking the scaffolds was derived from no more than two sequence paths through the Bloom filter de Bruijn graph (see the Experimental Procedures section for details). Further, assembly QC by mapping cDNA clones and CEGMA analysis rescues 550 and seven additional, complete cDNA and CEGs, respectively (Tables S and S9). We therefore conclude that, despite known polymorphisms between the white spruce genotypes, PG29 V4 which has been re-structured on WS77111 V1, is at the very least an improved assembly in the genic space.

**MAKER-P annotation of the white spruce PG29 V3 genome assembly**

We used MAKER-P (Campbell et al., 2014; Law et al., 2015) to generate an automated gene annotation of the PG29 genome assembly. MAKER-P generated more than...
105 000 transcript models in the PG29 V3 assembly (Holt and Yandell, 2011, annotations available on http://www.congenie.org). MAKER-P utilized assembled transcript sequences along with 27 720 white spruce cDNA sequences to help predict gene models. We derived a high-confidence set of 16 386 white spruce genes based on their expression in eight different tissue or organs exclusively. Direct sequence alignment of putative transcripts with identified coding potential (i.e., containing complete open reading frames (ORFs)) produced a shortlist of genes whose expression have also been measured by RNA-seq in those different white spruce tissues, an approach described in previous reports (Haas et al., 2002; International Peach Genome et al. 2013). The sequence identity of high-confidence genes to ORF-containing Trans-AbYSS scaffolds was high at 98.11 ± 3.80% (average ± SD). The annotation edit distance (AED) is a measure of the annotation to supporting evidence goodness of fit (Campbell et al., 2014) computed by MAKER-P and was used as an orthogonal method for validating alignment-based classification of the high-confidence gene set. AED values range from 0 to 1 with 0 representing complete concordance and 1 representing lack of supporting evidence. Over 75% of the 16 386 identified PG29 high-confidence genes had an AED lower than 0.2, indicating strong supporting evidence, at the expression level, for the majority of the high-confidence gene models. Only a minority (25%) of genes has less significant supporting evidence, as suggested by the AED metric (<1, Figure S1). These result could include a single line of evidence, RNA-seq transcript or cDNA clone, for instance.

To infer intron lengths in the PG29 genome V3, we used both the MAKER-P derived automated gene annotations and white spruce cDNA sequences. The largest intron length derived from the MAKER-P transcript models was approximately 44 kbp, but included only 10 kbp of known bases, consistent with the upper bound set for MAKER gene predictions. Keeping this value at 10 kbp helps minimize spurious gene annotations that would otherwise result from linking distant exons from adjacent genes. The largest measured intron size from experimental cDNA alignments was in excess of 370 kbp (Figure S2). We observed that MAKER-P had a propensity to minimize intron sizes when fitting models. In total, 60 265 introns were derived from alignments with the 42 440 cDNA sequence (downloaded from GCAT https://web.gydle.com/smartforests/gcat), whereas roughly double (n = 124 951) the number of introns contributed to the MAKER estimates, suggesting that in silico annotation could be improved upon in the future. The average intron length calculated here was nearly double the 2.4 kbp figure reported for loblolly pine (Wegrzyn et al., 2014) and may be attributable to differences between the spruce and pine genomes, or due to the lower number of pine transcripts sampled (n = 15 653).

Tissue sample grouping was assessed by non-negative matrix factorization, looking for positive linear combination of genes in the expression matrix of the high-confidence genes. A four-cluster solution emerged as the optimum (Figure S3a) and corroborated gene expression in biologically related tissue samples (Figure S3b). The needle tissue was most divergent from the other tissues, with genes having a measured expression up five-fold (>1000 RPKM) compared with that of other tissues. The needle cluster was enriched for genes with electron carrier activity, (MGSA analysis GO:0009055, enrichment score 0.5134, P < 0.004) and guanyl nucleotide binding (MGSA analysis, GO:0032561, enrichment score 0.3794, P < 0.004), supporting a role in photosynthesis, as one would expect to occur in this tissue. When looking at any of the top 5% of genes that were discriminatory (based on NMF score) for their respective cluster and looking at gene enrichment based on Gene Ontology (GO) classification, we found an over-representation of genes in the xylem/bark/young bud cluster with lyase activity (GO:0016829, score = 0.1240 P < 0.005) and more specifically carbon-oxygen lyase activity (GO:0016835, mgsa score = 0.1270 P < 0.005). Among P450s, 65 showed a broad profile of expression across the tissue types (Figure S3c), with only a 10 P450s discriminatory within a tissue sample (Figure S3d).

Annotation of select gene families and pathways in conifer defense

MAKER-P analysis provided a genome-wide annotation of the PG29 V3 genome assembly, which we complemented with expert annotations of target biological processes of conifer defense metabolism. Although the PG29 V3 assembly statistics were remarkable for such a large, highly repetitive genome, and are continually improving, the annotation of conifer genomes is currently still hampered by sub-optimal assembly contiguity. Potential gene models are often broken across multiple scaffolds due to long introns and repetitive sequence, or contain internal gaps. To manually identify gene models, we therefore used a combination of genomic (PG29 V3 assembly and genomic sequence capture) and transcriptomic (ESTs, fully sequenced cDNA clones, and Trinity and Trans-AbYSS assemblies of RNA-seq reads) resources as well as BLAST, exonerate, and meticulous manual examination. We found instances where the genome assembly contained partial gene fragments not found in the other data sources. Some of these fragmented genes may be due to insufficient assembly contiguity. However, sometimes several gene fragments on different scaffolds spanned the same small portion of a gene (such as one exon), suggesting that specific regions of a gene may have been duplicated in the genome. Our manual annotation focused on the TPS and...
P450 gene families and genes of the mevalonate pathway, methylerythritol phosphate pathway, and phenylpropanoid pathway. These gene families and pathways are important for the secondary (i.e. specialized) metabolism of conifer defense against biotic stressors. In particular, both the TPSs and the P450s are major drivers of the chemical diversity of terpenoids, and in case of the P450s also other metabolites, in specialized metabolism (Zerbe et al., 2013; Boutanaev et al., 2015). Approximately 50% of the gene models of these pathways and gene families identified in the manual annotation of the genome assembly appeared to be putative pseudogenes (containing one or more stop or frameshift mutations in the CDS). The automated annotation with MAKER-P often annotated the good portions of pseudogenes as several shorter gene models rather than identifying the full sequence as one putative pseudogene.

Annotation of the terpene synthase gene family

Across the transcriptome and genome sequence resources, we identified 83 unique TPSs that had at least 400 amino acids of CDS, including 28 (34%) putative pseudogenes. The TPS family is schematically shown in Figure 2 and shown with a detailed phylogeny in Figure S4. Eleven TPS genes clustered with diterpene synthases (di-TPSs) from other gymnosperm species, 32 clustered with known gymnosperm sesquiterpene synthases (sesqui-TPSs), 39 clustered with monoterpene synthases (mono-TPSs), and one clustered with hemiterpene synthases (hemi-TPSs). From our analysis of the PG29 genome assembly, we identified 726 putative TPS sequence fragments with at least 100 amino acids of CDS. Within these, 425 (59%) were putative pseudogenes. Many of the shorter fragments in the genome that do not contain stop codons or frameshifts could be degraded gene fragments, rather than broken gene models due to assembly contiguity, potentially increasing the proportion of putative pseudogenes observed.

