BMC Genomics

RESEARCH

Open Access

Long-insert sequence capture detects high copy numbers in a defence-related beta-glucosidase gene *βglu-1* with large variations in white spruce but not Norway spruce

Tin Hang Hung[1*], Ernest T. Y. Wu[1], Pauls Zeltiņš[2], Āris Jansons[2], Aziz Ullah[3], Nadir Erbilgin[3], Joerg Bohlmann[4,5,6], Jean Bousquet[7], Inanc Birol[8], Sonya M. Clegg[1] and John J. MacKay[1*]

Abstract

Conifers are long-lived and slow-evolving, thus requiring effective defences against their fast-evolving insect natural enemies. The copy number variation (CNV) of two key acetophenone biosynthesis genes *Ugt5*/*Ugt5b* and *βglu-1* may provide a plausible mechanism underlying the constitutively variable defence in white spruce (*Picea glauca*) against its primary defoliator, spruce budworm. This study develops a long-insert sequence capture probe set (Picea_hung_p1.0) for quantifying copy number of *βglu-1*-like, *Ugt5*-like genes and single-copy genes on 38 Norway spruce (*Picea abies*) and 40 *P. glauca* individuals from eight and nine provenances across Europe and North America respectively. We developed local assemblies (Piabi_c1.0 and Pigla_c.1.0), full-length transcriptomes (PIAB_v1 and PIGL_v1), and gene models to characterise the diversity of *βglu-1* and *Ugt5* genes. We observed very large copy numbers of *βglu-1*, with up to 381 copies in a single *P. glauca* individual. We observed among-provenance CNV of *βglu-1* in *P. glauca* but not *P. abies*. *Ugt5b* was predominantly single-copy in both species. This study generates critical hypotheses for testing the emergence and mechanism of extreme CNV, the dosage effect on phenotype, and the varying copy number of genes with the same pathway. We demonstrate new approaches to overcome experimental challenges in genomic research in conifer defences.

Keywords Acetophenone pathway, Targeted capture, Picea, CNV, Secondary metabolism, Conifer genomics

*Correspondence:
Tin Hang Hung
tin-hang.hung@biology.ox.ac.uk
John J. MacKay
john.mackay@biology.ox.ac.uk
[1] Department of Biology, University of Oxford, Oxford OX1 3RB, UK
[2] Latvian State Forest Research Institute "Silava", Salaspils 2169, Latvia
[3] Department of Renewable Resources, University of Alberta, Edmonton, AB T6G 2E3, Canada
[4] Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[5] Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[6] Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[7] Canada Research Chair in Forest Genomics, Forest Research Centre, Université Laval, Québec, QC G1V 0A6, Canada
[8] Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC V5Z 4S6, Canada

Introduction

Conifers are keystone species in many terrestrial biomes [1]. They play an important role in global biogeochemical cycles, including the carbon, nutrient, and water cycles. They are also of large economic importance globally as a resource, derived from naturally regenerated forests and plantations, for timber and non-timber products [2]. However, forests and tree populations are facing threats from environmental change, most prominently from warmer climates [3] as well as pest and pathogen outbreaks [4, 5]. As part of their sessile lifestyle, longevity, and slow evolutionary response, conifers have developed effective defence systems against diverse faster-evolving natural enemies, which are key attributes to the forest health [6–8].

One such defence system relies on acetophenones and their glucosides in white spruce (*Picea glauca*) across northern North America. Acetophenones and their glucosides are the most abundant soluble phenolic and glucosidic compounds in their shoots [9]. Variation in the constitutive accumulation of two particular acetophenones, piceol and pungenol, has been demonstrated to confer defensive properties and resistance against spruce budworm (*Choristoneura fumiferana*, Lepidoptera: Tortricidae), which is largely sympatric with white spruce [10]. Higher levels of these acetophenones reduce the budworm defoliation, which has been linked to decreased larval survival and pupal mass and delayed development in female moths, and increased fitness in the tree host [11]. Resistant white spruces have a high level of foliar piceol and pungenol, while both resistant and non-resistant trees can accumulate high levels of their glycosylated forms, picein and pungenin [12]. Phenology studies of white spruce and spruce budworm showed the peak of piceol and pungenol accumulation in hosts temporarily matches with the larval stage that is most damaging [10]. These acetophenones are conserved to some extent in Pinaceae and most *Picea* species accumulate at least one acetophenone glucoside in their foliage. Besides *P. glauca*, Norway spruce (*P. abies*), which is distributed across Europe, is the only spruce species known to accumulate both picein and pungenin [13].

The biosynthesis and function of these acetophenones has been characterised by metabolite analyses, gene discovery and gene expression analysis, biochemical characterisation of biosynthetic enzymes, and insect in vivo assays. The formation of pungenin and isopungenin from pungenol is catalysed by UDP-sugar dependent glucosyltransferases, encoded by the *PgUgt5b* and *PgUgt5* genes respectively [14], while genes and enzymes for formation of picein from piceol are still unknown. The glucosidic bond can be hydrolysed by a β-glucosidase encoded by the *Pgβglu-1* gene to release the active piceol and pungenol [15]. Substantial geographical and seasonal variations exist in the expression of *Pgβglu-1* linked to the phenotypic variation of acetophenone accumulation and tree survival [16], with up to 1,000-fold difference in the gene expression between resistant and non-resistant trees.

However, the genetic underpinning of this phenotypic variation remains largely unknown and copy number variation (CNV) could be a plausible mechanism. Copy number variation was first studied in humans [17, 18] and was found to occur in 4.8–9% of the human genome, which is more frequent than single nucleotide polymorphisms [19, 20]. There has been a lack of consensus on the definition of CNV, for example, the size of CNVs has been reported to range from 50 bp to several Mbps, but shorter ones are sometimes classified as indels [21]. CNV may be best defined as 'the relative difference in copy numbers of particular DNA sequences among conspecific individuals' [22]. CNV can lead to gene expression differences reaching several orders of magnitude [23] brought about by various mechanisms, including simple dosage effects, changes in gene regulatory regions, and alterations to the physical proximity of genes in relation to their regulatory elements [24].

Although CNV may explain phenotypic variation in important adaptive traits, there are very few studies that have investigated intraspecific or population level CNV in conifers. These studies have mostly focused on individual loci, such as the thaumatin-like protein gene in *Pinus sylvestris* (*PsTLP*) [25, 26] and the (+)-3-carene synthase gene in *Picea sitchensis* (*PsTPS-3car*) [27, 28] or mapped genomic CNV at the species level [29–31]. In *P. glauca* two genomic scans detected more than 1,300 and 3,900 CNVs, representing around 10% and 32% of the genes tested, respectively and, significant over-representation of defence response genes [32, 33].

The very large size of conifer genomes, between 18–35 Gbp for most species [34], has raised evolutionary questions regarding their ability to adapt to changing biotic pressures and has also posed experimental limitations and challenges. While gene duplication has a central role in genome evolution, conifers generally retain more non-functional gene copies than other species and accumulate as high as 4% of gene-like sequences [35]. They also have a larger amount of intronic sequences containing more repetitive elements compared to flowering plants [36]. The abundance of pseudogenes also has challenged a contiguous assembly for their genomes and curation of a well-characterised gene catalogue [1].

The sheer size and complexity of conifer genomes make whole genome sequencing impractical for studying conifer populations, therefore, CNV studies in conifer trees have utilised either array Comparative

Genome Hybridization (aCGH) or quantitative PCR (qPCR). However, these methods have a significant limitation due to the lack of sequence data on the signals detected, which may result from non-targets, and unquantified variation among individuals in on-target signals. Alternatively, shotgun sequencing with short reads could confound non-functional gene-like sequences or pseudogenes from their highly similar parent genes [37], and has substantial amplification bias such as GC content [38]. Therefore, a new method to bridge this gap in precision requires incorporating both long-read sequencing, which enhances the confidence of sequence validation, and targeted capture, which reduces the cost significantly for gigagenomes like spruces', and makes experiments scalable for large population studies.

Here, we aim to develop a long-insert sequence capture protocol for quantifying copy number based on known gene targets in acetophenone biosynthesis in *P. glauca*. Our objectives are three-fold: (1) to design a probe set suitable across *Picea* spp. that captures the target genes *βglu-1* and *Ugt5*, which regulate the accumulation and release of acetophenone defence compounds, and single-copy genes as internal standards; (2) to characterise the identity and diversity of target gene models by integrating genomic sequence capture, transcriptome assemblies, and local gene assemblies; (3) to quantify the copy numbers of *βglu-1* and *Ugt5* and compare them among populations of *P. abies* and *P. glauca* using a coverage-based algorithm. The development of this protocol with the probe set particularly addresses the gaps in methods and resources in studying the genetic underpinning of insect resistance in conifer species.

Methods

Plant materials and nucleic acids extraction

Foliage samples for *Picea abies* and *Picea glauca* were collected in September 2021 from two common gardens located in Saldus, Courland, Latvia (56.85°, 22.52°), established in 1975 [39], and Calling Lake, Alberta, Canada (55.28°, −113.17°), established in 1982 [39, 40] respectively. We collected foliage samples from the top-half crown of 38 and 40 individual trees, which represented 8 and 9 provenances respectively (Fig. 1 and Supplementary Table 1), compliant with relevant institutional guidelines and national legislations. The plant materials were identified by the sampling teams: P. Z. and A.J. for *P. abies* and A.U. and N.E. for *P. glauca*. Foliage samples were snap-frozen immediately after removal from the tree and stored at −80 °C to preserve the integrity of nucleic acids until analyses.

We extracted genomic DNA and total RNA using the DNeasy Plant Mini Kit (Qiagen, United Kingdom) and the NEB Monarch Total RNA Miniprep Kit (#T2010, New England Biolabs, United Kingdom) respectively, determined their quantity on a Qubit 4 Fluorometer (Thermo Fisher Scientific, United Kingdom), and assessed their purity using a NanoDrop One Spectrophotometer (Thermo Fisher Scientific), with A260/280 and A260/230 above 1.80. Integrity was verified on a 0.5% agarose gel for genomic DNA and a 2100 Bioanalyzer (Agilent Technologies, United States) for total RNA.

