











## RESOURCE

# Spruce giga-genomes: structurally similar yet distinctive with differentially expanding gene families and rapidly evolving genes

Kristina K. Gagalova<sup>1</sup> , René L. Warren<sup>1</sup> , Lauren Coombe<sup>1</sup> , Johnathan Wong<sup>1</sup>, Ka Ming Nip<sup>1</sup> , Macaire Man Saint Yuen<sup>2</sup> , Justin G. A. Whitehill<sup>2</sup> , Jose M. Celedon<sup>2</sup> , Carol Ritland<sup>2</sup>, Greg A. Taylor<sup>1</sup>, Dean Cheng<sup>1</sup>, Patrick Plettner<sup>1</sup>, S. Austin Hammond<sup>1,3</sup>, Hamid Mohamadi<sup>1</sup>, Yongjun Zhao<sup>1</sup>, Richard A. Moore<sup>1</sup>, Andrew J. Mungall<sup>1</sup>, Brian Boyle<sup>4</sup>, Jérôme Laroche<sup>4</sup>, Joan Cottrell<sup>5</sup>, John J. Mackay<sup>6</sup>, Manuel Lamothe<sup>7</sup>, Sébastien Gérardi<sup>4,8</sup>, Nathalie Isabel<sup>7,8</sup>, Nathalie Pavy<sup>4,8</sup>, Steven J. M. Jones<sup>1</sup> , Joerg Bohlmann<sup>2</sup> , Jean Bousquet<sup>4,8</sup> and Inanc Birol<sup>1,\*</sup> 

<sup>1</sup>Canada's Michael Smith Genome Sciences Centre, Vancouver, BC V5Z 4S6, Canada,

<sup>2</sup>Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada,

<sup>3</sup>Next-Generation Sequencing Facility, University of Saskatchewan, Saskatoon, SK S7N 5E5, Canada,

<sup>4</sup>Institute for Systems and Integrative Biology, Université Laval, Québec, QC G1V 0A6, Canada,

<sup>5</sup>Forest Research, U.K. Forestry Commission, Northern Research Station, Roslin, EH25 9SY, Midlothian, UK,

<sup>6</sup>Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK,

<sup>7</sup>Natural Resources Canada, Canadian Forest Service, Laurentian Forestry Centre, Québec, QC G1V 4C7, Canada, and

<sup>8</sup>Canada Research Chair in Forest Genomics, Forest Research Centre, Université Laval, Québec, QC G1V 0A6, Canada

Received 11 August 2021; revised 22 June 2022; accepted 27 June 2022; published online 4 July 2022.

\*For correspondence (e-mail [ibirol@bcgsc.ca](mailto:ibirol@bcgsc.ca)).

## SUMMARY

Spruces (*Picea* spp.) are coniferous trees widespread in boreal and mountainous forests of the northern hemisphere, with large economic significance and enormous contributions to global carbon sequestration. Spruces harbor very large genomes with high repetitiveness, hampering their comparative analysis. Here, we present and compare the genomes of four different North American spruces: the genome assemblies for Engelmann spruce (*Picea engelmannii*) and Sitka spruce (*Picea sitchensis*) together with improved and more contiguous genome assemblies for white spruce (*Picea glauca*) and for a naturally occurring introgress of these three species known as interior spruce (*P. engelmannii* × *glauca* × *sitchensis*). The genomes were structurally similar, and a large part of scaffolds could be anchored to a genetic map. The composition of the interior spruce genome indicated asymmetric contributions from the three ancestral genomes. Phylogenetic analysis of the nuclear and organelle genomes revealed a topology indicative of ancient reticulation. Different patterns of expansion of gene families among genomes were observed and related with presumed diversifying ecological adaptations. We identified rapidly evolving genes that harbored high rates of non-synonymous polymorphisms relative to synonymous ones, indicative of positive selection and its hitchhiking effects. These gene sets were mostly distinct between the genomes of ecologically contrasted species, and signatures of convergent balancing selection were detected. Stress and stimulus response was identified as the most frequent function assigned to expanding gene families and rapidly evolving genes. These two aspects of genomic evolution were complementary in their contribution to divergent evolution of presumed adaptive nature. These more contiguous spruce giga-genome sequences should strengthen our understanding of conifer genome structure and evolution, as their comparison offers clues into the genetic basis of adaptation and ecology of conifers at the genomic level. They will also provide tools to better monitor natural genetic diversity and improve the management of conifer forests. The genomes of four closely related North American spruces indicate that their high similarity at the morphological level is paralleled by the high conservation of their physical genome structure. Yet, the evidence of divergent evolution is apparent in their rapidly evolving genomes, supported by differential expansion of key gene families and large sets of genes under positive selection, largely in relation to stimulus and environmental stress response.

**Keywords:** conifers, divergent adaptive evolution, genetic map, genome sequence, non-synonymous SNPs, phylogeny, *Picea* species, positive selection, super-scaffolds.

## INTRODUCTION

Spruces (*Picea* spp.) are coniferous evergreens widely distributed across the northern hemisphere with their origin dating back to the diversification of Pinaceae during the Cretaceous era (Savard et al., 1994). Due to their abundance, spruces are keystone species of many boreal and mountain ecosystems and they play a major role in supporting the forest industry in northern countries (Mullin et al., 2011). In Canada alone, more than 300 million spruce seedlings are planted every year (<http://nfdp.ccfm.org/en/data/regeneration.php>), and mature spruces are an important source of lumber and wood fiber, contributing a substantial portion of the country's GDP (Natural Resources Canada, 2021).

Adaptation to environment is one of the major driving forces in evolution that modifies the genetic makeup of species. In the mountainous landscape of western North America, high diversity of spruce species is found where they have adapted to a variety of climates, including glaciations of the Pleistocene repeatedly reshaping their geographic distributions (Jaramillo-Correa et al., 2009). Genomics has been key to better understanding the evolutionary biology of spruces, characterizing their genetic diversity in relation to response to biotic and abiotic factors, and providing tools to accelerate spruce breeding programs (Bousquet et al., 2021). In the context of climate change, the search for ecologically relevant markers has taken increasing relevance in the past years and several large-scale studies have contributed to better understanding some of the linkages between the genetic diversity of spruce species and climate (Hornoy et al., 2015; Yeaman et al., 2016).

The sequencing and assembly of conifer genomes requires major computational efforts due to their large genome size, ranging between approximately 7 and 37 billion base pairs (Gbp) (Ahuja & Neale, 2005). One of the major forces driving spruce genome expansion has been the accumulation of repeat DNA of diverse nature (De La Torre et al., 2014), but much remained to be deciphered in terms of the divergent evolution of conifer genomes.

Several conifer draft genome assemblies have been reported, including those of Norway spruce (*Picea abies*) (Nystedt et al., 2013), loblolly pine (*Pinus taeda*) (Zimin et al., 2017), Douglas fir (*Pseudotsuga menziesii*) (Neale et al., 2017), sugar pine (*Pinus lambertiana*) (Crepeau et al., 2017), silver fir (*Abies alba*) (Mosca et al., 2019), giant sequoia (*Sequoiadendron giganteum*) (Scott et al., 2020), and coast redwood (*Sequoia sempervirens*) (Neale et al., 2022). Here we introduce the genome

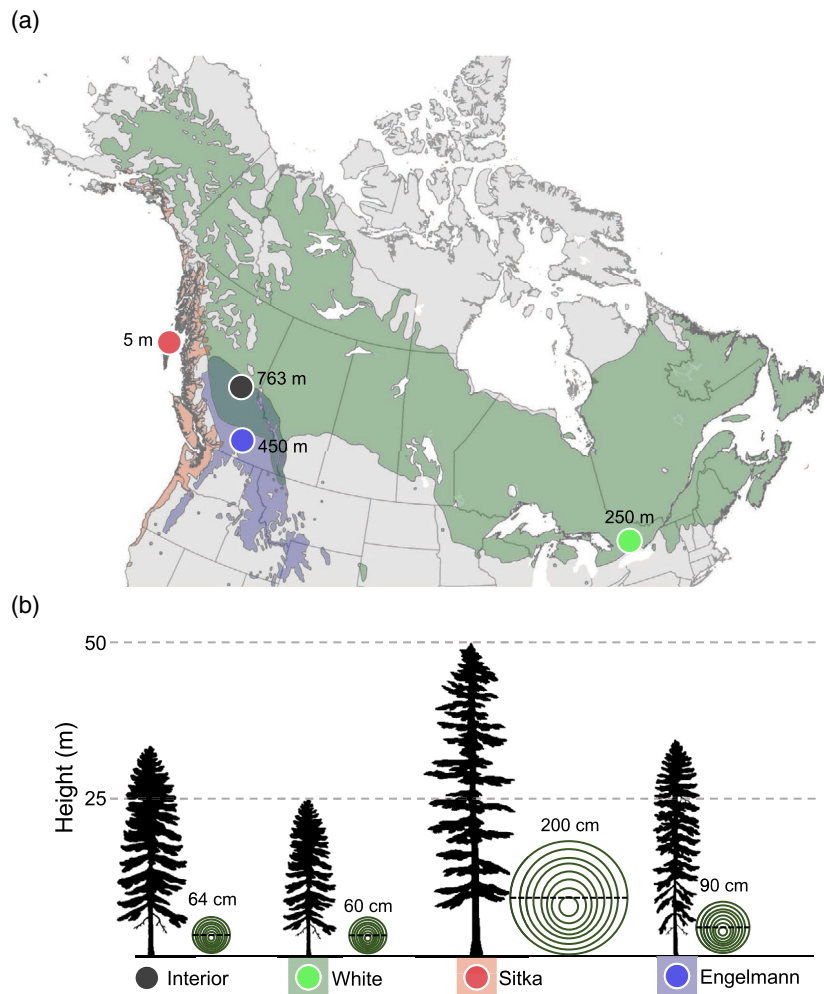
sequences of Engelmann spruce (*Picea engelmannii* [genotype Se404-851]) and Sitka spruce (*Picea sitchensis* [Q903]) along with improved genome assemblies of white spruce (*Picea glauca* [WS77111]) (Warren, Keeling, et al., 2015) and interior spruce (*Picea engelmannii* × *glauca* × *sitchensis* [PG29]) (Birol et al., 2013), a naturally occurring ingress of the former three. Engelmann spruce, which can live for up to 300 years, has a scattered and continental distribution in western North America confined to the east of the Coastal Mountains of the Rocky Mountains (Figure 1a). Sitka spruce is one of the typical trees of the Pacific coastal forests, with a natural range from Northern California to Alaska, and a life span reaching 700–800 years (Farrar, 1995). Its mature trees are the largest among the species represented here, reaching up to 55 m in height and 200 cm in diameter. Sitka spruce is well adapted to the temperate rainforest climatic conditions of the Pacific Northwest (Figure 1b). The more cold-tolerant white spruce has a vast continental range that spreads across the North American boreal forests, and it can reach 200 years of age. Interior spruce is widely used in managed forests in western Canada and the United States and it corresponds to the large area of sympatry between the previous three species.

Given the largely different distributions and ecology of these closely related spruce species, the analysis of their assembled genomes provided a unique opportunity for the discovery of features and mechanisms implicated in their divergent evolution, which could be of potential relevance for differential adaptation. We present a comparative analysis of their genomes and identify conserved features as well as elements indicative of the diversifying evolution of their genomes.