Diterpene synthases of general and specialized metabolism

The di-TPSs identified in the PG29 genome and transcriptome include members involved in general (i.e. primary) metabolism, such as ent-copalydiphosphate synthase (ent-CPS) and ent-kaurene synthase (ent-KS) involved in gibberellic biosynthesis (Keeling et al., 2010), as well as specialized (i.e., secondary) metabolism, such as levopimaradiene/abietadiene synthase (LAS) and isopimaradiene synthase (Isos) involved in diterpene resin acid biosynthesis (Keeling and Bohlmann, 2006a; Keeling et al., 2011a; Zerbe and Bohlmann, 2014). We examined the number of genes and pseudogenes in greater detail for the di-TPSs to determine whether there was evidence for greater diversity of genes and pseudogenes in specialized compared to general metabolism. In general metabolism, we found three unique gene models for ent-CPS with at least 400 amino acids of CDS: one 401 amino acid long fragment, one full-length (equivalent to ACY25274, (Keeling et al., 2010)), and one putative pseudogene. Within the genome assembly, the full-length model had no representation, the 401 amino acid fragment was the only representation, and there were five putative pseudogenes with at least 100 amino acids. We found only one unique gene model for ent-KS with at least 400 amino acids of CDS (equivalent to ADB55711, (Keeling et al., 2010)). Within the genome assembly, two
partial sequence fragments represented the full-length gene model, and there were four putative pseudogenes of at least 100 amino acids. In secondary metabolism, we identified seven unique di-TPS representing genes in diterpene resin biosynthesis, including two putative pseudogenes. Within the genome assembly, there were 51 putative sequence fragments of at least 100 amino acids, including 23 (45%) putative pseudogenes.

Annotation of the cytochrome P450 gene family

Across the white spruce transcriptome and PG29 genome sequence resources, we identified 307 unique P450s of at least 400 amino acids in length, including 43 (14%) putative pseudogenes (Figures 3 and S5). In comparison with P450s in other plant species, several conifer- or gymnosperm-specific subfamilies (CYP76AA, CYP736B-C, and CYP750A in the CYP71 clan; CYP716B and CYP720B in the CYP85 clan; and CYP88K-P in the CYP86 clan) were apparent (Figures 3, 4 and S5). The distribution of the white spruce P450s within the families of the 11 plant P450 clans (Nelson and Werck-Reichhart, 2011) is shown in Figure S6. Analysis of the genome assembly alone identified 2203 putative P450 sequence fragments with at least 100 amino acids of CDS, including 1103 (50%) putative pseudogenes.

P450s of diterpenoid biosynthesis of general and specialized metabolism

As with the di-TPSs, different P450s are involved in general metabolism of gibberellin biosynthesis (ent-kaurene oxidase CYP701 and ent-kaurenolic acid oxidase CYP88) and specialized diterpene resin acid biosynthesis (CYP720Bs) (Ro et al., 2005; Hamberger et al., 2011). We identified one CYP701 gene model (CYP701A24), three CYP88 gene models (CYP88A28, CYP88A63P, and CYP88A64P), including two putative pseudogenes, and eight CYP720B gene models (CYP720B12v1, CYP720B12v2, CYP720B15, CYP720B2, CYP720B20P, CYP720B4, CYP720B7, and CYP720B8) including one putative pseudogene, across the white spruce datasets. Looking only at the genome assembly for sequence fragments with at least 100 amino acids of CDS, we found nine CYP701 gene models (including five putative pseudogenes), three CYP88s (including two putative pseudogenes), and 50 CYP720Bs (including 22 putative pseudogenes). Consistent with the pattern observed between general and specialized metabolism in the di-TPS, the P450s in diterpene oxidation of specialized metabolism (CYP720Bs) were more abundant, and included more putative pseudogenes, than those for general metabolism (CYP88 and CYP701).

Figure 3. Phylogenetics of the white spruce cytochrome P450 family. A phylogenetic tree of 307 white spruce P450s is shown with CYP51G used as the root. The phylogeny was created with FastTree 2 after protein alignment with MAFFT, and visualized with FigTree. The 10 plant P450 clans are labeled in black. Areas of conifer- or gymnosperm-specific expansion are labeled in color: CYP76AAs, blue; CYP736s, red; CYP750s, orange; CYP720Bs, green; CYP88s, olive; and CYP716Bs, purple. Figure S5 shows this phylogeny with the details of all P450 designators.
The mevalonate and methylerythritol phosphate pathways of isoprenoid biosynthesis

The mevalonate (MEV) and methylerythritol phosphate (MEP) pathways provide the isoprenoid precursors for terpenoid biosynthesis in spruce oleoresin defense, volatile emissions, and general metabolism. We identified full-length representatives of all genes in these two pathways with the exception of mevalonate kinase (MK) and phosphomevalonate kinase (PMK) genes in the mevalonate pathway (Figure 5). We found that the genome assembly typically contained multiple sequence fragments for the gene at each pathway step, including MK and PMK, and usually included one or more putative pseudogenes. However, although present in the transcriptome data, we did not find a sequence fragment in the genome for 4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase (CMK).

Phenylpropanoid pathway

The phenylpropanoid pathway is the origin of lignin as well as many other specialized metabolites that are important for conifer defense. We searched the white spruce genome and transcriptome resources for representatives of 17 phenylpropanoid pathway genes (Figure 6). Four of the phenylpropanoid pathway genes studied are P450s (C3H, C4H, F3H, and F5H), which were included in the P450 gene family analysis. We found putative full-length gene models for all pathway genes examined except for pinosylvin synthase (PS) and ferulic acid 5-hydroxylase (F5H). FSH would not have been expected, as spruce does not use F5H activity in the formation of lignin monomers. All pathway genes had representative sequence fragments in the genome with at least 100 amino acids of CDS, including one or more putative pseudogenes, except for PS and F5H. We found no evidence for PS or F5H in the white spruce genome.

Using genomic and transcriptomic resources for manual annotations

We used multiple sources of genomic and transcriptomic data to identify gene models to avoid limiting the annotations to those that are complete or nearly complete in PG29 V3 assembly only. For the gene families and pathways we examined, the origin of the best representative chosen for each gene model could have been from the genome assembly, genomic sequence capture data, ESTs, cDNA clone sequences, or the Trinity or Trans-ABySS RNA-seq assemblies. For the gene families and specific pathways examined, we found that the genome and transcriptome data contributed approximately equally to providing the most complete representative sequences for the non-redundant gene set. The best representative sequence for a gene model came from the genome assembly 36% of the time, from the sequence capture data 20% of the time, from the Trinity transcriptome assembly 22% of the time, from the Trans-ABySS transcriptome assembly 8% of the time.
Figure 5. White spruce genes annotated in the mevalonate (left) and methylerythritol phosphate (right) pathways. Green check mark or red ‘X’ indicates presence/absence of at least one ‘full-length’ (at least 75% the CDS length of the N50 of the known proteins) gene model in the genome and/or transcriptome sequences. Adjacent numbers (a/b/c) indicate: a) the total number of unique full-length gene models; b) the total number of sequence fragments in the genome assembly with at least 100 amino acids of CDS; c) the number from (b) that are putative pseudogenes (CDS includes one or more frameshift or in-frame stop).

Abbreviations in red are enzyme names: AACT, acetyl-CoA C-acetyltransferase (EC: 2.3.1.9); CMK, 4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase (EC: 2.7.1.148); DXR, 1-deoxy-D-xylulose-5-phosphate reductoisomerase (EC: 1.1.1.267); DXS, 1-deoxy-D-xylulose-5-phosphate synthase (EC: 2.2.1.7); FPPS, (2E,6E)-farnesyl diphosphate synthase (EC: 2.5.1.10); GPPS, dimethylallyltransferase/geranyl diphosphate synthase (EC: 2.5.1.29); HDR, 4-hydroxy-3-methylbut-2-enyl-diphosphate reductase (EC: 1.17.1.2); HDS, (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (EC: 1.17.7.1); HMG-R, hydroxymethylglutaryl-CoA reductase (EC: 1.1.1.34); HMG-S, hydroxymethylglutaryl-CoA synthase (EC: 2.3.3.9); IPPI, isopentenyl diphosphate Δ-isomerase (EC: 5.3.3.2); MCT, 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (EC: 2.7.7.60); MDS, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (EC: 4.6.1.12); MK, mevalonate kinase (EC: 2.7.1.36); MPDC, diphiphosphomevalonate decarboxylase (EC: 4.1.1.33); PMK, phosphomevalonate kinase (EC: 2.7.4.2).
Figure 6. White spruce genes annotated in the phenylpropanoid pathway. Green check mark or red ‘X’ indicates presence/absence of at least one ‘full-length’ (at least 75% the CDS length of the N50 of the known proteins) gene model in genomic and/or transcriptomic data. Adjacent numbers (a/b/c) indicate: *the total number of unique full-length gene models; †the total number of sequence fragments in the genome assembly with at least 100 amino acids of CDS; ‡the number from (b) that are putative pseudogenes (CDS includes one or more frame-shift or in-frame stop). An asterisk indicates that this is a P450 and the value came from the P450 gene family analysis.

