Molecular Ecology Resources

A reference genome sequence for the European silver fir (*Abies alba* Mill.): a community-generated genomic resource

RUNNING TITLE: SILVER FIR GENOME ABAL 1.1

Elena Mosca [1], Fernando Cruz [2], Jèssica Gómez Garrido [2], Luca Bianco [3], Christian Rellstab [4], Eric Bazin [5,] Sabine Brodbeck [4], Katalin Csilléry [4,6], Bruno Fady [7], Matthias Fladung [8], Barbara Fussi [9], Dušan Gömöry [10], Santiago C. González-Martínez [11], Delphine Grivet [12], Marta Gut [2, 13], Ole Kim Hansen [14], Katrin Heer [15], Zeki Kaya [16], Konstantin V. Krutovsky [17, 18,19], Birgit Kersten [8], Sascha Liepelt [15], Lars Opgenoorth [15], Christoph Sperisen [5], Kristian K. Ullrich [20], Giovanni G. Vendramin [21], Marjana Westergren [22], Birgit Ziegenhagen [14], Tyler Alioto [2, 13], Felix Gugerli [4], Berthold Heinze [23], Maria Höhn [24], Michela Troggio [3], David B. Neale [25*]

[1] C3A - Centro Agricoltura Alimenti Ambiente, University of Trento, via E. Mach 1, 38010 S. Michele a/Adige (TN); Italy (elena.mosca@unitn.it); [2] CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, BaldiriReixac 4, Barcelona 08028; Spain (fernando.cruz@cnag.crg.eu; jessica.gomez@cnag.crg.eu; marta.gut@cnag.crg.eu; tyler.alioto@cnag.crg.eu); [3] Fondazione Edmund Mach, Via Mach 1, 38010 S. Michele a/Adige (TN); Italy (michela.troggio@fmach.it; luca.bianco@fmach.it); [4] Swiss Federal Research Institute WSL, Zürcherstrasse 111, 8903 Birmensdorf; Switzerland (felix.gugerli@wsl.ch; christian.rellstab@wsl.ch; christoph.sperisen@wsl.ch; sabine.brodbeck@wsl.ch; katalin.csillery@wsl.ch); [5] Laboratoire d'Ecologie Alpine, Université Grenoble Alpes (LECA),Université Grenoble Alpes CS 40700; 38058 Grenoble cedex 9; France; [6] University of Zürich, Department of Evolutionary Biology and

Environmental Studies, Winterthurerstrasse 190, CH-8057 Zurich; [7] Institut National de la Recherche Agronomique (INRA), Unité de Recherche Ecologie des Forêts Méditerranéennes (URFM), Site Agroparc, Domaine Saint Paul, 84914 Avignon; France (Bruno.fady@inra.fr); [8] Thünen-Institute of Forest Genetics, Sieker Landstr. 2, 22927 Grosshansdorf; Germany (matthias.fladung@thuenen.de; birgit.kersten@thuenen.de); [9] Bavarian Office for Forest Seeding and Planting (ASP), Applied Forest Genetics, Forstamtsplatz 1, 83317 Teisendorf; Germany (barbara.fussi@asp.bayern.de); [10] Technical University in Zvolen, TG Masaryka 24, 96053 Zvolen; Slovakia (gomory@tuzvo.sk); [11] Institut National de la Recherche Agronomique (INRA), UMR1202 Biodiversity, Genes & Communities (BIOGECO), University of Bordeaux, 69, route d'Arcachon, 33610 Cestas; France (santiago.gonzalez-martinez@pierroton.inra.fr); [12] INIA Forest Research Centre, Carretera de la Coruña km 7.5, 28040 Madrid; Spain (dgrivet@inia.es); [13] Universitat Pompeu Fabra (UPF), Plaça de la Mercè, 10, 08002 Barcelona, Spain; [14] Department of Geosciences and Natural Resource Management (IGN), University of Copenhagen, Rolighedsvej 23, 1958 Frederiksberg C; Denmark (okh@ign.ku.dk); [15] Philipps-Universität Marburg, Faculty of Biology (PUM), Karl-von-Frisch-Str. 8, 35032 Marburg; Germany (katrin.heer@biologie.uni-marburg.de; liepelt@biologie.uni-marburg.de; lars.opgenoorth@uni-marburg.de; birgit.ziegenhagen@biologie.uni-marburg.de); [16] Department of Biological Sciences (METU), Middle East Technical University, 06800 Çankaya/Ankara; Turkey (kayaz@metu.edu.tr); [17] Department of Forest Genetics and Forest Tree Breeding, Georg-August University of Göttingen, Büsgenweg 2, 37077 Göttingen; Germany (konstantin.krutovsky@forst.uni-goettingen.de); [18] Laboratory of Population Genetics, Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina Str. 3, Moscow 119991, Russia; [19] Laboratory of Forest Genomics, Genome Research and Education Center, Institute of Fundamental Biology and Biotechnology, Siberian Federal University, 50a/2 Akademgorodok, Krasnoyarsk 660036, Russia; [20] Max Planck Institute for Evolutionary Biology, Department for Evolutionary Genetics (MPI), August Thienemann Str. 2, 24306 Ploen; Germany (ullrich@evolbio.mpg.de); [21] Institute of Biosciences and BioResources, National Research Council, Via Madonna del Piano 10,50019 Sesto Fiorentino (Firenze); Italy (giovanni.vendramin@ibbr.cnr.it); [22] Slovenian Forestry Institute (SFI),