## RESULTS AND DISCUSSION

### Genome sequencing and assembly

The different spruce genomes were sequenced using a combination of short, linked, and long sequencing reads (Appendix S1). The short and linked reads were sequenced at approximately 80–110 fold coverage for all the genomes and were used for the initial *de novo* assembly. The Oxford Nanopore ONT long reads were sequenced at 2–4-fold coverage for Engelmann and Sitka spruce, and were used for scaffolding their draft genome assemblies. For each of the four spruces, the total reconstructed genome assembly size was approximately 21 Gbp, closely matching their estimated genome sizes (Table 1; Appendix S2). The scaffold NG50 length, a metric for contiguity of genome assemblies



**Figure 1.** Geographical distribution and characteristic features of spruce species of this study. (a) Geographical distribution of Engelmann spruce (*P. engelmannii*) in green, Sitka spruce (*P. sitchensis*) in red, white spruce (*P. glauca*) in green, and the range for interior spruce (*P. engelmannii* × *glauca* × *sitchensis*) in gray as indicated by the overlap of Engelmann and white spruce distributions. Colored dots on the geographic map indicate the locations of the specific Engelmann spruce (Se404-851 – blue), Sitka spruce (Q903 – red), white spruce (WS77111 – green), and interior spruce (PG29 – gray) genotypes sampled for genome sequencing, with numbers indicating the location elevations. The Sitka, white, and interior spruce trees sequenced are from their native ranges. The Engelmann spruce originated from a seed collection in New Mexico, USA, and was grown in a comparative provenance field test in British Columbia. The genotypes sequenced for Engelmann, Sitka, and white spruce were from allopatric populations distant from the area of sympatry where interior spruce is found. (b) Spruce dendrometric attributes. Maximum height of spruce species: 55 m for Sitka spruce (the tallest of the three), 36 m for interior spruce, 35 m for Engelmann spruce, and 25 m for white spruce. Maximum diameter is 200 cm for Sitka, 90 cm for Engelmann, 64 for interior, and 60 cm for white spruce.

normalized for genome size, was 355, 38, 131, and 122 kbp for Engelmann, Sitka, white, and interior spruce, respectively (Table 1, Appendix S3). All four genome assemblies

**Table 1** Genome assembly statistics and reconstruction size. The final genome assembly statistics are shown for scaffolds of ≥1 kbp. NG50 was calculated for an assembly size of 21 Gbp

	Number of scaffolds	Longest scaffold (kbp)	Scaffolds NG50 (kbp)	Genome size (Gbp)
Engelmann	946 053	6646.0	355.4	20.75
Sitka	1 770 974	1973.1	38.4	18.22
White	2 443 500	4209.0	131.3	21.58
Interior	2 064 648	3589.0	121.7	20.14

have similar completeness in the genic space, as measured by the number of reconstructed single-copy orthologs reported by BUSCO analysis (Simao et al., 2015): the ‘complete – single copy’ BUSCO ranges from 29.1 to 41.1% across the four spruces. The gene space completeness levels were comparable to those of other published conifer genomes (Crepeau et al., 2017; Neale et al., 2017; Neale et al., 2022; Nystedt et al., 2013; Scott et al., 2020; Zimin et al., 2017) (Appendix S4).

**Linkage groups assignment**

To organize the four spruce genomes into super-scaffolds matching linkage groups representative of chromosomes,

we assembled an improved white spruce genetic map that orders the relative positions of 14 727 genes represented by cDNAs along the 12 white spruce linkage groups (Pavy et al., 2008, 2012). Integrating the genetic map and the genome assemblies provided a coordinate scheme that associates centimorgan positions on the genetic map and the nucleotide positions on assigned scaffolds, building one super-scaffold per linkage group for each genome (Figure 2a). The map and assembly coordinate systems were largely collinear, as evidenced by the majority of the cDNA positions falling on a continuous, main diagonal. The few off-diagonal cDNA points represented less than 2% of the total assigned sequence for all four genomes. In contrast, up to 35% of the genetic map cDNAs co-occur with at least one other cDNA on the same scaffold (Appendix S5.1). The off-diagonal cDNA points included off-target alignments, alignments to paralogs, and possible misassemblies in either the map or genome sequences. The overall consistency between the genome sequences and the genetic map highlighted the quality and correctness of the genome assemblies, and confirmed at a fine scale the high synteny between phylogenetically distant spruce (Pavy et al., 2008) and Pinaceae genomes (Pavy et al., 2012, 2017). Up to approximately 31% of scaffolds from a single assembly (*P. engelmannii*, 6.12 Gbp) were assigned to the genetic map (Table 3, Appendix S5.2).

### Interior spruce genomic composition

The interior spruce PG29 genotype is an elite tree in a breeding program for insect resistance and other traits (Celedon et al., 2020; Warren et al., 2015a). It was previously reported as white spruce (Warren et al., 2015a), but was then hypothesized to represent an introgress of white spruce and other spruce species. We used shared single-nucleotide polymorphisms (SNPs) from the super-scaffolds to infer its genome composition (Figure 2b) and confirmed its hybrid nature but with asymmetrical contributions from three parental species. Its ancestral contributions were predominantly of white spruce genomic background (68.4%), with lesser contributions from the Engelmann (16.1%) and Sitka (12.9%) spruce genomes (Appendix S5.2). The detection of such three-species hybrids is rare (Hamilton

et al., 2015) and noteworthy, given that interior spruce is most often described as a two-species hybrid in the zone of contact between *P. glauca* and *P. engelmannii* (e.g., Haselhorst & Buerkle, 2013), and that natural two-species hybrids have also been frequently reported between *P. glauca* and *P. sitchensis* where they come into contact (e.g., Hamilton et al., 2015). As our results indicate, it is likely that small gene leakage from a third parental species has been difficult to detect without assessing natural hybrids with an informative genome-wide detection method.

### Genome annotation

All four genomes were annotated with the MAKER2 (Holt & Yandell, 2011) pipeline, and only gene models supported by direct evidence such as RNA-seq, cDNA, and manually annotated predicted protein sequences were considered. We derived 34 365, 30 324, 30 410, and 28 943 high-confidence genes for Engelmann, Sitka, white, and interior spruce, respectively (Table 2), selected for protein completeness and gene length to filter putative pseudogene annotations. The high-confidence genes contained known protein domains selected based on Pfam (Finn et al., 2014) analysis and BLAST (Shiryev et al., 2007) alignments against evidence-based proteins. The content and completeness of functional domains were compared between the protein sequences from the predicted gene models (Appendices S6.1 and S6.2). We observed little variation in the number of BUSCO 'complete' core genes and Pfam domains between the spruce genome annotations reported herein. We plotted the length distributions of exons and introns found in our gene annotations for genes longer than 10, 25, 50, 100, and 250 kbp (Appendix S6.3). We observed longer intron and gene sizes in white and interior spruce when compared to Engelmann and Sitka spruce. The distribution of exon lengths remained similar for the gene lengths examined. Gene homology was tested through reciprocal best BLAST hit (RBH) in each pair of spruce taxa; approximately 30–40% of the annotated proteins had a protein homolog in each comparison (Appendix S6.4).

Overall, repeats accounted for approximately 70% of each of the four genomes. Repeat composition was largely

**Figure 2.** Genomic structure and sequence similarity. (a) Collinearity of super-scaffolds and the genetic map. The mapped cDNAs were aligned to the genome assembly of each species, and the scaffold that best aligned to each cDNA was identified. These scaffolds were stitched together in the order dictated by the genetic map, yielding one super-scaffold per linkage group representative of each of the 12 chromosomes. The cDNAs were then realigned to the scaffolds. The plot shows the start positions of the aligned cDNAs versus their positions on the genetic map. (b) Shared SNP composition in introgressed interior spruce genotype PG29. For each of the interior spruce PG29 linkage group super-scaffolds, we plotted the proportion (0–100%) of SNPs unique to interior spruce (gray) and those shared with Sitka (red), Engelmann (blue), and white spruce (green) within each 1-Mbp tile, while moving the frame over by 100-kbp window increments (track 2 from the rim). Shared SNP composition densities in white spruce (green, track 3), Engelmann spruce (blue, track 4), and Sitka spruce (red, track 5) were plotted using the `circos.genomicDensity` function of the `circize` R package (v0.4.8, `window.size = 10E6`). Ideograms (track 1) show sections of the linkage groups having the highest shared SNP composition within each overlapping 5-Mbp tile. Regions of identical densities between two or more species were not assigned and are shown as white gaps in the ideogram. By varying the window size, we estimated the base contribution proportions of white, Engelmann, and Sitka spruce to be approximately (average  $\pm$  SD)  $68.4 \pm 11.2$ ,  $16.1 \pm 5.0$ , and  $12.9 \pm 5.1\%$ , respectively, as indicated in the center track (track 6). Unassigned portions ( $2.6 \pm 2.5\%$ ) are shown in light gray.



**Table 2** Genome annotation statistics for the high-confidence genes from MAKER (Holt & Yandell, 2011). The annotated mRNAs are shown as total annotated and as single-exon mRNAs. The length is calculated for the total mRNAs, exons, and total proteins, shown as an average (bp or aa)  $\pm$  standard deviation

	Total annotated genes	Total annotated mRNAs	Single-exon mRNAs	Average total mRNA length (bp)	Average total exon length (bp)	Average total protein length (aa)
Engelmann	34 365	60 224	14 804	1284 $\pm$ 857	272 $\pm$ 346	338 $\pm$ 246
Sitka	30 324	58 175	13 002	1347 $\pm$ 900	264 $\pm$ 326	345 $\pm$ 256
White	30 410	56 535	12 833	1275 $\pm$ 838	257 $\pm$ 320	323 $\pm$ 234
Interior	28 944	62 397	13 043	1250 $\pm$ 865	261 $\pm$ 315	311 $\pm$ 240

**Table 3** Genetic map integration. Total number of scaffolds, scaffolds NG50, and genome size assigned to the genetic map, part of the super-scaffolds assignments, for the four genomes

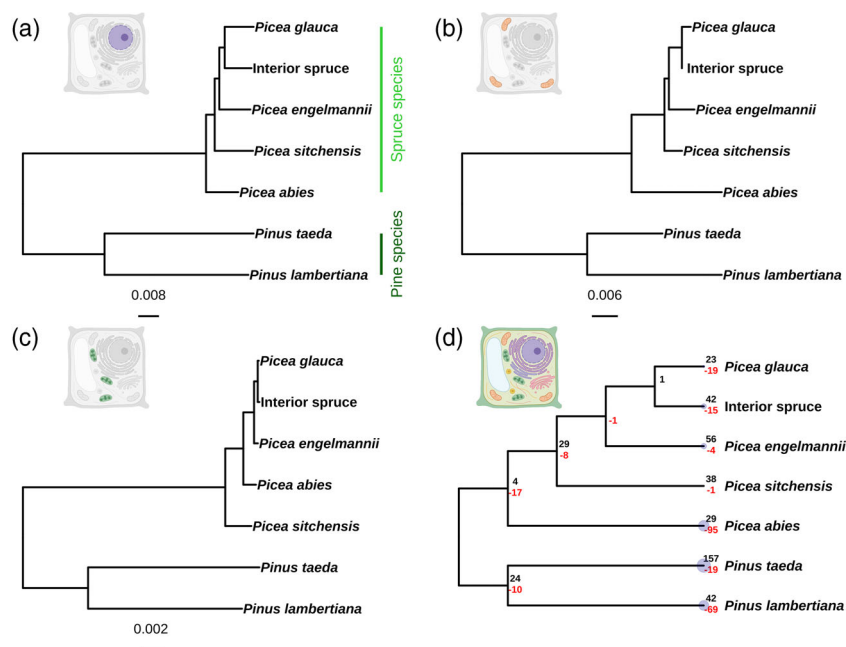
	Number of scaffolds	Scaffolds NG50 (kbp)	Total genome size (Gbp)
Engelmann	11 262	920.29	6.12
Sitka	12 897	246.81	1.94
White	12 627	664.97	3.86
Interior	12 483	482.68	3.25

consistent across the four spruces (Appendix S7) and other published conifer genomes (Nystedt et al., 2013; Stevens et al., 2016). Long terminal repeat (LTR) transposons (LTR-gypsy, LTR-copia, and unclassified LTR) covered more than 60% of the assemblies, similar to the high interspersed repeat content in the sugar pine (*P. lambertiana*,

approximately 67%) and Norway spruce (*P. abies*, approximately 50%) genomes (Nystedt et al., 2013; Stevens et al., 2016).

### Phylogenetic analysis

Whole organellar and nuclear (nc) genome sequences were used to construct phylogenies comprising five spruce and two pine species (Figure 3). In agreement with previous analyses relying on rRNA and plastid (pt) genes (Chaw et al., 1995; Wang et al., 2000), spruces and pines formed distinct clades separated by the longest internode in each phylogeny. Notably, the topologies of the spruce clades in the nc and mitochondrial (mt) phylogenies were congruent, substantiating previous results that were based on individual gene sequences (e.g., intron of mt *NADH* genes, the nc *4CL* gene) or transcriptome data (Bouillé et al., 2011; Feng et al., 2019; Lockwood et al., 2013).



