Supplementary Material Section 1: Sequencing Norway spruce

1.1 Genome size estimations

Flow cytometry was used to estimate the DNA C-values of Norway spruce, and two related species. Nuclei were isolated and purified from fresh intact young needles of Norway spruce (clone Z4006) and from 23 other unrelated *P. abies* clones of Swedish origin from a provenance trial in Sävar, Sweden. Samples were also prepared from *Abies sibirica* (Umeå University Campus, location: 63°49´08.93´´N/20°18´40.95´´E) and *Taxus baccata* (Bergianska Gardens Stockholm. (originally from Ramlösa, 1951), location: Parken 36: 5, named f. columna-suecica). All samples were chopped simultaneously with two internal standards using a razor blade according to Obermayer & Greilhuber[1] (1999). Two of the internal standards recommended for plant DNA flow cytometry (*Pisum sativum* L. cv. Ctirad, 2C= 9·09 pg, and *Allium cepa* L. cv. Ailsa Craig, 2C= 33.55 pg) were chosen as optimal for this study. The samples were stained with propidium iodide, which intercalates into double-stranded DNA. Measurements of DNA content were performed with a PAII flow cytometer (Partec, Münster), using a 20 mW argon ion laser light source (488 nm wavelength) with an RG 590 longpass filter. Each sample was analyzed three times, with at least 5000 nuclei assayed per run. The genome size was determined from the C-value according to the formula

genome size (Gbp) = (0.978) x 1C DNA-value (pg)

Supplementary Table 1.1. Genome sizes estimated by flow cytometry

Species	Accession	LAT/LONG	n	Population	1C DNA (pg±SD)	Genome size (Gbp)
Picea abies	clone Z4006	63º18´N/15º59´E	30	-	20.02 ± 0.95	19.6 ± 0.9
	Block 6-001	67º18´N/23º18´E	1	Pajala	20.04 ± 0.55	19.6 ± 0.5
	Block 11-001	67º18´N/23º18´E	1	Pajala	20.01 ± 0.67	19.6 ± 0.7
	Block 8-002	67º15´N/23º03´E	1	Tärendö	20.03 ± 0.87	19.6 ± 0.9
	Block 3-012	65º54´N/21º50´E	1	Tärendö	20.03 ± 0.35	19.6 ± 0.3
	Block 6-012	65º54´N/21º50´E	1	Tärendö	20.03 ± 0.81	19.6 ± 0.8
	Block 8-012	65º54´N/21º50´E	1	Tärendö	20.04 ± 0.74	19.6 ± 0.7
	Block 1-17	65º40´N/21º37´E	1	Fällträsk	19.98 ± 0.56	19.5 ± 0.5
	Block 3-17	65º40´N/21º37´E	1	Fällträsk	20.02 ± 0.86	19.6 ± 0.8
	Block 11-17	65º40´N/21º37´E	1	Fällträsk	20.02 ± 0.59	19.6 ± 0.6
	Block 12-17	65º40´N/21º37´E	1	Fällträsk	20.01 ± 0.90	19.6 ± 0.9
	Block 1-037	63º31´N/17º50´E	1	Bredbyn Långsele	20.05 ± 0.24	19.6 ± 0.2
	Block 3-037	63º31´N/´17º50´E	1	Bredbyn Långsele	20.00 ± 0.58	19.6 ± 0.6

	Block 5-037	63°31´N/´17°50´E	1	Bredbyn Långsele	19.98 ± 0.98	19.5 ± 1.0
	Block 3-059	61°12´N/´12°41´E	1	Åmots. Förv.	20.06 ± 0.77	19.6 ± 0.8
	Block 7-059	61°12´N/´12°41´E	1	Åmots. Förv.	20.01 ± 0.69	19.6 ± 0.7
	Block 4-71	58°00´N/15°18´E	1	Åsbo	20.01 ± 0.78	19.6 ± 0.8
	Block 6-71	58°00´N/15°18´E	1	Åsbo	20.03 ± 0.78	19.6 ± 0.8
	Block 7-71	58°00´N/15°18´E	1	Åsbo	20.01 ± 0.89	19.6 ± 0.9
	Block 11-71	58°00´N/15°18´E	1	Åsbo	19.99 ± 0.87	19.6 ± 0.9
	Block 1-80	56°42´N/´12°22´E	1	Härryda	19.99 ± 0.67	19.6 ± 0.7
	Block 6-80	56°42´N/´12°22´E	1	Härryda	19.99 ± 0.76	19.6 ± 0.7
	Block 10-80	56°42´N/´12°22´E	1	Härryda	20.00 ± 0.87	19.6 ± 0.9
	Block 11-80	56°42´N/´12°22´E	1	Härryda	20.02 ± 0.49	19.6 ± 0.5
Abies sibirica	-		1	-	15.56 ± 0.84	15.2 ± 0.8
Taxus baccata	-	.	1	-	11.23 ± 0.76	11.0 ± 0.7

1.2 Shotgun sequencing and assembly

Haploid whole genome shotgun sequencing and assembly

Seeds from the *P. abies* clone Z4006 were stored at -20° C at Skogforsk Sävar, then soaked in water overnight, and manually dissected under a microscope to free the haploid megagametophyte tissue. Total DNA was isolated from a single, dissected megagametophyte (seed identification: 466) using a DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocols. Shotgun Illumina paired-end (PE) libraries with insert sizes of 180 bp, 300 bp and 625 bp were made from a total of 600 ng haploid DNA preparation (Supplementary Table 1.2), and sequenced on an Illumina HiSeq 2000. The reads were trimmed based on their quality scores, as follows: each read was cut off at the first base (from the 5' end) that had Q<10, and the remaining read was kept only if it had a length of >50 bp and >95% of the bases had Q>20. Overlapping reads from the 180 bp library were aligned to form longer single-end reads using a custom tool provided by CLCbio ("join-pairs"; CLCbio, Aarhus, Denmark). All quality filtered reads (760 Gbp; ~38X) were assembled using CLC Assembly Cell (Beta-4.0.6)(CLCbio, Aarhus, Denmark), on a 2TB RAM computer in ~5 days, with the scaffolding option turned off, thus utilizing paired read information, but disallowing scaffolding where the sequence between read pairs could not be fully resolved. Due to the low quantity of DNA used for library construction, the haploid libraries approached saturation, resulting in some read redundancy (in particular for the 625 bp library).

Diploid whole genome shotgun sequencing

Genomic DNA was extracted from diploid leaf tissue using a DNeasy Plant Maxi Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocols. DNA concentration was measured using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). Shotgun Illumina PE libraries with insert sizes of 180 bp, 300 bp and 625 bp were prepared, as were jumping libraries with insert sizes of 2.4 kbp, 4.4 kbp and 10.4 kbp (Supplementary Table 1.2). The 2.4 kbp and 4.4 kbp jumping libraries were produced using a 454-derived in-house circularization protocol (resulting in broad ranges of insert size, but very low fractions of PE reads) and sequenced on an Illumina HiSeq 2000. Multiple sequencing libraries were constructed for each insert size of jumping library to avoid library depletion due to PCR redundancy. The 10.4 kbp jumping libraries were constructed and Illumina sequenced by BGI (Beijing, China). The diploid PE reads were quality filtered in the same way as the haploid PE reads (described above). The diploid jumping reads were trimmed for internal linker sequences and reverse complemented prior to use.

Supplementary Table 1.2. Whole genome shotgun sequence data after quality filtering

	Data type	**Insert size (bp)**	**Avg read length (bp)**	**Coverage**
DNA from haploid tissue (seed ID: 466)	SE	NA	97	5X
	STITCH	NA	162	6X
	PE	180	2 x 100	1X
	PE	300	2 x 101	15X
	PE	625	2 x 101	11X
				38X
DNA from diploid tissue	SE	NA	96	9X
	STITCH	NA	159	14X
	PE	180	2 x 98	14X
	PE	300	2 x 92	11X
	PE	625	2 x 96	7X
				55X
	JUMP	2.4k	2 x 85	25X*
	JUMP	4.4k	2 x 75	10X*
	JUMP	10.4k	2 x 45	20X*
				55X*

SE: Single-end reads, STITCH: Overlapping paired-end reads aligned into single-end reads, PE: Paired-end reads, JUMP: Jumping libraries for long-insert paired reads, NA: Not applicable, k: thousand
* For the jumping libraries, physical span coverage (rather than read coverage) is given

Fosmid pool shotgun sequencing and assembly

Fosmid libraries were constructed from DNA preparations from diploid leaf tissue, grown on agar plates and assigned to 450 pools. Based on a small pilot study of 5 pools of varying sizes, it was decided to aim for approximately 1000 random clones per pool (theoretically equivavlent to 0.2% of the genome per pool) as a reasonable trade-off between time and cost on one hand, and assembly performance on the other (Supplementary Table 1.3). In addition, the final set of pools also included a few other later test pools of varying sizes (Supplementary Table 1.4).

All fosmid pools were constructed at Lucigen (Middleton, USA) or at the Children's Hospital Oakland Research Institute (Oakland, USA). Shotgun PE libraries of 300 bp were prepared from each pool, and sequenced on an Illumina HiSeq 2000. All reads were quality trimmed in the same way as the haploid PE reads (described above), and each pool was assembled separately using CLC Assembly Cell (v. Beta-4.0.6), with the scaffolding option enabled. Further scaffolding was performed on each pool individually with BESST[2] (https://github.com/ksahlin/BESST, https://pypi.python.org/pypi/BESST), using the mapped WGS diploid 625 bp and 2.4 kbp libraries, allowing a maximum of one mismatch per read in BWA[3]. BESST is a stand-alone scaffolder specifically developed and tailored for this genome project, capable of 1) efficient handling of large datasets, 2) accurate gap size estimations, 3) robust handling of variations in library insert size, and 4) flexible parameter settings for various scaffolding scenarios. Scaffolding increased the amount of data in >10 kbp scaffolds from 2.0 Gbp in the merged genome to 4.2 Gbp in the scaffolded genome. Due to the presence of repeats in the genome, there is a risk that some WGS reads could be mapped to a pool even if they were not originally from that part of the genome. Hence, special care was taken with the scaffolding parameters to avoid false joins; edges were disallowed at any ambiguity in the scaffolding graph. Additional jumping libraries were not included in this step, as overlong scaffolding gaps impaired the hierarchical genome assembly (below).

Mitochondrial, chloroplast, *Escherichia coli* and fosmid vector sequences were removed from the assemblies, both by read mappings prior to assembly and by blasting contigs post-assembly. The resulting N50 and the total assembly size of the nuclear inserts varied substantially across pools because: (1) varying levels of *E. coli* contamination (~1% - 50%) caused lower than expected fosmid read coverage, and (2) varying fractions of mitochondrial fosmid inserts (~5% - 50%) lowered the effective fosmid pool size. Hence, while the sequenced fosmid library represented almost 1X genome coverage in theory, the total assembly size represented ~0.5-fold genome coverage, some of which was still very fragmented due to low read coverage. The sequencing depth after quality filtering and contamination screening differed substantially between pools (avg: 54X, sd: 47X, excluding test pools), explaining to a large extent the variation in assembly success between pools (Supplementary Figure 1.1). Despite these technical limitations, this set of fosmid pools contained more data in scaffolds over 10 kbp in length than was achieved with the haploid WGS assembly, indicating the power of the method (Supplementary Table 1.5).

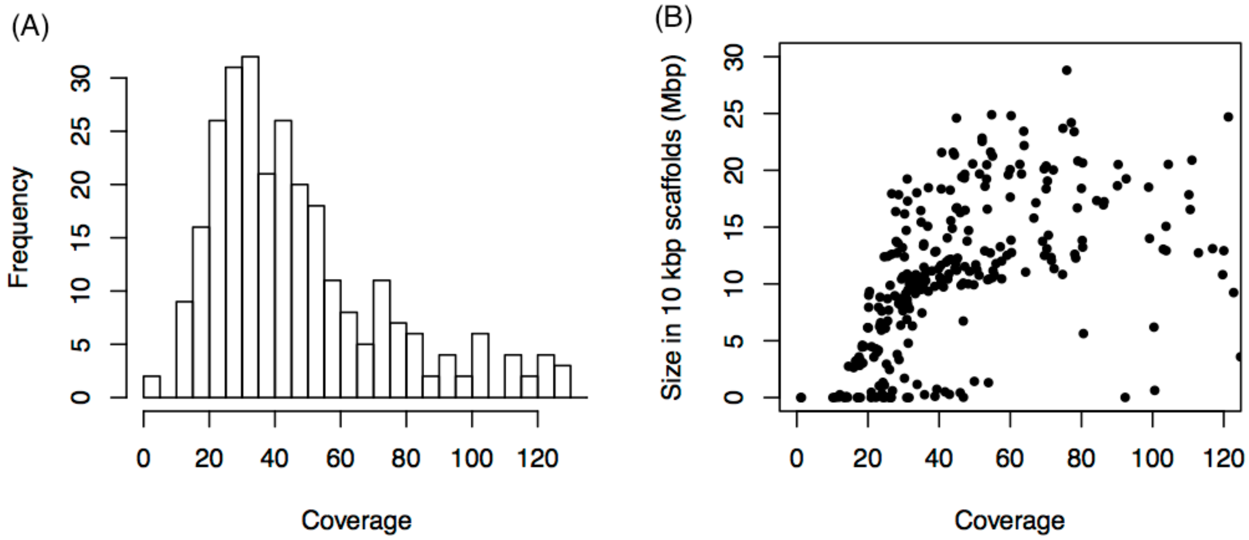
Supplementary Table 1.3. Pilot study of 5 fosmid pools of varying sizes

Fosmids per pool	Theoretical total size	Theoretical percentage of the genome	Percentage assembled in scaffolds >10 kbp
100	4 Mbp	0.02 %	85 %
500	20 Mbp	0.1 %	76 %
1000	40 Mbp	0.2 %	76 %
2500	100 Mbp	0.5 %	65 %
5000	200 Mbp	1.0 %	56 %

Supplementary Table 1.4. Target number of fosmids per pool

Fosmids per pool	Number of pools
100*	1
192*	2
384*	1
500*	1
768*	1
1000	425
2500	1
4000	1
5000	1
6000	16
Total	450

* The exact number of fosmid clones were counted before pooling



Supplementary Figure 1.1. Variation of read coverage across fosmid pools (A). Fosmid pool assembly size in scaffolds >10 kbp as a function of coverage (B). Data only shown for pools with 1000 fosmids per pool.

1.3 Hierarchical genome assembly

The seed storage tissue of conifers, the megagametophyte, is haploid, thereby representing favourable material for genome assembly since each locus is only represented by one allele. However, too little DNA could be extracted from a single *P. abies* megagametophyte for construction of all the sequencing libraries we required. Therefore, to enable assembly of the *P. abies* genome, we developed a strategy combining hierarchical assembly of fosmid pools with both haploid and diploid Whole Genome Shotgun (WGS) data (Supplementary Figure 1.2a). As existing assembly software proved inadequate for fosmid pool integration[4], a pipeline composed of several customized solutions was developed. The large datasets provided a particular challenge, and we were unable to find any overlay graph assembler capable of adequately handling our set of fosmid pool scaffolds. We opted instead for a strategy based on merging the fosmid pool scaffolds into a haploid WGS assembly, which required development of new stand-alone tools[2,5] that could efficiently handle datasets considerably larger than the human genome.

Assembly merging

Fosmid pool scaffolds (see above) were merged into the haploid WGS assembly using GAM-NGS[5]. The WGS assembly (a total of 9.8 Gbp before scaffolding) was used as the GAM-NGS master assembly, while the complete set of fosmid pool scaffolds of at least 1 kbp (a total of 6.7 Gbp) was offered to GAM-NGS for merging. In brief, diploid PE reads (55X genome coverage, Supplementary Table 1.2) were mapped to both the WGS assembly and the fosmid pool scaffolds, and syntenic blocks were formed based on uniquely mapping reads that were shared between the two data sets. Merging was performed using a semi-global Smith-Waterman alignment between the contigs and scaffolds. Fosmid pool scaffolds not merged by GAM-NGS and with no megablast hit of more than 95% over 30% length against the merged assembly (i.e. sequences recovered only in the fosmid pools; 1.1 Gbp) were also added to the assembly post merging. The merged assembly comprised 12.0 Gbp, with 1.9 Gbp of this being in scaffolds of at least 10 kbp (Supplementary Table 1.4). A substantial set of >1 kbp fosmid pool scaffolds (totalling 2.7 Gbp) could not be used reliably in the WGS assembly, probably due to ambiguities resulting from repeats and highly heterozygous alleles. The complete set of fosmid pool scaffolds, as well as the corresponding subsets from the merging process, can be accessed at the ConGenIE website (http://congenie.org).

Scaffolding with paired WGS reads

The 300 bp, 625 bp, 2.4 kbp, 4.4 kbp and 10.4 kbp libraries derived from diploid tissue (Supplementary Table 1.2), were mapped to the merged assembly using BWA[3], and scaffolded with BESST[2].

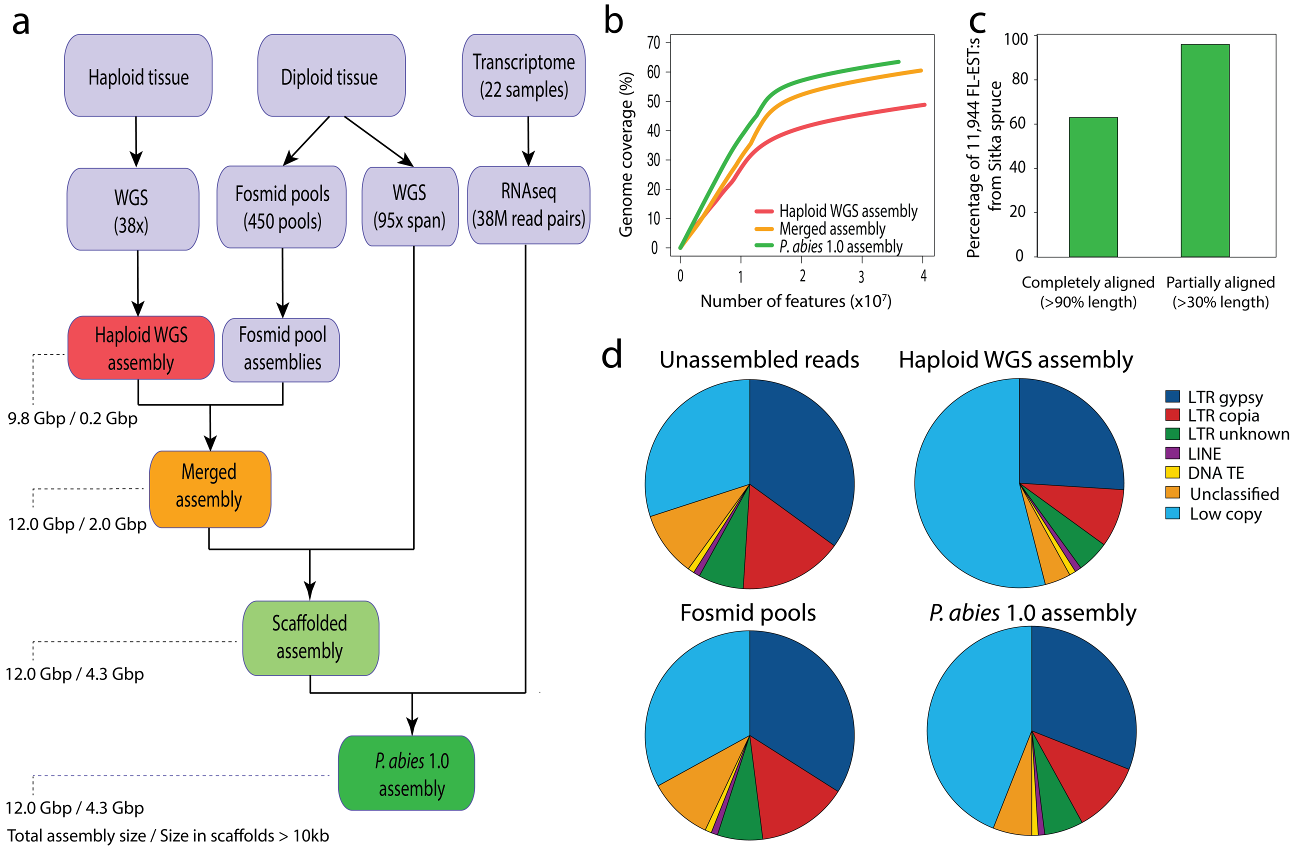
Scaffolding with paired RNA-Seq reads

Paired end Illumina RNA sequencing (RNA-Seq) data from 22 samples (Supplementary Table 2.1) were digitally normalized (Supplementary Information 2.10) resulting in a set of ~38 M read pairs that were mapped to the scaffolded assembly (above). The paired end information was used by BESST to link scaffolds containing successive exons of the same gene. This transcriptome-based scaffolding improved the contiguity of the gene-space, but did not provide any estimation of intron sizes. Ambiguous scaffolding signals were omitted from this step by only utilizing unique mapping read alignments. In total, 11528 new scaffolds (13811 new edges) were produced during the RNAseq scaffolding.

Contamination and redundancy screening

Initial analysis indicated that amounts of contaminating bacteria and fungi differed between samples (data not shown). The diploid WGS data could be used to effectively screen away contaminants from the final assembly, and thus all scaffolds with <1X coverage of mapped diploid reads were removed (~49 Mbp). Scaffolds representing the chloroplast were also removed (those with >99% sequence identity over >85% of their length), while 14 chloroplast-like scaffolds of lower identity were kept in the assembly and flagged as potential nuclear-chloroplast integration events. A further 10254 potential non-plant scaffolds (top megablast hits to non-plant sequences in nt), and scaffolds with aligned potential non-plant transcriptome sequences (top blastx hit to non-plant sequences in nr) were kept but flagged. Lastly, 12500 redundant scaffolds occurring as occasional artifacts of the GAM-NGS merging were removed (100% identical copies within the assembly; 29 Mbp in total), while a further 915 scaffolds were flagged as potentially redundant.

The resulting genome assembly**,** which was used for downstream annotation and analyses**,** was named "*P. abies* 1.0" (Supplementary Figure 1.2a; Supplementary Table 1.5).



Supplementary Figure 1.2. Hierarchical assembly strategy and characteristics of the *Picea abies* 1.0 asssembly. (**a**) Schematic representation of the strategy used to construct the *P. abies* 1.0 assembly. A set of 450 assembled fosmid pools were merged to obtain a whole genome shotgun (WGS) assembly derived from haploid tissue. This was followed by scaffolding using five insert libraries created from diploid tissue. Finally, the assembly was scaffolded using digitally normalised RNA sequencing read-pairs. (**b**) Feature Response Curve evaluation of the main steps of the assembly pipeline. The haploid WGS assembly (red line), merged assembly (yellow line) and *P. abies* 1.0 assembly (green line) showed progressively steeper FRC curves, demonstrating assembly improvements as a balanced measure of quality and contiguity. (**c**) Alignment of 11,944 Sitka spruce (*P. sitchensis*) full-length ESTs[8] to the P. abies 1.0 assembly. Only the best hit scaffold for each EST was considered. (**d**) Repeat content of various assemblies compared to an assembly-independent estimation. Genomic repeat content was estimated from unassembled reads (top left). Fosmid pool assemblies (bottom left) and the final *P. abies* 1.0 assembly (bottom right), were more representative of true genomic repeat content than the haploid WGS assembly (top left), which was depleted of high-copy repeats.

Supplementary Table 1.5. Assembly statistics (Gbp)

	Haploid WGS assembly	Fosmid pool assemblies	Merged assembly	*P. abies* 1.0
>200 bp scaffolds (total)	9.8	10.2	12.0	12.0
>1000 bp scaffolds	5.8	6.7	8.4	9.2
>5000 bp scaffolds	1.4	5.0	3.5	6.0
>10000 bp scaffolds	0.2	3.8	2.0	4.3

1.4 Assembly evaluation

Assembly validation presents a challenging problem, and contiguity based statistics, such as N50, can be poor predictors of quality[6]. Hence, the final assembly was validated in three different ways, as follows.

Assembly quality metrics (FRCurve)

FRCurve[7] is a tool designed to capture the trade-off between contiguity and correctness in de novo assembly projects. FRCurve analyses are based on a set of well-chosen features (derived from mapping reads back onto the assembly) representing potential problems in the assembled sequences. Here, we used FRCurve to evaluate each step in the assembly process described in the previous section, based on mapping information from the diploid 300 bp and 2.4 kbp libraries. FRCurve indicated clear assembly improvements (characterized by increasingly steeper curves) during both the merging and the scaffolding steps, supporting the validity of the assembly strategy (Supplementary Figure 1.2b).

Full-length protein coding genes

From a set of 13,197 random full-length cDNA sequences (including 5' and 3' UTRs) from *P. sitchensis*[8], 11,944 full-length coding regions (CDSs), were recovered and mapped to the *P. abies* 1.0 assembly using GMAP[9], with a specified requirement of 95% nucleotide identity. 7261 CDSs (63%) were aligned over at least 90% of their length within single scaffolds, while 11281 CDSs (96%) were aligned over at least 30% of their length (Supplementary Figure 1.2c). Without RNA-Seq scaffolding, 59% of the CDS:s were aligned over at least 90% of their length within single scaffolds.

Repeat content

A library of annotated sequences representing high-copy repeats was constructed from WGS long sequence reads (Supplementary Information S3). This repeat library was used to estimate the repeat content in the haploid WGS assembly, the 450 fosmid pool assemblies and the final draft genome, using RepeatMasker (v. open-3.3.0), and compared to the assembly-free repeat content estimated from 454 reads. Inferences from unassembled reads indicated that ca. 70 % of the *P. abies* genome was composed of high-copy repeats. Due to complexity-reduction, the fosmid pool assemblies (and thus also the *P.abies* 1.0 assembly) represented the genomic repeat content substantially better then the WGS assembly (Supplementary Figure 1.2d; Supplementary Table 1.6).