GozdarskiinštitutSlovenije),Večna pot 2, 1000 Ljubljana; Slovenia (marjana.westergren@gozdis.si); [23] Federal Research and Training Centre for Forests, Natural Hazards and Landscape (BFW), Seckendorff-Gudent Weg 8, 1130 Wien; Austria (berthold.heinze@bfw.gv.at); [24] Faculty of Horticultural Science, Department of Botany (SZIU/FHS), Szent Istvan University, 1118 Budapest; Hungary (Hohn.Maria@kertk.szie.hu); [25] Department of Plant Sciences, University of California at Davis (UCD), Davis 95616; USA (dbneale@ucdavis.edu)

* Corresponding Author

Abstract (246 words)

Silver fir (*Abies alba* Mill.) is widespread in Central, Eastern and Southern Europe. In Southern Europe, its distribution has increased overall during the 20th century due to land-use change and recolonization from refugial, over-logged populations. During recent decades, its distribution has decreased in most of its distributional range, mainly due to extreme temperature events, forest management practices and ungulate browsing. To forecast its future distribution and survival, it is important to investigate the genetic basis of its adaptation to environmental change, notably extreme events. Here, we provide a first draft genome assembly and annotation of the silver fir genome. DNA obtained from haploid megagametophyte and diploid needle tissue was used to construct and sequence Illumina paired-end (PE) and mate-pair (MP) libraries, respectively, to high depth. The assembled *A. alba* genome sequence accounted for over 37 million scaffolds corresponding to 18.16 Gb, with a scaffold N50 of 14,051 bp. Despite the fragmented nature of the assembly, a total of 50,757 full-length genes were functionally annotated in the nuclear genome. The chloroplast genome was also assembled into a single scaffold (120,908 bp) that shows a high collinearity with both the *A. koreana* and *A. sibirica* complete chloroplast genomes. This first genome assembly of silver fir is an important genomic resource that is now publicly available in support of a new generation of research. By genome-enabling this important conifer, this resource will be opening the gate for new experiments and more precise genetic monitoring of European silver fir forests.

Keywords: *Abies alba*, annotation, conifer genome, genome assembly, genomic resource

Word counts excluding references 7,060

1. INTRODUCTION

Conifers represent the dominant trees in some temperate and all boreal ecosystems and have important economic value, especially in timber production. They are also facing the effect of the current climate change, with an increase in temperature and lower precipitation particularly in Southern Europe, and increased frequency of extreme events, to which some species may be unable to adapt at sufficient pace. Silver fir (*Abies alba* Mill.) is a keystone conifer of European montane forest ecosystems, which is dominant in cool areas of the temperate zone (Ellenberg, 2009). It can live up to 500–600 years, mark late stages of forest succession and reach up to 60 m in height (Wolf, 2003). It grows on different soil types, but requires high soil moisture during the growing season, preferring places with a mean annual precipitation ranging from 700 to 1800 mm (Tinner et al., 2013). Its distribution ranges from the Pyrenees (up to 2100 m a.s.l.), to the Alps (300-1800 m a.s.l.) and the Carpathians where it reaches its easternmost range edge (100-1500 m a.s.l.; Fig. S1 Supplemental Information). Growing interest in silver fir has emerged because of its potential vulnerability to climate change, which could change conditions for sustainable use and economic value of the species. In turn, this species is more drought-resistant than other economically important species for timber production, such as Norway spruce (Vitali, Büntgen, & Bauhus, 2017), at least in parts of its range, which could turn out to be beneficial under the expected increase in extended future drought periods. During the mid-1970s, several stands in Central Europe showed crown dieback and declining tree growth that were mainly due to air pollution (Kandler & Innes, 1995) that also increased the species' drought susceptibility (Elling, 2009). Currently several stands in southern parts of the silver fir distribution have shown symptoms of crown die back (Cailleret, Nourtier, Amm, Durand-

Gillmann, & Davi, 2014), which were due to drought and heat waves. The species' sensitivity to extreme events was confirmed in mixed temperate forests in southern Europe (Lebourgeois, Rathgeber, & Ulrich 2010). As a consequence of climate change, a shift toward higher elevation and northern latitude is expected as well as die back at lower elevations (Cailleret & Davi, 2011, Cailleret et al., 2014; Tinner et al., 2013; Büntgen et al., 2014). While the species is not endangered, its distribution has decreased over the last century. In the Mediterranean area, the distribution of silver fir is highly fragmented, resulting in small stands, which are the forests of priority for conservation according to the European Habitat Directive (92/43/CEE Habitat). Several studies investigated the environmental effect on silver fir genetic diversity across the Italian Alps, showing the association between silver fir genetic diversity and seasonal minimum temperature (Mosca et al., 2012) as well as between genetic diversity and both temperature and soil type (Mosca, Gonzáles-Martínez, & Neale, 2014). Recent studies confirmed the environmental effect local adaptation of silver fir, which was shaped by winter drought in marginal silver fir populations (Roschanski et al., 2016). Local adaptation was also investigated combining genetic data and common gardens, showing selection on height driven by thermal stability and on growth phenology driven by precipitation seasonality (Csilléry, Sperisen, Ovaskainen, Widmer, & Gugerli, 2018). Another study investigated the association between genetic diversity and dendro-phenotypic information (Heer et al., 2018), while Piotti et al. (2017) confirmed the importance of the Apennines as a refugium of genetic diversity of the species. However, all these studies were based on a modest number of genetic markers (several hundreds of single-nucleotide polymorphisms, SNPs, or tens of simple sequence repeats, SSRs) due to the lack of genomic resources.