**Figure 3.** Conifer genome phylogenies. The sequence divergence between the (a) nuclear (nc), (b) mitochondrial (mt), and (c) plastid (pt) genome assemblies for each conifer species was estimated by comparing the k-mer content of the respective genome assemblies obtained using the neighbor-joining method. Panel (d) shows the topology of the summary tree obtained by analysis of 780 orthologs with STAG and ASTRAL-III, with numbers on interior nodes and terminal branch tips indicating the number of rapidly evolving gene families ( $P < 0.001$ ) in black and red for expanding and contracting gene families, respectively. Cell diagrams were made in ©BioRender (biorender.com).

The topology of the pt phylogeny differed from those of the nc and mt genomes in the relative positions of Norway and Sitka spruce. In particular, Sitka spruce appeared as an outgroup to all other spruce sequences including Norway spruce. This result is consistent with an earlier report, which showed that spruce phylogenies based on maternally inherited mitochondrial DNA resulted in topologies that are more consistent with geographical distributions, while the paternally inherited chloroplast DNA may yield different phylogenetic trees due to ancient recombination, which for Sitka spruce may result from long-distance gene flow driven by pollen dispersal (Bouillé et al., 2011). The discordant pt phylogeny obtained here was also confirmed with an orthogonal phylogeny method (Appendix S8). Engelmann spruce appeared as a sister group to white and interior spruce in the nc, mt, and pt phylogenies (Figure 3a–c), reflecting the close genetic proximity between Engelmann and white spruce which naturally hybridize and introgress in their large area of sympatry (Hamilton et al., 2015; Haselhorst & Buerkle, 2013). In the nc and mt phylogenies, the lineage leading to Sitka spruce appeared as a close sister group to Engelmann and white spruce, reflecting their parapatric distributions and the reported natural introgression between Sitka spruce and the two other species (Hamilton et al., 2015). This is also in agreement with the sizeable genomic contribution of Sitka spruce to the genome of the introgressed interior spruce described above. The lineage leading to Norway spruce was more remote and a sister group to North American spruces in all but the pt genome sequences. This is consistent with the estimates of an ancient divergence time in excess of 10 million years between the lineages leading to white and Norway spruce (Bouillé & Bousquet, 2005). The close position of interior spruce with white spruce in all tree topologies reflects the large genomic contribution of white spruce to the genome of the introgressed interior spruce as described above. Overall, the phylogenetic patterns observed here reflected the fraction of shared SNPs at orthologous nc genes between Engelmann/white spruce, Sitka/white spruce, and Norway/white spruce pairs, with respective proportions of 64%, 22%, and 12% (Pavy, Gagnon, et al., 2013), thus showing a declining pattern with increasing phylogenetic distance (as depicted in Figure 3a).

Predicted protein sequences were clustered in orthogroups (gene families) for comparative analysis across different spruce and pine species. Out of a total of 22 397 orthogroups, 3165 were shared between the different species, 907 were shared between all spruces, and 1215 were shared between the North American spruce species (Appendix S9). A set of 780 proteins from the 3164 orthogroups represented in all species was used to reconstruct phylogenetic relationships based on protein sequences inferred from gene models. The protein family

phylogeny obtained (Figure 3d) had the same topology as we observed for nc and mt phylogenies (Figure 3a,b).

### Expansions and contractions of gene families

To first test for divergent genome evolution, changes in gene families were compared against the phylogenetic tree topology based on protein sequences (Figure 3d) to assess possible links between long-term evolutionary change and differential adaptation of the different spruce species to their environments. To this end, we first identified gene families evolving at a rate significantly different between the parent and child nodes (Table S1). Contrary to the common ancestor of North American spruces, we observed a substantial number of gene family contractions in the lineage leading to Norway spruce (+29/–95) (Figure 3d), where +/- numbers in the parentheses indicate the number of expanding/contracting families. Among the phylogenetically closely related North American spruces, the taxon with the largest number of rapidly evolving gene families was Engelmann spruce with 60 (+56/–4) gene families, followed by interior, white, and Sitka spruce with 57 (+42/–15), 42 (+23/–19), and 39 (+38/–1) gene families, respectively (Figure 3d). As the assembly contiguities improve, it should be possible to assess whether expanding gene families are physically tandemly arrayed (Pavy et al., 2017) and if other molecular mechanisms are involved in their expansion, such as translocations (Guillet-Claude et al., 2004). It is also noteworthy that many more gene family expansions versus contractions were detected in Engelmann, Sitka, and interior spruce than in white spruce. The former taxa may have been under additional selective pressures since their separation, given that they have historically been facing a more heterogeneous landscape in western North America (Figure 1a). This is in sharp contrast with what is typically observed east of the Rocky Mountains on the continental Canadian shield where white spruce is mostly found, including the individual representative tree used herein for genome sequencing, which pertains to the eastern North American phylogeographic lineage (Figure 1a).

Among the top expanding gene families, several were of unknown functions and represented by several domains of unknown function (DUFs; DUF4283, DUF4219, DUF659), several were related to retroelements (transposase family tnp2, GAG-pre-integrase domain, reverse transcriptase), and the others represented various functions well known in plants (Table S1). These proteins of known function were associated to gene ontology (GO) terms, and enrichment tests highlighted some terms found enriched only among the expanding gene families and not among the contracting gene families (Table S2). For example, at a *P*-value of <1E–5, the 18 GO terms enriched among expanding families included eight terms related to metabolism or its regulation and six terms related to responses to various

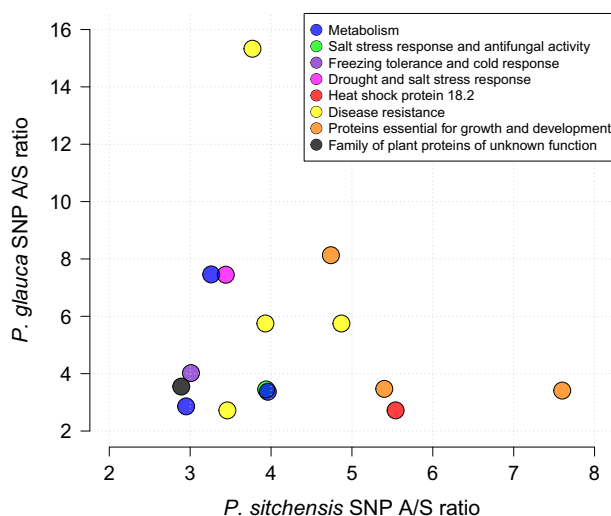
stimuli, along with four other terms (embryo development, protein-containing complex subunit organization, multi-organism process, reproduction). Interestingly, among families related to environmental responses, several families were involved in various stress responses (universal stress protein family), the heat stress response (EF-Tu elongation factor thermo unstable and heat shock proteins), resistance to pathogens (NB-ARC), or ubiquitination, which regulates protein degradation and turnover. In expanding gene families noted in Engelmann, interior, and Sitka spruce, the two biological processes enriched with the most significant  $P$ -value were response to abiotic stress and reproduction. The same two processes were still significantly enriched but were ranking lower for expanding gene families in white spruce, for which the top-ranking processes were related to nitrogen compounds and macromolecule localization.

### Rapidly evolving genes under positive and balancing selection

To further test for genome-wide divergent evolution, we used transcriptome SNP datasets from resequencing for two spruce species, Sitka spruce native to the temperate rain forests of the Pacific Northwest coast and the continental white spruce, adapted to much harsher climate conditions. Among the different spruce species included in this study, they occupy the most disjunct geographic distributions and the most contrasted environmental conditions (Figure 1a). To detect signatures of relaxed selective constraints, we estimated the SNP A/S ratio for each gene in each species. The SNP A/S ratio is the rate of non-synonymous SNPs (A), corresponding to amino acid polymorphisms, over the rate of synonymous SNPs (S), which are assumed to be of more neutral nature. A ratio below 1 is indicative of variable levels of purifying selection while a ratio over 1 indicates relaxed selective constraints involving positive selection (Fay et al., 2001; Liu et al., 2008). We then identified outlier genes with high SNP A/S ratios well above 1, and compared trends between these two ecologically contrasting species. Out of 17 352 genes in Sitka spruce and 11 008 genes in white spruce, a set of 913 genes (3.2%) were found as statistically significant outliers ( $P < 0.05$ ) based on their high SNP A/S ratios (threshold values of 2.51 for Sitka spruce and 2.85 for white spruce) (Appendix S10 and Figure S10.1). In these genes for both species, we found a low occurrence of synonymous SNPs (Appendix S10 and Figure S10.2), which is indicative of hitchhiking effects from positive selection on non-synonymous SNPs (Jensen & Bachtrog, 2010; Wiehe & Stephan, 1993). This may further implicate codon bias related to increased translational efficiency (Guo et al., 2017). More frequent in Sitka spruce (748 genes, 4.3%) than in white spruce (165 genes, 1.5%), these outlier genes represented a great diversity of predicted functions

(Appendix S10 and Figures S10.3–S10.5) and gene families (Appendix S10 and Figure S10.6). Notably, 36% of outlier genes were annotated as involved in response to biotic or abiotic stress, which were the two most represented GO classes (Appendix S10 and Figure S10.3). Second, 23% of outlier genes were annotated as having a role in developmental processes, with reproductive structure development, post-embryonic development, and cell differentiation being the most represented among these.

Only 15 gene outliers were shared by both Sitka spruce and white spruce, indicating that a small proportion of outlier genes could be the result of convergent adaptive evolution. Of these, 14 could be assigned GO terms and a majority (eight) was related to resistance to biotic and abiotic stress (Figure 4 and Table S3). Notably, the pair of genes exhibiting the highest SNP A/S values for both species showed high similarity to the *rps2* gene, which confers pathogen resistance in *Arabidopsis thaliana* (Appendix S10). Non-synonymous SNPs in the spruce *rps2*-like genes were specific to each species but in both of them, they resulted in amino acid polymorphisms located in coils or alpha-helices outside of the hydrophobic clusters forming the core of the protein, thus leading to similar functional protein changes (Appendix S10 and Figure S10.7). Such a pattern is strongly indicative of convergent evolution at the molecular level, which could be linked to balancing selection. To assess this, we looked at the distribution of allelic frequencies of non-synonymous SNPs of this gene in both Sitka and white spruce. Intermediate allelic frequencies could be detected (Appendix S10



**Figure 4.** SNP A/S ratios for 14 pairs of outlier genes with high ratio values ( $P < 0.05$ ) shared by white spruce (*P. glauca*) and Sitka spruce (*P. sitchensis*) (Table S3 and Appendix S10). Threshold values for outlier significance ( $P < 0.05$ ) were SNP A/S ratio  $> 2.85$  for white spruce and SNP A/S ratio  $> 2.51$  for Sitka spruce. Proteins were classified according to functions reported in other plant species through sequence and literature searches.



and Figure S10.8), which are a hallmark signature of balancing selection (Siewert & Voight, 2017). Also, synonymous SNPs were absent from this gene in both species, which indicates strong hitchhiking effects from positive selection (Jensen & Bachtrog, 2010; Wiehe & Stephan, 1993). Altogether, these results indicate that positive and balancing selection should not be considered as mutually exclusive and that they could evolve in a concerted manner at the molecular level.

A further comparison of our results with a restricted set of highly significant outlier genes detected in a previous population genomics study in white spruce also highlighted a set of six outlier genes identified as top candidates for adaptation to climate after a genome-wide association genetics scan (Hornoy et al., 2015) (Appendix S10 and Table S4). Of those, a CBL-interacting protein kinase and S-Phase Kinase-Associated Protein 2A (SKP2A) were also implicated in local adaptation in lodgepole pine (*Pinus contorta*) and interior spruce (Yeaman et al., 2016). The CBL-interacting protein kinase family has also been found to be involved in response to stresses in angiosperms (Singh et al., 2020), while SKP2A is an auxin-binding protein that connects auxin signaling and cell division (Jurado et al., 2010). These annotations are indicative of potentially important adaptive roles which appear to be largely shared across plants.

