ARTICLE

OPEN

doi:10.1038/nature12211

The Norway spruce genome sequence and conifer genome evolution

Lists of authors and their affiliations appear at the end of the paper

Conifers have dominated forests for more than 200 million years and are of huge ecological and economic importance. Here we present the draft assembly of the 20-gigabase genome of Norway spruce (*Picea abies*), the first available for any gymnosperm. The number of well-supported genes (28,354) is similar to the >100 times smaller genome of *Arabidopsis thaliana*, and there is no evidence of a recent whole-genome duplication in the gymnosperm lineage. Instead, the large genome size seems to result from the slow and steady accumulation of a diverse set of long-terminal repeat transposable elements, possibly owing to the lack of an efficient elimination mechanism. Comparative sequencing of *Pinus sylvestris*, *Abies sibirica*, *Juniperus communis*, *Taxus baccata* and *Gnetum gnemon* reveals that the transposable element diversity is shared among extant conifers. Expression of 24-nucleotide small RNAs, previously implicated in transposable element silencing, is tissue-specific and much lower than in other plants. We further identify numerous long (>10,000 base pairs) introns, gene-like fragments, uncharacterized long non-coding RNAs and short RNAs. This opens up new genomic avenues for conifer forestry and breeding.

Gymnosperms are a group of land plants comprising the extant taxa, cycads, *Ginkgo*, gnetophytes and conifers. Gymnosperms first appeared more than 300 million years ago (Myr ago)[1], well before the angiosperm lineage separated from the stem group of extant gymnosperms[2]. The major radiation of conifer families occurred 250–65 Myr ago[3], and during their evolution the morphology of conifers has changed relatively little. There are approximately 630 conifer species, representing about 70 currently recognized genera, which dominate many terrestrial ecosystems, especially in the Northern Hemisphere. Conifers also dominated both before and after the major mass extinction events at the end of the Permian and Cretaceous periods, around 250 and 65 Myr ago, respectively. Conifers are of immense ecological and economic value; coniferous forests cover enormous areas in the Northern Hemisphere, and conifers are keystone species in many other ecosystems. Conifers contribute a large fraction of terrestrial photosynthesis and biomass, and the cultural and economic values of conifers are also paramount; early civilizations used conifers for firewood, tools and artefacts and today several national economies depend on commodities produced from conifers. However, despite their abundance and importance, our understanding of conifer genomes is limited. Most conifers have 12 ($2n = 24$) chromosomes, probably reflecting the ancestral karyotype[4], which are typically of similar size, each being roughly comparable to the size of the human genome, and containing high proportions of repetitive elements[5,6]. The gene space of conifer genomes has not been well characterized, although several reports have suggested that gene families in conifers may be larger than their angiosperm counterparts[7] and that conifer genomes contain numerous pseudogenes[8].

Because their genomes are among the largest—typically 20–30 gigabases pairs (Gb)—of all organisms, genome-wide analyses of conifers are particularly challenging. Thus, no full genome sequence of a gymnosperm species is available at present, whereas 30 angiosperm and more basal plant genomes have been sequenced. However, size is not the only challenge to sequencing presented by conifer genomes. Conifers are typically outbreeding, produce wind-dispersed pollen, have very large effective population sizes, and their genomes are highly heterozygous, although their nucleotide substitution rates are lower than those of most angiosperms[8,9], perhaps owing to long lifespan (decades to centuries). Furthermore, inbreeding depression negates the production of inbred lines that could facilitate genome assembly.

The availability of conifer genome sequences would enable comparative analyses of genome architecture and the evolution of key traits for seed plants, including flower or fruit development and life history (perennial versus annual), and help to determine how and why conifer genomes became so large. To address these issues and aid forest tree breeding, biodiversity and conservation studies by, for example, enabling the genome-wide design of genetic markers, we used data from massively parallel DNA sequencing to assemble a draft of the 20-Gb nuclear genome of Norway spruce (*Picea abies* (L.) Karst), one of the most widespread, ecologically and economically important plants in Europe. We analysed the protein-coding and non-coding fractions of the genome and compared them to the low-coverage draft genome assemblies of five other gymnosperms—Scots pine (*P. sylvestris*), Siberian fir (*A. sibirica*), juniper (*J. communis*), yew (*Taxus baccata*) and *Gnetum gnemon*—to gain insight into conifer genome evolution.

Sequencing and assembly

We sequenced a 43-year-old root-grafted copy of the *P. abies* clone Z4006, which originated from a tree in Ragunda, central Sweden, collected in 1959. Many copies of this clone are available in clone archives and seed orchards, and it has been extensively used in Swedish breeding programs. We estimated its genome size to be 19.6 Gb ($C = 20.02 \pm 0.95$ pg (mean $\pm$ s.d.); Supplementary Information 1.1), in accordance with previous reports[10].

De novo sequencing and assembly of large, repeat-containing, heterozygous genomes remains challenging. To assemble the *P. abies* genome, we developed a hierarchical strategy combining fosmid pools[11] with both haploid and diploid whole genome shotgun (WGS) data, and RNA sequencing (RNA-Seq) data[12–14] (Supplementary Information 1.2–1.3). The resulting assembly (P.abies 1.0) included 4.3 Gb in >10-kilobase (kb) scaffolds (Table 1), and we estimated that approximately 63% of protein-coding genes[15] were fully covered (>90% of their length), and 96% partially covered (>30% of their length) within single scaffolds (Supplementary Information 1.4). By mapping diploid reads to the P.abies 1.0 assembly, the single nucleotide

polymorphism (SNP) frequency was estimated to be 0.77% and the short insertion/deletion (indel) frequency to be 0.05% (Supplementary Information 1.5).

The chloroplast genome (124 kb) revealed considerable structural variation within the genus *Picea* (Supplementary Information 1.6). The draft mitochondrial genome (>4 Mb) was among the largest reported for plants and was rich in short open-reading frames (ORFs), which appeared to be gene remnants derived from repeat-driven mitochondrial rearrangements[16] (Supplementary Information 1.7).