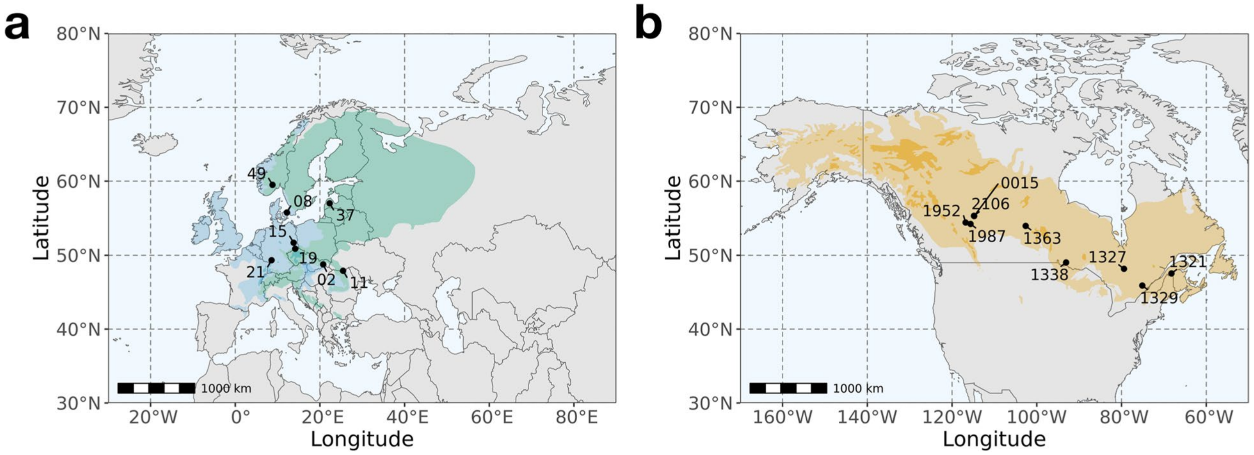

Fig. 1 The localities of provenances (circles) of (**a**) *P. abies* and (**b**) *P. glauca*. For *P. abies*, the green polygon shows its native distribution, and the blue polygon shows its introduced and naturalised distribution. For *P. glauca*, the orange polygon shows its native distribution. Species distribution maps are obtained from [41] and [42] respectively

Selection of target regions and bait development

Full-length cDNA sequences of *Pgβglu-1* (NCBI GenBank Accession: KJ780719.1) [12], *PgUgt5* (KY963363.1), and *PgUgt5b* (KY963364.1) [14] were searched against the gene models of *P. glauca* WS77111_v2 [2] using BLAST 2.11.0 + [43]. Their corresponding genomic coordinates were padded for 1,000 bp in both upstream and downstream directions, resulting in 24 target genomic sequences with a total length of 108,485 nt. We designed 80-nt baits at 2X tiling density to cover these sequences and produced a total of 2,427 baits. We filtered the baits according to number of BLAST hits against reference genome per bait and the predicted melting temperature between the baits and the BLAST hits, and retained 1,836 baits. We performed BUSCO analysis [44] on the gene models to predict single-copy genes in the assemblies indicated above, 134 of which were randomly selected to serve as internal standards. Using the same design approach, we produced 21,951 baits. After filtering, 18,164 baits were retained to cover 913,726 nt. The final bait set covering acetophenone biosynthesis target genes and single-copy internal standard genes contained 20,000 baits and covered 158 target regions of 1,022,211 bp. The probe set was named Picea_hung_p1.0 (Supplementary Data 1, meta data in Supplementary Table 2).

The alignment of the probe set was tested on the reference genomes of five *Picea* species, *P. abies* [45], *P. engelmannii* [2], *P. engelmannii* × *P. glauca* × *P. sitchensis* [2], *P. glauca* [2], and *P. sitchensis* [2], using minimap 2.22 [46].

A graphical abstract of the experimental and bioinformatic pipeline is presented in Fig. 2.

Sequence capture and sequencing of genomic DNA

For each sample, 500 ng of genomic DNA was randomly fragmented to achieve a median size of ~ 15 Kbp, using 1/8 reaction of NEBNext dsDNA Fragmentase (#M0348, New England Biolabs) incubated at 37 °C for 5 min. Fragmented DNA was purified, end-repaired, dA-tailed, and ligated with adaptors for Illumina (#E6609, New England Biolabs) using NEBNext Ultra II Library Prep Kit for Illumina (#E7645, New England Biolabs). The sub-libraries were amplified and barcoded using a ProFlex PCR System (Thermo Fisher Scientific). The 25-µl reactions contained: 12.5 µl Kapa HiFi HotStart ReadyMix (Kapa Biosystems, United Kingdom), 2.5 µl NEBNext Multiplex Oligos for Illumina (10 µM) (#E6609, New England Biolabs), 10 µl adaptor-ligated DNA. The thermal cycling profile was: 94°C 3 min, 10 × [94 °C 30 s, 65°C 4 min 30 s], 65°C 5 min, with ramp rate of each step at 3°C/s, to amplify fragments roughly from 1 to 10 Kbp. For all purification steps, a

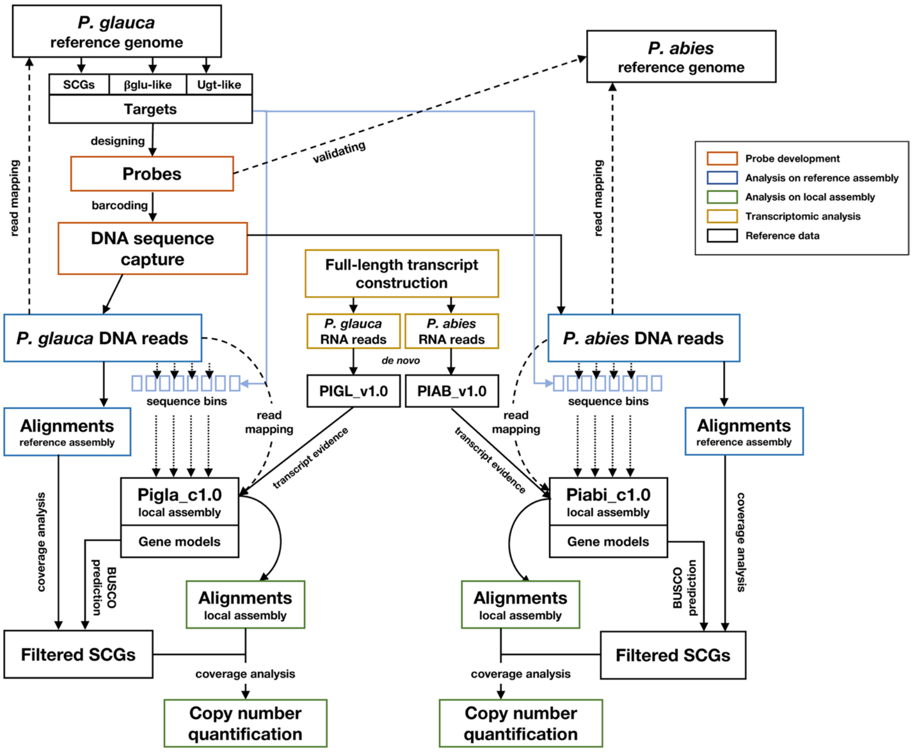

Fig. 2 Experimental and bioinformatic pipeline of copy number quantification

0.4 × AMPure XP (Beckman Coulter, United States) clean-up was used for size selection above ~ 2 Kbp.

Hybridisation capture was performed with myBaits Custom 1–20 K Kit (Daciel Arbor Biosciences, United States), using the Long Insert Protocol (manual version 5.02). The target-enriched sub-library was amplified using the same profiles as the barcoding amplification above, except using 25 cycles. All sub-libraries were pooled and normalised.

Nanopore libraries were constructed using the Ligation Sequencing Kit Chemistry 14 (SQK-LSK114, Oxford Nanopore Technologies, United Kingdom) using ~ 200 fmol pooled library. Nanopore libraries were then sequenced on a R10.4.1 flow cell (FLO-PRO114M) at 400 bps ('default' mode) on a PromethION system (Oxford Nanopore Technologies) at the DNA Technologies & Expression Analysis Core Laboratory, UC Davis Genome Center.

All nanopore reads in this study were basecalled and demultiplexed from raw electrical signals using Guppy v6.0.0, trimmed for Nanopore and/or Illumina adaptors and split for chimeras using Porechop 0.2.4.

Full-length cDNA (fl-cDNA) sequencing and transcriptome assembly

All total RNA samples were pooled to 1 µg for each species. First-strand synthesis, tailing, template switching, and extension were all completed with the SMARTer PCR cDNA Synthesis Kit (Takara Bio Europe, France). Full-length cDNA (fl-cDNA) library was constructed using the primers provided and the thermal cycling profile was: 94°C 3 min, 18 × [94°C 30 s, 65°C 4 min 30 s], 65°C 5 min. For all purification steps, a 1.2 × AMPure XP (Beckman Coulter, United States) clean-up was used. Nanopore libraries were constructed using the Ligation Sequencing Kit Chemistry 14 (SQK-LSK114, Oxford Nanopore Technologies) using ~ 200 fmol pooled library. Nanopore libraries were then sequenced on a R10.3 flow cells (FLO-MIN111) on a GridION system (Oxford Nanopore Technologies, United Kingdom).

Fl-cDNA sequences were filtered for quality (Q score > 10) and trimmed for Nanopore adaptors and SMARTer PCR primers using Porechop 0.2.4 [47] with manual configuration. Filtered reads were used to construct the transcriptome using RNA-Bloom2 [48]. Completeness of the transcriptomes were assessed using BUSCO v5.1.2 [44] with the embryophta_odb10 database. The transcriptome assemblies were named PIAB_v1 and PIGL_v1 respectively. Abundance of each transcript was quantified using minimap 2.22 [46] for mapping the raw fl-cDNA sequences to PIAB_v1 and PIGL_v1.