Conifer genomes are often very large (mean 17.4 ± 7.5 G bp), ranging from 4 to 35 giga base pairs (Gb) as taken from KEW Database in August 2018 (Bennett & Leitch, 2012;

Grotkoppet, Rejmánek, Sanderson, & Rost, 2004; Zonneveld, 2012), but their gene content is similar to that of other vascular plants (Leitch, Soltis, Soltis, & Bennett, 2005). Conifer genomic resources have grown in recent years due to the application of Next Generation Sequencing technologies. To date, only a few conifer genomes have been fully sequenced, including: *Picea abies* (L.) Karst (Nystedt et al., 2013), *Picea glauca* (Moench) Voss (Birol et al., 2013), *Pinus taeda* L. (Neale et al., 2014), *Pinus lambertiana* Dougl. (Stevens et al., 2016), *Pseudotsuga menziesii* (Mirb.) Franco (Neale et al., 2017), and *Larix sibirica* Ledeb. (Kuzmin et al., 2018). Until now, *Abies* species have lacked a whole reference genome. This is understandable, as the sequencing of conifer genomes is still a challenge due to their large size, the presence of interspersed repetitive sequences, the high frequency of genome duplication events and Long Terminal Repeats (LTR) retrotransposon bursts (Stevens et al., 2016).

In contrast to most of these sequenced conifers, silver fir, as a late successional species, has a peculiar life-history strategy. Saplings of silver fir are able to survive long periods of shading in the understory, and then to grow quickly when light conditions are favorable. Once available, the whole-genome sequence of silver fir offers the opportunity to study genes underlying traits like shade tolerance and regeneration capacity that are characteristic of silver fir. The elucidation of the genomic basis of these traits in silver fir has the potential to make a large impact on conifer ecological research. The silver fir genome sequence can also be used to assist genomic selection (Grattapaglia et al., 2018), as well as forest management and conservation strategies through well-selected source stands for assisted migration. Furthermore, the development of this genetic resource could help to characterize and certify the origin of forest reproductive material (FRM) used in reforestation, and to effectively conserve genetic resources in natural forests. Selecting FRM from the northern edge of the distribution range

depends on late-frost tolerant material, while at the southern edge, drought tolerance becomes important.

The aim of this project was to sequence and assemble the silver fir genome and to compare this resource with other conifer genomes (Nystedt et al., 2013; Birol et al., 2013; Neale et al., 2014; Stevens et al., 2016; Neale et al., 2017; Kuzmin et al., 2018). This study also provides more information on the *Abies* chloroplast genome in relation to closely related taxa. A long-term perspective related to other *Abies* taxa is to identify gene regions involved in drought resistance and late flushing, which are traits found in Mediterranean firs that hybridize with *A. alba* in both natural forests at range margins and in plantations (George et al., 2015).

2. MATERIALS AND METHODS

2.1 Reference tree for genome sequencing

Tissue samples for sequencing were obtained from an adult silver fir tree (AA_WSL01) located in a public forest next to the institute of WSL Birmensdorf, Switzerland (47.3624°N, 8.4536°E; Supplemental Information). Seeds were collected directly from the selected tree in November 2016, dried at ambient temperature and stored at -5°C. Fresh needles were harvested shortly after flushing in May 2017. A multilocus SNP analysis across the species range in Switzerland placed the sampled tree mainly within the genetic cluster of the Swiss plateau (Fig. S2 Supplemental Information), with ancestry proportions similar to populations of the Jura Mountains and Central Alps. This was confirmed using nuclear microsatellites (C. Rellstab, personal communication).

2.2 DNA preparation

2.2.1 Haploid megagametophyte DNA isolation for paired-end (PE) sequencing

Seeds of the reference tree were incubated in tap water for 24 h at room temperature. Seeds were dissected in a sterile 0.9% sodium-chloride solution under a stereo lens in an environment cleaned with bleach, using micro scissors and forceps. The embryo and all seed skins were carefully removed. The retained megagametophyte tissue was rinsed with fresh sterile 0.9% sodium-chloride solution, immediately transferred to a 2 mL Eppendorf tube and stored at -80°C. Megagametophyte tissue was lyophilized for 16 h prior to extraction and homogenized for 30 s using a mixer mill (Retsch MM 300, Haan, Germany). DNA extraction was performed with a customized sbeadex kit (LGC Genomics, Berlin, Germany), which included all used chemicals and reagents as mentioned below. 500 µL LP-PVP, 5 µL Protease, 1 µL RNAse and 20 µL debris capture beads were added as lysis buffer to the ground tissue and the mix was incubated at 50°C and 350 rounds per minute (rpm) in a heating block for 30 min. After brief centrifugation, 400 µL cleared lysate was added to 400 µL binding buffer SB and 10 µL sbeadex beads. After 15 min binding at room temperature with shaking at 850 rpm, magnetic beads were collected on a magnetic stand for 2 min, and the supernatant was discarded completely. Beads were successively washed with the following buffers: 400 µL BN1, 400 µL TN1, 400 µL TN2, and 400 µL PN2. Washing time was 7 min for all four steps, with shaking at 850 rpm, followed by a short spin, 2 min of bead collection on a magnetic stand, and careful discarding of wash buffer. DNA was finally eluted in 100 µL elution buffer AMP at 60°C and 850 rpm on a heating block for 10 min. After a short spin and 3 min of magnetic bead collection on a magnetic stand, DNA was transferred into a new tube, centrifuged at 21,000 x g for 2 min, and transferred without pellet into a new tube.