Supplementary Table 1.6. Repeat content (%)

	Unassembled reads	**Haploid WGS assembly**	**Fosmid pool assemblies**	***P. abies* 1.0**
LTR gypsy	35	26	34	31
LTR copia	16	9	14	11
LTR unknown	7	5	7	6
LINE	1	1	1	1
DNA TE	1	1	1	1
Unclassified repeats	10	4	10	6
Single/low copy	30	54	33	44

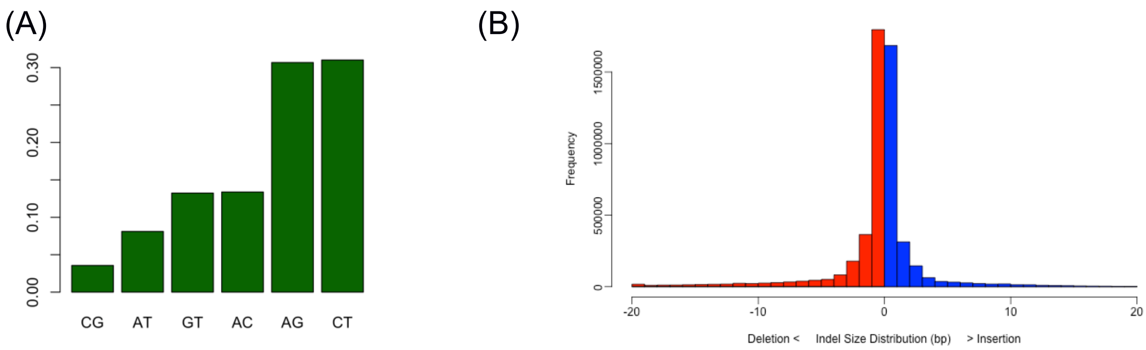
1.5 Heterozygosity and indel determination

To estimate variation in heterozygosity within the spruce genome, we mapped the PE reads originating from diploid tissue (Supplementary Table 1.2) to the master assembly using BWA with default settings[3]. We determined the pipeline that produced the best correspondence to expectations from direct examination of raw read alignments; it proved to be:

```
samtools mpileup -BAu -q 1 -d 250 … | bcftools view -Ncgv …
```

Results were filtered to include sites having coverage of non-duplicate-mapped reads **of** between 15× and 150× and an estimated allele frequency of between 0.25 and 0.75. Because of the large size of the read alignment files, BED files were constructed which divided the genome into 13 ~1 Gbp regions (samHeader2Bed.pl, https://github.com/douglasgscofield/bioinfo). Results for all regions were merged prior to further analysis. There were 9.79 Gbp of sites within the read coverage range, and 75.195 Mbp of heterozygous sites**,** giving a heterozygosity of 7.68/kbp. To check for potential biases introduced by low- or high-coverage sites, we restricted the range of coverage values to only those close to the expected input coverage value (range 55× to 65×, 1.67Gbp sites) and found an almost identical heterozygosity estimate (7.55/kbp). The ratio of transitions to transversions in heterozygous sites in *P. abies* was found to be 1.61:1 (Supplementary Figure 1.3A). A very low fraction (0.093%) of the heterozygous sites showed evidence of >2 alleles, indicating minimal bias introduced by mapping of reads from multiple repetitive regions to collapsed repetitive regions.

Indels were determined by direct examination of pileups from mapped reads using the same read selection criteria as described above for heterozygosity and a custom script that summarized short indel descriptions (pileVar.pl, https://github.com/douglasgscofield/bioinfo). As above, sites were treated as eligible if they had coverage between 15X and 150X, and indels were further filtered to include only those having a single indel operation among all indel-containing reads and an indel frequency of between 0.1 and 0.9. There were 5.32 M indels identified between alleles in the spruce genome (Supplementary Figure 1.3B). Indels are likely to result in reduced mapping rates for reads derived from the alternate haplotype to a much greater extent than reads lacking indels, leading to higher indel rates in sites with lower coverage. Indeed, indel rates were found to be greater in lower-coverage sites (0.38/kbp, 15X to 54X) than in intermediate-coverage (0.38/kbp, 55X to 65X) and higher-coverage (0.33/kbp, 66X to 150X) sites.



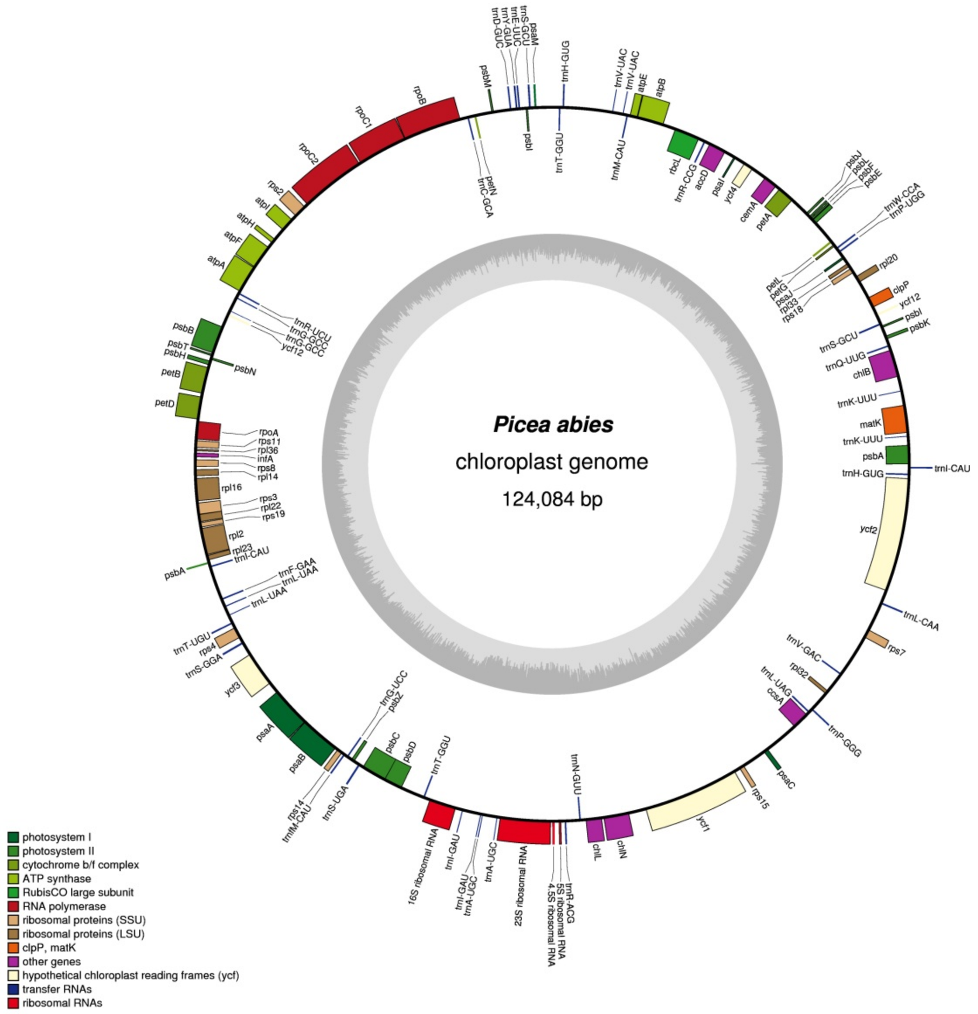
Supplementary Figure 1.3. Heterozygosity and indels **in** spruce. **(A)** Proportion of genotypes at heterozygous sites. **(B)** Distribution of insertions and deletions ≤20bp.

1.6 The chloroplast genome

The chloroplast genome (acc nb: HF937082) was assembled using Newbler 2.6 (default parameters) starting with approximately 580,000 454 Titanium reads (mode read length = 707 bp) from a whole genome shotgun library. Two chloroplast contigs were identified by blastn against the chloroplast genome sequence of *Picea sitchensis* (GenBank accession NC_011152). These contigs were approximately 53 kbp and 71 kbp in length, with a read coverage of 36X and 38X respectively, corresponding to ~1.1% chloroplast DNA in the WGS library.

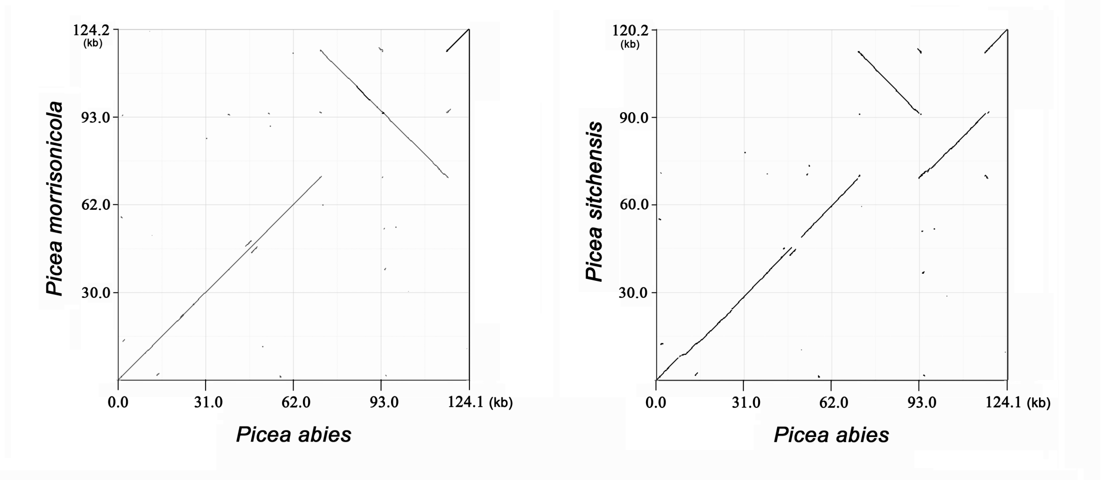
Polymerase chain reactions (PCR) and Sanger sequencing allowed us to verify the assemblies and join the contigs, yielding one circular molecule representing the finished chloroplast genome. Contig assemblies were validated by means of 11 PCR amplifications of randomly chosen regions and contig joins were verified by six PCRs, with primers designed using Primer Premier 5.0 (Premier Biosoft International, Palo Alto, CA, USA). Amplifications were each carried out in a total volume of 25μL containing 5–50 ng of genomic DNA, 5.0 pM of each primer, 0.25 mM of each dNTP, 2.0 mM $MgCl_2$ and 0.75 U Top*Taq* DNA Polymerase (Qiagen, Toronto, ON, Canada) in an Eppendorf thermocycler (Hamburg, Germany). Amplification conditions were 3 min at 95°C, followed by 36 cycles of 1 min at 94°C, 1 min at 52°C and 1 min 30 s at 72°C, with a final elongation of 10 min at 72°C. For validation of the contig assemblies, gel electrophoresis was used to confirm that the amplified fragment sizes matched the lengths of the corresponding regions of the assemblies in all cases. For the contig joins, PCR products were purified and sequenced on an ABI 3730 by BGI (Beijing, China) according to their internal protocols. Contigs were visually compared with the chromatograms and manually curated, filling the joined gaps to form a complete circular molecule.

The chloroplast genome was 124,084 bp in total length, with a GC content of 38.7%. Annotation with DOGMA[14] showed that the chloroplast genome contained 109 genes, including 72 protein genes, 4 ribosomal RNA genes and 33 tRNA genes (Supplementary Figure 1.4). In common with other Pinaceae, the inverted repeat region in *P. abies* was highly reduced[11], with a 440 bp IRA region separating the two single-copy regions. This reduction was validated by PCR amplification and Sanger sequencing of both copies of the inverted repeat (methodology described above).



Supplementary Figure 1.4. The *P. abies* chloroplast genome and its 109 genes, characterized by functional group**,** drawn using OGDraw[12].

Major rearrangements among spruce chloroplast genomes were detected by comparisons with the other two spruce chloroplast genomes published to date**,** those of *P. sitchensis* (NC_011152) and *P. morrisonicola* (NC_016069)[10,11]. The three chloroplast genomes were of similar sizes (*P. sitchensis* 120,176 bp*, P. morrisonicola* 124,168 bp) and gene content. Three different structural conformations were identified among the three species. Dot-plots made using zPicture[13] revealed that a 46 kbp inversion (from clpP to trnR-UCU) differentiated the chloroplast genomes of *P. abies* and *P. morrisonicola*, while a 23 kbp translocation (between clpP and trnS-GCU) plus a 23 kbp inversion (from psbI to trnR-UCU) distinguished *P. abies* from *P. sitchensis* (Supplementary Figure 1.5). The inversion between *P. abies* and *P. morrisonicola* was confirmed by means of four PCRs (conditions described previously). This is the first report of major genome rearrangements within the chloroplasts of the genus *Picea* and the result shows that structural variants should be taken into account in phylogenetic reconstructions**.**



Supplementary Figure 1.5. Dot-plots demonstrate a 46 kbp inversion, and a 23 kbp translocation plus a 23 kbp inversion, between *P. abies* and *P. morrisonicola* and *P. sitchensis*, respectively.

1.7 The mitochondrial genome

Putative mitochondrial contigs were apparent in GC%-vs-coverage plots as a discrete set of longer (≥ 5kbp) contigs having GC% ~ 42-48%, unlike the average 37.9% of the nuclear genome, and coverage ~20-100 times higher than the median coverage of the single-copy fraction of the nuclear genome assembly. We defined all contigs >1kbp in the final assembly with read coverage >20X over average and GC content >40% as being putatively mitochondrial. The putative mitochondrial genome had a GC content of ~44.7%, similar to values for other vascular plant mitochondria, and totalled 4.3 Mbp in size (which is an expected underestimation of the true size due to unidentified mitochondrial contigs in the assembly), suggesting a very large mitochondrial genome exceeded in size to date by only two angiosperm species[15]. Blast searches were performed for mitochondrial genes from the nonvascular plants *Physcomitrella patens*, *Megaceros aenigmaticus*, *Pleurozia purpurea*, *Phaeoceros laevis*, and *Marchantia polymorpha*, the angiosperms *Arabidopsis thaliana* and *Cucumis sativus*, and the only other gymnosperm mitochondrion available, *Cycas taitungensis*[16]. In total 41 complete or nearly-complete mitochondrial genes were found, including all 39 of those reported for *Cycas taitungensis*[16]. Hits were also found for 18 annotated ORFs of unknown function from *Arabidopsis* and 5 from *Physcomitrella*. As revealed by *de novo* scans[17], the scaffolds were rich in ORFs (median 68 aa), with 0.66 ORFs/kbp totaling 15% of the scaffold length. Such ORFs can arise during repeat-driven mitochondrial rearrangements and are often chimeric[18]; indeed, among these ORFs we found fragments of at least 28 of the 41 identified genes. There was no evidence of heterozygosity or sequence heteroplasmy between tissues; polymorphism within genes across pooled reads from two separate megagametophytes (data not shown) and diploid leaf tissue (mean $1.5\times10^{-3} \pm 0.2\times10^{-3}$/bp) was at or below the sequencing error rate.

1.8 References

1. Obermayer R, Greilhuber J (1999) Genome size in Chinese soybean accessions: stable or variable? Annals of Botany 84: 259-262.
2. Sahlin K, Vezzi F, Nystedt B, Lundeberg J, Arvestad L (2013) BESST - Scaffolding large fragmented assemblies efficiently. Submitted
3. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60
4. Zhang et al. (2012) The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490:49-54
5. Vicedomini R, Vezzi F, Scalabrin S, Arvestad L, Policriti A (2013) GAM-NGS: Genomic Assemblies Merger for Next Generation Sequencing. BMC Bioinformatics 14:S6.
6. Vezzi F, Narzisi G, Mishra B. (2012a) Feature-by-feature: evaluating de novo sequence assembly. PLoS One. 7:e31002
7. Vezzi F, Narzisi G, Mishra B. (2012b) Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons. PLoS One. 7:e52210
8. Ralph SG, Chun HJ, Kolosova N, Cooper D, Oddy C, Ritland CE, Kirkpatrick R, Moore R, Barber S, Holt RA, Jones SJ, Marra MA, Douglas CJ, Ritland K, Bohlmann J. (2008) A conifer genomics resource of 200,000 spruce (Picea spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (Picea sitchensis). BMC Genomics 9:484
9. Wu D.T. and Watanabe C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859-1875.
10. Cronn R, A Liston, M Parks, DS Gernandt, R Shen, T Mockler (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 36: e122.
11. Lin, C-P, J-P Huang, C-S Wu, C-Y Hsu, S-M Chaw (2012) Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. Genome Biol. Evol. 2:504–517
12. Lohse, M, O Drechsel, R Bock (2007) OrganellarGenomeDRAW (OGDRAW) - a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr. Genet. 52: 267-274
13. Ovcharenko I, GG Loots, RC Hardison, W Miller, L Stubbs (2004) zPicture: Dynamic Alignment and Visualization Tool for Analyzing Conservation Profiles Genome Research 14: 472-477
14. Wyman SK, RK Jansen, JL Boore (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252-3255
15. Sloan D, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer J, Taylor DR (2012) Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. PLoS Biol. 10:e1001241
16. Chaw SM, Shih AC, Wang D, Wu YW, Liu SM, Chou TY. (2008) The mitochondrial genome of the gymnosperm *Cycas taitungensis* contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. Mol Biol Evol. 25:603-15
17. Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 26:1107-15
18. Shedge V, Arrieta-Montiel M, Christensen AC, Mackenzie SA (2007) Plant mitochondrial recombination surveillance requires unusual RecA and MutS homologs. Plant Cell. 19:1251-64

Supplementary Material Section 2: The transcriptome and gene space

2.1 Tissue sampling

Tissues for RNA extraction were obtained from 22 samples that included needles, stems, and cones collected at different developmental stages and at different time-points during the 2010 growing season. Full sample details are given in Supplementary Table 2.1. Samples were collected from mature (40 years+) clonal copies growing at the Skogforsk research station, Sävar, Sweden (see above for details).
Brief details for some sample types are given:
Wood: For the wood samples, an approximately 5 x 10 x 3 cm (width x length x depth) section of the trunk of a mature clonal copy of Z4006 was removed at breast height using a hammer and chisel and immediately flash frozen in liquid nitrogen. For RNA extraction, the wood piece was slightly thawed to allow separation of the bark from the wood across the cambium. Xylem cells were defined by examination of cross sections under a microscope after nitroblue tetrazolium (NBT) staining. Tissue was scraped from the xylem and phloem sides and pooled for RNA extraction. On the phloem side, tissue was removed with a scalpel until the secondary phloem fibres were reached. This layer consisted of the cambial zone and primary phloem. On the xylem side, tissues were scraped with a scalpel until the layer of dead xylem (tracheid cells) was reached. This layer consisted of living xylem cells. Sample Z4006TR24 included early wood, while in sample Z4006TR25 late wood formation was observed but cell walls were not fully matured. Cell death could not be determined but was likely to have started in some of the tracheids. In both samples no dead tracheid/xylem cells were sampled, as assessed by microscopic observations.
Infected needles: Needles produced in the 2009 growing season were sampled. These had visible signs of fungal rust infection. Although the pathogen was not diagnosed, it was likely to have been *Chrysomyxa abietis* or *Lirula macrospora.*
Pineapple Galls: 10 galls from the 2009 growing season were sampled. Galls most likely resulted from infection by the adelgid blade spruce gall (*Adelges laricis*). The galls no longer contained live adelgid, however after the live adelgids leave the galls, the gall structures can remain on the tree for a few years.
Male cones: Immature male cones between 0.7 and 1 cm in length were sampled from multiple branches of a clonal copy of Z4006.
Female cone: An immature female cone of 5 cm in length was sampled from a clonal copy of *P. abies* clone Z3001 growing at the same location as the Z4006 clones. This was the only sample not collected from the Z4006 clone also used to produce the *P.abies* 1.0 assembly (see above).
Girdled twig: A single branch was removed and girdled (a strip of bark was removed from the perimeter of the branch) approximately 1 m from the apical end of the branch. Samples from shoots produced during the 2010 growing season were collected six days after girdling, and frozen in liquid N_2. RNA was subsequently isolated from needles (Z4006TR18) and stem tissue (Z4006TR19) harvested from multiple side shoots of the main stem. The shoots were fully matured and were approximately 10 cm long.

The needles and stems for samples Z4006TR11 and Z4006TR12 were collected from the same stems. All samples except Z3001TR10 represent multiple, pooled biological samples.

Supplementary Table 2.1 Overview of samples collected for Illumina short read mRNA sequencing and used to create pooled, normalised mRNA and total RNA cDNA libraries for 454 EST sequencing. %Q20 plus indicates the percentage of reads with a quality score >20.

Sample	Tissue	Sample description	Sample date	Total Reads	Total bp	% Q20+	GC %
Z4006TR01	male cones	Immature male cones	May 27, 2010	24,827,234	4,965,446,800	97.81%	45.91%
Z4006TR02	shoots	Vegetative shoots produced in 2010	May 27, 2010	23,797,169	4,759,433,800	97.80%	45.92%
Z4006TR03	needles	Needles produced in 2009	May 27, 2010	24,188,191	4,837,638,200	97.55%	46.49%
Z4006TR04	needles	Needles produced in 2008	May 27, 2010	23,095,399	4,619,079,800	97.47%	46.31%
Z4006TR05	needles	Infected needles produced in 2009	May 25, 2010	23,504,948	4,700,989,600	97.94%	46.60%
Z4006TR07	shoots	Vegetative shoots produced in 2010	June 21, 2010	24,025,688	4,805,137,600	98.05%	46.00%
Z4006TR08	pineapple galls	Pineapple galls	June 21, 2010	23,345,334	4,669,066,800	97.54%	45.29%
Z4006TR09	buds	Buds, early season developing	August 12, 2010	24,594,322	4,918,864,400	97.55%	46.34%
Z3001TR10	female cone	Immature female cone	June 9, 2010	24,142,621	4,828,524,200	97.51%	45.90%
Z4006TR11	needles	Needles from vegetative shoots produced in 2010	August 12, 2010	24,537,590	4,907,518,000	95.48%	44.39%
Z4006TR12	stem	Stem from vegetative shoots produced in 2010	August 12, 2010	23,287,473	4,657,494,600	97.15%	45.34%
Z4006TR13	needles	Needles from vegetative shoots produced in 2010	September 7, 2010	24,077,054	4,815,410,800	94.81%	45.11%
Z4006TR15	buds	Buds, late season developed	September 7, 2010	23,846,759	4,769,351,800	97.13%	45.69%
Z4006TR16	needles	Needles from dried twig (2 days on bench)	September 9, 2010	23,178,302	4,635,660,400	97.09%	44.86%
Z4006TR18	needles	Needles from girdled twig (sampled 1 week after girdling)	September 13, 2010	24,061,272	4,812,254,400	98%	45.00%
Z4006TR19	stem	Stem from girdled twig (sampled 1 week after girdling)	September 13, 2010	21,961,993	4,392,398,600	94.66%	45.33%
Z4006TR20	needles	Early morning (05:30)(dawn) needles from 2010	September 13, 2010	23,174,557	4,634,911,400	93.78%	44.38%
Z4006TR21	needles	Mid-day (12:00) needles from 2010	September 13, 2010	23,465,880	4,693,176,000	97.08%	45.93%
Z4006TR22	needles	Late afternoon (19:30)(dusk) needles from 2010	September 13, 2010	23,309,640	4,661,928,000	97.29%	45.16%
Z4006TR23	needles	Night (23:30) needles from 2010	September 13, 2010	23,795,641	4,759,128,200	97.35%	45.15%
Z4006TR24	wood	Wood (phloem+cambium+xylem, early)	June 21, 2010	23,352,912	4,670,582,400	97.36%	45.41%
Z4006TR25	wood	Wood (phloem+cambium+xylem, late)	August 12, 2010	24,830,162	4,966,032,400	97.82%	44.60%

2.2 RNA extraction

Total RNA was extracted from 0.5 g tissue using the CTAB method[1] Precipitated RNA was further purified using an RNeasy Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. RNA concentration and purity were measured using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and its integrity was analysed on an Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbronn, Germany).

2.3 454 EST sequencing and assembly

Equal quantities of RNA from the 21 Z4006 samples detailed above (*i.e.* excluding the single female cone sample) were pooled and used to produce two normalised cDNA libraries, one from polyA-selected RNA and one from total RNA. Normalised libraries were produced by Evrogen (Moscow, Russia) using a modified template-switching (SMART) approach (Cat # CS010) and normalised cDNA was sequenced using 454 pyrosequencing by SciLifeLab, (Science for Life Laboratory, Stockholm, Sweden).

Sequences were assembled using GS De Novo Assembler v2.6 (A.K.A Newbler, Roche 454 Life Sciences, Brandford, USA) using the settings `-sio -cpu 8 -urt -minlen 45 -tr -het -cdna –vt`. The file specified for vector trimming included adapter sequences used during library normalisation. Assembly metrics for the two libraries are shown in Supplementary Table 2.2. Assemblies were corrected using FrameDP [2] due to the tendency of 454 transcript assemblies to contain homopolymer errors that may introduce erroneous frameshifts.

