Supplementary Information

1. Supplementary Methods

All references to Figshare refer to the associated data repository at doi: 10.17044/scilifelab.28737623

All references to custom scripts used or to the tallgran github refer to the repository available at: https://github.com/tallgran

Genome sequencing and assembly

Material for genome sequencing

All material for the genome sequencing of Norway spruce was prepared from a root-grafted clonal copies of the genotype Z4006 (Nystedt et al., 2013) growing at Skogforsk, Sävar, Sweden. The original tree was sampled in Ragunda, central Sweden, collected in 1959. All Scots pine material was obtained from a clonal copy of genotype Y3088 growing at the same location (Skogforsk, Sävar). For both species, young expanding needles were sampled from multiple branches of a single clonal replicate just after initiation of bud break in the spring. This pool of needles was used for isolation of HMW gDNA, nuclei isolation, preparation of the optical map and RNA extraction.

Nuclei isolation

To isolate high quality high molecular weight genomic DNA (HMW gDNA), we followed the method described in (Zhang *et al.*, 2012) with certain modifications. In brief, the young expanding needles were powdered by grinding in liquid nitrogen. The powdered material was homogenised in nuclei isolation buffer in a ratio of 1:10 (w/v) on ice and incubated for 15 minutes with gentle rotation at 4 °C. The homogenates were gradually filtered through two layers of pre-wetted cheesecloth and one layer of Miracloth under gravity. The homogenates were centrifuged 1900 x g for 10 min at 4 °C to collect the nuclei, and pellets were resuspended in 2 ml NIB with help of paint brush and transferred to 15 ml falcon tubes for further washing with NIB. The wash process consists of resuspension with help of paintbrush in NIB followed by centrifugation (1900xg) for 4-5 times until supernatant becomes clear from any green colour and turbidity. The purity of nuclei preparation was accessed under Axioplan 2 (Ziess) microscope after DAPI staining.

DNA isolation, RNA and short fragment removal

The nuclei were resuspended in lysis buffer (200 mM Tris, 50 mM EDTA, 1.4 M NaCl, 1% (w/v) PVP-40, 2% (w/v) CTAB) supplements with Proteinase-K (2.5µg/µl) and incubated at 65 °C for 40-50 min with intermittent inversion of tubes. After observed lysis, one volume Chloroform:Isoamyl alcohol (C:I) [24:1] was added and incubated at RT for 15 min with

shaking the mixture was centrifuged at 8000 rpm 20 min 4 °C for phase separation. The aqueous phase without disturbing the interphase was transferred to fresh falcon tube and RNAase A (10-20ug/ml) was added and incubated for 15 min at 37 °C. The C:I step was repeated and the aqueous phase containing purified DNA was collected. The DNA was precipitated by adding 2 volumes of absolute ethanol and 0.3M of sodium acetate to the collected aqueous phase, the tubes were swirled for 3-5 min and the DNA threads were collected by spooling. The spooled DNA is washed in 70 % ethanol just by cycling, dipping and air-drying cycle for 40 sec. each, 3-times in different tubes. The dried DNA was rehydrated in 1X Tris-EDTA buffer for 36-48h before proceeding. To avoid short fragment contaminations and loss of HMW gDNA, we used our own Polyethylene glycol (PEG) based solution optimised from Lis and Schleif (1975). The high quality HMW gDNA was used for long read (PacBio and Oxford Nanopore Technology) and short read (NovaSeq, Illumina) sequencing.

Oxford nanopore technology library preparation and sequencing

Long reads for samples from both the species were achieved by sequencing on the Oxford Nanopore (ONT) PromethION system. The libraries were prepared using the ONT Ligation Sequencing gDNA kit (SQK-LSK110), during the process whenever needed size selection was performed using the Circulomics Short Read Eliminator (SRE) Kit. For each library 4-5µg HMW-DNA was used, starting with shearing of HMW-DNA using Megarupor 3 (Diagenode), speed 34, targeting 14-15Kbp DNA fragment lengths. After shearing 3µg of size selected 14-15Kbp DNA was used for further library preparation as per ONT kit instruction. Quality of the libraries was assessed using the Qubit dsDNA broad range kit and sequencing using ONT Flongle flow cell (FLO-FLG001). Final sequencing was done on four PromethION flow cells (FLO-PRO002), with 300ng library loaded per flow cell and 72 hours run time.

PacBio WGS library preparation and sequencing

Multiple PacBio SMRTbell™ libraries were constructed and sequenced at Uppsala Genome Centre (National Genomics Infrastructure, Uppsala university) following the instructions described in "Procedure & Checklist - Preparing HiFi SMRTbell® Libraries using SMRTbell Express Template Prep Kit 2.0" (PN 101-853-100 Version 03 (January 2020)). In brief, a total of 67 µg of genomic DNA was sheared into 20 – 30 Kbp fragments using the Megaruptor 3 (Diagenode). After removal of ssDNA overhangs, DNA damage repair and end-repair/A-tailing, DNA fragments were ligated to hair-pin adaptors to generate SMRTbell™ libraries for circular consensus sequencing. The libraries were then subjected to nuclease treatment before they were size selected using the SageELF system (SageScience). The fractions with desired fragment length obtained during size selection went on to sequencing. In total 25 SMRTcells™ were sequenced on the Sequel® II system using Sequel® II Binding Kit 2.0, Sequencing Primer v2 and Sequencing Plate 3.0. Each cell was sequenced following PacBio's sequencing recommendations and had an On-Plate Loading Concentration of 80 - 90 pM. For Norway spruce, this resulted in an average yield of 24 Gbp per cell, totalling 601 Gbp PacBio HiFi read data in 36.9 million reads with an N50 of 16.7 Kbpp. For Scots pine, the average yield was 25 Gbp per cell, totalling 678 Gbp PacBio HiFi read data in 37.3 million reads with an N50 of 18.7 Kbpp.

PacBio IsoSeq library preparation and sequencing

In total six PacBio SMRTbell™ libraries were prepared as described in "Procedure & Checklist – Iso-Seq™ Express Template Preparation for Sequel® and Sequel II Systems" (PN 101-763-800 Version 02 (October 2019) using the NEBNext® Single Cell/Low Input cDNA Synthesis & Amplification Module, the Iso-Seq Express Oligo Kit, ProNex beads (Promega) and the SMRTbell Express Template Prep Kit 2.0. In brief, for each library preparation 300 ng of total RNA was used for cDNA Synthesis followed by cDNA Amplification. In the purification step of amplified cDNA the standard workflow was applied (sample is composed primarily of transcripts centered around 2 Kbp) for the libraries that are going to be sequenced on Sequel™ SMRT® Cell 8M v3, whereas for the libraries that are going to be sequenced on Sequel™ SMRT® Cell 1M v3 LR a modified workflow was applied (sample is enriched for short transcripts <2Kbp). After purification the amplified cDNA went into SMRTbell library construction. The ready SMRTbell libraries were then prepared for sequencing using Sequencing Primer v4 and Sequel® II Binding Kit 2.0 or Sequel® II Binding Kit 2.1 respectively. In total 4 Sequel™ SMRT® Cell 1M v3 LR and 6 Sequel™ SMRT® Cell 8M v3 were sequenced on Sequel II or Sequel IIe System using Sequel® II Sequencing Plate 2.0. The on-Plate Loading Concentration was 50 pM for Sequel™ SMRT® Cell 1M v3 LR and 80 – 110 pM for Sequel™ SMRT® Cell 8M v3. Movie time was 24 hours with a pre-extension time of 2 hours.

Hi-C, Micro-C, ATAC-Seq and ChIP-Seq assay and library preparation

For Hi-C, ATAC-seq and Micro-C, needles were ground to a fine powder in liquid nitrogen using a cooled pestle and mortar. The nuclei were purified using the Percoll gradient method as detailed in (Sikorskaite *et al.*, 2013). In brief, 2 g each of grounded tissues were resuspended in 20 ml of 1X nuclei isolation buffer (NIB) [10 mM MES-KOH (pH 5.4), 10 mM NaCl, 10 mM KCl, 2.5 mM EDTA, 250 mM sucrose, 0.1 mM spermine, 0.5 mM spermidine, 1 mM DTT] supplemented with 2 % (w/v) Polyvinylpyrrolidone-40 (PVP-40, Sigma-Aldrich) and 0.1 % Protease inhibitor cocktail (Roche) just before extraction. The homogenates were gradually filtered through two layers of pre-wetted cheesecloth and one layer of Miracloth. In filtrate, Triton-X100 was added in a dropwise manner to make the concentration 1% (v/v) and incubated for 20 min at 4 °C under gentle rotation. The homogenates were centrifuged 1900 x g for 10 min at 4 °C to collect the nuclei, and pellets were resuspended in 5 ml 1X NIB with help of paintbrush. The crude preparation of nuclei suspension was carefully loaded on the top of the pre-assembled density gradient (3 ml 60% Percoll in 1x NIB and 3 ml of 2.5 M Sucrose) by slowly releasing the solution onto the side of the tube above the 60% Percoll layer with the help of Pasteur pipette. The assembled gradient is subject to centrifugation in a swinging bucket rotor at 1200 × g for 30 min at 4 °C without breaks. The rest of the nuclei purification steps were kept as described for *Rosaceae* in (Sikorskaite *et al.*, 2013). The DAPI (0.5 µg/ml) stained nuclei were counted using Neubauer chamber (Sigma-Aldrich) under Axioplan 2 (Ziess) microscope with DAPI filter, and requisite replicates of nuclei were used for Hi-C, Micro-C and ATAC -seq library preparation.

For Micro-C data, frozen nuclei (10^6) were sent to Scilife laboratory, Stockholm and the Micro-C library was prepared using the Dovetail® Micro-C Kit according to the manufacturer's protocol. Briefly, the chromatin was fixed with disuccinimidyl glutarate (DSG) and

formaldehyde in the nucleus. The cross-linked chromatin was digested in situ with micrococcal nuclease (MNase). Following digestion, the cells were lysed with SDS to extract the chromatin fragments, and the chromatin fragments were bound to Chromatin Capture Beads. Next, the chromatin ends were repaired and ligated to a biotinylated bridge adapter, followed by proximity ligation of adapter-containing ends. After proximity ligation, the crosslinks were reversed, the associated proteins were degraded, and the DNA was purified and then converted into a sequencing library using Illumina-compatible adaptors. Biotin-containing fragments were isolated using streptavidin beads before PCR amplification. The library was sequenced on an Illumina NovaSeq 6000 platform to generate around 3000 million 2 x 150 bp read pairs.

To generate Hi-C data, the Arima High Coverage HiC kit was used following manufacturers protocols with certain modifications. In brief, 10 million plant nuclei were crosslinked following Arima-HiC workflow for mammalian cell lines in 15ml falcon tubes with 2% of formaldehyde for 10 min at RT and inverting 10 times. The crosslinking was stoped by adding 460 µl of stop solution-1, mixing by inverting 10 times and incubation on RT 5min followed by incubation on ice for 10 min. The fixed nuclei were harvested by centrifugation and resuspended in 1xPBS, and aliquoted to 1 million nuclei per tube for restriction digestion, 5'-overhangs filling and labelling with biotinylated nucleotides and proximal ligation for capturing the sequence and structure of the genome. The ligated DNA is then purified, producing pure proximally ligated DNA. The proximally ligated DNA is then fragmented by ultrasonication (Covaris E220) targeting 200bp DNA fragments, and the biotinylated fragments are enriched. The enriched fragments are then subjected to a custom library preparation protocol using Swift Biosciences Accel-NGS 2S Plus DNA Library Kit and indexes. All the library quantification and amplifications were done by using kits from KAPA biosciences for Illumina platforms (KAPA Cat # KK4824, KAPA Cat # KK2620), and sequenced in pair end mode on NovaSeq 6000.

For the ATAC-seq, libraries were prepared after optimising the amount of Tn5, number of nuclei and time of incubation. The final ATAC-libraries were prepared using 30000 nuclei tagmented for 30 min at 37 °C with 3 µl of adapter loaded-Tn5, tubes were flipped after each 10 min of incubation. The tagmented DNA was purified using MinElute PCR Purification Kit (Qiagen Cat. No. / ID: 28004) and eluted in 12 µl nuclease free water. The eluted DNA was used for sequencing library preparation using indexing primers Illumina Nextera Library Prep Kits and the size selected libraries were sequenced on NovaSeq 6000 in pair end mode.

For ChIP-seq, nuclei purification was performed as in Zhang *et al.* (2012) using 2g of frozen Spruce needle powder per histone marks. The nuclei integrity and crosslinking were performed as explained in Hi-C section.

Chromatin isolation and fractionation by sonication

The ChIP-assay was done following method developed for plants (Gendrel et al. 2005) with modifications. Fixed nuclei were thawed, and 150µL of nuclei lysis buffer (Tris-buffer, 50mM; EDTA, 10mM; SDS, 1%v/v; PMSF, 1mM and Protinase inhibitor cocktail) was added. The chromatin targeting 200bp in size were obtained by ultarsonication with Covaris E220 focused-ultrasonicator (PerkinElmer), in 1ml tube (milliTUBE 1ml AFA Fiber, Covaris) having settings of Peak power,150; Duty factor, 30; Cycles/Burst, 200 for 300 seconds.

After sonication samples were kept on ice, the chromatin was collected in fresh tube as supernatant after centrifugation at 16000g for 5 min, 4^{o}C. To access fragment size and yield of chromatin aliquot of 30µL was reverse crosslinked by adding 70µL 1XTE and 4µl, 5M NaCl solution and 1 µL, Protinase K (25mg/ml) and incubated for overnight at 65 ^{o}C, followed by purification with the ChIP DNA clean and concentrator kit according to the manufacturer's protocol (Zymo Research Corp. D5205) and eluted in 20 µl water. The concentration was measured by qubit HS-assay and the fragment distribution by running 1.5 % agarose gel.

Chromatin immunoprecipitation

Approximately 20-25ug of chromatin were used per histone marks (H3, H3K4me2, H3K4Me3, H3K27Me3 and H3K9ac). The chromatin concentration was estimated from the above aliquoted material. The chromatin samples were diluted 10 times with the ChIP dilution buffer (Tris-buffer, 16.7 mM; NaCl, 167mM; EDTA, 1.2 mM; Triton-x100, 1.1% v/v; PMSF, 1mM and Protinase inhibitor cocktail) and added to the tubes having 5µg of respective antibody (anti-H3 antibody, abcam-ab1791; anti-H3K4Me2 antibody, ab7766; anti-H3K27Me3 antibody, Merck- 07-449; anti-H3K4Me3 antibody, Catalog No: 39159 and anti-H3K9ac antibody, Catalog No: 39137 from Active motif) for the immunoprecipitation reaction and with same volume of dilution buffer for input control, and rotated overnight at 4oC. Forty microliters of Protein A-magnetic beads (Dynabeads, Invitrogen) equilibrated with Tris-buffered saline-Tween (TBST) [Tris-buffer, 50mM; NaCl, 150mM; Tween-20, 0.05% v/v] were added to test samples and same amount of TBST is added into tube containing input chromatin and rotated for 2h at 4 oC to bind with antibody. The bead bound co-immunoprecipitants were washed twice with 800 µl of low salt buffer (Tris-buffer, 20 mM; NaCl, 150 mM; EDTA, 2 mM; Triton-x100, 1 % v/v; SDS, 0.1 %v/v; H2O), once each with 500µl each of high salt buffer (Tris-buffer, 20 mM; NaCl, 500 mM; EDTA, 2 mM; Triton-x100, 1 % v/v; SDS, 0.1 %v/v; H2O) and LiCl-buffer (Tris-buffer, 10 mM; EDTA, 1 mM; LiCl, 250 mM; Igepal Ca-630, 1% v/v; Sodium deoxycholate 1% w/v; H2O) and finally 2x with Tris-EDTA buffer (Tris-buffer, 10 mM; EDTA, 1 mM; H2O). The elution solution (NaHCO3, 8.4mg/ml; SDS, 1% v/v) was freshly prepared and kept at room temperature. The elution was done in two steps, first by adding 150µl of elution buffer and incubating at room temperature for 20min, supernatant was collected in new tube. Second elution was performed by adding 50 µl of elution buffer and incubation at 65 °C for 20 min, and supernatant was added to previous collection.

Deproteination and reverse crosslinking

To reverse the crosslinking and digest proteins, 8µl of 5M NaCl solution and 2µl proteinase K (25mg/ml) was added to supernatant and incubated at 65 °C overnight. The DNA was purified by using Qiagen minielute PCR cleanup kit according to the manufacturer's protocol and eluted with reasonable amount of nuclease free water. The DNA concentration was estimated using the Qubit HS-dsDNA kit.

ChIP-seq library preparation

The ChIP-seq libraries were prepared, cleaned and size selected as per instruction for Accel-NGS 2S Plus DNA Library Kit Workflow. Equal amount of ChIP-ed dsDNA (3.75ng) from each reaction and input control were resuspended in 40 µl of Low EDTA TE buffer as input DNA for

library preparation. The library preparation was carried out as explained in kit manual in following steps, end repair-I, repair-II, ligation-I and ligation-II. The libraries were indexed by ligation and PCR-amplified to produce enough library for sequencing. The number of PCR-cycles required for specific library were estimated by running qPCR on aliquoted (10µl) PCR reactions. All the libraries were quality accessed before sequencing by measuring library concentrations using Qubit DNA-high sensitivity assays and fragment size distribution using DNA high sensitivity assay on Agilent 2100 Bioanalyzer system using Agilent High Sensitivity DNA Kit (Agilent Technologies) and sequenced in pair end mode on NovaSeq 6000.

Genome assembly

Genome assembly was performed on 30X PacBio HiFi reads using HiFiasm version 0.12-r304 (Cheng *et al.*, 2022). Trial runs using the built-in purging did not result in any significant haplotig removal and was disabled at this step (parameter: -l 0). The resulting Norway spruce assembly had a total of 33.6 Gbp in 15,222 contigs with an N50 of 10.8 Mbp representing an almost fully diploid genome. The corresponding values for Scots pine were 38,7 Gbp, 13,384 contigs and N50 of 14.4 Mbp. An initial purging of un-collapsed haplotypes was done using purge_dups version 1.2.3 (https://github.com/dfguan/purge_dups), using an all-vs-all alignments of the HiFiasm contigs by minimap2 version 2.16 (Li, 2018). Due to the characteristics of the genome, the default settings resulted in an unmanageable number of alignments, so more stringent thresholds were applied (option -m 250). Other non-default parameters were -x asm5 -I 100G -DP. The parameters set for purge_dups were manually set based on the read coverage profile.

Haplotig removal based on mapping of haploid tissue

Despite exploring various mapping and purging parameters using purge_dups, we noted a relatively large number of remaining haplotigs causing artificial duplications and gene copies to still be present in the draft purged assembly. To address this, we developed an in-house purging method based on genomic coverage of aligned reads from 30X haploid (ENA study accession PRJEB1891, experiment accession ERX242655) and 50X diploid tissue (new data from this study). All reads were aligned to the primary purged assembly using bwa-mem v0.7.17, and coverage was calculated in windows of 10 Kbp along all contigs using samtools depth v1.16. Next, we identified all windows with 20%-60% (10X-30X) of the expected diploid coverage and with low or no (less than 2X) haploid coverage. We marked these windows as haplotigs and removed them from the assembly. Contigs for which over 90% of the 10Kbpp windows were marked as haplotigs were entirely removed from the primary assembly. The other contigs were purged by splitting them into smaller contigs and removing the haplotig marked windows in between the split contigs. The coverage purge resulted in a significantly reduced “double-coverage-peak” when mapping the diploid sequence data back to the final purged assembly, indicating effective removal of additional haplotigs using the coverage-based method.

For Norway spruce, we noted a potential issue in over-purging in the draft purged assembly, with some contigs assigned by purged_dups to the REPEAT category of the alternative assembly. Thus, in a second round we again aligned all haploid and diploid reads to the updated purged assembly plus all 4874 contigs in the purge_dups REPEAT category. In this

step, contigs in the REPEAT category were moved back into the primary assembly if they displayed the expected 100% coverage of both haploid and diploid data, indicating they represented haplotigs not previously present in the updated draft purged assembly. Only complete contigs (not contig sections) were considered for recovery in this step. Following these two rounds of additional haplotig purging the updated draft haploid assembly contained 3291 contigs.

Hi-C scaffolding

Hi-C reads (described above) were mapped to the purged assembly using BWA version 0.7.17 (Li and Durbin, 2010), followed by filtering and processing by Pairtools version 1.0.2 (Open2C *et al.*, 2023). The spruce assembly was scaffolded using Salsa2 version 2.2 (Ghurye *et al.*, 2017; Ghurye *et al.*, 2019), resulting in an intermediate assembly with 1303 scaffolds. The pine assembly was scaffolded using YAHS (https://github.com/c-zhou/yahs) resulting in a total of 8544 scaffolds.

Linkage map scaffolding and manual edits

This step was only performed for Norway spruce. To further process the assembly into super-scaffolds, we utilised a previously published ultra-dense haploid genetic map of Norway spruce with more than 21,000 genetic markers (Bernhardsson *et al.*, 2019), of which 97% mapped uniquely to our primary assembly. The genetic markers were highly consistent with the Salsa2 scaffolds, with a within-scaffold Spearman's rank regression value between r=0.995 and r=1.0 for the different linkage groups. Using Allmaps (Tang *et al.*, 2015), the Salsa2 output contigs/scaffolds were linked and oriented into 12 large pseudo-chromosomes, representing about 90% of the assembly size.

The Hi-C data (see above) was aligned to the draft scaffolded assembly (processing as for the Salsa2 input, see above) and the contact map was visualised with juicer_tools version 1.14 for two rounds of manual curation. All chromosomes were manually inspected and edited to correct scaffolding errors such as contig translocations and inversions based on the combined information from the Hi-C visualisations and the genetic map information. Notably, several haplotigs with a total size of 1.0 Gbp were removed from the assembly also in this step, despite previously applied automated haplotig purging methods (see above). A few cases of complex signals in the Hi-C contact maps, or conflicting signals between the Hi-C contact maps and the genetic map, were not feasible to resolve and were noted down as regions of low confidence.

For Norway spruce, BioNano data was generated to independently validate the assembly. A BioNano assembly (Bionano Access) was largely congruent with the final reference genome assembly, and hybrid NGS/BioNano scaffolds contained 98.4% of the original sequencing data (data not shown).

Annotation of transposable elements

An improved Norway spruce transposable element (TE) library was constructed based on the previously published library (Nystedt *et al.*, 2013). Additional repeat sequences were

identified from the current assembly using RepeatModeler2 (Flynn *et al.*, 2020), followed by manual curation and annotation. The TE library was further curated to remove redundancy and mis-annotated sequences. Some sequences in the library were annotated as No Hit Found (NHF) indicating their origin is unknown and might represent strongly diverged TE families or other types of unknown repeat elements. The curated TE library was used to produce a soft masked version of the genome assembly using RepeatMasker v. 4.0.8 (https://www.repeatmasker.org/), which was used for subsequent analyses. The analysis was performed using the tool Scan for Matches (https://blog.theseed.org/servers/2010/07/scan-for-matches.html). Similarly, a Scots pine transposable element (TE) library was constructed as above.

The ratio of solo to complete LTR-TE elements was analysed by considering the nine most abundant LTR-TE elements, representing ~46,000 elements. The results confirmed a strong suppression of unequal recombination with ~6.7 complete elements to solo. There was large variation between different elements, however, there was consistent and distinct underrepresentation of solo-LTRs.

Annotation of endogenous caulimovirid sequences

Endogenous caulimovirid sequences were annotated using the branch A of the Caulifinder package (Vassilieff *et al.*, 2022). The attached libraries were complemented with 545 RT and pol domains from *Metaviridae* and *Retroviridae* retrieved from the Gypsy database (Llorens *et al.,* 2011), and with new tentative *Caulimoviridae* genome sequences identified in gymnosperm genomes and their translated proteins (unpublished). The consensus sequences generated were manually curated to filter out false positives. The remaining consensus sequences were masked with the TE library obtained from the respective genome assemblies using RepeatMasker 4.0.9 (Tarailo-Graovac & Chen, 2009) with the "-nolow" option and a maximum divergence set to 20%. The masked consensus libraries were then used to annotate their respective genome using RepeatMasker.

Genome annotation

RNA isolation

Total RNA (totRNA) and short RNA (sRNA) from different conditions and tissues of Norway spruce as well as of Scot Pine were extracted using CTAB based method (Chang *et al.*, 1993) and purified using Qiagen RNeasy Mini Kit (Cat. No: 74004) and Qiagen RNeasy MinElute Cleanup Kit (Cat. No: 74204), respectively. A 700 µl aliquot of pre-warmed extraction buffer (2% CTAB, 100mM Tris-HCl pH 8.0, 25 mM EDTA, 2.0 M NaCl, and 0.5g/L spermidine) was added to 2ml microfuge tubes containing 200-300 mg of powdered samples, mixed immediately by brief vertexing. Equal volume of Chloroform:isoamylalcohol (24:1) [C:I] was added and mixed by inverting for 20-25 times and centrifuged for 10 minutes at 10, 000 rpm, RT. The aqueous phase was transferred carefully to a new tube without disturbing the interphase, and the C:I step was repeated. The RNA precipitation was carried out by adding 2.0 volume ethanol and incubating for 1 hour at 4^{o}C, followed by centrifugation at 14 000 rpm

for 20 minutes at 4^{o}C, pellet was collected. The RNA was further purified and fractionated into totRNA and sRNA with help of kits mentioned above following manufactures instructions except initial steps, the RNA pellet was dissolved in 350 µL of RLT buffer from the Qiagen kit, added 1 volume of 70% ethanol, transferred to the column tube and spun at 10 000 rpm for 15 seconds, and the flow through containing sRNA were collected in 2ml microfuge tubes. The column was then washed with 350 µL of buffer RWT and proceeded to in-column DNase treatment following instruction of RNase-Free DNase Set (Cat no. 79254). The column was again washed with 350 µL of buffer RWT and rest of steps were kept as kit manual. The flowthrough collected from first part was added with 0.65 volume of ethanol and loaded to MinElute Cleanup column and sRNA were obtained following Qiagen RNeasy MinElute Cleanup Kit manual.

The totRNA and sRNA from each sample were quantified using the RNA High sensitivity programme in the Qubit 2.0 fluorometer (Life Technologies, Carlsbad, California, USA). The integrity and quality of the tRNA was determined with the Bioanalyzer (Agilent 2100; Agilent Technologies, Palo Alto, California, USA). The method for RNA isolation remains same throughout the manuscript until stated further. The RNA samples having integrity above 7.0 were used for transcriptome sequencing, as detailed in the supplementary methods appendix.

Full-length transcriptome PacBio sequencing (IsoSeq)

Full-length transcriptome PacBio IsoSeq libraries were prepared from equimolar pools of total RNA from different tissues of both the species Norway Spruce (Pollen, Embryo, PEM, Needle, Xylem and Phloem) and Scot pine (Pollen, Root, Needle and Seed) from same individuals as of used for genomic DNA isolation. To increase the robustness of IsoSeq full length transcript annotation total RNA from Pollen, Embryo, Needle and Xylem of Norway Spruce were sequenced separately. Different sequencing libraries were prepared for PacBio IsoSeq sequencing according to the manufacturer's recommendation. Sequencing was performed on a PacBiso Sequel II, generating non-redundant sets of IsoSeq full-length transcripts.

Illumina RNA sequencing (RNAseq)

RNAseq libraries were constructed using Illumina TrueSeq mRNA protocols according to the manufacturer's recommendation. Sequencing was performed on an Illumnia Novaseq (Illumina Inc.), producing paired 2 x 150 bp reads with an insert size of 350 bp and a target 10M paired-end reads per library.

Transcriptome reconstruction

In total, 1800 RNA-seq samples were used in this study (Table S9). The RNAseq data was grouped to construct 27 batches, with each batch representing similar tissues, experimental conditions, and developmental stages. Sequences from each RNAseq batch were aligned to the Norway spruce genome assembly and digitally normalised using Trinity v2.14.0 (Grabherr *et al.*, 2011). For each batch, the aligned reads were used for transcriptome assembly using Genome Guided Trinity Transcriptome Assembly).

Annotation of protein coding genes

An initial draft annotation was constructed to be used as a seed annotation for PASA v2.5.1 (Haas *et al.*, 2003). The IsoSeq data from the pooled tissue sample was aligned to the Norway spruce genome assembly using GMAP version 2021-07-23 (Wu and Watanabe, 2005), and a non-redundant set of non-overlapping transcripts was selected from the initial GMAP results. To complement this draft annotation, IsoSeq transcripts from *Pinus sylvestris* were aligned to the Norway spruce genome using GMAP, providing an additional 3862 potential gene models to the seed annotation. These models were selected in such a way as to only add a *Pinus sylvestris* model if there was not already a spruce model located in the same area of the genome. The final non-overlapping seed annotation contained 29226 protein coding transcripts, and was used for the subsequent PASA runs.

Each IsoSeq (n=5) and RNAseq (n=27) transcriptome set was processed using PASA, resulting in 32 separate protein coding gene annotations. All annotations were merged using AGAT version 1.0.0 (https://agat.readthedocs.io/en/latest/index.html) and in-house scripts to construct a single annotation file. As the merged annotation contained a large number of near-identical isoforms, we used indexing with Salmon v1.10.2 (https://salmon.readthedocs.io/en/latest/index.html) to create a non-redundant set of isoforms with potential to be used for isoform-specific gene expression studies. In addition, a number of unexpected artefacts produced in the PASA runs were detected and processed by custom scripts, including i) transcripts from unrelated loci being combined in the same gene feature, ii) intronic (presumably TE) genes being included in the "parent" gene feature, iii) tandemly duplicated genes being included in the same gene feature, iv) transcripts with no valid coding region (including some in anti-sense orientation to other genes).

In addition to the genes identified by PASA, 250 exact gene duplications that were not present in the initial annotation were detected and added, as well as a set of six previously published MADS-box genes that had not been identified in the PASA annotation process.