Abbreviations in red are enzyme names: 4CL, 4-coumarate-CoA ligase (EC: 6.2.1.12); C3H, trans-cinnamate 4-monooxygenase (EC: 1.14.13.11); CAD, cinnamyl-alcohol dehydrogenase (EC: 1.1.1.195); CCoAOMT, caffeoyl-CoA O-methyltransferase (EC: 2.1.1.104); CCR, cinnamoyl-CoA reductase (EC: 1.2.1.44); CHI, chalcone isomerase (EC: 5.5.1.6); CHS, naringenin-chalcone synthase (EC: 2.3.1.74); DCS, 6'-deoxychalcone-4'-cinnamate 4-monooxygenase (EC: 1.14.13.11); COMT, caffeate O-methyltransferase (EC: 2.1.1.68); PAL, phenylalanine ammonia-lyase (EC: 4.3.1.24); PS, pinosylin synthase (EC: 2.3.1.146); STS, trihydroxystilbene synthase (EC: 2.3.1.95). The pathway leading to syringyl lignin is not shown because we do not expect this route to be found in conifers. The first enzyme leading to syringyl lignin from ferulic acid, feruloyl-CoA, coniferaldehyde, or coniferyl alcohol in the pathway shown is F5H, coniferaldehyde/ferulate 5-hydroxylase (EC: 1.14.13.-). No putative gene models for F5H were identified in any of the PG29 sequence resources examined.
those resources has become our biggest challenge in a shared resource environment, taking upwards of 1 month to format and move files and coordinate assembly runs where computing time for the white spruce WS77111 genome assembly did not exceed 5 days (Table S2). Recent efforts to minimize the genome assembly resource footprint have led to the implementation of several memory-efficient assemblers (Simpson and Durbin, 2010; Conway and Bromage, 2011; Chikhi and Rizk, 2012; Ye et al., 2012), but usually at the expense of time and accuracy. We have been preoccupied by the scale problem for some time (Simpson et al., 2009) and have recently outlined and presented the theory behind assembly by spaced seeds, a re-design of the traditional k-mer that, even in current data structure implementations, has potential for an over two-fold speed-up and a four-fold reduction in memory without compromising on assembly correctness (Birol et al., 2014).

Here we introduce re-scaffolded improved draft genome assemblies V3 and V4 of the western white spruce PG29 genome, first assembled in 2013 (Birol et al., 2013) and the additional genome assembly of eastern white spruce genotype WS77111. Having genomes of two white spruce individuals represented a unique opportunity to leverage the information of WS77111 to conservatively re-scaffold the PG29 V3 draft, especially as we observed no major structural genome rearrangement between the two genotypes. Large community resources, such as the Sanger sequences of over 42 440 cDNA clones (GCAT; https://web.gydle.com/smartforests/gcat), have been instrumental for assembly assessment providing good quality, contiguous assemblies of the spruce draft genomes. In fact, less than 2.5% or 1061 of cDNA clones of the catalogue are not found in either PG29 V3 or WS77111 V1 genome drafts. Consistent with this observation, automated gap-filling of over 38% newly linked scaffolds post PG29 re-scaffolding (V4), a process that stringently restricted a maximum of two similar assembly paths between scaffold end points, further attests to the quality of the latest PG29 draft genome.

Conifer genomes are becoming available owing to the cost-effectiveness of sequencing and enabling bioinfor-
matics technologies (Birol et al., 2013; Nystedt et al., 2013; De La Torre et al., 2014a; Wegryzn et al., 2014). Comparative genomics on the 20 Gbp scale is still out of reach, however. Towards this end, we have developed a scalable technology to analyze and compare entire genomic contents between any two sets of sequences. Comparing the sequence divergence between various conifer genomes reveals decreasing similarity that is consistent with the evolutionary distance between species, with 2.2% sequence dissimilarity between the two white spruce genotypes. Such a dissimilarity is likely higher than that from a strict comparison at the intraspecies level, given that PG29 results from a genetic admix of white spruce with the hybridizing species Sitka spruce and Engelmann spruce in British Columbia (De La Torre et al., 2014b; Hamilton et al., 2014). The dissimilarity was higher between Norway spruce and white spruce, reflecting their larger phylogenetic distance and quite ancient split beyond 10 Myr (Bouillé and Bousquet, 2005; Bouillé et al., 2011). And the dissimilarity was even higher between loblolly pine and white spruce, reflecting an ancient split of more than 100 Myr between spruce and pine lineages (Savard et al., 1994).

The genetic admix of white spruce PG29 with Engelmann spruce and Sitka spruce (De La Torre et al., 2014b; Hamilton et al., 2014) was assessed in previous work using a SNP panel that is discriminatory for these species. In the absence of Engelmann spruce and Sitka spruce genome sequences and due to the high sequence similarity across these species as gleaned from EST and full-length cDNA sequences, it was not possible to identify specific sequences in the PG29 genome that originated from Engelmann or Sitka spruce ancestry.

### Annotation of gene families and pathways in conifer defense

Beyond the MAKER-P annotation of PG29, we focused a more detailed annotation on gene families and pathways of conifer chemical defense systems. A major part of the constitutive and inducible chemical defenses in conifers are terpenoids (Keeling and Bohlmann, 2006b; Zerbe and Bohlmann, 2014), produced by the TPS gene family, and formed from precursors derived from the mevalonate and methylerythritol phosphate pathways. In addition, we annotated the PG29 genes of the phenylpropanoid pathway, which also plays important roles in the constitutive and inducible defenses in spruce (Franceschi et al., 2005; Hammerbacher et al., 2013, 2014; Mageroy et al., 2015). The large plant P450 gene family contains genes that act on terpenoids as well as phenolics. Identification of genes in these pathways and in the TPS and P450 families provides a large resource for future research in conifer defense. When comparing the automated and manual annotations, we found that MAKER-P annotations often did not encompass the full gene model as identified by manual annotations, especially for the TPS gene family (Figure 7d). As detailed below, manual annotations identified a large percentage of putative pseudogenes in all pathways and gene families examined. A preliminary examination of these pathways and gene families in the Norway spruce (Nystedt et al., 2013) and loblolly pine (Neale et al., 2014) genome assemblies indicated similar proportion of pseudogenes, suggesting that these pathways are actively evolving in conifers.

### Annotation of white spruce terpene synthases

While conifer TPSs of specialized metabolism are members of a gymnosperm-specific TPS-d subfamily, spruce TPSs of the general gibberellin metabolism cluster with angiosperm di-TPSs in the TPS-c and TPS-e subfamilies (Keeling et al., 2010, 2011b; Chen et al., 2011; Hall et al., 2013b). We identified 83 unique white TPS gene models in the PG29 genome, including 28 putative pseudogenes. This number is comparable with the number of TPS genes in several angiosperm genomes, e.g. 32 in Arabidopsis (Aubourg et al., 2002), 31 in rice (Oryza sativa) (Goff et al., 2002), 32 in poplar (Populus trichocarpa) (Tuskan et al., 2006), 113 in Eucalyptus (Myburg et al., 2014), and 69 in grapevine (Vitis vinifera) (Jaillon et al., 2007; Martin et al., 2010).