Local sequence assembly and gene characterisation

To circumvent the extreme sequencing and computational requirements of whole-genome global assembly in spruce giga-genomes, local assembly could be used to assemble reads from the sequence capture and improve the characterisation of gene models [49], which may be incomplete in the reference genome.

Non-uniform sequence coverage due to the nature of sequence capture, rich repeat content, and biological variation had to be taken account into the local sequence assembly. For both species, all filtered reads were first searched against the gene models using BLAST + and binned accordingly. Sequences in each bin were assembled using Canu 2.2 [50] with the parameters 'maxInputCoverage = 10000 corOutCoverage = 10000 corMhap Sensitivity = high corMinCoverage = 0 redMemory = 32 oeaMemory = 32 batMemory = 200' according to the developers' recommendations to retain as many reads as possible. The local sequence assemblies were named Piabi_c1.0 and Pigla_c1.0 respectively.

Gene models were characterised using both ab initio prediction and transcript evidence of PIAB_v1 and PIGL_v1 from the transcriptome assemblies above using MAKER 3.01.03 [51]. The gene models were named Piabi_c1g and Pigla_c1g respectively.

Gene analysis of *βglu-1* and *Ugt5*

Full-length cDNA sequences of *Pgβglu-1* (NCBI GenBank Accession: KJ780719.1) [12], *PgUgt5* (KY963363.1), and *PgUgt5b* (KY963364.1) [14] were searched against the Piabi_c1g and Pigla_c1g using BLAST 2.11.0 + [43]. Homologous sequences of the cDNA and corresponding protein were subjected to local alignment with generalised affine gap costs (Altschul method) for *βglu-1* and *Ugt5* by using E-INS-i in MAFFT v7.490 [52] with 1,000 iterations. ModelTest-NG 0.17 [53] was used to select the best evolutionary model. A maximum likelihood phylogenetic tree of the cDNA sequences was built using RAxML-NG v. 1.0.2 [54] with the GTR + G4 substitution model with bootstrapping. The phylogenetic tree was midpoint-rooted and visualised using ggtree [55], along with the protein alignment using ggmsa.

Single-copy gene validation

The 139 putative single-copy genes were validated before use as internal standards. First, filtered genomic DNA reads were mapped to the reference genomes of *P. abies* and *P. glauca* (WS77111_v2) using minimap 2.22 with the specific option '-I 100 g' in both the indexing and mapping step given the large genome size. The alignment BAM files were converted to BED files using bam2bed 2.4.39 in the BEDOPS package [56]. Coverage statistics were calculated using the package TEQC [57] in R. Normalised coverage for each target gene was calculated as the average coverage over the target bases divided by the average coverage over bases of all target genes. Second, BUSCO analysis was performed on the local gene models Piabi_c1.0 g and Pigla_c1.0 g to differentiate single-copy, duplicated, fragmented, and missing models.

Copy number quantification

Most of the available CNV algorithms could only report relative copy number in forms of ratios [58]. We used single-copy genes as internal standards and were able to quantify absolute copy number by comparing coverage statistics of each gene model with that of the single-copy genes.

Filtered genomic DNA reads were mapped to the local assemblies Piabi_c1.0 and Pigla_c1.0. Normalisation of average coverage of the gene models was performed using the trimmed mean of M value (TMM), which accounted for factors of variations, such as sequencing depth, library composition, and gene length, with the edgeR package [59]. Copy numbers of *βglu-1* and *Ugt5b* were then calculated as their normalised counts divided by the geometric mean of the normalised counts of high-confidence single-copy genes. We only considered *Ugt5b* but not *Ugt5* because only *Ugt5b* was responsible for the synthesis of the biologically prevalent pungenin [14]. High-confidence single-copy genes were defined as those both within one standard deviation from mean for the log-normalised coverage in the reference assembly and classified as single-copy in the local assembly. CNV was detected by two-way ANOVA to test the significance of the effects of provenance, gene models, and their interaction on the total copy number.

Validation of copy number quantification

We ran an independent copy number prediction with CNVPanelizer [60] as parallel analysis to our copy number quantification using single-copy-gene reference. It was performed in *P. glauca* using a subsampling strategy similar to random forest for 10,000 replicates, and in particular compared the mean ratios of copy number of the complete gene form Pigla_c1g_00044 among provenances.

Quantitative PCR (qPCR) was also used to validate the copy number results and detection of CNV. The target amplicon started at exon 12 and ended at exon 13, which was unique in complete gene forms of *βglu-1*. The forward primer was 5'–CACAACCCCGCTTGAAGAAG–3' and the reverse primer was 5'–TAACCTCGGACGTCTGCTCC–3'. The 10-μl reactions contained: 5 μl Luna Universal qPCR Master Mix (New England Biolabs), 0.25 μl forward primer (10 μM), 0.25 μl reverse primer (10 μM), and 4.5 μl of 20 ng genomic DNA. The thermal cycling profile was: 95°C 1 min, 45 × [95°C 15 s, 60°C 30 s with plate read]. The threshold cycle, known as C-half or $C_{1/2}$, for each reaction was determined based on inflection point of the sigmoidal curve of the amplification profile. The amplicons were then validated by Sanger sequencing. Copy number quantification was calculated from a standard curve based on a serial dilution of a standard synthetic 2,000-bp oligomer of *Pgβglu-1* (Twist Bioscience, United States) which was conserved among all predicted gene forms.

Results

Performance of probe set and sequence capture

We developed a probe set targeting 158 genic sequences in *Picea glauca* including the sequences of the acetophenone biosynthesis genes *Pgβglu-1*, *PgUgt5*, and *PgUgt5b*, and 139 putative single-copy genes. The probes aligned well to the homologous sequences in the reference genomes of all five *Picea* species (Supplementary Fig. 1). *P. glauca* had a 99.9% alignment rate, followed by *P. engelmannii* (98.13%), *P. abies* (97.79%), *P. sitchensis* (97.74%), and *P. engelmannii* × *P. glauca* × *P. sitchensis* (97.66%).

This newly designed probe set served for sequence capture of genomic DNA in individuals of *P. abies* ($N=38$) and *P. glauca* ($N=40$), which yielded a total 51.0 M reads with an N50 of 1,777 bp. The mean read number ± standard error (SE) for each sample was 357.49 K ± 16,397.46. The mean total sequence length ± SE was 551 ± 24.84 Mbp. The mean N50 ± SE was 1,691.79 ± 2.47 bp. We aligned the reads to the *P. abies* and *P. glauca* genomes and the mean alignment rate was 99.97%. The mean enrichment factor (target-to-background ratio) ± SE was 156.69 ± 4.64. The mean sensitivity ± SE was 69.88% ± 1.38%.

Local sequence assembly and gene models

We produced local assemblies (Piabi_c1.0 and Pigla_c1.0) with the genomic sequence capture reads using a metagenomic approach. The two assemblies shared similar statistics: they contained 3,549 and 3,369 contigs and had a total read length of 9.28 and 9.11 Mbp, with a N50 of 2,644 and 2,726 bp, respectively. In silico gene prediction and transcript evidence of PIAB_v1 and PIGL_v1 produced 1,443 and 1,552 gene models (Piabi_c1.0 g and Pigla_c1.0 g) comprising of a total of 1.38 and 1.48 Mbp, with an average size of 957.5 and 951.7 bp, respectively. The statistics for the local genome assemblies, transcriptome assemblies, and the gene models can be found in Supplementary Table 3.