Filtering of likely transposable element genes from protein coding gene models

To filter out TE-derived gene models, the CDS sequences were run against the manually curated repeat library using RepeatMasker. All gene models with a masked fraction of 75% or more were flagged as TEs and removed. After manual inspection of the results further sequences were removed using blast with the following settings:

Functional annotation of protein coding genes

The predicted protein sequences were used as input for eggnog-mapper (version 2.1.12, Huerta-Cepas et al., 2019). In addition to eggnog-mapper, interproscan version 5.52-86.0 was used to functionally annotate the predicted proteins.

Ribosomal and transfer RNA annotation

The Norway spruce genome was searched for rRNA using barrnap (version 0.9, https://github.com/tseemann/barrnap) with '--kingdom euk'. The Norway spruce genome, including unplaced scaffolds and contigs was searched for rRNA using barrnap (version 0.9, https://github.com/tseemann/barrnap) with '--kingdom euk'. tRNA genes were identified with tRNAscan-SE V2.0.9 (Chan et al. 2021) with the setting '-I -E'. Redundancy in the fasta files of rRNA and tRNA sequences were then removed by merging sequences with greater than a 90% identity match by using CD-HIT-EST V4.8.1 (Li and Godzik 2006) with default parameters.

In parallel to the rRNAs annotated by barrnap, the NCBI was mined to retrieve rRNA sequences using the query: ("Pinus"[Organism] OR "Picea"[Organism]) AND (5S[All Fields] OR 5.8S[All Fields] OR 18S[All Fields] OR 25S[All Fields] OR 28S[All Fields]) AND rRNA[All Fields] NOT (mRNA[filter] OR cRNA[filter] OR ncRNA[filter] OR tRNA[filter]) NOT (chloroplast[filter] OR mitochondrion[filter] OR plastid[filter]). The resulting sequences were merged using CD-HIT-EST with default parameters.

Duplication patterns among genes and pseudogenes in *Picea abies* and *Pinus sylvestris*

Inference of orthogroups

Genome assemblies in fasta format and gene annotations in gff3 format were downloaded from PlantGenIE (Sundell *et al.*, 2015) for all 27 species used in the ortholog analysis (Details available at the associated Figshare resource: /Genome/comparative_genomics_accessions.tsv), except for *Pinus tabuliformis*, *Sequoiadendron giganteum* and *Taxus chinensis*. The longest CDS transcript of each gene was extracted and translated to a protein sequence using AGAT version v0.7.0 (https://github.com/NBISweden/AGAT) on each genome and annotation pair. A rooted species tree of the 27 species was determined using timetree.org (Kumar *et al.*, 2022). Groups of orthologs between and within the 24 species were identified using OrthoFinder v2.5.2 (Emms and Kelly, 2019), with the individual species protein sets and the rooted species tree as inputs. Orthofinder was run with default parameters.

Synteny and K_s analysis

To generate one-to-one orthologues, a separate run of orthofinder v 2.5.5 was performed to find the one-to-one orthologues for the subset of species of interest, which comprised *Picea abies, Pinus sylvestris, Torreya grandis, Sequoiadendron giganteum and Pinus densiflora*. Synteny plotting was performed using the R-package syntenyplotteR (Quigley *et al.*, 2023). K_s plots based on the set of one-to-one orthologs were generated using WGD2 (Chen *et al.*, 2024) with modelled distributions to detect putative large-scale duplication events (Fig S6). The lack of a systematic pattern of syntenic regions between chromosomes in *P. abies*, the lack of a clear recent Ks peak, and the clear macro-synteny patterns between *P. abies* and *T.*

grandis comprise evidence that there has been no whole-genome duplication (WGD) in the *Picaeae* lineage since the split from *T. grandis*.

Pseudogene analysis

To detect the pseudogenes, we created a set of query sequences comprising all exons from the annotated protein coding genes of each species. We extracted all the intergenic regions and aligned the exons against the intergenic regions. The alignment was performed using BLAT (Kent, 2002) using default parameters. Any repeat regions were filtered out from the blat hits. We added another layer of filtering using the following criteria, which we termed *coverage-identity*. *Coverage:* the length of the exon sequence covered or involved in the alignment. *Identity:* the sequence identity of the exon in the hit with the intergenic region. In the downstream analysis, we considered 50-50 as a coverage-identity cutoff to obtain a set of pseudogenes. We believe the cutoff is not strict enough to miss potential pseudogenes and at the same time, it is not a relaxed cutoff, so we avoid including random sequences as pseudogenes.

To identify a copy of a gene as a pseudogene, we classified any incomplete copy of a protein-coding source gene to be a pseudogene. Incomplete was defined in terms of short alignment hit length of the source genes exons and identity <100. For multi-exonic source genes, we chained the exons representing a source gene if the exons were detected within the same intergenic region. If only some of the exons of a multi-exonic source gene were detected in the same intergenic region, we considered that as a partial pseudogene, otherwise is all exons of the source gene were represented but the copy was incomplete compared to the source gene, we classified a 'complete' pseudogene. We considered complete duplication of a single exonic source gene if its only exon was duplicated but if the hit was fragmented/incomplete with respect to the source gene.

We further investigated the junction of the duplicated exons in the intergenic regions. We checked if the duplicated exons were separated by the same source gene intron length/s or larger or a shorter length. We checked if the duplicated exons were joined to each other in their target space. If they were joined, we called them retrotransposed pseudogenes. Otherwise, they were classified as a segmental duplication event. We consider a pseudogene to be retrotransposed if at least one junction is joined.

Clustering analysis of protein coding and pseudogenes

The tendency of gene duplications to occur on the same chromosome was investigated using the entropy package in R to measure the divergence of the expected distribution of duplications from random placement. To emphasize the increase of same-chromosome duplications in recent gymnosperms we categorized gene duplication events as early, middle, and late, by approximate time intervals. A duplication was considered "late" if it occurred on a terminal species branch, or on the branch immediately before a speciation event that was at most 60 MY old. "Early" duplications were those that occurred on the long branches to *Physcomitrella patens*, and the split between angiosperms and gymnosperms. Gene duplication events that occurred on any other branch of the species tree were considered "middle". A list of species tree branches and categories can be found at the associated Figshare repository (located at /Genome/Species_tree_time_table.xlsx).

Protein coding genes were observed to cluster with their paralogs more often than pseudogene copies clustered with each other. To quantify this across the genome and for all multi-copy genes and pseudogenes that diverged since the split of spruce and pine, orthologs were grouped at the level of N10 hierarchical ortholog groups (HOGs). These genes in each HOG were analysed with DBSCAN (Ester et al., 1996) with a reachability distance of 10 MB. Where the number of copies within a HOG was exactly 2, the duplication event was called local if both copies were on the same chromosome and within 10MB. Conversely the duplication was dispersed if the copies were >10MB apart or on separate chromosomes. In cases where the HOG contained > 2 genes, the number of local duplication events was equal to *n(genes) - n(clusters)*, and the number of dispersed duplication events was equal to *n(clusters) - 1*.

Syntenic block identification indicative of multi-gene duplication

In addition to identifying locally and dispersed copies of individual genes, we sought evidence of larger segments of the genome which were duplicated and contained multiple genes and/or pseudogenes. On each chromosome all non-TE protein coding genes and pseudogenes were assigned to sliding window groups of size 2-12 loci with a step of 1 locus. Each of the genes/pseudogenes in those blocks had their HOG10 membership recorded. If two or more gene/pseudogene blocks contained the identical HOG10 sequence in either forward or reverse order, then those blocks were considered to be syntenic. Syntenic blocks that were subsumed by larger blocks, and blocks that overlapped were removed. What remained were blocks of genes and pseudogenes which were represented by an identical sequence of orthologous genes and pseudogenes elsewhere in the genome. The required software used for the clustering analysis and syntenic block identification was: R 4.4.2, RStudio 2024.12.1+554, fpc 2.2-13, IRanges 2.38.0.

Detailed analysis of duplicated segments

Regions spanning 50 Kbp upstream and downstream of a duplicated gene were extracted and compared using dotplot analysis with the tool Dotter (Sonnhammer et al., 1995). Manual inspection of the results revealed sequence features characteristic of DNA transposable elements (DNA-TEs), allowing us to identify the TE carrying the gene copy. To assess the abundance of these TEs, we designed sequence patterns that summarize their key genomic features (e.g., length, target site duplications, inverted repeats). We then used these patterns to scan the entire genome with the tool Scan for Matches (https://blog.theseed.org/servers/2010/07/scan-for-matches.html). Below are the patterns for the three most abundant DNA-TEs:

pattern CACTA:
CACTACTACATTTGGGT[1,0,0] 500...40000 ACAATATGTAGTAGTG[1,0,0]
Find any CACTACTACATTTGGGT with up to one mismatch ([1,0,0]) followed by any tract with a length comprised between 500 and 40,000 bp (500...40000) followed by ACAATATGTAGTAGTG with up to one mismatch ([1,0,0])

pattern MARINER:
p1=4...4 p2=29...31 5000...90000 ~p2[1,0,0] p1

This defines an element having 4 bp long direct repeats (p1=4...4 and p1) followed by an inverted repeat (p2 and ~p2) with a length comprised between 29 and 31 bp (p2=29...31), allowing 1 mismatch ([1,0,0]). The region between the two inverted repeats has a length comprised between 5,000 and 90,000 bp (5000...90000)

pattern hAT:
p1=7...8 CATAGATGATTTTAGG[1,0,0] 1000...25000 TTAAAATCACCCCTGC[1,1,1] p1
This defines an element having direct repeats from 7 to 8 bp long (p1=7...8 and p1) followed by CATAGATGATTTTAGG with up to one mismatch ([1,0,0]) followed by any sequence with a length comprised between 1,000 and 25,000 bp (1000...25000) followed by TTAAAATCACCCCTGC with up to one mismatch, and/or insertion and/or deletion ([1,1,1])

Analysis of *FLOWERING LOCUS T (FT)/TERMINAL FLOWER 1 (TFL1)* gene family

Phylogenetic analysis of the *FLOWERING LOCUS T (FT)/TERMINAL FLOWER 1 (TFL1)* gene family included annotated *PHOSPHATIDYLETHANOLAMINE-BINDING PROTEIN (PEBP)* genes from *P. abies*, *P. sylvestris* and *A. thaliana.* For each species, the gene coding sequences were translationally aligned using the Clustal Omega module with in Geneious (Geneious prime 2023.0.1 Biometters Ltd) and subsequently trimmed using the Block Mapping and Gathering with Entropy (BMGE) method to retain the phylogenetically informative regions of the aligned sequences (Criscuolo & Gribaldo, 2010). A phylogenetic analysis of the resulting nucleotide alignment was carried out using MrBayes version 3.2.6 (Huelsenbeck & Ronquist, 2001) and the selected model of evolution was GTR + I + G. Four chains of the Markov Chain Monte Carlo were run in parallel, sampling one tree every 500 generations for 2.5 million generations starting with a random tree. The search reached stationarity after approximately 100,000 generations. The first 500,000 generations (the burn in) were omitted in generating the consensus phylogeny.

Chromatin structure and epigenetic assays in *P. abies*

ATAC-seq and ChIP-seq pre-processing

Raw reads were trimmed and quality filtered using Trimmomatic V0.39 with the setting 'ILLUMINACLIP:$TRIMMOMATIC_HOME/adapters/NexteraPE-PE.fa:2:30:10:1:TRUE SLIDINGWINDOW:5:20 MINLEN:38' for ATAC-seq reads and 'ILLUMINACLIP:$TRIMMOMATIC_HOME/adapters/TruSeq2-PE.fa:2:30:10:1:TRUE SLIDINGWINDOW:5:20 MINLEN:50' for ChIP-seq reads. Bowtie2 V2.4.5 was then used to align cleaned and trimmed reads to the Norway spruce reference genome with the setting '--very-sensitive --dovetail --maxins 1000'. Alignments were then filtered using SAMtools V1.16 to remove reads with a MAPQ score lower than 20 and discordantly paired reads by using the filtering setting '-f 3 -F 12 -q 20' with SAMtools view followed by removal of PCR duplicates with SAMtools markdup.

Peaks were then called using MACS3 V3.0.0b1 with the setting '--gsize 17e9 --keep-dup all --format BAMPE' for both the ATAC-seq and H3K4me2, H3K4me3 and H3K9ac ChIP-seq data but with the addition of the input and H3 control supplied with the '--control' flag for calling peaks on the ChIP-seq data. For calling peaks on the H3K27me3 ChIP-seq data epic2 V0.0.52

was used with the setting '--keep-duplicates' and supplying the input and H3 control with the '--control' flag for peak detection. The '--keep-dup' (MACS3) and '--keep-duplicates' (epic2) flags were used for both tools as duplicate reads had already been removed in a prior step using Samtools.

Oxford Nanopore Technologies sequence data pre-processing

Reads were base called using Oxford Nanopore Technologies Guppy V5.0.17. The nanopore current signal of base called reads classified as passed with a minimum q-score of 7 were then mapped to the Norway spruce reference using the Tombo V1.5.1 (https://github.com/nanoporetech/tombo) re-squiggle algorithm with the setting '--dna --signal-matching-score 2 --signal-align-parameters 4.2 4.2 300 3000 20.0 40 750 2500 250'. The signal-matching score was set to 2 (above the DNA default of 1.1), as recommended in Tombo's optimized settings, to retain reads during mapping, since modified base calling was performed in a subsequent step. 5mC modified bases in the DNA motif contexts CG, CHG and CHH were then identified with Deepsignal plant V0.1.6 using default settings and the signal current model 'model.dp2.CNN.arabnrice2-1_120m_R9.4plus_tem.bn13_sn16.both_bilstm.epoch6.ckpt' supplied through the developers Github page (https://github.com/PengNi/deepsignal-plant).

To assess cytosine methylation coverage and signal depth across the genome, motifs across the genome, CG, CHG, and CHH motifs were identified in the Norway spruce reference genome using Modkit V0.2.3 (https://github.com/nanoporetech/modkit) motif-bed. For each motif, we counted how many times it was overlapped by a 5mC call using BEDTools V2.31.0 intersect '-wao', considering both same strand overlaps '-s' and, for CG and CHG contexts, overlaps on either DNA strand. We then calculated the proportion of motifs that were covered by at least one 5mC call (≥1). For CG motifs, 73.23% were overlapped by ≥1 same strand 5mC call, and 83.37% were overlapped by ≥1 call on either strand. For CHG motifs, the values were 78.87% (same strand) and 86.15% (either strand). For CHH motifs, 81.63% of same strand motifs were covered by ≥1 call.

To evaluate methylation signal depth, we further calculated the proportion of motifs overlapped by increasing numbers of 5mC calls (≥1 to ≥5). For CG motifs, 73.23%, 54.49%, 37.73%, 24.32%, and 14.86% were covered by ≥1 to ≥5 calls on the same strand, while 83.37%, 73.58%, 63.96%, 53.89%, and 43.84% were covered when considering either strand. For CHG motifs, 78.87%, 61.99%, 45.32%, 30.76%, and 19.63% were covered on the same strand, and 86.15%, 77.02%, 67.84%, 58.22%, and 48.53% on either strand. For CHH motifs (same strand only), 81.63%, 66.14%, 49.80%, 34.72%, and 22.65% were covered by ≥1 to ≥5 calls, respectively.

Micro-C pre-processing

Trimmomatic V0.39 was used to trim and filter out low quality reads with the setting 'ILLUMINACLIP:$TRIMMOMATIC_HOME/adapters/TruSeq3-PE-2.fa:2:30:10:1:TRUE SLIDINGWINDOW:5:20 MINLEN:75'. Reads were then processed as outlined in the Dovetail Genomics Micro-C documentation (https://micro-c.readthedocs.io/en/latest/). In short, reads were aligned to the Norway spruce reference genome using BWA-MEM V0.7.17 set

with the '-5 -S -P' parameters. Aligned reads were then processed with Pairtools V0.3.0 framework with the following pairtools operations and settings: pairtools parse '--min-mapq 20 --walks-policy 5unique --max-inter-align-gap 30', pairtools sort (default settings), pairtools dedup '--mark-dups' and pairtools split (defaults settings). Juicer Tools V1.22.01 was then used to generate .hic files with Juicer Tools pre, using default settings for all analyzed resolutions. Cooler V0.9.3 was used to generate .cool files at all analyzed resolutions using cooler cload with default settings and the contig lengths of the Norway spruce reference followed by matrix balancing using cooler balance with the flags --cis-only --mad-max 20.

Processing of needle RNA-Seq data

RNA-Seq reads were cleaned using Sortmerna V4.3.4 and a fasta with the tRNA and rRNA sequence annotation supplied with '--ref' and '--idx-dir' and then trimmed with Trimmomatic V0.39 was used to trim and filter out low quality reads with the setting 'ILLUMINACLIP:$TRIMMOMATIC_HOME/adapters/TruSeq3-PE-2.fa:2:30:10:1:TRUE SLIDINGWINDOW:5:20 MINLEN:50'. Gene expression levels were then quantified using Salmon V1.9.0 salmon quant with the settings '--dumpEq --numGibbsSamples 100 --gcBias --posBias --seqBias'. Gene expression values estimated by Salmon were transformed using variance-stabilized transformation (VST) with DESeq2 V1.46.0. The median VST expression across replicates was assigned to each gene for downstream analyses.

Region calls of annotations, open chromatin and epigenetic marks and 5mC

Region calls over 25 Kbp and 250 Kbp, based on the sum of bases overlapping genes, intergenic LTR-TEs, and pseudogenes, were determined by creating genome tiles using BEDTools v2.29.2 (Quinlan and Hall, 2010) with BEDTools V2.31.0 makewindows '-w 25000 and -w 250000'. To generate intergenic LTR-TE annotations, LTR-TEs overlapping gene annotations were removed using BEDTools subtract with default parameters. The 25Kbp and 250Kbp tiles were then intersected with gene, intergenic LTR-TE, and pseudogene annotations using BEDTools intersect '-wao', followed by BEDTools groupby '-o sum'. Similarly, region calls for bases identified as open chromatin (ATAC-seq) or bases enriched for H3K4me2, H3K4me3, H3K9ac, and H3K27me3 histone modifications (ChIP-seq), i.e., peak calls, were determined by intersecting the same 25 Kbp and 250 Kbp tiles with ATAC-seq and ChIP-seq peaks using BEDTools intersect '-wao', followed by BEDTools groupby '-o sum'. Additionally, 5mC methylation levels for CG, CHG, and CHH contexts were determined by intersecting the tiles with 5mC methylation frequencies using BEDTools intersect '-wo', and regional averages were calculated with BEDTools groupby '-o mean'.

Identifcation and analysis of A/B, sub-A/B and sub a/b compartments

A/B, sub-A/B and sub-a/b compartments were identified using Cooltools V0.6.1 using cooltools eigs-cis. ATAC-seq signal was supplied as the '--phasing-track' to guide compartment orientation. A/B compartments were called at 250 Kbp resolution, in 100 Mbp tiles at 250Kb and 25 Kbp resolution for the sub AB and sub ab compartments respectively by providing cooltools eigs-cis with 100 Mbp window tiles with '--view-bed'.

Sub-A/B and sub-a/b compartments were intersected to evaluate sub compartmentalisation using BEDTools V2.31.0 intersect '-wo'. Tiles containing sum of bases annotated as genic, intergenic TE and pseudogene were merged with compartment calls based on the genomic

coordinates of the tile for analysis of feature overlap. For counting gene abundance and comparing gene expression the location of the TSS was used to assign the genes location using BEDTools intersect '-wo'.

Genomic feature overlap and gene expression in chromatin compartments were analyzed; see ab_and_AB_analysis.R for full details. The source code for visualizing chromatin compartment contact frequencies and saddle plots for the Norway spruce genome were obtained from the Cooltools documentation found at https://cooltools.readthedocs.io/en/latest/notebooks/compartments_and_saddles.html.

TAD calling and analysis of TADs and their border regions

Topologically Associated Domains (TADs) were called with Juicer Tools V1.22.01 Arrowhead algorithm with the '-k KR --ignore-sparsity' parameters and resolutions '-r' 25 Kbp, 50 Kbp, 100 Kbp, 250 Kbp, 500 Kbp and 1 Mbp.

For analysis of chromatin compartmentalization, open chromatin, epigenetic marks and genic features in TAD regions. 25 Kbp and 250 Kbp were selected as representative resolutions. The corresponding data tracks were converted to bigWig format with UCSC utilities 1.04.00 (https://hgdownload.soe.ucsc.edu/admin/exe/) provided with the Norway spruce contig lengths and run with default parameters. deepTools v3.3.2 was then used for visualizing the aggregated signals across TADs and TAD borders by running deepTools computematrix scale-regions '--averageTypeBins mean' and '--binSize' set to the same as the resolution used for the TAD call and the coordinates of the annotated TADs and the bigWig feature tracks. Final figures were generated with deepTools plotHeatmap '--averageTypeSummaryPlot mean' and plotProfile '--averageType mean'.

TADs were then intersected all vs all using BEDTools V2.31.0 intersect '-wao'. TADs with a 90–90% overlap across multiple resolutions, identified with the flags '-f 0.9 -F 0.9', were considered to be the same TAD identified at more than one resolution and were filtered to retain only the highest-resolution instance. TADs found to be inside of a larger TAD post removal of TADs identified at more than one resolution were labelled as "nested", see TAD_descriptive_stats.R.

TAD borders were defined as the 250 Kbp region flanking a TAD to assess whether the gene signal observed in the heatmaps originated from one or more genes, with BEDTools V2.31.0 flank '-b 250000'. TAD regions and the windows flanking borders were intersected with the locations of all annotated genes using BEDTools intersect '-wao'. For the next steps, the frequency of genes in TADs and the distribution of gene locations relative to TADs and their borders were analyzed in TAD_descriptive_stats.R. GO enrichment and TAD co-expression analyses were performed only at 250 Kbp resolution to avoid overrepresentation of regions due to nested TADs and were conducted in chromatin_structure_and_GO.R and chromatin_structure_and_coexp.R, respectively. Gene expression levels were analyzed in chromatin_structure_and_expression_specificity.R, with genes showing zero or near-zero expression filtered out specifically for the gene expression specificity analysis, see chromatin_structure_and_expression_specificity.R for details.

To test for non-random overrepresentation of long genes (>15894 bp) in the 250 Kbp windows flanking TAD boundaries, we compared the observed overlap to a null model assuming no association between gene length and TAD border proximity. We performed 1,000 permutations by shuffling gene positions to the coordinates of other genes, preserving gene length and the natural genomic distribution. For each permutation, we counted long genes overlaps and compared this to the observed count using BEDTools V2.31.0 intersect '-u'.

An empirical p-value was calculated as (r+1)/(n+1), where r is the number of permutations with values ≥ the observed, and n=1000, giving p=0.00013, with none of the random permutations exceeding the observed.

To quantify the difference between observed and expected overlap, we calculated the observed odds (~1.92:1) and the expected odds (~0.68:1) as the mean of odds across permutations. The observed-to-expected odds ratio was ~2.82, indicating nearly a threefold enrichment in the odds of long genes being at TAD boundaries relative to expectation. See obs_to_exp_long_gene_TAD_borders.R, obs_to_exp_long_gene_TAD_borders.sh.

Loop calling and analysis

Chromatin loops were identified with Juicer Tools V1.22.01 HiCCUPS algorithm with '--ignore-sparsity -k KR -r 25000 -f .1 -p 1 -d 50000' and a range of window sizes set with '-i' (3, 5, 7, 10, 12 and 15). Loops with identical coordinates detected at multiple window sizes were filtered to retain only one instance. Loop start and end coordinates were intersected with the locations of all annotated genes using BEDTools V2.31.0 intersect '-wao'. For profiling genomic features, chromatin compartmentalization, open chromatin and epigenetic the same steps were followed, using the start and end loop coordinates for intersecting with bigWig tracks and generating aggregated signal visualizations with deepTools v3.3.2.

For further analysis, loop length and loop types were examined in loop_descriptive_stats.R, gene co-expression was analyzed in chromatin_structure_and_coexp.R, gene expression levels were analyzed in chromatin_structure_and_expression_specificity.R, where genes with zero or near-zero expression were filtered out specifically for expression specificity analysis, following the same approach used for TADs.

Epigenetic clustering and analysis of open chromatin and epigenetic marks

The 25 Kbp tiles created to describe levels of open chromatin (ATAC-seq), histone modifications (ChIP-seq), and 5mC (CG, CHG, and CHH) modifications were used in downstream analysis for determining typical epigenetic regions. Regions lacking 5mC methylation calls were excluded. K-means clustering was applied to evaluate the optimal number of clusters, testing values of k from 2–10 and identifying the elbow point of the total within-cluster sum of squares. This indicated 4 as the optimal number of clusters, and clustering assuming 4 means was performed. 25 Kbp tiles intersected with gene TSS and region calls for bases annotated as gene, intergenic TE pseudogene and sub-ab signal were then integrated with the locations of categorised epigenetic clusters and analysed, see epigenetic_tiles.R.

Per chromosome visualizations of chromatin compartments

Visualisation of chromatin compartments, TADs, loops, and genomic features was performed using the R package plotgardener V1.12.0. plotgardener was also used to generate per-chromosome views of chromatin compartmentalization. See plotgardener_main_fig.R and plotgardener_all_chromosomes.R.

RNAseq processing of seasonal needle data

To evaluate the raw sequence data, FastQC version 0.11.9 was used (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Residual ribosomal RNA (rRNA) contamination was filtered out using SortMeRNA (v4.3.4; Kopylova et al., 2012) with the settings (--paired_in --fastx -a 20 --sam --num_alignments 1). Adapter sequences were removed, and quality trimming was carried out using Trimmomatic (v0.39) with settings (TruSeq3-PE.fa:2:30:10:1 SLIDINGWINDOW:5:20 MINLEN:50). FastQC was rerun after both filtering steps for quality control. The reads were quantified and aligned to the *P. abies* genome assembly using Salmon (v1.9.0; Patro et al., 2017), applying non-default parameters (--dumpEq --numGibbsSamples 100 --gcBias --posBias --seqBias --validateMappings).

Co-expression calculation by Pearson

A quality assessment was performed according to Bag et al., (2021). Using DESeq2 (v1.44.0; Love et al., 2014) a DESeqDataSet (dds) object was made taking the seasons into account for the design (design = ~Season). A variance stabilizing transformation (VST) matrix was calculated being aware of the dds model. The transformed data were then filtered to remove genes without any expression. This filtered VST was used to calculate the Pearson correlation between genes for co-expression analysis using the Seidr toolset v0.14.2 (Schiffthaler *et al.*, 2023).

A comparative wood expression atlas between *P. abies* and *P. sylvestris*

Sampling for wood RNA-Seq

Wood samples from *Pinus sylvestris* and *Picea abies* were obtained at clonal field trials at Skogforsk, Sävar in northern Sweden (63º 89564800461675’ N, 20º 549005496922252’ E). *P. sylvestris* samples from genotype 29 were taken on July 8th, 2020, between 13:30-14:30. *P. abies* samples from genotype K27-2125 were taken on July 10th, between 10:30-12:00. Rectangular wood blocks were collected at breast height from each tree, immediately placed in dry ice and subsequently stored at -80 ºC. No obvious abiotic or biotic stresses were observed during the process. Prism-shaped pieces, containing the main layers of wood (cortex, phloem, cambium, early/developing xylem and mature xylem), were cut from the wood blocks using an electric saw. Those pieces with clearly discernible developmental layers were placed into a cryo-microtome and multiple longitudinal sections (15 µm thick) were cut from each layer (Uggla et al., 1996). During the cutting process, cross sections from the pieces were regularly prepared and examined under a light microscope to characterise the distinct layer each section was cut from. In total, between 120 and 150 sections were prepared from

each prism-shaped piece, covering the entire current year growth and the four developmental layers: phloem, cambium, early/ developing xylem and mature xylem. Each section from those layers was stored in a separate 1.5 ml Eppendorf tube and stored at -80 ºC. This was repeated three times for a total of three replicates per tree. Sections were pooled as detailed in Supplementary Data 1 and an overview of the sampling approach is provided in Fig S34.

RNA extraction and RNA library preparation

Total RNA and microRNA were extracted separately from each section pool following the protocol detailed in the above annotation section. For each replicate, sections were pooled and thawed in prewarmed extraction buffer and homogenised using glass beads. The rest of the steps were as detailed above.

The Universal Plus mRNA-Seq with NuQuant protocol (Tecan, Männedorf, Switzerland) was used to prepare cDNA libraries from the extracted RNA of each of the section pools. These libraries would later be used separately for mRNA sequencing of each of the section pools. Using the protocol, first, polyadenylated RNA was enriched from the RNA samples using oligo(dT) magnetic beads, selectively capturing mRNA. The purified mRNA was then treated with a fragmentation buffer to achieve 300 base-pair fragments. Reverse transcription was then performed using random primers, followed by second-strand cDNA synthesis to generate double-stranded cDNA. The resulting cDNA was subjected to end repair to generate blunt ends. Next, unique dual-index adapters with barcodes were ligated to the cDNA fragments, to enable sample multiplexing. Following adapter ligation, the cDNA libraries were amplified by PCR and purified using Agencourt AMPure XP beads (Beckman Coulter, Brea, California, USA). Final concentration was measured using Qubit RNA High sensitivity and quality was studied with the Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, California, USA) using Pico chips.

Gene expression quantification and comparative co-expression analysis

RNA-Seq data was processed and quantified as detailed above. For each tree species: Read counts were subjected to a moving average smoothing (within each replicate tree) before normalisation using the Variance Stabilizing Transformation (VST) as implemented in DESeq2. Samples were inspected using hierarchical clustering and PCA, and outliers were removed iteratively (i.e. normalisation was redone, removed samples are marked in Supplementary Data 1: 15 samples in *P. abies*). Genes with VST values > 3 in two samples in three replicate trees (two replicate trees for *P. abies* due to the large number of samples removed from one tree) were defined as expressed. Hierarchical clustering was conducted using the R-function *hclust* with parameter *method="ward.D2"* and Person correlation. PCA was done using the R-function *prcomp*.

Co-expression networks were generated for each species using Pearson's correlation and mutual rank normalisation. The networks were constrained to only include the 3% most highly co-expression links. Co-expression neighbourhoods of orthologs were compared using the ComPlEx method[41].