DNA concentration was measured using the QuantiFluor dsDNA System (Promega, Madison, WI, USA). 260/280 and 260/230 ratios were measured using a Nanodrop 1000 (Thermo Fisher Scientific, Waltham, MA, USA; Table S1 Supplemental Information), and DNA integrity was visualized by running 5 µL of DNA on a 1% agarose gel. Nuclear and chloroplast microsatellites were used to exclude the contamination of the haploid maternal DNA with diploid DNA deriving from the surrounding tissue and to confirm the presence of only one maternal haplotype (C. Rellstab, personal communication). Because different megagametophytes from the same tree represent different haplotypes, only one DNA sample with high DNA quality and quantity was chosen for PE sequencing. DNA from a single megagametophyte (3.6 µg at 40 ng/µL; Table S1) was transferred to CNAG-CRG for PE library preparation and sequencing.

2.2.2 Diploid needle DNA isolation for mate-pair (MP) sequencing

Young, bright green needles of the reference tree were collected, frozen at -80 °C and lyophilized for 24 h. For DNA extraction, 25 mg of tissue were ground in a 2 mL Eppendorf tube with two steel balls (d = 3.1 mm) for 1.5 min, using a mixer mill MM300 (Retsch). DNA was extracted with the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany), starting with 600 µL AP1, 1 µL RNAse and 1 µL DX reagent. Then, DNA extraction was carried out according to the manufacturer's protocol, with an additional washing step with washing buffer AW2. DNA was eluted in 2x 100 µL nuclease-free water. DNA concentration was measured using QuantiFluor dsDNA System (Promega), 260/280 and 260/230 ratios were measured using a Nanodrop 1000 (ThermoFisher), and DNA integrity was visualized by running 0.6 µL of DNA on a 1 % agarose gel. DNA samples were verified using nuclear and chloroplast microsatellite markers as mentioned above, in order to exclude contamination (C. Rellstab, personal

communication), and one sample (24.5µg at 136 ng/µL; Table S1) was used to prepare for MP sequencing.

2.3 Sequencing

2.3.1 Whole-genome sequencing (WGS) library preparation and sequencing

Haploid DNA material from the single megagametophyte was used to construct three 300 bp-insert paired-end libraries at the CNAG-CRG Sequencing Unit. The short-insert PE libraries for the whole-genome sequencing were prepared with KAPA HyperPrep kit (Roche-Kapa Biosystems) with some modifications. In short, 1.0 µg of genomic DNA was sheared on a Covaris™ LE220 (Covaris Woburn, Massachusetts, USA) in order to reach fragment sizes of ~500 bp. The fragmented DNA was further size-selected for fragment sizes of 220-550 bp with AMPure XP beads (Agencourt, Beckman Coulter). The size-selected genomic DNA fragments were end-repaired, adenylated and ligated to Illumina sequencing compatible indexed paired-end adaptors (NEXTflex® DNA Barcodes). The adaptor-modified end library was size selected and purified with AMPure XP beads to eliminate any not ligated adapters. The ligation product was split into three samples and in three separate reactions enriched with 12 PCR cycles and then validated on an Agilent 2100 Bioanalyzer with the DNA 7500 assay (Agilent) for size and quantity. The resulting libraries had estimated fragment sizes of 304 bp, 307 bp and 311 bp. These are referred to as PE300-1, PE300-2, and PE300-3 in Table 1.

All three libraries were sequenced in equal proportions on HiSeq 4000 (Illumina, Inc, San Diego, California, USA) in paired-end mode with a read length of 2 × 151 bp using a HiSeq 4000 PE Cluster kit sequencing flow cell, following the manufacturer's protocol. Image analysis, base calling and quality scoring of the run were processed using the manufacturer's

software Real Time Analysis (RTA 2.7.6) and followed by generation of FASTQ sequence files by CASAVA.

2.3.2 Mate-pair library preparation and sequencing

DNA extracted from the diploid needle material was used to build three mate-pair (MP) libraries of increasing insert size: 1,500 bp (MP1500), 3,000 bp (MP3000) and 8,000 bp (MP8000). Libraries were prepared using the Nextera Mate Pair Library Prep Kit (Illumina) using the gel-plus protocol selecting for three different distribution sizes according to the manufacturer's instructions. After fragmentation, bands of 1.5, 3 and 8 Kb were selected for circularization. The following amounts of size-selected DNA were used for the circularization reaction: 270 ng (1.5 kb), 285 ng (3 kb), and 97.4 ng (8 kb).

All three MP libraries were sequenced on HiSeq2000 (Illumina, Inc) in paired-end mode with a read length of 2 × 101 bp using TruSeq SBS Kit v4. Image analysis, base calling and quality scoring of the run were processed using the manufacturer's software Real Time Analysis (RTA 1.18.66.3) and followed by generation of FASTQ sequence files by CASAVA.