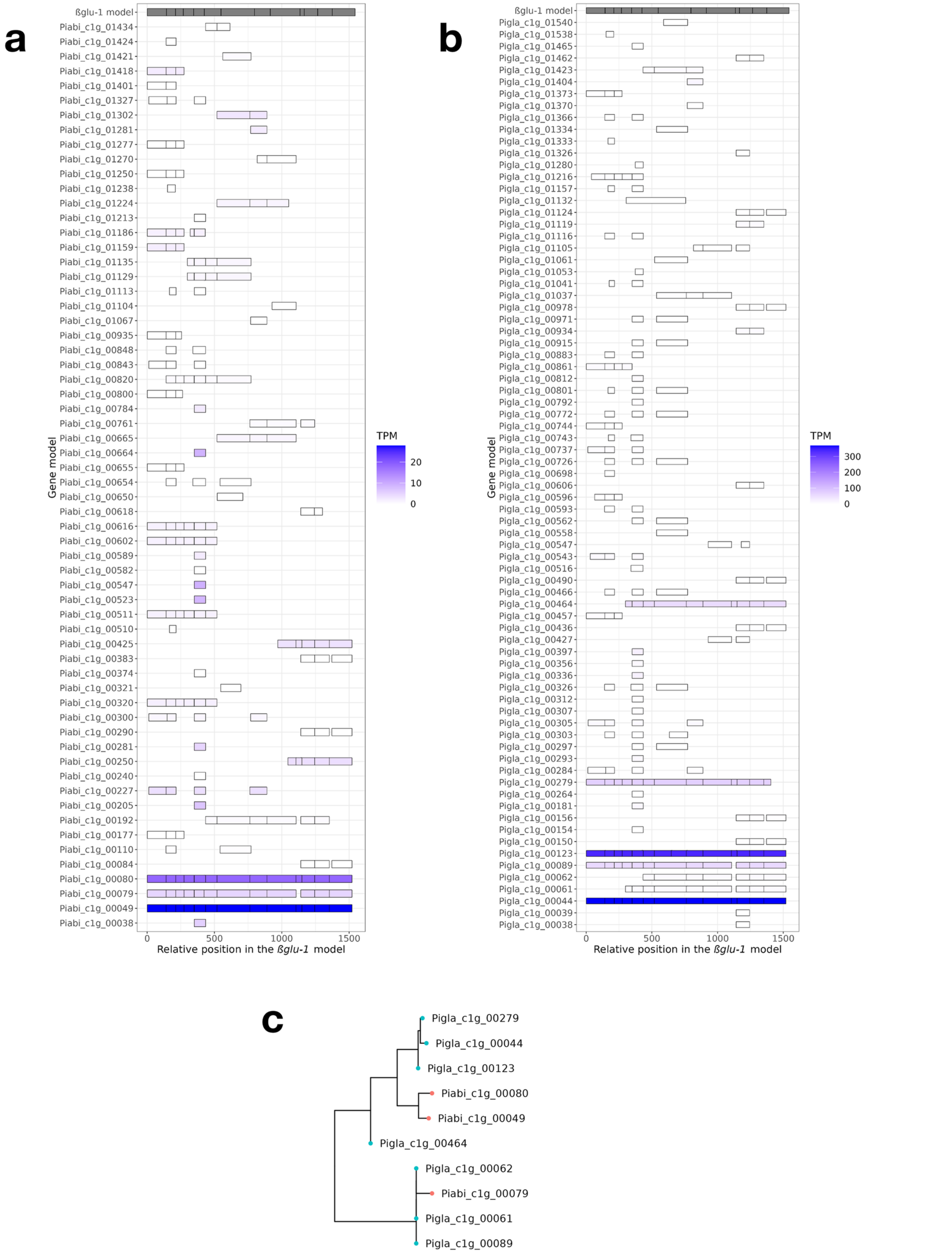

Fig. 3 Exonic correspondence of *βglu-1*-like gene models in the local assemblies of (**a**) *P. abies* and (**b**) *P. glauca* against the reference *βglu-1* gene model (KJ780719.1). The colour shows the transcript per million (TPM) of the fl-cDNA reads mapped against the gene models. (**c**) Reduced phylogenetic tree of the 10 complete and near-complete *βglu-1* gene models in the local assemblies of *P. abies* (red tips) and *P. glauca* (blue tips). The complete phylogenetic tree of all 139 *βglu*-1-like gene models can be found in Supplementary Fig. 3

Gene characterisation of *βglu-1* and *Ugt5*

Sequence similarity analysis of the gene models resulting from the local assemblies with the *βglu-1* model (KJ780719.1) gave 62 and 77 targets for *P. abies* and *P. glauca* respectively, which had substantial variation in terms of transcript abundance in the sample pool (Fig. 3). A few of the gene models resembled the gene structure predicted in the *βglu-1* model (Supplementary Fig. 2) [12], which was located on the contig Pg-03r170320s1672760 and linkage group LG02 in the WS77111_v2 assembly of *P. glauca* [2]. For *P. abies*, Piabi_c1g_00049 and Piabi_c1g_00080 retained all 13 exons, while Piabi_c1g_00079 only missed the exon 10. For *P. glauca*, Pigla_c1g_00044 retained all 13 exons and Pigla_c1g_00123 had a split exon 7, while Pigla_c1g_00089 missed exon 10, Pigla_c1g_00279 had a truncated exon 13, and Pigla_c1g_00464 started with a truncated exon 4. All these complete or nearly complete gene models had considerably higher transcript levels compared to the other gene models, and transcripts in *P. glauca* were generally around tenfold more abundant than those in *P. abies*.

Sequence similarity analysis of the gene models with the *Ugt5* (KY963363.1) and *Ugt5b* (KY963364.1) resulted in 2 and 33 targets for *P. abies* and 2 and 40 targets for *P. glauca* respectively, which had substantial variation in terms of transcript abundance (Fig. 4). Many of the assembled sequences showed full-length or near-full-length correspondence with the *Ugt5b* model, which was located on the contig Pg-03r170320s1673621 and linkage group LG11 in the WS77111_v2 assembly of *P. glauca*. However, some of the *Ugt5b* models had much higher transcript abundance than other models, such as Piabi_c1g_00253, Piabi_c1g_00368, Piabi_c1g_00738, and Piabi_c1g_01026 in *P. abies*, and Pigla_c1g_00566 and Pigla_c1g_00787 in *P. glauca*.

Gene evolution of *βglu-1*

We generated a rooted phylogenetic tree of 139 gene models that had a BLAST match (E < 1e–06) with the *βglu-1* model (KJ780719.1) to reconstruct the evolution of the *βglu-1*-like genes (Supplementary Fig. 3). In particular, we investigated the complete *βglu-1* models which had 13 exons (Fig. 3c), namely Piabi_c1g_00049, Piabi_c1g_00080 in *P. abies* and Pigla_c1g_00044 and Pigla_c1g_00123 in *P. glauca*. They formed a monophyletic clade in the reduced tree and a paraphyletic clade in the full tree, closely related to an ancestral Pigla_c1g_00464 that contained an exon 4 with a 5'-end truncation. All other gene models formed a separate cluster and were highly similar to each other as revealed by the near-zero branch lengths.

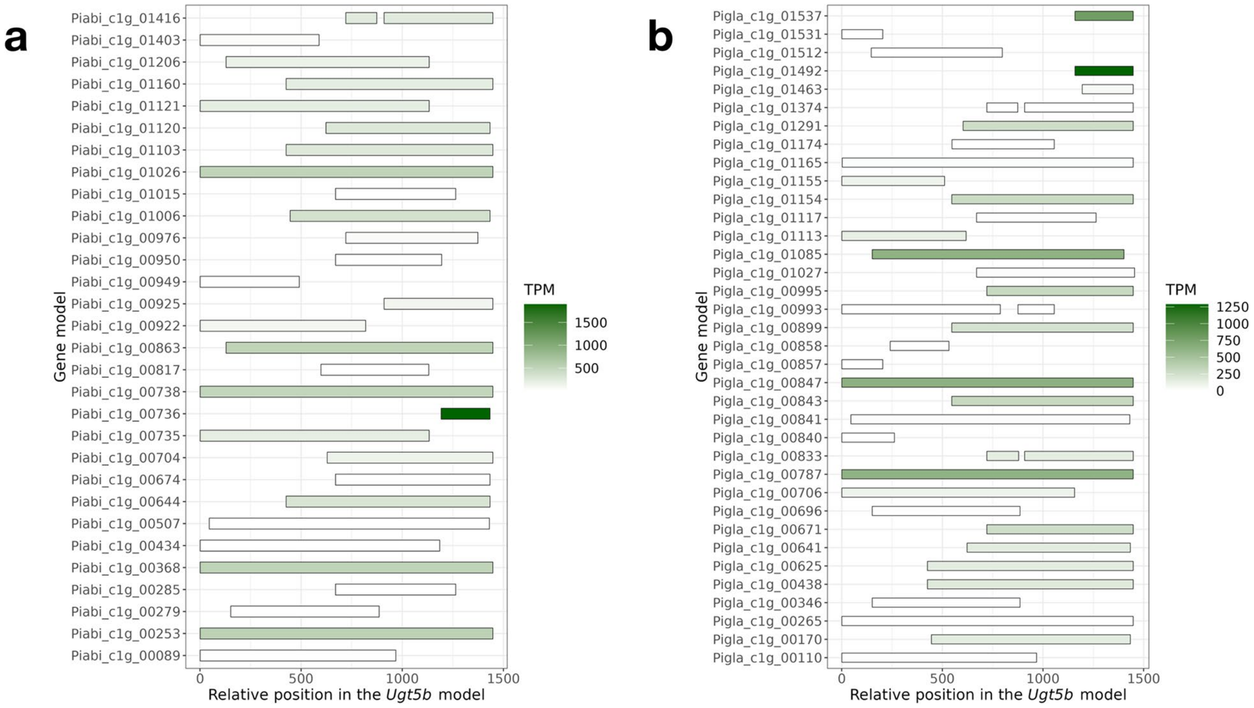

Fig. 4 Exonic correspondence of *Ugt5b*-like gene models in the local assemblies of (**a**) *P. abies* and (**b**) *P. glauca* against the reference *Ugt5b* gene model (KY963364.1). The colour shows the transcript per million (TPM) of the fl-cDNA reads mapped against the gene models