Whole genome re-sequencing of 1000 *P. abies* trees

Material for whole-genome re-sequencing

A total of 1023 new samples of Norway spruce were re-sequenced for this study, and, combined with 33 previously published samples (PRJEB34927) and the reference individual, for a total of 1057 spruce samples used for variant detection. In addition, four different outgroup species were added: *Picea engelmannii*, *Picea glauca*, *Picea sitchensis* and two samples of *Picea obovata* (Supplementary Data 2). The *P. abies* samples comprised one individual from each of the 279 range-wide provenances selected from the IUFRO 1964/68 provenance test site in Lisjö, central Sweden, and one individual from each of the 749 open-pollinated families selected from a large northern progeny trial located in Haggenås, Sweden. Additionally, 10 mutant genotypes were selected from the clonal archive in Sävar, Sweden.

DNA isolation and sequencing

Total genomic DNA was extracted from vegetative buds or needles from 1023 *P. abies* samples using the E.Z.N.A. SP Plant DNA Kit (Omega Bio-Tek, Norcross, GA, USA) following the manufacturer's instructions. Extracted DNA concentration was measured with NanoDrop 2000 (Thermo Fisher Scientific, Waltham, MA, USA). DNA library preparation was done using the Illumina TruSeq DNA PCR-free library preparation kit (Illumina, USA), and sequencing performed on the NovaSeq 6000 Sequencing System using the NovaSeq 6000 v1.0 and v1.5 reagents (Illumina, USA) at SciLifeLab (Stockholm, Sweden).

Mapping and SNP calling.

For downstream processing reasons, we partitioned the reference sequence into 100 Mbp chunks, such that the 12 autosomes were split into 165 sequences. Samples were sequenced in multiple batches and mapped independently. We used bwa v0.7.17-r1188 (Li 2013) to map Illumina paired end reads to the chunked reference sequence, marking shorter split hits as secondary and discarded mappings with mapping quality <= 30 with samtools v1.12 (Danecek et al, 2021) to generate unsorted BAM files. The BAM files were sorted with samtools v1.12 and indels were realigned with GATK IndelRealigner VN:3.7-0-gcfedb67 (DePristo et al, 2011) using default options. For each sample, we merged batch BAM files with samtools v1.12 to produce final BAM files for variant calling. Finally, variants were called with bcftools v1.14 (Danecek et al, 2021) using options --multiallelic-caller --variants-only on autosomes to generate 165 separate output VCF (Variant Call Format) files.

Coverage analyses and accessibility masks

We generated coverage profiles of individual BAM files with Mosdepth v0.3.3 (Pedersen, B. S., & Quinlan, A. R., 2018) with the command options --d4 -–fast-mode -–mapq 10. The output was stored in the Dense Depth Data Dump (D4) (Hou et al, 2021) format and further processed with d4tools to generate histograms and coverage summaries. Coverage files were aggregated in D4 format in two ways to provide data for generation of accessibility masks. First, we summed individual coverage files to generate a coverage depth profile for all

samples. Second, we generated an aggregate measure for missing data by counting the number of samples with a coverage >= 3X for each position. We then generated histograms with d4tools stat –hist and plotted the output files to determine thresholds for masking, following criteria suggested by Lou et al 2023.

A base was masked as inaccessible if 1) the number of missing data points was >= 50% and 2) the aggregate sequencing depth was less than 0.7X * mean coverage or larger than the median coverage + 2 standard deviations, or by manual inspection. The second criterium serves to exclude sites that are either difficult to sequence or represent unresolved repeats that therefore suffer excessive mapping. Masks generated from these criteria were combined to produce a final accessibility mask. This was done separately for each sample set in the study. All steps were integrated into a Snakemake (Mölder et al, 2021) pipeline (github.com/tallgran/conifer-percyfal-ms/src/conifer/workflow/snakemake/commands/smk-aggregate-coverage.smk).

Nucleotide diversity

We calculated nucleotide diversity (π, the average number of pairwise nucleotide differences per site between a pair of randomly sampled chromosomes in the population) and Watterson's theta (θ_W, the number of segregating sites) for all segregating sites over all VCF files. In all diversity calculations we discarded variants with quality <= 20. Nucleotide diversity was calculated with vcftools 0.1.16 (Danecek et al, 2011), option –site-pi. Watterson's theta was calculated as the number of segregating sites, divided by the harmonic number of the sample size. We used a corrected sample size for the harmonic number as an average of the number of present calls at a locus. We used the accessibility masks (see above) to filter segregating sites based on sequence coverage and call rate and similarly corrected the total sequence length to include only accessible bases for the calculation of per-site statistics. The accessible bases were further partitioned into genomic features, such as CDS, UTR, intron, and intergenic, to obtain diversity statistics for different sequence contexts (Table S7). The proportion of accessible bases was in the range 40-80%, depending on genomic feature and sample set, with pseudogenes corresponding to the lower and CDS the upper value. Aggregated diversity statistics were calculated with a custom script (github.com/tallgran/conifer-percyfal-ms/ src/conifer/tools/summarize_diversity.py).
We used SnpEff v5.2 (Cingolani et al., 2012) to tally the number of synonymous and non-synonymous segregating sites for the longest CDS regions per gene model. The CDS sequences were self-compared with KaKs_Calculator v3.0 (Zhang 2022) to estimate the total number of synonymous and non-synonymous sites across gene bodies. We corrected for the proportion of accessible sites as above and used the corrected sample size for an independent calculation of Watterson's theta.

We furthermore investigated diversity proximal to transcription start/end sites by generating 1 Kbp windows 200 Kbp up- and down-stream of annotated genes. We selected genes separated by more than 400 Kbp (n=20,061) to avoid overlapping windows and plotted the diversity measures as a function of distance from TSS with 95% confidence intervals estimated from 200 bootstrap replicates.

Structural variation analysis

To assess structural variability between geographically isolated populations, we performed a coverage-based analysis on northern (n=26) and southern (n=26) individuals. The analysis was performed in a Jupyter notebook (github.com/tallgran/conifer-percyfal-ms/notebooks/presabs.ipynb) using machine learning algorithms from Scikit-learn (Pedregosa et al, 2011). Individual coverages per sample were aggregated to feature-based coverages; here, features were based on the annotation and could be anything from genes, exons, CDS or pseudogenes. We applied a coverage threshold (>= 3X) to each base, thereby scoring a base as present (=1) or absent (=0), and calculated the sum over each individual feature (e.g., in the case of genes, we retrieved an aggregate score for each individual gene). Given that features have different lengths, the aggregate scores are higher for longer features. We therefore applied different normalisations to the data to mitigate the length bias. First, we normalised a feature score by its length to produce a score in the range 0-1. Second, we applied a discretised normalisation, in which a feature was score as present if it had more than 20% present bases (normalised score > 0.2).

We performed exploratory data analyses with PCA which indicated that samples clustered according to sample set; Fig S40c shows an example PCA based on the structural variation segregating patterns in genes and Fig S40d for pseudogenes. To identify genes driving the separation, we applied PLS-DA to the discretized data. We applied K-fold cross validation to select a 3-component model for fitting the data. The top 1% loadings were selected and plotted as a hierarchically clustered heat map (Fig 3d) with individuals grouped by population.

Population structure

To look at population structure in *Picea abies* the Rangewide dataset was used. We estimated pairwise relatedness between all genomes directly from the genotype calls using NgsRelate v2.0 on default parameters (Hanghøj et al., 2019). 34 samples were excluded due to high relatedness to another sample resulting in a total of 263 samples. Additionally, 28 samples from the Northern breeding population were added to adjust for the sparse number of samples in the north of Sweden in the Range wide dataset. The 28 samples from the Northern breeding population were chosen by assigning samples to a 1° by 1° (Longitude, Latitude) grid and randomly select 1 sample with a genome wide mean coverage > 3x within each cell. The two *P. obovata* samples were also added as outgroup information. This resulted in a dataset of 291 *P. abies* samples and two *P. obovata* samples to be used in the population structure analysis.

The raw genome wide SNP calls were filtered to only include bi-allelic sites with average mapping quality (MQ) >= 55, alternative allele frequency >= 0.01, minimum of 3 reads covering a SNP in at least 50% of the samples and a coverage across all samples between 145 and 1070. To remove highly correlated SNPs, LD pruning was preformed using the locate_unlinked function (scikit-allele 1.3.13) in sliding windows of 500 SNPs with a 200 SNP and removing SNPs with an LD (r^2) value >= 0.1. After filtering and LD pruning a total of ~12.7 million SNPs were retained, out of which 1.5 million were randomly selected genome wide to base the population structure analysis on.

The filtered VCF file with 1.5 million SNPs were converted into a Zarr file using the command vcf_to_zarr (python package scikit-allel 1.3.13 (Miles et al., 2024)). The genotype matrix was converted into 0, 1 and 2 representing homozygote reference, heterozygote and homozygote alternative allele respectively using the to_alt_allel function (scikit-allel 1.3.13), the default to fill missing values with 0 (homozygous reference) was used. The genotype matrix was further scaled to have the same mean and variance for each sample using StandardScaler().fit_transform() (scikit-learn v.1.5.2 (Pedregosa et al., 2011)). Two methods were then applied to look at genetic clustering, PCA (scikit-learn v.1.5.2) and UMAP (umap-learn v.0.5.7 (McInnes et al., 2020, Dalmia et al., 2021)).

Genetic clustering with PCA was performed in a Jupyter notebook (github.com/tallgran/conifer-mceriksson/population_structure/notebooks/PCA_pabies-pobovata_genome-wide.ipynb). The 293 samples were divided into a 'core' set, samples with average genome wide coverage >= 5x, and a 'low coverage', samples with average genome wide coverage < 5x. For clustering the samples, we wanted to capture 80% of the variance within that data. Therefore, we first constructed a PCA using the 'core' sample set without specifying the number of PCs. Then calculated the cumulative expressed variance and extracted the PC where we reached 80%, PC 91. After finding the number of PCs that explain 80% of the variance, we constructed a PCA again using the 'core' sample set and specifying the number of PCs to 91. We then projected the 'low coverage' sample set into the PCA based on the 'core' sample set. To find the optimal number of clusters the PCA matrix (n_PC (91) x n_samples (293)) was first clustered using K-means clustering then two methods to determine the number of clusters were applied. First, the Elbow method, which is based on within-cluster sum of squares score calculated for each cluster size (k). The within-cluster sum of squares (WCSS) score is calculated as the sum of squared distances of samples to their closest cluster centre. Second, Silhouette clustering, which calculates a mean distance score per cluster based on a mean intra-cluster distance and a mean nearest-cluster distance for each sample. Higher scores indicate better fit of the clustering. Both methods were run for each k up to 10 and iterated 10000 times to get the variance within each estimated k. The WCSS scores obtained from the elbow method showed a weak elbow at k2, similarly, the k2 was the k with the best silhouette score, therefore, two clusters were chosen when running K-means clustering on the PCA matrix.

Examination of genetic clustering using UMAP was performed in a Jupyter notebook (github.com/tallgran/conifer-mceriksson/population_structure/notebooks/umap_clustering.ipynb) where the transposed genotype matrix was clustered using the UMAP function (umap-learn v.0.5.7). To further determine number of clusters HDBSCAN (hdbscan v.0.8.39 (Moulavi et al., 2014, McInnes et al., 2017)) was used to iterate over a range of minimum samples (1 to 50) and minimum cluster (2 to 49). For each iteration, the relative validity or DBCV score was extracted and the combination with the highest score was then chosen as settings when running HDBSCAN for determining the number of clusters, resulting in two separate groups. By comparing the genotype clustering by PCA and UMAP most samples were clustered similarly by PCA K-means and UMAP HDBscan. The clusters were labelled north and south based on the geographical origin of most samples within each cluster. Only a few samples deviated between the two methods and were labelled inconclusive. The geographical location of the deviating samples was in Lativa, Lithuania and Belarus.

Two additional PCAs were conducted following the same protocol, the first using the same set of 1.5 million genome wide SNPs but excluding the outgroups species *P. obovata.* The second, also excluding *P. obovata* but only using exonic SNPs (n=90398) extracted from the filtered and LD pruned set of ~12.7 million SNPs. The two PCAs (genome wide and exonic SNPs) showed very similar patterns and to see if the pattern in the genome wide PCA were driven by genic SNPs we extracted the top 1% outlier loadings (two-sided) for PC1 and PC2 for the genome wide PCA. Each SNP was classified as genic or intergenic and the total amount and genic SNPs were counted in the full set, 1% outliers for PC1 and PC2. To test for an enrichment in genic SNPs in the top loadings a Fishers exact test can conducted for each PC, resulting in an underrepresentation of genic SNPs (PC1: odd ratio=1.473, p-value = 1.63e-40; PC2: odd ratio= 1.47, p-value = 4.75e-40). However, doing the same but only considering exonic SNPs there was a significant enrichment with Fishers exact test (PC1: odd ratio= 0.666, p-value = 1.194e-06; PC2: odd ratio= 0.617 p-value = 3.28e-09), although the SNP counts were rather small (PC1: 165 exonic SNPs out of 15002; PC2: 178 exonic SNPs out of 15002) (code: github.com/tallgran/conifer-mceriksson/population_structure/notebooks/PCAs_genome-wide_vs_exonic_SNPs.ipynb).

As a complement to the genotype clustering, we also performed an admixture analysis using ADMIXTURE v.1.3.0 (Alexander et al., 2009). First looking at the cross-validation error to find the optimal K by running admixture with the --cv flag for 1000 iterations for each k between 1 and 10. Then running admixture for the best K with 1000 bootstraps.
For more details and the code see: github.com/tallgran/conifer-mceriksson/population_structure/population_structure_pipeline.md.

Ancestral state

To be able to polarise the SNP data, the software est-sfs v2.04 (Keightley & Jackson, 2018) was used to predict the ancestral state. It was run using 296 *P. abies* samples with a genome wide mean coverage >= 6.5 and two outgroup species, *P. obovata* (sample: P.obovata_P24355) and *P. engelmannii*. Criteria that were applied to filter SNPs were, one alternate allele, minimum mean mapping quality >= 55, not indel, each sample call to be covered by at least 3 reads in 50% of the samples. To run est-sfs each chromosome was divided in chunks containing 1 Millon SNP using custom scrip (conifer-mceriksson/selection/scripts/est-sfs_make_input-file.py) and each chunk was run using the kimura model, i.e., the config-kimura file that accompanies the software was used and was kept as provided except for the n_outgroup parameter that was changed to two. The output file containing the ancestral allele probabilities was further processed using a custom python script (github.com/tallgran/conifer-mceriksson/selection/scripts/est-sfs_convert_probability_output.py) to extract the most likely ancestral state at each site based on the probabilities for the major allele and/or each base (A, C, G or T). The assignment of an ancestral state was done based on three different criteria.

1. If the base probability (for A, C, G or T) >= 0.95.
2. If the major allele probability >= 0.95 and the base probability (for the same base as the major allele) >= 0.85.
3. If the major allele probability == 1 and the base probabilities (A, C, G, T) was missing.

Sites that do not pass any of the criteria were left empty. For more details and the code see: github.com/tallgran/conifer-mceriksson/selection/ancestral_state_and_recombination_map_pipelines.md

Recombination map

For the recombination map, we used samples with sufficient coverage that were within the same/similar population. Therefore, we selected 18 samples from two locations in Västerbotten, Sweden with an average genome wide coverage around 12.4 (min: 10.7, max: 13.3, std: 0.74; Table S10). Before estimating the recombination rates, the genotype data were phased to know which variant of a heterozygous locus belongs to which chromosome, using SHAPEIT v5.1 (Hofmeister et al. 2023). We used an additional 112 samples (in total 130 samples), all with an average genome wide coverage >= 10, to have better power for phasing. Only SNPs with a minor allele frequency >= 0.05 were phased. After phasing, the 18 samples for the recombination map estimation were extracted as well as positions with an ancestral state prediction, since the ancestral probabilities for each base (A, C, G, T) were used as input for the recombination map estimation. The recombination map was made using LDhelmet v.1.10 (Chan et al., 2012). For more details and the code see: github.com/tallgran/conifer-mceriksson/selection/ancestral_state_and_recombination_map_pipelines.md.

PCA based test to identify signals of local adaptation

To identify genes with signals of local adaptation across all 291 range wide *P. abies* samples we used pcangsd (Meisner *et al.*, 2021) on SNPs filtered to include only bi-allelic sites and sites of expected coverage (above 1/3 and below 3 times the average genome coverage across all individuals). pcangsd was run using:

```
pcangsd -p filtered.contig.vcf --pcadapt --out contig --threads 8 --minMaf 0.01
```

Then, we calculated averages for the selection score (PC1, pcangsd output) in 10 Kbp windows and extracted all genes containing a window with an average selection score (>=3.23) corresponding to the top 1% of the genome wide distribution, resulting in 679 genes (Supplementary Data 4).

Haplotype based tests to identify signals of positive selection

Given that there was little evidence of population structure but a clear latitudinal cline, selection scans focusing on the two extremes of the cline (north/south) were conducted as a compliment to the PCadapt selection scan. To represent the two extremes, 26 samples from the north and 26 from the south were chosen out of the 291 samples used for population structure based on three criteria, 1) a min coverage of 5x, 2) concordant cluster assignment by PCA and UMAP (Fig S36a and c), 3) a min proportion of alleles to either of the two ancestry pools of 0.98 in the admixture analysis (Fig S36e). Filters applied to the VCF files were to include only biallelic SNPs (i.e. no indels), minimum mean mapping quality of 55, minor allele frequency of 0.1 and a coverage of 3 reads from at least 50% of individuals. After filtering, the

SNPs were phased using SHAPEIT v5.1 (Hofmeister et al. 2023) and the predicted ancestral allele added to the INFO field as AA using annotate in BCFtools v1.17 (Danecek et al., 2021). As a last step the chunked chromosome VCFs were concatenated into full chromosome VCFs using concat in BCFtools v1.17 (Danecek et al., 2021). Chunked chromosome names and positions were changed using a custom python script (github.com/tallgran/conifer-mceriksson/selection/scripts/change_subchrm-pos_2_chrm-pos.py).

To find signals of positive selection differing between the two extremes we first ran a cross-population extended haplotype homozygosity test (XP-EHH) using xpehh from scikit-allel v1.3.5 (Miles et al., 2024) to contrast the north and south sample sets. Since the xpehh function returns unstandardized XP-EHH scores, the output values were standardized following Sabeti et al. 2007 by subtracting the genome wide mean and then dividing by the genome wide standard deviation. In addition to contrasting north and south we also preformed tests for positive selection by computing integrated haplotype score (iHS) for the north and south sample set separately using ihs from scikit-allel v1.3.5 (Miles et al., 2024). Similarly to xpehh, ihs also returns unstandardized scores. The iHS scores were standardized following Voight et al. 2006 by grouping SNPs in allele frequency windows and each score was subtracted by the window mean and then divided by the standard deviation for the window. For both iHS and XP-EHH the genotypes were polarized towards the predicted ancestral state, in cases without a predicted ancestral state the reference was kept. The estimates from the recombination map were used for genetic map positions with the map_pos option when running both ihs and xpehh. As potential signals of positive selection values above and belove the genome wide 1% outlier thresholds (iHS: low = -2.97 and high = 2.87, XP-EHH: low = -3.42 and high = 3.51) was considered, for iHS the thresholds were calculated for both sample sets combined. For more details and the code see: github.com/tallgran/conifer-mceriksson/selection/selection_scan_pipeline.md.

To identify genes under positive selection we first looked for genes with a contrasting signal between the north and south sample set. The top 1% of SNPs from the genome wide distribution in the cross population (XP-EHH) method was used as a baseline. As additional information all outliers XP-EHH SNPs were cross-referenced between the sample set specific tests (iHS). To be classified as a supporting iHS SNP the scores between north and south had to be either, 1) a negative 1% outlier in test A and a score > 0 or missing in test B, or 2) a positive 1% outlier in test A and a score < 0 or missing in test B. For each gene the number of classified outliers was counted and genes with >= 5 outliers was considered as potential genes showing contrasting signs of positive selection between the north and south sample sets, resulted in 4202 genes. We then looked for genes with a similar selection signal between the north and south samples sets by extracting the top 1% of SNPs from the genome wide distribution in both the iHS north and iHS south (that was not classified as an outlier in XP-EHH). To be classified as an outlier SNP with similar signal between north and south the SNP had to be either, 1) a negative 1% outlier in test A and a score < 0 or missing in test B, or 2) a

positive 1% outlier in test A and a score > 0 or missing in test B. For each gene the number of classified outliers was counted and genes with >= 5 outliers was considered as potential genes showing similar signs of positive selection between the north and south sample sets, resulted in 7769 genes. To get genes that show signs of positive selection the contrasting gene set (4202) and similar (7769) was combined and resulted in a list of 8995 genes (FigShare). The code to classify SNPs and extract genes under positive can be seen here: github.com/tallgran/conifer-mceriksson/selection/notebooks/classify_selected_xpehh-ihs_genes.ipynb. Details of the selection test scores for each gene are available in Supplementary Data 3. All scores (intergenic and genic) calculated using xpehh are available from the Figshare repository (file: selection_xpehh-scores.tab.gz).

We extracted the genes with a signal of local adaptation and a signal of positive selection by intersecting the list of 679 genes with the list of 8995 genes resulting in 319 genes (Supplementary Data 4). We then preformed GO enrichment analysis for each of the four gene sets (local adaptation: (n=679), positive selection contrasting north and south (n=4202), positive selection unique or similar north and south (n=7769) and overlap between all tests (n=319) using the R package TopGO v2.58.0 (Alexa & Rahnenfuhrer, 2024) with the "parentchild" method and fishers test for significance. P values were adjusted using “weight01” and GO terms with an adjusted p-value >= 0.05 was taken as significantly enriched (Figshare). All was run in RStudio v2024.9.1.394 (Posit team, 2024) using R v4.4.3 (R Core Team, 2025), for full code and session information: github.com/tallgran/conifer-mceriksson/selection/notebooks/GO-enrichment_top-GO.nb.html.

Inbreeding coefficient and population differentiation

To get an idea about inbreeding and population differentiation despite the lack of population structure, we used the north and south sample sets (also used for the tests for positive selection) to represent the two extremes as a proxy for population. SNPs were first filtered to keep only biallelic SNPs (i.e. no indels), minimum mean mapping quality of 55, minor allele frequency of 0.05. For population differentiation the samples in the north and south sample set were kept in the same VCF and filtered for sites with at least 3x for 50% of samples and a coverage across all samples between 285 and 936. F_{st} was calculated as chromosome wide averages (Table S6) using two different method, Weir & Cookham (Weir & Cockerham, 1984) and Hudson (Hudson et al., 1992, Bhatia et al., 2013) both methods were run using in scikit allel v1.3.13 (Miles et al., 2024) (functions: average_weir_cockerham_fst and average_hudson_fst).

To calculate the inbreeding coefficient (F_{is}), we additionally, to the north and south sample sets, used the high coverage (296 samples) as a representation to cover the whole geographical range. Each sample set was filtered with the intersect of their own to coverage maps that represents a minimum of 3x coverage for 50% of samples for a SNP and a coverage

across all samples between a minimum and maximum value (north: min=166, max=512; south: min=159, max=514; highcov: min=2000, max=4500). Genome wide observed (H_o) and expected heterozygosity (H_e) was calculated based on per SNP estimations using the functions heterozygosity_observed and heterozygosity_expected from scikit-allel v.1.3.13 (Miles et al., 2024) which were summed across all chromosomes and divided by the number of assessable sites. The genome wide inbreeding coefficient (F_{is}) was then calculated using the genome wide observed and expected heterozygosity's (Table S8) using equation: $1-(H_o/H_e)$ (Weir & Cockerham, 1984). For full code: github.com/tallgran/conifer-mceriksson/f_statistics/f-statistics_pipeline.md

2. Supplementary Tables

Table S1 Genome assembly and annotation statistics of Norway spruce (*Picea abies*) and Scots pine (*Pinus sylvestris*).

	Norway spruce (*P. abies*)	**Scots pine (*P. sylvestris*)**
Reference individual, clone ID	Z4006	Y3088
Genome assembly size	17.7 Gbp	20.3 Gbp
N50, contig	12.2 Mbp	7.8 Mbp
N90, contig	3.1 Mbp	1.2 Mbp
GC content	38.4%	38.0%
Repeat content	78.6%	78.2%
BUSCO completeness (genome)	91.0%	90.4%
Chromosomal scaffolds	12	12
Telomere ends	18	14
Protein coding genes	43,410	49,387
Longer than 50 Kbpp	8,492	8,296
Mean exons per gene	4.0	3.7
Functional protein-coding LTR-TE genes	11,100	9,642
Identified pseudogenes	131,696	155,941

Table S2 Manually curated repeat libraries and their corresponding masking of the genome assemblies of Norway spruce (*Picea abies*) and Scots pine (*Pinus sylvestris*). TE, transposable element.

Repeat type	**Number of sequences in the repeat library (*P. abies*)**	**Masked fraction of the assembly (*P. abies*)**	**Number of sequences in the repeat library (*P. sylvestris*)**	**Masked fraction of the assembly (*P. sylvestris*)**
LTR Gypsy	619	41.0%	691	35.7%
LTR Copia	318	17.9%	497	15.6%
LTR, unclassified	272	9.8%	227	6.9%
LINE	104	2.4%	208	1.3%
DNA and MITES	157	3.9%	362	4.4%
TE, unclassified	3	<0.1%	33	0.5%
Unclassified repeats (No Hit Found)	187	2.5%	1125	12.3%
Low complexity/simple repeat	-	1.0%	-	0.9%
Caulimovirid sequences	349	0.6%	829	2.0%
Total		**79.2%**		**80.2%**

Table S3. Solo to complete Long Terminal Repeat (LTR) ratio calculations for various LTR elements.

Element(s)	Solo LTRs	Complete LTRs	Ratio (C/S)
LTR_11_8060470, LTR_ 11_19174909, LTR_ 11_5047755	36739	297965	3.56
LTR_11_42242	602	25380	20.58
LTR_11_18702631	2496	31582	5.83
LTR_11_2927923	4628	150061	15.71
LTR_11_8985340	2607	89733	16.71
LTR_11_61097641	1590	34519	10.36
LTR_11_7739391	6822	58512	3.79
LTR_11_6081655	2159	85289	19.25
LTR_11_30956144	2122	88351	20.32
Total	59765	861392	6.71

Table S4. Dispersed duplications of genes and pseudogenes in Norway spruce and Scots pine.

	Norway spruce (*P. abies*)	**Scots pine (*P. sylvestris*)**
Orthogroups with recent in-paralogs, all protein coding genes	16,509	18,006
Dispersed duplications (>10 Mbp apart)	3,622	4,879
Exon/intron structure (partially) retained	1,236	1,408
Exon/intron structure not retained	715	801
Undefined (single-exon genes)	1,641	2,651
Pseudogenes, all	131,696	155,941
Dispersed duplications (>10 Mbp apart)	122,833	146,999
Exon/intron structure (partially) retained	25,608	27,591
Exon/intron structure not retained	14,130	17,400
Undefined (no exon/intron junction duplicated)	83,095	102,008
Pseudogenes with detectable promoter region, all	19,693	21,981

Table S5 Number of called SNPs and number of samples for each *Picea abies* sample set included in the different analyses.

Sample set	**N samples**	**Raw SNP count**
All, *P. abies*	1,056	3,461,153,343
High coverage	296	3,024,792,074
Rangewide + *P. obovata*	293	3,073,829,835
North/South pop. extremes	52	2,003,262,421
Northern pop. extremes	26	1,448,994,476
Southern pop. extremes	26	1,439,171,520

Table S6 Average F_{st} estimates per chromosome between the north and south sample sets, representing the two genetic extremes.

Chromosome	Method	F_{st}	Standard error
PA_chr01	Weir & Cockerhams	0.07409	0.00108
PA_chr01	Hudson	0.07503	0.00107
PA_chr02	Weir & Cockerhams	0.08275	0.00122
PA_chr02	Hudson	0.08369	0.00122
PA_chr03	Weir & Cockerhams	0.10820	0.00192
PA_chr03	Hudson	0.10883	0.00191
PA_chr04	Weir & Cockerhams	0.07827	0.00119
PA_chr04	Hudson	0.07902	0.00119
PA_chr05	Weir & Cockerhams	0.07007	0.00118
PA_chr05	Hudson	0.07087	0.00117
PA_chr06	Weir & Cockerhams	0.08773	0.00170
PA_chr06	Hudson	0.08844	0.00169
PA_chr08	Weir & Cockerhams	0.09420	0.00148
PA_chr08	Hudson	0.09501	0.00148
PA_chr09	Weir & Cockerhams	0.11417	0.00194
PA_chr09	Hudson	0.11474	0.00193
PA_chr10	Weir & Cockerhams	0.11911	0.00183
PA_chr10	Hudson	0.11960	0.00183
PA_chr11	Weir & Cockerhams	0.08025	0.00179
PA_chr11	Hudson	0.08082	0.00179
PA_chr12	Weir & Cockerhams	0.07634	0.00150
PA_chr12	Hudson	0.07719	0.00149

Table S7. Nucleotide diversity across different genomic features in Norway spruce (*Picea abies*).

Feature	Nucleotide diversity (π)
CDS	0.0060
UTR	0.0078
Intron	0.0096
Functional LTR-TE genes	0.0112
Pseudogene	0.0142
Intergenic	0.0158
Genome (all features)	0.0151

Table S8 Genome wide observed heterozygosity (H_o), expected heterozygosity (H_e), and inbreeding coefficient (F_{is}).

Sample set	H_o	H_e	F_{is}
High coverage	0.00928	0.00985	0.0578
Northern pop. extremes	0.00956	0.00970	0.0147
Southern pop. extremes	0.00885	0.00879	-0.0076

Table S9 Overview of the hierarchy of accession IDs at the European Nucleotide Archive (ENA). All datasets are part of the umbrella ID PRJEB88492. There is accession for all for all *Picea abies* (PRJEB88468) and *Pinus sylvestris* (PRJEB88485) datasets. The table provides project accession IDs per analysis type per species, within which there are specific accession IDs for each data type. Existing datasets have an associated DOI (Digital Object Identifier) referring to the original publication. Datasets with no DOI were newly generated for this project.