Presence of long introns and gene-like fragments

We generated >1 billion RNA-Seq reads and used transcript assemblies of these in combination with public expressed sequence tags (ESTs) and transcripts to perform *ab initio* prediction of protein-coding genes, which identified a high confidence set of 28,354 loci with >70% coverage by supporting evidence from the total set of 70,968 predicted loci. A notable characteristic of the predicted gene structures was the presence of numerous long introns (Fig. 1b), with mean intron length being higher than in most available plant genomes, although similar to the repeat-rich genomes of *Vitis vinifera* and *Zea mays*[17,18]. The longest intron in the high-confidence genes was 68 kb (Supplementary Table 2.6), and 2,384 high-confidence genes contained 2,880 longer than 5-kb introns (20 of which we confirmed by PCR amplification; Supplementary Information 2.14), 2,679 of which contained a repeat, suggesting that repeat insertions account for intron expansion. By contrast, exon size was consistent among the species considered (Supplementary Information 2.6.3). Numerous genes (~30%) remained split across scaffolds owing to assembly fragmentation, and as such, the longest introns were not represented in the P.abies 1.0 assembly. Long introns (either individual or cumulative intron length) did not influence expression levels (Fig. 1c) and introns containing repeats have not contracted despite a lack of recent repeat activity (see below).

Analysis of gene families in the high-confidence gene set and seven sequenced plant genomes (five angiosperms: *Arabidopsis thaliana*, *Populus trichocarpa*, *Vitis vinifera*, *Oryza sativa* and *Zea mays*, and two basal plants: *Selaginella moellendorffii* and *Physcomitrella patens*) identified 1,021 *P. abies*-specific gene families (Fig. 1a and Supplementary Information 2.8). *P. abies*-specific families included over-representation of Gene Ontology categories involved in DNA repair and methylation of DNA and chromatin (Supplementary Information 2.8). As for most draft genomes, these results probably overestimate gene numbers[19] and will be refined as we further improve the genome assembly.

Table 1 | Characteristics of the *P. abies* genome

Genome	
Size (1*n*)	19.6 Gb
Karyotype	2*n* = 24
GC content	37.9%
High-copy repeat content*	
LTR_*Gypsy*/*Copia*/unclassified	35%/16%/7%
LINE	1%
DNA transposable element	1%
Unclassified	10%
Genes and gene-like fragments†	2.4%
Assembly	
Size in scaffolds >200 bp/>10 kb	12 Gb/4.3 Gb
N50/NG50	4,869 bp/721 bp
Annotation	
High confidence gene set	28,354
Genes with >5-kb introns	8.4%
Avg. exon/intron size	312 bp/1,017 bp
Avg. gene density	1 gene in 705 kb
Transposable element genes	284,587
Non-coding loci	
lncRNA (unique/conserved)	13,031/9,686
miRNA (*de novo* predicted)	2,719

* Inferred from unassembled reads. †Including pseudogenes, excluding transposable elements.

A common mechanism leading to genome size expansion is the occurrence of a whole genome duplication (WGD) event. We calculated the number of synonymous substitutions per synonymous site (K_s) of paralogues within the high-confidence genes but found no evidence for any recent WGD; there was a clear, exponential decay in the number of retained paralogues with increasing K_s values (Supplementary Information 2.9 and Supplementary Fig. 2.6). However, a population dynamics model that takes into account both small- and large-scale modes of gene duplication[20] suggested the presence of a small peak (around K_s of 1.1), which, considering the slow substitution rate of conifers, might represent the ancient WGD predating the divergence of angiosperms and gymnosperms (350 Myr ago[21]).

Previous examinations of small genomic subsets indicated that conifer genomes contain numerous pseudogenes[5,6,22,23]. The gene-like fraction of the P.abies 1.0 assembly was identified by alignment of RNA-Seq reads and *de novo* assembled transcripts (Supplementary Information 2.10). Within this subset of the genome, loci with valid spliced alignments of *de novo* assembled transcripts or the presence of a high-confidence gene were also identified. The high-confidence gene set represented 27 Mb of protein-coding sequence, whereas 72 Mb of regions were identified with a valid spliced alignment or a high-confidence gene. In stark contrast, 524 Mb of gene-like regions were identified by less stringent alignments. The presence of such a large gene-like fraction lacking predicted gene structures supports the presence of numerous pseudogenes.

Recent ENCODE publications[24,25] characterized numerous long non-coding RNA (lncRNA) loci in the human genome, but this class of RNA remains largely uncharacterized in plants. Using short-read *de novo* transcript assemblies, 13,031 spruce-specific and 9,686 conserved intergenic lncRNAs were identified (Supplementary Information 2.4.3). In common with the ENCODE results, *P. abies* lncRNA loci contained fewer exons, were shorter (Fig. 1c), and had more tissue-specific expression than protein-coding loci (Supplementary Fig. 2.8).

There has been conflicting evidence about the presence of 24-nucleotide short RNAs (sRNAs) in gymnosperms[26–29], a class of sRNA that silence transposable elements by the establishment of DNA methylation[30]. Across 22 samples, we identified numerous 24-nucleotide sRNAs, but these were highly specific to reproductive tissues, largely associated with repeats but present at substantially lower levels than in angiosperms (Fig. 1d and Supplementary Fig. 2.10). By contrast, 21-nucleotide sRNAs were associated with genes, repeats and promoters/untranslated regions (UTRs) (Fig. 1d). *De novo* microRNA (miRNA) prediction identified 2,719 loci, including 20 known miRNA families, with target sites predicted within the high-confidence gene set for 1,378 of these (Supplementary Information 2.13). Furthermore, 55 known miRNA families had >5 aligned sRNA reads and mature miRNAs, representing 49 known families aligned to the genome (Supplementary Information 2.13).