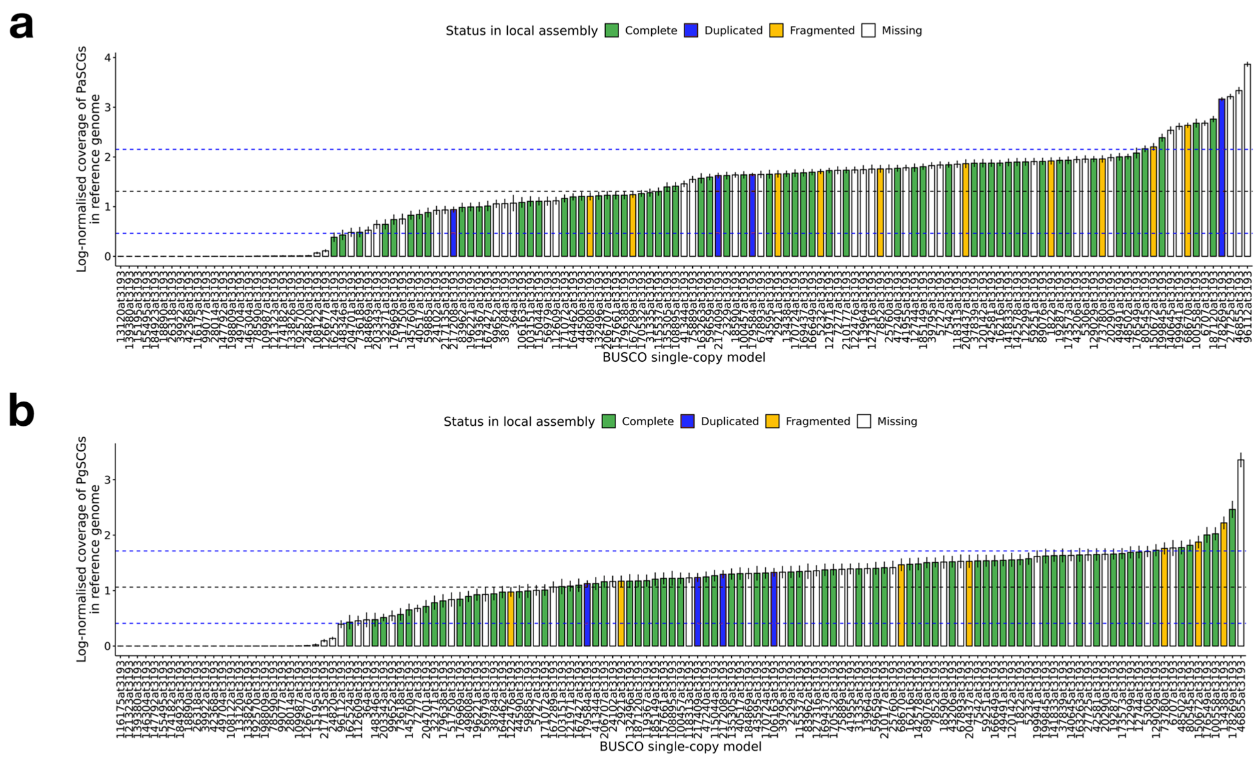

Fig. 5 Log-normalised coverage of putative single-copy genes in the reference assemblies of (**a**) *P. abies* and (**b**) *P. glauca*. The genes are arranged in ascending order of their mean log-normalised coverage. The black and blue dashed lines indicate the mean ± 1 standard error. The colour of the bars represents their status in the local assemblies Piabi_c1.0 and Pigla_c1.0 respectively using BUSCO embryophyta_odb10 database as the reference

Validation of single-copy genes

We analysed the putative single-copy gene set using cross-references from the reference and local assemblies with the BUSCO embryophyta_odb10 database as the standard. We observed that their log-normalised coverage in the reference genomes showed substantial variation in both species, with means of 1.31 and 1.06 and standard deviations of 0.84 and 0.65, respectively (Fig. 5). On the other hand, we recovered 72 single-copy, 9 duplicated, and 14 fragmented genes in Piabi_c1g, and 79, 11, 11 respectively in Pigla_c1g. We retained 54 and 62 high-confidence single-copy genes in *P. abies* and *P. glauca*, respectively, which are within 1 standard deviation in the log-normalised coverage in the reference assembly and classified as complete in the local assembly. The two species shared 48 of these high-confidence single-copy genes (Supplementary Table 4).

CNV of *βglu-1*

We observed very high copy numbers and large CNV of *βglu-1* among populations of both *P. abies* and *P. glauca* (Fig. 6a and b). The total copy number of all *βglu-1* gene models ranged between 56 copies in a *P. glauca* individual from the provenance 1338 (Twist Lake, Canada) and 381 copies in a *P. glauca* individual from the provenance 1987 (Woodlands County, Canada), representing a 6.8-fold difference between the two most extreme cases. In *P. abies*, the total copy number of *βglu-1* was significantly different among the gene models (Two-way ANOVA, $P < 2e{-}16$), but not among provenances ($P = 0.552$) and for their interaction ($P = 0.282$) (Supplementary Table 5). In *P. glauca*, it was significantly different among the provenances ($P < 2e{-}16$), among the gene models ($P < 2e{-}16$), and also for their interaction ($P = 0.000281$) (Supplementary Table 6). Overall, the mean total copy number of *βglu-1* was higher in *P. glauca* than *P. abies*, which were 160 and 127 respectively (Wilcoxon rank sum exact test, $P = 0.001535$). The mean ratios of copy number of Pigla_c1g_00044 predicted using CNVPanelizer showed high congruence with our method using single-copy-gene references ($r = 0.86$, $P = 2e{-}11$, Fig. 6c and d).

We observed a similar scale of copy number and CNV of *βglu-1* in *P. glauca* using qPCR validation (Supplementary Fig. 4). The copy number of *βglu-1* was significantly different among provenances (One-way ANOVA, $P = 6.82e{-}5$). The copy numbers were not significantly different between that assessed using sequence analysis and that using qPCR ($P = 0.065$, Welch two sample t-test).

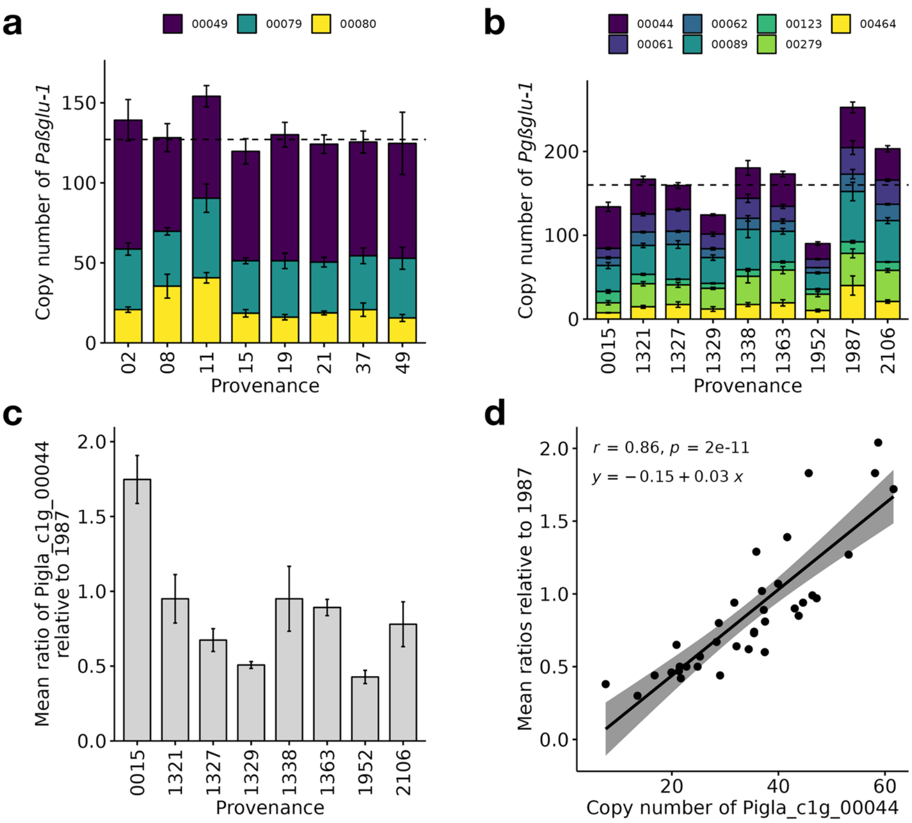

Fig. 6 Copy number of complete and near-complete *βglu-1* gene models determined by targeted sequencing in (**a**) *P. abies* and (**b**) *P. glauca* across their provenances. The error bar shows mean ± 1 standard error for each provenance. The dotted line shows the median of the total copy number. **c** Mean ratio of copy number of the complete gene form Pigla_c1g_00044 relative to provenance 1987 predicted using CNVPanelizer. **d** Correlation between the copy number of Pigla_c1g_00044 predicted using the single-copy-gene reference against the mean ratios predicted using CNVPanelizer

We did see discrepancies in the trends as generally fewer significant pairwise differences were found using qPCR than sequence analysis (Supplementary Tables 7 and 8), for example, provenance 0015 was estimated to have a higher copy number using qPCR, but the general trend and difference for other provenances were similar to than found using sequence analysis.

CNV of *Ugt5b*

In contrast to *βglu-1*, the *Ugt5b* gene models were detected as single-copy in most provenances in both species, with the medians of total copy number close to 1 (Fig. 7) and no significant differences (Wilcoxon rank sum test, $P = 0.135$). However, there were a few clear outliers in both species. Specifically, the two *P. abies* provenances (08 and 11) had copy numbers of 14 and 12, respectively, and the two *P. glauca* provenances (1329 and 1952) had copy numbers of 11 and 4, respectively. The effects of provenance, gene model, and their interaction were all significant for both *P. abies* ($P < 0.05$) (Supplementary Table 9) and *P. glauca* ($P < 0.05$) (Supplementary Table 10).

Discussion

A novel method and probes for studying CNV in spruce gigagenomes

We have developed a long-insert target capture protocol to assess copy number of genes, which we optimised for a high throughput and quantitative comparison across samples. Most available copy number quantification algorithms have been developed for human disease diagnostics, particularly in cancers, where pairing of a control reference and an affected sample is the standard method of analysis [58, 61]. A few recent studies have used coverage-based approaches for reference-free comparisons across samples and even populations [62, 63]. Our method extended these approaches by incorporating probes that target single-copy genes, which serve as internal standards for copy number quantification of target genes in contrast to relative variations. We also implemented a bioinformatic pipeline incorporating local