Supplementary Table 2.2 Summary metrics of EST assemblies of 454-sequenced normalised mRNA and totalRNA cDNA libraries created from a pool of 21 samples. Assemblies were performed using the 'Newbler' assembler.

	mRNA	total RNA
# input reads	1,731,386	1,717,974
Input data (Mbp)	458.36	509.07
# isogroups/isotigs	26,364/36,069	23,876/33,426
Mean isotigs per isogroup	1.4	1.4
Isotig N50 (bp)	1,455	1,233
% aligned reads	81.855	81.42

2.4 Illumina transcriptome sequencing and assembly

Total RNA preparations from each of the 22 samples (for the 21 Z4006 samples the same starting RNA was used to create the normalised pools described above) were sent to the Beijing Genome Institute (BGI, Shenzhen, China). Paired-end (2 x 100 bp) RNA Sequencing (RNASeq) data were generated using standard Illumina protocols and kits (TruSeq SBS KIT-HS v3, FC-401-3001; TruSeq PE Cluster Kit v3, PE-401-3001) and all sequencing was performed using the Illumina HiSeq 2000 platform. Briefly, the sequencing protocol involved DNase 1 digestion of total RNA, mRNA isolation by use of oligo(dT) beads, mRNA fragmentation, first and second strand cDNA synthesis, end-repair, A-tailing, bar-coded adapter ligation and PCR amplification. Sequencing libraries were quality checked using an Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbronn, Germany) before sequencing. Raw read data was quality control filtered using the BGI in-house filtering pipeline. An overview of raw sequencing read counts per sample is included in Supplementary Table 2.1.

De novo transcriptome assembly was performed using Trinity[3] (release 2012-06-08). Sequencing reads from all 21 Z4006 RNASeq libraries (*i.e.* not including the Z3001TR10 female cone sample) were combined (totalling 522,400,141 read pairs) and assembled using the settings `--seqType fq --min_kmer_cov 10 --group_pairs_distance 500 --path_reinforcement_distance=50 --bfly_opts --edge-thr=0.20`. Assembly metrics are given in Supplementary Table 2.3. A high `min_kmer` value was used to reduce noise in the assembly and to identify only transcripts that were relatively highly expressed. At lower `min_kmer` values, intron retention was observed, possibly from incomplete splicing of pre-mRNAs. The vast majority of examples of potential intron retention showed very low expression levels based on post alignment of the RNASeq reads. As such, increasing the `min_kmer` value prevented many such cases from being included in the initial inchworm assembly step. The primary drawback of this approach was that some lowly expressed transcripts were no longer assembled and other transcripts with low expression became more fragmented. Hereafter we refer to transcripts assembled using Trinity as 'Trinity transcripts'.

Supplementary Table 2.3 Summary statistics of Trinity transcripts assembled using non-normalised paired-end (2x100 bp) Illumina RNASeq data from 21 samples. ORF denotes Open Reading Frame and FPKM Fragments Per Kilobase per Million mapped reads. ORFs and FPKM values were obtained using support scripts included with Trinity.

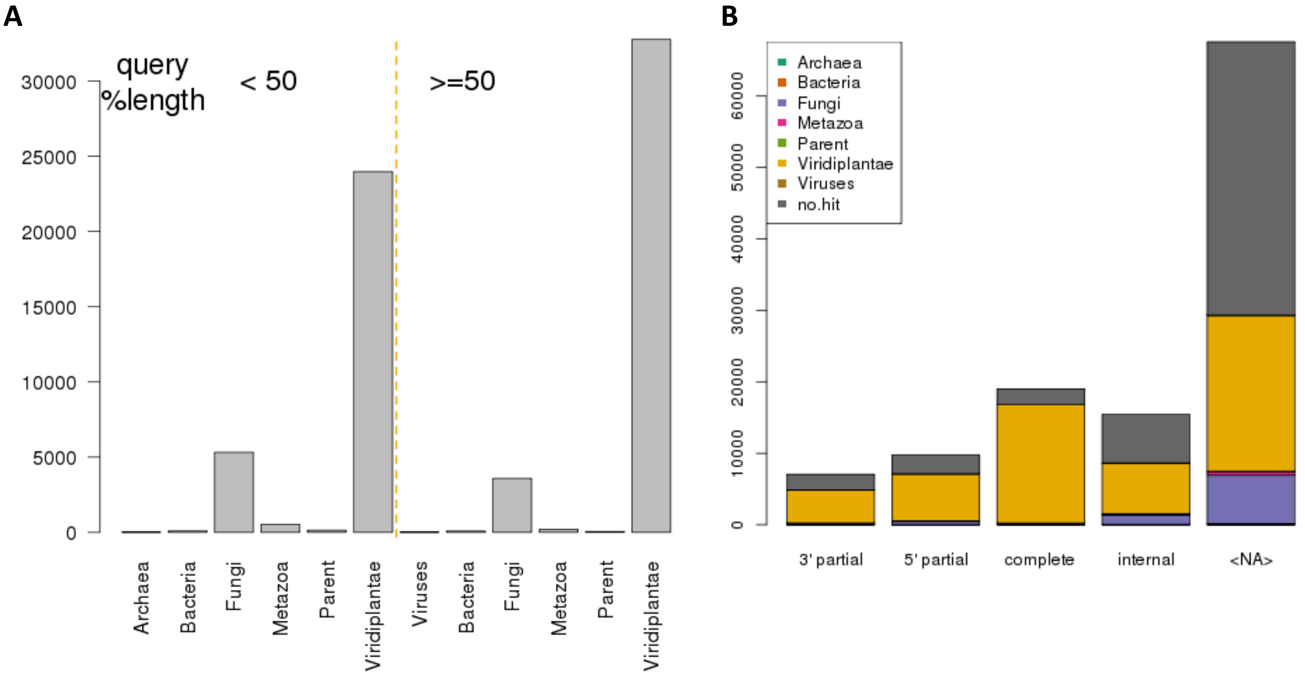
# components/clusters/sequences	77,189/88,098/118,799
Average sequences per cluster	1.35
Average component size	1.54
Mean sequence length (bp)	783
Mean ORF length	265
# Mean FPKM >1	80,633
# ORF >50 amino acids	51,197

2.4.1 Similarity searches and contaminant screening

As all samples used for RNA extraction were collected from mature trees in the field, some degree of contamination was expected. To identify potential contaminants, all Trinity transcripts were aligned to the NCBI (National Center for Biotechnology Information) nt (nucleotide) database (downloaded September 13th, 2012) using BLAST[4] and the best `blastn` (e-value <1^{e-5}) hit result stored. Taxonomic information corresponding to NCBI gi identifiers was retrieved from the NCBI Taxonomy database (downloaded September 13th, 2012). These results, in addition to GC (Guanine Cytosine) content, were used to identify transcripts as likely contaminants as well as to detect homology to existing plant sequences. The distribution of best BLAST hits is shown in Supplementary Figure 2.1. All assembled transcripts were aligned to the gene-containing subset of the genome (see Supplementary Table 2.7 below), using GMAP[5] (version 2012-07-20, `-B 5 -t 8 -K 50000 -w 50000 -f 2 -n 1`). Open Reading Frames (ORFs) were detected using the Trinity utility script `transcripts_to_best_scoring_ORFs.pl`. An ORF >50 amino acids was detected for 51,197 transcripts and ORFs were classified as 3' incomplete, 5' incomplete, internal or complete on the basis of stop and/or start codon presence or absence (Supplementary Figure 2.1). A summary of the classification counts is given in Supplementary Table 2.4.

Supplementary Table 2.4 Category classification count for assembled trinity transcripts. *Fungi* indicates all Trinity transcripts with a best BLAST hit to fungi; *No Hit* means all Trinity transcripts with no BLAST hit to sequences available in NCBI nt database; *Contaminant – others* means all Trinity transcripts with a best BLAST hit to an organism other than plants or fungi; *Plants* means all Trinity transcripts with a best BLAST hit to a plant species. The number of sequences in each category is subdivided into sequences with a GMAP alignment to the *P.abies* 1.0 genome assembly (*with GMAP*) and those without (*No GMAP*). The number of sequences with a GMAP alignment is further subdivided into sequences with an Open Reading Frame (*with ORF*) and those without (*No ORF*). The *No ORF* column thus reflects the number of putative long non-coding RNAs (lncRNA) in the trinity assembly (> 200bps).

Category	Components	Sequences	with GMAP	No GMAP	with ORF	No ORF
Fungi	7,478	8,267	526	7,741	159	367
No Hit	45,314	49,602	33,898	15,704	8,861	25,037
Contaminant-other	950	1,017	418	599	131	287
Plant	34,353	50,311	49,059	1,252	30,817	18,242



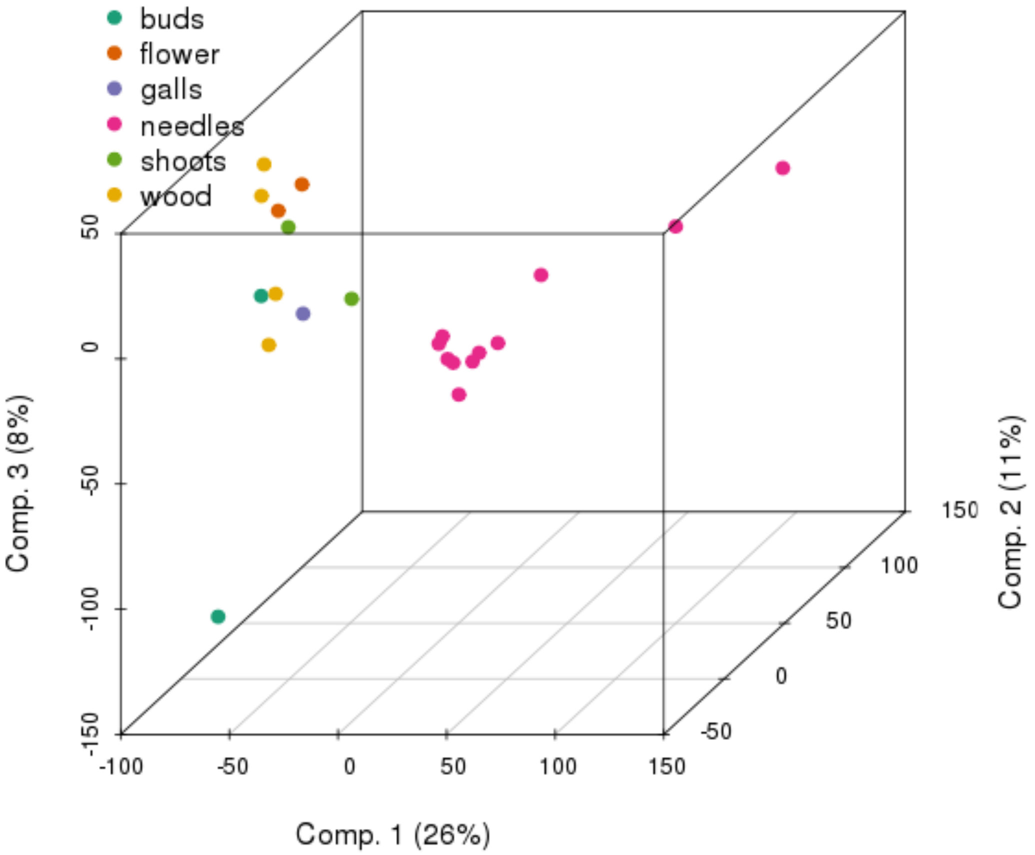
Supplementary Figure 2.1 A Distribution of best BLAST hit results of Trinity transcripts aligned against the NCBI nt database. **B** Counts of different Open Reading Frame (ORF) categories detected with the distribution of BLAST hit categories per ORF-type indicated. ORFs were classified as 3' incomplete, 5' incomplete, internal or complete on the basis of stop and/or start codon presence or absence. Parent indicates anything higher in the taxonomic tree (or closer to the root) than the selected divisions.

In addition, the Trinity transcripts were aligned using BLAST (blastn, e-value <1^{e-5}) against the assembled *P. abies* chloroplast sequence (see above) and flagged as putative chloroplast transcripts whenever >80% of the Trinity transcript aligned with >85% identity (a total of 76 transcripts).

2.4.2 Trinity transcripts expression value (FPKM) calculation

Using the Trinity `alignReads.pl` utility (Trinity version r2012-10-05), read data from the 22 samples were aligned to the Trinity transcripts using Bowtie[6] (version 0.12.8). The Trinity `runRSEM.pl` utility script, a wrapper around RSEM[7] was used to perform multi-mapping alignments and estimate expression values. The obtained FPKM (Fragments Per Kilobase per Million mapped reads) values so obtained were used in all analyses using gene expression values.

To validate the biological relevance of expression data generated based on the Trinity transcript assembly, a Principal Component Analysis (PCA) was performed, which showed that as expected samples clustered primarily according to tissue type (Supplementary Figure 2.2).



Supplementary Figure 2.2 Principal Component Analysis performed using the Trinity transcript Fragments Per Kilobase per Million mapped reads (FPKM) expression values. With the exception of an outlier bud sample, all samples group primarily by tissue type.

2.4.3 Long non-coding RNA classification

Long non-coding RNAs (lncRNA) were classified on the basis of having no detectable ORF >50 amino acids in length (identified using the Trinity `transcripts_to_best_scoring_ORFs.pl` script), no ORF occurring by chance at a 1 % certainty threshold, no evidence of being a Transposable Element (TE), not being a chloroplast-derived transcript, having no BLAST hit to the *ab initio* gene prediction set (see below), having a GMAP alignment to the genome (to screen out potential contaminants) and being intergenic between loci predicted *ab initio* to be protein coding gene (*i.e.* we do not here consider potential intronic lncRNAs). The probability of having a stop codon by chance given the estimated % GC of the genome (~40 %) is assumed to be

UAG + UGA + UAA
(0.3 x 0.3 x 0.2) x 2 + (0.3^3) = 0.063

Then, assuming a random selection of the codon (no bias) and the probability of a binomial distribution of 0 stop codon occurrences, there is still a 5 % chance that a sequence of 47 amino acids will occur randomly. A 70 amino acid sequence has a 1% chance of occurring randomly.

In total, 22,868 lncRNA transcripts were found using these criteria. 13,011 of these lncRNA candidates had no homology to any sequences available in the NCBI nt and protein databases (as identified using `blastn` and `blastp` searches). There were 9,707 candidate lncRNAs with identified homology (`blastn` e-value <1^{e-5}, >50 % coverage by aligned HSPs). An overview of the candidates identified is provided in Supplementary Table 2.7.

2.4.4 Assessing *de novo* transcriptome contiguity

In order to estimate the degree of contiguity and the extent to which full-length transcripts were reconstructed in the Newbler (454) and Trinity (Illumina) *de novo* transcriptome assemblies, we compared them to a set of *P. sitchensis* full-length ESTs[8]. Each set of *de novo* assembled transcripts was aligned to the *P. sitchensis* ESTs using BLAST (`blastn`, e-value 1^{e-5}). Alignments were filtered to remove redundancy and the cumulative HSP coverage of *P. sitchensis* ESTs by the *de novo* transcripts was calculated. In the Trinity transcript set, 20 % (1791 of 8968) of *P. sitchensis* ESTs with aligned Trinity transcripts had >1 transcript non-redundantly accounting for the aligned coverage. In comparison, 16 % (723 of 4669) of *P. sitchensis* ESTs were covered by >1 Newbler transcript, suggesting that the Newbler transcriptome assembly resulted in a higher percentage of full-length transcripts being assembled but with significantly lower total representation of the transcriptome.

2.4.5 Estimating gene number using *de novo* transcriptome assemblies

Employing the methodology presented in Rigault et al.[9] we used the *P. abies* Trinity and Newbler *de novo* transcriptome assemblies presented above and the below-detailed *P. abies* 454 Newbler transcriptome assembly from UC Davis (described below) to estimate gene number. The transcriptome datasets were filtered to remove redundancy due to the presence of assembled splice variants using USEARCH (v 6.0.307, using the setting `-id 0.85`, http://drive5.com/usearch/). Comparing the UC Davis assembly to the *P. abies* mRNA assembly described above resulted in an estimate of 39,811 genes; comparing the UC Davis assembly to the *P. abies* Trinity assembly detailed above resulted in an estimate of 42,907; comparing the two *P. abies* assemblies presented here resulted in an estimate of 72,075 genes. As the assemblies presented here are based on RNA collected across 22 samples they are likely to represent a greater proportion of the transcriptome than the UC Davis assembly, which is based on a single sample. Such transcript assembly based estimates are inherently fraught with complications, including the potential inclusion of non-collapsed splice variants, alleles and the presence of fragmented transcripts and so must be treated with caution.

2.5 Reference-guided transcript assembly

Sequencing reads from all 22 RNASeq libraries were individually aligned to the genome assembly using OSA[10] with the settings `-i 220 -s 40 -e TPM`. Read alignments from OSA were used as input to the Cufflinks[11,12] suite cufflinks tool (v2.0.2) to perform reference-guided transcript detection using the settings `--multi-read-correct --frag-bias-correct --library-type fr-unstranded --max-intron-length 50000`. The Cufflinks cuffmerge tool was subsequently run with default settings using the output generated by cufflinks to merge the 22 cufflinks predictions. As there were no spliced alignment programs available that could index a genome >4 Gbp, all reference alignments and reference-based assembly steps were performed using only scaffolds for which there was evidence that they contained gene-like fragments (see below)

2.6 *Ab initio* gene prediction

Ab initio predictions of coding loci were performed using AUGUSTUS[13] and EuGene[14]. A number of sources of extrinsic evidence were used to improve predictions, as described in the following section. All input evidence sources were aligned to the genome using GMAP. To allow a*b initio* predictions to be performed in a reasonable time, only scaffolds with evidence that they contained gene-like fragments (see below) plus all scaffolds >10 kbp (kilobase pairs) were used.

Based on contaminant-free, full-length cDNA ESTs identified from the Trinity assembled transcripts, *P. sitchensis*[8] and *P. glauca*[9] ESTs, a subset of 256 genes was carefully annotated. To enhance the sensitivity and specificity of splice site detection in EuGene, SpliceMachine[15] was trained based on the GMAP alignment of Trinity transcripts. AUGUSTUS and EuGene were trained with the manually annotated gene models to determine optimal parameters (*e.g.* donor and acceptor splice sites, coding potential). In addition to the Trinity transcript full-length cDNA evidence, publicly available conifer EST transcript assemblies (PUTs) from PlantGDB[16] and 454 transcript assemblies from eight gymnosperm species[17] (http://loblolly.ucdavis.edu/bipod/ftp/conifers_454_assembly/) were included to facilitate gene calling. To overcome the presence of long introns in some gene models, intron regions from GMAP alignments of full-length ESTs were provided as additional supporting evidence to AUGUSTUS.

The prevalence of TEs in the genome, and the correspondingly high chance of a TE insertion occurring within an intron, presents a major challenge for gene prediction. Due to the different handling of repeat masked regions by the gene prediction algorithms implemented in AUGUSTUS and EuGene, two complementary approaches were used to deal with TEs. In AUGUSTUS, the genome assembly was not masked for repeats before gene prediction. Consequently, a substantial number of TE loci were predicted *ab initio*. In order to identify those TEs and to differentiate them from other protein coding genes, an extensive filtering procedure was used. Firstly, predicted CDS (Coding DNA Sequences) were repeat-masked using RepeatMasker applied to the custom repeat library detailed below. The search engine `cross_match` was used to maximise sensitivity. Predicted loci were classified as TEs if >20 % of the sequence was masked by the repeat library. Secondly, genes showing obvious, identifiable TE-related Pfam[18] (Protein family) hits were excluded from the predicted genes. The Pfam domains taken into consideration are available at the ConGenIE ftp site (ftp://spruce.plantphys.umu.se:24). The genome sequence used as input for EuGene prediction was extensively masked. In addition to the custom repeat library, TEs identified by AUGUSTUS were included as an additional repeat library for RepeatMasker.

2.6.1 Overview of protein coding loci predicted *ab initio*

In total 70,968 protein coding loci were predicted. Of these, 28,354 were well supported (>70 % coverage) by either ESTs or UniProt proteins and were classified as High Confidence (HC) genes. The HC gene set comprised 7,681 predicted only by AUGUSTUS, 6,195 only by EuGene and 14,478 by both. The remaining genes were classified as Medium Confidence (MC) where EST and/or protein alignment coverage was between 30-70 % (33,039 genes) or Low Confidence (LC, 9,765 genes) where a locus was commonly predicted by both AUGUSTUS and EuGene but with poor supporting evidence (<30% alignment coverage). Excluding loci on putative mitochondrial scaffolds on located in the chloroplast there were 67054 loci in total consisting of 26,597 HC, 32,263 MC and 8,197 LC genes. The ConGenIE.org ftp site contains subsets of loci specific to the nuclear, chrloroplast and mitochondrial genomes.

The HC gene set was used to perform gene family and expression analyses. A summary of the HC genes is given in Supplementary Table 2.5. There were 19,123 and 3,885 loci in the MC and LC subsets, respectively, with matches to Pfam domains, and there was an over-representation of zinc-finger and Leucine-Rich Repeat (LRR) domains. As it is well known that gene prediction using draft genome assemblies is prone to over-estimations of gene numbers, especially the MC and LC gene sets should be used with sensible caution. As we further improve the genome assembly we fully expect the revise and improve the quality of the gene prediction and many of the current MC and LC loci will likely be found to be spurious. It is particularly likely that the MC and LC sets contain many potential pseudogenes. Indeed, one of the major future challenges facing the conifer community will be the dissection and real from pseudogene. Given the current level of fragmentation that remains in the *P.abies* 1.0 genome assembly, this is currently extremely challenging and as such, this provides a major motivation for further assembly improvement.

Supplementary Table 2.5 Summary statistics for the High Confidence gene models predicted *ab initio*.

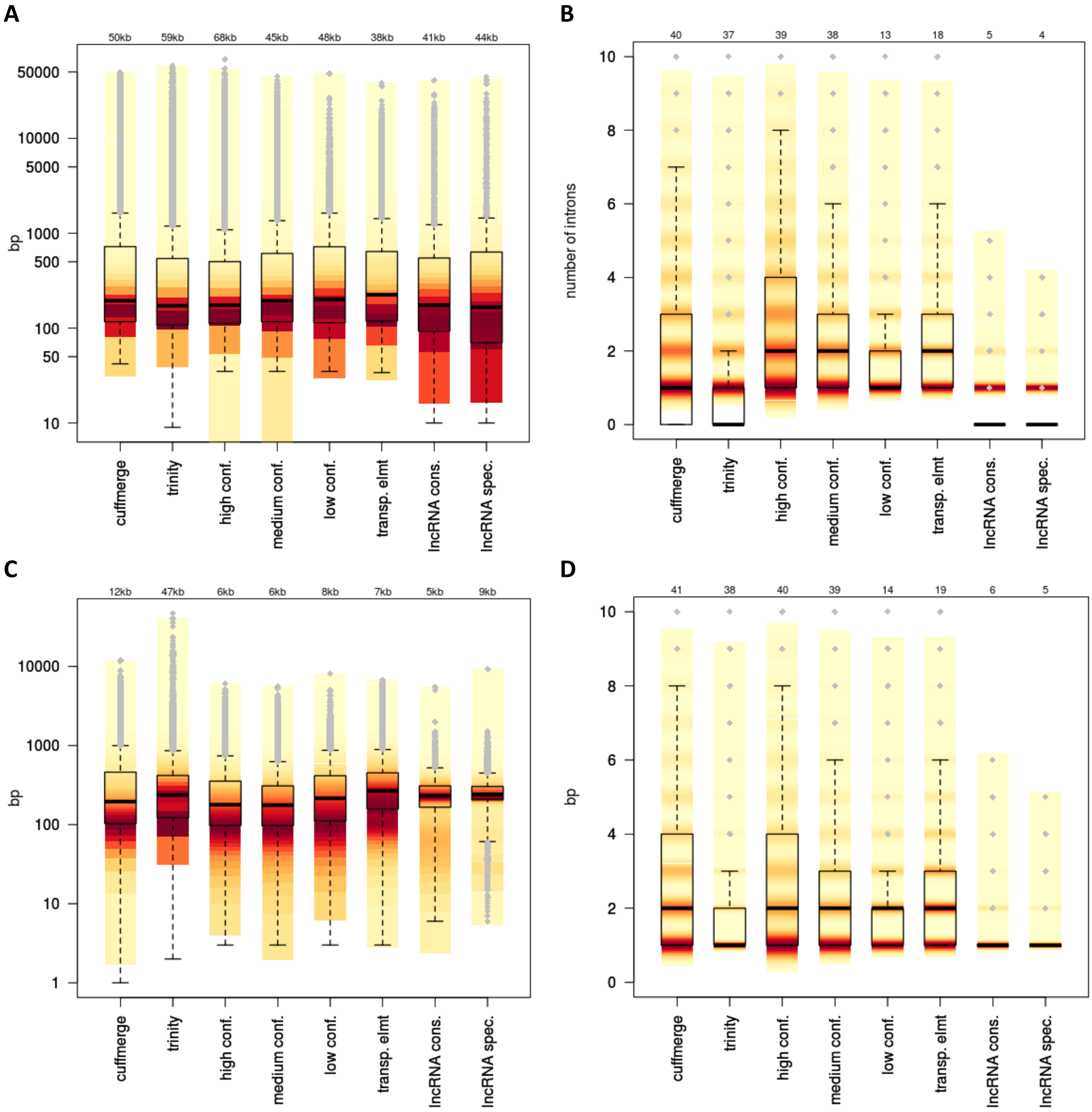
	High Confidence
Predicted genes	28,354
Mean total gene length (bp)	3,148
Mean CDS length (bp)	941
Mean exon length (bp)/number	312/3
Max/min exon length (bp)	6,069/3
Max/min intron length (bp)	68,269/34
Mean intron length	1,017
Single exon genes	11,573
# FPKM > 1	21,505
UniProt database support >50/70 %	12,737/8,342

2.6.2 Functional annotation of genes predicted *ab initio*

Functional annotation information of predicted genes was assigned on the basis of sequence similarity searches against a number of public databases. The gene description line from the best `blastp` hit (>30% sequence similarity and >50% alignment coverage) against the UniProt database was used as the gene description. Gene Ontology (GO) terms were obtained using the Blast2GO pipeline[19]. Motifs and domains of genes were determined by searching the InterPro database[20] including the Pfam, PRINTS, PROSITE, ProDom and SMART databases. Prediction of signal peptides was based on the SignalP results and transmembrane domains were predicted by TMHMM.