We identified 39 mono-TPS gene models, including 15 putative pseudogenes (Figures 2 and S4), only four of which have known functions (Keeling et al., 2011b). Some of the mono-TPSs may be functionally similar or identical to those of characterized orthologues from other conifers; however, it is important to note that a single amino acid difference can change TPS product profiles (e.g. Keeling et al., 2008; Roach et al., 2014). A cluster of putative white spruce hemi-TPSs was identified within the mono-TPS clade. Conifer hemi-TPSs appear to have evolved independently from their angiosperm counterparts (Sharkey et al., 2013). Gray et al. (2011) showed that certain conifer mono-TPSs [(–)-linalool synthases] and hemi-TPSs (3-methyl-2-buten-3-ol synthases) cluster together in the TPS-d phylogeny. Eight white spruce TPSs, including two pseudogenes, belong to this linalool/methylbutenol synthase clade. Only one of these has been characterized as a (–)-linalool synthase (Keeling et al., 2011b), and these genes might also have hemi-TPS activity. Two genes appeared immediately next to the linalool/methylbutenol synthase clade. Another gene model was orthologous to the methylbutenol synthases from Norway spruce and blue spruce (P. pungens) (AFJ73582 and AFJ73583), while the other four genes appeared orthologous to (–)-linalool synthases.

White spruce TPSs are most abundant in the sesqui-TPS clade with 32 gene models, including 10 putative pseudogenes. However, only one white spruce sesqui-TPS has been functionally characterized (Keeling et al., 2011b). Compared with mono- and diterpenes, sesquiterpenes con-
tribute only a small amount to conifer oleoresin, but they may contribute the most to the oleoresin chemical complexity (Keeling and Bohlmann, 2006b). Due to difficulties with sesquiterpene identification and due to the multiple products of many sesqui-TPSs (Steele et al., 1998), research into biological roles of conifer sesquiterpenes and sesqui-TPSs has been lacking. Two regions of the sesquiterpene clade are particularly bare of functional information: the eight white spruce genes surrounding the white spruce α-humulene synthase (Keeling et al., 2011b) and Sitka spruce δ-selinene-like synthase (Byun-McKay et al., 2006); and a clade of 12 white spruce genes including seven putative pseudogenes surrounding a germacradiene synthase from *Pinus sylvestris* (Köpke et al., 2008). (E, E)-α-farnesene synthases appeared as a small clade within the mono-TPS clade. One of these sesqui-TPSs also has mono-TPS (ocimene synthase) activity (Keeling et al., 2011b).

We identified 11 di-TPS gene models, including three putative pseudogenes, with seven in specialized metabolism and four in general metabolism. One di-TPS gene had closest similarity to the monofunctional class I isopimaradiene synthases from *Pinus contorta* and *P. banksiana* (Hall et al., 2013b), and contains a DIDV motif instead of the DXDD signature class II active site motif. Priscic et al. (2007) showed that a mutation of the second or third aspartate of the DXDD motif reduces class II activity. The discovery of a second gene model for a class II *ent-CP5* was unexpected. Whether this gene encodes a functional (±)-copalyl diphosphate synthase, as postulated to exist by Hall et al. (2013b) to complement monofunctional class I di-TPSs, requires future functional characterization.

In summary, the present TPS gene family annotation in white spruce is the most comprehensive to date for any gymnosperm. Functional characterization of conifer TPS genes has long been an active area of research (Stofer Vogel et al., 1996; Bohlmann et al., 1997; Steele et al., 1998; Martin et al., 2004; Keeling et al., 2011b); however, the white spruce genome and transcriptome annotation highlights that many TPS genes remain to be functionally characterized toward a comprehensive understanding of the chemical diversity of conifer terpenoids.

### Annotation of white spruce cytochrome P450s

The P450s form one of the largest gene families in plants and are important for chemical diversity of specialized and general metabolism (Nelson and Werck-Reichhart, 2011). We identified 307 P450s in white spruce including 43 putative pseudogenes, which may be the most comprehensive annotation of P450s in any gymnosperm. The number of P450 genes in the white spruce genome is in the same order of magnitude as the number of P450s found in angiosperm genomes, that is 272 (including 26 pseudogenes) in *Arabidopsis thaliana*, 455 (including 99 pseudogenes) in *Oryza sativa*, and 312 in *Populus trichocarpa* (Nelson et al., 2004; Nelson, 2006).

In contrast with the TPSs, which evolved as a large gymnosperm-specific TPS-d gene family (Chen et al., 2011; Keeling et al., 2011b), there is no deep separation of gymnosperm and angiosperm P450s. However, a few P450 families appear to be expanding differently in angiosperms and gymnosperms (Figure 4). More than 50% of the P450s identified in white spruce belong to the CYP71 clan, and were dominated by expansions of the CYP736 family, and the gymnosperm-specific CYP76AA and CYP750 subfamilies (Figures 3 and 4). We identified 42 CYP736 members, including three putative pseudogenes, 30 CYP76AA members, including five putative pseudogenes, and 43 CYP750 members, including five putative pseudogenes (Figure S7). Functions of these subfamilies are unknown. In the CYP85 clan, which contains many P450s for terpenoid modification (Hamberger and Bohlmann, 2006; Zerbe et al., 2013), the CYP720Bs form a conifer-specific subfamily with eight members in white spruce, including one putative pseudogene. A few members of the conifer CYP720B subfamily have been functionally characterized in diterpene resin acid biosynthesis (Ro et al., 2005; Hamberger et al., 2011). CYP716B is another gymnosperm-specific subfamily (Hamberger and Bohlmann, 2006) in the CYP85 clan, with 12 members identified in white spruce, including two putative pseudogenes. The only functionally characterized CYP716B-like gene is a taxoid 9α-hydroxylase from *Ginkgo biloba* (Zhang et al., 2014). In the CYP86 clan, white spruce is the only species present in the CYP86K, L, M, N, and P subfamilies, with 34 members including seven putative pseudogenes (Figure S7). Other conifer species appear to have orthologs in these subfamilies but have not yet been well annotated. In angiosperms, CYP86 members typically hydroxylate or epoxidize fatty acids, fatty alcohols, or alkane and their derivatives (Pinot and Beisson, 2011). Although the diversification of P450s in conifers may have been an important driver for species-specific diversity of specialized metabolites (Mizutani, 2012; Hamberger and Bak, 2013), there are few gymnosperm P450s that have been functionally characterized. The annotation of the large P450 gene family in white spruce identified many apparently gymnosperm- or conifer-specific subfamilies, highlighting where future efforts at identifying functions should be concentrated.

### Comparison of di-TPS and P450s of general and specialized metabolism

We compared the number of gene and putative pseudogenes in the TPS and P450 gene families associated with the similar diterpenoid biosynthetic pathways in the general (gibberellins) and specialized (diterpene resin acids) metabolisms (Keeling and Bohlmann, 2006a). TPS and P450 genes involved in the gibberellin pathway had fewer (or no
paralogues and fewer (or no) putative pseudogenes compared with TPS and P450 genes involved in diterpene resin acid biosynthesis (Figure 8). This finding is consistent with a previous analysis of ESTs and full-length cDNAs (Hamberger and Bohlmann, 2006; Keeling et al., 2010, 2011b) and a recent examination across several plant genomes, not including gymnosperms (Chae et al., 2014). These observations suggest that chemical diversity of terpenoid specialized metabolism originates, in part, from increased number and divergence of TPS and P450 genes. The smaller number of genes and pseudogenes for general metabolism may indicate less abundant gene duplication and/or that pseudogenes were not retained. In contrast, the diversity of gene models in specialized metabolism suggests increased rate of duplication and/or retention of functionally diversifying TPS and P450 paralogues, contributing to adaptive capacity of conifers as long-lived organisms.

Annotation of the mevalonate and methylerythritol phosphate pathways

The mevalonate and methylerythritol phosphate pathways provide the 5-carbon building blocks, DMAPP and IPP, for terpenoids in plants (Hemmerlin et al., 2012). However, only few genes from these pathways have been identified previously or characterized functionally in conifers or other gymnosperms. For those genes that have been studied previously, we found good concordance in the white spruce genome and transcriptome. The single gene model for HMG-S in PG29 (Figure 5) had 95% identity to the HMG-S protein from P. sylvestris (Wegener et al., 1997). Li et al. (2014) described an expansion of HMG-R genes in plants, not including a sequenced gymnosperm genome. We found 15 white spruce HMG-R genes, including five putative pseudogenes. The number in other plant genomes ranges from one in Selaginella moellendorfii to nine in Gossypium raimondii (Li et al., 2014). All but one of the white spruce HMG-Rs clustered distinctly between the angiosperm and lower plant clades, while one appeared at the base of the angiosperm clade (Figure S8). Three DXS genes are described in Norway spruce (Phillips et al., 2007); of the corresponding three genes in white spruce each share 99% protein identity with their Norway spruce orthologue. Two copies of CMK exist in G. biloba, GbCMK1 and GbCMK2, associated with general and specialized metabolism, respectively (Kim et al., 2008). Only one CMK was found in white spruce, with 75 and 67% protein identity to GbCMK1 and GbCMK2, respectively. An HDS gene model in white spruce is orthologous to a gene in G. biloba (Kim and Kim, 2010) with 87% protein identity.