#### Relationships between expanding gene families and rapidly evolving genes

Similarly as for expanding gene families, our results indicate that the sets of rapidly evolving genes were largely related to stimulus and stress response. Proportionally more rapidly evolving genes were observed in the coastal Sitka spruce than in the continental white spruce genome, paralleling the finding of many more gene family expansions than contractions observed in Sitka compared to white spruce. These patterns could be related to different levels of environmental pressures and biotic interactions since species separation. They indicate that much divergent evolution has occurred between these ecologically contrasting species, likely related to differential adaptation.

A limited set of 12 rapidly evolving outlier genes also belonged to expanding gene families. Among them, a gene encoding a molybdate transporter was an outlier shared between Sitka and white spruce, belonging to a gene family significantly expanded in white spruce and to one of the superfamilies that were most represented among outlier genes (MFS transporter in Figure S10.6 in Appendix S10). At the same time, the limited overlap observed between expanding gene families and rapidly evolving genes indicates that they would represent largely non-redundant thus complementary molecular mechanisms of adaptive evolution in spruces.

## CONCLUSIONS

The spruce genome assemblies provided here represent the most contiguous spruce giga-genomes to date with super-scaffolds now matching chromosomal linkage groups. Our analyses indicate that overall, much conservation exists in the structure and components of spruce genomes, which should facilitate comparative analyses at the micro- and macroscale with other conifer genomes and the monitoring of population genetic processes such as introgression. At the same time, much variation was observed among these closely related species in terms of gene family contractions and expansions, and in terms of high rates of non-synonymous SNPs indicative of positive selection in specific genes. The limited overlap observed between expanding gene families and rapidly evolving genes indicates that they represent largely non-redundant thus complementary mechanisms of adaptive genomic evolution. Given that their annotations indicated relationships with key metabolic and physiological processes largely related to response to biotic and abiotic stress, such divergent evolution of spruce species appears to be indicative of adaptation to their different environmental niches mostly defined by climate and biotic interactions. Mapping the geographical variation of SNP allelic frequencies in rapidly evolving genes and expanding gene families, as well as conducting association studies with key phenotypic traits related to biotic (e.g. Lamara et al., 2018; Whitehill et al., 2019) and abiotic stress (e.g. De La Torre et al., 2022; Depardieu et al., 2021), should help identify more distinctive trends of adaptation in this group of long-lived woody plants. At the more fundamental level, these studies should also assist in further characterizing the role of balancing selection in maintaining diversity at the molecular and genomic levels (Fijarczyk & Babik, 2015). In the context of climate change, the analyses enabled by these genome sequences will facilitate the further delineation of the key features and mechanisms of tree adaptive evolution, and help improve our monitoring of forest health and its management for the future.

## EXPERIMENTAL PROCEDURES

### Sample collection and sequencing

Apical shoot tissues were collected from a single individual as representative genotype for each species. The locations and the year of collection are listed in Appendix S11. For each spruce genotype, genomic DNA was extracted from 60 g tissue by BioS&T (<http://www.biost.com/>, Montreal, QC, Canada) using an organelle exclusion method yielding 300 µg of high-quality purified ncDNA. The DNA samples were used to build sequence libraries as described below for each platform and protocol.

**Illumina HiSeq and MiSeq.** DNA quality was assessed by spectrophotometry and gel electrophoresis before library construction. DNA was sheared for 45 sec using an E210 sonicator

(Covaris, Woburn, MA, USA) and then analyzed on 8% PAGE gels. The 200–300-bp (for libraries with an insert size of 250 bp) or 450–550-bp (for libraries with an insert size of 500 bp) DNA fractions were excised and eluted from the gel slices overnight at 4°C in 300 µl of elution buffer and purified using a Spin-X Filter Tube (Fisher Scientific, Waltham, MA, USA) and ethanol precipitation. Genome libraries were prepared using a modified paired-end tag (PET) protocol supplied by Illumina Inc. This involved DNA end repair and formation of 3' adenosine overhangs using the Klenow fragment of DNA polymerase I (3'-5' exonuclease minus) and ligation to Illumina PE adapters (with 5' overhangs). Adapter-ligated products were purified on QIAquick spin columns (Qiagen, Germantown, MD, USA) and amplified using Phusion DNA polymerase (NEB) and 10 PCR cycles with PE primers 1.0 and 2.0 (Illumina, San Diego, CA, USA). PCR products of the desired size range were purified from adapter ligation artifacts using 8% PAGE gels. DNA quality was assessed using an Agilent DNA 1000 series II assay (Agilent, Santa Clara, CA, USA) and DNA was quantified using a Nanodrop 7500 spectrophotometer (Nanodrop, Waltham, MA, USA). DNA was subsequently diluted to 8 nM. The final concentration was confirmed using a Quant-iT dsDNA HS assay kit and a Qubit fluorometer (Invitrogen, Waltham, MA, USA). The Engelmann spruce DNA samples were processed with a long-insert PCR protocol with a 900-bp whole genome library constructed following a previously described protocol with minor modifications (Jones et al., 2016; Tsang et al., 2019). Briefly, 5 µg of genomic DNA was subjected to shearing by sonication (Covaris LE220) using a duty factor of 5 and a peak incident power of 450 W for 70 sec. The sonicated DNA products were fractionated in a 6% PAGE gel to recover fragments greater than 700 bp for library preparation. These PCR-free libraries were sequenced with paired-end 150-base reads on an Illumina HiSeqX platform using V4 chemistry according to the manufacturer's recommendations.

**MPET.** The mate paired-end tag (MPET, a.k.a. jumping) libraries were constructed using 4 µg of genomic DNA with the Illumina Nextera Mate-Pair library construction protocol and reagent (FC-132-1001). The genomic DNA sample was simultaneously fragmented and tagged with a biotin-containing mate pair junction adapter, which left a short sequence gap in the tagged DNA. The gap was filled by a strand displacement reaction using polymerase to ensure that all fragments were flush and ready for circularization. After an AMPure Bead cleanup, size selection was done on a 0.6% agarose gel to excise 6–9-kbp and 9–13-kbp fractions, which were purified using a ZymoClean Large Fragment DNA Recovery Kit. The fragments were circularized by ligation, followed by digestion to remove any linear molecules, and left circularized DNA for shearing. The sheared DNA fragments that contain the biotinylated junction adapter (mate pair fragments) were purified by means of binding to streptavidin magnetic beads, and the unwanted unbiotinylated molecules were washed away. The DNA fragments were then end repaired and A-tailed following the protocol and ligated to indexed TruSeq adapters. The final library was enriched by a 10-cycle PCR and purified by AMPure bead clean-up. Library quality and size were assessed by an Agilent DNA 1000 series II assay and the KAPA Library Quantification protocol. The two fractions were pooled for paired-end 100-bp sequencing using an Illumina HiSeq2000 platform.

The construction of the 12-kbp mate pair libraries was achieved by a hybrid 454/Illumina procedure. Briefly, 50 µg of genomic DNA was fragmented for 20 cycles at speed code 12 using a Hydroshear (Digilab, Marlborough, MA, USA) equipped with a large assembly module. The fragmented DNA was loaded on a 1% agarose gel, and fragments from 12 to 18 kbp were extracted.

Biotinylated circularization adapters from the GS Titanium Paired-end Adaptor set (454 Life Sciences/Roche, Branford, CT, USA) were added to ends of the gel-extracted fragments. Homologous recombination of the ends was performed with Cre recombinase (New England Biolabs, Ipswich, MA, USA), and linear molecules remaining in solution were removed with Plasmid Safe (Epicenter, Madison, WI, USA). Circular molecules were fragmented using GS Rapid Library Nebulizers (454 Life Sciences/Roche, Branford, CT, USA), and fragment end repair followed by A-tailing was performed with the GS Rapid Library preparation kit (454 Life Sciences/Roche, Branford, CT, USA). TruSeq Adaptors (Illumina, San Diego, CA, USA) were ligated to the repaired/A-tailed ends. Biotinylated fragments were enriched using Streptavidin-coupled Dynabeads (Life Technologies, Grand Island, NY, USA) and amplified by PCR using Illumina-specific primers.

**10× Genomics Chromium.** High-molecular weight (HMW) DNA was extracted using the Nucleospin 96 plant kit (MACHERY-NAGEL) or the cetyl trimethylammonium bromide (CTAB) method and the HMW genomic DNA extraction protocol as detailed in the Chromium Genome Reagent Kits Version 2 User Guide (PN-120229). Integrity of the DNA was assessed by pulsed field gel electrophoresis with the majority of DNA fragments over 50 kbp in length. The fragment size was confirmed *in silico* after assembly; the weighted mean molecule length was 46 kbp. A microfluidic partitioned library was created using the Chromium system from 10× Genomics (10× Genomics, Pleasanton, CA, USA). Gel beads-in-EMulsion (GEMs) were produced by combining DNA, Master Mix, and partitioning oil in the 10× Genomics Chromium Controller instrument with the microfluidic Genome Chip (PN-120216) (10× Genomics). The DNA in each GEM underwent isothermal amplification as a barcode was added to each fragment. Barcoded fragments then underwent Illumina library construction (as per the Chromium Genome Reagent Kits Version 2 User Guide [PN-120229]). The resulting library was assessed for quality using an Agilent 2100 Bioanalyzer (Santa Clara, CA, USA) and a DNA 1000 assay.

**Oxford Nanopore technologies.** The genomic DNA library was constructed using the SQK-LSK109 Ligation Library Kit from Oxford Nanopore. Liquid handling was performed using wide bore tips to avoid physically breaking the DNA. Six micrograms of HMW DNA was gently sheared using 10 passes up and down through a 26-gauge needle (cat. no. 309625, BD medical, Franklin Lakes, NJ, USA) and a size selection step was completed using a 0.35:1 ratio of PCRclean DX magnetic beads to DNA (cat. no. C-1003-450, ALINE Biosciences, Woburn, MA, USA). NEB Ultra II (cat. no. E7646A, New England Biolabs, Ipswich, MA, USA) was used for end repair and 30 A-tailing. NEB Blunt/TA Ligation Master Mix (M0367S) was used to ligate the Oxford Nanopore adapters. A final size selection step at a ratio of 0.4:1 (magnetic beads to library) was performed to eliminate smaller molecules. MinION sequencing proceeded using the FLO-MIN106 (R9 Version) flow cell and the software programs MinKnow 1.13.1 and GUI 2.0.13.

## Genome assembly

**Assembly of the Engelmann spruce (*Picea engelmannii*; genotype Se404-851) genome.** Prior to the *de novo* genome assembly, the paired-end short reads were merged using Konnector and ABySS-mergепairs v2.0.1 in the ABySS package (Jackman et al., 2017). First, Bloom filters with k-mer sizes of 75–245 (step = 10) were built using ABySS-Bloom and all short reads. Cascading Konnector runs were then performed using these

Bloom filters, starting at the highest k-mer size. All reads not merged by Konnector at any k were then input to ABySS-mergepairs, where merges were attempted by trimming the reads progressively from the 3' end (0–40 bp, step = 10) in successive runs. Read pairs that remained unmerged were used in the assembly at their original lengths.

Following read merging, all merged and unmerged reads were assembled using ABySS v2.1.4 (Jackman et al., 2017) (B = 500G, n = 5, N = 5–20, l = 50, kc = 4) with a variety of k-mer sizes (k = 112, 128, 144, 160). The k = 144 assembly was then scaffolded using the k = 112, 128, and 160 assemblies as long-range sequence and linkage evidence for Cobbler v0.5.1, RAILS v1.4.1 (Warren, 2016) (–d500, –i0.99), and LINKS v1.8.5 (Warren et al., 2015b) (–k26, –l10, –a0.3; –d1000, 2500, 5000, 7500, 10 000, 15 000, 20 000). Next, MPETs were emulated from the Nanopore data by extracting fragments of various sizes (5, 10, 15 kbp) from the Nanopore reads with a shift of one eighth the fragment size, and extracting 500-bp pseudoreads from the fragment ends. Then, abyss-scaffold v2.1.4 (–n = 5–20) was used to scaffold the assembly with the simulated MPET reads. Next, the linked read data were used to first break the assembly at putative misassemblies using Tigmint v1.1.2 (Jackman et al., 2018) (span = 5, window = 1000) and then to scaffold the resulting assembly using ARCS v1.0.6 (Yeo et al., 2018) (–c3, –m50, –20 000, –z500, –a0.9, –l3). Finally, gap filling was performed using Sealer v2.1.4 (Paulino et al., 2015) (–L150, –P10, –k75-115 [step = 10]).