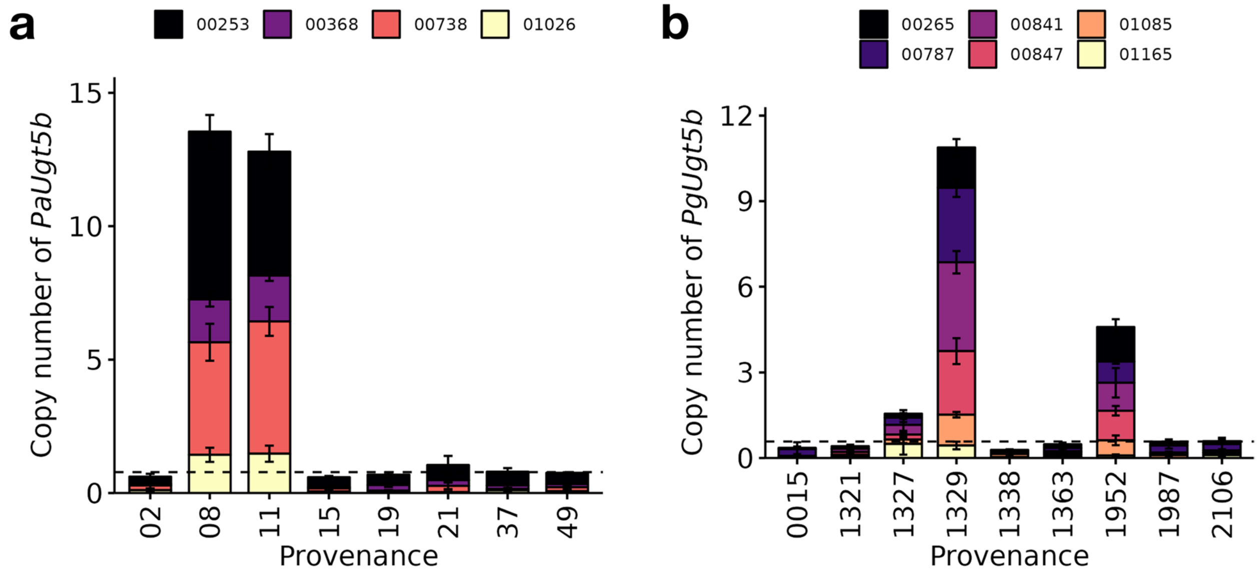

Fig. 7 Copy number of complete and near-complete *Ugt5b* gene models in (**a**) *P. abies* and (**b**) *P. glauca* across their provenances

genome assembly for each target gene, which overcomes the biases from non-uniform coverage, as assembly algorithms tend to either collapse highly similar paralogues or exclude outlier-coverage reads [64].

We assessed copy numbers of *βglu-1*-like and *Ugt5*-like genes in two keystone spruce species, *P. abies* and *P. glauca*. We also found near-perfect mapping rates of the probe set with other *Picea* spp. genomes. Therefore, the protocol is likely to be transferable to similar studies in any other conifer species, enabling large-scale, population-wide analyses that would be difficult to with other methods.

Evolution and diversity of *βglu-1*

We detected a total of five complete gene forms of *βglu-1* across the two species (Piabi_c1g_00049, Piabi_c1g_00080, Pigla_c1g_00044, Pigla_c1g_00123, Pigla_c1g_00279), which have all 13 exons and formed a paraphyletic clade. Therefore, our phylogenetic analysis suggests that an ancestral complete *βglu-1* gene emerged prior to the split of the *P. abies* and *P. glauca* lineages and only once, with gene duplication appearing to have occurred subsequently and independently within each lineage. One explanation for this observation is that the whole genome duplication in family Pinaceae around 200 to 342 million years ago, around the end-Permian extinction, may have generated new genes by sub- or neo-functionalisation [65].

The absence of exon 1–5 (1–434) and 10 (369–380) in *Pgβglu-1* still yields gene transcripts. The structural model of the PgβGLU-1 enzyme predicted that the catalytic site residues E199 and E413 form hydrogen bonds with oxygen atoms O1 and O5 of the picein substrate, while residues T202, Y206, and R293 form hydrogen bonds with the phenolic moiety of picein [12]. While the near-complete gene forms retained these functionally important residues, it is likely that the missing exons impact the enzymatic activity of the encoded protein, and these shortened gene forms also have much lower expression than the complete forms. The absence of exon 1 (1–49) would also exclude a predicted N-terminal signal peptide (1–25), which is thought to target the PgβGLU-1 protein to the vacuole with release to the lumen after cleaving the hydrophobic signal peptide [12]. This predicted protein targeting is in general agreement with the histological observation of phenolic compounds in the vacuoles of mesophyll cells in spruce trees [66].

The diversity of *βglu-1* gene copies discovered here was not captured in the most recent reference genome assemblies and gene model predictions of *P. abies* [45] and *P. glauca* [2], which only had one and two *βglu-1* gene models respectively. In *P. glauca*, the *βglu-1* gene models DB47_00003133 and DB47_00003134 had 10 and 6 exons respectively, implying errors in gene prediction and annotation. The genome of WS77111 v2 (*P. glauca*) has a BUSCO score of ~ 50%, and its gene models only ~ 20%, while that of the full-length transcriptome PIGL_v1 is ~ 85%.

Predicting single-copy genes in conifers

Surprisingly, we identified some very large copy numbers of up to 100 copies among genes that were assumed to be single-copy genes based on BUSCO prediction. This discrepancy between BUSCO prediction and our results may be explained by genomic difference between gymnosperms and other land plants. BUSCO is regarded as a reputable benchmarking tool to assess the completeness of genome assemblies and gene catalogues, with the assumption that a set of universal genes is expected to be found in a genome and in single-copy [67]. While the expectation of single-copy BUSCOs evolving under single-copy control has an evolutionary justification when the associated database (OrthoDB) was established [68], none of the 61 species used to curate the embryophyte database until OrthoDB v10 is a gymnosperm [69]. Therefore, applying BUSCO predictions in gymnosperms would assume a conserved evolutionary history of these single-copy genes among gymnosperms and angiosperms, which is not supported by our results. Some of the main differences include the rate of molecular evolution in gymnosperms being seven times slower than in angiosperms [70], Pinaceae having higher gene turnover rate than angiosperms [71], fewer polyploid species among extant gymnosperms [65], and highly distinct evolution of gene duplication and tandem-arrayed genes in conifers [72–77].

Rare extreme copy number and its variation of *βglu-1*

The scale of up to 381 gene copies in a *P. glauca* individual and the level of variation we have observed in *βglu-1* are very rare in plants. The largest genic copy number known to date was 250–300 in a protein-coding sequencing encoding a tRNA ligase gene *RLG1a* in monkeyflower (*Mimulus guttatus*) from a population-pooled sequencing [62], followed by 5–160 copies of 5-enolpyruvylshikimate-3-phosphate synthase (*EPSPS*) in carelessweed *Amaranthus palmeri* [78]. In these examples, expansion of *EPSPS* conferred herbicide resistance via a gene dosage effect, and a triple copy of *RLG1a* was also associated with elevated *RLG1a* mRNA dosage but inconclusive for the *RLG1a* hundreds-copy extreme. In *P. glauca*, the conserved *R2R3-MYB* gene duplications may contribute to ~ 100-fold increase in *PtMYB14* expression, which serves as important transcription factors in conifers, in the transgenic plantlets with a ubiquitous gene promoter than with a cinnamyl alcohol dehydrogenase tissue-preferential promoter [72]. Therefore, the contribution of CNV towards gene dosage could also be a plausible hypothesis for *βglu-1* with respect to insect resistance.

The mechanism underpinning extreme CNV remains poorly understood in plants. In the *EPSPS* case, cytogenetic analysis revealed that *EPSPS* clusters form extra-chromosomal circular DNA (eccDNA). These eccDNA are not integrated into a linear chromosome, are autonomously replicating, and transmittable to the next generations by tethering to mitotic and meiotic chromosomes [79]. Similarly, a tandem array of hundreds of *RLG1a* copies was unlinked to its single-copy locus, which leads to a working hypothesis that the cluster may also originate via rapid amplification in eccDNA and re-insertion into the nuclear genome [62].

While eccDNAs may also provide a plausible mechanism to explain the emergence of large and variable copy numbers of *βglu-1* in *P. abies* and *P. glauca*, it remains a challenge to confirm. Many conifer genome assemblies are still highly fragmented and poorly annotated, due to their large size, long introns, and repetitivenes [71]. For example, ~ 30% and ~ 44% of the initially annotated genes in *P. abies* and loblolly pine (*Pinus taeda*) were fragmented or split into multiple scaffolds [45, 80]. Therefore, it is challenging to align the extreme CNV of *βglu-1* in this study to the reference genomes and gene models, which may be collapsed during assembly, and to pinpoint their location. While conifer genome assemblies are expected to improve in the short future due to recent genomic advances, fluorescence in situ hybridisation probes mapping of *βglu-1* may be another approach to explore the origin of the very large number and variation of *βglu-1* copies.

Among-provenance CNV in P. glauca but not in P. abies

CNV among provenances was only observed in *P. glauca* but not in *P. abies*. At the same time, the level of gene expression is on average more than tenfold higher in *P. glauca* than in *P. abies*. Although fitness and survival are not directly tested in this study, these two findings indicate natural selection acting on *Picea* species.

The *βglu-1* gene contributes to resistance against spruce budworm *C. fumiferana* in *P. glauca* [10, 11, 16], but *P. abies* is not affected by *C. fumiferana* in its natural species range [81]. Therefore, absence of CNV and low expression of *βglu-1* in *P. abies* may imply the absence of dosage effect of *βglu-1*, as it does not confer evolutionary advantage to the species. Meanwhile, low dosage of *βglu-1* may limit the autotoxicity and allelopathy of pungenol from foliage leachate, which could potentially decrease seed germination, in *P. abies* [82]. In contrast, the divergent evolution of copy number of *βglu-1* in *P. glauca* may correspond to the varying abundance and genetic structure of *C. fumiferana* across its distribution range [10, 83].

The contrast of a very high copy number of *βglu-1* and single-copy *Ugt5b*

We analysed and observed large differences in gene copy numbers between *Ugt5b* and *βglu-1*, which encode enzymes that catalyse consecutive reactions in the spruce acetophenone pathway [12, 14–16]. UGT5B glycosylates pungenol to form pungenin, and *βglu-1* releases piceol and pungenol from picein and pungenin, respectively. Tandem repeats of genes within the same pathway tend to possess a high degree of conservation in co-expression patterns, as found in the CNV in the benzylisoquinoline alkaloid pathway in opium poppy (*Papaver somniferum*) [84], proteasome and ribosome pathways in *Arabidopsis thaliana* [85], and duplicated genes in *Caenorhabditis elegans* [86]. Such a conserved co-expression may contribute to an increased flux through a metabolic pathway.