2.6.3 Structural characteristics of predicted *ab initio* and Cufflinks gene models

In both the *ab initio* HC gene set and the Cufflinks predicted loci, a notable feature was the presence of loci containing very long introns, with a maximum intron size of 68 kbp in the HC set. The maximum allowed intron size was set to 50 kbp in Cufflinks, and therefore no introns longer than this were reported. Intron and exon size and number distributions are shown in Supplementary Figure 2.3 (A-D). A comparison to other sequenced plant genomes is given in Figure 2b (in the main text) and Supplementary Table 2.6 (exons) and Supplementary Table 2.7 (introns). In contrast to introns, exon sizes were far more consistent among all species. The corresponding genomic feature sizes were extracted from GFF files retrieved from the Phytozome resource[21] for five angiosperms species (*Arabidopsis thaliana, Populus trichocarpa, Vitis vinifera, Oryza sativa* and *Zea mays*) and two basal plant species (*Selaginella moellendorffii* and *Physcomitrella patens*). The analysis was performed using R[22] and Bioconductor[23] and figures were created using the R LSD package.

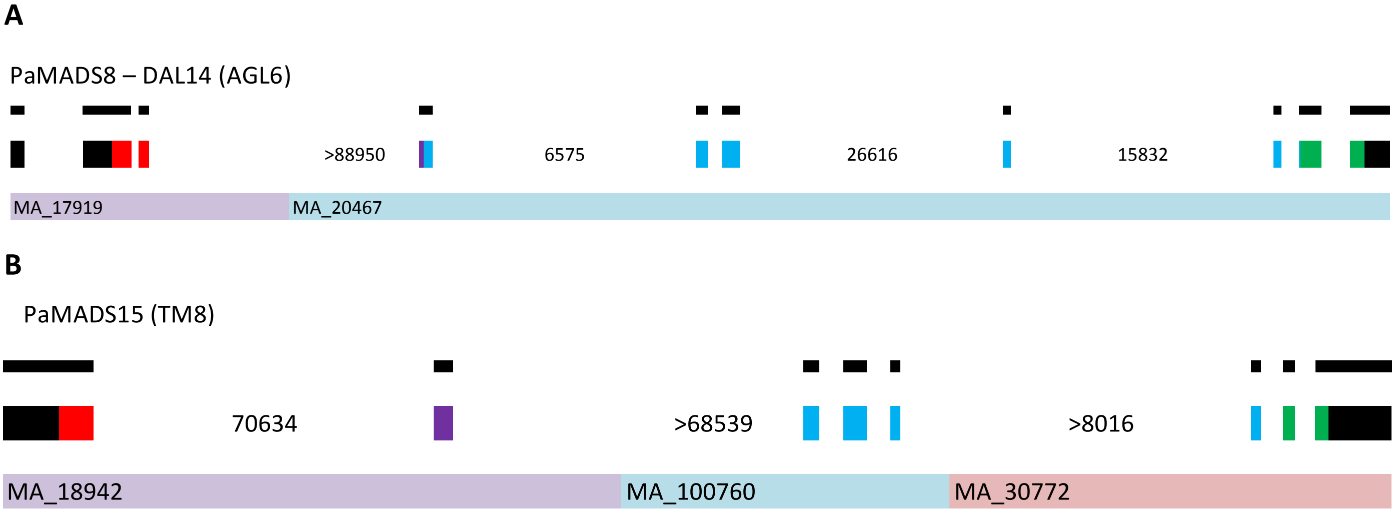


Supplementary Figure 2.3 Boxplot representations of **(A)** Intron size distribution in different transcript sets. The values above each box indicate maxima. Density is shaded from red (high) to yellow (low). **(B)** Intron number distribution. **(C)** Exon size distribution. (**D)** Intron number distribution.

GMAP alignment of the Trinity transcripts and other EST datasets indicated that a number of genes remain fragmented across >1 scaffold in the *P.abies* 1.0 genome assembly. These are cases where the RNA-Seq-based scaffolding approach was not able to resolve a unique path through scaffolds to reconstruct a contiguous series of scaffolds covering the complete gene structure. This is an expected feature of the conservative approach taken to avoid introducing assembly errors. Within the HC genes, 2,568 spanned a total of 3,142 scaffolding gaps created by the RNA-Seq scaffolding stage. Using GMAP alignments of the *P. sitchensis* full-length ESTs (detailed above) with the genome assembly, an estimated 33 % of genes (2,926 of 9,158 *P. sitchensis* ESTs had alignments fragmented across >1 scaffold) could be expected to remain fragmented across two or more scaffolds (assuming no unaccounted for bias). Similarly, alignment of the Trinity transcripts, which had a smaller mean length than that of the *P. sitchensis* full-length ESTs, resulted in an estimate of 27 % fragmentation. The same analysis also revealed that a significantly smaller set of genes are fragmented into >1 loci within the same genomic scaffold. Manual inspection of such cases revealed fragmentation to often result from the presence of long annotated repeats. Numerous cases of fragmented genes represent potentially extremely long introns and, accordingly, the intron size metrics reported here should be regarded as lower bounds. Examples of such fragmented gene structures identified from manual annotation of the MADS-box gene family are shown in Supplementary Figure 2.4.

Supplementary Table 2.6 Cross-species exon size statistics for seven sequenced plant genomes and the *P. abies* High Confidence *ab initio* prediction gene set. –ile indicates percentile, s.d. indicates standard deviation and mad indicates mean average difference. # > 95%-ile represents the number of introns with length greater than the 95%-ile size. The seven publicly available genomes include five angiosperms: *Arabidopsis thaliana, Populus trichocarpa, Vitis vinifera, Oryza sativa* and *Zea mays* and two basal plants: *Selaginella moellendorffii* and *Physcomitrella patens.*

	P. patens	*S. moellendorffii*	*P. abies*	*Z. mays*	*O. sativa*	*V. vinifera*	*P. trichocarpa*	*A. thaliana*
total	162052	85595	331223	460072	312497	174956	215909	124643
mean	241	312	295	277	363	246	296	213
s.d.	438	391	373	348	498	282	482	282
median	143	178	153	149	178	154	150	128
mad	110	148	125	114	156	111	116	86
min	4	3	1	2	3	1	1	1
max	40410	6069	13430	7911	15363	9023	15195	20532
90%-ile	514	720	685	637	856	520	609	440
95%-ile	702	1066	993	921	1281	742	975	696
99%-ile	1380	2060	1869	1760	2583	1428	2376	1428
# > 95%-ile	8115	4282	16565	23012	15650	8750	10810	6254
# > 99%-ile	1622	856	3315	4604	3135	1752	2160	1248



Supplementary Figure 2.4 Representative gene structures for two MADS-box genes where exons are fragmented across numerous genomic scaffolds. Narrow black lines indicate aligned Trinity transcripts and thicker coloured lines indicate domains where red represents the MADS-, purple the I-, blue the K- and green the C-domain, respectively. Longer coloured lines labelled with MA_XXXXX indicate genomic scaffold IDs. Where an intron spans a scaffold-scaffold boundary the minimal inferred intron size is indicated by a > sign. Other intron sizes are given.

Supplementary Table 2.7 Cross-species intron size statistics for seven sequenced plant genomes and the *P. abies* High Confidence *ab initio* prediction gene set. –ile indicates percentile, s.d. indicates standard deviation and mad indicates mean average difference. # > 95%-ile represents the number of introns with length greater than the 95%-ile size. The seven publicly available genomes include five angiosperms: *Arabidopsis thaliana, Populus trichocarpa, Vitis vinifera, Oryza sativa* and *Zea mays* and two basal plants: *Selaginella moellendorffii* and *Physcomitrella patens.*

	P. patens	*S. moellendorffii*	*P. abies*	*Z. mays*	*O. sativa*	*V. vinifera*	*P. trichocarpa*	*A. thaliana*
total	139017	102358	57241	267683	246158	135706	387059	175164
mean	309	111	1018	640	400	966	380	165
s.d.	741	518	2841	2444	640	2243	496	188
median	206	59	174	152	166	212	180	100
mad	122	13	117	107	129	191	141	30
min	10	10	35	1	5	9	1	8
max	42114	48860	68270	169080	18327	39916	10053	11602
90%-ile	536	172	2268	1170	948	2233	882	334
95%-ile	769	279	5040	2063	1367	4397	1236	461
99%-ile	1776	678	14761	8100	2739	11547	2408	846
# > 95%-ile	6957	5145	2864	13386	12309	6787	19371	8770
# > 99%-ile	1393	1024	573	2681	2462	1358	3875	1758

2.6.4 Using the Core Eukaryotic Genes Mapping Approach to assess gene-space coverage

Of the 248 core eukaryotic genes in the CEGMA[24] (Core Eukaryotic Genes Mapping Approach) the number of core genes present in the *P.abies* 1.0 genome assembly was assessed prior to gene prediction. In total, 86% (213 of 248) core genes were partially (>30 % coverage) identified and 53 % (132 of 248) were covered >80 %. The completeness of the HC gene prediction set was similarly assessed, and here 93 % (207 of 248) of core genes were identified with coverage >30 % and 40 % (98 of 248) with coverage >80 %

2.7 Gene prediction expression value (FPKM) calculation

The same approach was used as for the Trinity transcript expression value calculation (above). The complete gene prediction set of 70,968 loci was used and the FPKM values obtained were used in all the analyses requiring expression values from the gene prediction.

2.8 Gene family analysis

Genes in the HC gene set were used alongside protein sequences from seven publicly available genomes (five angiosperms: *Arabidopsis thaliana, Populus trichocarpa, Vitis vinifera, Oryza sativa* and *Zea mays* and two basal plants: *Selaginella moellendorffii* and *Physcomitrella patens*) to construct gene families on the basis of sequence similarity (all-all `blastp`, e-value 1^{e-5}) using TribeMCL[25] using an inflation value of 4 (`-I 4.0`). These seven genomes were selected to represent basal, monocotyledonous and dicotyledonous angiosperm plant species. The seven genome annotations were obtained form the PLAZA resource (REF) as these have been filtered to remove TEs, which would be detrimental if included in a gene family analysis. The results included 17,930 multi-gene families and 34,728 orphans (*i.e.* genes without no homology to other sequences in the dataset). Excluding orphan genes, *P. abies* had 6,615 gene families with a mean family size of 3.9 genes per family, which is slightly larger than in the other sequenced plant genomes used in this comparison (3.4 genes per family for the included angiosperm species and 2.8 for basal plant species). The higher family sizes are not unexpected given the above-detailed level of fragmentation that remains in the *ab initio* gene prediction set, as explained above, and the draft nature of the *P.abies* 1.0 genome assembly.

2.8.1 Identification of gene family expansion and contraction

To further understand gene family expansion or contraction in the *P. abies* in comparison to the seven other species considered, the standard deviation and mean gene family size were calculated for all gene families (excluding orphans and species specific families). The number of genes by species for each family was transformed into a matrix of z-scores to centre and normalize the data. The z-score was calculated as

$$z = \frac{x - \mu}{\sigma}$$

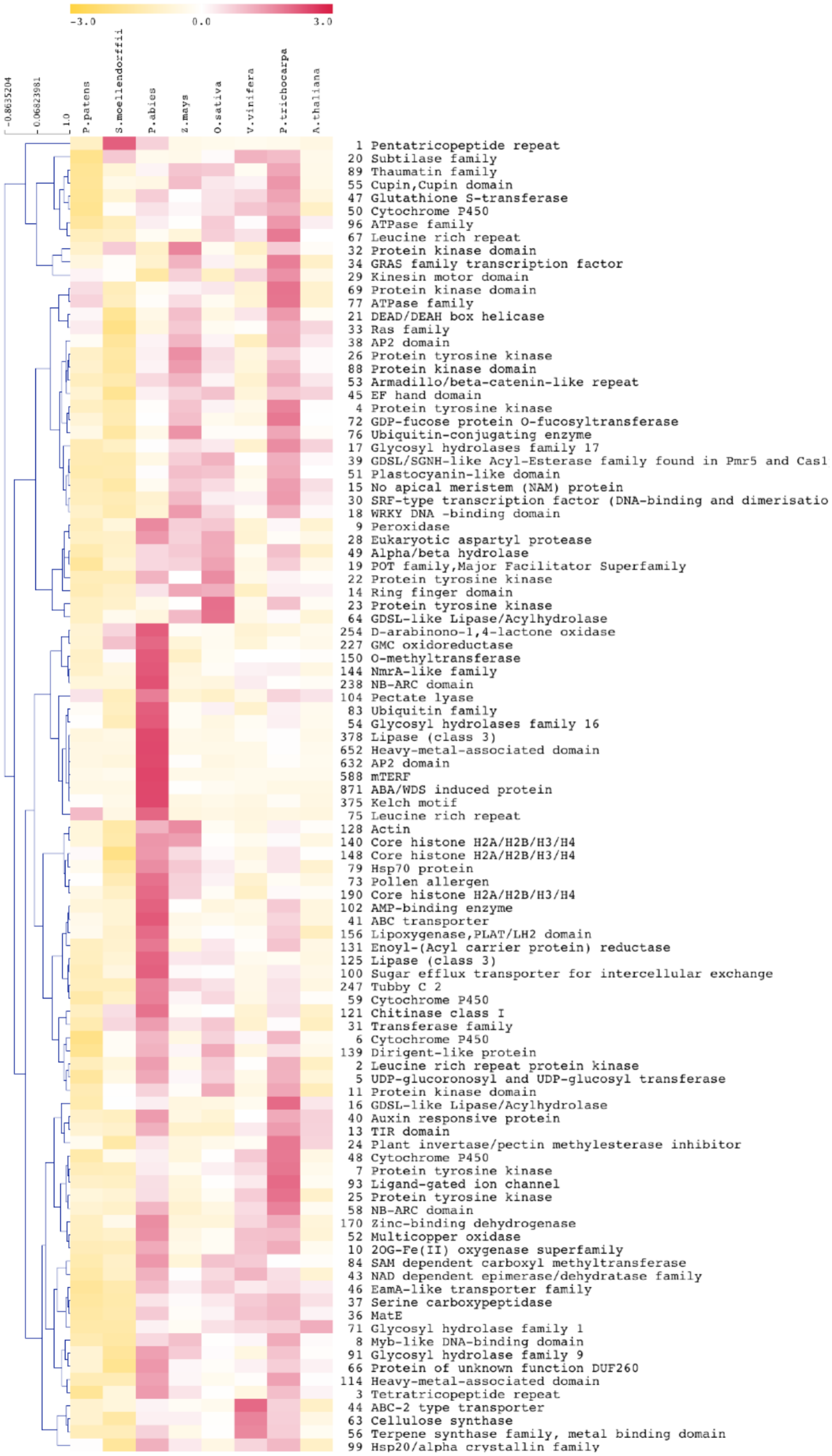
where x is the number of genes within a family for the given species, μ is the population mean and σ the population standard deviation.

The first 100 families with the largest gene family size in *P. abies* were selected. The z-score profile was hierarchically clustered (complete linkage clustering) using Pearson correlation as a distance measure. The biological function of each family was predicted based on sequence similarity to entries in the Pfam protein domain database where more than 30% of proteins in the family share the same protein domain (Supplementary Figure 2.5).

2.8.2 Phylogeny based gene family analysis and identification of gene families unique to *P. abies*

The phylogenetic profile and phylogenetic tree topology provided at the Phytozome resource[21] were used to reconstruct the most parsimonious series of gene gain and loss events. The DOLLOP program from the PHYLIP package[26] was used to determine the minimum gene set for ancestral nodes of the phylogenetic tree. The DOLLOP program is based on the Dollo parsimony principle, which assumes that genes arise exactly once on the evolutionary tree and can be lost independently in different evolutionary lineages[27]. The results are presented in Figure 2a (main text), showing that there are 6,615 *P. abies* gene families, 1021 of which were unique to *P. abies* and these may represent, for example, Norway spruce, conifer or gymnosperm specific families. There are 2,840 conserved orthologous families with genes present in all eight species, 95 of which are single copy gene families where each species has exactly one copy of the gene representing those families. A larger set of 6,451 ancestral orthologous families was identified where at least one species of the basal, gymnosperm and angiosperm species analysed contained a gene. The biological function of gene families unique to *P. abies* was explored on the basis of assigned GO information (see above).

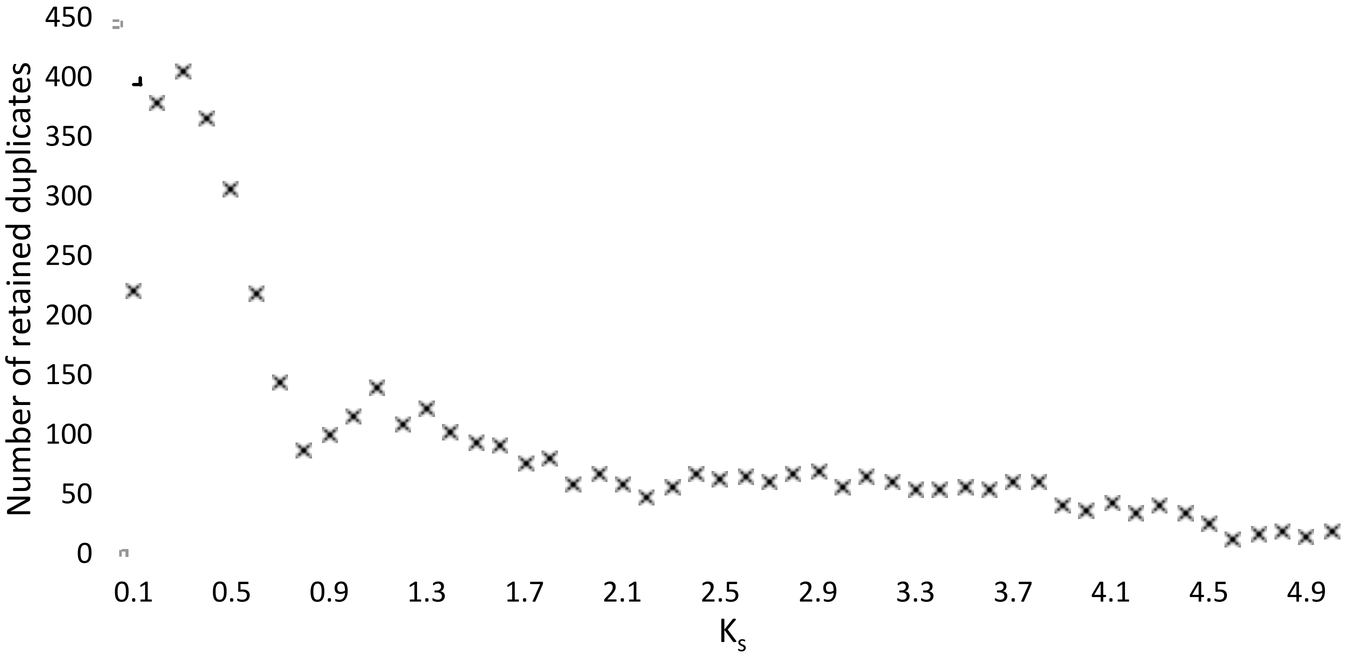
Gene families with a clear signature of expansion in *P. abies* reflected the unique aspects of conifer development and their response to environmental cues. Genes involved in lignin biosynthesis, xylem development and cellulose degradation (Families 54, 104, 150 and 652) represent a *P. abies* unique class of genes families and as such may contain genes explaining key biological differences between gymnosperms and angiosperms. On the basis of functional annotation inferred from Pfam domain presence, families 104, 238, 254 and 227 are likely involved in response to environmental and biological stresses such as interaction with fungi and insect attack, water deficit. Floral organ determining genes (Family 632) were functionally conserved among seed plants but showed low sequence similarity between *P. abies* and their *A. thaliana* homologous. The exact function of an identified family of leucine zipper genes (Family 588) remains unclear, but is of interest as there has been a clear expansion of this family in *P. abies*, which contained 44 genes in contrast to only a single gene in *A. thaliana*, *P. trichocarpa*, *V. vinifera* and *S moellendorffii*). Family information and functional analyses are available from the ConGenIE ftp site (http://congenie.org).



Supplementary Figure 2.5 Hierarchical clustering of the 100 largest *P. abies* gene families. For each gene family in each species the z-score was calculated to indicate the degree of family expansion/contraction and the resultant z-score matrix was clustered using Pearson correlation as a distance measure. Each row represents a gene family with the numerical gene family ID and corresponding functional description based on presence of a Pfam domain shown. Colour indicates z-score with yellow indicating contraction and red expansion.

2.9 Construction of empirical K_s age distributions

An all-against-all protein sequence similarity search was performed (blasp, E-value cutoff of 1^{e-10}). Gene families were then built using Markov Clustering[25] using the mclblastline pipeline (v10-201, micans.org/mcl). For each gene family, a protein alignment was constructed using MUSCLE[28] (v3.8.31). This alignment was used as a guide for aligning the DNA sequences of gene family pairs. Only gene pairs with a gap-stripped alignment length >100 amino acids were considered for further analyses. K_s estimates were obtained through maximum likelihood estimation (MLE) using the CODEML program[29] from the PAML package[30] (v4.4c). Codon frequencies were calculated based on the average nucleotide frequencies at the three codon positions (F3x4), and a constant K_a/K_s (reflecting selection pressure) was assumed for every pairwise comparison (codon model 0), because a single pair of sequences does not generally provide sufficient information to detect variability in selection pressure. For each pairwise comparison, K_s estimation was repeated five times to avoid inconsistent estimates due to MLE entrapment in local maxima. Only K_s estimates <5 were considered. Gene families were subdivided into subfamilies for which K_s estimates between genes did not exceed a value of 5. To correct for the redundancy of K_s values (a gene family of n members produces n[n – 1]/2 pairwise K_s estimates for n − 1 retained duplication events), an average linkage clustering approach was used[31]. Briefly, for each gene family, a tentative phylogenetic tree was constructed by average linkage hierarchical clustering, using K_s as a distance measure. For each split in the resulting tree, corresponding to a duplication event, all K_s estimates between the two child clades were added to the K_s distribution with a weight 1/m, so that the weights of all K_s estimates for a single duplication event sum up to one. The resulting K_s age distribution is shown in Supplementary Figure 2.6.



Supplementary Figure 2.6 K_s age distribution of *P. abies* calculated using single linkage clustering of genes within the High Confidence gene set.

2.10 Identifying the gene-like fraction of the *P.abies* 1.0 genome assembly

Regions in the *P.abies* 1.0 genome assembly with any similarity to expressed genes were identified by BLAST (blastn) alignment of assembled Trinity transcripts and by BWA[6] (v0.6.1-r104-tpx; bwa aln –o 0 –n 0, bwa samse –n 0) alignment of normalised RNA-Seq data (reads were aligned as single end data and all multiple mapping locations were reported). RNA-Seq data were first normalised using the normalize_by_kmer_coverage.pl script included with Trinity[3]. All raw Illumina RNA-Seq data from the 22 samples detailed above were combined and used as input to the normalisation procedure using the settings --PARALLEL_STATS --min_kmer_cov 2 –max_cov 80 –pairs-together, reducing the input set of 522,400,141 million read pairs to 37,820,814 million pairs. This alignment method cannot align reads crossing exon-exon boundaries: however since the aim was to select scaffolds showing any evidence of similarity to transcribed loci, maximal sensitivity was not required as it is highly unlikely that the only mapping read to a scaffold would be an exon-exon spanning one.

2.11 Tissue specificity of expression

The degree of specificity of expression within sampled tissues represented by the 22 RNA-Seq samples (see above) was analysed using $\log_2$(FPKM) expression values. FPKM <1 was set to $\log_2$(FPKM) = 0 and transcripts with $\log_2$(FPKM) = 0 in all 22 samples were removed. This resulted in the transcript counts summarised in Supplementary Table 2.8.

Supplementary Supplementary Table 2.8 Transcript counts per category defined for tissue-specificity tau-score calculations. *with* and *no GMAP* means with and without a GMAP alignment to the genome assembly. *Contaminants – fungi* means all Trinity transcripts with a BLAST hit to fungi, *Contaminants – others* means all Trinity transcripts with a BLAST hit to something other than than plants or fungi, *Spruce* means all High Confidence (HC) *ab initio*-predicted genes, *Spruce - all plants* means HC *ab initio*-predicted genes that are in TribeMCL families containing genes in all species included in the gene-family analysis, *Spruce – unique* means HC *ab initio*-predicted genes that are in TribeMCL families containing only spruce genes, *lncRNA – no hit* means Trinity transcripts with no ORF, > 200bps and no BLAST hit, and *lncRNA – plant* means Trinity transcripts with no ORF, > 200bps and a BLAST hit to plants. Note that Trinity transcripts overlapping with HC genes were removed. Spruce genes have NA for GMAP columns as they *per se* are located within the genome sequence.