Projects	**Analysis**	**Members**	**Data type**	**Publication DOI**
Picea abies				
PRJEB69221	genome			
PRJEB88135	epigenetics	PRJEB87632	ONT methylation	
		PRJEB87714	Micro-C	
		PRJEB87765	ATAC-Seq	
		PRJEB87785	ChIP-Seq	
		PRJEB87786	RNA-Seq	
PRJEB88455	transcriptome	PRJEB1795	RNA-Seq	doi.org/10.1038/nature12211
		PRJEB8220	RNA-Seq	10.1093/treephys/tpae139
		PRJEB9578	RNA-Seq	10.1104/pp.17.01590
		PRJEB10305	RNA-Seq	10.1104/pp.19.00743
		PRJEB12752	RNA-Seq	10.1093/treephys/tpx078
		PRJEB13079	RNA-Seq	10.1104/pp.17.00085
		PRJEB15530	RNA-Seq	10.1111/nph.14458
		PRJEB19683	RNA-Seq	10.1371/journal.pone.0219272 10.1007/s00425-019-03160-z
		PRJEB22154	RNA-Seq	10.1371/journal.pone.0192945
		PRJEB26453	RNA-Seq	10.1111/tpj.15530
		PRJEB26399	RNA-Seq	10.1128/mSystems.00884-20
		PRJEB26933	RNA-Seq	10.1093/treephys/tpaa178
		PRJEB26934	RNA-Seq	10.1111/pce.14241
		PRJEB35823	RNA-Seq	10.1073/pnas.2118852119
		PRJEB45942	RNA-Seq	10.1111/nph.18449
		PRJEB71586	RNA-Seq	
		PRJEB72619	RNA-Seq	
		PRJEB88392	Iso-Seq	
PRJEB88137	resequencing	PRJEB1891	genome	10.1038/nature12211
		PRJEB34927	resequencing	10.1093/gbe/evaa005
		PRJEB69221	genome	
		PRJEB85275	resequencing	
		PRJNA83435	genome	10.1111/tpj.12886
		PRJNA304257	genome	10.1111/tpj.15889
		PRJNA504036	genome	10.1111/tpj.15889
Pinus sylvestris				
PRJEB77112	genome			

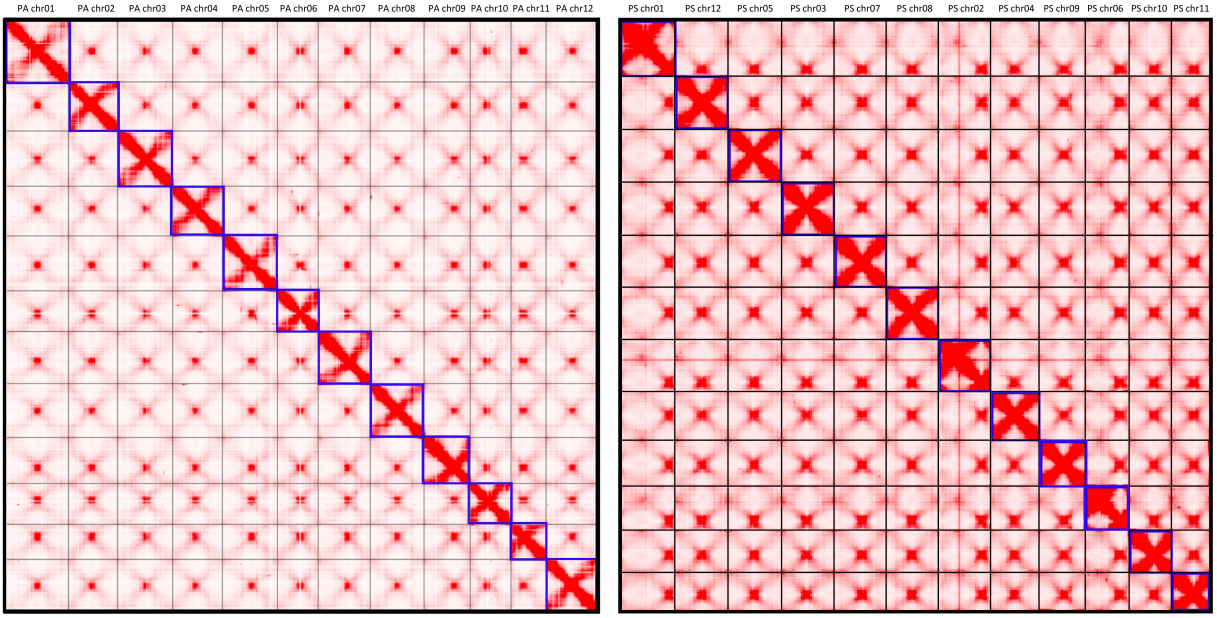
PRJEB88469	transcriptome	PRJEB88462	RNA-Seq	
		PRJEB88465	Iso-Seq	
		PRJNA497687	RNA-Seq	10.1016/j.envpol.2019.04.064
		PRJNA526608	RNA-Seq	10.3390/plants9070913
		PRJNA526785	RNA-Seq	10.3390/plants9070913
		PRJEB19683	RNA-Seq	10.1007/s00425-019-03160-z
		PRJNA384502	RNA-Seq	10.3390/ijms18061199

Table S10 Geographical and genome wide coverage statistics for the sample groups used to compute the recombination map. N = Number of samples.

Location	**N**	**Latitude**	**Longitude**	**Min**	**Max**	**Median**	**Mean**	**Standard deviation**
Långrumpskogen	9	63.7	19.5	10.7	13.3	12.0	12.1	0.89
Marsfjället	9	65.1	15.4	12.4	13.2	12.8	12.8	0.26
Across all	18	-	-	10.7	13.3	12.6	12.4	0.74

3. Supplementary Figures

Figure S1 Hi-C chromatin conformation capture contact maps for Norway spruce (*Picea abies*; left) and Scots pine (*Pinus sylvestris*; right). Note that pine sequences are renamed based on similarity to the spruce chromosomes.



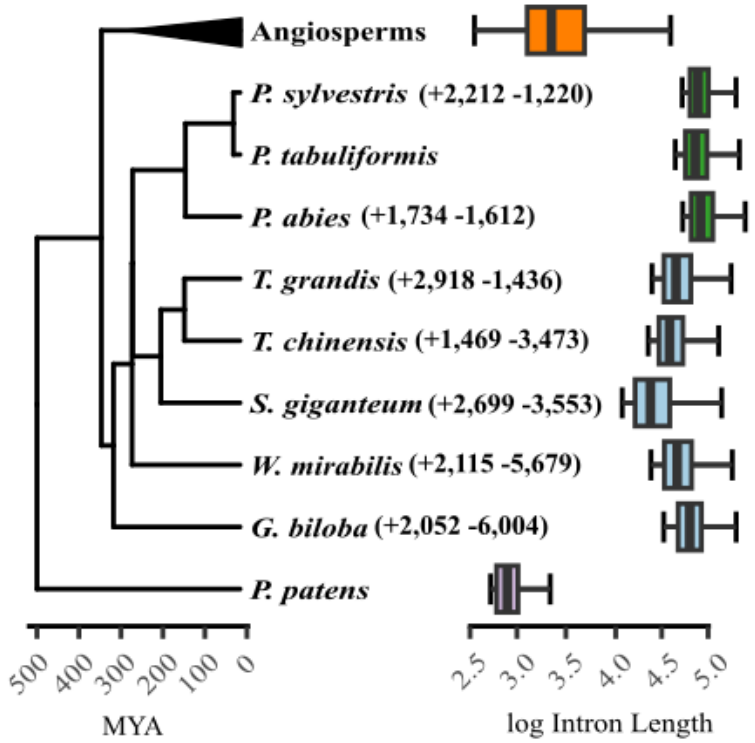


Figure S2 A phylogeny of plant species used in the ortholog identification (left). Angiosperms have been collapsed and the ancestral outgroup used in the analysis is displayed *(Physcomitrella patens).* The total number of orthogroups which gained or lost genes as determined by CAFE5 is shown next to the species name. *P. patens* was excluded due to tree distance and *Pinus tabuliformis* was excluded due to annotation quality. The log10 size of the 10% longest introns are plotted for each species, or the average for all included angiosperms (left), box edges represent quartiles and whiskers 1.5×IQR.

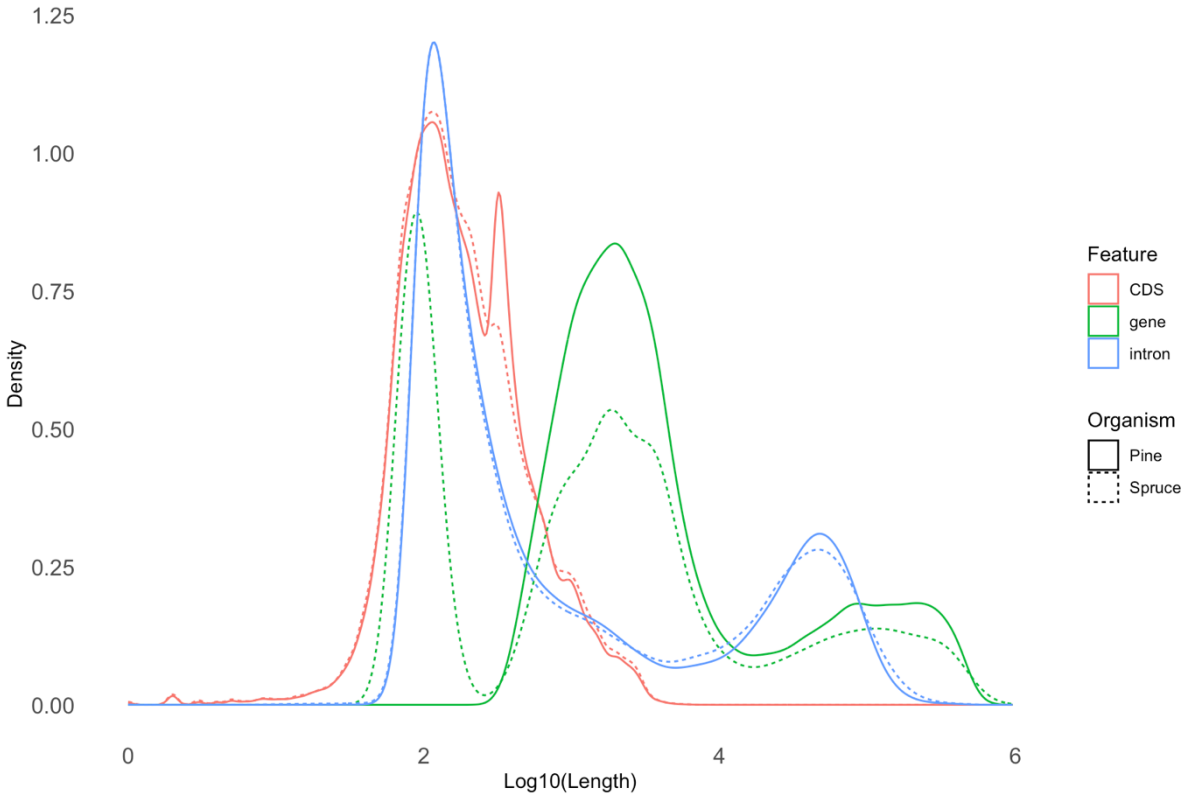


Figure S3 Density distributions of coding sequence (CDS; red), gene length (green) and intron length (blue) in Norway spruce (*Picea abies*; dashed lines) and Scots pine (*Pinus sylvestris*; solid lines).

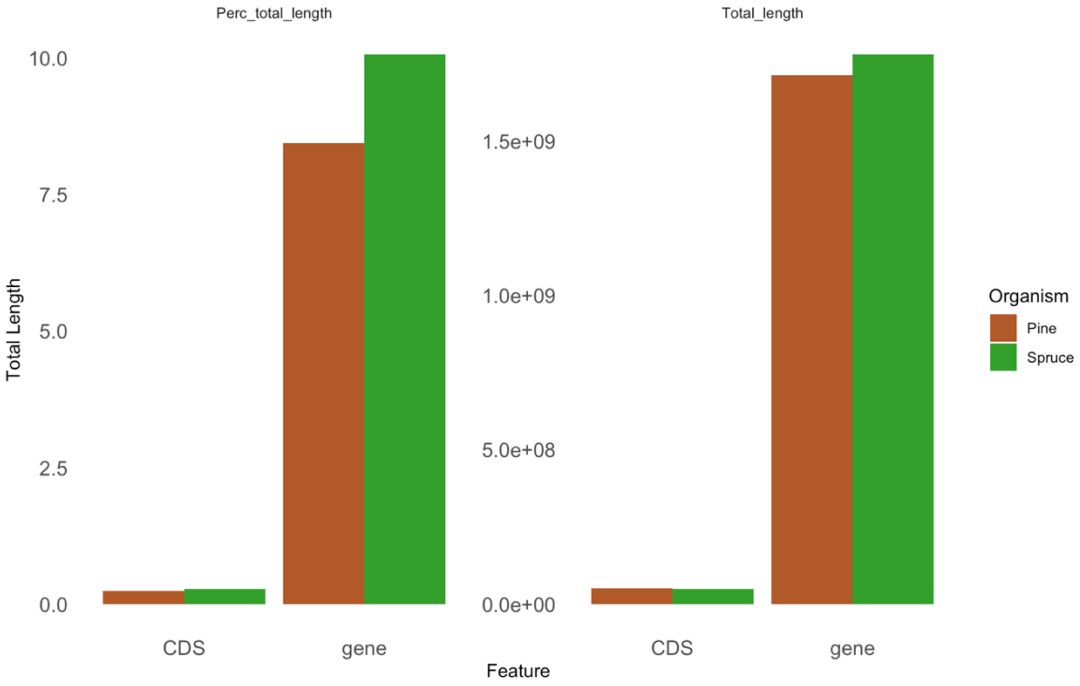
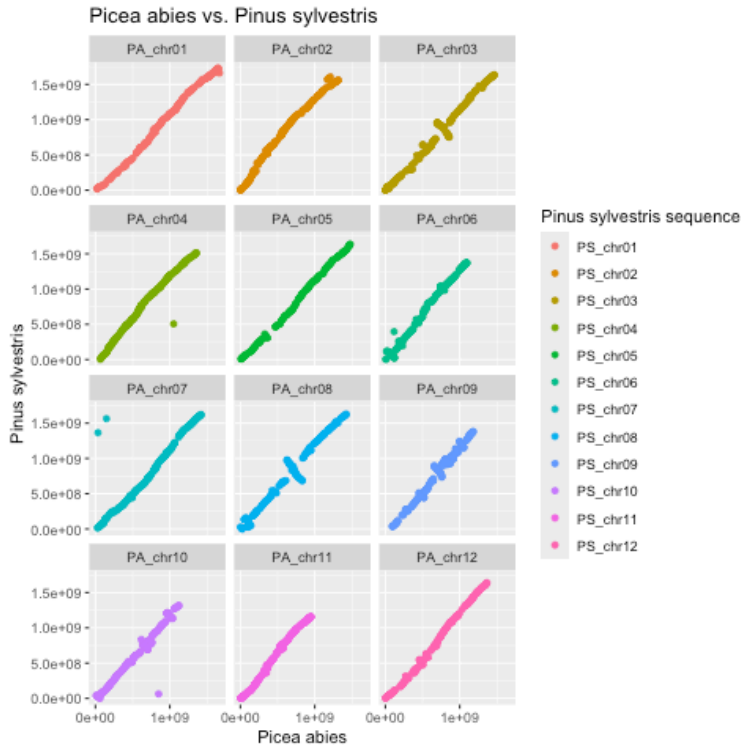
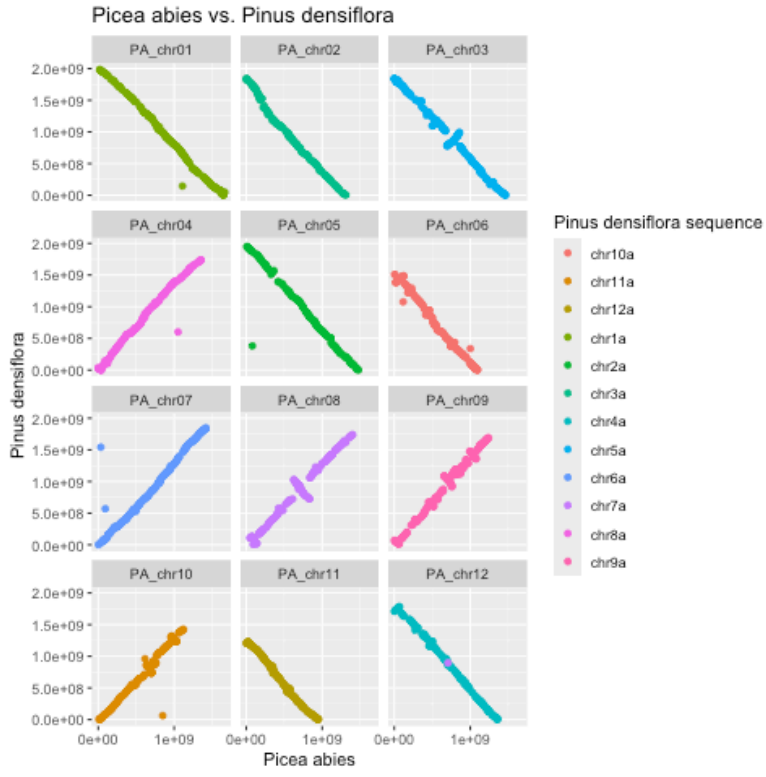


Figure S4 Percentage (left) and length (right) of the genome represented by genes and by coding sequences in Norway spruce (*Picea abies*; green) and Scots pine (*Pinus sylvestris*; brown).

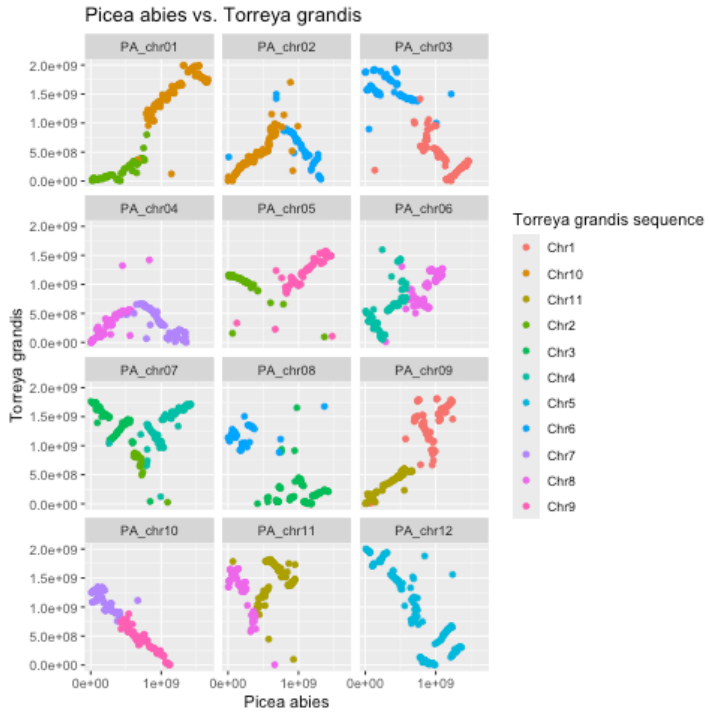
a


Picea abies vs. Pinus sylvestris
PA_chr01
PA_chr02
PA_chr03
PA_chr04
PA_chr05
PA_chr06
PA_chr07
PA_chr08
PA_chr09
PA_chr10
PA_chr11
PA_chr12
1.5e+09
1.0e+09
5.0e+08
0.0e+00
0e+00
1e+09
Pinus sylvestris
Picea abies
Pinus sylvestris sequence
PS_chr01
PS_chr02
PS_chr03
PS_chr04
PS_chr05
PS_chr06
PS_chr07
PS_chr08
PS_chr09
PS_chr10
PS_chr11
PS_chr12

b


Picea abies vs. Pinus densiflora
PA_chr01
PA_chr02
PA_chr03
PA_chr04
PA_chr05
PA_chr06
PA_chr07
PA_chr08
PA_chr09
PA_chr10
PA_chr11
PA_chr12
2.0e+09
1.5e+09
1.0e+09
5.0e+08
0.0e+00
0e+00
1e+09
Pinus densiflora
Picea abies
Pinus densiflora sequence
chr10a
chr11a
chr12a
chr1a
chr2a
chr3a
chr4a
chr5a
chr6a
chr7a
chr8a
chr9a

c



d

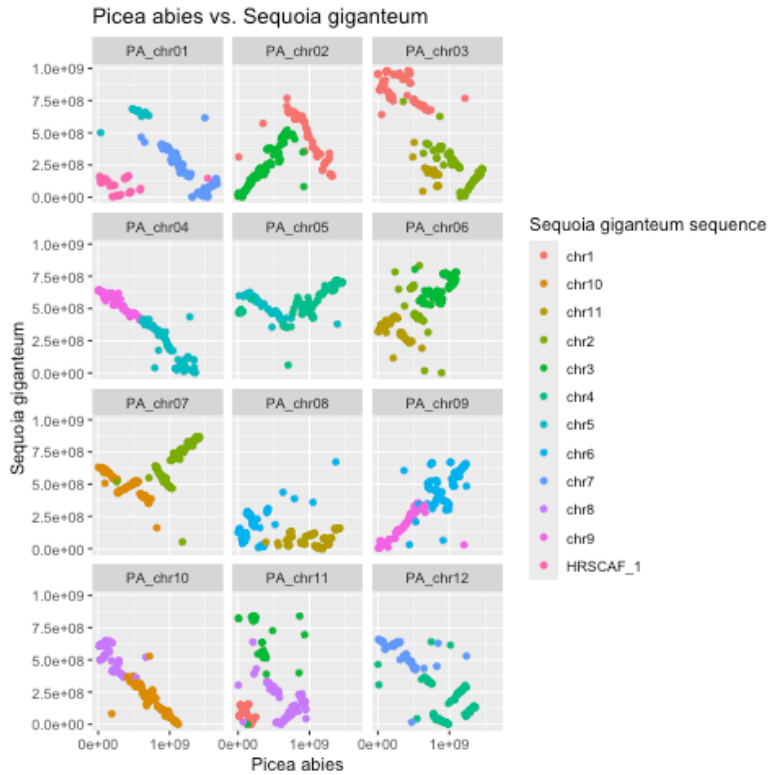


Figure S5: Pairwise species comparisons indicating chromosomal locations of 1-to-1 orthologs between *Picea abies* and A) *Pinus sylvestris*, B) *Pinus densiflora*, C) *Torreya grandis*, and D) *Sequoia giganteum*.

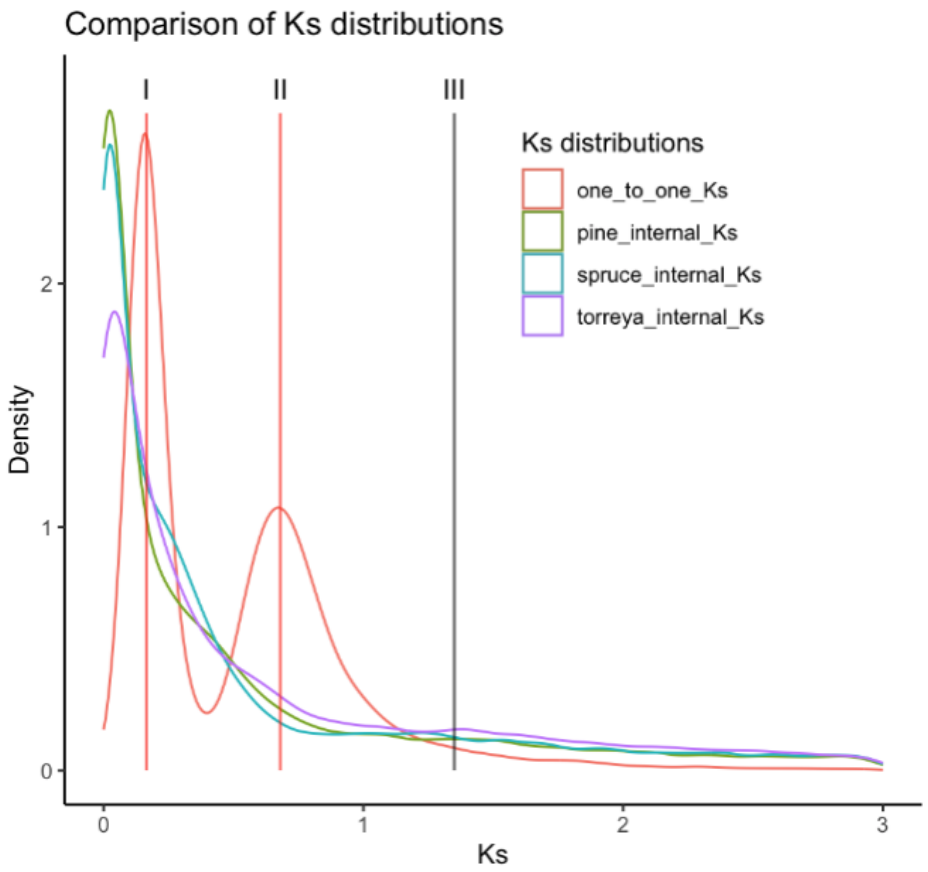


Figure S6: Ks distributions for Norway spruce (*picea abies*), Scots pine (*Pinus sylvestris*) and Torreya (*Torreya grandis*). Peaks indicated in the plot are the following: I = Ks values representing the best reciprocal hits between Norway spruce and Scots pine, i.e. the Norway spruce/Scots pine speciation event, II = values representing the split between Norway spruce/Scots and Torreya and III = individual peaks from internal pairwise Ks calculations for all three species indicating traces of earlier duplications compatible with previously described ancient WGD events (Lou, H. *et al.*, 2023; Li Z. *et al.*, 2015; Stull GW. *et al.*, 2021). The dating of the speciation events denoted by "I" and "II" are 81.0 - 140.0 MYA and 222.0 - 298.0 MYA, respectively (https://timetree.org).

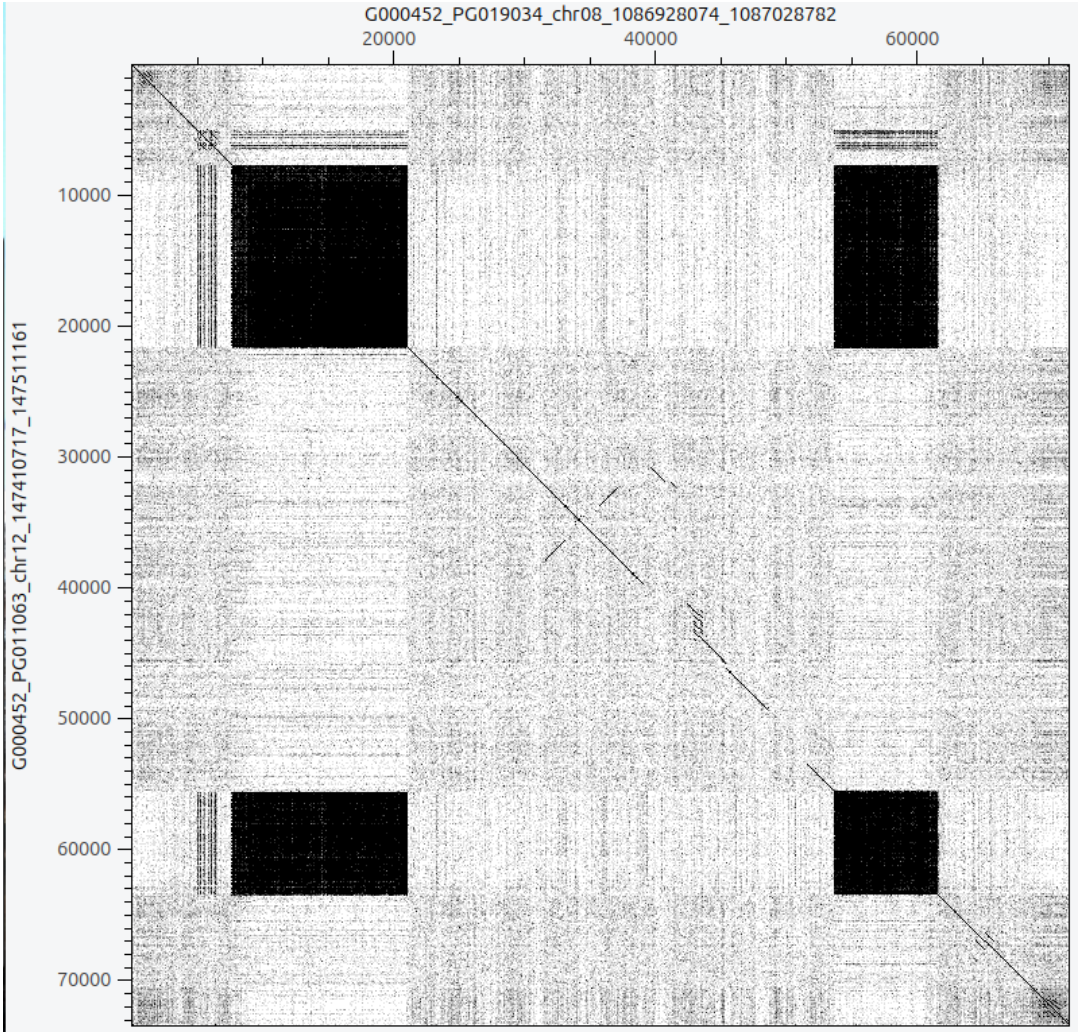


Figure S7 Dot plot alignment of two copies of an example mariner DNA element.