**Assembly of the Sitka spruce (*Picea sitchensis*; genotype Q903) genome.** All short read data were merged as described above for the Engelmann spruce assembly. Then, all merged and unmerged short reads were assembled with ABySS v2.0.1 (Jackman et al., 2017) (k = 96, 128, 144, 176; kc = 4, n = 5, l = 50). The assembly with the highest N50 (k = 144) was then scaffolded using the other three ABySS assemblies as sources of long-range linkage evidence. First, gap filling and scaffolding were performed using Cobbler v0.3 and RAILS v1.2 (Warren, 2016) (–d500, –i0.99). Next, the same long sequences from the three ABySS assemblies (k = 96, 128, 176) were used to scaffold the assembly with LINKS v1.8.5 (Warren et al., 2015b) (–k26, –a0.3, –l10; –d1000, 2500, 5000, 7500, 10 000; –t100, 50, 30, 20, 15). The resulting draft assembly was then scaffolded again using LINKS v1.8.6, this time using the corrected Nanopore reads as the long-range linkage evidence (–k23, –l10, –a0.3, –d2500-10 000 [step = 2500], 10 000–50 000 [step = 5000]; –t50, 20, 18, 10, 5, 2, 1). Following the LINKS scaffolding steps, the Nanopore reads were also used as long-range sequence and linkage evidence for Cobbler v0.5 and RAILS v1.4 (–d500, –i0.85, –g500). Following scaffolding using the long sequences, the assembly was broken at potentially misassembled regions with Tigmint v1.1.0 (Jackman et al., 2018), using the linked reads as evidence (span = 2, window = 1000). The linked reads were then used for scaffolding with ARCS v1.0.3 (Yeo et al., 2018) (–c4, –l4, –a0.9, –z500, –s90, –m30-20 000, –e30000). Finally, gap filling was performed on the draft assembly using Sealer v2.0.1 (Paulino et al., 2015) (–L150, –P10, –k75-115 [step = 10]).

### Improving the white spruce (*Picea glauca*; genotype WS77111) genome assembly

Using linked reads from the 10× Genomics Chromium platform, Tigmint v1.1.2 (Jackman et al., 2018) (span = 2, window = 1000) was used to break the v1 WS77111 assembly (GCA\_000966675.1) at putative misassemblies. Then, this corrected assembly was scaffolded using ARCS v1.0.1 (Yeo et al., 2018) (–c3, –l3, –a0.9, –z3000, –s90, –m50-20 000, –e30000). Following ARCS scaffolding,

gap filling was performed using Sealer v2.0.1 (–L150, –P10, –k75-115 [step = 10]) (Paulino et al., 2015). Next, using the Kollecator (Kucuk et al., 2017) reconstructions of target genes (–d500, –i0.99), Cobbler v0.3 and RAILS v1.2 (Warren, 2016) were run for gap filling and scaffolding, respectively, to refine these genic regions. The resulting assembly was then scaffolded with full-length cDNA sequences using ABySS v2.0.1 (Jackman et al., 2017). To correct misassemblies introduced by the final scaffolding steps, a final Tigmint (span = 2, window = 1000) run was performed.

### Improving the interior spruce (*P. engelmannii* × *glauca* × *sitchensis*; genotype PG29) genome assembly

To improve the v4 PG29 genome assembly (GCA\_000411955.4), Tigmint v1.1.2 (Jackman et al., 2018) was used with Chromium linked reads to break the assembly at putative misassemblies (span = 2, window = 1000), followed by scaffolding with ARCS v1.0.3 (Yeo et al., 2018) (–c3, –l4, –a0.9, –z500, –m50, –20 000, –e30000, –s90). Next, automated gap filling was performed using Sealer v2.0.1 (Paulino et al., 2015) (–L150, –P10, –k75-115 [step = 10]).

### Construction of custom repeat libraries and repeat masking

Masking repetitive DNA elements prior to performing gene predictions helps to minimize spurious predictions. A custom repeat library was built for each spruce genome by combining *de novo* identified elements with curated elements from RepBase (Bao et al., 2015). LTR elements were identified using LTR\_retriever v1.3 (Ellinghaus et al., 2008) with candidate sequences provided by LTRharvest v1.5.9 (GenomeTools; –similar90, –vic10, –seed20, –minlenltr100, –maxlenltr7000, –mintsd4, –maxtsd6, –motifmis1, with and without –motifTGCA) (Ellinghaus et al., 2008; Gremme et al., 2013) and LTR\_FINDER v1.06 (Xu & Wang, 2007) (–D15000, –d1000, –L7000, –l100, –p20, –M0.9). Redundant elements from LTR\_retriever were removed by cd-hit-est v4.6.6 (Fu et al., 2012) (–c0.8, –G0.8, –s0.9, –aL0.9, –aS0.9, –M0). Additional repeat elements were predicted by RepeatModeler v1.0.8 (<http://www.repeatmasker.org/RepeatModeler/>). These *de novo* elements were combined with RepBase v22.08 (Bao et al., 2015) to yield each final custom library of repeat elements.

Instances of the repeat elements from the custom library were detected in the genome assemblies using RepeatMasker open-4.0.7 and were annotated according to the labels provided by the LTR prediction tools, RepBase, or RepeatMasker itself. The genomes were annotated for their repeat content using custom repeat libraries.

### Quality assessment of genome assemblies and genome annotations

Genome assembly quality was assessed by mapping the reads to the respective genome assemblies with BWA-mem (Li, 2013) and estimating the percentage of mapped reads (Appendix S13.2). Depending on the species, the reads mapping rate ranged between 96 and 98%. The genome completeness in the gene space was assessed with BUSCO v5.1.2 (Simao et al., 2015) using the Embryophyta library odb10 (n = 1614) with the –long option.

Genome completeness was further assessed by mapping the RNA-seq reads with hisat2 (Kim et al., 2015) to the respective genome assemblies (Appendix S13.3). The mapping rate was above 85% for the Sitka and white spruce read libraries and approximately 75–85% for interior spruce. We also mapped the assembled RNA-seq transcripts with length longer than 200 bp to the genome assemblies with GMAP v2017-11-15 (Wu et al., 2016) using the pooled assemblies from species-specific studies. A

transcript was considered as mapped to the genome when aligning with  $\geq 95\%$  identity and  $\geq 95\%$  query coverage (Appendix S13.4). The resulting alignment rate was  $< 50\%$  for the most of the samples. The genome annotation was assessed with BUSCO v5.1.2 executed in protein mode ( $-m$  protein option) and DOGMA v3.4 (Dohmen et al., 2016), evaluating conserved Pfam domains. The genome annotation assessed with DOGMA used 948 single-domain conserved domain arrangements (CDAs) and 491 multiple-domain CDAs across eukaryotes.

### Genome annotation

The genome annotation was limited to contigs containing at least 1 kbp of non-repeat sequences and at least one putative complete gene based on alignment with spruce transcriptomics sequences.

The genomes were annotated with the MAKER v2.31.10 pipeline (Holt & Yandell, 2011) run with *ad hoc* trained parameters for gene predictors together with transcriptomics and proteomics evidence. Augustus metaparameters were optimized with a semi-automatic training protocol and evaluated by splitting into training and test sets. SNAP and Augustus were trained with high-quality gene models generated by a preliminary run of MAKER and selected by exon annotation edit distance (eAED) score and QI tag. GeneMark was self-trained as GeneMark-ES with an unsupervised procedure where the algorithm parametrization is solved automatically. More details about the annotation steps can be found in Appendix S14.1.

In annotating each species, we used common input evidence from full-length cDNA and SwissProt (<https://www.uniprot.org/>) plant proteins. RNA-seq assemblies were used for interior and Sitka spruce: short reads RNA-seq libraries were assembled with a pooled assembly approach in RNA-Bloom v0.9.8 (Nip et al., 2020). More details about the assembly parameters can be found in Appendices S14.2 and S14.3. The transcripts were screened for contaminants, and only transcripts with putative coding sequence were used for annotation selected through EvidentialGene v2017.12.21 (Gilbert, 2013).

MAKER was run iteratively in two steps with subsequent manual review to improve on the gene model prediction. The first run of MAKER included species-specific evidence together with RNA-seq, and the second run used combined evidence from the four genotypes. Repetitive elements were identified with the repeat library described before, and used as a customized library during the annotation process.

Genes were assigned to the high-confidence gene set according to stringent criteria: (i) having an annotated Pfam domain or BLAST hit to SwissProt with  $e$ -value  $< 1E-5$ , (ii) being assigned an eAED score of  $< 1$  by MAKER, (iii) having a gene length of  $> 1$  kbp, (iv) having an intron length of  $> 10$  bp, (v) having a complete coding sequence (start and stop codons), and (vi) having annotated start and stop codons  $> 500$  bp from scaffold ends.

### GO terms and Pfam analysis of annotated genes

The genes were functionally annotated with InterProScan v5.30-69 (Jones et al., 2014) with functional protein domains derived from Pfam v31 (Finn et al., 2014). We used InterProScan to infer protein superfamilies based on models and assignments available in the SUPERFAMILY database (Gough et al., 2001; Pandurangan et al., 2019; Wilson et al., 2009). These superfamilies provide protein domain assignment at the structural classification level of proteins. InterProScan was also run to annotate the corresponding GO terms with PANTHER GO (Thomas et al., 2003) and metabolic pathways.

Gene set enrichment analysis was performed with the AgriGO toolkit (Tian et al., 2017) and a complete set of GO terms using,

the hypergeometric statistical test and Hochberg multiple test adjustment. The annotated Pfam terms were further used for domain enrichment analysis with dcGO (Fang & Gough, 2013).

*In silico* annotation of the genes with outlier SNP A/S ratios (see section Estimation of SNP A/S ratios) was performed under the Blast2GO environment (Conesa et al., 2005) based on the protein sequence. Blastp searches were conducted against SwissProt ( $e$ -value  $< 1E-15$ ). GO mapping was conducted with the plant GO-Slim terms. Classification into functional categories was checked manually based on several sources: the Arabidopsis database (<https://www.arabidopsis.org/>), UniProtKB (<https://www.uniprot.org/>), QuickGO GO and annotation (<https://www.ebi.ac.uk/QuickGO/>), and literature searches. Protein secondary structure elements were predicted by running several methods implemented in the Network Protein Sequence Analysis (Combet et al., 2000) and by using the consensus predicted elements. Methods included in the prediction were DPM (Deleage & Roux, 1987), DSC (King & Sternberg, 1996), HNNC (Guermeur, 1997), PHD (Rost et al., 1994), and SOPM (Geourjon & Deleage, 1994). A hydrophobic cluster analysis (Callebaut et al., 1997) was also run to visualize the secondary structures and hydrophobic clusters in the protein sequences (<http://bioserv.rpbis.univ-paris-diderot.fr/services/HCA/>).

### Genome size estimates

Genome sizes were estimated using the k-mer frequency histograms computed by ntCard (Mohamadi et al., 2017). The software was run on the complete set of raw genomics reads. After excluding the effect of erroneous k-mers from the histogram, the homozygous k-mer (k-mers common in both parental alleles) was identified, which is usually the maximum peak in the histogram (Appendix S15). Although the estimation of abundances can be refined using distribution mixture models, we note that this first-order approximation works well for the range of experiments analyzed here. The genome size estimation was then performed by integrating the error-free k-mer frequency histogram curve. The final value of the genome size was estimated by averaging the values for the range of k-mer lengths of 30, 40, 50, 60, 70, 80, and 90. More details about the process can be found in (Biroi et al., 2018).