DMAPP and IPP are condensed by isoprenyl diphosphate synthases (GPPS, FPPS, and GGPPS) to form the 10-, 15-, and 20-carbon substrates of conifer TPSs. GPPS, FPPS, and GGPPS genes have been identified in Norway spruce, white spruce and grand fir (Abies grandis) (Burke and Croteau, 2002a,b; Schmidt and Gershenzon, 2007, 2008; Schmidt et al., 2010; Nagel et al., 2014). Plant genomes typically contain families of GGPPS genes (Coman et al., 2014). In white spruce we identified one GPPS, three FPPS, including one putative pseudogene, and 16 GGPPSs, including seven putative pseudogenes (Figure S9).

Annotation of the phenylpropanoid pathway

Metabolites originating from the phenylpropanoid pathway have important roles across the plant kingdom.
including conifer defense (Tohge et al., 2013). For example, in white spruce the phenolic acetophenones pungenin and picein confer resistance against the spruce budworm (Choristoneura fumiferana) when released from the corresponding glucosides by β-glucosidases (Delvas et al., 2011; Mageroy et al., 2015). Initial work identified a set of ESTs of the phenylpropanoid pathway in white spruce (Porth et al., 2011). In the white spruce genome and transcriptome sequences, we found full-length gene models for all phenylpropanoid pathway genes examined except for PS and ferulic acid 5-hydroxylase (Figure 6). To date, pinosylvin and ferulic acid 5-hydroxylase (Figure 6). To date, pinosylvin has only been identified in pine species (Hovestad et al., 2006), suggesting that the white spruce genome may not contain a PS gene. We also did not find any sequence fragment in the white spruce genome with similarity to pine PS (Flegmann et al., 1992; Schwekendiek et al., 1992; Raiber et al., 1995; Kodan et al., 2002). Ferulic acid 5-hydroxylase (F5H, a CYP84) is also not expected to be present in white spruce because conifers do not produce syringyl lignin (Sarkanen and Ludwing, 1971; Neale et al., 2014). We found no evidence for a CYP84 gene or pseudogene. However, we identified multiple white spruce gene models for other P450s in the phenylpropanoid pathway, F3H (CYP75B1), C3H (CYP98A), and C4H (CYP73A). Stilbene synthase (STS) is important for the biosynthesis of diberzylen polyphenolic stilbenes in conifer defense (Hammerbacher et al., 2011). Hammerbacher et al. (2011) identified two STSs in Norway spruce with orthologs in white, and Sitka spruce. Consistent with these findings, we identified two white spruce STSs and two putative pseudogenes (Figure 6). Hammerbacher et al. (2014) functionally characterized four Norway spruce leucoanthocyanidin reductase genes (LAR1-LAR4) in flavan-3-ol biosynthesis. Transcript levels of these genes, and the monomeric and polymeric flavan-3-ols, increased after inoculation with the bark beetle-associated fungus Ceratocystis polonica, suggesting a role for these flavan-3-ols in conifer defense. In white spruce, we found one orthologue for each of LAR1, LAR2, and LAR4, and two orthologues for LAR3. Identity with the Norway spruce proteins were: LAR1, 99.2%; LAR2, 96.2%; LAR3, 82.7 and 99.7%; LAR4, 100%.

In summary, annotation of the phenylpropanoid pathway in white spruce with focus on defense will enable functional characterization of this metabolic system and its possible roles in conferring insect and pathogen resistance traits in white spruce.

CONCLUSION

We report the improved (PG29) and new (WS77111) genome assemblies for two genotypes of white spruce. We also present the associated white spruce genome and transcriptome sequence resources, new bioinformatics tools and their applications for the assembly of very large genomes from short read sequences only, and the annotation of a complete set of genes and pathway systems for conifer defense metabolism. These new resources will allow a more exhaustive investigation of adaptation in long-lived conifers. The present results uncovered an exceptional diversity of genes involved in chemical defense systems, which appear to be critical to the exceptional resilience of conifers over geological eras.

EXPERIMENTAL PROCEDURES

WS77111 library construction, sequencing and assembly

A single, diploid, tissue source (gDNA from white spruce genotype WS77111 needles) was used to build 14,000 bp, 600 bp and 15 kbp fragments) random WGS sequence libraries, as described previously (Birol et al., 2013). Overall, 3.8 and 1.0 billion sequence reads were generated from these libraries using the Illumina sequencing platforms at the McGill University Génome Québec Innovation Centre and Canada’s Michael Smith Genome Sciences Centre in Vancouver, respectively, and provided an approximately 48-fold coverage of the WS77111 genome and 1.2 T nucleotide bases sequenced (Table S1). White spruce WS77111 sequence data are available in GenBank under accession number PRJNA242552.

We assembled the WS77111 reads using the same assembly paradigm published for white spruce genotype PG29 (Birol et al., 2013), which we briefly summarize here. In the initial assembly step, we pre-processed libraries containing overlapping reads (HiSeq 400 bp and MiSeq 600 bp; Table S1) by merging the read pairs with the ‘abyss-mergepairs’ tool packaged in ABySS v1.3.7 (Simpson et al., 2009, latest version available at http://www.bcgsc.ca/platform/bioinfo/software/abyss). We ran abyss-mergepairs -p0.75 -m10 -q15 -v reads1.fastq reads2.fastq where ‘-p0.75’ indicates a minimum percent of 75% sequence identity in the overlap alignment, ‘-m10’ indicates a minimum of 10 bases matching exactly in the overlap alignment, and ‘-q15’ indicates that read tails should be quality trimmed up to the first occurrence of a Phred score of 15 or greater (prior to merging). This served to generate longer, higher quality sequencing reads for a large proportion (71.0 and 60.5% of the PE400 and PE800 libraries, respectively) of the read pairs in these libraries. With the exception of the read alignments, the remainder of the assembly process was coordinated using the ‘abyss-pe’ script and consisted of the four assembly stages, pre-unitig, unitig, contig, and scaffold, which we have previously described in detail (Birol et al., 2013). Briefly, the pre-unitig stage performs the initial assembly of sequences using a distributed, Message Passing Interface (MPI)-based de Bruijn graph assembly process and includes basic error correction algorithms and bubble popping. The unitig stage performs additional post-processing of the pre-unitigs including removal of redundant pre-unitigs, popping of large bubbles, and merging of overlapping pre-unitigs. The contig stage uses paired-end (PE) alignments of the PE reads to link unitigs into contigs, and similarly the scaffold stage uses paired-end alignments of the MPET reads to link contigs into scaffolds.

Sequence read alignments at the ABySS contig and scaffold assembly stage is a significant challenge because the construction of the FM-index for the corresponding unitig and contig assemblies required more RAM than was available on our largest memory machine (120 GB). To address this issue and to improve the speed of the read alignment stages, we developed a novel distributed framework for generic NGS alignments against large targets.
called Distributed Indexing and Dispatched Alignments (Mohamadi et al., 2015). Simplistically, DIDA partitions a set of target reference sequences and using a Bloom filter data structure (Bloom, 1970), assigns short reads to their correct partition for alignment. Alignments from each partition are then merged according to a set of rules that produces the best alignment or set of alignments for a given read.

The resource requirements for performing the full WS77111 assembly with ABySS were substantial and are detailed in Table S2, with the most expensive stage being the MPI-based pre-unigig assembly. We estimated the optimal k-value for the assembly based on the assembly statistics for the pre-unigigs across a range of k-values, as shown in Table S3 and Figure S10.