Perhaps surprisingly, we found that in both *P. abies and P. glauca*, the upstream gene *Ugt5b* is predominantly single-copy, while the downstream gene *βglu-1* is present with very high copy number. This implies that the acquisition and maintenance of additional copies in *βglu-1* does not effect a similar increase in copy number of *Ugt5b*. The two genes *Pgβglu-1* and *PgUgt5b* are located on different linkage groups (LG11 and LG02 respectively in *P. glauca*) and do not form operon-like gene clusters that might facilitate co-inheritance and co-regulation [87]. This is in contrast with some known gene clusters in plants which encode secondary metabolites for defence, such as benzoxazinoids, cyanogenic glucosides, terpenoids, and alkaloids [88].

We propose three different hypotheses to explain the difference of copy number between genes that encode consecutive steps in the spruce acetophenone pathway. First, the two enzymes may have very different levels of enzymatic activity, including different substrate affinity and turnover, and thus the biochemical pathway may be well balanced despite a very different gene dosage. Second, it is possible that the majority of the *βglu-1* copies are not expressed, or their transcripts are not translated, and thus the effective dosage may be similar to that of *Ugt5b*. Third, it is possible that the two consecutive steps of the acetophenone pathway may not be acting in synchrony. The glucoside pungenin may accumulate slowly and constitutively over a period of time during the early summer based on the expression of a single-copy *Ugt5b*, while a larger gene dosage of *βglu-1* may be required later in mid-summer to release the defensive acetophenone, when the larval stage of spruce budworm is the most damaging [10].

Besides the observation on the drastic difference in copy numbers, provenances 1329 and 1952 in *P. glauca* have higher copy numbers of *PgUgt5b* than other provenances, yet lower copy numbers of *Pgβglu-1*. This may imply an accumulation of glucosidic forms without effective synthesis of the acetophenones, and thus may lead to higher susceptibility in these provenances. However, without gene expression and metabolic data, we caution that we are not able to draw a conclusive relationship between the genotypes and phenotypes.

Implications for genomic research in conifers

Our study provides a detailed characterisation of the gene forms, their sequences, and CNVs of a beta-glucosidase gene *βglu-1* in *P. abies* and *P. glauca*, with implications for understanding insect resistance in keystone spruce species and the evolution of complex conifer genomes. We generate original hypotheses for testing the emergence and mechanism of large and variable copy numbers, the potential dosage effect on phenotype, and the varying copy number of other genes within the same pathway. The next steps along this line of research would thus involve testing relationship between the absolute copy number of these defence-related genes, their expression, and the products using a bigger sample size and wider geographic coverage.

The challenge of incomplete and uncertain genome assemblies, despite advances in sequencing technologies, is likely to continue for conifer gigagenomes. We have developed experimental and bioinformatic approaches that overcome the limits of many current genome assemblies, and also developed new resources for genomic research in conifers.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-09978-6.

Additional file 1. Supplementary Table 1. Provenances of the samples analysed in this study for *P. abies* and *P. glauca*. **Supplementary Table 2.** Details of the genomic regions targeted by the sequence capture bait set Picea_hung_p1.0. **Supplementary Figure 1.** Alignment rates of the probe set Picea_hung_p1.0 on the reference genomes of 5*Picea* species. **Supplementary Table 3.** Statistics of local genome assembly, transcriptome assembly, and gene models of*P. abies* and *P. glauca*. **Supplementary Figure 2.** Syntenic relationship of 7 complete or near-complete *Pgβglu-1* gene forms in *P. glauca* with the reference *Pgβglu-1* gene model (KJ780719.1). **Supplementary Figure 3.** Phylogenetic tree of 139 *βglu-1*-like gene models in the local assemblies of *P. abies* (red tips) and *P. glauca* (blue tips) and their protein multiple sequence alignment. The bold tip labels denote the complete and near-complete gene forms of*βglu-1*. **Supplementary Table 4.** Details of the high-confidence single-copy genes shared by *P. abies* and*P. glauca*. **Supplementary Table 5.** Analysis of variance (ANOVA) table of the total copy number of *Paβglu-1* in *P. abies*. **Supplementary Table 6.** Analysis of variance (ANOVA) table of the total copy number of *Pgβglu-1* in *P. glauca*. **Supplementary Figure 4.** (a) Standard curve of *Pgβglu-1* synthetic oligomer. (b) Copy number of *Pgβglu-1* in *P. glauca* across their provenances estimated by qPCR. The error bar shows mean± 1 standard error. **Supplementary Table 7.** TukeyHSD pairwise comparison table of the total copy number of *Pgβglu-1* in *P. glauca* estimated with sequence analysis. **Supplementary Table 8.** TukeyHSD pairwise comparison table of the total copy number of *Pgβglu-1* in *P. glauca* estimated with qPCR. **Supplementary Table 9.** Analysis of variance (ANOVA) table of the total copy number of *PaUgt5b* in *P. abies*. **Supplementary Table 10.** Analysis of variance (ANOVA) table of the total copy number of *PgUgt5* in *P. glauca*.

Additional file 2. Supplementary Data 1. Probe sequences of Picea_hung_p1.0.

Acknowledgements

Not applicable.

Authors' contributions

T.H.H.: designed the study, processed the DNA samples, conceived and conducted the experimental and bioinformatic analyses, drafted the manuscript, and secured funding for the project;E.T.Y.W.: processed the RNA samples, and revised the manuscript;P.Z.: collected the samples and revised the manuscript;A.J.: collected the samples, revised the manuscript, and secured funding for the project;A.U.: collected the samples and revised the manuscript;N.E.: collected the samples, revised the manuscript, and secured funding for the project;J.Boh.: developed and provided the initial genomic data for P. glauca, and revised the manuscript;J.Bou.: developed and provided the initial genomic data for P. glauca, and revised the manuscript;I.B.: developed and provided the initial genomic data for P. glauca, and revised the manuscript;S.M.C.: supervised the study, revised the manuscript, and secured funding for the project;J.J.M.: designed and supervised the study, revised the manuscript, and secured funding for the project.

Funding

This work was supported by funding to T.H.H., E.T.Y.W., A.J., N.E., S.M.C., J.J.M. from the Leverhulme Trust Research Project Grant (RPG-2021–264). The contributions by J.Boh., J.Bou., and I.B. were supported through the 'Spruce-Up' project with funds from Genome Canada, Genome British Columbia, Génome Québec, and the Natural Sciences and Engineering Research Council of Canada (NSERC).

Availability of data and materials

Sequence data generated in this project, including the raw reads of genomic sequence capture and full-length transcriptome sequencing, local genome assemblies (Piabi_c1.0 and Pigla_c1.0), and full-length transcriptome assemblies (PIAB_v1 and PIGL_v1), can be publicly accessed in NCBI GenBank under the BioProject PRJNA940966 (http://www.ncbi.nlm.nih.gov/bioproject/940966). BioSamples are also registered under this BioProject.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 18 August 2023 Accepted: 5 January 2024
Published online: 27 January 2024

References

1. De La Torre AR, et al. Insights into Conifer Giga-Genomes. Plant Physiol. 2014;166:1724–32.
2. Gagalova KK, et al. Spruce giga-genomes: structurally similar yet distinctive with differentially expanding gene families and rapidly evolving genes. Plant J. 2022;111:1469–85.
3. Hanewinkel M, Cullmann DA, Schelhaas MJ, Nabuurs GJ, Zimmermann NE. Climate change may cause severe loss in the economic value of European forest land. Nat Climate Change. 2012;3(3):203–7.
4. Brasier C, Webber J. Sudden larch death. Nature. 2010;466(7308):824–5.
5. Raffa KF, Powell EN, Townsend PA. Temperature-driven range expansion of an irruptive insect heightened by weakly coevolved plant defenses. Proc Natl Acad Sci U S A. 2013;110:2193–8.
6. Wiggins NL, Forrister DL, Endara MJ, Coley PD, Kursar TA. Quantitative and qualitative shifts in defensive metabolites define chemical defense investment during leaf development in Inga, a genus of tropical trees. Ecol Evol. 2016;6:478–92.
7. Whitehill JGA, Bohlmann J. A molecular and genomic reference system for conifer defence against insects. Plant Cell Environ. 2019;42:2844.
8. Celedon JM, Bohlmann J. Oleoresin defenses in conifers: chemical diversity, terpene synthases and limitations of oleoresin defense under climate change. New Phytol. 2019;224:1444–63.
9. Kraus C, Spiteller G. Comparison of phenolic compounds from galls and shoots of *Picea glauca*. Phytochemistry. 1997;44:59–67.
10. Parent GJ, et al. Insect herbivory (*Choristoneura fumiferana*, Tortricidea) underlies tree population structure (*Picea glauca*, Pinaceae). Sci Rep. 2017;1(7):1–11.
11. Delvas N, Bauce É, Labbé C, Ollevier T, Bélanger R. Phenolic compounds that confer resistance to spruce budworm. Entomol Exp Appl. 2011;141:35–44.
12. Mageroy MH, et al. Expression of the β-glucosidase gene *Pgβglu-1* underpins natural resistance of white spruce against spruce budworm. Plant J. 2015;81:68–80.
13. Parent GJ, Giguère I, Mageroy M, Bohlmann J, MacKay JJ. Evolution of the biosynthesis of two hydroxyacetophenones in plants. Plant Cell Environ. 2018;41:620–9.
14. Mageroy MH, et al. A conifer UDP-sugar dependent glycosyltransferase contributes to acetophenone metabolism and defense against insects. Plant Physiol. 2017;175:641.
15. Mageroy MH, et al. In vivo function of *Pgβglu-1* in the release of acetophenones in white spruce. PeerJ. 2017;7(5):e3535.
16. Parent GJ, et al. Hydroxyacetophenone defenses in white spruce against spruce budworm. Evol Appl. 2020;13:62–75.
17. Iafrate AJ, et al. Detection of large-scale variation in the human genome. Nat Gen. 2004;36(9):949–51.
18. Sebat J, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;1979(305):525–8.
19. Redon R, et al. Global variation in copy number in the human genome. Nature. 2006;444(7118):444–54.
20. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Gen. 2015;16(31 6):172–83.
21. Werdyani S, et al. Germline INDELs and CNVs in a cohort of colorectal cancer patients: their characteristics, associations with relapse-free survival time, and potential time-varying effects on the risk of relapse. Cancer Med. 2017;6:1220–32.
22. Pös O, et al. DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. Biomed J. 2021;44:548–59.
23. Mileyko Y, Joh RI, Weitz JS. Small-scale copy number variation and large-scale changes in gene expression. Proc Natl Acad Sci U S A. 2008;105:16659–64.
24. Gamazon ER, Stranger BE. The impact of human copy number variation on gene expression. Brief Funct Genomics. 2015;14:352.
25. Šķipars V, Rauda E, Snepste I, Krivmane B, Rungis D. Assessment of gene copy number variation of Scots pine thaumatin-like protein gene using real-time PCR based methods. Tree Genet Genomes. 2017;13:1–13.
26. Šķipars V, Krivmane B, Ruņģis D. Thaumatin-like protein gene copy number variation in Scots pine (*Pinus sylvestris*). Environ Exper Biol. 2011;9:75–81.
27. Roach CR, Hall DE, Zerbe P, Bohlmann J. Plasticity and evolution of (+)-3-carene synthase and (−)-sabinene synthase functions of a sitka spruce monoterpene synthase gene family associated with weevil resistance. J Biol Chem. 2014;289:23859.
28. Hall DE, et al. An integrated genomic, proteomic and biochemical analysis of (+)-3-carene biosynthesis in Sitka spruce (*Picea sitchensis*) genotypes that are resistant or susceptible to white pine weevil. Plant J. 2011;65:936–48.
29. Neves LG, Davis JM, Barbazuk WB, Kirst M. A high-density gene map of loblolly pine (*Pinus taeda* L.) based on exome sequence capture genotyping. G3. 2014;G3(4):29–37.
30. Pinosio S, et al. Characterization of the poplar pan-genome by genome-wide identification of structural variation. Mol Biol Evol. 2016;33:2706.
31. Aoyagi Blue Y, Kusumi J, Satake A. Copy number analyses of DNA repair genes reveal the role of poly(ADP-ribose) polymerase (PARP) in tree longevity. iScience. 2012;24:102779.
32. Prunier J, Caron S, MacKay J. CNVs into the wild: screening the genomes of conifer trees (*Picea* spp.) reveals fewer gene copy number variations in hybrids and links to adaptation. BMC Genomics. 201;18.
33. Prunier J, et al. Gene copy number variations in adaptive evolution: The genomic distribution of gene copy number variations revealed by