Conifer genomes grew by insertion of LTR–RT elements

We constructed a manually curated library of 1,773 repetitive sequences, approximately half of which could be assigned to known transposable element repeat families (Supplementary Information 3.1–3.3). Long terminal repeat-retrotransposons (LTR-RTs) comprised the most abundant fraction of transposable elements, with the *Ty3*/*Gypsy* superfamily being more abundant than the *Ty1*/*Copia* superfamily (Fig. 2a and Table 1). We also identified and characterized transposable elements using 454 reads from randomly sheared genomic DNA in five other gymnosperms (*P. sylvestris*, *A. sibirica*, *J. communis*, *T. baccata* and *G. gnemon*) and, in all six species, LTR-RTs were the most abundant class (Fig. 2a, Supplementary Information 4.1 and Supplementary Table 3.1).

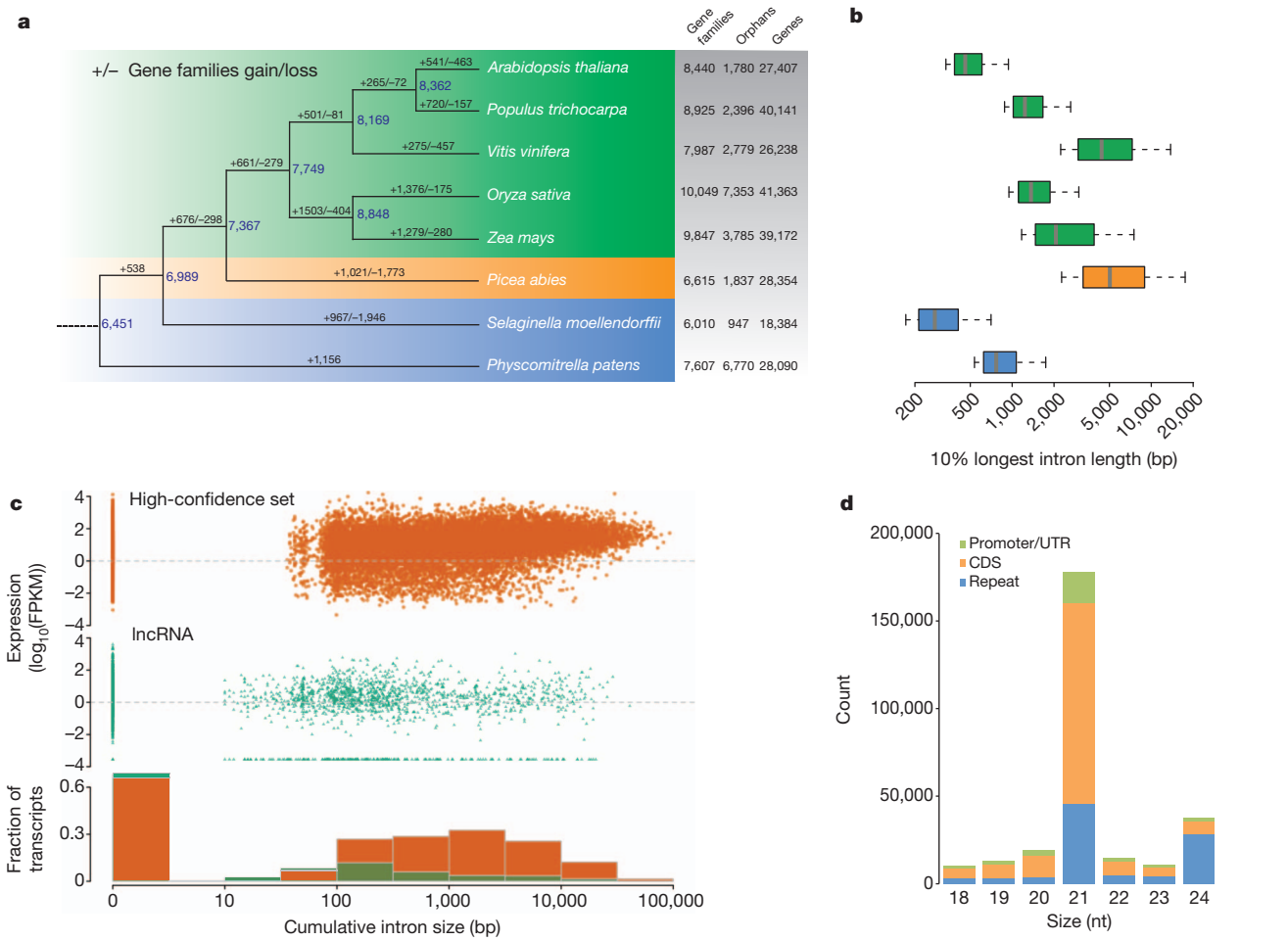

Figure 1 | The gene-space and transcribed fraction of the P.abies 1.0 assembly. **a**, Gene family loss and gain in eight sequenced plant genomes (*Arabidopsis thaliana*, *Populus trichocarpa*, *Vitis vinifera*, *Oryza sativa*, *Zea mays*, *Picea abies*, *Selaginella moellendorffii* and *Physcomitella patens*). Gene families were identified using TribeMCL (inflation value 4), and the DOLLOP program from the PHYLIP package was used to determine the minimum gene set for ancestral nodes of the phylogenetic tree. We used plant genome annotations filtered to remove transposable elements. 'Orphans' refers to gene families containing only a single gene. Blue numbers indicate the number of gene families. **b**, Boxplot representation of length distribution for the 10% longest introns in the same eight genomes. **c**, Scatter plots of cumulative intron length against $\log_{10}$ expression calculated as fragments per kilobase per million mapped reads (FPKM) for high-confidence gene loci (top, coloured orange) and green for lncRNA loci (middle, shaded green). The bottom panel shows a histogram of cumulative intron size in the two sets of loci. **d**, Distribution of small (18–24-nucleotide (nt)) RNAs and their co-alignment-based colocation to genomic features (repeats, high-confidence genes and their promoter/UTRs). CDS, coding sequence.