**PG29 genome assembly re-scaffolding**

We aligned the transcriptome assembly (see below) along with the cDNA sequences in the GCAT database (Rigault et al., 2011 and https://web.gydle.com/smartforests/gcat) to the PG29 V2 assembly (Biroi et al., 2013) using BWA-MEM (Li, 2013, version 0.7.5a; parameters -a -S -P -k75). From these alignments we extracted putative exon regions that mapped to different contigs, created links between all contigs that each RNA sequence mapped to, and supplied the resulting links into the remaining scaffolding algorithm of ABySS. The application that generates the links, abyss-longestdist, is released in ABySS version 1.3.7 onward.

After scaffolding the V2 assembly with cDNAs and RNA-seq contigs, we re-aligned the MPET data using DIDA (Mohamadi et al., 2015), see Table S4 for example commands and benchmark to further re-scaffold the assembly. This process was also performed with ABySS-1.3.7 executables, but with identical parameters and MPET sequence data used in scaffolding the PG29 V1 assembly. The PG29 V3 assembly is available for download at NCBI-GenBank under accessions ALW00000000; PId: PRJNA834345 and is published on our ftp site (ftp://ftp.bcgsc.ca/public/Picea_Glaucia/Release_3).

**PG29 genome assembly quality assessment**

We aligned 42 440 cDNA sequences (Rigault et al., 2011 and GCAT v3.3, https://web.gydle.com/smartforests/gcat) onto the PG29 V3 assembly using the cDNA-to-genome aligner GMAP (Wu and Watanabe, 2005, version: 2014-06-10, parameters: -f samse -t 12) and kept the five best alignments for each corresponding cDNA sequences. Each alignment was then filtered requiring a percent identity of 90% or more. Base deletions of 9 bp and longer were considered by GMAP as intronic sequences. The percent identity value takes into account insertions and deletions, but not introns, and was calculated as follows:

(i) alignment_length = sum(matches) + sum(insertions) + sum (deletions);
(ii) percent_identity = (sum(matches) - edit_dist)/alignment_length.

Where matches refer to exact matching bases and edit_dist is the edit distance, the minimum number of operations required to transform one string into the other.

For this dataset, we used the best alignment reported by GMAP to identify cDNA-containing PG29 scaffolds. As an orthogonal data set for assembly validation, we aligned gene capture contigs to the assembly using BWA-MEM (Li, 2013; version 0.7.6a-r4335a; parameters: -t 12). Sequence alignments with less than 90% identity were filtered out. With default parameters, BWA-MEM finds the best possible match for each segment of each gene capture contig. From the alignments in each category (cDNA and gene capture) we tallied complete, partial and missed sequences as those aligning to a single genomic scaffold and covering >80, >20% and >80% and >20% of the sequence query.

Large (>1 kbp) repeated sequences in the PG29 V3 assembly were identified by alignment to self, using BWA-MEM and samtools (version 0.7.5a and 0.1.19; command:bwa mem -t12 -a PG29-12_500.fa PG29-12_500.fa | samtools view -Sb -l samtools sort - self.sorted). When reading the alignments, only secondary alignments (sub-alignments) were considered to avoid counting legitimate scaffolds as repeats. Repeats larger than 1, 2, 5, 10, 20 and 30 bp were tallied, counting each scaffold-coordinate pairs once and summarizing the underlying bases (Table S10).

**White spruce genotypes PG29 and WS77111 assembly comparison**

Genome rearrangements between the two white spruce genotypes and assessment of WS77111 suitability for further re-scaffolding the PG29 V3 assembly was done by first aligning the new WS77111 assembly scaffolds (this study) onto the PG29 genome using bwa mem and samtools (Li, 2013; version: 0.7.5a and 0.1.19, command: bwa mem -t12 -a PG29-12.fa WS77111-b-500.fa | samtools view -Sb - l samtools sort - WS77111_vs_PG29.sort). Structural variants 100 nt or larger were detected with and reported by PAVfinder (J. Chu, in preparation; Python find_sv.py WS77111_vs_PG29.sorted.bam bwa mem WS77111-b-500.fa PG29-12.fa sv --min_size 100) (Table S10).

**Automated genome finishing**

Sequence gaps within assembly scaffolds were closed with Sealer, a scalable gap-closing pipeline for finishing draft genomes (https://github.com/bcgsc/abyss/tree/sealer-prelease). Briefly, regions with Ns are identified from the scaffold files of any given de novo assembler run. Flanking nucleotides (2 × 100 bp) are extracted from those regions while respecting the strand direction (5′→3′) on the sequence immediately downstream of each gap. Each flanking sequence pair is used as input to Konnector, a de novo PE assembler with memory-efficient de Brujin graph representation with a Bloom filter (Vandervalk et al., 2014). Instead of populating a two-level cascading Bloom filter with the input flank sequences, we use next-generation WGS reads, and populated the filter at a range of k-values, typically k = 30 to k = L/2 where k is the k-mer length and L the read length. With the ‘k sweep’ complete, successfully merged sequences are inserted into the gaps of the original scaffold file and Sealer outputs a new, gap-filled, genome assembly (abyss-sealer -ro run -S PG29.v3.fa -v -j 12 -B 300 -F 700 -P10 -k96—input-bloom = <(zcat PG29-bloom-k96.bloom.gz) -k80—input-bloom = <(zcat PG29-bloom-k80.bloom.gz)).

**Re-scaffolding the assembly PG29 V3 using the WS77111 V1 assembly**

The sealed WS77111 draft genome assembly was used to re-scaffold the PG29 V3 assembly (scaffolds ≥ 500 nt). We first indexed the gap-filled WS77111 V1 draft (NCBI BioProject PRJNA242552) with bwa index (version 0.7.5a; bwa index ws77111seal ed1_500.fa). We then aligned the gap-filled PG29 V3 assembly (GenBank assembly accession: GCA_000411955.2) onto the WS77111 V1 assembly and sorted the alignments (bwa mem -t4 ws77111sealed500.fa pg29sealed_500.fa | samtools view -Sb - l samtools sort - pg29_vs_WS77111-500.sorted). Using custom scripts, we converted the PG29.sam alignments into ordered and directed PG29 scaffold graph paths, which in turn were used to inform
AbYSS in making new scaffold merges. The resulting re-scaffolded PG29 assembly (version V4, GenBank assembly accession: GCA_000411955.4) was gap-filled with Sealer following the method and using the k-values described above.

**Genome analyses with bloom filters**

We developed a novel resource-efficient and scalable Bloom filter (Bloom, 1970) based approach to estimate the sequence identity between any two genome assemblies. For each genome, we constructed a Bloom filter representing the set of k-mers contained in the published assembly. Then, for each pair of Bloom filters, we counted the number of overlapping k-mers and estimated the percent sequence identity. To construct the Bloom filters for the comparison, we further developed the ‘abyss-bloom’ utility packaged with AbYSS 1.5.2, specifying an output Bloom filter size of 40 GB. We filtered out all sequences from the input assemblies that were shorter than 500 bp and constructed each Bloom filter using the following command:

```
$ cat <genome.fasta> | \
  fasta-minlen 500 - | \
  abyss-bloom build -v -k<k> -b40G - - | \
  gzip -c > <output_bloom.gz>
```

To calculate the intersection between Bloom filters, we again used the ‘abyss-bloom’ utility:

```
$ cat <bloom_gz_1> <bloom_gz_2> | \
  abyss-bloom intersect -v -k<k> - - | \
  gzip -c > <output_bloom.gz>
```

AbYSS-Bloom reports the number of true bits on unix standard error after creating a Bloom filter file; this provides the values for o, n1, and n2 necessary to estimate the number of overlapping k-mers in the genome comparisons. We conducted AbYSS-Bloom k-mer intersect analyses, comparing the sequence content between each white spruce PG29 V3, white spruce WS77111 V1, Norway spruce V1 (Nystedt et al., 2013), loblolly pine v1.01 (Neale et al., 2014) and Arabidopsis thaliana (Arabidopsis Genome Initiative 2000) genome pairs (k = 26) as well as validated our techniques on genomes with published estimates of sequence divergence, that of human and chimpan and human and macaque (Chimpanzee Sequencing and Analysis Consortium 2005, Rhesus Macaque Genome Sequencing and Analysis Consortium et al., 2007, Scally et al., 2012) at k = 24.