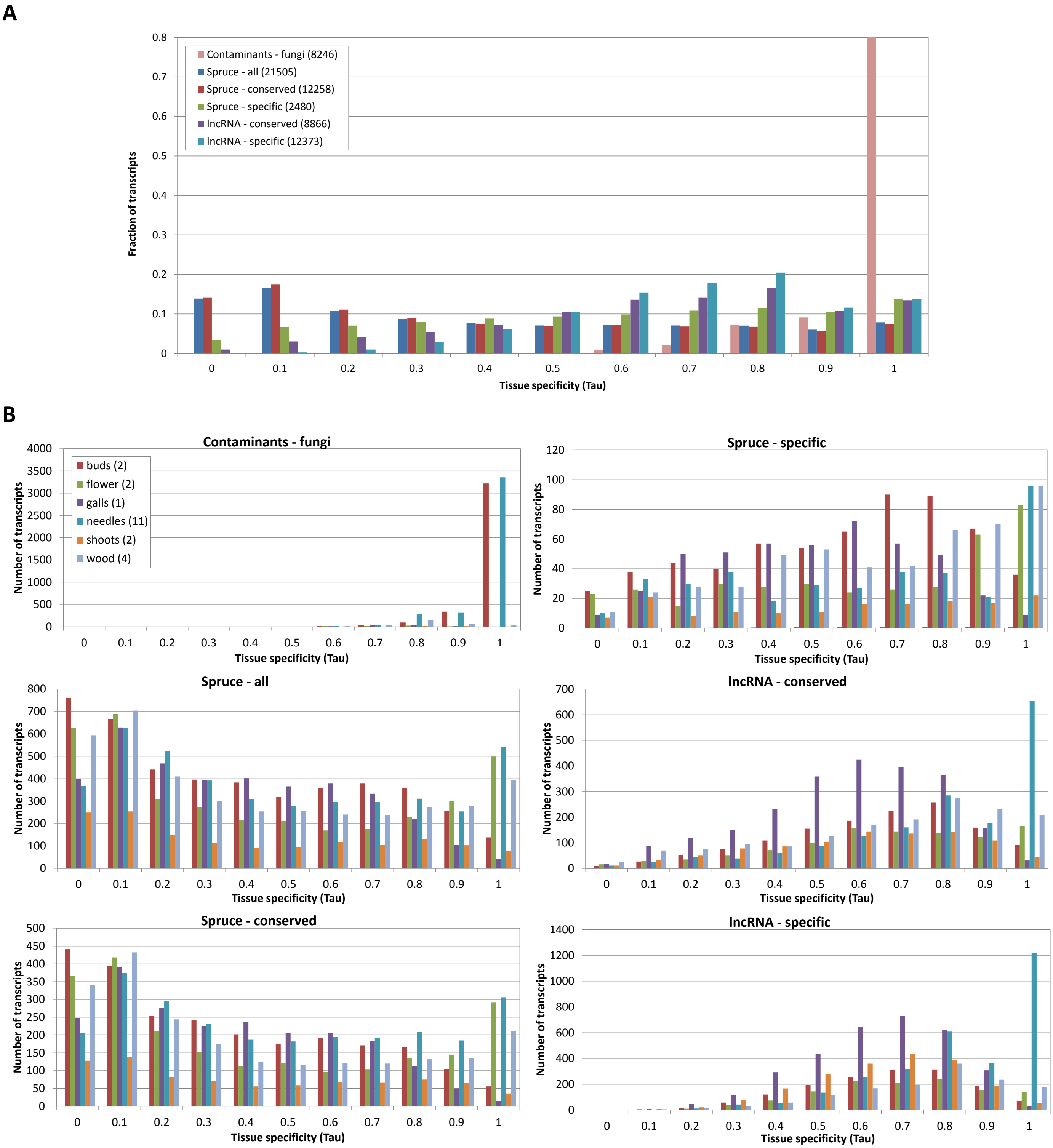
Category	Sequences	with GMAP	no GMAP
Contaminants - fungi	8246	513	7733
Contaminants - others	1001	402	599
Spruce	21505	NA	NA
Spruce - all plants	12258	NA	NA
Spruce - unique	3558	NA	NA
lncRNA - no hit	36101	24631	11470
lncRNA - plant	16601	15916	685

To measure tissue specificity, the tau score[32] was computed. Let a_{ij} be the average expression of gene *i* in tissue *j*. Then the tissue specificity of gene *i* is given by

$$\tau_i = \frac{1}{n-1}\sum_{j=1}^{n}(1 - \frac{a_{ij}}{\max_j(a_{ij})})$$

where *n* is the number of tissues. Thus, if the average expression of a gene is the same in all tissues the tau score is zero, and if a gene is only expressed in only one tissue the tau score is one.

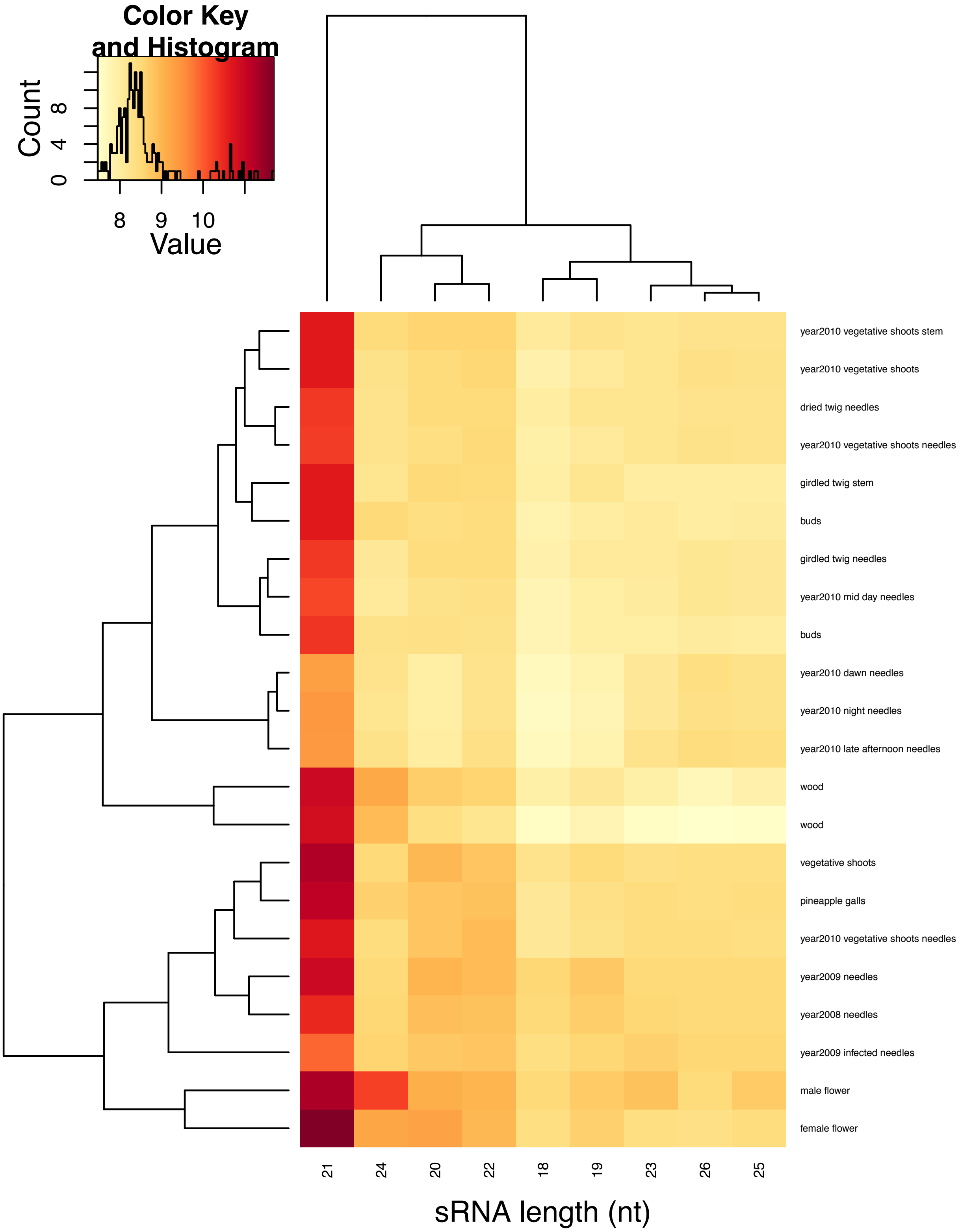
Supplementary Figure 2.8 shows the distribution of tissue specificity scores for some categories of transcripts. Transcripts classified as originating from contaminant species were extremely tissue specific to buds and needles. *P. abies* genes, on the other hand, were more uniformly distributed across all tau scores in all tissues. Interestingly, *P. abies* genes with orthologs in the seven other plants had very similar tissue specificity to that of the full set of *P. abies* genes while genes unique to *P. abies* were clearly more tissue specific. lncRNAs were expressed more broadly than contaminants, but were more tissue specific than *P. abies* genes. Conserved lncRNAs (the lncRNA-c subset) were somewhat more broadly expressed than those unique to *P. abies* (the lncRNA-s subset).



Supplementary Figure 2.8 A Normalised histogram of the tissue specificity scores (tau scores) for a subset of defined transcript categories. The number of transcripts in each category is given in parentheses in the legend. **B** Histograms of tissue specificity scores (tau scores) subdivided according to the tissue with the highest average expression. The number of samples from each tissue is given in parentheses in the common legend (upper left corner).

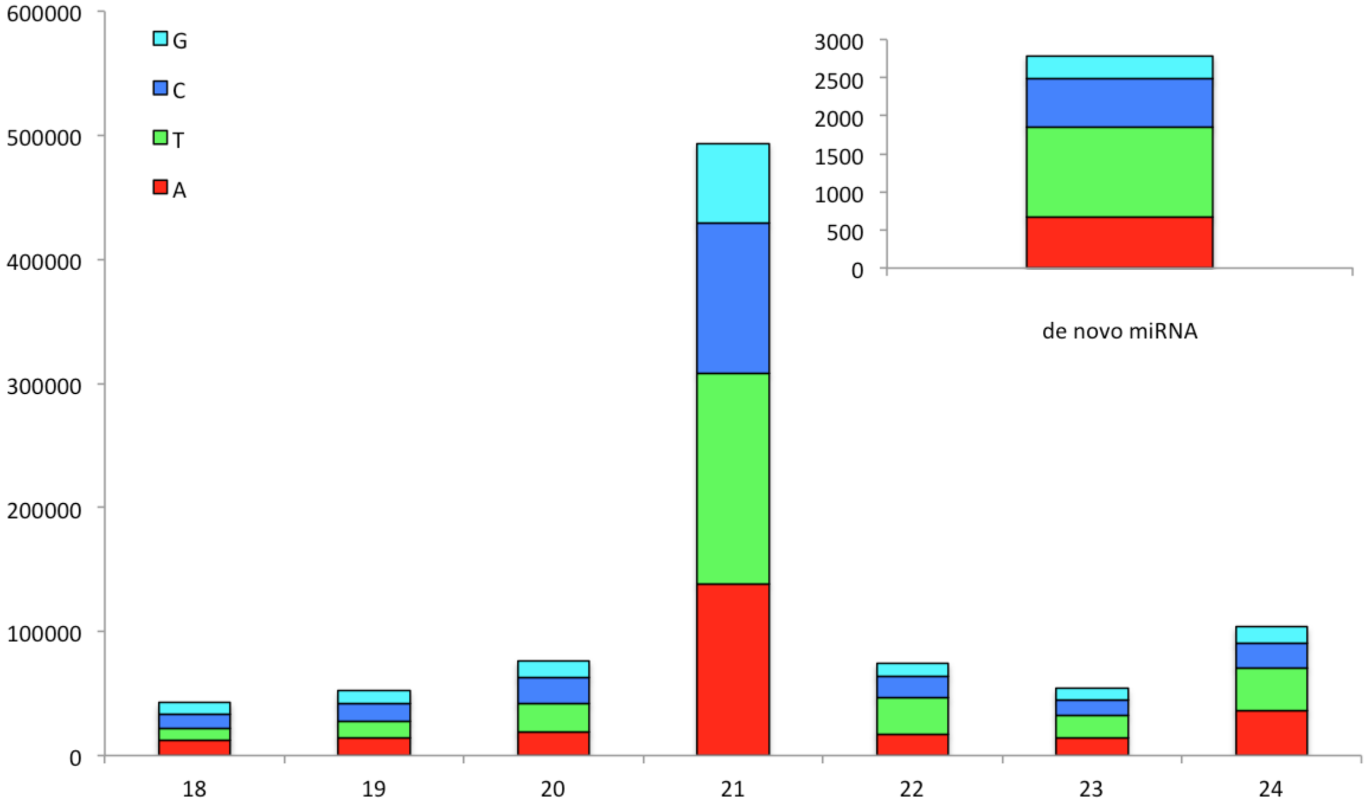
2.13 Short RNA analysis

The short RNA (sRNA) fraction (<200 bp) was isolated via agarose gel size selection from the above detailed 22 total RNA extractions described above. Size selected sRNA was sequenced using the Illumina HiSeq 2000 platform by the Science for Life Laboratory (SciLifeLab, Stockholm, Sweden). Libraries were prepared using the TruSeq Small RNA Sample Prep Kit (RS-200-0012, Illumina) and two libraries were prepared and sequenced from each RNA sample. The data from both libraries were pooled for subsequent analysis. Sequencing adapters were trimmed using cutadapt[33] (v1.0). All subsequent analysis steps were performed using the UEA sRNA Workbench[34] (v2.4.2, hereafter abbreviated to the prefix UEA with tool names from the workbench given). Reads were filtered using the UEA `Filter` tool to remove matches to rRNA/tRNA, low complexity sequences and sequences outside the 18-26 bp range. In addition, during the process, read sets were made non-redundant but original abundance was tracked to allow recreation of the redundancy information (*i.e.* expression values). The presence in the 22 sequenced samples of different sRNA sizes was examined (Supplementary Figure 2.10) revealing that expression of 24 nt sRNAs expression was extremely tissue-specific, being observed only in the sexual reproductive tissue samples at appreciable levels and, in particular, in the male cone sample.



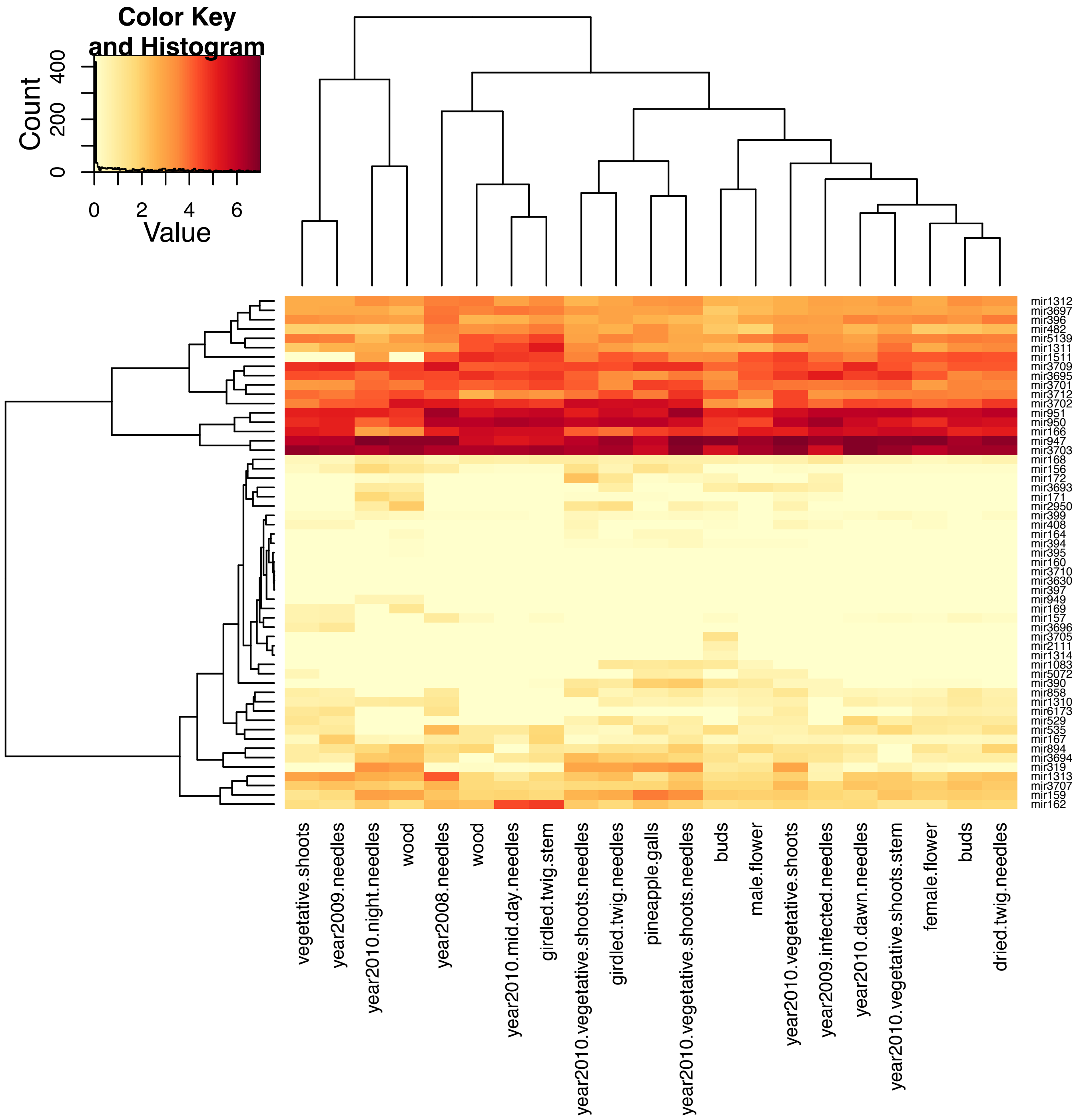
Supplementary Figure 2.10 Heatmap representation of unique sequence counts for 18-26 nt sRNA sequences across 22 samples. Samples and sRNA sizes are clustered. Plotted values are $\log_{10}$(number of reads). The density scales runs from low abundance (yellow) to high abundance (dark brown).

The first base composition of the different size classes was examined, revealing the commonly observed bias towards 21 nt sRNAs starting with a Uracil (Supplementary Figure 2.11), although this was less pronounced than is commonly observed in angiosperms (*e.g.* Klevebring *et al.*[35]).



Supplementary Figure 2.11 First base pair composition of sRNA sequences of size 18-24 nt. The inset shows first base composition of all *de novo* detected miRNAs as identified using the miRCat tool for the UEA sRNA Workbench.

The non-redundant, filtered read sets were aligned to the genome using the UEA `Sequence Alignment` tool. A minimum read abundance of 5 was required and 0 mismatches were allowed. Conserved miRNAs were identified using the UEA `miRProf` tool and miRBase[36] (v19). Expression of identified known and conserved miRNAs was examined across the 22 sequenced samples revealing that many of the previously identified spruce/conifer specific miRNAs present in miRBase (v19) were amongst those most highly expressed (Supplementary Figure 2.12).



Supplementary Figure 2.12 Heatmap representation of miRNAs identified using miRBase (v19) across the 22 sequenced samples. Values plotted are $\log_{10}$(Reads per Million Reads). The density scale runs from low abundance (yellow) to high abundance (dark brown).

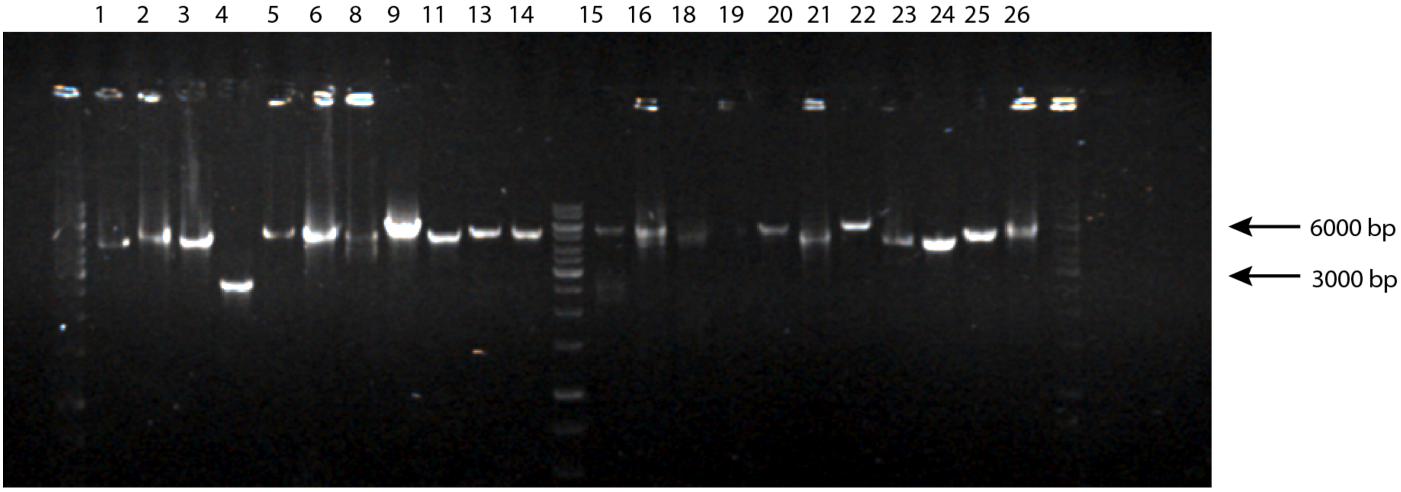
De novo miRNAs were predicted using the UEA `miRCat` tool with default plant settings applied. *De novo* predicted miRNAs with no match to miRBase are referred to as novel. The details, including mature miRNA sequences of all predicted miRNAs are available at the ConGenIE ftp site. Target prediction was performed using the UEA `Target Prediction` tool with default plant settings. The intersect function of `BedTools`[37] (v2.17.0) was used to identify overlap between aligned sRNAs and other genomic features. The overlap of different sRNA size classes with genomic features is shown in Figure 2d (main text).

2.14 Verification of long introns

To verify the presence of long introns, 20 ESTs from the *P. sitchensis* full-length ESTs that, within the *ab initio* HC gene set, contained at least one predicted intron >10 kbp were selected. PCR primers were designed to amplify a genomic fragment of the predicted long introns. One primer was anchored in the closest exon and the second primer was designed so that a fragment with a target size of 5 kbp would be amplified. Primers were designed using Primer3 (v0.4.0, http://frodo.wi.mit.edu) with default parameters. A total of 26 fragments were targeted, as some genes contained more than one intron >10 kbp. The average expected fragment size was 5.28 kbp (min: 2812 bp, max: 6760 bp, fragments are summarised in Supplementary Table 2.9). Fragments were amplified using a "Long-range PCR kit" from QIAGEN (cat. no. 206402) on a BioRad S1000 Thermo Cycler with DNA extracted from the *P. abies* Z4006 clone. Of the 26 fragments subjected to PCR, 22 were successfully amplified to give a fragment of the predicted size (Supplementary Figure 2.13). There was no significant difference in size between fragments that failed and those that were successfully amplified (t=0.813, df=3.3, p=0.471).

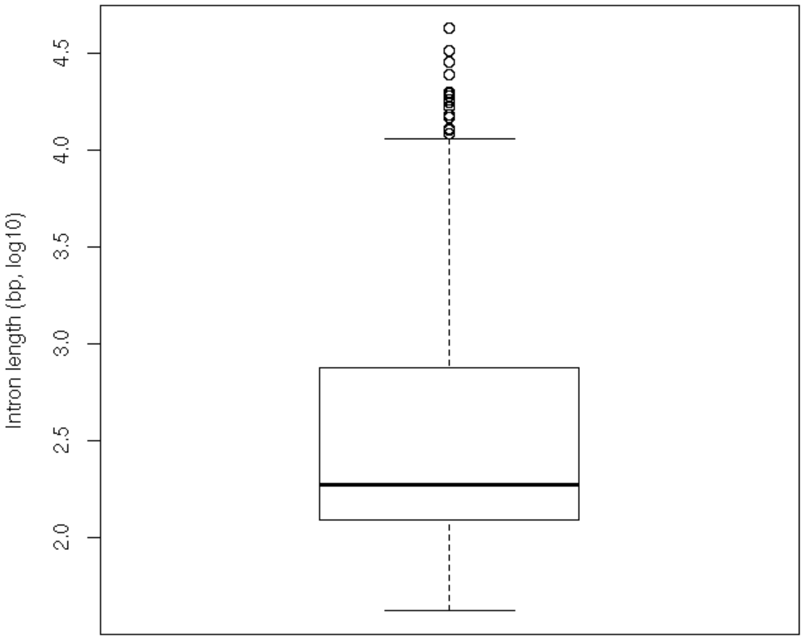
Supplementary Table 2.9 Summary of intron fragments tested using long-range PCR. [1]+; successful amplification, -; failed amplification

Fragment	Gene model	Exon	Fragment size (bp)	PCR[1]
1	Pa_S10735	4	4603	+
2	Pa_S10735	3	5508	+
3	Pa_S11869	3	5119	+
4	Pa_S12160	6	2941	+
5	Pa_S15175	5	5776	+
6	Pa_S15175	6	6238	+
7	Pa_S17699	3	5260	-
8	Pa_S17834	7	5103	+
9	Pa_S17834	1	6361	+
10	Pa_S19116	4	2812	-
11	Pa_S19189	3	5084	+
12	Pa_S19380	5	5293	-
13	Pa_S19380	2	5315	+
14	Pa_S20245	4	5420	+
15	Pa_S20557	3	6760	+
16	Pa_S42108	7	5728	+
17	Pa_S5185	1	5845	-
18	Pa_S577	5	5166	+
19	Pa_S6412	6	4984	+
20	Pa_S6475	2	5472	+
21	Pa_S7362	3	5692	+
22	Pa_S7732	5	5456	+
23	Pa_S9564	3	5188	+
24	Pa_S9564	6	5524	+
25	Pa_S97559	5	5299	+
26	Pa_S97559	3	5333	+



Supplementary Figure 2.13 PCR based confirmation of predicted long introns within the *P.abies* 1.0 genome assembly. See Supplementary Table 2.8 for details of the 26 fragments represented.