To trace the history of transposable elements in vascular plants we inferred phylogenies of a domain of the reverse transcriptase genes of both *Ty1/Copia* and *Ty3/Gypsy* elements. The phylogenies revealed several diverse and ancient transposable element subfamilies, present in almost all of the examined conifer genera, whereas only a few subfamilies were expanded in the angiosperm genomes (Fig. 2b, c and Supplementary Information 3.11). Most internal clades with significant bootstrap support were consistently species-specific, indicating that most expansions of extant transposable element families occurred after divergence. Two species-specific amplification bursts were evident: a *Ty1/Copia* family in *J. communis* and a *Ty3/Gypsy* family in *T. baccata*. We used complete LTR-RTs from *P. abies* and *P. glauca* to investigate further the timing of conifer transposable element insertions[31] (Supplementary Information 3.4–3.8). In contrast to a similar set of elements identified in *Oryza sativa* and *O. glaberrima* (Fig. 2d), we detected no evidence of recent activity (that is, less than 5 Myr ago) in *P. abies*. Instead, insertions seem to have occurred over several tens of millions of years (older insertions are more likely to escape detection). Analysis of 68 orthologous transposable element insertions in *P. abies* and *P. glauca* further supported this: 63 insertions apparently predated divergence, and only five occurred after the lineages separated 13–20 Myr ago (Supplementary Information 3.9).

We clustered LTRs of complete elements to identify transposable element families[32]. More than 86% of the elements remained as singletons, indicating that LTR-RTs are quite divergent and that there are several low-abundance families. We searched three LTR-RT families for signatures of unequal intra-element recombination events in scaffolds >50 kb and 20 complete fosmids[33]. For families ALISEI, 3K05 and 4D08_5 we identified 21, 22 and 39 complete elements, and four, five and no solo LTRs, respectively (Supplementary Information 3.10). Although this data set is limited, the analysis suggested that LTR-RT-related sequences might be removed less frequently by unequal recombination than in other plant genomes. The ratio of solo-LTRs to complete elements in *P. abies* is ~1:9, whereas in *A. thaliana*, rice and barley it is 1:1 (ref. 33), 0.6:1 (ref. 34) and 16:1 (for the abundant BARE 1 element[35]), respectively. Taken together, these findings indicated that the extant set of transposable elements in *P. abies* accumulated slowly over tens or hundreds of millions years, mainly by the insertion of LTR-RT elements with limited transposable element removal.

An analysis of introns across taxa provided further insight into the genome of the last common ancestor to the conifers. We identified orthologues of normal sized (50–300 bp) and long (1–20 kb) introns in