**PG29 gene annotation, transcriptome sequencing and assembly**

The PG29 V3 genome assembly was annotated using the MAKER-P (Campbell et al., 2014) pipeline, limiting annotations to contigs over 1 kb in length and with maximum intron lengths of 10 000 bp. MAKER runs third-party gene predictors and, together with experimental gene evidence, produces a gene annotation summary. Within this framework, RepeatMasker (http://www.repeatmasker.org/) was used to mask low complexity genomic sequence (Jurka et al., 2005). Also within MAKER-P, AUGUSTUS (Stanke et al., 2006) was run to produce gene predictions based on the Arabidopsis training set of genes, SNAP (Korf, 2004) gene predictions were based on a P. glauca EST training set and GeneMark (Lukashin and Borodovsky, 1998) was self-trained to produce its predictions. These three sets of predictions were combined with BLASTX (Altschul et al., 1997), BLASTN (Altschul et al., 1997) and exonerate (Slater and Birney, 2005) alignments of 27 720 white spruce cDNA sequences, eight assembled PG29 RNA-seq libraries (Table S7, data deposited at SRA under accession SRR026551, more details below), and all Swiss-Prot proteins (UniProt Consortium 2014) to produce the final annotations. Known protein domains were then further annotated using InterProScan (Hunter et al., 2009) and GO classifications linked to transcript models, when available. A reference PG29 transcriptome was established by assembling the RNA-seq reads from 8 PG29 tissue sources separately using every even k-value from 38 to 74. The resulting 19 sub-assemblies were merged to yield each tissue-specific assembly and combined to produce a single RNA-seq assembly that was used to aid in the annotation of the genome (NCBI BioProject PRJNA210511). The PG29 transcriptome assembly comprised 41 253 710 sequence scaffolds, 193 949 of which harbored complete ORF, as detected by TransDecoder (http://transdecoder.sourceforge.net/; default parameters) a tool for identifying candidate coding regions within transcript sequences. This more manageable set of sequence transcripts was provided to MAKER-P to be used as direct evidence for annotation, as described above.

A high-confidence coding gene set of 16 386 sequences was obtained by first aligning the 193 949 ORF-containing transcriptome scaffolds to the 105 724 MAKER-P predicted transcript models with NCBI-BLASTn (version –2.2.28+; parameters e-value <1 x 10^-20; Figure S11), identifying 70 184 predictions with alignments to our reference transcriptome. However, we observed high gene redundancy and several cases where a single transcriptome scaffold aligns to many predicted gene models. We collapsed the 70 184 predicted genes further to 16 386 by keeping MAKER-P predicted genes having the best hit to a given RNA scaffold, ensuring uniqueness of our high-confidence gene set. RNA-seq reads from eight PG29 tissue libraries (Table S7) were aligned onto the high-confidence gene set with BWA (Li and Durbin, 2009) and normalized to RPKM i.e. to reads per 1000 bp per million reads aligned. InterProScan runs (Hunter et al., 2009) yielded GO annotation for 9306 of the high-confidence genes. Blast2GO (Conesa et al., 2005) was used to rescue the remaining genes. After execution of BLAST, GO-mapping, Annotation and InterPro annotations within the Blast2GO interface, 319 genes were further annotated, bringing the total of high-confidence genes associated with one or more GO terms to 9625.

**General approach developed for manual annotation of gene families and pathways in conifer chemical defense**

We used the PG29 V3 genome assembly scaffolds, PG29 genomic sequence capture contigs, white spruce ESTs, fully sequenced white spruce cDNA clones, and PG29 transcriptome assembly contigs to BLASTx (Camacho et al., 2009) search against relevant protein sequences from plants. Putative gene models were identified via exonerate (protein2genome model, version 2.2.0, (Slater and Birney, 2005)) using the protein sequence databases. In the case of transcriptome data, redundancy was reduced by using CD-HIT (version 4.6.1, (Fu et al., 2012)) at 98% amino acid identity on the protein predictions from exonerate. Gene models were flagged as putative pseudogenes if the predicted CDS contained at least one internal stop codon or frameshift. Final gene models were manually edited in CLC Bio Main Workbench 7.5 (http://www.clcbio.com) after exonerate was used to create the initial gene models. In the case of pseudogenes, where possible we corrected frameshifting and continued the gene model after the internal stop (s) to the full length of the orthologous proteins. For manual annotation, phylogenies and counting of ‘near-full-length’ genes, we filtered for gene models with at least 400 amino acids of CDS for the TPS and P450 gene families (approximately 75% of the length of the shortest conifer sesquiterpene synthases and plant P450s), and calculated a cut-off value for each of the other gene families based on 75% of the N50 length of bait protein sequences used in the BLASTx searches. The putative mevalonate, methylyerythritol diphosphate,
and phenylpropanoid pathway gene models were searched with BLASTp against the KEGG database (Kanehisa and Goto, 2000) and the gene models annotated in Hammerbacher et al. (2011) to confirm which of these gene models had homology to the correct EC number or description for each gene of interest. Sequence redundancy was reduced by collapsing sequences that shared at least 98% protein sequence identity.

**Genomic sequence capture**

Sequence capture aimed at the gene space of PG29 was carried with the SeqCap EZ developer capture procedure (Roche NimbleGen, Madison, WI, USA). We used 462 160 probes targeting 23 084 genes described previously (Stival Sena et al., 2014) and followed the general guidelines of the manufacturer. Briefly, 1 μg of genomic DNA was used to generate a FLX+ rapid library (Roche 454, Branford, CT, USA) with a 1.2 kb mean insert size according to the manufacturer’s guidelines. The library was amplified by ligation-mediated PCR using 454 A and B primers as described in the NimbleGen SeqCap EZ Library LR User’s guide; 1 μg of amplified library was combined with 10 μl of plant capture enhancer (Roche NimbleGen) and 5 μl of 100 μM hyb enhancing A and B primers. The mixture was dried and resuspended in 7.5 μl of 2× SC hybridization buffer and 3 μl of SC component A and heated to 70°C for 10 min. After a quick spin, the mixture was added to 4.5 μl of capture oligonucleotides solution. The hybridization mixture was heated to 95°C for 10 min followed by 64 h at 47°C. Streptavidin-coated Dynabeads M-270 (Life Technologies; www.lifetechnologies.com) were used to pull out captured material and non-captured material was washed away according to the NimbleGen SeqCap EZ user’s guide. The captured library was amplified by ligation-mediated PCR using 454 A and B oligos. The capture efficiency was assessed by qPCR comparing pre- and post-capture libraries using four white spruce genes, and was 100-fold on average. Emulsion PCR and GS-FLX + sequencing were performed according to manufacturer’s instructions at the Institute for Systems and Integrative Biology (Univ. Laval, QC, Canada). Three full GS-FLX+ sequencing plates generated a total of 3.7 M raw sequencing reads with an average read length of 629 nt that were assembled with Newbler V2.8. Post assembly analysis showed that 22 184 (97%) of the cDNAs mapped to at least one of the contigs and a total of 126 508 contigs mapped to one or more cDNAs targets; the set of contigs could be reduced by collapsing anything identical within the same species. In addition, TPS and P450 database proteins less than 400 amino acids were excluded from the alignment and phylogeny.

**Picea glauca sequence resources used for manual annotation**

We used the following sequence resources for identifying unique genes:

(i) **Genome assembly sequence resource:** The PG29 V3 assembly reported here.

(ii) **Sequence capture resource:** Genomic sequence capture data from PG29 described above.

(iii) **EST and cDNA sequence resources:** 27 720 cDNA clusters (Rigault et al., 2011) and 47 492 Sanger ESTs (Pavy et al., 2005; Ralph et al., 2008).