2.4 Assembly

2.4.1 Genome assembly

Given the nearly equivalent estimated fragments sizes, the reads from the three paired-end libraries (PE300-1, PE300-2, and PE300-3) were joined into one library for assembly and collectively referred to as PE300. Before assembling the genome, its size and its complexity were evaluated using *k*-mer analyses. Jellyfish v2.2.0 (Marçais & Kingsford, 2011) was run on the sequence reads of this PE library to obtain the distribution of 17 *k*-mers. SGA preqc

(Simpson & Durbin, 2011; Simpson, 2014) was then used to estimate the mean fragment size and standard deviation of the PE300 library.

First, an initial assembly of the PE300 reads was performed with MaSuRCA v3.2.2 (Zimin, Marçais, Puiu, Roberts, Salzberg, & Yorke, 2013). MaSuRCA was run using default parameters, choosing SOAPdenovo for faster contig and light scaffold assembly. A *k*-mer of 105 was chosen by MaSuRCA for *de Bruijn* graph construction. The initial assembly was run for 33 days on a single 48-core node (4 Intel(R) Xeon(R) CPU E7-4830 v3 at 2.10GHz and 2TB of RAM) and with a maximum memory usage of 1.22 TB.

Second, the PE300 and the three MP libraries (MP1500, MP3000 and MP8000) were used to scaffold the initial assembly with BESSTv2.5.5 (Sahlin, Vezzi, Nystedt, Lundeberg, & Arvestad, 2014). It was run with options *--separate_repeats, -K*=105 *-max_contig_overlap*=115 and *–k*=466. Briefly, –K specifies the k-mer size used in the *de Bruijn* graph for the input assembly to be scaffolded. As 90 % of the input "contigs" were longer than 115 bp, this length was selected, instead of the default value of 200 bp, as the maximum identical overlap to search (*k*). Given the fragmented input assembly, the idea was to avoid using contigs smaller than the original genomic fragment. Therefore, the contig size threshold for scaffolding was set to 466 bp, 10 bp greater than the mean (294) plus two times the standard deviation (81) of the PE300 fragment size as estimated by mapping. The scaffolded genome assembly is referred to as ABAL 1.0. Moreover, an analysis of the spectra copy number (KAT; Mapleson, Garcia Accinelli, Kettleborough, Wright, & Clavijo, 2016) of the assemblies was done before and after scaffolding using the PE300 library.

2.4.2 Chloroplast genome assembly and annotation

All of the 100 bp reads from the MP1500 library (the library with the tightest size distribution and highest complexity) were mapped to the closest complete reference chloroplast sequence available in NCBI, i.e. from *Abies koreana* (NC_026892.1, Yi et al., 2015), using BWAmem (Li & Durbin, 2010) in paired mode and option –M to discard short split mappings. The mapped reads were then extracted from the alignment using BAM2FASTQ v1.1.0 (Alpha GSLaH). Both the linker sequence and the Nextera adapters present in the MP sequences were removed with Cutadapt (Martin, 2011). Finally, they were reversed-complemented in order to obtain an artificial PE library with insert size of 1,387 ±327 bp.

The FAST-PLAST pipeline was run producing SPAdes (Bankevich et al., 2012) assemblies using a range of *k*-mers (55, 69, 87). Afterwards, Ragout (Kolmogorov, Raney, Paten, & Pham, 2014) was used to obtain a reference-assisted assembly. In this case, *A. sibirica* (NC_035067.1) was used as chloroplast reference to place and orient all the *A. alba* contigs. Finally, Gapfiller (Boetzer & Pirovano, 2012) was used to close gaps in the chloroplast genome. DNA diff module - from MUMMER 3.22 package (Kurtz et al., 2004) - was run to compare the intermediate SPases assembly with the *A. koreana* (NC_026892.1) and *A. sibirica* (NC_035067.1) complete chloroplast sequences. Finally, the annotation of the chloroplast was carried out with DOGMA (Wyman, Jansen, & Boore 2004).

2.4.3 Genome quality assessment

The final nuclear assembly was evaluated for gene completeness using CEGMA v2.5 (Parra et al., 2007), which searches for 248 ultra-conserved core eukaryotic genes (CEGs), and BUSCO v3.0.2 (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov 2015), using 956 single-copy orthologues from plants (BUSCO v1 plantae database).

To obtain a more comprehensive estimate of genes present in the genome assembly, the STAR software package (Domin & Gingeras, 2015) was used to map the genome assembly with the silver fir RNA-seq produced by Roschanski et al. (2013) (GenBank accession numbers JV134525– JV157085) as well as 12 transcriptomes originating from Mont Ventoux (France) and the Black Forest (District Oberharmersbach, Germany), as reported in Roschanski et al. (2013) and available in the Dryad Digital Repository (Roschanski et al., 2015; 2016). In addition, the transcripts from *P. taeda* were aligned to the genome using GMAP with default options (Wu, Reeder, Lawrence, Becker, & Brauer 2016).

2.5 Annotation

2.5.1 Protein-coding gene annotation

Repeats were identified, annotated and masked in the silver fir genome assembly following three sequential steps. First, RepeatMasker (http://www.repeatmasker.org) v4.0.6 was run using the Pinaceae-specific repeat library included in the RepeatMasker release. Then, repeats annotated in *P. taeda* and *P. menziesii* were used in a second run of RepeatMasker. Finally, *Abies alba*-specific repeats were detected with RepeatModeler and masked with RepeatMasker. An annotation of the genes present in the assembly was further obtained by combining transcript alignments, protein alignments and *ab initio* gene predictions as follows.