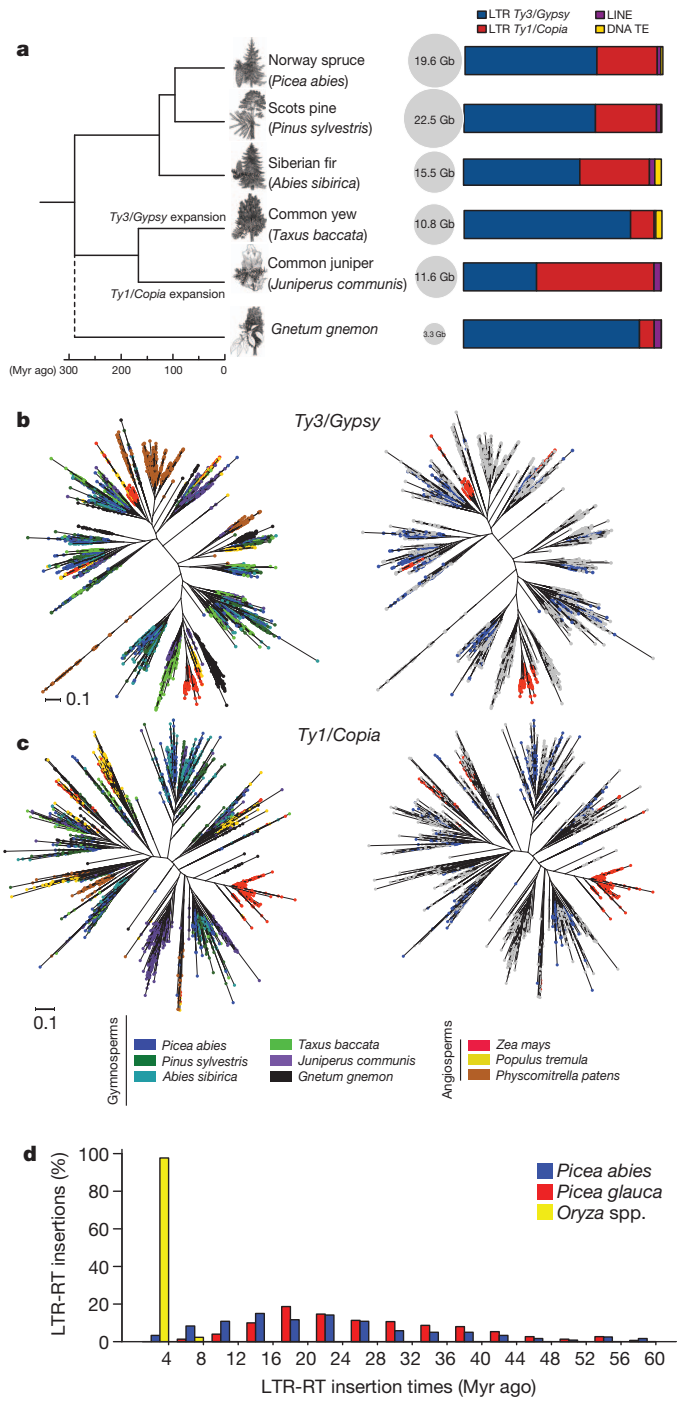

Figure 2 | Conifer genomes contain expansions of a diverse set of LTR-RTs. **a**, Distribution of different classes of transposable elements from six gymnosperm species. The figure is based on the total fraction of transposable elements (TE) identified and grouped into different classes from the different species. Genome sizes of the six species are given in circles and their phylogenetic relationship is shown, with tentative dating of divergence times (*x*-axis) based on 64 chloroplast genes over 39 species and five fossil calibration points. **b**, **c**, Heuristic neighbour-joining trees constructed from 5,922 sequences similar to the *Ty3/Gypsy* (**b**) and 3,052 sequences similar to the *Ty1/Copia* (**c**) reverse transcriptase domain from nine plant species. The trees to the right have only sequences from *P. abies* and *Z. mays* coloured, whereas the grey dots are the uncoloured versions of the other species represented on the left. **d**, Distributions of insertion times calculated for LTR-RTs in *Picea abies*, *Picea glauca* and *Oryza glaberrima*/*O. sativa*, using mutation rates (per base per year) of 2.2×10^{-9} for the *Picea* spp. and 1.8×10^{-8} for *O. glaberrima*[50].

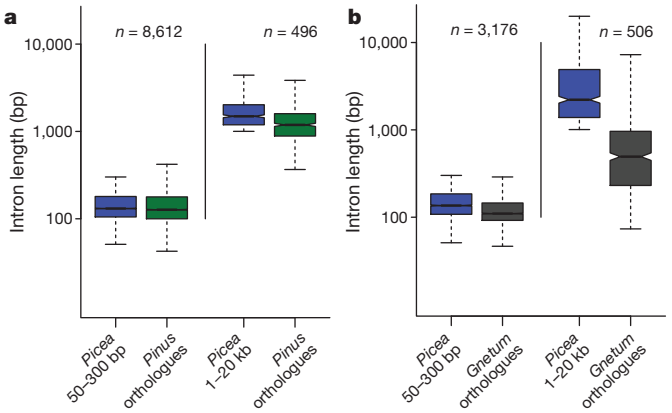

Figure 3 | Intron sizes are conserved among gymnosperms. **a**, **b**, Intron size comparisons between *P. abies*, *P. sylvestris* (**a**) and *G. gnemon* (**b**), respectively. Orthologues of introns that were categorised as short (50–300 bp) or long (1–20 kb) in *P. abies* were identified in *P. sylvestris* and *G. gnemon*, and the corresponding intron size was scored.

spruce within draft genome assemblies of *P. sylvestris* and *G. gnemon* (Supplementary Information 4.2). Introns identified as orthologous to a long intron in *P. abies* were also atypically long (Fig. 3a, b), suggesting that intron expansions started early in the history of conifers.

The evolution of important conifer traits

Two major differences between angiosperms and gymnosperms are their contrasting reproductive development and the development of water-conducting xylem cells. We therefore manually identified *P. abies* loci homologous to genes known to be centrally involved in these processes in angiosperms.

In angiosperms, homologues of the *A. thaliana* phosphatidylethanolamine-binding protein (PEBP) FLOWERING LOCUS T (FT) are key activators of flowering. It has been suggested that gymnosperms lack orthologues of *FT* genes, instead containing a group of *FT*/*TFL1*-like genes that probably act as flowering repressors[36,37]. We identified four putative *FT*/*TFL1*-like genes in the P.abies 1.0 genome that have not been previously described and confirmed that the genome does indeed lack *FT*-like genes (Supplementary Information 5.1).

MADS-box genes are involved in controlling most aspects of angiosperm development[38]. A total of 278 sequences with clear homology to MADS boxes were identified in the P.abies 1.0 assembly (Supplementary Information 5.2), 41 of which had transcript support. Type I and II MADS-box genes are distinguished in plants. Only 5% of the identified MADS boxes were of type I (Supplementary Fig. 5.2.), the lowest percentage of potential type I genes recorded in any plant genome. Type II MADS-box genes are subdivided into about a dozen ancient clades. We observed remarkable expansions in the *TM3*-like (or *SOC1*-like), *STMADS11*-like and *TM8*-like gene clades in *P. abies*. Because members of these gene clades are involved in vegetative development and phase changes such as the floral transition in angiosperms[39], we propose that the expansion of these gene clades has contributed to the evolution of developmental phase changes in gymnosperms.

The xylem tissue of most gymnosperms comprises a single water-transporting cell type, tracheids. By contrast, the xylem tissue of angiosperm species contains fibres, originating from tracheids that have to a large extent lost the capacity to conduct mass water flows, and vessels that have taken over the water-transport function in the stem. Formation of vessels is controlled by the *VASCULAR NAC DOMAIN* (*VND*) gene family, which has seven members in *A. thaliana*[40]. We detected two *VND* genes in *P. abies* (Supplementary Information 5.3), suggesting that co-option and expansion of the *VND* gene family in vessel formation might have been important for angiosperm evolution.

A model for conifer genome evolution

We propose the following model of conifer genome evolution. After the lineage that led to angiosperms had branched off and the most recent common ancestor of extant conifers had been established, the 12 ancestral chromosomes expanded at a slow and steady rate due to the activity of a diverse set of *Gypsy* and *Copia* LTR transposable elements that are largely shared among extant conifers. The expansion started early and, in contrast to angiosperms genomes in which this has been counteracted by efficient recombination mechanisms[33] resulting in only smaller transposable element subsets remaining following recent bursts of activity[41,42], these elements have remained in the genome. We propose that mechanisms for transposable element removal (for example by unequal recombination) have been less active in conifers than in most other organisms[43], and our data suggest that the insertion of transposable elements into genes gave rise to large introns, and (combined with other mechanisms) abundant pseudogenes. Each chromosome has grown to a similar size—perhaps limited by physical constraints on, for instance, chromosomal replication—with genes separated by large regions of transposable-element-rich, highly polymorphic non-protein-coding regions with low recombination frequencies. The gradual increase in size, the lack of WGDs and a predominately out-crossing mating system have probably also buffered conifer genomes against chromosomal rearrangements (WGD reduces sensitivity to aneuploidy), thereby maintaining synteny over large phylogenetic distances[44].