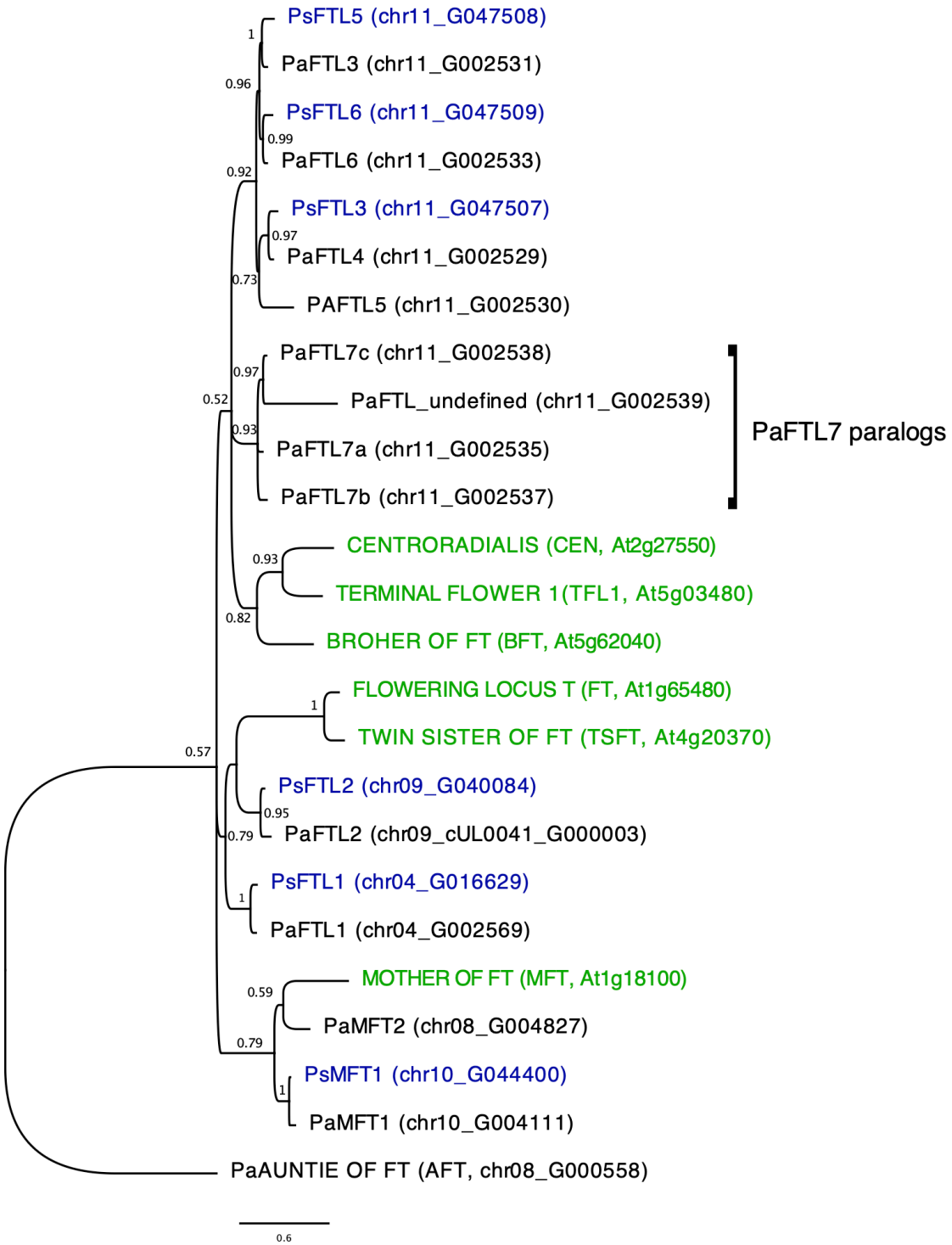


Figure S8 Phylogenetic relationship of *FT/TFL1* family genes from the the angiosperm *Arabidopsis thaliana* and the conifers *Picea abies* and *Pinus sylvestris*. Shown is a 50% majority rule tree derived using Bayesian phylogenetics. Numbers beside each node indicate posterior probabilities. *A. thaliana* (At) gene names are coloured green *P. abies* (Pa) gene names black and, *P. sylvestris* (Pt) gene names blue. Locus IDs (in bractes) are provided after each gene name.

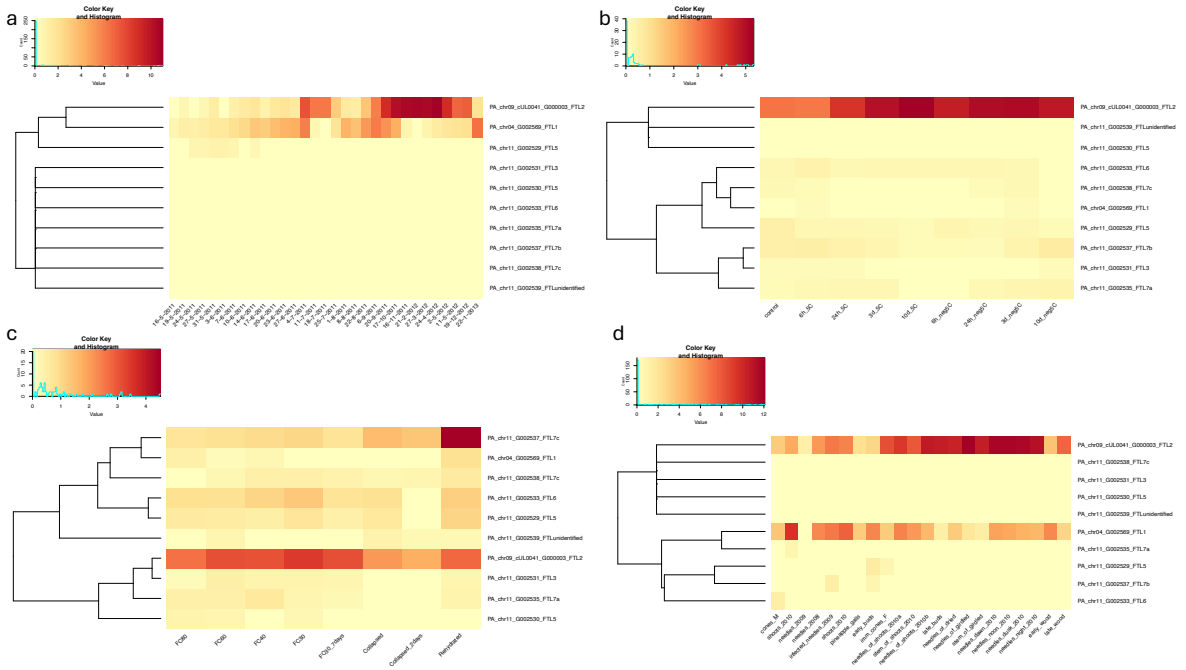


Figure S9 Gene expression of FTL-like genes in Norway spruce (*Picea abies*) **a** needles sampled throughout a growing season (sample labels indicate sampling date), **b** cold stress (sample labels indicate temperature and time duration at that temperature), **c** drought stress (sample labels indicate percentage of water field capacity (FC) or sampling at, or two days after a collapse of photosynthesis), **d** an expression atlas of different tissue types. Values are variance stabilizing transformation (VST) values calculated using DESeq2. Genes are clustered based on expression values.

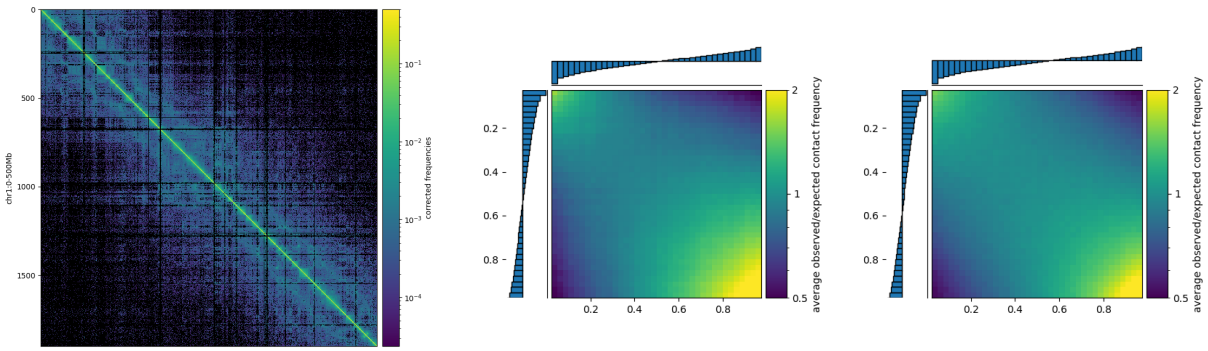
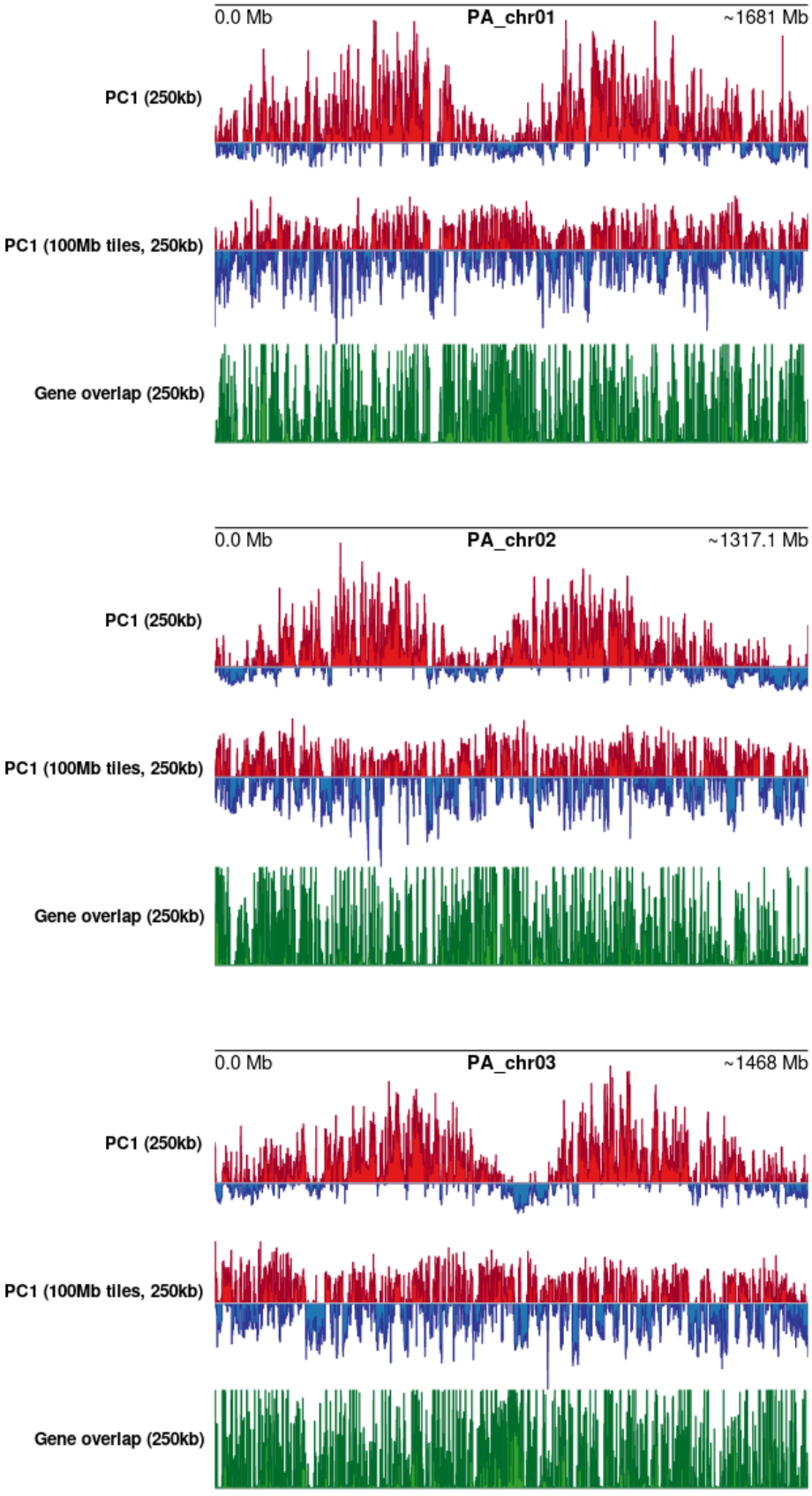
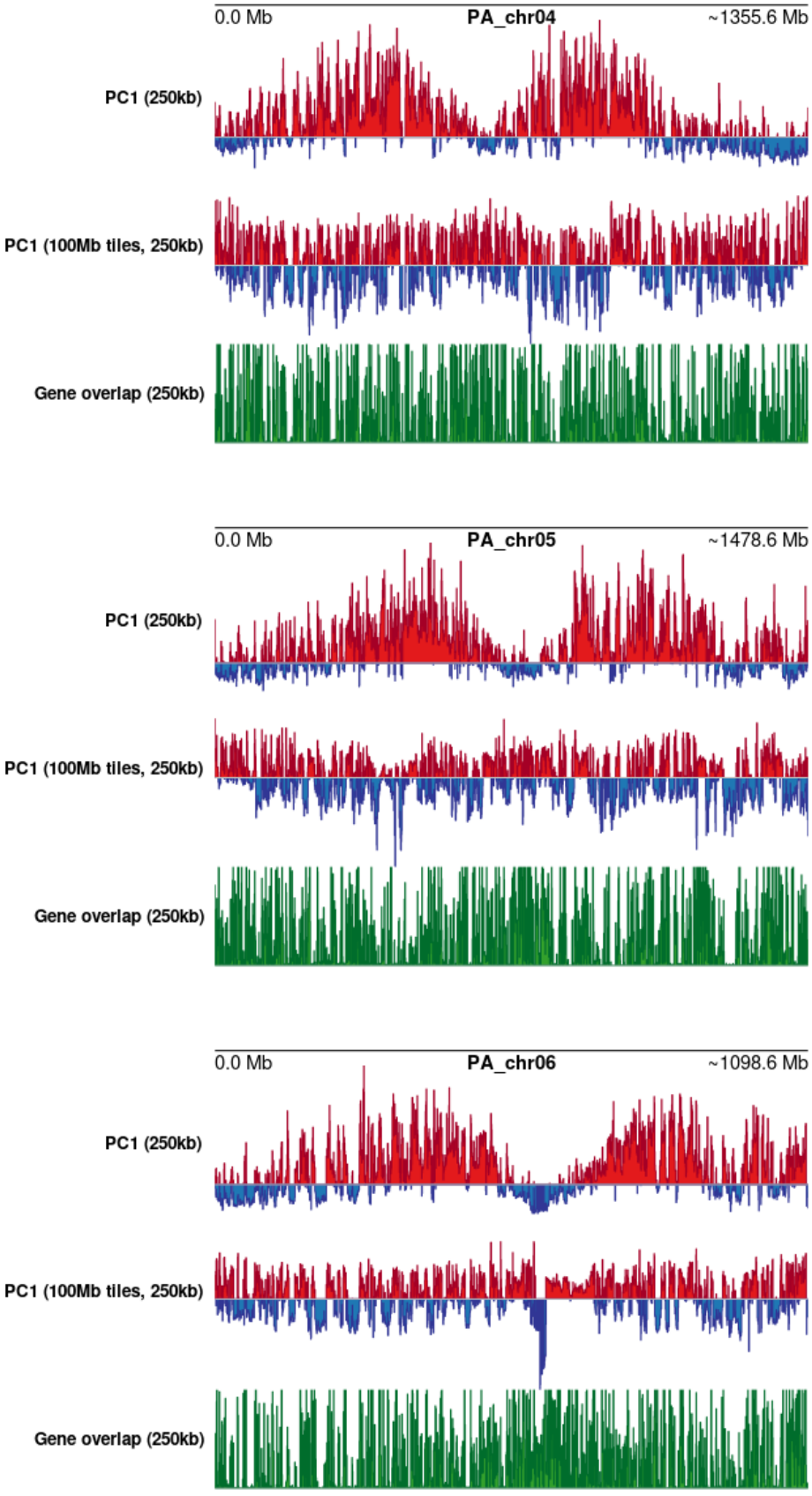
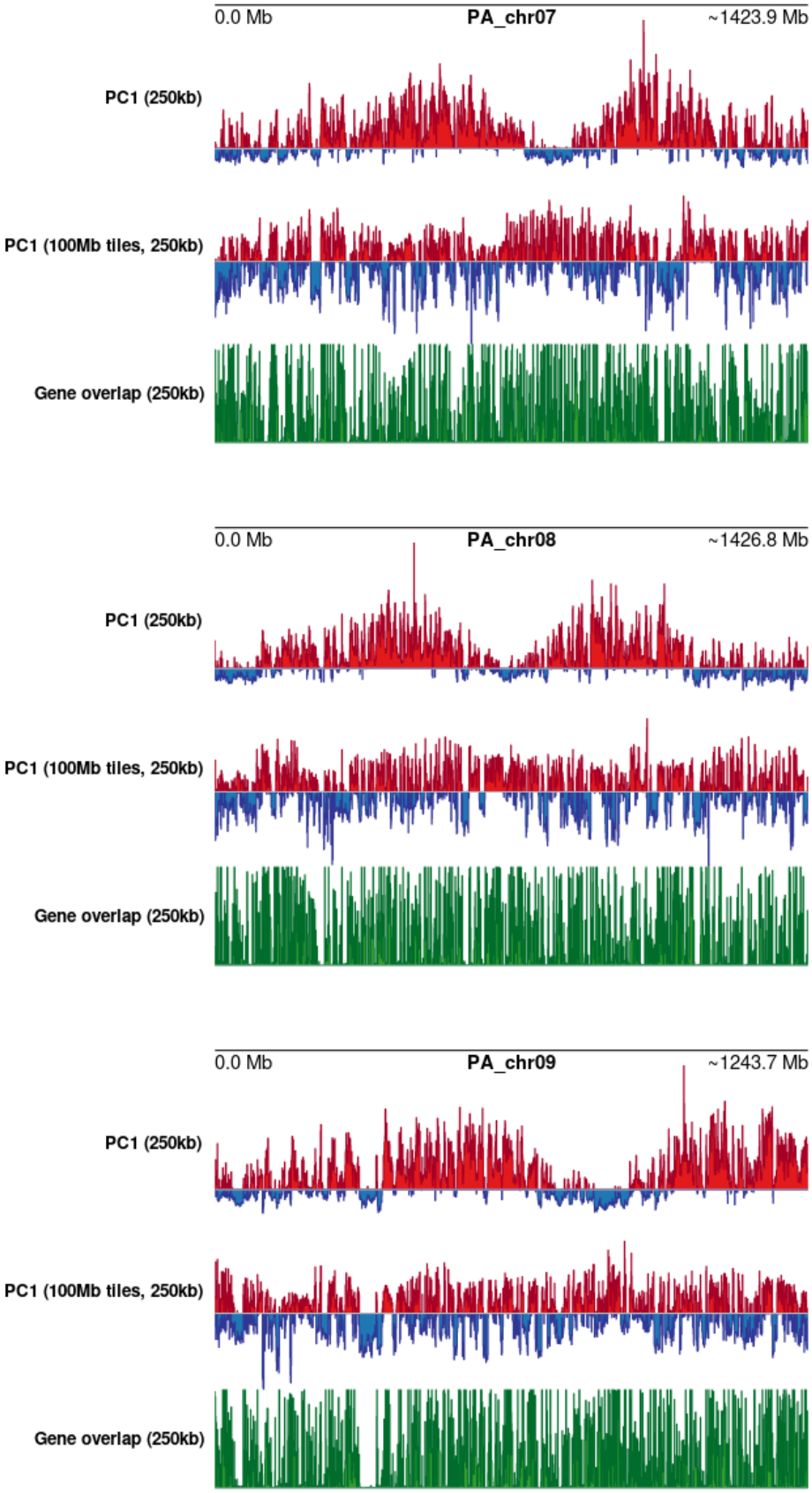


Figure S10 a Contact (corrected) frequencies between 250 Kbp bins in the first 500 Mbp of chromosome 1 (left) and saddleplots showing contact frequencies (observed/expected) between AB compartments binned by PC1 scores (high negative to high positive), **b** 250 Kbp resolution (middle), **c** 25 Kbp resolution (right).


0.0 Mb
PA_chr01
~1681 Mb
PC1 (250kb)
PC1 (100Mb tiles, 250kb)
Gene overlap (250kb)
0.0 Mb
PA_chr02
~1317.1 Mb
PC1 (250kb)
PC1 (100Mb tiles, 250kb)
Gene overlap (250kb)
0.0 Mb
PA_chr03
~1468 Mb
PC1 (250kb)
PC1 (100Mb tiles, 250kb)
Gene overlap (250kb)


0.0 Mb
PA_chr04
~1355.6 Mb
PC1 (250kb)
PC1 (100Mb tiles, 250kb)
Gene overlap (250kb)
0.0 Mb
PA_chr05
~1478.6 Mb
PC1 (250kb)
PC1 (100Mb tiles, 250kb)
Gene overlap (250kb)
0.0 Mb
PA_chr06
~1098.6 Mb
PC1 (250kb)
PC1 (100Mb tiles, 250kb)
Gene overlap (250kb)


0.0 Mb
PA_chr07
~1423.9 Mb
PC1 (250kb)
PC1 (100Mb tiles, 250kb)
Gene overlap (250kb)
0.0 Mb
PA_chr08
~1426.8 Mb
PC1 (250kb)
PC1 (100Mb tiles, 250kb)
Gene overlap (250kb)
0.0 Mb
PA_chr09
~1243.7 Mb
PC1 (250kb)
PC1 (100Mb tiles, 250kb)
Gene overlap (250kb)

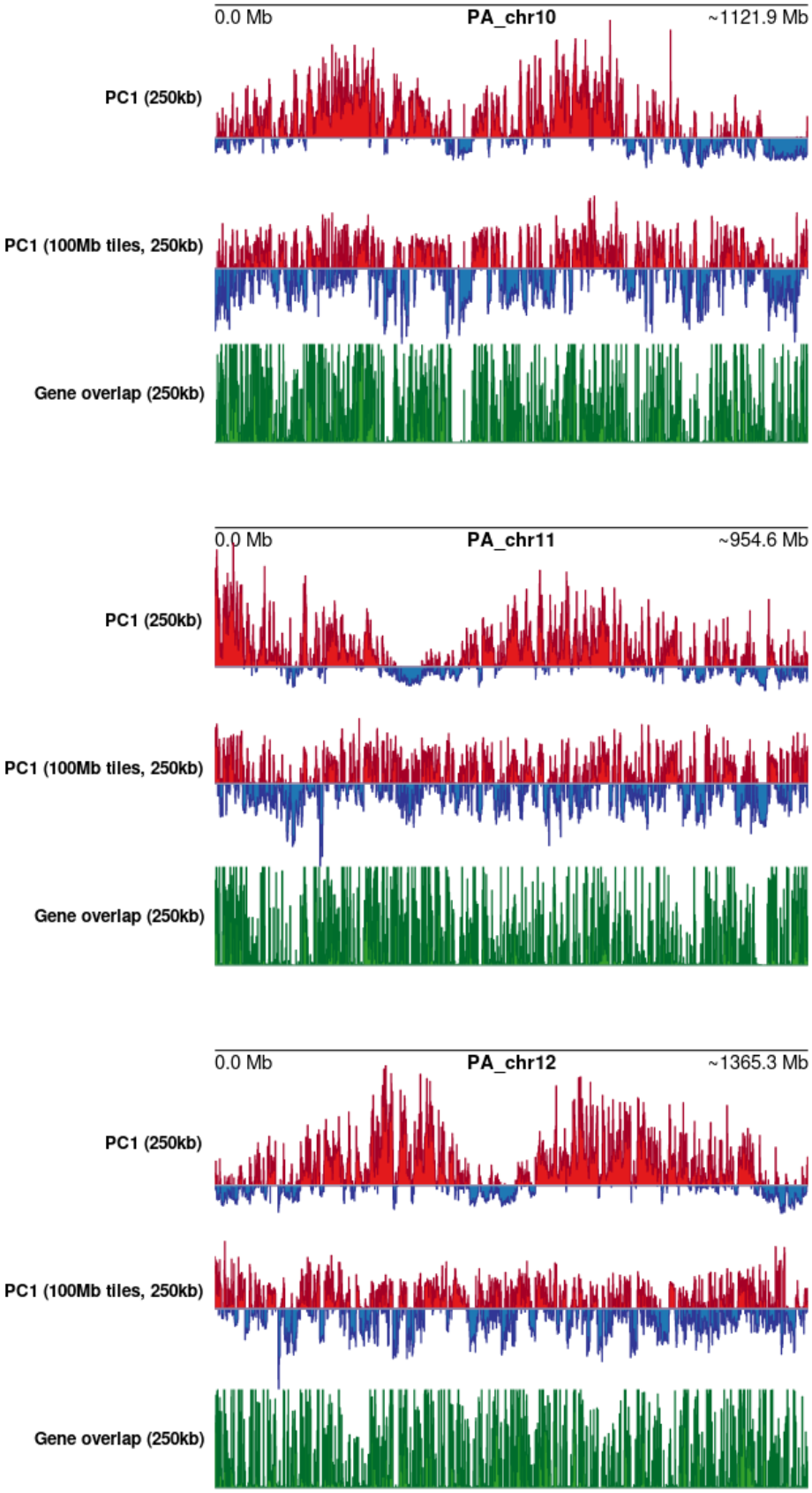


Figure S11 Score values for principal component (PC) 1 of a principal component analysis of chromatin contact frequency identified using Micro-C data from young needles in Norway spruce (*Picea abies*) for each chromosome (as indicated by

PA_chrX). The PC orientation was adjusted so that positive values corresponded to regions with higher open chromatin content. Positive values (red) are termed A compartments while negative values (blue) and termed B compartments. The analysis was performed chromosome-wide and in 100 Mbp tiles using 250 Kbp windows. Gene overlap (the percentage of bases covered by a gene feature) within each 250 Kbp is indicated as a track (green) below the PC1 score values.

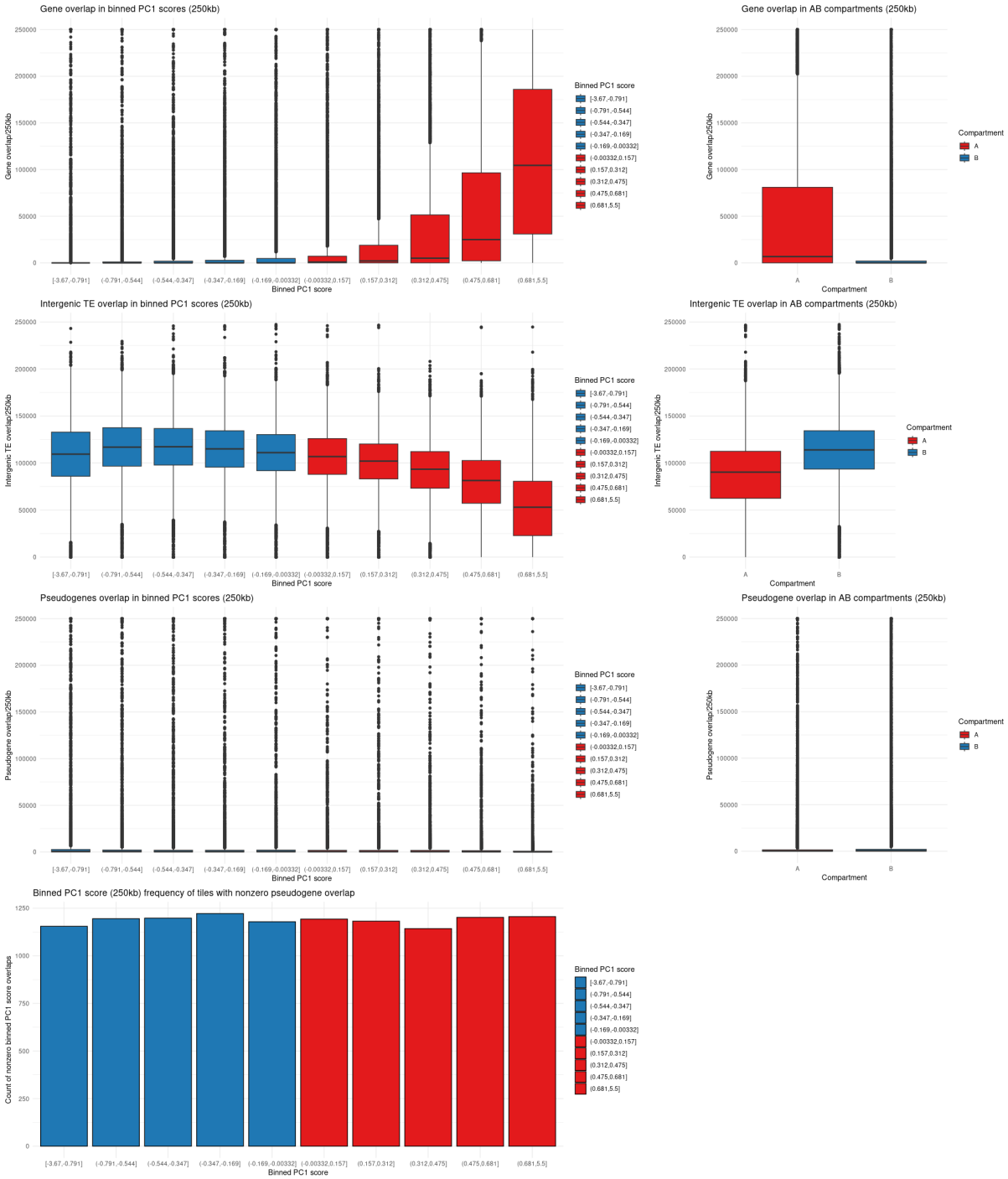


Figure S12 Gene, intergenic transposable element (TE) and pseudogene overlap split by AB compartment sign and spectrum of binned PC1 scores (250 Kbp resolution in 100 Mbp tiles).