**RNA-seq transcriptome sequence resource.** Eight samples of different PG29 tissues (megagametophyte, embryo, seedling, young buds, xylem, mature needles, flushing buds, and bark) were separately PE sequenced using the Illumina HiSeq 2000 (NCBI PRJNA210511). We used both Trinity and Trans-AbYSS assemblers separately to assemble the RNA-seq data for transcriptome mining.

**Trinity transcriptome assembly.** Sequences from the young buds, xylem, mature needles, flushing buds, and bark libraries were used. Quality control of the sequences was assessed with FastQC (version 0.10.1, (Andrews, 2014)). Sequences were filtered and trimmed with Trimmomatic (version 0.30, (Lohse et al., 2012)), and bases with quality <10 were trimmed from the 3’ end of each read. Sequences were dropped if the final trimmed length was less than 70 bp, or contained only Illumina TruSeq sequence. The resulting trimmed sequences were pooled and assembled with Trinity (version 2013-02-25, (Grabherr et al., 2011)). The assembly generated a total of 565 628 contigs with an average length of 701 bp. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GCHX00000000. The version described in this paper is the first version, GCHX01000000.

**Trans-AbYSS transcriptome assembly.** We also used AbYSS version abyss-1.3.4 (Simpson et al., 2009) and Trans-AbYSS version 1.4.6 (Robertson et al., 2010) to assemble the RNA-seq data from all eight libraries. This assembly generated a total of 339 058 contigs with an average length of 1639 bp (NCBI BioProject PRJNA210511).

**Protein databases used for the manual annotation**

For the TPSs, we used a set of 298 predominantly gymnosperm TPS protein sequences extracted from NCBI. For the P450s, we used a set of over 7000 curated plant P450 protein sequences (http://drnelson.utsc.edu/CytochromeP450.html). For the mevalonate, meyletherthiol phosphate, and phenylpropanoid pathways we used plant protein sequences available from PlantCyc release 9.0 (http://www.plantcyc.org/). In the case of naringenin-chalcone synthase, trihydroxystilbene synthase, and PS, we included the gymnosperm protein sequences reported by Hammerbacher et al. (2011), and we used the protein sequences reported by Hammerbacher et al. (2014) to identify leucoanthocyanidin reductase (LAR) genes.

**Alignments and phylogenies used for the manual annotation**

Alignments were prepared with MAFFT (version 7.123, options—maxiterate 50000—reorder-auto, (Katoh and Standley, 2013)). Phylogenies were prepared with FastTree 2 (version 2.1.6, options: -boot 1000 -wag -gamma, (Price et al., 2010)) and displayed with either CLC Main Workbench 7.5 (http://www.clc-bio.com) or FigTree (version 1.4.2, http://tree.bio.ed.ac.uk/software/figtree/). Redundancy in TPS and P450 database proteins were reduced with CD-HIT by collapsing anything identical within the same species. In addition, TPS and P450 database proteins less than 400 amino acids were excluded from the alignment and phylogeny.

**Comparison of gene models to PG29 V3 and WS77111 V1 genome assemblies**

We used the CDSS of the manually derived gene models to blast against the entire PG29 V3 and WS77111 V1 genome assemblies, and the MAKER-P transcripts from PG29 V3, with BLASTn (megablast, e-value ≤1 x 10^-7) and identified the high-
est query coverage of those matches that had ≥95% nucleotide identity.

Accession numbers
The NCBI accession numbers for the genome and transcriptome resources described in this paper are as follows:

(i) The PG29 genome project is contained in BioProject PRJNA83435.
(ii) PG29 V3 genome assembly accession number: GCA_0004 11955.3 PG29. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession ALWZ00000000. The V3 version described in this paper is version ALWZ04000000.
(iii) V3 MAKER-P annotations: pending – currently being processed by NCBI.
(iv) PG29 V4 genome assembly accession number: GCA_0004 11955.4. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession ALWZ00000000. The V4 version described in this paper is version ALWZ04000000.
(v) PG29 genome sequence capture: This assembly is contained in BioProject PRJNA83435. The sequence reads are deposited under NCBI accession SRR1982100. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession LDPM00000000. The version described in this paper is version LDPM01000000.
(vi) PG29 RNA-seq reads: BioProject PRJNA210511, SRA SRX318118–SRX318125.
(vii) Trinity RNA-seq transcriptome assembly of PG29 RNA-seq reads: GCHX00000000. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GCHX00000000. The version described in this paper is the first version, GCHX01000000. This assembly is contained in BioProject PRJNA210511.
(viii) Trans-AbYSS RNA-seq transcriptome assembly of PG29 RNA-seq reads: This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GCZO00000000. The version described in this paper is the first version, GCZO01000000. This assembly is contained in BioProject PRJNA210511.
(ix) The WS77111 genome project is contained in BioProject PRJNA242552. WS77111 V1 genome assembly accession number: GCA_000966675.1. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession JZKD00000000. The V1 version described in this paper is version JZKD01000000.

The assemblies and Maker-P annotations reported in this paper are also available at ConGenIE (http://congenie.org). In addition, the manually annotated genes are shown in ConGenIE mapped to all three genome assemblies.

ACKNOWLEDGEMENTS

We thank Ms Karen Reid (UBC) for excellent assistance with project and laboratory management, and the BC Cancer Agency Michael Smith Genome Sciences Centre and the McGill University and Genome Québec Innovation Centre for sequencing support. This work was supported by Genome Canada, Genome Quebec, Genome British Columbia and Genome Alberta (to I.B., J. Boh., J. Bou., S.J.M.J., J.M., and A.Y.) as part of the SMarTForests Project (www.smartforests.ca) and with funds (to J. Boh.) from the Natural Sciences and Engineering Research Council of Canada (NSERC). J. Boh. is a University of British Columbia Distinguished University Scholar.

CONFLICT OF INTEREST

None declared.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Evidence supporting MAKER-P gene annotations in PG29.
Figure S2. Introns lengths derived from MAKER-P transcript models and experimental transcripts.
Figure S3. NMF consensus clustering of high-confidence genes.
Figure S4. Phylogeny of gymnosperm terpene synthases.
Figure S5. Phylogeny of the white spruce cytochrome P450 family.
Figure S6. Distribution of white spruce P450s across the eleven plant P450 clans.
Figure S7. Gymnosperm- and conifer-specific cytochrome P450 families.
Figure S8. Phylogeny of HMG-R proteins in plants.
Figure S9. Phylogeny of isoprenyl diphosphate synthase proteins in plants.
Figure S10. N20 and N50 length statistics for WS77111 V1 pre-unитig assemblies across a range of k-values.
Figure S11. Derivation of a high-confidence gene set based on detected gene expression in eight white spruce tissue and organ samples.

Table S1. White spruce WS771111 sequence data.
Table S2. AbYSS resources required at various assembly stages of building the 20 Gbp White Spruce WS77111 V1 draft genome.
Table S3. Statistics for WS77111 V1 pre-unitig assemblies across a range of k-values.
Table S4. Execution, runtime and resources required at different stages of the DIDA read alignment framework.
Table S5. White spruce WS77111 V1 AbYSS v1.3.7 assembly statistics.
Table S6. N50 length (kbp) comparisons between white spruce individuals PG29 and WS77111 at various stages of the AbYSS assembly.
Table S7. White spruce PG29 RNA-seq transcriptome sequence reads.
Table S8. White spruce assembly completeness, as measured by CEGMA analyses.
Table S9. Assembly quality control (QC) using cDNA and sequence capture resources.
Table S10. Exact repeat content in the PG29 V3 genome assembly.
Table S11. AbYSS-Bloom sequence identity calculations between various draft genome assemblies.
Table S12. Structural variation (S.V.) in WS77111 relative to PG29 V3 by PAVfinder (Chiu, et al. in preparation) analysis of whole-scaffold alignments.
Table S13. White spruce genome re-scaffolding-scaffold assembly statistics.

REFERENCES


Hamberger, B. and Bohlmann, J. (2008) Cytochrome P450 mono-oxygenases in conifer genomes: discovery of members of the terpenoid oxygen-
ase superfamily in spruce and pine. Biochim. Soc. Trans. 34, 1209–1214.


White spruce genome


© 2015 The Authors