Some angiospems, such as cereals, also have large genomes but it seems as if the "one way ticket towards genome obesity"[45] that is barely recognizable in angiosperms prevails in conifers. The underlying mechanism remains unclear, but the low frequency of 24-nucleotide sRNAs, their role in methylation of repeats and their restriction to reproductive tissues may have influenced the process. However, considering the effect of methylation patterns on recombination rates[46] and the fact that 24-nucleotide sRNAs trigger methylation, such low recombination frequencies would more likely result from hypermethylation[47]. A state of 'genome paralysis' could potentially have been triggered once an obesity threshold was reached. In the angiosperm lineage, the occurrence of a number of WGDs probably increased diversification potential, allowing morphological innovation (for example, the origin of flowers and fruits) and facilitating speciation[46,48,49]. By contrast, the conserved genome structure resulting from the paucity of genome rearrangements and lack of WGDs in conifers probably limited the evolution of reproductive barriers (resulting in relatively low rates of speciation), and may explain the high degrees of conservation through time and low morphological diversity. Nevertheless, these processes do not seem to affect fitness as conifers dominate many ecosystems, probably because they contain high degrees of standing genetic variation, allowing them to occupy very wide ecological niches in climatic regions where other plant species are less competitive. The future availability of additional gymnosperm genome sequences will allow further exploration of the unique processes that have driven their evolution and facilitate improvement of this important species.

METHODS SUMMARY

We shotgun-sequenced 450 fosmid pools containing around 100–6,000 fosmids per pool (Supplementary Table 1.4). Each fosmid pool was assembled and scaffolded individually. Fosmid pool scaffolds larger than 1 kb (~6.7 Gb in total) were merged[12] (Supplementary Information 1.3) with a 38× haploid WGS assembly (~9.8 Gb in total, derived from ~600 ng of DNA extracted from a single megagametophyte). We subsequently performed scaffolding[13] using WGS libraries of five different insert sizes (0.3, 0.65, 2.4, 4.4 and 10.4 kb) from diploid tissue. We further increased assembly contiguity of protein-coding regions by scaffolding using a set of ~38 million unassembled (after digital normalization) RNA-Seq read-pairs generated from 22 samples (Supplementary Information 1.3).

Ab initio prediction of protein-coding genes was performed using ESTs from numerous conifer species, our own short-read *de novo* transcript assemblies and proteins from other plant species as supporting evidence (Supplementary Information 2.6). Predicted loci were used to perform gene family analysis and to examine the K_s substitution rates of paralogues to identify evidence for a recent WGD event. *De novo* transcript assemblies were used to identify lncRNA, and sRNA sequencing was performed and used for *de novo* miRNA prediction.

Repeated sequences were identified *de novo* using 454 reads longer than 700 bp generated from randomly sheared genomic DNA. Candidates were characterized using similarity searches at the nucleotide and amino acid level against public and custom collections of plant transposable element sequences. Complete LTR-RTs were identified using a combination of *de novo* searches and manual inspection.

WGS assemblies from shallow sequencing (3.8–12.5×) of *P. sylvestris*, *A. sibirica*, *J. communis*, *T. baccata* and *G. gnemon* were produced using the CLC Bio *de novo* assembler.

For website and accession number information, see Supplementary Information 6.

Received 10 February; accepted 22 April 2013.
Published online 22 May 2013.

1. Stewart, W. N. & Rothwell, G. W. *Paleobotany and the Evolution of Plants* (Cambridge Univ. Press, 1993).
2. Savard, L. *et al.* Chloroplast and nuclear gene sequences indicate late Pennsylvanian time for the last common ancestor of extant seed plants. *Proc. Natl Acad. Sci. USA* **91,** 5163–5167 (1994).
3. Leslie, A. B. *et al.* Hemisphere-scale differences in conifer evolutionary dynamics. *Proc. Natl Acad. Sci. USA* **109,** 16217–16221 (2012).
4. Flory, W. S. Chromosome numbers and phylogeny in the gymnosperms. *J. Arnold Arb.* **17,** 82–87 (1936).
5. Morse, A. M. *et al.* Evolution of genome size and complexity in Pinus. *PLoS ONE* **4,** e4332 (2009).
6. Kovach, A. *et al.* The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. *BMC Genomics* **11,** 420 (2010).
7. Ahuja, M. R. & Neale, D. B. Evolution of genome size in conifers. *Silvae Genet.* **54,** 126–137 (2005).
8. Buschiazzo, E., Ritland, C. E., Bohlmann, J. & Ritland, K. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evol. Biol.* **12,** 8 (2012).
9. Jaramillo-Correa, J. P., Verdu, M. & González-Martínez, S. C. The contribution of recombination to heterozygosity differs among plant evolutionary lineages and life-forms. *BMC Evol. Biol.* **10,** 22 (2010).
10. Murray, B. G. Nuclear DNA amounts in gymnosperms. *Ann. Bot. (Lond.)* **82,** 3–15 (1998).
11. Zhang, G. *et al.* The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490,** 49–54 (2012).
12. Vicedomini, R., Vezzi, F., Scalabrin, S., Arvestad, L. & Policriti, A. GAM-NGS: genomic assemblies merger for next generation sequencing. *BMC Bioinformatics* **14,** S6 (2013).
13. Sahlin, K., Street, N., Lundeberg, J. & Arvestad, L. Improved gap size estimation for scaffolding algorithms. *Bioinformatics* **28,** 2215–2222 (2012).
14. Vezzi, F., Narzisi, G. & Mishra, B. Feature-by-feature–evaluating de novo sequence assembly. *PLoS ONE* **7,** e31002 (2012).
15. 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).
16. Mackenzie, S. A. in *Plant Mitochondria* (ed. Logan, D. C.) 36–49 (Blackwell, 2007).
17. Messing, J. *et al.* Sequence composition and genome organization of maize. *Proc. Natl Acad. Sci. USA* **101,** 14349–14354 (2004).
18. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449,** 463–467 (2007).
19. Bennetzen, J. L., Coleman, C., Liu, R., Ma, J. & Ramakrishna, W. Consistent over-estimation of gene number in complex plant genomes. *Curr. Opin. Plant Biol.* **7,** 732–736 (2004).
20. Vanneste, K., Van de Peer, Y. & Maere, S. Inference of genome duplications from age distributions revisited. *Mol. Biol. Evol.* **30,** 177–190 (2013).
21. Jiao, Y. *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* **473,** 97–100 (2011).
22. García-Gil, M. R. Evolutionary aspects of functional and pseudogene members of the phytochrome gene family in Scots pine. *J. Mol. Evol.* **67,** 222–232 (2008).
23. Magbanua, Z. V. *et al.* Adventures in the enormous: a 1.8 million clone BAC library for the 21.7 Gb genome of loblolly pine. *PLoS ONE* **6,** e16214 (2011).
24. Bánfai, B. *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res.* **22,** 1646–1657 (2012).
25. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22,** 1775–1789 (2012).
26. Dolgosheina, E. V. *et al.* Conifers have a unique small RNA silencing signature. *RNA* **14,** 1508–1515 (2008).
27. Morin, R. D. *et al.* Comparative analysis of the small RNA transcriptomes of *Pinus contorta* and *Oryza sativa*. *Genome Res.* **18,** 571–584 (2008).
28. Wan, L.-C. *et al.* Identification and characterization of small non-coding RNAs from Chinese fir by high throughput sequencing. *BMC Plant Biol.* **12,** 146 (2012).
29. Zhang, J. *et al.* Dynamic expression of small RNA populations in larch (*Larix leptolepis*). *Planta* **237,** 89–101 (2013).