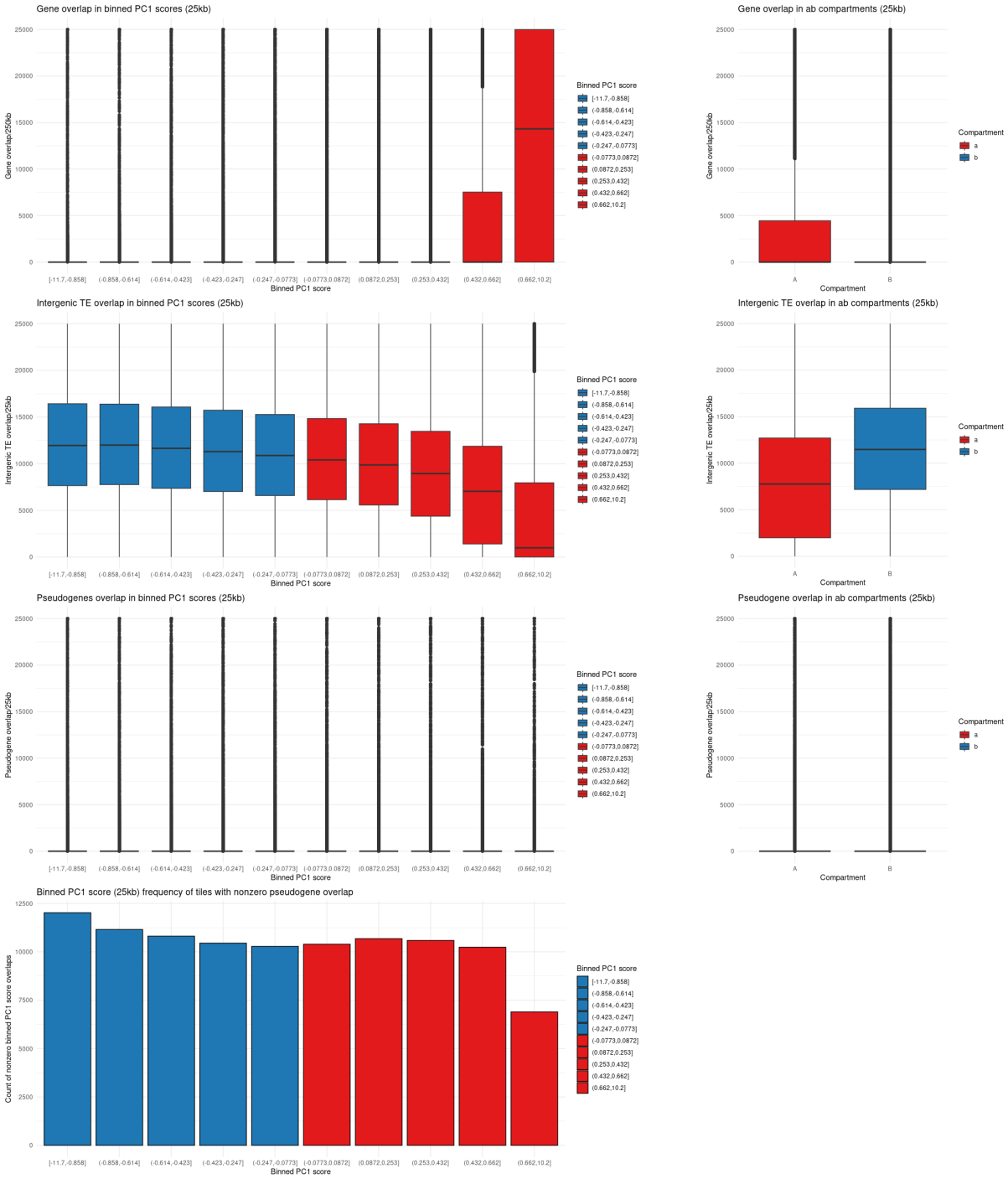


Figure S13 Gene, intergenic transposable element (TE) and pseudogene overlap split by AB compartment sign and spectrum of binned PC1 scores (25 Kbp resolution in 100 Mbp tiles).

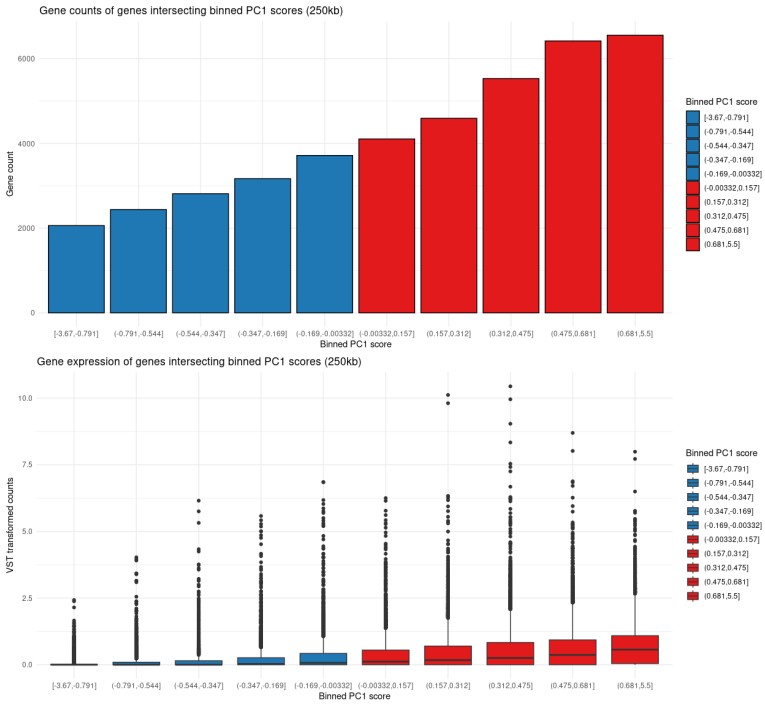


Figure S14 Gene count and gene expression levels of TSS intersecting tiles with binned PC1 scores 250 Kbp resolution in 100 Mbp tiles.

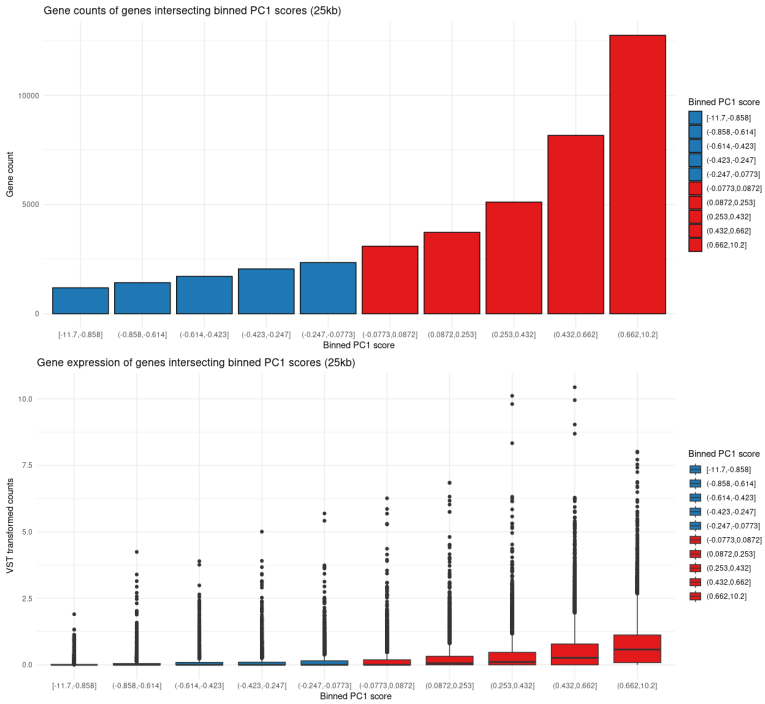


Figure S15 Gene count and gene expression levels of TSS intersecting tiles with binned PC1 scores 25 Kbp resolution in 100 Mbp tiles.

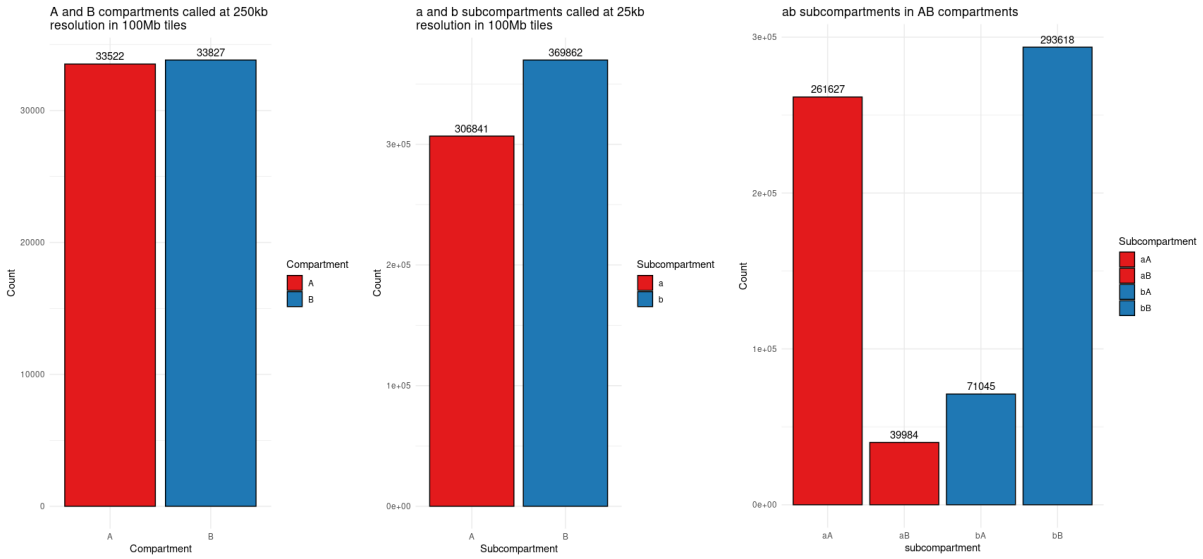


Figure S16 Counts of AB compartments called at 250 Kbp resolution in 100 Mbptiles (left), counts of ab sub-compartments called at 25 Kbp resolution in 100 Mbp tiles and the frequency of ab subcompartments (25Kbp) called within AB compartments (250 Kbp).

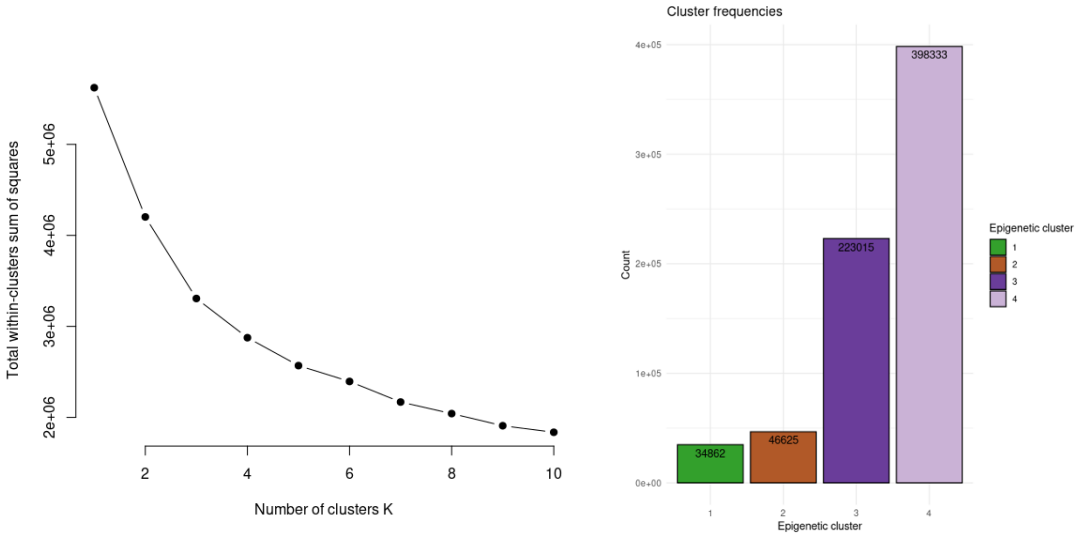


Figure S17 K-means clustering of epigenetic signals identifies an optimum of four clusters. Epigenetic signals were analysed within 25 Kbp windows.

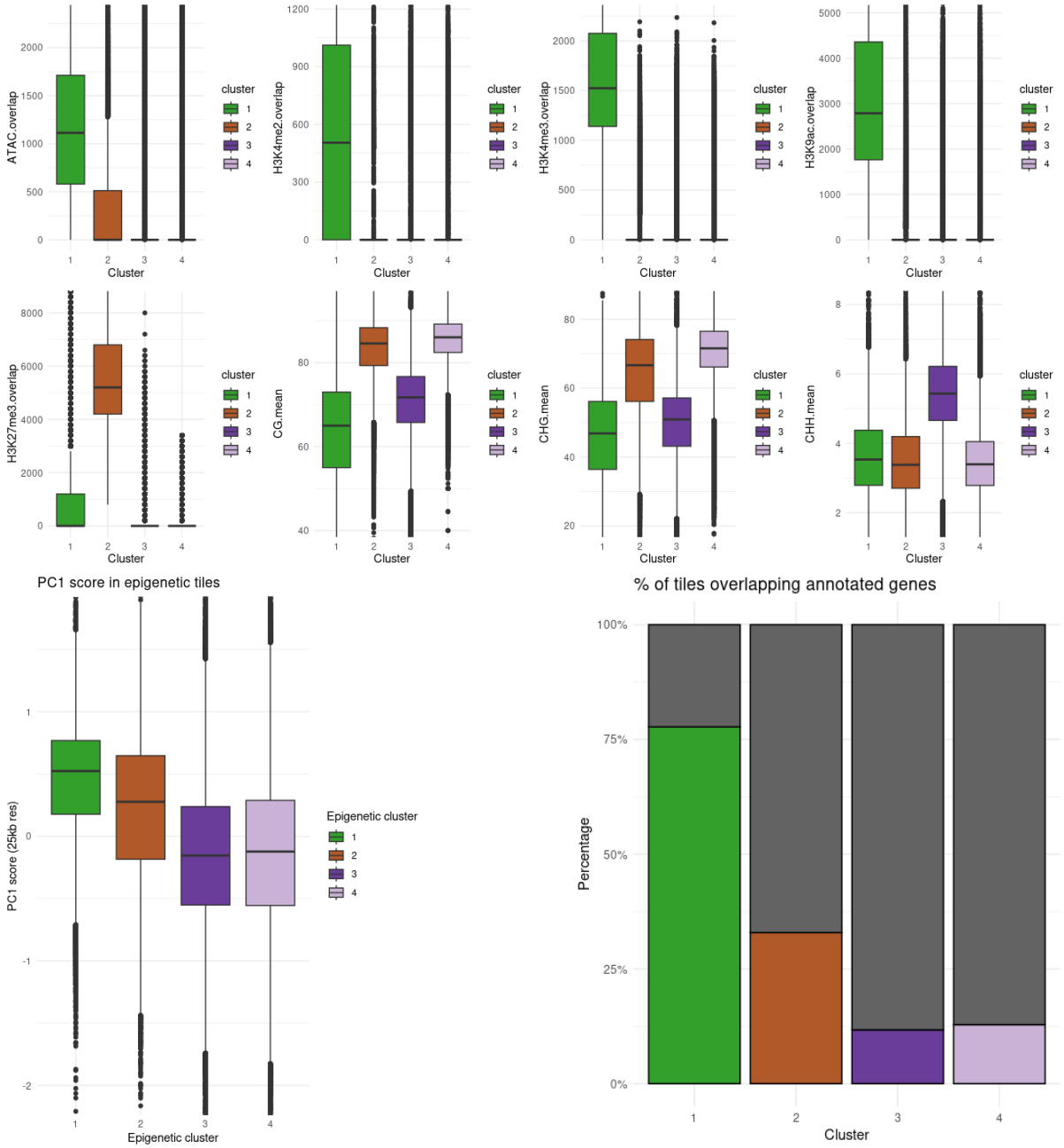


Figure S18 Box plot distributions of epigenetic signals analysed in 25 Kbp windows used as input to k-means clustering, where an optimum of four clusters was identified, 1st - 99th percentile shown for each epigenetic feature. PC1 scores (25 Kbp resolution in 100 Mbp tiles) of regions annotated to each cluster type in bottom left figure. The percentage of 25 Kbp windows containing genes in each cluster is indicated bottom right.

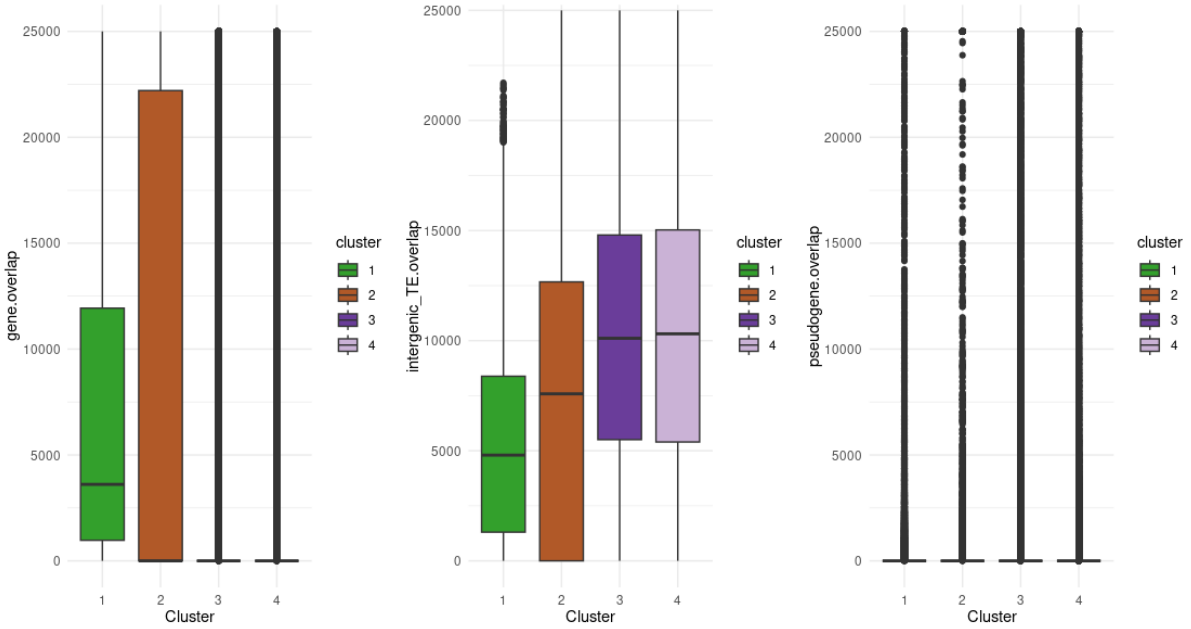


Figure S19 Box plot distributions of gene, intergenic transposable element (TE) and pseudogene feature overlap (the number of bases covered by each feature type) in 25 Kbp windows within four clusters identified from k-means clustering of epigenetic signals.

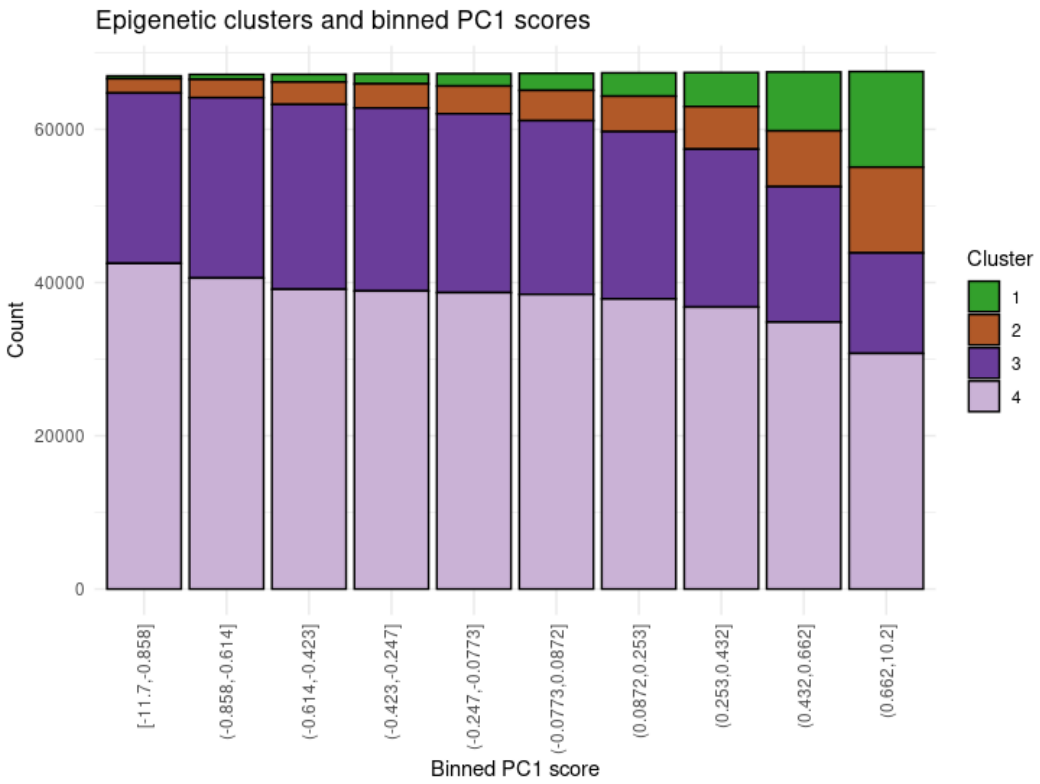


Figure S20 Binned principal component 1 score values of a principal component analysis of Micro-C chromatin contact frequency in young needles. The stacked bars indicate the count of 25 Kbp windows of each cluster type. Clusters identified by k-means clustering of epigenetic signals analysed in 25 Kbp windows.

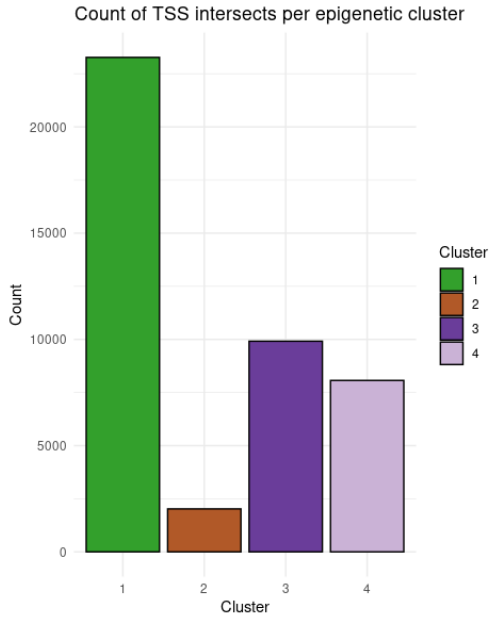


Figure S21 Counts of transcription start sites (TSS) within four clusters identified by k-means clustering of epigenetic signals analysed in 25 Kbp windows.

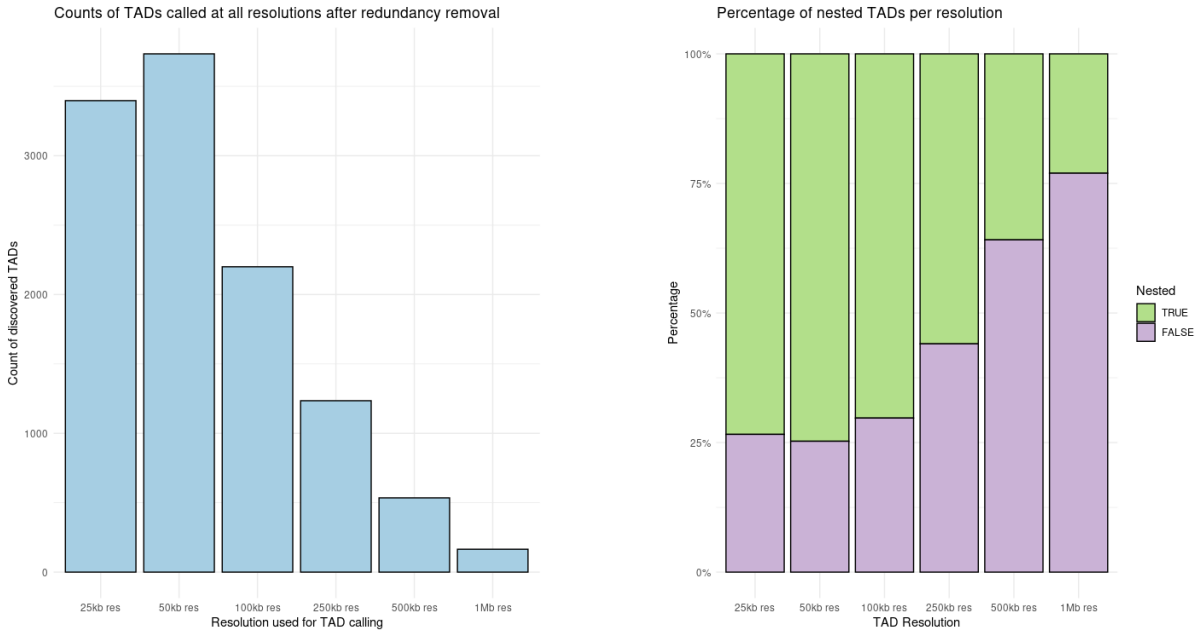


Figure S22 a Count of topologically associating domains (TADs) found at each resolution after redundancy removal (left) and **b** the percentage of TADs identified as nested inside a larger TAD (right).

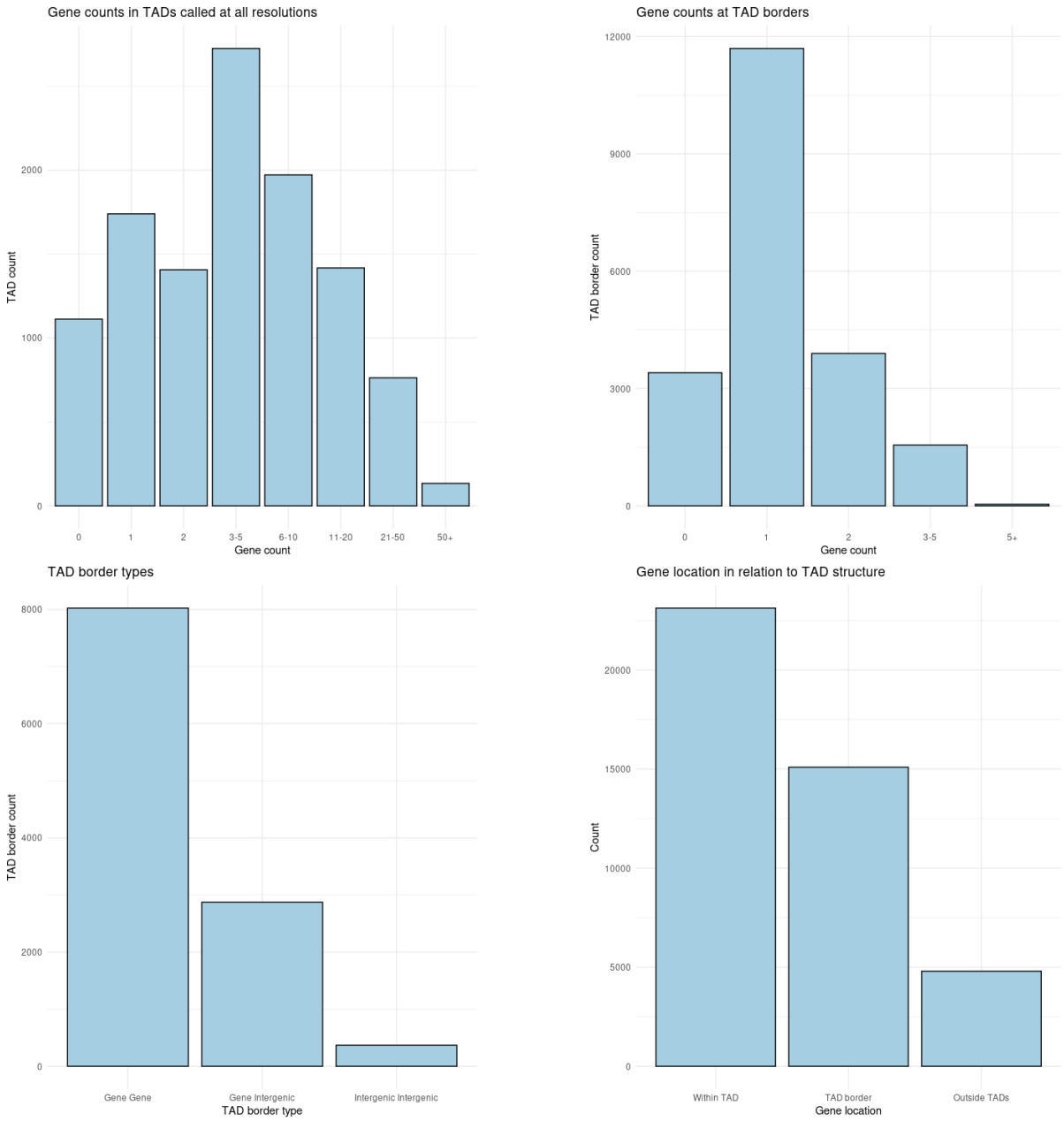


Figure S23 Topologically associating domains (TADs) called at any resolution, gene counts inside TAD (top left), the count of genes flanking the border (top right), the count of TAD border types (bottom left) and the number of genes found inside a TAD, the regions flanking a TAD border or outside of TADs and not at a border (bottom right).

Figure S24 Expression level and gene location in relation to topologically associating domains (TADs).