### Constructing chromosome-level super-scaffolds using an augmented white spruce genetic map

The genetic map used in this study contains positions for a total of 14 727 expressed genes as per the catalog of white spruce expressed genes (GCAT\_ID genes) (Rigault et al., 2011). The augmented map was built by adding three additional sets of gene SNPs to a slightly modified version of the previous most saturated white spruce genetic map (Pavy et al., 2017). The original genotyping of gene SNPs (Pavy, Gagnon, et al., 2013) was revisited using an in-house script in order to validate the SNP calling from 1959 progeny of the base mapping population and separate unambiguous SNPs from those that likely included erroneous genotypes for a few progeny (accessory markers). A base map was then built with 7868 unambiguous SNPs (one SNP per gene locus) using JoinMap v4.1 (Van Ooijen & Voorrips, 2006) and the multipoint maximum likelihood algorithm (Van Ooijen, 2011), to which 1304 accessory SNPs were added by fixing both the order and positions of the base map SNPs, for a total of 9172 mapped gene loci.

A first set of additional 1223 SNPs representing as many unmapped gene loci were added to the base map above from genotyping a subset of 156 progeny from the main set of 1959 progeny using the AdapTree Affymetrix Axiom 50 K interior

spruce SNP array (MacLachlan et al., 2018). The gene loci were then added to the base map similarly as for the accessory markers above. A second set of 1834 unmapped gene loci were added to the base map by merging the data from (Verta et al., 2013). For this, we considered the genotypes obtained from their RNA-seq on 60 megagametophytes from the female parent WS77111 of the base mapping population also used to generate the white spruce genome sequence assembly herein. The synteny observed for 2619 out of the 2629 gene loci in common between this dataset and the base map was used as anchors to position the unmapped genes. Finally, a third set of 2533 other unmapped gene loci were added by determining the synteny of the three genetic maps recently produced for Norway spruce (Bernhardsson et al., 2019). In this case, 4380 Norway spruce gene models in common with the mapped white spruce gene loci of the base map (Blastn identity > 98%) were used as anchors to include updated gene models matching the cDNAs of the catalog of white spruce expressed genes (Rigault et al., 2011) with identity  $\geq$  85%. This map was further filtered to only include genes with a single position assignment in the genetic map, resulting in a grand total of 15 750 mapped expressed genes on the 12 spruce linkage groups, which were used for the subsequent analyses.

The scaffolds from each spruce assembly were further joined using the augmented genetic map. Briefly, the 15 750 cDNAs positioned on the genetic map were aligned to the assemblies using GMAP, and the best scaffold hit for each cDNA was identified. Then, the scaffolds were stitched together using both the information from these alignments and the ordering of the cDNAs in the genetic map. This stitching resulted in 12 super-scaffolds per assembly, that is, one per linkage group.

### Analysis of single-nucleotide polymorphisms to estimate the genomic composition of interior spruce introgress genotype PG29

We first built separate k-mer Bloom filters from 40-fold coverage sequencing reads of each spruce species using ntHits v0.0.1 (<https://github.com/bcgsc/nthits>; --outbloom --solid, -b36, -k50, -t48). In three separate runs, we used a modified version of nEdit (Warren et al., 2019) on the interior spruce genome draft using Engelmann, Sitka, or white spruce primary Bloom filters to detect SNPs and, using the interior spruce secondary Bloom filter, categorically identified and reported homozygous variant bases (v1.0 nEditBF2.pl -f PG29v5.fa -r (engelmann40x\_k50.bf/sitka40x\_k50.bf/white40x\_k50.bf), -sinterior40x\_k50.bf, -k50, -d0, -i0). Variant bases were tallied for each interior spruce scaffold, and the data were reorganized into 12 linkage groups using the white spruce genetic map. We then calculated, within each 1-Mbp tile, the proportion of variant bases in Engelmann, Sitka, and white spruce that were shared with interior spruce, while shifting the frame over by 100 kbp. The circos.genomicDensity function of the circize R package v0.4.8 (Gu et al., 2014) was used to compute shared variant composition densities (window.size = 10E6) and create an ideogram for each interior spruce linkage group where overlapping 5-Mbp blocks were assigned to either Engelmann, Sitka, or white spruce.

### Phylogenetic analysis

The sequence divergence between the nc, mt, and pt genomes of Engelmann, Sitka, white, interior, and Norway spruce and loblolly and sugar pine (Appendix S16) was estimated by pairwise comparisons of the k-mer content as described previously (De La Torre et al., 2014). Briefly, each genome assembly was decomposed into k-mers ( $k = 26$ ) and loaded into individual Bloom filters using

ABYSS-Bloom (Jackman et al., 2017). For each species pair, the intersection of the respective Bloom filters approximates the sequence identity. The estimated sequence identities were represented as a distance matrix, and these distances were used to construct the nc, mt, and pt phylogenetic trees using MegaX and the neighbor-joining method (Kumar et al., 2018; Saitou & Nei, 1987), which does not assume a molecular clock. Because full pt genomes are available for these species, we also constructed a phylogenetic tree for the pt genomes using multiple sequence alignment for comparison. The pt genomes from each species were aligned using ClustalW2 (Larkin et al., 2007) using default parameters. Then, the phylogenetic tree was constructed using multiple sequence alignment and RAxML (Kozlov et al., 2019) (Appendix S8). The phylogenetic trees from these two approaches yielded identical topologies.

The protein divergence was additionally evaluated on single-copy genes defined by the OrthoFinder tool suit running with the STAG (Emms & Kelly, 2019) method for phylogeny inference. A set of 780 orthogroups containing single-copy genes and missing at most one gene from the annotation was used to estimate the species phylogenetic tree. The guide tree in STAG was supported by ASTRAL-III v5.6.3 (Zhang et al., 2018), which was used to infer the species tree topology.

### Analysis of gene family expansions and contractions

Annotations of *P. engelmannii*, *P. sitchensis*, *P. glauca*, interior spruce, *P. abies*, *P. taeda*, and *P. lambertiana* genomes were scanned to select the longest isoforms as representatives. The sequences were grouped in gene families with OrthoFinder v2.3.1 (Emms & Kelly, 2019), and homology was inferred through the last common ancestor and orthogroups reported by the tool. Models that estimate the gene family expansions and contractions (gene turnover rate) may overestimate both types of changes in draft genome assemblies; together with the original gene family dataset (OGF) we also created a filtered dataset (F50) to check the degree of overestimation in gene turnover rate. F50 is a modified version of the OGF in which proteins that were at least 50% of the length of the longest protein from the same species were filtered from each gene family, as described previously (Casola & Koralewski, 2018).

Significant orthogroup expansions were identified with CAFE v4.2.1 (Han et al., 2013) based on birth and death process models. A total of 9464 orthogroups present in at least five species were used in CAFE to estimate gene turnover using the maximum likelihood inference method. The distances in the rooted tree obtained from the single-copy genes were transformed to ultrametric units by r8s v1.81 (Sanderson, 2003). The average species divergence time used to calibrate the tree was 116 million years ago between spruce and pine as reported by (Wang et al., 2000). The estimated average gene turnover rate parameter ( $\lambda$ ) in CAFE was +0.0053 in the original dataset (OGF) and further used to define expanding/contracting gene families. The value was estimated based on the less numerous gene families ( $\leq 100$  genes in total), and the same  $\lambda$  was applied to the larger gene families ( $> 100$  genes in total) in order to avoid non-informative parameter estimates. After removing 21 713 putative misannotations from OGF, corresponding to approximately 11% of the total genes, the test dataset F50 yielded a slightly lower  $\lambda$  score of +0.0049. OGF and filtered F50 estimates were similar, supporting the notion that the bias of draft genome assemblies on the estimate of  $\lambda$  is marginal, as reported by (Casola & Koralewski, 2018). The number of significantly expanding/contracting gene families was reduced by approximately 30% in the F50 dataset comparing to the OGF dataset, thus indicating the contribution of a number of putative

misannotations in the analysis of gene families turnover on individual gene families. We report the gene family expansions and contractions for those genes that are expanded or contracted with  $P < 0.001$ .

### Detection of single-nucleotide polymorphisms in Sitka and white spruce genes

Plant material consisted of 212 white spruces from natural populations and germplasm collections (Pavy, Deschenes, et al., 2013) and 152 Sitka spruces from various coastal provenances.

For white spruce, sequences were derived from 48 cDNA libraries prepared from different tissues and after different treatments; each library was prepared from as many as 40 unrelated individuals. For Sitka spruce, six libraries were prepared for an exome capture, based on *P. glauca* probes as previously described (Azaiez et al., 2018).

In total, 64 million high-quality reads were obtained with a standard Sanger sequencing protocol and next-generation sequencing technologies for white spruce (Pavy et al., 2005; Ralph et al., 2008; Rigault et al., 2011) and 325 million paired-end reads were obtained for Sitka spruce by HiSeq sequencing (Genome Quebec Expertise and Services Center, Montreal, Canada). The reads were mapped with BWA-MEM v0.7.17 (Li, 2013) against the annotated transcripts (longest isoform) with default settings. SNPs were called with the GATK v4.0.11.0 HaplotypeCaller (DePristo et al., 2011; McKenna et al., 2010) and filtered using the following parameters:  $DP < 20$ ,  $QD < 2.0$ ,  $FS > 60.0$ ,  $MQ < 40.0$ ,  $MQRankSum < -12.5$ ,  $ReadPosRankSum < -8.0$ . For white spruce, given the lesser sequencing effort, only SNPs at positions where depth was below five were excluded. To reduce the number of false positives and because of variable sequencing effort leading to different SNP abundance between species, minimum allele frequency cut-off values of 0.20 for Sitka and 0.10 for white spruce were applied to further filter SNPs.

### Estimation of SNP A/S ratios for Sitka and white spruce genes

To detect rapidly evolving genes under positive selection, SNP A/S ratios were estimated (Fay et al., 2001; Liu et al., 2008). Non-synonymous and synonymous SNPs were identified and annotated using an in-house script. First, sequences involving no SNPs after SNP filtering were removed. Thus, 28 856 and 29 293 sequences remained for Sitka and white spruce, respectively. A perl script was developed to calculate the numbers of non-synonymous (La) and synonymous (Ls) sites in each coding sequence. La was defined as the number of non-degenerate sites plus two-thirds of the 2-fold degenerate sites. Similarly, Ls was defined as the number of 4-fold degenerate sites plus one third of the 2-fold degenerate sites. Then for each gene in each species, the rate of non-synonymous SNPs (A) was estimated as the number of SNPs observed at non-synonymous sites (Na) divided by La. Similarly, the rate of synonymous SNPs (S) was estimated as the number of SNPs at synonymous sites (Ns) divided by Ls. The SNP A/S ratio was then estimated for each gene in each species by calculating an adjusted SNP A/S ratio to include genes with no synonymous SNPs following the empirical logit principle (Agresti, 2002):

$$\text{Adjusted SNP A/S} = \frac{([Na + 0.5]/[La + 1])}{([Ns + 0.5]/[Ls + 1])} \quad (1)$$

The adjusted SNP A/S ratios were log-transformed to normalize their distribution (Appendix S10 and Figure S10.1). These

adjusted ratios are simply referred to as the SNP A/S ratios in the various sections of this report, where  $A/S < 1$  indicates variable intensity of purifying selection and  $A/S > 1$  indicates variable levels of positive selection (Fay et al., 2001).

The conventional Ka/Ks ratio estimated in a pairwise fashion at the interspecific level and the intraspecific SNP A/S ratio are strongly and positively correlated (Liu et al., 2008). One important and well-established advantage of the SNP A/S ratio over the Ka/Ks ratio is that it is more precise at the intraspecific level and thus more sensitive to detect genes under purifying selection or under positive selection within species, given that it is only based on intraspecific SNP variation (Fay et al., 2001; Liu et al., 2008).