genetic mapping and their adaptive role in an undomesticated species, white spruce (*Picea glauca*). Mol Ecol. 2017;26:5989–6001.

34. Leitch, I. J., Johnston, E., Pellicer, J., Hidalgo, O. & Benneett, M. D. Plant DNA C-values database (release 7.1). Royal Botanic Gardens, Kew https://cvalues.science.kew.org/ (2019).
35. Prunier J, Verta JP, Mackay JJ. Conifer genomics and adaptation: at the crossroads of genetic diversity and genome function. New Phytol. 2016;209:44–62.
36. Stival Sena J, et al. Evolution of gene structure in the conifer *Picea glauca*: A comparative analysis of the impact of intron size. BMC Plant Biol. 2014;14:1–16.
37. Cheetham SW, Faulkner GJ, Dinger ME. Overcoming challenges and dogmas to understand the functions of pseudogenes. Nat Rev Gen. 2019;21(3):191–201.
38. Amarasinghe SL, et al. Opportunities and challenges in long-read sequencing data analysis. Gen Biol. 2020;21(1):1–16.
39. Zeltiņš, et al. Adaptation capacity of norway spruce provenances in Western Latvia. Forests. 2019;10:840.
40. Sebastian-Azcona J, Hamann A, Hacke UG, Rweyongeza D. Survival, growth and cold hardiness tradeoffs in white spruce populations: Implications for assisted migration. For Ecol Manage. 2019;433:544–52.
41. Caudullo G, Welk E, San-Miguel-Ayanz J. Chorological data for the main European woody species. Mendeley Data. 2022;15:2021.
42. Little, E. L. Jr. Atlas of United States Trees. 1999.
43. Camacho C, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
44. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2012;38:4647–54.
45. Nystedt B, et al. The Norway spruce genome sequence and conifer genome evolution. Nature. 2013;497:579–84.
46. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
47. Wick, R. R. Porechop. Preprint at https://github.com/rrwick/Porechop (2018).
48. Nip, K. M. et al. Reference-free assembly of long-read transcriptome sequencing data with RNA-Bloom2. bioRxiv 2022.08.07.503110 (2022) https://doi.org/10.1101/2022.08.07.503110.
49. Wala JA, et al. SvABA: Genome-wide detection of structural variants and indels by local assembly. Genome Res. 2018;28:581–91.
50. Koren S, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:gr.215087.116.
51. Holt C, Yandell M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.
52. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. 2018;34:2490–2.
53. Darriba DI, et al. ModelTest-NG: a new and scalable tool for the selection of dna and protein evolutionary models. Mol Biol Evol. 2020;37:291–4.
54. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019;35:4453–5.
55. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8:28–36.
56. Neph S, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012;28:1919–20.
57. Hummel M, Bonnin S, Lowy E, Roma G. TEQC: an R package for quality control in target capture experiments. Bioinformatics. 2011;27:1316–7.
58. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. BMC Bioinformatics. 2013;14:1–16.
59. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:1–9.
60. Oliveira, C. & Wolf, T. CNVPanelizer: Reliable CNV detection in target sequencing applications. Preprint at https://doi.org/10.18129/B9.bioc.CNVPanelizer (2023).
61. Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinformatics. 2017;18:1–13.
62. Nelson TC, et al. Extreme copy number variation at a tRNA ligase gene affecting phenology and fitness in yellow monkeyflowers. Mol Ecol. 2019;28:1460–75.
63. Schiessl S, Huettel B, Kuehn D, Reinhardt R, Snowdon R. Post-polyploidisation morphotype diversification associates with gene copy number variation. Sci Rep. 2017;7(1):1–18.
64. Hahn MW, Zhang SV, Moyle LC. Sequencing, assembling, and correcting draft genomes using recombinant populations. G3. 2014;G3(4):669–79.
65. Li, Z. et al. Early genome duplications in conifers and other seed plants. Sci Adv 1, (2015).
66. Kettrup AAF, Kicinski HG, Masuch G. Investigating the effect of hydrogen peroxide on Norway spruce trees. Anal Chem. 1991;63:1047–56.
67. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
68. Waterhouse RM, Zdobnov EM, Kriventseva EV. Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. Genome Biol Evol. 2011;3:75.
69. Kriventseva EV, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2019;47:D807–11.
70. De La Torre AR, Li Z, Van De Peer Y, Ingvarsson PK. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. Mol Biol Evol. 2017;34:1363–77.
71. Casola C, Koralewski TE. Pinaceae show elevated rates of gene turnover that are robust to incomplete gene annotation. Plant J. 2018;95:862–76.
72. Bedon F, et al. Subgroup 4 R2R3-MYBs in conifer trees: gene family expansion and contribution to the isoprenoid- and flavonoid-oriented responses. J Exp Bot. 2010;61:3847–64.
73. Sena JS, Giguère I, Rigault P, Bousquet J, Mackay J. Expansion of the dehydrin gene family in the Pinaceae is associated with considerable structural diversity and drought-responsive expression. Tree Physiol. 2018;38:442–56.
74. Guillet-Claude C, Isabel N, Pelgas B, Bousquet J. The evolutionary implications of knox-i gene duplications in conifers: correlated evidence from phylogeny, gene mapping, and analysis of functional divergence. Mol Biol Evol. 2004;21:2232–45.
75. Van Ghelder C, et al. The large repertoire of conifer NLR resistance genes includes drought responsive and highly diversified RNLs. Sci Rep. 2019;9:1–13.
76. Warren RL, et al. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. Plant J. 2015;83:189–212.
77. Pavy N, et al. A high-resolution reference genetic map positioning genes for the conifer white spruce: structural genomics implications and correspondence with physical distance. Plant J. 2017;90:189–203.
78. Gaines TA, et al. Gene amplification confers glyphosate resistance in *Amaranthus palmeri*. Proc Natl Acad Sci U S A. 2010;107:1029–34.
79. Koo DH, et al. Extrachromosomal circular DNA-based amplification and transmission of herbicide resistance in crop weed Amaranthus palmeri. Proc Natl Acad Sci U S A. 2018;115:3332–7.
80. Wegrzyn JL, et al. Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. Genetics. 2014;196:891–909.
81. Rummukainen A, Julkunen-Tiitto R, Räisänen M, Lehto T. Phenolic compounds in Norway spruce as affected by boron nutrition at the end of the growing season. Plant Soil. 2007;292:13–23.
82. Ruan X, et al. Autotoxicity and allelopathy of 3,4-Dihydroxyacetophenone isolated from *Picea schrenkiana* needles. Molecules. 2011;16:8874–93.
83. Lumley LM, et al. Continent-wide population genomic structure and phylogeography of North America's most destructive conifer defoliator, the spruce budworm (*Choristoneura fumiferana*). Ecol Evol. 2020;10:914–27.
84. Li Q, et al. Gene clustering and copy number variation in alkaloid metabolic pathways of opium poppy. Nat Commun. 2020;11(1):1–13.
85. Williams EJB, Bowles DJ. Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Res. 2004;14:1060.

86. Lercher MJ, Blumenthal T, Hurst LD. Coexpression of neighboring genes in caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res. 2003;13:238.
87. Osbourn A. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. Trends Genet. 2010;26:449–57.
88. Boycheva S, Daviet L, Wolfender JL, Fitzpatrick TB. The rise of operon-like gene clusters in plants. Trends Plant Sci. 2014;19:447–59.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.