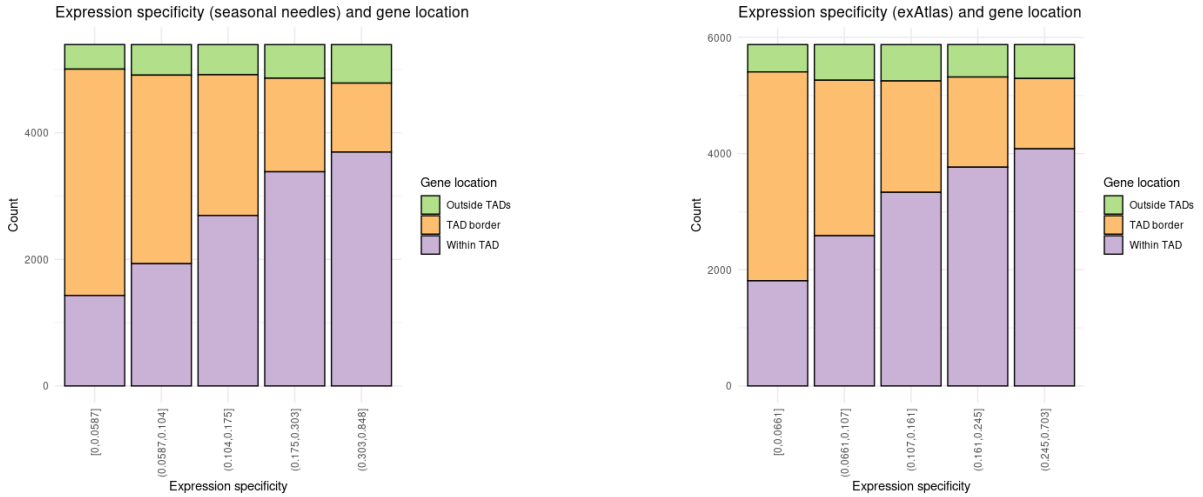


Figure S25 Expression specificity and gene location in relation to topologically associating domains (TADs), seasonal needle (left) and exAtlas (right).

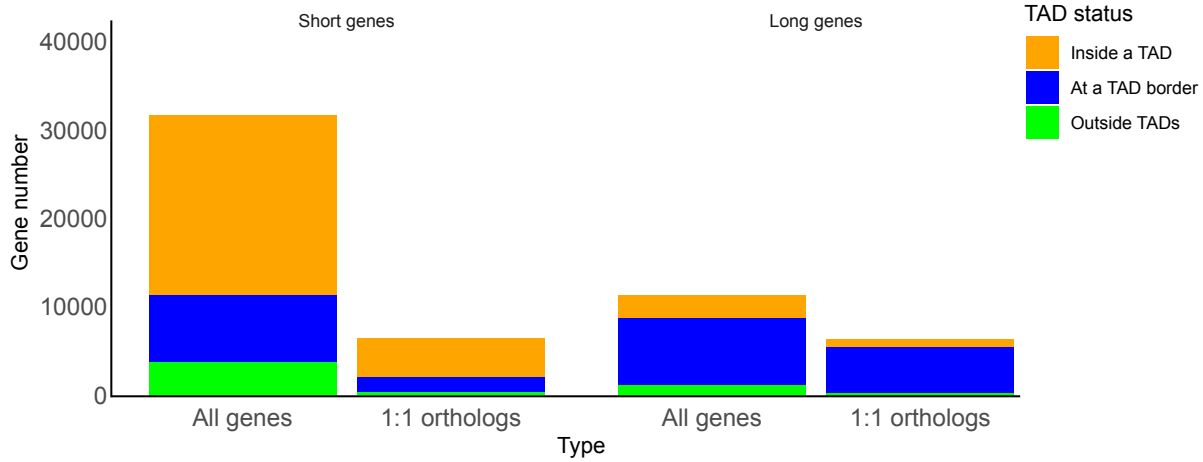


Figure S26 Gene counts of genes containing long or short introns and for all genes or genes that are also one-to-one orthologs between *Picea abies* and *Pinus sylvestris*. For each group, the count of genes in different topologically associating domain (TAD) contexts is indicated by colour.

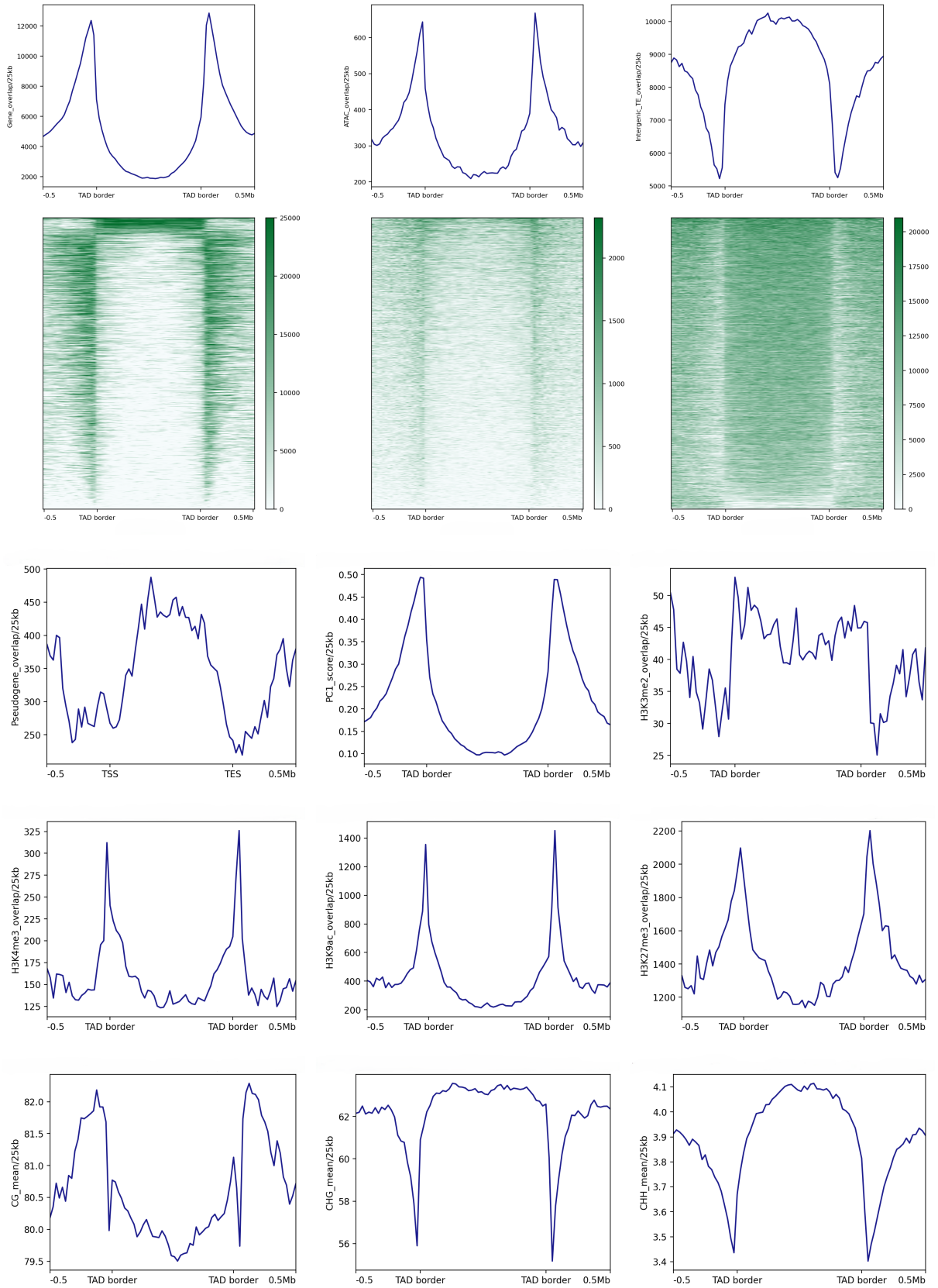


Figure S27 Gene, transposable element, pseudogene and various epigenetic features across topologically associating domains called at 25 Kbp resolution. Histone modifications were assayed using chromatin immunoprecipitation and DNA methylation using Oxford Nanopore sequencing. Overlap refers to the number of bases covered by a feature type.

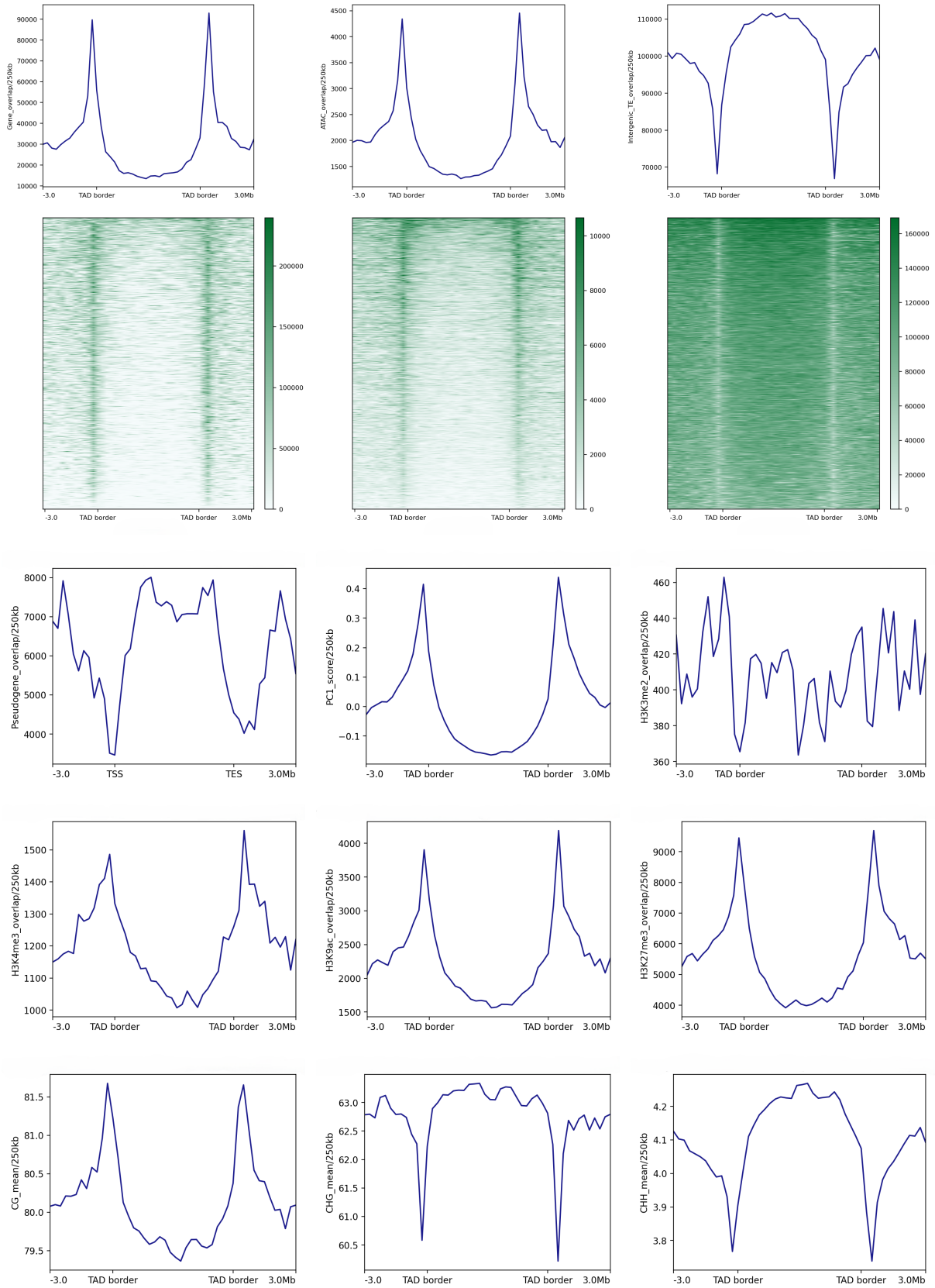


Figure S28 Gene, transposable element, pseudogene and various epigenetic features across topologically associating domains called at 250 Kbp resolution. Histone modifications were assayed using chromatin immunoprecipitation and DNA methylation using Oxford Nanopore sequencing. Overlap refers to the number of bases covered by a feature type.

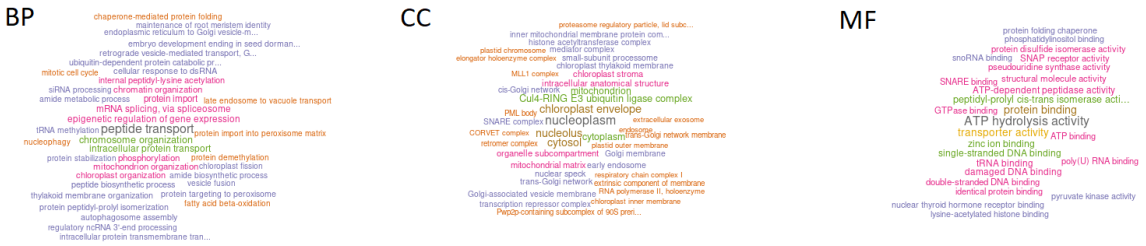


Figure S29 Gene Ontology term enrichment of biological process (BP), cellular component (CC) and molecular function (MF) terms of genes located at topologically associating domain boundaries.

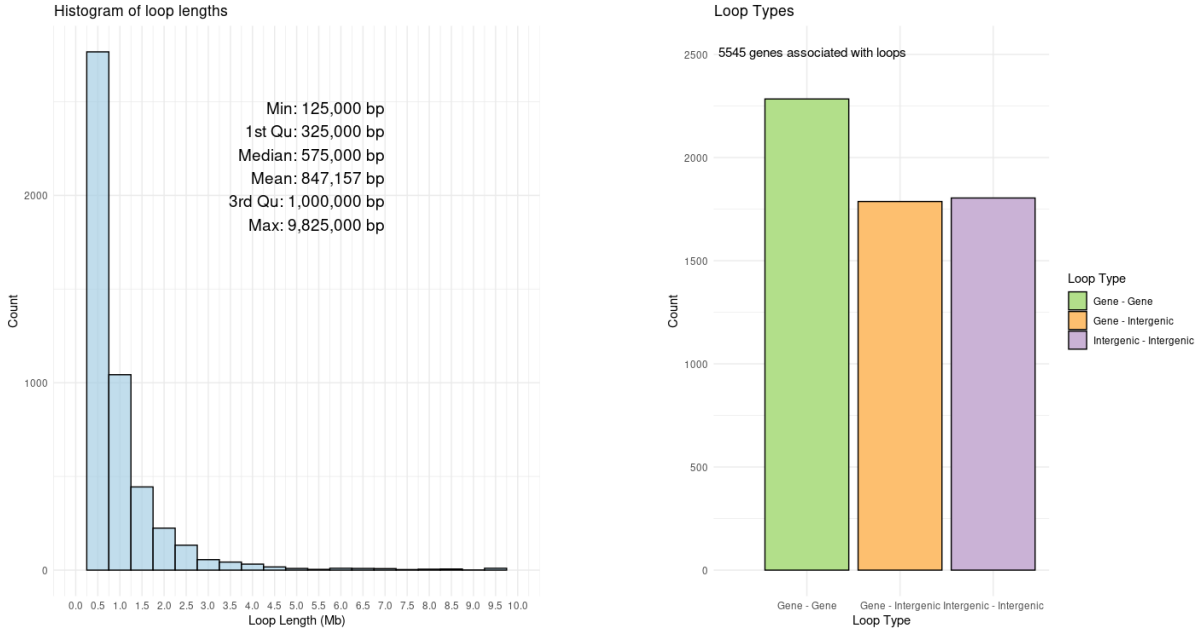


Figure S30 Length of loops called (left), and count of loop types (right).

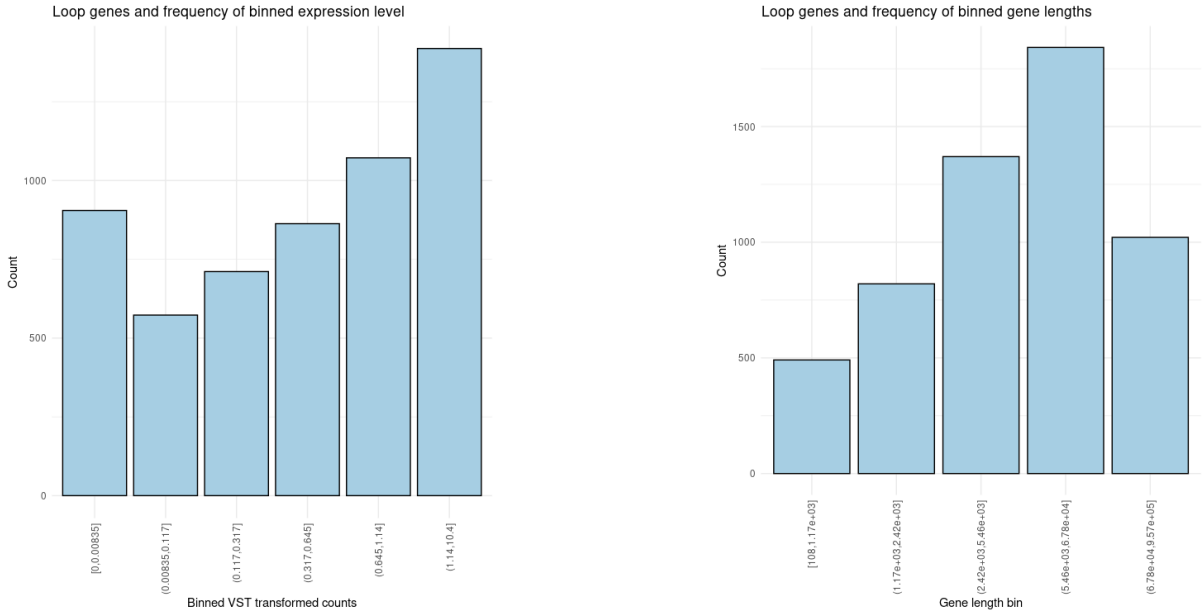


Figure S31 Chromatin loop genes and binned expression levels (left) and gene length (right).

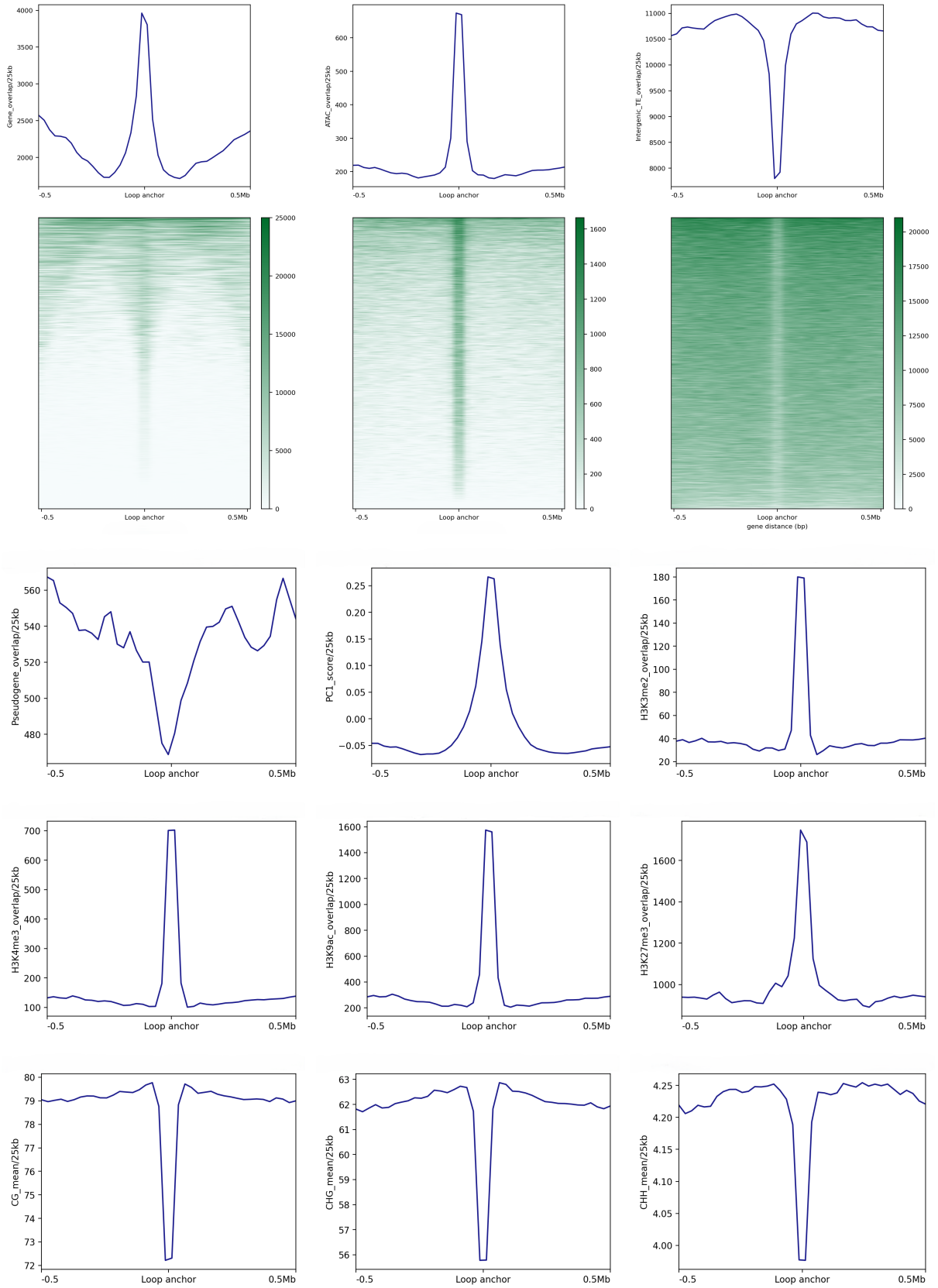


Figure S32 Gene, transposable element, pseudogene and various epigenetic features at chromatin loops called at 25 Kbp resolution. Histone modifications were assayed using chromatin immunoprecipitation and DNA methylation using Oxford Nanopore sequencing. Overlap refers to the number of bases covered by a feature type.

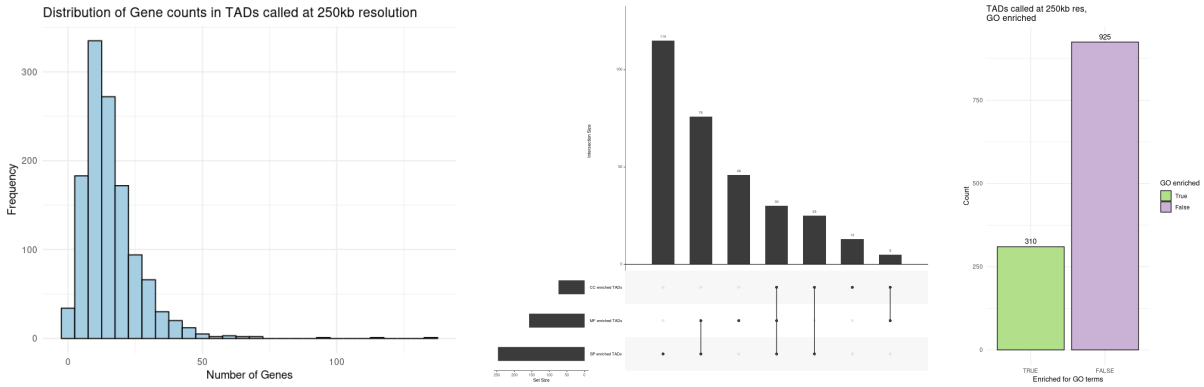


Figure S33 Topologically associating domains (TADs) called at 250 Kbp resolution, gene counts in TADs of 250 Kbp resolution (left), Gene Ontology (GO) term enrichment for biological process (BP), cellular component (CC) and molecular function (MF) (middle) and the total number of GO enriched (BP, CC or MF) TADs called at 250 Kbp resolution (right).

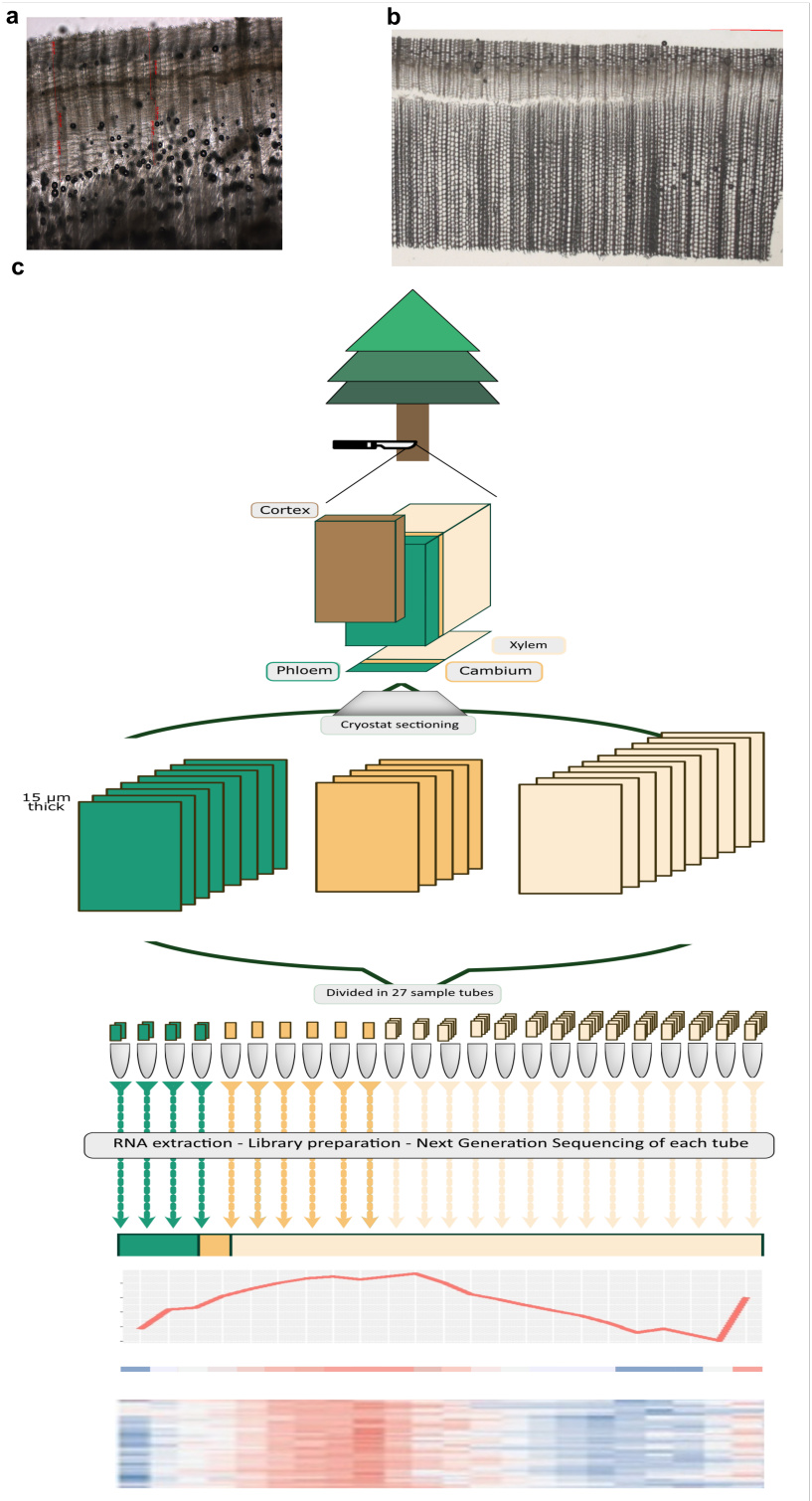


Figure S34 Representative cross sections and sampling schematic for RNA-Sequencing analysis of wood development. A Tangential cross section of Norway spruce (*Picea abies*) **b** Tangential cross section of Scots pine (*Pinus sylvestris*). **c** Overview schematic of the cryosection sampling strategy used to generate RNA-Sequencing data profiling transcript expression during wood development of *Picea abies* and *Pinus sylvestris*.

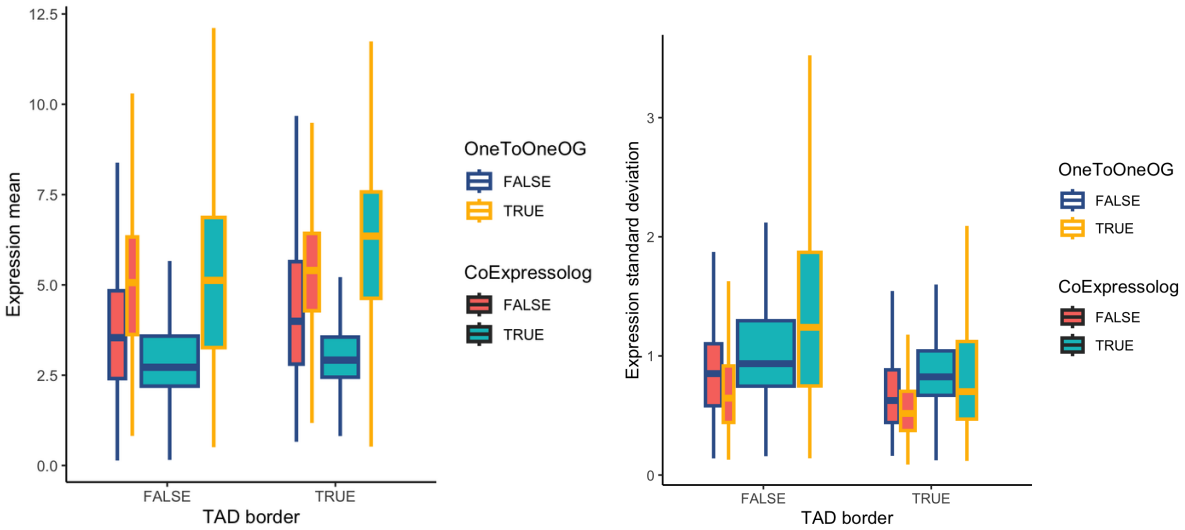


Figure S35 a Expression mean distribution in Norway spruce of genes at topologically associating domain (TAD) boundaries or not (indicated as True/False) and divided into gene classified as having conserved co-expression during wood development in *Picea abies* and *Pinus sylvestris* (co-expressologs; True/False) and as being one-to-one orthologs between the two species (True/False). **b** Expression standard deviation of gene expression values with categories as indicated in **a**.