30. Henderson, I. R. & Jacobsen, S. E. Epigenetic inheritance in plants. *Nature* **447,** 418–424 (2007).
31. Sanmiguel, P. & Bennetzen, J. L. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann. Bot. (Lond.)* **82,** 37–44 (1998).
32. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nature Rev. Genet.* **8,** 973–982 (2007).
33. Devos, K. M., Brown, J. K. M. & Bennetzen, J. L. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12,** 1075–1079 (2002).
34. Vitte, C. & Panaud, O. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L. *Mol. Biol. Evol.* **20,** 528–540 (2003).
35. Vicient, C. M. *et al.* Retrotransposon *BARE-1* and its role in genome evolution in the genus *Hordeum*. *Plant Cell* **11,** 1769–1784 (1999).
36. Karlgren, A. *et al.* Evolution of the PEBP gene family in plants: functional diversification in seed plant evolution. *Plant Physiol.* **156,** 1967–1977 (2011).
37. Klintenäs, M., Pin, P. A., Benlloch, R., Ingvarsson, P. K. & Nilsson, O. Analysis of conifer *FLOWERING LOCUS T/TERMINAL FLOWER1*-like genes provides evidence for dramatic biochemical evolution in the angiosperm *FT* lineage. *New Phytol.* **196,** 1260–1273 (2012).
38. Gramzow, L. & Theissen, G. A hitchhiker's guide to the MADS world of plants. *Genome Biol.* **11,** 214 (2010).
39. Smaczniak, C. *et al.* Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development* **139,** 3081–3098 (2012).
40. Kubo, M. *et al.* Transcription switches for protoxylem and metaxylem vessel formation. *Genes Dev.* **19,** 1855–1860 (2005).
41. Piegu, B. *et al.* Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **16,** 1262–1269 (2006).
42. Hawkins, J. S., Kim, H., Nason, J. D., Wing, R. A. & Wendel, J. F. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res.* **16,** 1252–1261 (2006).
43. Bennetzen, J. L., Ma, J. & Devos, K. M. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot. (Lond.)* **95,** 127–132 (2005).
44. Pavy, N. *et al.* A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. *BMC Biol.* **10,** 84 (2012).
45. Bennetzen, J. L. & Kellogg, E. A. Do plants have a one way ticket to genomic obesity? *Plant Cell* **9,** 1509–1514 (1997).
46. Van de Peer, Y., Fawcett, J. A., Proost, S., Sterck, L. & Vandepoele, K. The flowering world: a tale of duplications. *Trends Plant Sci.* **14,** 680–688 (2009).
47. Fedoroff, N. V. Presidential address. Transposable elements, epigenetics, and genome evolution. *Science* **338,** 758–767 (2012).
48. Van de Peer, Y. A mystery unveiled. *Genome Biol.* **12,** 113 (2011).
49. Soltis, D. E. *et al.* Polyploidy and angiosperm diversification. *Am. J. Bot.* **96,** 336–348 (2009).
50. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. USA* **101,** 12404–12410 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors would like to acknowledge support from the Knut and Alice Wallenberg Foundation. Additional funding was provided in particular by the Swedish Research Council (VR), the Swedish Governmental Agency for Innovation Systems (Vinnova), the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (Formas), the Swedish foundation for Strategic Research (SSF), the Government of Canada through Genome Canada, by Genome British Columbia, and by Genome Quebec, Science for Life Laboratory and the National Genomics Infrastructure (NGI), Sweden. We also acknowledge Skogforsk for the *P. abies* genetic material, UPPMAX for computational infrastructure, CLC Bio for assembly software development and Lucigen for fosmid-pool method development.