Robust distances from the log-transformed adjusted SNP A/S values to the center of mass were computed using the minimum covariance determinant method (Hubert et al., 2012; Rousseeuw & Driessen, 1999). The fastmcd command implemented in R was used. For each of Sitka and white spruce, genes with adjusted SNP A/S ratios with a distance greater than 1.645 (the 95th percentile of a standard normal distribution) were declared outliers. In total, 913 genes were declared as significant outliers ( $P < 0.05$ ) with the highest SNP A/S values. For Sitka spruce, the adjusted SNP A/S threshold was 2.51, resulting in the detection of 748 outlier genes (4.3% of the genes harboring SNPs); for white spruce, the adjusted SNP A/S threshold was 2.85, resulting in the detection of 165 outlier genes (1.5% of the genes harboring SNPs).

### AUTHOR CONTRIBUTIONS

Conceptualization: IB, JeB, JoB, SJMJ; data curation: DC, PP, HM, YZ, RAM, AJM, BB; formal analysis: KKG, RLW, LC, JW, KMN, MMSY, JGAW, JMC, GAT, SAH, JL, ML, SG, NP; funding acquisition: NI, JoB, JeB, IB; methodology: KKG, RLW, LC, ML, NP, SJMJ, JoB, JeB, IB; project administration: CR; resources: JC, JJM, NI; visualization: RLW, KKG, LC; writing – original draft: KKG, RLW, NP; writing – review & editing: IB, JeB, JoB.

### CONFLICTS OF INTEREST

Authors declare that they have no competing interests.

### DATA AVAILABILITY STATEMENT

The genome assemblies for Engelmann spruce, Sitka spruce, white spruce, and interior spruce are respectively available under the following NCBI IDs: WSFP000000000, SNQJ01000000, JZKD02000000, and ALWZ05000000. The pt genomes are available under MK241981, KU215903.2, MK174379, and KT634228.1. The mt genomes for Sitka spruce and interior spruce are available under MK697696-MK697708 and LKAM01000001.1-LKAM01000036. The genome annotations are labeled with the following locus tags in NCBI: EFE08 (Engelmann), E0M31 (Sitka), DB47 (white), and ABT39 (interior). Genome assemblies and annotations can be also found at <https://www.bcgsc.ca/downloads/btl/Spruce>. All other relevant data can be found within the manuscript and its supporting materials.

### SUPPORTING INFORMATION

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

**Appendix S1.** Sequencing data types for each genome assembly: sequencing read data and corresponding fold coverage (millions of generated reads).

**Appendix S2.** Genome size estimates based on k-mers: genome size estimates in Gbp at different k values.

**Appendix S3.** Genome assembly statistics: statistics at each stage of the Engelmann (v1), Sitka (v1), white (v2), and interior spruce (v5) genome assemblies.

**Appendix S4.** Genome assembly completeness estimates.

**Appendix S5.** Linkage group assignments.

**Appendix S6.** Annotation statistics.

**Appendix S7.** Repeat content in Engelmann, Sitka, white, and interior spruce.

**Appendix S8.** Validation of the plastid genome phylogenetic tree with multiple sequence alignments and maximum likelihood inference.

**Appendix S9.** Gene orthogroups.

**Appendix S10.** SNP detection and SNP A/S ratios in Sitka and white spruce.

**Appendix S11.** DNA samples' geographical origin and local climate of the collected representative spruce genotypes.

**Appendix S12.** Data accessions.

**Appendix S13.** Gene completeness in the interior spruce transcriptome.

**Appendix S14.** Genome annotation.

**Appendix S15.** k-mer coverage histograms.

**Appendix S16.** Phylogenetic analysis.

**Table S1.** Gene gain and loss in gene families – CAFE summary results and OG annotation.

**Table S2.** Gene gain and loss in gene families, GO term enrichment analysis.

**Table S3.** Outlier genes defined by their high SNP A/S ratios ( $P < 0.05$ ) shared by white and Sitka spruce as represented on Figure 4.

**Table S4.** Outlier genes defined by their high SNP A/S ratios ( $P < 0.05$ ) found similar to white spruce genes related to climate adaptation according to Hornoy et al. (2015).

## REFERENCES

- Agresti, A. (2002) *Categorical Data Analysis*, 2nd edition, Hoboken, NJ: John Wiley & Sons.
- Ahuja, M.R. & Neale, D.B. (2005) Evolution of genome size in conifers. *Silvae Genetica*, **54**, 126–137.
- Azaiez, A., Pavy, N., Gérardi, S., Laroche, J., Boyle, B., Gagnon, F. et al. (2018) A catalog of annotated high-confidence SNPs from exome capture and sequencing reveals highly polymorphic genes in Norway spruce (*Picea abies*). *BMC Genomics*, **19**, 942.
- Bao, W., Kojima, K.K. & Kohany, O. (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 11.
- Bernhardsson, C., Vidalis, A., Wang, X., Scofield, D.G., Schiffthaler, B., Baisson, J. et al. (2019) An ultra-dense haploid genetic map for evaluating the highly fragmented genome assembly of Norway spruce (*Picea abies*). *G3: Genes, Genomes, Genetics*, **9**, 1623–1632.
- Biról, I., Mohamadi, H. & Chu, J. (2018) ntPack: a software package for big data in genomics. *IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)*, 41–50.
- Biról, I., Raymond, A., Jackman, S.D., Pleasance, S., Coope, R., Taylor, G.A. et al. (2013) Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics*, **29**, 1492–1497.
- Bouillé, M. & Bousquet, J. (2005) Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer *Picea* (Pinaceae): implications for the long-term maintenance of genetic diversity in trees. *American Journal of Botany*, **92**, 63–73.
- Bouillé, M., Senneville, S. & Bousquet, J. (2011) Discordant mtDNA and cpDNA phylogenies indicate geographic speciation and reticulation as driving factors for the diversification of the genus *Picea*. *Tree Genetics & Genomes*, **7**, 469–484.
- Bousquet, J., Gérardi, S., de Lafontaine, G., Jaramillo-Correa, J.P., Pavy, N., Prunier, J. et al. (2021) Spruce population genomics. In: Rajora, O.P. (Ed.) *Population genomics: forest trees*. Switzerland: Springer Nature, pp. 1–64.
- Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J. et al. (1997) Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cellular and Molecular Life Sciences*, **53**, 621–645.
- Casola, C. & Koralewski, T.E. (2018) Pinaceae show elevated rates of gene turnover that are robust to incomplete gene annotation. *The Plant Journal*, **95**, 862–876.
- Celedon, J.M., Whitehill, J.G.A., Madilao, L.L. & Bohlmann, J. (2020) Gymnosperm glandular trichomes: expanded dimensions of the conifer terpenoid defense system. *Scientific Reports*, **10**, 12464.
- Chaw, S.M., Sung, H.M., Long, H., Zharkikh, A. & Li, W.H. (1995) The phylogenetic positions of the conifer genera *Amentotaxus*, *Phyllocladus*, and *Nageia* inferred from 18S rRNA sequences. *Journal of Molecular Evolution*, **41**, 224–230.
- Combet, C., Blanchet, C., Geourjon, C. & Deleage, G. (2000) NPS@: network protein sequence analysis. *Trends in Biochemical Sciences*, **25**, 147–150.
- Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. & Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.
- Crepeau, M.W., Langley, C.H. & Stevens, K.A. (2017) From pine cones to read clouds: resccaffolding the megagenome of sugar pine (*Pinus lambertiana*). *G3: Genes, Genomes, Genetics*, **7**, 1563–1568.
- De La Torre, A.R., Biról, I., Bousquet, J., Ingvarsson, P.K., Jansson, S., Jones, S.J. et al. (2014) Insights into conifer giga-genomes. *Plant Physiology*, **166**, 1724–1732.
- De La Torre, A.R., Sekhwal, M.K., Puiu, D., Salzberg, S.L., Scott, A.D., Allen, B. et al. (2022) Genome-wide association identifies candidate genes for drought tolerance in coast redwood and giant sequoia. *The Plant Journal*, **109**, 7–22.
- Deleage, G. & Roux, B. (1987) An algorithm for protein secondary structure prediction based on class prediction. *Protein Engineering*, **1**, 289–294.
- Depardieu, C., Gérardi, S., Nadeau, S., Parent, G.J., Mackay, J., Lenz, P. et al. (2021) Connecting tree-ring phenotypes, genetic associations and transcriptomics to decipher the genomic architecture of drought adaptation in a widespread conifer. *Molecular Ecology*, **30**, 3898–3917.
- DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C. et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.
- Dohmen, E., Kremer, L.P., Bornberg-Bauer, E. & Kemena, C. (2016) DOGMA: domain-based transcriptome and proteome quality assessment. *Bioinformatics*, **32**, 2577–2581.
- Ellinghaus, D., Kurtz, S. & Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18.
- Emms, D.M. & Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, **20**, 238.
- Fang, H. & Gough, J. (2013) DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic Acids Research*, **41**, D536–D544.
- Farrar, J.L. (1995) *Trees in Canada*. Markham, Ontario, Canada: Fitzhenry and Whiteside Ltd.
- Fay, J.C., Wyckoff, G.J. & Wu, C.-I. (2001) Positive and negative selection on the human genome. *Genetics*, **158**, 1227–1234.
- Feng, S., Ru, D., Sun, Y., Mao, K., Milne, R. & Liu, J. (2019) Trans-lineage polymorphism and nonbifurcating diversification of the genus *Picea*. *New Phytologist*, **222**, 576–587.
- Fijarczyk, A. & Babik, W. (2015) Detecting balancing selection in genomes: limits and prospects. *Molecular Ecology*, **24**, 3529–3545.
- Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R. et al. (2014) Pfam: the protein families database. *Nucleic Acids Research*, **42**, D222–D230.

- Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
- Geourjon, C. & Deleage, G. (1994) SOPM: a self-optimized method for protein secondary structure prediction. *Protein Engineering*, **7**, 157–164.
- Gilbert, D. (2013) *Gene-omes built from mRNA-seq not genome DNA*. Bloomington, IN: Biology Dept., Univ. of Indiana.
- Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology*, **313**, 903–919.
- Gremme, G., Steinbiss, S. & Kurtz, S. (2013) GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **10**, 645–656.
- Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. (2014) Circlize implements and enhances circular visualization in R. *Bioinformatics*, **30**, 2811–2812.
- Guermeur, Y. (1997) Combinaison de classifieurs statistiques, application à la prédiction de la structure secondaire des protéines. Ph.D. Thesis, Univ. Paris 6.
- Guillet-Claude, C., Isabel, N., Pelgas, B. & Bousquet, J. (2004) The evolutionary implications of *Knox-I* gene duplication in conifers: correlated evidence from phylogeny, gene mapping and analysis of functional divergence. *Molecular Biology and Evolution*, **21**, 2232–2245.
- Guo, Y., Liu, J., Zhang, J., Liu, S. & Du, J. (2017) Selective modes determine evolutionary rates, gene compactness and expression patterns in *Brassica*. *The Plant Journal*, **91**, 34–44.
- Hamilton, J.A., De La Torre, A.M. & Aitken, S.N. (2015) Fine-scale environmental variation contributes to introgression in a three-species spruce hybrid complex. *Tree Genetics and Genomes*, **11**, 817.
- Han, M.V., Thomas, G.W., Lugo-Martinez, J. & Hahn, M.W. (2013) Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution*, **30**, 1987–1997.
- Haselhorst, M.S.H. & Buerkle, C.A. (2013) Population genetic structure of *Picea engelmannii*, *P. glauca* and their previously unrecognized hybrids in the central Rocky Mountains. *Tree Genetics & Genomes*, **9**, 669–681.
- Holt, C. & Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, **12**, 491.
- Hornoy, B., Pavy, N., Gerardi, S., Beaulieu, J. & Bousquet, J. (2015) Genetic adaptation to climate in white spruce involves small to moderate allele frequency shifts in functionally diverse genes. *Genome Biology and Evolution*, **7**, 3269–3285.
- Hubert, M., Rousseeuw, P.J. & Verdonck, T. (2012) A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, **21**, 618–637.
- Jackman, S.D., Coombe, L., Chu, J., Warren, R.L., Vandervalk, B.P., Yeo, S. et al. (2018) Tigrint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics*, **19**, 393.
- Jackman, S.D., Vandervalk, B.P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S.A. et al. (2017) ABYSS 2.0: resource-efficient assembly of large genomes using a bloom filter. *Genome Research*, **27**, 768–777.
- Jaramillo-Correa, J.P., Beaulieu, J., Khalsa, D.P. & Bousquet, J. (2009) Inferring the past from the present phylogeographic structure of north American forest trees: seeing the forest for the genes. *Canadian Journal of Forest Research*, **39**, 286–307.
- Jensen, J.D. & Bachrog, D. (2010) Characterizing recurrent positive selection at fast-evolving genes in *Drosophila miranda* and *Drosophila pseudoobscura*. *Genome Biology and Evolution*, **2**, 371–378.
- Jones, M.R., Schrader, K.A., Shen, Y., Pleasance, E., Ch'ng, C., Dar, N. et al. (2016) Response to angiotensin blockade with irbesartan in a patient with metastatic colorectal cancer. *Annals of Oncology*, **27**, 801–806.
- Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C. et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
- Jurado, S., Abraham, Z., Manzano, C., Lopez-Torrejon, G., Pacios, L.F. & Del Pozo, J.C. (2010) The *Arabidopsis* cell cycle F-box protein SKP2A binds to auxin. *Plant Cell*, **22**, 3891–3904.
- Kim, D., Langmead, B. & Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, **12**, 357–360.
- King, R.D. & Sternberg, M.J. (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Science*, **5**, 2298–2310.
- Kozlov, A.M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. (2019) RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, **35**, 4453–4455.
- Kucuk, E., Chu, J., Vandervalk, B.P., Austin Hammond, S. & Warren, R.L. (2017) Collector: transcript-informed, targeted *de novo* assembly of gene loci. *Bioinformatics*, **33**, 1782–1788.
- Kumar, S., Stecher, G., Li, M., Nknyaz, C. & Tamura, K. (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, **35**, 1547–1549.
- Lamara, M., Parent, G.J., Giguère, I., Beaulieu, J., Bousquet, J. & Mackay, J.J. (2018) Association genetics of acetophenone defense against spruce budworm in mature white spruce. *BMC Plant Biology*, **18**, 231.
- Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H. et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
- Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*, 1303.3997.
- Liu, J., Zhang, Y., Lei, X. & Zhang, Z. (2008) Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. *Genome Biology*, **9**, R69.
- Lockwood, J.D., Aleksic, J.M., Zou, J., Wang, J., Liu, J. & Renner, S.S. (2013) A new phylogeny for the genus *Picea* from plastid, mitochondrial, and nuclear sequences. *Molecular Phylogenetics and Evolution*, **69**, 717–727.
- MacLachlan, I.R., Yeaman, S. & Aitken, S.N. (2018) Growth gains from selective breeding in a spruce hybrid zone do not compromise local adaptation to climate. *Evolutionary Applications*, **11**, 166–181.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytzky, A. et al. (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.
- Mohamadi, H., Khan, H. & Biroi, I. (2017) ntCard: a streaming algorithm for cardinality estimation in genomics data. *Bioinformatics*, **33**, 1324–1330.
- Mosca, E., Cruz, F., Gomez-Garrido, J., Bianco, L., Rellstab, C., Brodbeck, S. et al. (2019) A reference genome sequence for the European silver fir (*Abies alba* Mill.): a community-generated genomic resource. *G3: Genes, Genomes, Genetics*, **9**, 2039–2049.
- Mullin, T.J., Andersson, B., Bastien, J.-C., Beaulieu, J., Burdon, R.W., Dvorak, W.S. et al. (2011) Economic importance, breeding objectives and achievements. In: Plomion, C., Bousquet, J. & Kole, C. (Eds.) *Genetics, genomics and breeding of conifers*. New York: CRC Press & Science Publishers, pp. 40–127.
- Natural Resources Canada. (2021) *The State of Canada's Forests: 2020 Annual Report and National Inventory Report*. Ottawa: Canada.
- Neale, D.B., McGuire, P.E., Wheeler, N.C., Stevens, K.A., Crepeau, M.W., Cardeno, C. et al. (2017) The Douglas-fir genome sequence reveals specialization on the photosynthetic apparatus in Pinaceae. *G3: Genes, Genomes, Genetics*, **9**, 3157–3167.
- Neale, D.B., Zimin, A.V., Zaman, S., Scott, A.D., Shrestha, B., Workman, R.E. et al. (2022) Assembled and annotated 26.5 Gbp coast redwood genome: a resource for estimating evolutionary adaptive potential and investigating hexaploid origin. *G3: Genes, Genomes, Genetics*, **12**, jk-ab380.
- Nip, K.M., Chiu, R., Yang, C., Chu, J., Mohamadi, H., Warren, R.L. et al. (2020) RNA-bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes. *Genome Research*, **30**, 1191–1200.
- Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.C., Scofield, D.G. et al. (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**, 579–584.
- Pandurangan, A.P., Stahlhacke, J., Oates, M.E., Smithers, B. & Gough, J. (2019) The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. *Nucleic Acids Research*, **47**, D490–D494.
- Paulino, D., Warren, R.L., Vandervalk, B.P., Raymond, A., Jackman, S.D. & Biroi, I. (2015) Sealer: a scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics*, **16**, 230.
- Pavy, N., Deschenes, A., Blais, S., Lavigne, P., Beaulieu, J., Isabel, N. et al. (2013) The landscape of nucleotide polymorphism among 13,500 genes



- of the conifer *Picea glauca*, relationships with functions, and comparison with *Medicago truncatula*. *Genome Biology and Evolution*, **5**, 1910–1925.
- Pavy, N., Gagnon, F., Rigault, P., Blais, S., Deschenes, A., Boyle, B. et al.** (2013) Development of high-density SNP genotyping arrays for white spruce (*Picea glauca*) and transferability to subtropical and nordic congeners. *Molecular Ecology Resources*, **13**, 324–336.
- Pavy, N., Lamothe, M., Pelgas, B., Gagnon, F., Birol, I., Bohlmann, J. et al.** (2017) A high-resolution reference genetic map positioning 8.8 K genes for the conifer white spruce: structural genomics implications and correspondence with physical distance. *The Plant Journal*, **90**, 189–203.
- Pavy, N., Paule, C., Parsons, L., Crow, J.A., Morency, M.J., Cooke, J. et al.** (2005) Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters. *BMC Genomics*, **6**, 144.
- Pavy, N., Pelgas, B., Beauseigle, S., Blais, S., Gagnon, F., Gosselin, I. et al.** (2008) Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. *BMC Genomics*, **9**, 21.
- Pavy, N., Pelgas, B., Laroche, J., Rigault, P., Isabel, N. & Bousquet, J.** (2012) A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. *BMC Biology*, **10**, 84.
- Ralph, S.G., Chun, H.J., Kolosova, N., Cooper, D., Oddy, C., Ritland, C.E. et al.** (2008) A conifer genomics resource of 200,000 spruce (*Picea* spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (*Picea sitchensis*). *BMC Genomics*, **9**, 484.
- Rigault, P., Boyle, B., Lepage, P., Cooke, J.E., Bousquet, J. & MacKay, J.J.** (2011) A white spruce gene catalog for conifer genome analyses. *Plant Physiology*, **157**, 14–28.
- Rost, B., Sander, C. & Schneider, R.** (1994) PHD - an automatic mail server for protein secondary structure prediction. *Computer Applications in the Biosciences*, **10**, 53–60.
- Rousseeuw, P.J. & Driessen, K.V.** (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.
- Saitou, N. & Nei, M.** (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.
- Sanderson, M.J.** (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, **19**, 301–302.
- Savard, L., Li, P., Strauss, S.H., Chase, M.W., Michaud, M. & Bousquet, J.** (1994) Chloroplast and nuclear gene sequences indicate late Pennsylvanian time for the last common ancestor of extant seed plants. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 5163–5167.
- Scott, A.D., Zimin, A.V., Puiu, D., Workman, R., Britton, M., Zaman, S. et al.** (2020) A reference genome sequence for giant sequoia. *G3: Genes, Genomes, Genetics*, **10**, 3907–3919.
- Shiryev, S.A., Papadopoulos, J.S., Schaffer, A.A. & Agarwala, R.** (2007) Improved BLAST searches using longer words for protein seeding. *Bioinformatics*, **23**, 2949–2951.
- Siewert, K.M. & Voight, B.F.** (2017) Detecting long-term balancing selection using allele frequency correlation. *Molecular Biology and Evolution*, **34**, 2996–3005.
- Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M.** (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
- Singh, N.K., Shukla, P. & Kirti, P.B.** (2020) A CBL-interacting protein kinase AdCIPK5 confers salt and osmotic stress tolerance in transgenic tobacco. *Scientific Reports*, **10**, 418.
- Stevens, K.A., Wegrzyn, J.L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C. et al.** (2016) Sequence of the sugar pine megagenome. *Genetics*, **204**, 1613–1626.
- Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R. et al.** (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Research*, **13**, 2129–2141.
- Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z. et al.** (2017) agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*, **45**, W122–W129.
- Tsang, E.S., Shen, Y., Chooback, N., Ho, C., Jones, M., Renouf, D.J. et al.** (2019) Clinical outcomes after whole-genome sequencing in patients with metastatic non-small-cell lung cancer. *Cold Spring Harbor Molecular Case Studies*, **5**, a002659.
- Van Ooijen, J.W.** (2011) Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genetics Research*, **93**, 343–349.
- Van Ooijen, J.W. & Voorrips, E.E.** (2006) *JoinMap®4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations*. Wageningen: Kyazma BV.
- Verta, J.P., Landry, C.R. & Mackay, J.J.** (2013) Are long-lived trees poised for evolutionary change? Single locus effects in the evolution of gene expression networks in spruce. *Molecular Ecology*, **22**, 2369–2379.
- Wang, X.Q., Tank, D.C. & Sang, T.** (2000) Phylogeny and divergence times in Pinaceae: evidence from three genomes. *Molecular Biology and Evolution*, **17**, 773–781.
- Warren, R.L.** (2016) RAILS and cobbler: scaffolding and automated finishing of draft genomes using long DNA sequences. *Journal of Open Source Software*, **1**, 116.
- Warren, R.L., Coombe, L., Mohamadi, H., Zhang, J., Jaquish, B., Isabel, N. et al.** (2019) ntEdit: scalable genome sequence polishing. *Bioinformatics*, **35**, 4430–4432.
- Warren, R.L., Keeling, C.I., Yuen, M.M., Raymond, A., Taylor, G.A., Vandervalk, B.P. et al.** (2015a) Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *The Plant Journal*, **83**, 189–212.
- Warren, R.L., Yang, C., Vandervalk, B.P., Behsaz, B., Lagman, A., Jones, S.J. et al.** (2015b) LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience*, **4**, 35.
- Whitehill, J.G.A., Macaire, M.S.Y., Henderson, H., Madilao, L., Kshatriya, K., Bryan, J. et al.** (2019) Functions of stone cells and oleoresin terpenes in the conifer defense syndrome. *The New Phytologist*, **221**, 1503–1517.
- Wiehe, T.H.E. & Stephan, W.** (1993) Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Molecular Biology and Evolution*, **10**, 842–854.
- Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M. et al.** (2009) SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Research*, **37**, D380–D386.
- Wu, T.D., Reeder, J., Lawrence, M., Becker, G. & Brauer, M.J.** (2016) GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods in Molecular Biology*, **1418**, 283–334.
- Xu, Z. & Wang, H.** (2007) LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, **35**, W265–W268.
- Yeaman, S., Hodgins, K.A., Lotterhos, K.E., Suren, H., Nadeau, S., Degner, J.C. et al.** (2016) Convergent local adaptation to climate in distantly related conifers. *Science*, **353**, 1431–1433.
- Yeo, S., Coombe, L., Warren, R.L., Chu, J. & Birol, I.** (2018) ARCS: scaffolding genome drafts with linked reads. *Bioinformatics*, **34**, 725–731.
- Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S.** (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, **19**, 153.
- Zimin, A.V., Stevens, K.A., Crepeau, M.W., Puiu, D., Wegrzyn, J.L., Yorke, J.A. et al.** (2017) An improved assembly of the loblolly pine megagenome using long-read single-molecule sequencing. *Gigascience*, **6**, 1–4.