The RNAseq reads mentioned above (JV134525– JV157085 in Roschanski et al., 2013; 2015; 2016) were aligned to the genome using STAR v2.5.4a (Dobin et al., 2013) with default options and then transcript models were generated from Stringtie (Pertea et al., 2015) also with default options. The resulting models were given to PASA (Haas et al., 2008) v2.2.0 together with 2,806 *A. alba* Expressed Sequence Tags (ESTs) downloaded from NCBI on January 31st, 2018. Next, the TransDecoder program, which is part of the PASA package, was used to detect coding

regions in the PASA assemblies. A BLASTp (Altschul, Gish, Miller, Myers, & Lipman, 1990) search was performed on the Transdecoder predictions against the Swiss-Prot database (The UniProt Consortium, 2017). Sequences with a complete Open Reading Frame (ORF), a BLAST hit against Swiss-Prot (E-value < 1e-9), and not hitting any repeat were considered as potential candidates to train gene predictors. Of this list, the 500 sequences whose length differed the least from the length of their BLAST target were selected as the best candidate genes and used to train the parameters for three gene predictors: GeneID (Parra, Blanco, & Guigo, 2000) v1.4, Augustus (Stankeet, Schoffmann, Morgenstern, & Waack, 2006) v3.2.3 and Glimmer (Majors, Pertea, & Salzberg, 2004). These three gene predictors as well as GeneMark v2.3e (Lomsadze, Burns, & Borodovsky, 2014), which runs in a self-trained mode, were then run on the repeat-masked ABAL 1.0 assembly. Finally, an extra run of each GeneID, Augustus and GeneMark was performed using intron data extracted from the RNAseq mappings.

The complete Pinaceae protein sets present in PLAZA (https://bioinformatics.psb.ugent.be/plaza/versions/gymno-plaza/) in January 2018, were aligned to the repeat-masked genome using exonerate v2.4.7 (Slater & Birney, 2005). Moreover, all the data described above were provided as input to Evidence Modeler v1.1.1 (Haas et al., 2008) and combined into consensus coding sequence (CDS) models. These models were then updated with UTRs and alternative splice isoforms with two rounds of PASA updates.

To remove the potential presence of some bacterial genes in the genome annotation, a protein-based bacterial decontamination procedure was performed on the assembly and annotation. This process utilizes a BLASTp search of the annotated proteins against the bacterial non-redundant protein database from NCBI to detect genes likely to belong to bacteria. All the scaffolds containing more than 50% of bacterial genes and without conifer-specific

repeats and RNAseq mappings were removed from the assembly, resulting in the final assembly ABAL 1.1.

Finally, to check for the presence of the chloroplast genome in the nuclear genome assembly, the chloroplast assembly was mapped to ABAL 1.1 using Minimap2 (Li, 2018) with the parameter --asm10. Sixty-six unique mappings longer than 1 kb were found in the assembly (the longest being 18.49 kb) but they did not meet the threshold of at least 70% matches. Therefore, these regions were considered as nuclear sequence homologous to chloroplast and were kept in the ABAL_1.1 assembly.

The proteins resulting from the structural annotation process described above were functionally annotated using the Blast2GO v4.1 (Conesa et al., 2005) pipeline with default parameters. The annotated proteins were first scanned for InterProScan patterns and profiles. Next, a BLASTp search against the NCBI RefSeq database (Uniprot and Swissprot databases) was performed, inheriting the functional annotations of the top-20 BLAST hits with an e-value < 1e-06. Finally, Blast2GO produced a consensus annotation.

In addition, the software CateGOrize (Zhi-Liang, Bao, & Reecy, 2008) was run to assign all genes to the main Gene Ontology (GO) categories. The software provides the count and percentage of the GO term assigned in each category. Two classification lists (slim2 and myclass2) were used in the analysis. The slim2 list is a subset of gene ontology terms (http://www.geneontology.org/GO.slims.shtml). Myclass2 classification list is based on slim2 with 50 additional GO term categories (Table S2 Supplemental Information). The percentages across the two classification lists were visualised using the *geom_col* function of the "ggplot" package in R CRAN.

2.5.2 Comparison with other conifers

The summary statistics on the annotated genes were computed using a custom python script (available upon request). The same script was applied to calculate the length of exons, introns and genes in other conifer assemblies, such as *P. abies* v1.0, *P. glauca* v3.0, *P. lambertiana* v1.5, *P. taeda* v2.0 and *P. menziesii* v1.5. The distributions of the exon, intron, gene and transcript lengths across the genome were visualized using the *violinBy* function of the "psych" package in R CRAN (R version 3.3.3; 2017-03-06).

3. RESULTS

3.1 Genome sequencing and genome size estimation

PE and MP sequencing produced a total of 1,880,827 and 765,104 Mb, respectively (Table 1). The mean fragment size of the PE300 estimated using *SGA preqc* was 294 bp with a standard deviation of 81 bp.

The estimate of the silver fir genome size, using the distribution of 17-mers (Figure 1) is 17.36 Gb. The plot of all 17-mers present in the PE300 aggregated library that were counted and the number of distinct 17-mers (*k*-mer species) for each depth from 1 to 600 shows the existence of a considerable amount of two-, three- and four-copy repeats (17-mers) in this large genome (Figure 1). The main peak at depth 91X corresponds to unique haploid sequences, while the right-most peaks at depths 182, 273, and 364 correspond to considerable fractions of multi-copy repeat sequences (Figure 1).