We additionally examined intron size in a set of homologs of 177 genes that have been shown to be strictly single copy in other plant genomes[38]. Homologs were identified in the HC gene set for 157 of these genes and the intron size distribution of these was similar to those for all HC genes, with a number of these single copy genes containing long (>2 Kbp) introns (Supplementary Figure 2.14). As these genes are present in all plant species and, in those other species are strictly single copy, we believe that the homologs we identified in the HC gene set represent the true functional homologs. It is therefore unlikely that the long introns contained within this set of genes are artefacts (*e.g.* due to pseudogenisation). As the HC genes required >70% reciprocal alignment of supporting evidence (aligned supporting evidence must cover >70% of the predicted gene locus and that alignment must cover >70% of the aligned supporting evidence source), we are also reasonably confident that these long introns do not represent potential assembly artefacts such as false fusions due to repeat presence (in such cases, supporting evidence would not along across the intron with proper splice site presence *etc.*).



Supplementary Figure 2.14 Box plot representation of Intron size distribution within homologs of the 177 strictly single copy genes identified by Smet *et al.*[38]. Homologs were identified within the HC gene set for 155 of the 177 single copy genes. Mean intron size was 1643.

2.15 Data availability

The raw RNA-Seq data (454 and Illumina) detailed abovehas been deposited at the ENA (www.ebi.ac.uk/ena) under the accessions ERP002475 (mRNA) and ERP002476 (sRNA).

2.15 References

1. Chang, S., Puryear, J. & Cairney, J. A simple and efficient method for isolating RNA from pine trees. *Plant Molecular Biology Reporter* **11**, 113–116 (1993).
2. Gouzy, J., Carrere, S. & Schiex, T. FrameDP: sensitive peptide detection on noisy matured sequences. *Bioinformatics* **25**, 670–671 (2009).
3. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech* **29**, 644–652 (2011).
4. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–10 (1990).
5. Wu, T. & Watanabe, C. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
6. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009).
7. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323 (2011).
8. Ralph, S. G. *et al.* A conifer genomics resource of 200,000 spruce (Picea spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (Picea sitchensis). *BMC genomics* **9**, 484 (2008).
9. Rigault, P. *et al.* A white spruce gene catalog for conifer genome analyses. *Plant physiology* **157**, 14–28 (2011).
10. Hu, J., Ge, H., Newman, M. & Liu, K. OSA: a fast and accurate alignment tool for RNA-Seq. *Bioinformatics* **28**, 1933–1934 (2012).
11. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech* **28**, 511–515 (2010).
12. Roberts, A. *et al.* Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology* **12**, 1–14 (2011).
13. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* **33**, W465–W467 (2005).
14. Schiex, T., Moisan, A. & Rouzé, P. Eugène: An Eukaryotic Gene Finder That Combines Several Sources of Evidence. *Computational Biology SE - Lecture Notes in Computer Science* **2066**, 111–125 (2001).
15. Degroeve, S., Saeys, Y., De Baets, B., Rouze, P. & Van de Peer, Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* **21**, 1332–1338 (2005).
16. Duvick, J. *et al.* PlantGDB: a resource for comparative plant genomics. *Nucl. Acids Res.* **36**, D959–965 (2008).
17. Lorenz *et al.* Conifer DBMagic: a database housing multiple de novo transcriptome assemblies for 12 diverse conifer species. *Tree Genetics & Genomes* **8**, 1477–1485 (2012).
18. Finn, R. *et al.* The Pfam protein families database. *Nucleic Acids Research* **38**, D211–D222 (2010).
19. Conesa, A. & Götz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International journal of plant genomics* **2008**, 1–13 (2008).
20. Zdobnov, E. M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics (Oxford, England)* **17**, 847–848 (2001).
21. Goodstein, D. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic acids research* **40**, D1178–D1186 (2012).
22. R-Core-Team *R: A language and environment for statistical computing*. (Vienna, Austria OR - R Foundation for Statistical Computing, 2012).

23. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80 (2004).

24. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics (Oxford, England)* **23**, 1061–7 (2007).

25. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research* **30**, 1575–1584 (2002).

26. Felsenstein, J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164–166 (1989).

27. Farris, J. Phylogenetic Analysis Under Dollo's Law. *Systematic Zoology* **26**, 77–88 (1977).

28. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792–1797 (2004).

29. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**, 725–736 (1994).

30. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586–1591 (2007).

31. Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 5454–5459 (2005).

32. Liao, B.-Y. & Zhang, J. Evolutionary Conservation of Expression Profiles Between Human and Mouse Orthologous Genes. *Molecular Biology and Evolution* **23**, 530–540 (2006).

33. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Bioinformatics in Action* **17**, 10–12 (2012).

34. Stocks, M. *et al.* The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics* **28**, 2059–2061 (2012).

35. Klevebring, D. *et al.* Genome-wide profiling of populus small RNAs. *BMC genomics* **10**, 620 (2009).

36. Griffiths-Jones, S., Grocock, R. J., Dongen, S. van, Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research* **34**, D140–D144 (2006).

37. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

38. De Smet, R. *et al.* Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 2898–903 (2013).

Supplementary Material Section 3: Repeats

3.1 Repeat identification:

A random sample of 100,000 *Picea abies* 454 randomly sheared reads longer than 700 bp was searched de novo for repeats using the software RepeatScout [1]. The results were filtered to remove low complexity sequences and sequences shorter than 100 nt and to retain only repeats having at least 10 matches when mapped onto the original 454 set using RepeatMasker [2]. Since RepeatScout is tailored for analysis of complete genomes or at least large scaffolds, its output is usually fragmented when the program is run on randomly sheared reads. In order to reduce fragmentation we merged the repeats belonging to the same element by running cap3 [3] under relaxed settings (-o 30, -p 80, -s 500) on the RepeatScout output. This repeat dataset was completed by adding 124 putative Long Terminal Repeat Retrotransposons (LTR-RTs) identified by using LTR-FINDER [4] on *P. abies* contigs longer than 10,000 nt and another 54 LTR-RT elements (both complete and fragmented) identified from 4 completely sequenced *P. abies* BACs [5]. Finally, the whole set of repeated sequences was clustered using the software cd-hit-est [6], collapsing all the repeats that shared at least 80% similarity to remove redundancies. We hereafter refer to the *P. abies* repeat library constructed using this method as the Repeats2.0 library. For the remaining 6 gymnosperm datasets (*Abies sibirica, P. glauca, Pinus sylvestris, Juniperus communis, Taxus baccata* and *Gnetum gnetum*), 454 randomly sheared reads were produced; a subset of 100,000 reads for each species was analyzed separately using the same de novo strategy (RepeatScout + cap3 + cd-hit) adopted for Norway spruce.

3.2 Repeat characterization:

Similarity searches were used:
a) to associate candidate repeats with known TE families;
b) to remove repeats showing similarity to genes and therefore possibly members of gene families. To do this, the repeat candidates from each species were searched against RepBase [7] using TBLASTX [8] and setting an E-value of 1e-5 as the significance threshold. Repeats that did not yield significant hits were used as queries in BLASTX searches against the nr division of GenBank. Those giving significant hits with genes were removed from the library, those having significant hits with TEs were labeled accordingly, and the remainder are considered to be “Unclassified” repeats. The latter are likely to be parts of non coding portions of LTR-RTs, such as the LTRs, or parts of highly diverged TEs. In either case, any attempt to further characterize these sequences by similarity searches would be inconclusive unless a complete copy of the element is present in the database (Supplementary Table 3.1)

Supplementary Table 3.1. Repeat library composition and masking of randomly sheared sets of 454 reads. Each random sheared library contains 100,000 non-redundant 454 reads. "No" denotes number of elements in each repeat library, "%" is the percentage of sequences identified by repeat masking using the custom library.

	Abies sibirica		*Gnetum*		*Juniperus*		*Pinus sylvestris*		*Picea abies*		*Picea glauca*		*Taxus baccata*	
	No	%	No	%	No	%	No	%	No	%	No	%	No	%
LTR_gypsy	139	13.56	101	22.41	93	6.83	143	15.80	384	35.38	367	31.38	176	19.81
LTR_copia	75	8.01	32	2.09	66	10.90	85	7.38	216	16.13	151	9.72	43	2.80
LTR_unk	7	0.49	10	0.48	6	0.19	9	0.21	164	6.82	110	5.39	5	0.09
LTR total	221	22.06	143	24.98	165	17.92	237	23.39	764	58.33	628	46.49	224	22.70
LINE	14	0.65	22	0.99	24	0.63	18	0.52	25	0.96	25	0.76	15	0.20
DNA_TE	17	0.75	0	0	4	0.06	14	0.12	23	0.57	10	0.14	19	0.73
Unclassified	911	30.44	923	36.87	820	27.64	1158	28.23	961	9.67	821	13.31	1119	26.83
Total	1163	53.9	1231	62.84	1013	46.25	1427	52.26	1773	69.53	1484	60.7	1377	50.46
average read L	614.31		590.82		624.71		585.97		745.60		519.99		620.55	
Genome size (Mbp/1C)	15,560		3,785		11,712		22,474		20,020		19,756		11,230	

3.3 Validation of *P. abies* repeat library:

To evaluate the extent to which the repeat library created for *P. abies* is representative, we produced a second one using the software Trinity [9]. If the initial library is representative, this second set should contain all the repeats present in the first library. The Trinity based library includes 265,971 sequences with a total of 31,682,977 nucleotides whereas the library produced using the aforementioned strategy (Repeats_2.0 library) includes 1,773 sequences with a total of 2,035,418 nucleotides. When RepeatMasker was fed with the two libraries to mask the original set of 100,000 *P. abies* 454 reads, the Repeats_2.0 library based library masked 68.98% of the set whereas the Trinity based one masked 76.16%. When the two libraries were combined, the software masked 78.62% of the set of 454 reads. The Repeats2.0 library covers 88.43% of the total repeats identified by both libraries. The Repeats_2.0 library sequences not shared with those included in the Trinity based library (1,105 sequences; totaling less than 500 kb) mask 2.11% of the original 100,000 454 set whereas the sequences unique to the Trinity repeats (113,610 sequences totaling more than 13.5 Mbp) mask 9.09% of the same set.

In evaluating the results of this comparisons it is important to note:
a) the magnitude of the size difference between the two sets: the Trinity library has 15 times more nucleotides than the Repeats_2.0 library
b) the fact that the Repeats_2.0 library has been curated removing:
-all the sequences likely to have come from gene families;
-those of low complexity (such as microsatellites);
-those shorter than 100 nucleotides.
Despite these factors, the Trinity based library masks only 9% more of the total sequence and Repeats_2.0 still covers almost 90 % of the whole complement. For these reasons we consider the Repeats_2.0 library to be a complete representation of the most abundant TE related repeats in the Norway spruce genome.

3.4 Identification of complete LTR-RT elements in *Picea glauca*

24 *P. glauca* bacterial artificial chromosome (BAC) sequences with an average length of 95,681 bp (Supplementary Table 3.2) were screened, identifying 120 *P. glauca* LTR-RTs (average length = 14,583 bp) (Supplementary Table 3.3.xls). Identification of complete LTR-RTs was done using a combination of de novo searches and manual inspection using the programs LTR_Finder [4] and DOTTER [10] respectively. All putative transposable elements were searched for the presence of specific signatures of LTR-RT elements: terminal TG/CA inverted repeats in the LTRs and Target Site Duplication (TSD). They were also classified as Ty1-copia or Ty3-gypsy according to the results of BLAST [8] searches against the nr division of GenBank.

Supplementary Table 3.2. Picea glauca BAC sequences used for repeat analysis.

BAC sequence	Length (bp)
323	39,891
328	199,405
333	7,462
338	85,519
341	137,610
342	141,300
343	10,588
352	76,606
363	193,756
364	137,259
asn1-322	130,199
asn1	144,302
c3h	162,229
cesa2	150,478
cesa3	11,438
comt	10,565
hdzip	10,248
icdh	11,860
korrigan	83,371
myb14-331	162,405
myb8	93,010
pal	148,428
sad	11,244
susy	137,170

3.5 Identification of complete LTR-RT elements in *Picea abies*

Adopting the same strategy as was used for *P. glauca* BACs, we mined *P. abies* assembly scaffolds with length ≥ 50 kb (9,438 in total), identifying a subset of 150 complete LTR-RTs (average length = 8,732 bp) (Supplementary Table 3.4.xls). 20 fully sequenced fosmid sequences (Supplementary Table 3.5) (average length = 38,498 bp) were also searched, allowing the identification of 39 LTR-RTs ranging in size from 1.9 to 29.3 kb (Supplementary Table 3.6).

Supplementary Table 3.5: *Picea abies* fully sequenced fosmid sequences.

Fosmid sequence	Length (bp)
LuSPFOS_1	36,775
LuSPFOS_3	37,684

LuSPFOS_4	36,212
LuSPFOS_8	35,463
LuSPFOS_9	37,484
LuSPFOS_12	37,568
LuSPFOS_16	36,039
LuSPFOS_25	42,937
LuSPFOS_26	43,588
LuSPFOS_28	39,517
LuSPFOS_29	28,415
LuSPFOS_31	33,581
LuSPFOS_33	45,858
LuSPFOS_36	44,176
LuSPFOS_42	44,924
LuSPFOS_43	39,623
LuSPFOS_45	35,584
LuSPFOS_47	37,044
LuSPFOS_48	39,923
LuSPFOS_49	37,572

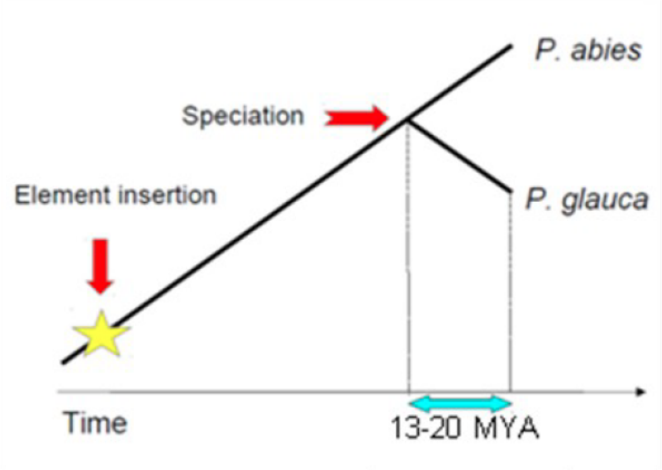
Supplementary Table 3.6: 39 *Picea abies* LTR-RT elements identified across 20 *P. abies* fosmid sequences. For each element (LTR-RT element) within the corresponding fosmid sequence: LTR5'S = start position of the LTR at the 5' end; LTR5'E = end position of the LTR at the 5' end; LTR3'S = start position of the LTR at the 3' end; LTR3'E = end position of the LTR at the 3' end; Classification = annotation of the element based on BLAST search results (Unclassified = no specific match was found).

LTR-RT element	Fosmid sequence	LTR5'S	LTR5'E	LTR3'S	LTR3'E	Classification
151	LuSpFOS_16	7,863	8,287	13,669	14,093	Gypsy
152	LuSpFOS_25	27,691	28,327	32,155	32,791	Copia
153	LuSpFOS_26	4,559	4,847	10,166	10,454	Gypsy
154	LuSpFOS_28	2,270	3,200	10,760	11,683	Gypsy
155	LuSpFOS_28	24,398	24,686	29,906	30,194	Gypsy
156	LuSpFOS_29	2,448	3,752	8,527	9,771	Gypsy
157	LuSpFOS_29	10,916	11,572	13,527	14,183	Gypsy
158	LuSpFOS_31	7,036	9,767	14,786	17,517	Gypsy
159	LuSpFOS_33	8,250	9,232	21,775	22,757	Gypsy
160	LuSpFOS_33	955	3,780	23,296	26,121	Gypsy
161	LuSpFOS_36	10,249	13,831	16,729	20,307	Unclassified
162	LuSpFOS_36	26,062	26,721	30,757	31,388	Copia
163	LuSpFOS_3	3,612	6,944	9,346	12,680	Unclassified
164	LuSpFOS_3	29,494	30,092	34,717	35,315	Gypsy
165	LuSpFOS_3	2,812	3,457	14,476	15,117	Gypsy
166	LuSpFOS_42	4,791	5,720	15,045	15,941	Gypsy
167	LuSpFOS_47	16,904	18,602	25,014	26,709	Gypsy
168	LuSpFOS_48	15,995	16,283	20,832	21,120	Gypsy
169	LuSpFOS_48	3,096	5,913	7,712	10,558	Unclassified
170	LuSpFOS_49	21,938	23,436	33,454	34,951	Gypsy
171	LuSpFOS_4	26,305	28,269	30,106	32,070	Unclassified
172	LuSpFOS_4	20,600	22,847	33,664	35,866	Gypsy
173	LuSpFOS_8	6,466	6,671	11,929	12,218	Gypsy
174	LuSpFOS_9	523	1,509	14,057	15,043	Gypsy
175	LuSpFOS_1	15,310	20,021	27,120	31,856	Gypsy
176	LuSpFOS_16	17,755	18,691	27,204	28,142	Gypsy
177	LuSpFOS_26	10,727	11,665	37,898	38,834	Gypsy
178	LuSpFOS_26	24,941	26,239	34,944	36,252	Gypsy
179	LuSpFOS_28	31,701	32,015	33,303	33,617	Gypsy
180	LuSpFOS_42	17,668	19,199	25,869	27,400	Gypsy
181	LuSpFOS_45	656	2,132	2,790	4,332	Copia
182	LuSpFOS_45	4,336	5,219	18,722	19,605	Gypsy
183	LuSpFOS_45	2,130	2,749	19,605	20,223	Gypsy
184	LuSpFOS_47	10,192	11,035	21,160	22,003	Gypsy
185	LuSpFOS_49	7,823	9,155	36,183	37,187	Gypsy
186	LuSpFOS_36	3,993	4,309	7,504	7,846	Copia
187	LuSpFOS_36	38,254	38,788	41,965	42,508	Copia
188	LuSpFOS_43	20,372	20,796	24,903	25,436	Copia

189	LuSpFOS_4	11,299	14,254	17,593	20,511	Copia

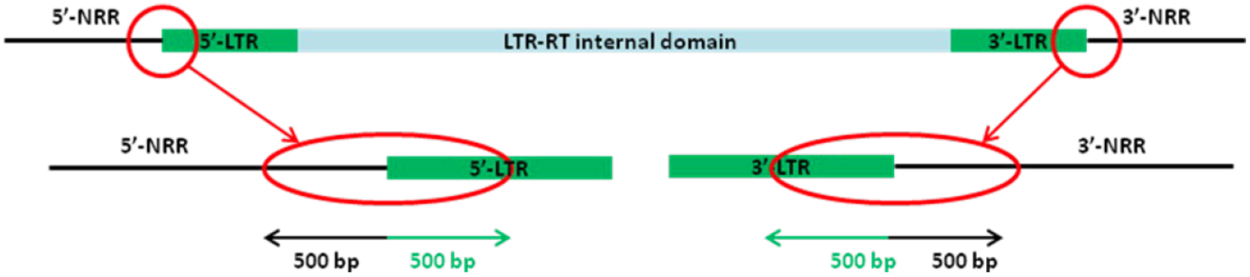
3.6 Estimating LTR substitution rate

To attempt an initial estimate of the LTR substitution rate in conifers, we identified orthologous insertions of these elements in *P. abies* and *P. glauca*. Since these orthologous LTR-RTs must have been present in a common ancestor pre-dating the divergence of the two species, nucleotide differences identified by comparing these regions must correspond to mutations accumulated post-speciation (Supplementary Figure 3.1).



Supplementary Figure 3.1: An LTR-RT element present in the common ancestor can be found as an orthologous insertion in the *Picea abies* and *Picea glauca* genomes.

Our strategy to detect orthologous insertions was as follows: for each LTR pair of the 120 P. glauca LTR-RT elements, the first 500 nucleotides of the 5'-LTR were extracted from the BAC sequences along with 500 upstream flanking nucleotides (not repeat related flanking sequence - NRR) (Supplementary Figure 3.2). Similarly, the last 500 nucleotides of the 3'-LTR were extracted along with 500 downstream flanking nucleotides. The tracts (LTR+NRR) were then compared by means of BLASTN [8] similarity searches against the *P. abies* whole haploid genome shotgun assembly in order to identify candidate orthologous insertions. 52 LTR-RT elements out of 120 were classified as "non-mapped" because their NRR tracts did not map to an unequivocal position on the *P. abies* assembly (Supplementary Table 3.7), therefore we were unable to draw any conclusions regarding their possible orthology. 63 were uniquely mapped (classified as "orthologous" insertions) while the results for 5 showed that they were inserted after speciation between *P. abies* and P. glauca had taken place, as these elements are present in *P. glauca* but not in *P. abies* (for each of these elements, the 5'-NRR tract maps consecutively to the 3'-NNR tract on the *P. abies* haploid genome sequence).



Supplementary Figure 3.2: Strategy used to identify *Picea glauca* orthologous LTR-RT elements in the *Picea abies* assembly.

Supplementary Table 3.7: Classification of Picea glauca LTR-RT elements as orthologous insertions.

Insertion classification	Abundance
Orthologous	63
Non-mapped	52
After speciation	5

The 63 *P. glauca* LTR-RT elements identified as putative orthologous insertions were manually confirmed by aligning each to its matching *P. abies* contig using dot-plot analysis. Orthologous LTR and NRR sequences were extracted from both species; extracted NRR sequences had lengths comparable to those of the corresponding flanking LTRs. *P. abies* NRRs were masked using RepeatMasker [2] and a *P. abies*-specific collection of repeats (Repeats_2.0 library, see paragraph "Validation of *P. abies* repeat library") to remove detectable transposable element insertions. In particular, we discarded from our analysis all the LTR-RTs that were nested into other TEs. Orthologous LTR and NRR sequences were aligned using the program "Stretcher" (part of the EMBOSS package [11]) and the nucleotide distance was estimated using the Kimura two-parameter (K2p) (transition-transversion ratio) criterion [12] as implemented in the program "Distmat" (EMBOSS package [11]). Substitution rates (μ) (Supplementary Table 3.8) were inferred using the formula:

$$\mu = d / 2T$$

where d is the nucleotide distance (K2p) for all orthologous *P. abies* and *P. glauca* LTR and NRR alignments, and T is the divergence time between *P. glauca* and *P. abies.* Divergence time deduced from fossil records is between ~13 and ~20 MYA [13].

Supplementary Table 3.8: Average substitution rates obtained from orthologous *Picea abies* and *Picea glauca* alignments.

Genomic portion	μ (13 MYA)	μ (20 MYA)
LTR	2.20×10^{-9}	1.43×10^{-9}
NRR	0.99×10^{-9}	0.64×10^{-9}

3.7 Base pair changes in LTRs and NRRs

We determined the proportion of each type of base pair change that had occurred in each orthologous LTR (Supplementary Table 3.9) and NRR sequence (Supplementary Table 3.10).

Supplementary Table 3.9: Base pair changes within LTR sequences.

Base pair change in LTRs	%
G<-->T (A<-->C)	26.31
A<-->T	7.61
A<-->G (C<-->T)	62.12
C<-->G	3.97

Supplementary Table 3.10: Base pair changes within NRR sequences.

Base pair change in NRRs	%
G<-->T (A<-->C)	26.22
A<-->T	8.09
A<-->G (C<-->T)	61.66
C<-->G	4.03

3.8 Estimation of insertion times for LTR-RT elements

We calculated the insertion time (T) for each of the 150 *P. abies* and 120 *P. glauca* LTR-RT elements using the molecular paleontology approach described by SanMiguel et al. [14]. Specifically, we used the formula:

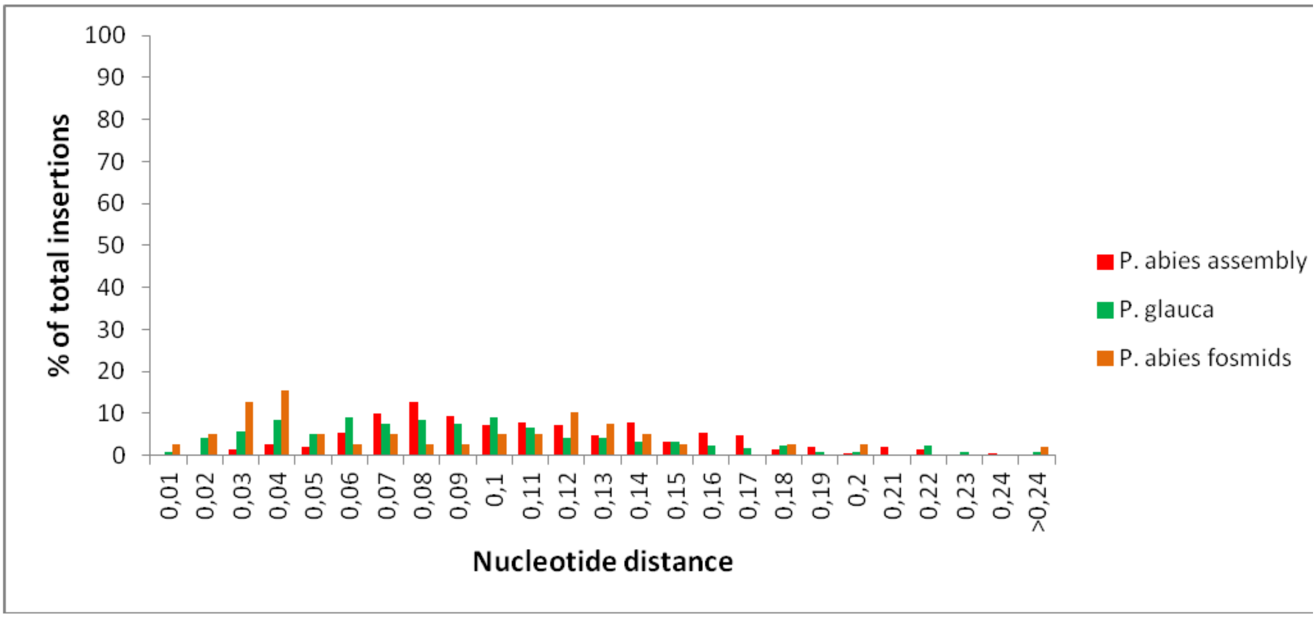
$$T = d / 2\,\mu$$

where d is the 2Kp nucleotide distance for each pair of LTRs and μ is the nucleotide substitution rate estimated as previously described (Supplementary Table 3.8). We used the fastest substitution rate estimated for each pair. The average nucleotide distances of *P. glauca*, *P. abies* and *P. abies* fosmid LTR sequence pairs Supplementary to be, respectively, 0.0885, 0.0981 and 0.0894 (Supplementary Table 3.11.xls, Supplementary Table 3.12.xls and Supplementary table 3.13; Supplementary Figure 3.3). The insertion time distributions of the LTR-RT elements detected for *P. glauca* and *P. abies* (in the master assembly) show the largest number of insertions between 16 and 32 MYA (Supplementary Figure 3.4); however, distributions for the LTR-RTs in the two *Picea* spp are significantly different when tested using the two-tailed Wilcoxon rank sum test with continuity correction (W = 6836.5, p-value =0.001293). The insertion time distribution for the *P. abies* LTR-RT elements detected across the fosmid sequences also shows a large number of insertions between 8 and 12 MYA (Supplementary Figure 3.4). This is probably because the complete fosmid sequence dataset can be used to reconstruct recent TE insertions in an even more efficient manner than is possible with the assembly.