Author Contributions B.N. and N.R.S. are joint first authors, and A.W., A.Z., Y-C.L. and D.G.S. are joint second authors, who contributed to most parts of the work. F.V., A.A., N.D., R.V., K.S. and E.S. contributed to the assembly and sequence analysis; S.G. to repeat analysis; N.D., M.E., L.G., M.Ku., T.N., Å.O., G.T., H.T., P.Z. and B.Z. to quality control of the assembly and analysis of gene families; K.H., J.H., O.K., M.Kä. and T.S. to the sequencing; L.K. to analysis of the mitochindrial genome; M.Ko. and N.R. to generation of fosmid pools; J.Lut., F.L., C.T.-L. and K.V. to analysis of the sequences of *P. abies* and other conifers; C.R. and J.S. to production of BAC sequences; Z.-Q.W. to analysis of the chloroplast genome and J.A.R. determined the genome size. L.A., R.B., J.Boh., J.Bou., R.G.G., T.R.H., P.d.J., J.M., M.M., K.R., B.S., S.L.T., Y.V.d.P. and B.A. contributed to the design and supervision of various parts of the research. O.N. headed and P.K.I. managed the project, J.Lun. coordinated the sequencing and assembly activities, and S.J. the bioinformatics activities. B.N., N.R.S., A.Z., O.N., P.K.I., J.Lun. and S.J. wrote and edited most of the manuscript. All authors commented on the manuscript.

Author Information Raw data and assemblies are available from the ConGenIE (Conifer Genome Integrative Explorer) web resource (http://congenie.org), as well as the European Bioinformatics Institute (EMBL) and European Nucleotide Archive (ENA); see Supplementary Information 6.1 for accession numbers. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to O.N. (ove.nilsson@slu.se), P.K.I. (par.ingvarsson@emg.umu.se), J.Lun. (joakim.lundeberg@scilifelab.se) or S.J. (stefan.jansson@umu.se).

Björn Nystedt[1], Nathaniel R. Street[2], Anna Wetterbom[3], Andrea Zuccolo[4,5], Yao-Cheng Lin[6], Douglas G. Scofield[2,7], Francesco Vezzi[8], Nicolas Delhomme[2], Stefania Giacomello[4,9], Andrey Alexeyenko[10], Riccardo Vicedomini[4,9], Kristoffer Sahlin[8], Ellen Sherwood[1], Malin Elfstrand[11], Lydia Gramzow[12], Kristina Holmberg[10], Jimmie Hällman[10], Olivier Keech[2], Lisa Klasson[13], Maxim Koriabine[14], Melis Kucukoglu[15], Max Käller[10], Johannes Luthman[3], Fredrik Lysholm[3], Totte Niittylä[15], Åke Olson[11], Nemanja Rilakovic[10], Carol Ritland[16], Josep A. Rosselló[17,18], Juliana Sena[19], Thomas Svensson[20], Carlos Talavera-López[3], Günter Theißen[12], Hannele Tuominen[2], Kevin Vanneste[6], Zhi-Qiang Wu[7], Bo Zhang[2], Philipp Zerbe[21], Lars Arvestad[8,22], Rishikesh Bhalerao[15], Joerg Bohlmann[16,21], Jean Bousquet[19], Rosario Garcia Gil[15], Torgeir R. Hvidsten[2,23], Pieter de Jong[14], John MacKay[19], Michele Morgante[4,9], Kermit Ritland[16], Björn Sundberg[15], Stacey Lee Thompson[7], Yves Van de Peer[6], Björn Andersson[3], Ove Nilsson[15], Pär K. Ingvarsson[7], Joakim Lundeberg[10] & Stefan Jansson[2]

[1]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Box 1031, 171 21 Solna, Sweden. [2]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, 901 87 Umeå, Sweden. [3]Department of Cell and Molecular Biology, Science for Life Laboratory, Karolinska Institutet, Box 1031, 171 77 Stockholm, Sweden. [4]Istituto di Genomica Applicata, Via J. Linussio 51, 33100 Udine, Italy. [5]Institute of Life Sciences, Scuola Superiore Sant'Anna, 56127 Pisa, Italy. [6]Department of Plant Systems Biology (VIB) and Department of Plant Biotechnology and Bioinformatics (Gent University), Technologiepark 927, 9052 Gent, Belgium. [7]Umeå Plant Science Centre, Department of Ecology and Environmental Science, Umeå University, 901 87 Umeå, Sweden. [8]School of Computer Science and Communication, Science for Life Laboratory, KTH Royal Institute of Technology, Box 1031, 171 21 Solna, Sweden. [9]Università degli Studi di Udine, Via delle Scienze 208, 33100 Udine, Italy. [10]School of Biotechnology, Science for Life Laboratory, KTH Royal Institute of Technology, Box 1031, 171 21 Solna, Sweden. [11]Department of Forest Mycology and Plant Pathology, Uppsala Biocenter, Swedish University of Agricultural Sciences, Box 7026, 750 07 Uppsala, Sweden. [12]Department of Genetics, Friedrich-Schiller-University Jena, Philosophenweg 12, 07743 Jena, Germany. [13]Molecular Evolution, Department of Cell and Molecular Biology, Uppsala University, Husargatan 3, 752 37 Uppsala, Sweden. [14]BACPAC Resources, Children's Hospital of Oakland Research Institute, Bruce Lyon Memorial Research Building, Oakland, California 94609, USA. [15]Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, 901 83 Umeå, Sweden. [16]Department of Forest and Conservation Sciences, University of British Columbia, 2424 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada. [17]Jardí Botànic, Universitat de Valencia, c/Quart 80, 46008 Valencia, Spain. [18]Marimurtra Botanical Garden, Carl Faust Fdn, 17300 Blanes, Spain. [19]Canada Research Chair in Forest and Environmental Genomics, Centre for Forest Research and Institute for Systems and Integrative Biology, Université Laval, Québec, Québec G1V 0A6, Canada. [20]Department of Biosciences and Nutrition, Science for Life Laboratory, Karolinska Institutet, Box 1031, 171 21 Solna, Sweden. [21]Michael Smith Laboratories, University of British Columbia, 321-2185 East Mall, Vancouver, British Columbia V6T 1Z4, Canada. [22]Swedish e-Science Research Center, Department Numerical Analysis and Computer Science, Stockholm University, Box 1031, 171 21 Solna, Sweden. [23]Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1432 Ås, Norway.