3.2 Genome assembly and quality assessment

The silver fir genome sequence presented here accounts for 18.17 Gb, with 37 million scaffolds characterized by an N50 of 14.05 kb (Table 2). The scaffold size ranges between 106 bp and

297,427 bp with a mean size of 489.5 bp. The gaps constitute a total of 236.7 Mb and are relatively small on average (29.3 ± 46.8 bp). The assembly size is slightly higher than the C-value of 16.19 Gb (Roth, Ebert, & Schmidt, 1997) or the *k*-mer-based estimate of 17.36 Gb (Figure 1). However, a comparison of *k*-mer frequency in the PE300 reads and their corresponding copy number in the final assembly using KAT (Figure 2) indicates that most of the homozygous *k*-mers belonging to the haploid peak were assembled. The analysis also reveals only minor collapsing of 2-copy repeats and correct assembly of the remaining multi-copy repeats that are resolvable by this method.

Genome completeness was estimated with three methods based on the presence of conserved genes. CEGMA estimated 81.5% completeness using 248 conserved eukaryotic genes. Using larger gene sets, BUSCO estimated a completeness of 49%, whereas mapping to the *P. taeda* transcriptome resulted in a completeness estimate of 69%. The contiguity of the silver fir assembly was also compared to those of other available conifer genome assemblies (Tree Gene Database; https://treegenesdb.org/). The scaffold N50 (scfN50) of the silver fir assembly was 14.05 kb, almost double that of the 5.21 kb scfN50 of the latest *P. abies* assembly (Paab1.0b) and the 6.44 kb of the *L. sibirica* assembly (Table 3). However, it is still far below those of *P. lambertiana* (2,509.9 kb), *P. glauca* (110.56 kb), *P. taeda* (2,108.3 kb) and *P. menziesii* (372.39 kb; Table 3).

3.3 Chloroplast assembly

De novo assembly, using SPADes and the *A. koreana* complete chloroplast sequence as a reference for mapping, gave an assembly totaling 123,546 bp and contig N50 of 9,211 bp. The second reference-assisted assembly with Ragout using *A. sibirica* and Gapfiller produced a single scaffold of 120,908 bp, comprised of eleven contigs (Table 2). The estimated contig N50

was 15.8 kb. Using the DNAdiff module for genome alignment, a high collinearity was observed with the *A. koreana* and *A. sibirica* complete chloroplast sequences except for a region of~45 kb that align in the opposite direction to *A. koreana* due the presence of inverted repeats (Fig. S3 Supplemental Information). The size of the chloroplast assembly of silver fir was not only close to those of *A. sibiric*a and *A. koreana*, as expected, but also to the 124 kb estimated in *P. abies* (Nystedt et al., 2013), the 121.3 kb in *Abies nephrolepis* (Yi et al., 2015) and 122.6 kb in *L. sibirica* (Bondar, Putintseva, Oreshkova, Krutovsky, 2018). By using Dogma 85 protein coding genes, four rRNA genes and 39 tRNA genes have been annotated. With respect to the *A. koreana* and *A. sibirica* chloroplast genomes, the *A. alba* chloroplast assembly has four duplicated tRNAs (*trn*A-UGC, *trn*I-GAU, *trn*L-UAA and *trn*V-UAC) and *trn*S-UGA has been replaced by *trn*S-CGA.

3.4 Annotation

3.4.1 Protein-coding gene annotation

According to the repeat annotation performed, 78% (14.25 Gb) of the genome assembly correspond to repeats. In the non-repetitive fraction, 94,205 genes were annotated, whose 98,227 transcripts encode 97,750 proteins (Table 4). Of the 97,750 protein sequences, 39,420 (35.8%) were assigned to functional labels, while the rest (58,327 proteins) were analyzed with BLAST, but failed to return significant hits against the RefSeq database. In total, 21,612 of the proteins with complete ORFs were functionally annotated successfully. The number of distinct genes is inflated because many partial genes have been annotated due to the large fragmentation of the assembly. Supporting this assessment, the median gene length was 558 bp, half of the genes were mono-exonic and 47% of the genes had a partial CDS. Actually, approximately half

459 of the annotated proteins (44,646) contained only partial open reading frames (ORFs); they
460 were missing a start or stop codon.

461 Two types of gene models were used to calculate the genome annotation statistics: the
462 protein-coding genes and the full-length genes, respectively. The coding GC content was 46.4%
463 in the protein coding genes and 45.2% in the full-length genes. While the number of exons for
464 the protein-coding genes was 187,740 with a mean length of 327 bp, the number of introns was
465 89,618 (mean length: 320 bp). The number of full-length genes was 50,757 with a median gene
466 length of 804 bp. The number of exons was 118,168 with mean length of 352 bp, the number
467 of introns was 64,728 (mean length: 330 bp) (Table 4, Table S4 Supplemental Information).

468 The distributions of the transcript, intron and exon lengths across the silver fir genome
469 were similar in the protein coding genes and full-length genes (Figures 3A and S4 Supplemental
470 Information). The violin plot showed a different length distribution in the low part of the violin
471 between the two gene models, due to the lower number of short genes in the full-length gene
472 model than in all genes.