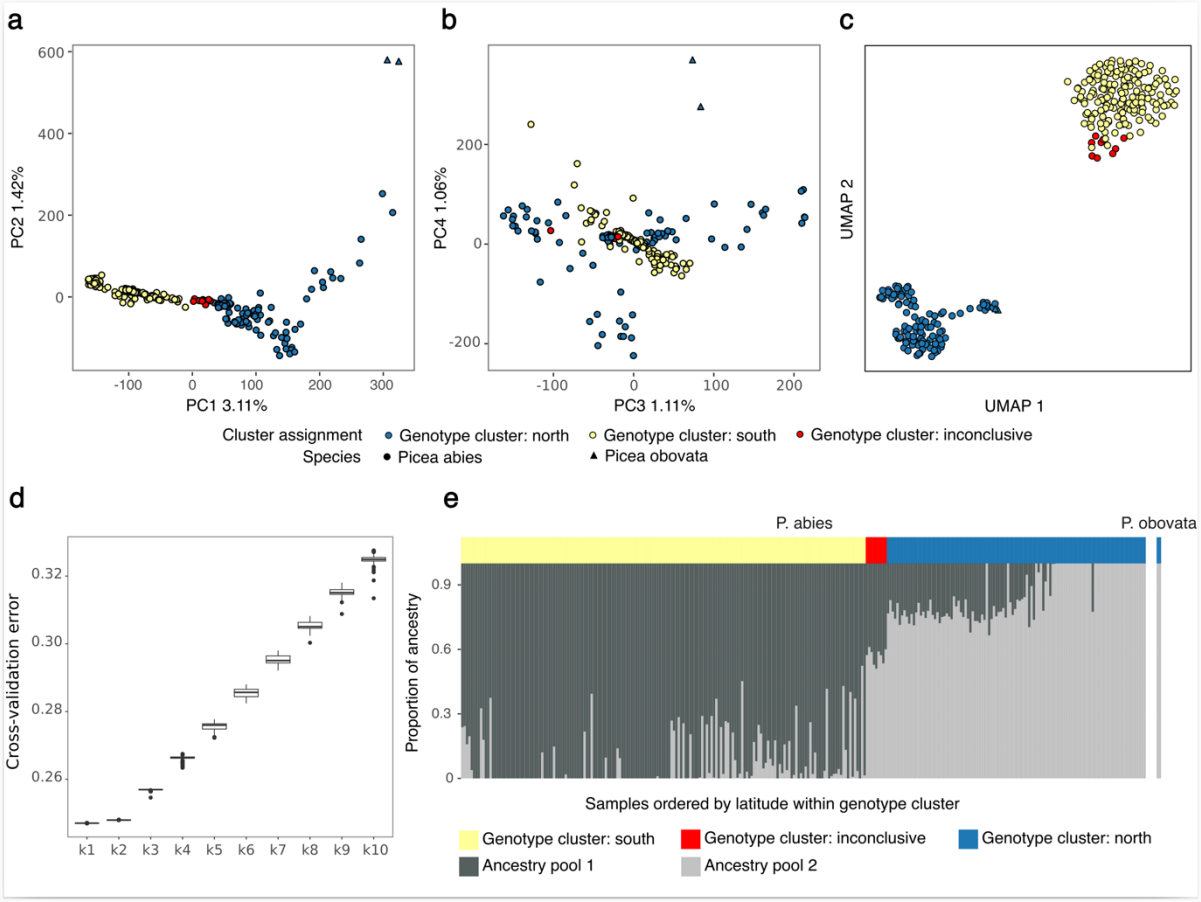


Figure S36 Population structure analysis using 1.5 million genome wide SNPs from 291 *Picea abies* and 2 *Picea obovate* individuals. Principal Component Analysis (PCA) clustering on genotypes showing Principal Component (PC) 1 and PC2 (**a**) PC3 and PC4 (**b**) and Uniform Manifold Approximation and Projection (UMAP) clustering of genotypes (**c**). Colours represent cluster assignment north (blue), yellow (south) or inconclusive (red), *i.e.,* samples assigned to different clusters when comparing PCA and UMAP clustering. Admixture analysis showing cross-validation error for 1000 iterations for K 1 to 10 to find the best K (**d**) and clustering result for K == 2 (**e**) with samples first sorted based on cluster assignment (PCA and UMAP) then latitude on the x-axis and the proportion of ancestry on the y-axis.

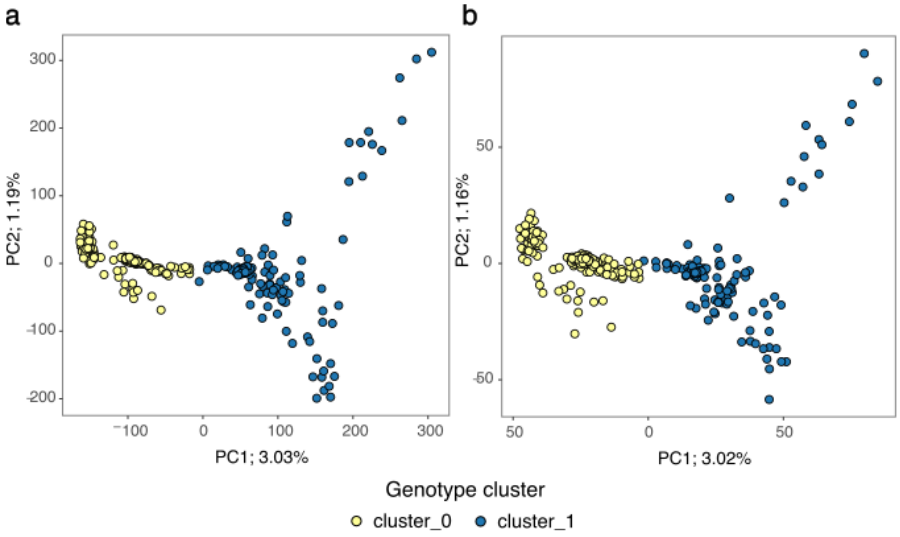


Figure S37 Principal Component Analysis clustering of 291 range-wide *Picea abies* samples comparing genome wide Single Nucleotide Polymorphisms (SNPs) (**a**) and exonic SNPs (**b**). Colours represent the two clusters made by K-means clustering of the genome wide SNPs set (**a**).

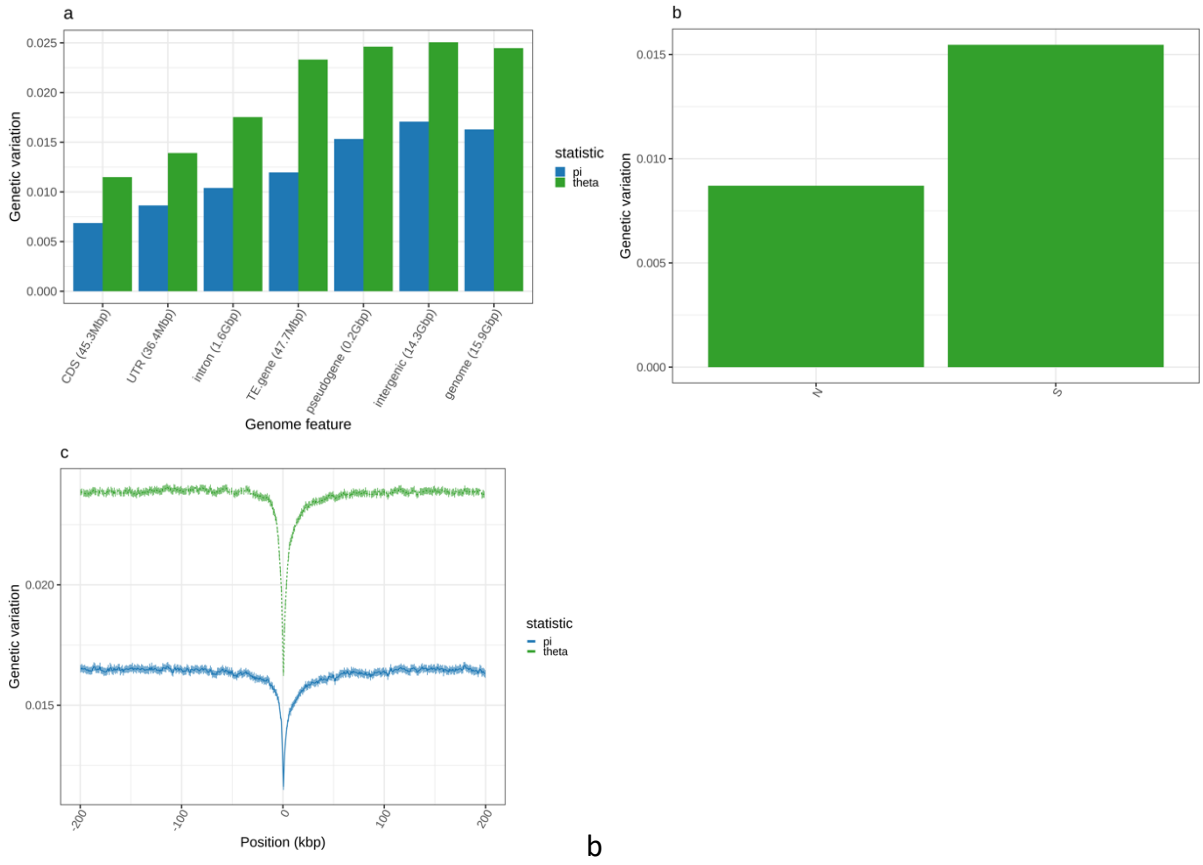


Figure S38 Mean levels of genetic variation (y-axis) **a** across genomic features **b** measured as Watterson's theta over non-synonymous and synonymous sites and **c** across 1Kbp nonoverlapping windows upstream or downstream of annotated CDS regions. Pi is nucleotide diversity and theta population mutation rate. Error bars in **c** are 95% confidence intervals (200 bootstrap replicates). Only unmasked accessible sites are considered for calculations.

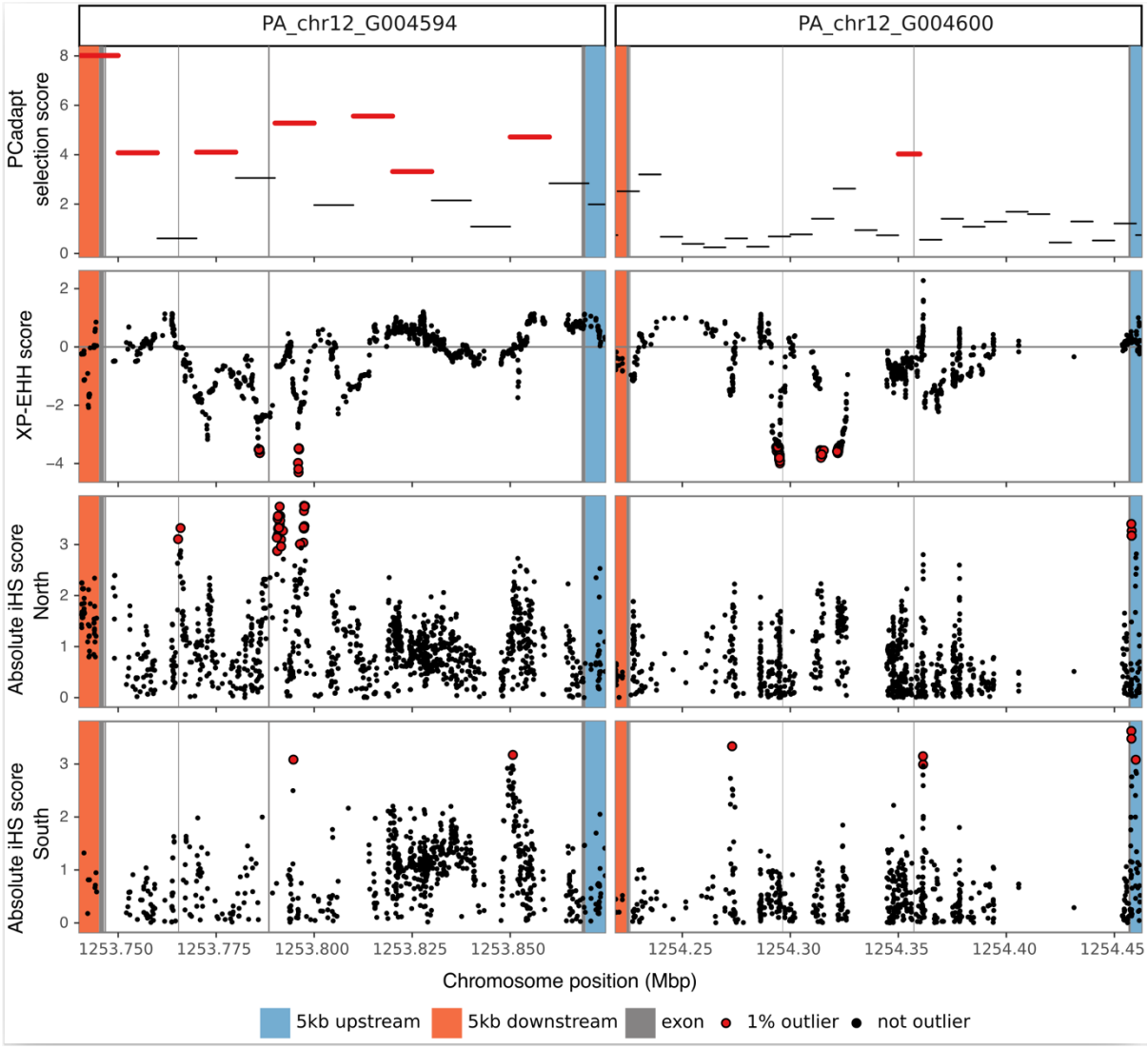


Figure S39 Selection signals across locally duplicated *CONSTANS-Like* (*COL*) genes, PA_chr12_G004594 (1st column) and PA_chr12_G004600 (2nd column). 5 Kbp upstream (blue), 5 Kbp downstream (orange) and exons (grey) are indicated as blocks while 1% outlier values (10 Kbp windows top row, Single Nucleotide Polymorphisms other rows) are indicated in red while non outlier values are black. The first row shows signs of local adaptation using a Principal Component Analysis based approach on all *Picea abies* samples. The second row shows signs of positive selection contrasting the north and south sample sets. The last two rows show signs of positive selection in the north (3rd row) or south (4th row) sample sets.

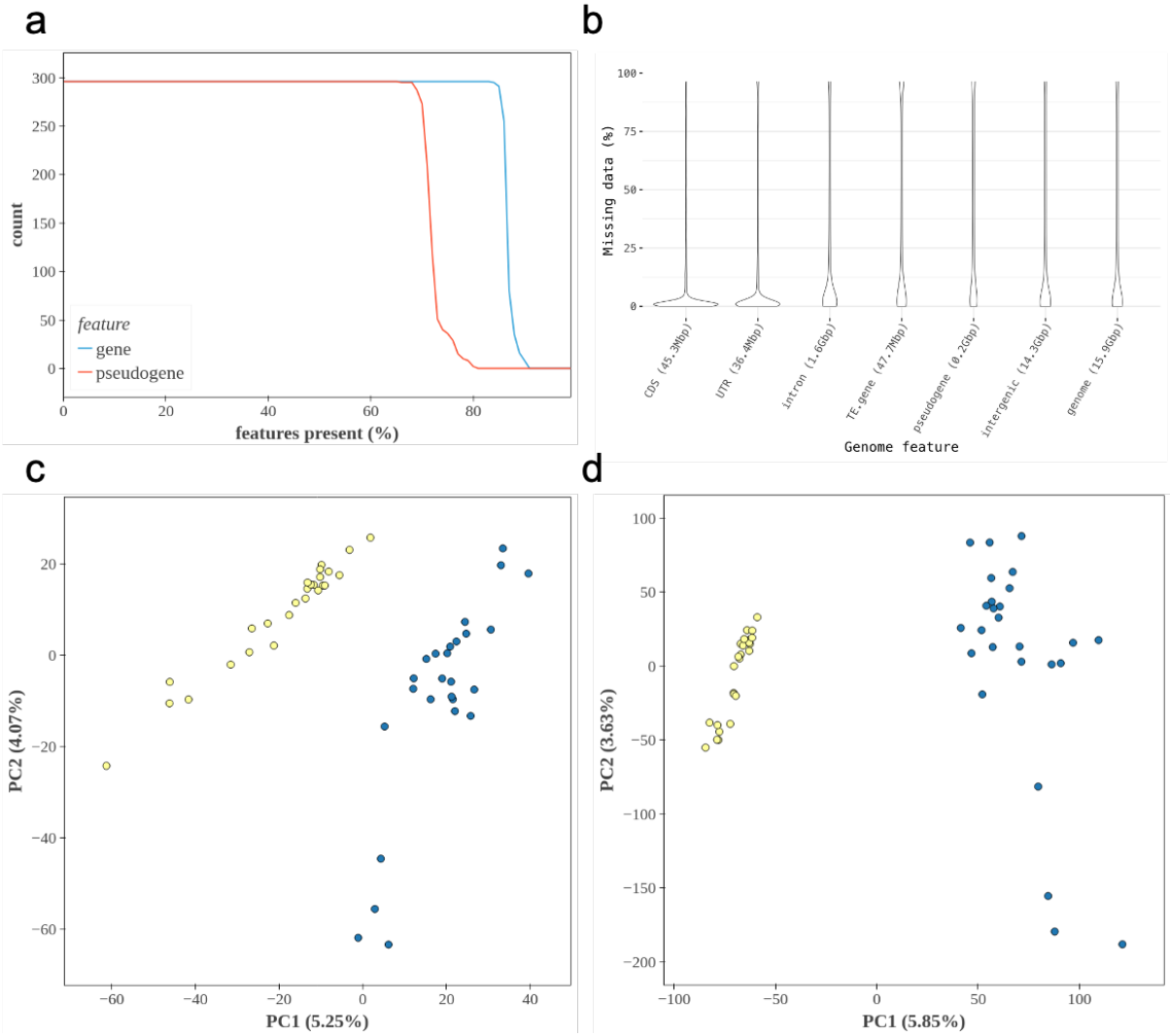


Figure S40 a The percentage of gene (blue) and pseudogene (red) features classified as present plotted against the number of individuals with that percentage of features present. The analysis was performed using a set of 296 individuals with high coverage resequencing data. Presence was defined as sequencing read coverage across >20% of gene/pseudogene sequence. **b** Percentage of missing data by genome feature. Each violin plot is consists of a subsample (n=10,000) of weighted coverages from mosdepth coverage profile histograms based on present/absent bases summed over all individuals. A base in a feature was coded as present if its sequencing coverage was >= 3 in an individual sample. **c** Principal Component Analysis plot of the first two principal components based on sequencing read coverage of gene coding sequence (CDS) regions for northern and southern individuals with the highest sequencing read depth. Each CDS was determined to be either present or absent in an individual if >20% of the gene had coverage. **d** As for **c** but representing pseudogenes.

Supplementary References

Adrian Alexa, J. R. (2017). *topGO* [Computer software]. Bioconductor. https://doi.org/10.18129/B9.BIOC.TOPGO

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Alistair Miles, pyup. io bot, Murillo F. Rodrigues, Peter Ralph, Jerome Kelleher, Max Schelker, Rahul Pisupati, Summer Rae, & Tim Millar. (2024). *cggh/scikit-allel: V1.3.13* (Version v1.3.13) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.13772087

Bag, P., Lihavainen, J., Delhomme, N., Riquelme, T., Robinson, K. M., & Jansson, S. (2021). An atlas of the Norway spruce needle seasonal transcriptome. *The Plant Journal*, *108*(6), 1815–1829. https://doi.org/10.1111/tpj.15530

Bernhardsson, C., Vidalis, A., Wang, X., Scofield, D. G., Schiffthaler, B., Baison, J., Street, N. R., García-Gil, M. R., & Ingvarsson, P. K. (2019). An Ultra-Dense Haploid Genetic Map for Evaluating the Highly Fragmented Genome Assembly of Norway Spruce *(Picea abies*). *G3 Genes|Genomes|Genetics*, *9*(5), 1623–1632. https://doi.org/10.1534/g3.118.200840

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting F_{ST}: The impact of rare variants. *Genome Research*, *23*(9), 1514–1521. https://doi.org/10.1101/gr.154831.113

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. https://doi.org/10.1093/molbev/msab293

Chan, A. H., Jenkins, P. A., & Song, Y. S. (2012). Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. *PLoS Genetics*, *8*(12), e1003090. https://doi.org/10.1371/journal.pgen.1003090

Chang, S., Puryear, J., & Cairney, J. (1993). A simple and efficient method for isolating RNA from pine trees. *Plant Molecular Biology Reporter*, *11*(2), 113–116. https://doi.org/10.1007/BF02670468

Chen, H., Zwaenepoel, A., & Van De Peer, Y. (2024). wgd v2: A suite of tools to uncover and date ancient polyploidy and whole-genome duplication. *Bioinformatics*, *40*(5), btae272. https://doi.org/10.1093/bioinformatics/btae272

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Criscuolo, A. and Gribaldo, S. (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology*, 10, 210. https://doi.org/10.1186/1471-2148-10-210

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w^{1118} ; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Dalmia, A., & Sia, S. (2021). *Clustering with UMAP: Why and How Connectivity Matters* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2108.05525

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems*, *3*(1), 95–98. https://doi.org/10.1016/j.cels.2016.07.002

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Gendrel, A.-V., Lippman, Z., Martienssen, R., & Colot, V. (2005). Profiling histone modification patterns in plants using genomic tiling microarrays. *Nature Methods*, *2*(3), 213–218. https://doi.org/10.1038/nmeth0305-213

Ghurye, J., Pop, M., Koren, S., Bickhart, D., & Chin, C.-S. (2017). Scaffolding of long read assemblies using long range contact information. *BMC Genomics*, *18*(1), 527. https://doi.org/10.1186/s12864-017-3879-z

Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A. M., & Koren, S. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Computational Biology*, *15*(8), e1007273. https://doi.org/10.1371/journal.pcbi.1007273

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Guppy protocol. (2018, November 16). Oxford Nanopore Technologies. https://nanoporetech.com/document/Guppy-protocol

Haas, B. J. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, *31*(19), 5654–5666. https://doi.org/10.1093/nar/gkg770

Hanghøj, K., Moltke, I., Andersen, P. A., Manica, A., & Korneliussen, T. S. (2019). Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *GigaScience*, *8*(5), giz034. https://doi.org/10.1093/gigascience/giz034

Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S., & Delaneau, O. (2023). Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nature Genetics*, *55*(7), 1243–1249. https://doi.org/10.1038/s41588-023-01415-w

Hou, H., Pedersen, B., & Quinlan, A. (2021). Balancing efficient analysis and storage of quantitative genomics data with the D4 format and d4tools. *Nature Computational Science*, *1*(6), 441–447. https://doi.org/10.1038/s43588-021-00085-0

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, *132*(2), 583–589. https://doi.org/10.1093/genetics/132.2.583

Huelsenbeck, J.P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17, 754-755. https://doi.org/10.1093/bioinformatics/17.8.754

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, *34*(8), 2115–2122. https://doi.org/10.1093/molbev/msx148

Keightley, P. D., & Jackson, B. C. (2018). Inferring the Probability of the Derived *vs.* the Ancestral Allelic State at a Polymorphic Site. *Genetics*, *209*(3), 897–906. https://doi.org/10.1534/genetics.118.301120

Kent, W. J. (2002). BLAT —The BLAST -Like Alignment Tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, *28*(24), 3211–3217. https://doi.org/10.1093/bioinformatics/bts611

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., & Hedges, S. B. (2022). TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, *39*(8), msac174. https://doi.org/10.1093/molbev/msac174

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, *9*(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1303.3997

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Li, Z., Baniaga, A. E., Sessa, E. B., Scascitelli, M., Graham, S. W., Rieseberg, L. H., & Barker, M. S. (2015). Early genome duplications in conifers and other seed plants. *Science Advances*, *1*(10), e1501084. https://doi.org/10.1126/sciadv.1501084

Llorens C., Futami R., Covelli L., Domínguez-Escribá L., Viu J.M., Tamarit D., Aguilar-Rodríguez J., Vicente-Ripolles M., Fuster G., Bernet G.P., Maumus F., Munoz-Pomer A., Sempere J.M., Latorre A. & Moya A. 2011. — The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Research 39 (Database issue): D70-74. https://doi.org/10.1093/nar/gkq1061

Lou, H., Song, L., Li, X., Zi, H., Chen, W., Gao, Y., Zheng, S., Fei, Z., Sun, X., & Wu, J. (2023). The Torreya grandis genome illuminates the origin and evolution of gymnosperm-specific sciadonic acid biosynthesis. *Nature Communications*, *14*(1), 1315. https://doi.org/10.1038/s41467-023-37038-2

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, *3*(29), 861. https://doi.org/10.21105/joss.00861

Meisner, J., Albrechtsen, A., & Hanghøj, K. (2021). Detecting selection in low-coverage high-throughput sequencing data using principal component analysis. *BMC Bioinformatics*, *22*(1), 470. https://doi.org/10.1186/s12859-021-04375-2
Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., & Sander, J. (2014). Density-Based Clustering Validation. *Proceedings of the 2014 SIAM International Conference on Data Mining*, 839–847. https://doi.org/10.1137/1.9781611973440.96
Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. *F1000Research*, *10*, 33. https://doi.org/10.12688/f1000research.29032.2
Ni, P., Huang, N., Nie, F., Zhang, J., Zhang, Z., Wu, B., Bai, L., Liu, W., Xiao, C.-L., Luo, F., & Wang, J. (2021). Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning. *Nature Communications*, *12*(1), 5976. https://doi.org/10.1038/s41467-021-26278-9
Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., … Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, *497*(7451), 579–584. https://doi.org/10.1038/nature12211
Open2C, Abdennur, N., Abraham, S., Fudenberg, G., Flyamer, I. M., Galitsyna, A. A., Goloborodko, A., Imakaev, M., Oksuz, B. A., Venev, S. V., & Xiao, Y. (2024). Cooltools: Enabling high-resolution Hi-C analysis in Python. *PLOS Computational Biology*, *20*(5), e1012067. https://doi.org/10.1371/journal.pcbi.1012067
Open2C, Abdennur, N., Fudenberg, G., Flyamer, I. M., Galitsyna, A. A., Goloborodko, A., Imakaev, M., & Venev, S. V. (2024). Pairtools: From sequencing data to chromosome contacts. *PLOS Computational Biology*, *20*(5), e1012164. https://doi.org/10.1371/journal.pcbi.1012164
Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197
Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: Quick coverage calculation for genomes and exomes. *Bioinformatics*, *34*(5), 867–868. https://doi.org/10.1093/bioinformatics/btx699
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, *12*(null), 2825–2830.
Quigley, S., Damas, J., Larkin, D. M., & Farré, M. (2023). syntenyPlotteR: A user-friendly R package to visualize genome synteny, ideal for both experienced and novice bioinformaticians. *Bioinformatics Advances*, *3*(1), vbad161. https://doi.org/10.1093/bioadv/vbad161
Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033
RStudio Team. (2020). *RStudio: Integrated development environment for R* [Manual]. RStudio, PBC. http://www.rstudio.com/
Schiffthaler, B., Van Zalen, E., Serrano, A. R., Street, N. R., & Delhomme, N. (2023). Seiðr: Efficient calculation of robust ensemble gene networks. *Heliyon*, *9*(6), e16811. https://doi.org/10.1016/j.heliyon.2023.e16811
Sikorskaite, S., Rajamäki, M.-L., Baniulis, D., Stanys, V., & Valkonen, J. P. (2013). Protocol: Optimised methodology for isolation of nuclei from leaves of species in the Solanaceae and Rosaceae families. *Plant Methods*, *9*(1), 31. https://doi.org/10.1186/1746-4811-9-31
Sonnhammer, E. L. L., & Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, *167*(1–2), GC1–GC10. https://doi.org/10.1016/0378-1119(95)00714-8
Stovner, E. B., & Sætrom, P. (2019). Epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics*, *35*(21), 4392–4393. https://doi.org/10.1093/bioinformatics/btz232
Stull, G. W., Qu, X.-J., Parins-Fukuchi, C., Yang, Y.-Y., Yang, J.-B., Yang, Z.-Y., Hu, Y., Ma, H., Soltis, P. S., Soltis, D. E., Li, D.-Z., Smith, S. A., & Yi, T.-S. (2021). Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. *Nature Plants*, *7*(8), 1015–1025. https://doi.org/10.1038/s41477-021-00964-4
Sundell, D., Mannapperuma, C., Netotea, S., Delhomme, N., Lin, Y., Sjödin, A., Van De Peer, Y., Jansson, S., Hvidsten, T. R., & Street, N. R. (2015). The Plant Genome Integrative Explorer Resource: PlantGen IE .org. *New Phytologist*, *208*(4), 1149–1156. https://doi.org/10.1111/nph.13557
Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J. C., Schnable, P. S., Lyons, E., & Lu, J. (2015). ALLMAPS: Robust scaffold ordering based on multiple maps. *Genome Biology*, *16*(1), 3. https://doi.org/10.1186/s13059-014-0573-1
Tarailo-Graovac M. & Chen N. 2009. — Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics Chapter 4: 4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s25
The International HapMap Consortium, Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., & Lander, E. S. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918. https://doi.org/10.1038/nature06250
Vassilieff H., Haddad S., Jamilloux V., Choisne N., Sharma V., Giraud D., Wan M., Serfraz S., Geering A.D.W., Teycheney P.-Y. & Maumus F. 2022. — CAULIFINDER: a pipeline for the automated detection and annotation of caulimovirid endogenous viral elements in plant genomes. Mobile DNA 13 (1): 31. https://doi.org/10.1186/s13100-022-00288-w
Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A Map of Recent Positive Selection in the Human Genome. *PLoS Biology*, *4*(3), e72. https://doi.org/10.1371/journal.pbio.0040072
Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358. https://doi.org/10.2307/2408641

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*, *20*(1), 129. https://doi.org/10.1186/s13059-019-1727-y

Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*(9), 1859–1875. https://doi.org/10.1093/bioinformatics/bti310

Zhang, M., Zhang, Y., Scheuring, C. F., Wu, C.-C., Dong, J. J., & Zhang, H.-B. (2012). Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. *Nature Protocols*, *7*(3), 467–478. https://doi.org/10.1038/nprot.2011.455

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), R137. https://doi.org/10.1186/gb-2008-9-9-r137

Zhang, Z. (2022). KaKs_Calculator 3.0: Calculating Selective Pressure on Coding and Non-Coding Sequences. *Genomics, Proteomics & Bioinformatics*, *20*(3), 536–540. https://doi.org/10.1016/j.gpb.2021.12.002