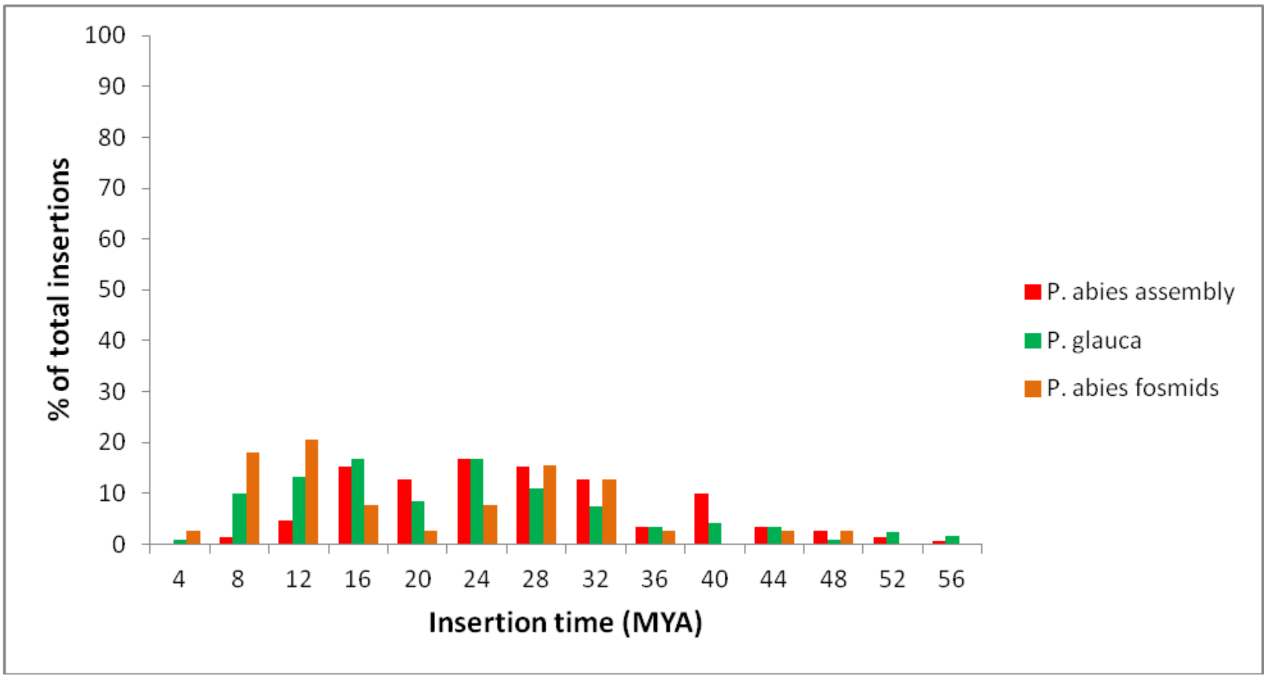
Supplementary Table 3.13: Nucleotide distance separating the LTRs of each LTR-RT element annotated in 20 *Picea abies* fosmid sequences.

LTR-RT element	Fosmid_name	Nucleotide distance
151	LuSpFOS_16	0.0095
152	LuSpFOS_25	0.1141
153	LuSpFOS_26	0.0175
154	LuSpFOS_28	0.0368
155	LuSpFOS_28	0.0356
156	LuSpFOS_29	0.1167
157	LuSpFOS_29	0.0264
158	LuSpFOS_31	0.0262
159	LuSpFOS_33	0.0333
160	LuSpFOS_33	0.0878
161	LuSpFOS_36	0.0320
162	LuSpFOS_36	0.0426
163	LuSpFOS_3	0.0770
164	LuSpFOS_3	0.0307
165	LuSpFOS_3	0.1259
166	LuSpFOS_42	0.0320
167	LuSpFOS_47	0.1425
168	LuSpFOS_48	0.0105
169	LuSpFOS_48	0.1386
170	LuSpFOS_49	0.0646
171	LuSpFOS_4	0.0222
172	LuSpFOS_4	0.1149
173	LuSpFOS_8	0.0933
174	LuSpFOS_9	0.0290
175	LuSpFOS_1	0.0421
176	LuSpFOS_16	0.1205
177	LuSpFOS_26	0.3493
178	LuSpFOS_26	0.1164
179	LuSpFOS_28	0.1281
180	LuSpFOS_42	0.1039
181	LuSpFOS_45	0.0506
182	LuSpFOS_45	0.0290
183	LuSpFOS_45	0.0660
184	LuSpFOS_47	0.3083

185	LuSpFOS_49	0.1039
186	LuSpFOS_36	0.1940
187	LuSpFOS_36	0.1380
188	LuSpFOS_43	0.1790
189	LuSpFOS_4	0.0987



Supplementary Figure 3.3: Nucleotide distance distribution of annotated *Picea abies* and *P. glauca* LTR-RT elements.



Supplementary Figure 3.4: Insertion time distribution of annotated *Picea abies* and *P. glauca* LTR-RT elements.

3.9 Identification of complete LTR-RTs in 43 *Pinus taeda* BACs

43 *Pinus taeda* bacterial artificial chromosome (BAC) sequences (accession numbers are reported in Supplementary Table 3.14.xls) were screened, identifying 104 LTR-RT elements. The *P. abies* master assembly was searched for insertions orthologous to these LTR-RT elements by exploiting the same strategy described in the section "Calculating LTR substitution rate" but using an upstream and downstream flanking region size of 250 bp due to the greater sequence divergence between *P. taeda* and *P. abies.* Only 16 flanking regions from the 104 LTR-RT elements were uniquely mapped onto the *P. abies* master assembly (Supplementary Table 3.15). Manual inspection of dot plots comparing *P. taeda* 5'NRR+LTR-RT element+3'NRR tracts and the corresponding *P. abies* scaffolds revealed that none of the P. taeda transposable elements was present in the *P. abies* master assembly.

Supplementary Table 3.15: *Pinus taeda* LTR-RTs whose flanking regions were mapped onto *Picea abies* scaffolds. For each element: start and end = coordinates of the complete element in the corresponding BAC; Region mapped onto = *P. abies* scaffold on which the regions flanking the *P. taeda* element were clearly mapped; Classification = annotation of the element based on BLAST search results; Status = presence or absence of the element in *P. abies*.

BAC accession number	Start	End	Region mapped onto	Classification	Status in *P. abies*
AC241268.1	18103	31085	scaffold_143562	copia	Absent
AC241276.1	13589	17560	scaffold_29270	copia	Absent
AC241263.1	50281	54500	scaffold_90450	gypsy	Absent
AC241265.1	54127	60849	scaffold_10128943	gypsy	Absent
AC241270.1	7608	14489	scaffold_159694	gypsy	Absent
AC241289.1	13888	19004	scaffold_386559	gypsy	Absent
AC241292.1	95193	100462	scaffold_10041	gypsy	Absent
AC241298.1	14587	19830	scaffold_143562	gypsy	Absent
AC241268.1	1972	12725	scaffold_1031947	copia	Absent
AC241276.1	58910	62118	scaffold_3815832	copia	Absent
AC241288.1	73722	89565	scaffold_72905	copia	Absent
AC241301.1	63108	77551	scaffold_143562	copia	Absent
AC241276.1	3949	7730	scaffold_270598	gypsy	Absent
AC241286.1	72826	78851	scaffold_195594	gypsy	Absent
AC241287.2	39109	45507	scaffold_335961	gypsy	Absent
AC241290.1	93875	104308	scaffold_1716	gypsy	Absent

3.10 Unequal Homologous Recombination (UHR) targeting complete LTR-RTs

Three different LTR-RT families (ALISEI, 3K05 and 4D08) were identified on the basis of different LTR sequences and analyzed to evaluate the occurrence of UHR events leading to the creation of solo LTRs [15]. We searched the *P. abies* assembly fraction ≥ 50 kb, 20 fosmids and 4 fully sequenced BACs for complete LTR-RT elements and solo LTRs belonging to each of the three families. For the assembly fraction the analysis was conducted by means of BLASTN similarity searches (using the threshold criteria E-value = e-5; identity percentage = 70%**) using as query a complete LTR sequence for each of the three families (Supplementary Table 3.16). For each LTR family, the BLASTN output was parsed to determine the number of complete LTR-RT elements detected and the number of putative solo LTR. Elements were identified as complete when the query gave a full-length match twice on the same scaffold within a specific distance range (from 1.5-2 kb to 15-20 kb) corresponding to the common LTR-RT element length. Solo LTRs were initially identified as single matches for the entire query. Putative solo LTRs were subsequently confirmed manually by inspecting the sequence for the presence of canonical flanking 5 bp target site duplications. In the case of fosmids and BACs, the identification of complete elements and solo LTRs for the three families was carried out by manual inspection of the results of dot plot analysis. In total, we identified 27 complete elements and 5 solo LTRs for the ALISEI family, 26 complete elements and 5 solo LTRs for 3K05, and 41 complete elements and no solo LTRs for the 4D08 family (Supplementary Table 3.17).

Supplementary Table 3.16: *Picea abies* LTR sequences used for the detection of solo LTRs and complete elements.

LTR-RT family	LTR sequence
ALISEI	TGTTAAGCCCTAATATTGGCTCATTACCCCTTAATTGGGGCATATTATATTAGTAATGTTATTTTATTTATGTCGGCTTTAGGCCATTAGTTGTAATTATACCTCCTAGTCTCCTATATATACCAAGGAGACATTTTCATTTTGAATCCTCTTGTACTTGAACATTTAAGTGGGGTAAAAGGAGGTTTCCCCCTCGAAGAGGCTTATTATCTGAAATTATCTCTCTTTCAAGGCATTGAGTACTTTTTCTGAGGATTTTCTATGCTATATTGATATTTGGTTCTTAACA
3K05	TGTTAATATTATGGTAGATGCGGGAGGTATTTACCTCTACATGTGAACTTGACGAGAATAGAAGACTCCGGAAAGCAAGAGAAACCCTAATGAGTTGGATAATAACCCTAGAAGCCTTATCGGCTCAAGAAGATAATTTATATTAGTCTATTAGCCTCCAAGAGAAGAAGTTAGCCAGAAATAATAATATTATACTGAACAAACACATACCCAGAAGGAGAAGTCGGCCCCAGCAAAACTTGCAGGATTACTATAAATACAGGGTGCCAGTTCTCATTTGACAACATCTCAGAATTTCAGTTCTATAGTTTGGTGTGCACTGCAGTTGCAGTTTGGTGCGCAGGTTTGGAGGAAGAGCTGGTGCAGTTGCAGTAGCAACAAGTTCTCAGTTTTGGCTTGAATATCTCTTGGCTAGGGTTTGGGACTATCTGATCAGCTTTTGTGATGCGCTATCCTTGGGGATTGATGTGCTACTCTTTGTTTGGCTATCTAGCCTAGGGTTTTCTTCCAGAGGAATTCTCAGTCAAGGTTAAGGTGCTTAGGGGGAGGCTTCGGCCTGTGGTTGGGTGACGAGGGTATCGTTAATGGTAAAACGAGCACACCATCTTCTATATTCTAACA
4D08_5	TGCAAGATACCAGATATATAAAATATGCAGAACAAAAATAAAATGCTAGAAATGGCAGA

	ATAACAGTAAAATATCCAGCATAATCTTCACAATACACAGAAATGATATAACAATAGAAATAGAAATAGCACTTTGATTCAGAAAGCAATGTTCATATGCCTCTTCGAGGGGGAAACCTCCAATCGAGCAACGCTCTGTTACAATGATATCCAAGCTAGGGTTCTCATACCCTCGACCCTTATTTATACCAAATATGTCTATTAAGAGAAGCATCTAGAGGAAGATCCTTCTAGAAGCGGAAGCTTCTATTAGGTCTTCTTACATCCACCTTATTCTATTTTATTGAAAGGCCTTATTAATATTGACCTTTATTACGTTTGACTACATTGATATACCTATGTGTGGGTTAGGACCTTAACA

Supplementary Table 3.17: Solo LTR and complete element abundance for three LTR-RT families across the *Picea abies* assembly, 20 fosmid sequences and 4 BAC sequences.

	P. abies assembly		*P. abies* fosmid sequences		*P. abies* BAC sequences	
LTR family	Complete LTR-RT element	Solo LTR	Complete LTR-RT element	Solo LTR	Complete LTR-RT element	Solo LTR
ALISEI	21	4	3	1	3	0
3K05	22	5	1	0	3	0
4D08_5	39	0	2	0	2	0

3.11 Phylogenetic analysis

In order to compare LTR-RT distribution and phylogeny, we analyzed genomic data from the 10 different plant species listed in Supplementary Table 3.18. We used sets of 100,000 454 reads for each of the species. 454 reads were available for the following species: *Abies sibirica, Gnetum genmon, Picea abies, Picea glauca, Pinus sylvestris, Taxus baccata, Juniperus communis, Populus tremula*. For the remaining species, simulated 454 reads were created starting from the available complete genome sequence using the perl script "fragsim". The perl script is available at the following link: http://www.bioinfo.ifm.liu.se/454tools/454sim.documentation.

Supplementary Table 3.18: Genomic data used for each species in the phylogenetic analysis.

Species	Genome size (Mbp)	Mean length 454 read (bp)	Total length considered (Mbp)
Picea abies	20,020	745.60	74.5
Picea glauca	19,800	519.99	51.9
Pinus sylvestris	22,400	585.97	58.5
Abies sibirica	15,560	614.31	61.4
Taxus baccata	11,230	620.55	62.0
Juniperus communis	11,700	624.71	62.4

Gnetum gnemon	3,700	590.82	59.0
Selaginella moellendorffii [17]	106	615	61.5
Physcomitrella [18]	479.9	615	61.5
Zea mays (B73) [19]	2,066	615	61.5
Oryza sativa [20]	374	615	61.5
Populus tremula	403	744	61.5

3.11.1 Identification of LTR-RT reverse transcriptase paralogs

The following strategy was adopted to detect paralogs of the Reverse Transcriptase domain specific for the Ty1-copia and Ty3-gypsy superfamilies among the datasets reported in Supplementary Table S3.18:

a) a representative 100 amino acid long sequence from the Reverse Transcriptase (RT) domain (Table S3.19) was used as query in TBLASTN searches (E-value 1e-5 or lower) against each 454 read dataset. The best significant hit covering at least 80% of the query sequence was extracted from the *P. abies* TBLASTN results;

b) the next best hit obtained in a) was used as query for a second TBLASTN search of all the 454 read datasets;

c) all positive hits from a) and b) covering at least 80% of the query were retrieved;

d) the paralogous sequences identified in all the datasets were aligned using MUSCLE [16] and used to build a RT Hidden Markov Model profile using the software HMMER (http://hmmer.org/);

e) the HMM profile was used to search each dataset using HMMER;

f) new positive hits (covering at least 80% of the HMM query) obtained from e) were added to those from step c). The number of total paralogs retrieved for each species is given in Supplementary Table 3.20.

g)

Supplementary Table 3.19: Sequences for Ty3-gypsy and Ty1-copia Reverse Transcriptase used as original queries in the phylogenetic analysis.

Superfamily	Reverse Transcriptase Sequence
Ty3-gypsy	EAYLDDLASRSRKRKDHPTHLRLIFERCRYFRIRLNPNKCSFCVTSGRLLGFIVS TTGIMVDPLKVGAIVQLPPPRTIVQLQSLQGKANFLRRFIANYAE
Ty1-copia	WKVYQMDVKSAFLNGYLEEEVYVQQPPRYEVRGQEDKVYRLKKALNGLKQ APRAWYSKIDSYMIKNEFIRSTSEPTLYTKVNEQGQILIVCLYVDDLIY

Supplementary Table 3.20: Reverse Transcriptase paralogs identified in each subset of 454 reads.

	Ty3-gypsy			Ty1-copia		
Species	TBLASTN	HMMER3	Total	TBLASTN	HMMER3	Total
Abies sibirica	874	44	918	498	12	510
Gnetum	511	192	703	73	3	76

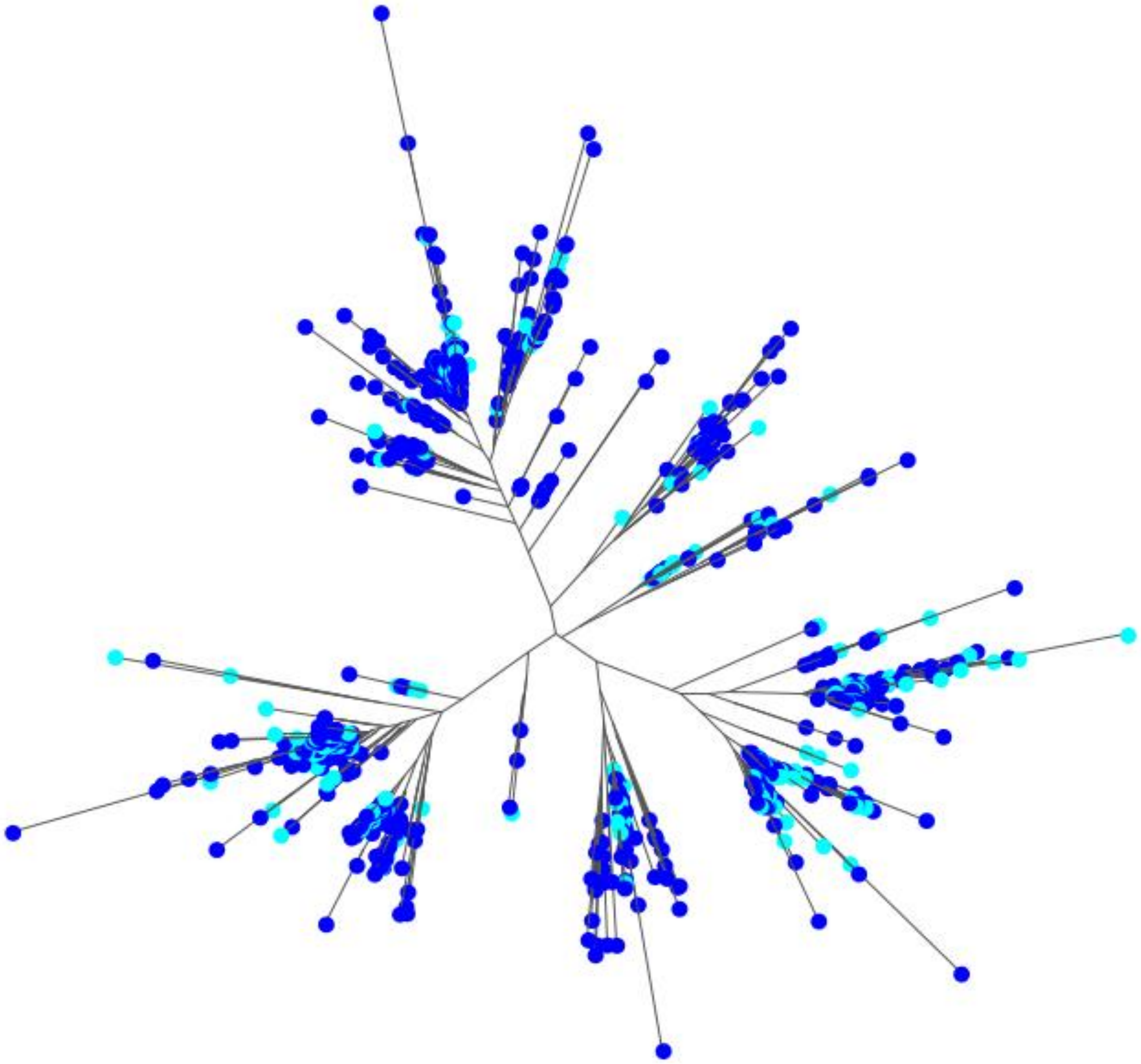
Juniper communis	500	20	520	497	49	546
Picea abies	1446	59	1505	636	50	686
Picea glauca	714	39	753	190	10	200
Pinus sylvestris	840	37	877	410	26	436
Taxus baccata	899	50	949	177	8	185
Populus tremula	222	10	232	203	13	216
Zea mays	1049	29	1078	818	28	846
Physcomitrella patens	1238	141	1379	167	21	188
Selaginella moellendorffii	541	45	586	35	1	36
Oryza sativa	314	18	332	113	3	116

3.11.2 Construction of phylogenetic trees

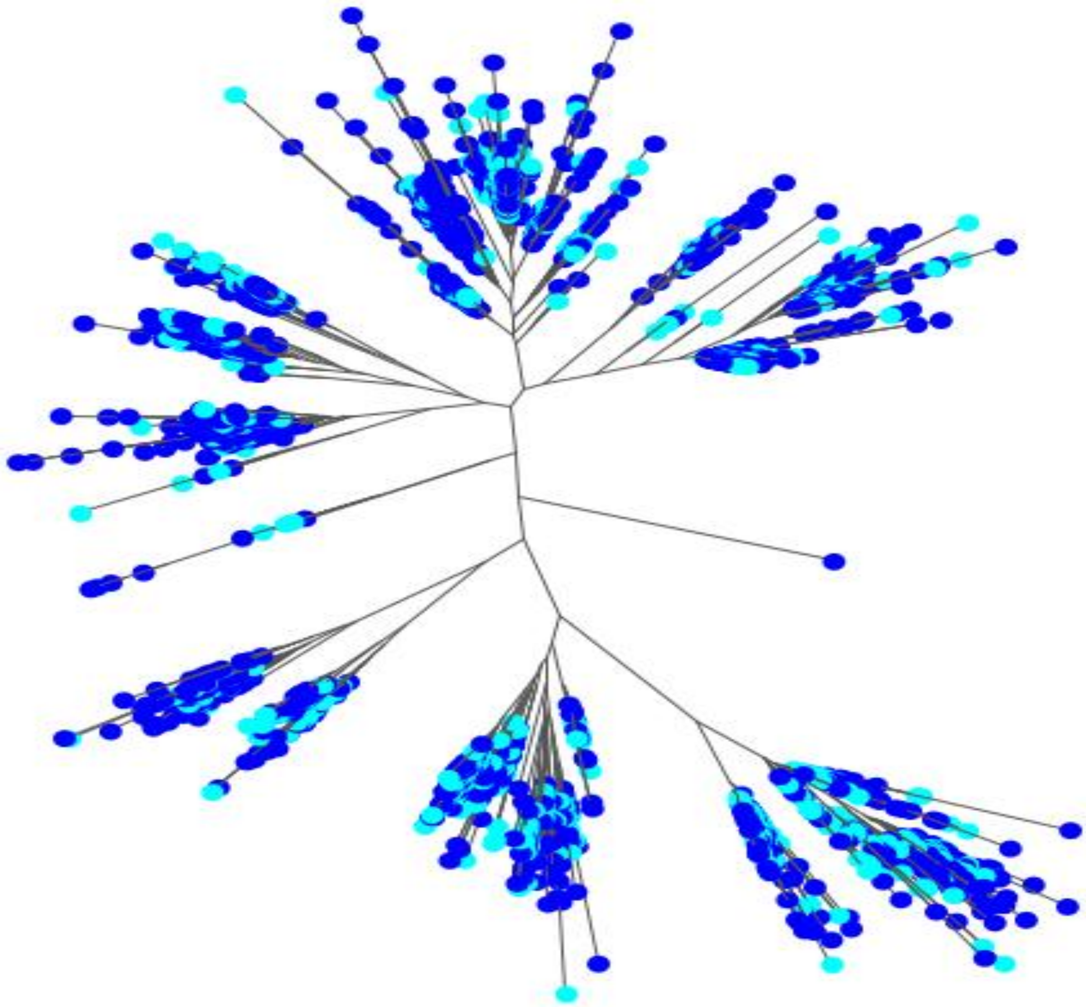
Multiple protein sequence alignments of paralogs obtained using the strategy described in the above paragraph were performed using MUSCLE and used to build (Heuristic)-NJ trees using FastTree [21] (default parameters).

(Heuristic)-NJ trees were also built for each single species for both Ty1-copia and Ty3-gypsy elements. Alignments in msf format and trees in Newick format are available at the ConGenIE ftp site (http://congenie.org). For all trees we provide the local bootstrap values calculated per 1000 replicates by FastTree.

Phylogenetic trees shown in Supplementary Figures 3.5 and 3.6 were obtained from multiple sequence alignments of RT paralogs belonging only to LTR-RTs identified in *P.abies* and *P. glauca* (the number of paralogous used are reported in Supplementary Table 3.17). The two phylogenetic trees are specific for the Ty1-copia and Ty3-gypsy superfamilies. For both superfamilies, most of the internal lineages with significant bootstrap values include elements from both species, indicating that they had mostly undergone amplification before speciation took place.



Supplementary Figure 3.5: Phylogenetic tree specific for *Picea abies* and *P. glauca* Ty1-copia RT paralogs.



Supplementary Figure 3.6: Phylogenetic tree specific for *Picea abies* and *P. glauca* Ty3-gypsy RT paralogs.

3.12 References

1. Price, A.L., Jones, N.C., Pevzner, P.A. 2005. *De novo* identification of repeat families in large genomes. *Bioinformatics,*Suppl 1:i351-8.

2. Smit, A.F.A., Hubley, R., Green, P. RepeatMasker Open-3.0, 1996-2010 http://www.repeatmasker.org.

3. Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.*, 9: 868-877.

4. Xu, Z., Wang, H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons, *Nucleic Acids Research,* 35(Web Server issue): W265-W268.

5. De Paoli, E. *et al.* 2012. Submitted.

6. Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26:680.

7. Jurka, J., Kapitonov, V., Pavlicek, A. *et al*. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogentic and Genome Research,* 110:462-467.

8. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool, *J. Mol. Biol*., 215: 403-410.

9. Grabher, M.G., Haas, B.J., Yassour. M., *et al*. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*., 15;29(7):644-52. doi: 10.1038/nbt.1883.

10. Sonnhammer, E.L. and Durbin, R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis, *Gene* 167: 1–10.

11. Rice, P., Longden, I., Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite, *Trends Genet.*, 16: 276-7.

12. Kimura, M. 1980. Simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences, *J. Mol. Evol.*, 16: 111-20.

13. Bouillé, M. and Bousquet, J. 2005. Trans-species shared polymorphisms at orthologous gene loci among distant species in the conifer *Picea* (*P*inaceae): implications for the long-term maintenance of genetic diversity in trees. *American Journal of Botany*, 92: 63-73.

14. SanMiguel, P., S. Gaut, B., Tikhonov, A., Nakajima, Y. & Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize. *Nature Genetics*, 20: 43-45.

15. Devos, K.M., Brown, J.K. and Bennetzen, J.L. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res*., 12: 1075–1079.

16. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* 32(5), 1792-97.

17. Banks, J.A., Nishiyama, T., *et al*. 2011. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science,* 332(6032):960-3.

18. Rensing, S,A., Lang, D., Zimmer, A.D., *et al*. 2008. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants*. Science*, 319(5859):64-9.

19. Schnable, P.S., Ware, D., Fulton, R.S., *et al*. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science,* 326(5956):1112-5.

20. Ouyang, S., Zhu, W., Hamilton, J., *et al*. 2007. The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res*., 35(Database issue):D883-7.

21. Price, M.N., Dehal, P.S., Arkin, A.P. 2009. FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, 26:1641-1650, doi:10.1093/molbev/msp077

Supplementary Material Section 4 Comparative analyses

4.1 Sequencing and assembly of the genomes of *Pinus sylvestris, Abies sibirica, Juniperus communis, Taxus baccata* and *Gnetum gnemon*

Five gymnosperm species were sequenced to a low-to-medium coverage to enable comparative analyses to *P. abies*. The species were selected to represent important branch points of the gymnosperm lineage as well as include a range of genome sizes. In all cases, mature needles were sampled and used to extract DNA to perform shotgun sequencing using the DNA extraction protocol described in Supplementary Information 1. The samples used were:

> *Abies sibirica:* A tree of unknown origin growing on the Umeå University Campus; 63°49´08.93´´N, 20°18´40.95´´E
>
> *Pinus sylvestris:* Clone W4009 from the Swedish pine breeding program growing at the Skogforsk tree archive, Uppsala, Sweden
>
> *Juniperus communis:* A wild tree growing on the Umeå University Campus; 63°49´08.94´´N, 20°18´40.96´´E.
>
> *Taxus baccata:* sampled from Bergianska Gardens Stockholm. (originally from Ramlösa, 1951) Location: Parken 36: 5, named f. columna-suecica. Received from from H. Wanntorp, 1981.
>
> *Gnetum gnemon:* Sampled from Bergianska Gardens Stockholm. Originally from P. Litfors, Botany Department, Stockholm University, 1997.

The five samples were sequenced using standard Illumina protocols and kits on the Illumina HiSeq 2000 platform at the Beijing Genome Institute (BGI, Shenzhen, China). Paired-end libraries (2 x 100 bp reads) with a mean insert size of approximately 300 bps were produced and raw read data was quality control filtered using the BGI in-house filtering pipeline. Four species were sequenced to a relatively low coverage (*A. sibirica, J. communis, T. baccata,* and *G. gnemon*), while *P. sylvestris* was sequenced to a higher coverage (Table 4.1). The genome assemblies were performed using CLC Assembly Cell (version 3.22.56754) with default parameters. As expected, the assemblies produced from such low coverage genome data were highly fragmented (Supplementary Table 4.1). The *P. sylvestris* assembly was larger overall, while the *G. gnemon* assembly produced longer contigs, most likely because *G. gnetum* has a substantialy smaller genome.

The five low coverage genome assemblies are available at the ConGenIE (http://congenie.org) ftp site.

Table 4.1: Assembly statistics for the five assembled gymnosperm genomes. NG50 indicates the N50 contig length of the expected genome size.

Species	Coverage	Assembly size (Gbp)	No. of contigs (million)	NG50 (bp)
Abies sibirica	3.8	2.183	4.6	556
Pinus sylvestris	12.5	6.795	16.1	447
Juniperus communis	4.5	1.861	4.7	406
Taxus baccata	4.0	3.001	5.9	611
Gnetum gnemon	5.5	1.837	1.9	1735

4.2 Intron length comparisons

Intron lengths were compared between *P. abies*, *P. sylvestris*, and *G. gnemon* using a procedure where exons representing the Coding DNA Sequences (CDSs) predicted *ab initio* of the *P. abies* High Confidence gene set (Supplementary Information 2) were aligned to the other two genomes to identify orthologous locations. The exons of *P. abies* genes containing >1 exon were used as query sequences with the program megablast in the BLAST+ suite, version 2.2.25[1], using the command "blastn -task dc-megablast" and 1e-20 as "evalue" parameter. Intron lengths were calculated as the distances between aligned contiguous exons in the subject genomes. Thus, if exons 1,2, and 4 of a gene were aligned, only the intron between exons 1 and 2 could be used in the analysis.

For each of the two subject species, the blast output was filtered to keep a maximum of one alignment per exon, with the requirement that the alignment to keep had a bit score >=20% higher than the second highest bit score. Additional requirements were that the alignment must be >=25 bps, with alignment identity >60 % and corresponding to >80 % of the exon length; finally, the length of an exon minus the length of its alignment must be <150 bps.

In the next step, the subject genome assembly contig with the largest number of such filtered alignments placed on a single strand was identified for each gene. The analysis pipeline then attempted to identify additional exons not yet aligned using a set of blast results using a more lenient e-value threshold (1^{e1}). These alignments were also filtered according to the same requirements with this analysis step often creating longer contiguous stretches of aligned exons (to refer back to the initial example, by providing the alignment of exon 3).

The intron lengths were calculated as the distances between aligned exons, minus the missing parts of the two exons flanking the intron (in the cases where alignment length was shorter than exon length; if *vice versa*, the length of the overhang was instead added).

4.3 Intron comparisons statistics

Introns of 50-300 bp are here named “short” and those of length 1-20 kbp “long”.

16,781 *P. abies* genes contained a total of 57,241 introns. 6,889 genes (24.3% of all genes) contained long introns; 14,176 (50.0%) contained short introns. Genes with long introns contained a mean of 1.43 long introns and a total of 4.62 introns of any length; genes with short introns contained a mean of 2.64 short introns a mean total of 3.78 introns.

The following ortholog statistics are lower bound estimates of true intron counts due to the limited number of cases where exons from a gene aligned within a single subject species genome assembly contig as a result of the high degree of fragmentation within those assemblies.

G. gnemon

3,176 short introns were identified in 1,471 genes; 506 long introns were identified in 409 genes. Thus, 10.4% of all genes that contained a short intron in *P. abies* also contained a short intron in *G.gnemon,* while 5.9% of all genes that contained a long intron in *P. abies* also contained a long intron in *G.gnemon*.

P. sylvestris

8,612 short introns were identified in 4,576 genes; 496 long introns were identified in 474 genes. Thius 32.3% of all genes containing a short intron in *P. abies* also contained a short intron in *P. sylvestris*; 6.9% of all genes that contained a long intron in *P. abies* also contained a long intron in *P. sylvestris.*

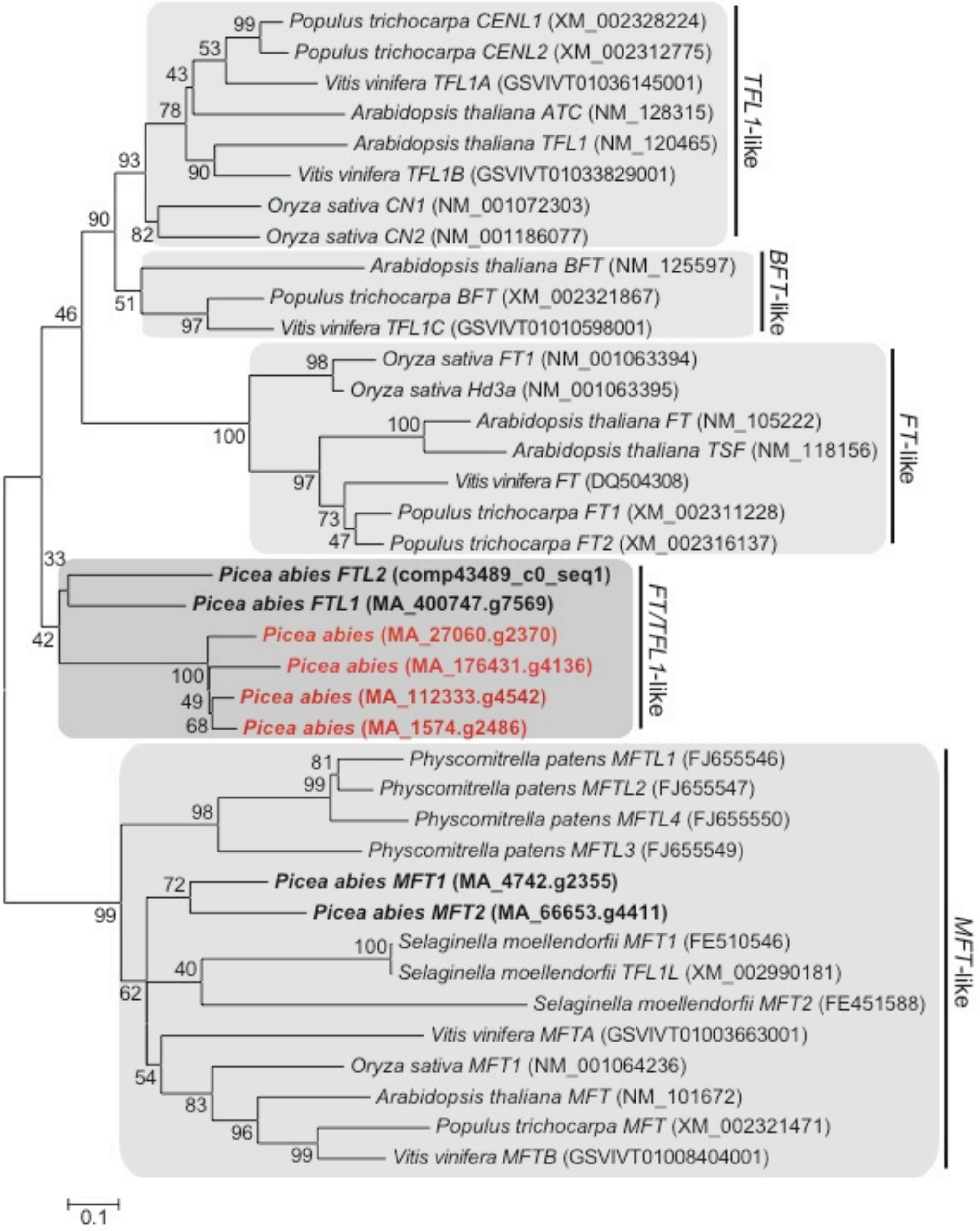
4.4 References

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J Mol Biol 215, 403-410 (1990).

Supplementary Material Section 5: Evolution of important conifer traits

5.1 Molecular phylogenetic analyses of phosphatidylethanolamine-binding protein (PEBP) genes in plants.

All analyses were conducted in MEGA5 [1] using full coding regions of *PEBP* genes from selected plant species (*Arabidopsis thaliana, Populus trichocarpa, Vitis vinifera, Oryza sativa, Picea abies, Selaginella moellendorffii* and *Physcomitrella patens*). Gene accession numbers for Genbank (via NCBI), *V. vinifera* genome (via Phytozome), and *P. abies* genome (via Congenie) are depicted next to the sequences. Newly identified putative *FT/TFL1*-like genes in the *P.abies* genome are marked in bold red, previously known *FT/TFL1*-like genes and *MFT* genes are marked in bold black. Multiple sequence alignments were created with MUSCLE [2]. The phylogenetic reconstruction was inferred by using the Maximum Likelihood method based on the best substitution model (Kimura 2-parameter model +G +I) [3] with 1000 bootstrap replications. All positions containing gaps were eliminated. The tree is drawn to scale and the percentage of trees in which the associated taxa clustered together is shown next to the branches.


Populus trichocarpa CENL1 (XM_002328224)
Populus trichocarpa CENL2 (XM_002312775)
Vitis vinifera TFL1A (GSVIVT01036145001)
Arabidopsis thaliana ATC (NM_128315)
Arabidopsis thaliana TFL1 (NM_120465)
Vitis vinifera TFL1B (GSVIVT01033829001)
Oryza sativa CN1 (NM_001072303)
Oryza sativa CN2 (NM_001186077)
Arabidopsis thaliana BFT (NM_125597)
Populus trichocarpa BFT (XM_002321867)
Vitis vinifera TFL1C (GSVIVT01010598001)
Oryza sativa FT1 (NM_001063394)
Oryza sativa Hd3a (NM_001063395)
Arabidopsis thaliana FT (NM_105222)
Arabidopsis thaliana TSF (NM_118156)
Vitis vinifera FT (DQ504308)
Populus trichocarpa FT1 (XM_002311228)
Populus trichocarpa FT2 (XM_002316137)
Picea abies FTL2 (comp43489_c0_seq1)
Picea abies FTL1 (MA_400747.g7569)
Picea abies (MA_27060.g2370)
Picea abies (MA_176431.g4136)
Picea abies (MA_112333.g4542)
Picea abies (MA_1574.g2486)
Physcomitrella patens MFTL1 (FJ655546)
Physcomitrella patens MFTL2 (FJ655547)
Physcomitrella patens MFTL4 (FJ655550)
Physcomitrella patens MFTL3 (FJ655549)
Picea abies MFT1 (MA_4742.g2355)
Picea abies MFT2 (MA_66653.g4411)
Selaginella moellendorfii MFT1 (FE510546)
Selaginella moellendorfii TFL1L (XM_002990181)
Selaginella moellendorfii MFT2 (FE451588)
Vitis vinifera MFTA (GSVIVT01003663001)
Oryza sativa MFT1 (NM_001064236)
Arabidopsis thaliana MFT (NM_101672)
Populus trichocarpa MFT (XM_002321471)
Vitis vinifera MFTB (GSVIVT01008404001)
TFL1-like
BFT-like
FT-like
FT/TFL1-like
MFT-like
99
53
43
78
93
90
90
82
51
97
46
98
100
100
97
73
47
33
42
100
49
68
81
99
98
72
99
100
40
62
54
83
96
99
0.1

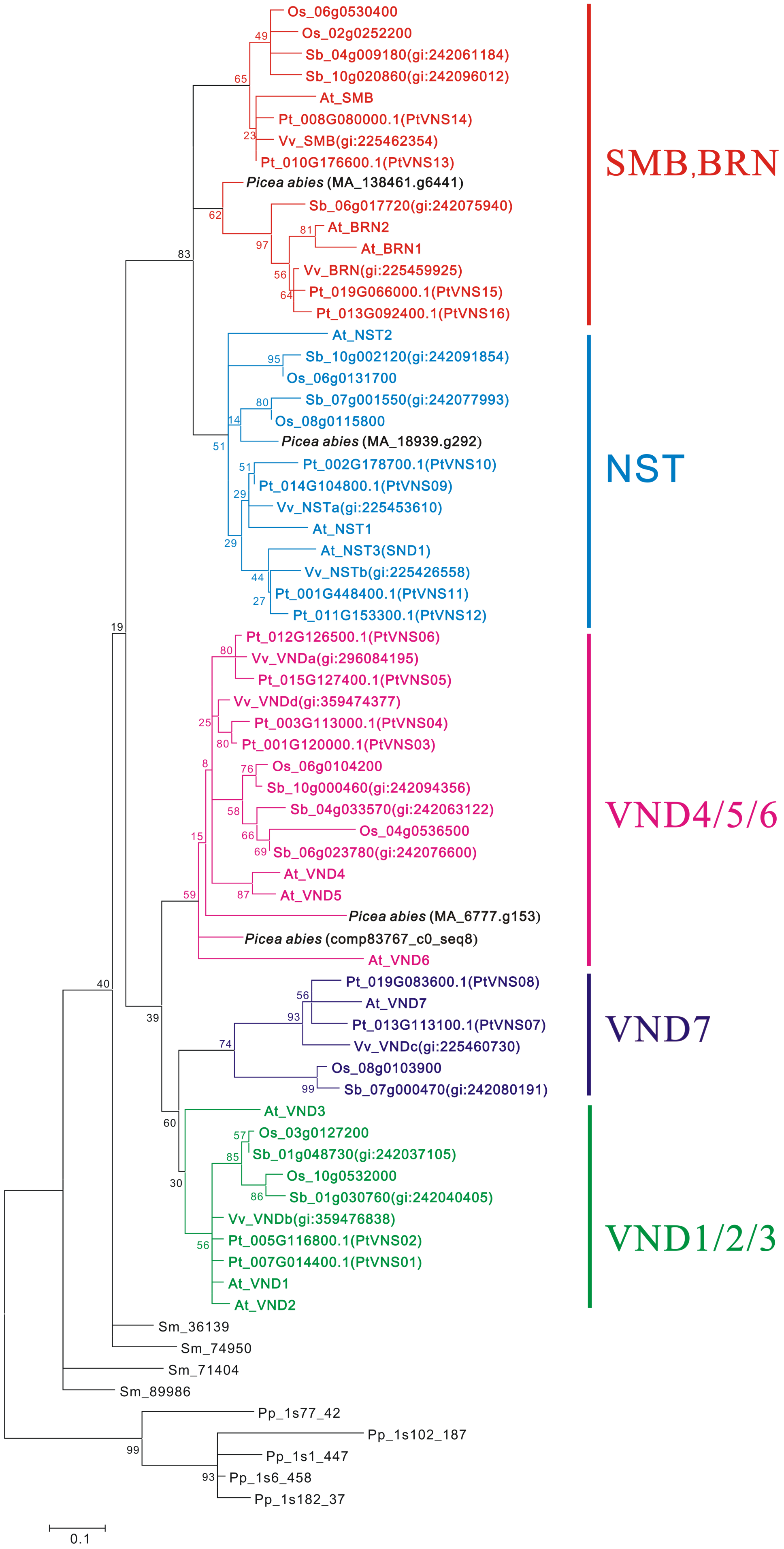
5.2 Molecular phylogenetic analyses of plant MADS domains.

Included are all MADS domain sequences from *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, *Vitis vinifera* and *Picea abies* (gene names starting with At, Os, Pt, Vv and Pa, respectively) and MADS domain sequences known from other gymnosperms (GGMxx from *Gnetum gnemon*, CjMADSx from *Cryptomeria japonica*, Prxxx from *Pinus radiata*, GpMADSxx from *Gnetum parvifolium*, GBM5 from *Gingko biloba*, CyAG from *Cycas edentate*, SAG1a and SMADS42B from *Picea mariana*). Clades belonging to Type I MADS domains are coloured in green and clades belonging to Type II MADS domains are coloured in blue. MADS domains from *Picea abies* are highlighted in bold. The alignment of the MADS domain sequences was conducted using MAFFT (Katoh *et al.* 2002) and the phylogeny was reconstructed using PhyML (Guindon *et al.* 2010).

[See external Supplementary Figure 5.2 - “Supplementary figure 5.2 - MADS domains.png”]

5.3 Molecular phylogenetic analyses of SMB/NST/VND genes in plants.

All analyses were conducted in MEGA5 [1]. Protein sequence of the Class IIB of the NAC transcription factor family (SMB/NST/VND) from selected plant species (*Arabidopsis thaliana* At*, Populus trichocarpa* Pt*, Vitis vinifera* Vv*, Oryza sativa* Os*, Sorghum bicolor* Sb, *Selaginella moellendorffii* Sm*,* and *Physcomitrella patens* Pp) were obtained from Phytozome (http://www.phytozome.net). Putative SMB/NST/VND orthogoues in the *P.abies* genome are marked in bold black. Multiple sequence alignments were created with MUSCLE [2]. The phylogenetic reconstruction was inferred by using the Maximum Likelihood method based on the Jones-Tarlor-Thorn substitution model with gamma rate distribution [3]. The numbers at the nodes indicate bootstrap support calculated by RAxML bootstrapping using 1000 replications.


Os_06g0530400
49
Os_02g0252200
Sb_04g009180(gi:242061184)
65
Sb_10g020860(gi:242096012)
At_SMB
Pt_008G080000.1(PtVNS14)
23
Vv_SMB(gi:225462354)
Pt_010G176600.1(PtVNS13)
Picea abies (MA_138461.g6441)
62
Sb_06g017720(gi:242075940)
81
At_BRN2
97
At_BRN1
83
56
Vv_BRN(gi:225459925)
64
Pt_019G066000.1(PtVNS15)
Pt_013G092400.1(PtVNS16)
SMB,BRN
At_NST2
95
Sb_10g002120(gi:242091854)
Os_06g0131700
80
Sb_07g001550(gi:242077993)
14
Os_08g0115800
Picea abies (MA_18939.g292)
51
51
Pt_002G178700.1(PtVNS10)
Pt_014G104800.1(PtVNS09)
29
Vv_NSTa(gi:225453610)
At_NST1
29
At_NST3(SND1)
Vv_NSTb(gi:225426558)
44
Pt_001G448400.1(PtVNS11)
27
Pt_011G153300.1(PtVNS12)
NST
19
Pt_012G126500.1(PtVNS06)
80
Vv_VNDa(gi:296084195)
Pt_015G127400.1(PtVNS05)
Vv_VNDd(gi:359474377)
25
Pt_003G113000.1(PtVNS04)
80
Pt_001G120000.1(PtVNS03)
8
76
Os_06g0104200
Sb_10g000460(gi:242094356)
58
Sb_04g033570(gi:242063122)
66
Os_04g0536500
15
69
Sb_06g023780(gi:242076600)
At_VND4
59
87
At_VND5
Picea abies (MA_6777.g153)
Picea abies (comp83767_c0_seq8)
At_VND6
VND4/5/6
40
Pt_019G083600.1(PtVNS08)
56
At_VND7
39
93
Pt_013G113100.1(PtVNS07)
74
Vv_VNDc(gi:225460730)
Os_08g0103900
99
Sb_07g000470(gi:242080191)
VND7
At_VND3
60
57
Os_03g0127200
85
Sb_01g048730(gi:242037105)
30
Os_10g0532000
86
Sb_01g030760(gi:242040405)
Vv_VNDb(gi:359476838)
56
Pt_005G116800.1(PtVNS02)
Pt_007G014400.1(PtVNS01)
At_VND1
At_VND2
VND1/2/3
Sm_36139
Sm_74950
Sm_71404
Sm_89986
Pp_1s77_42
Pp_1s102_187
99
Pp_1s1_447
93
Pp_1s6_458
Pp_1s182_37
0.1

5.4 References

Edgar R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32 (5): 1792-1797.

Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol **59**:307-321.

Katoh, K., K. Misawa, K. Kuma, and T. Miyata. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059-3066.

Kimura M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. of Mol. Evol. 16:111-120.

Tamura K., Peterson D., Peterson N., Stecher G., Nei M., and Kumar S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731-2739.

Supplementary Material Section 6: Data availability and accession numbers

All raw and processed data presented, including the *P.abies* 1.0 assembly, *ab initio* gene prediction sets and short-read *do novo* transcript assemblies, are available from the ftp site of the ConGenIE (Conifer Genome Integrative Explorer) web resource (http://congenie.org). ConGenIE includes a genome browser, BLAST server, and expression visualisation tools.

Data have also been deposited in the ENA database under these accession/project numbers:

6.1 Raw sequence data:

Picea abies DNA data: ERP002565

Picea abies mRNA data: ERP002475

Picea abies sRNA data: ERP002476

Abies sibirica DNA data: ERP002568

Gnetum gnemon DNA data: ERP002569

Juniperus communis DNA data: ERP002570

Picea glauca DNA data: SRX268449

Pinus sylvestris DNA data: ERP002572

Taxus baccata DNA data: ERP002571

6.2 Assemblies

The project PRJEB1822 contains:

- the *Picea abies* 1.0 assembly
- 20 individually sequenced *Picea abies* fosmids

The *Picea abies* chloroplast: PRJEB1695

10 individually sequenced *Picea glauca* BACs: GenBank Acc No KF008668 - KF008677