473

474 **3.4.2 Comparison with other conifers**

475 The comparison of silver fir genome metrics with other conifer species showed a genome size
476 similar to *P. menziesii* and *P. abies*. Moreover, the gene numbers (94,205) without filtering for
477 quality and completeness were similar to what was found in *P. abies* (70,968), *P. lambertiana*
478 (71,117), and *P. glauca* (102,915), but higher than in *P. menziesii* (54,830), *P. taeda* (47,602),
479 and *L. sibirica* (49,521). When applying a quality filter, more full-length genes (50,757) were
480 found in silver fir than high-confidence genes in *P. lambertiana* (13,936), *P. glauca* (16,386),
481 *P. abies* (28,354), and *P. menziesii* (20,616). The mean and maximum intron lengths were lower

than in the other conifers, while mean exon size was similar to that in *P. taeda, P. glauca, P. abies* and *L. sibirica* (Table 3).

While the distributions of gene length across the genome were similar between silver fir and *P. glauca* (Figure 3B), the mean length in *P. menziesii*, *P. taeda* and *P. lambertiana* was higher than in the other conifers (Table 3). In *P. abies*, the mean gene length was close to that in silver fir, whereas its distribution range was wider (Figure S5A Supplemental Information). The density plot using violin visualization confirmed these differences among species. In particular, the shape of this plot showed the distribution of the genes according to their lengths and highlighted the higher number of short genes in *P. abies*, *P. glauca* and silver fir than in the other conifers (Figure 3B).

The distribution of exon and intron lengths across the silver fir genome was also compared with those found in the other fully sequenced conifers. The exon distribution was similar across species (Figure S5B Supplemental Information), with *P. menziesii* and *P. glauca* showing a slightly lower mean value (Table 3). This was due to the short exons in *P. menziesii*, as it is visualized in the density plot (Figure 3C). The distribution of intron lengths was similar across all species (Figure 3D), with silver fir showing a narrower distribution range than the other conifer species (Figure S5C Supplemental Information).

Silver fir intron and exon statistics were compared to *P. menziesii*, which was sequenced, assembled and annotated using a similar approach (Table S4 Supplemental Information). For *P. menziesii*, the genes were classified into two categories that were based on gene quality and completeness (high-quality and high-quality full-length) and the counts were calculated for both categories. While the numbers of exons and their means were similar in the two species (187,740 for the protein-coding gene model in silver fir and 181,475 for the high-quality gene model in *P. menziesii*), a lower number of introns with a lower mean size was

found in silver fir than in *P. menziesii* (89,618 and 145,595, respectively). Moreover, a lower number of exons and introns per gene was found in silver fir (1.99 and 0.95) than in *P. menziesii* (2.33 and 4.25).

3.4.3 Functional annotation

The input file accounted for 462,216 GO terms that were mapped to the slim2 classification list categories. The total count (Table S5A Supplemental Information) was 27,723 terms corresponding to 32,272 genes, of which 12,221 unique terms belonged to at least one of the 110 slim2 classes. The rest of 1,313 odd terms were not assigned. The 462,216 GO terms were mapped to the myclass2 classification list categories. The total count (Table S5B Supplemental Information) was 31,839 terms corresponding to 32,275 genes, of which 12,361 unique terms belonged to at least one of the 162 myclass2 classes. The rest of 1,173 odd terms were not assigned.

In both classification lists, the main categories were metabolism (11.1% and 9.7% for slim2 and myclass2, respectively), catalytic activity (7.7%, 6.7%), cell (4.7%, 4.1%) and cell organization (4.3%, 3.7%; Table S5 Supplemental Information).

In general, a low percentage of GO terms was assigned to each class. The most abundant (with percentage higher than 0.2%) GO term categories were 61 for the slim2 classification list and 71 for myclass2 (Figure S6A Supplemental Information) and myclass2 classification list (Figure S6B Supplemental Information).

4. DISCUSSION

Here, we present the first *Abies* species whole-genome draft sequence, assembly and annotation. The sequencing strategy used in this project combined Illumina PE and MP libraries

following a protocol similar to that used to sequence other conifer genomes (Neale et al., 2017). The genome size using *k*-mers was estimated to be 17.36 Gb, slightly higher than previous empirical estimates of the haploid C-value of 16.19 Gb (Roth et al., 1997). The assembly comprises over 37 million scaffolds with a total length of 18.16 Gb. Its contiguity is characterized by a contig N50 of 2,477 bp and scaffold N50 of 14kb, and its completeness is estimated to be high with 81.5% of the Core Eukaryotic Genes and at least 69% *P. taeda* transcripts present in the assembly. While this first draft of the silver fir genome is highly fragmented, as were earlier conifer genome assemblies, it represents a very valuable reference resource to the community and can be used immediately to facilitate a broad spectrum of genetic and genomic studies in a demographic, evolutionary, and ecological context.

Given the size and complexity of the silver fir genome, the low contiguity of the assembly obtained with this sequencing approach was not surprising. However, a comparison of the *k*-mer spectra of the reads used to assemble contigs (from haploid material) with their copy number in the final assembly shows that we have obtained a fairly complete assembly. In fact, the majority of the *k*-mers belonging to the main haploid peak are contained in the assembly once and only once, while the peaks of double and triple *k*-mer depth are almost purely 2-copy and three-copy repeats. Only minor collapsing of repeats is observed. Given the haploid nature of the sample (conifer megagamethophyte), we consider these repeat tails to be real and they might contain repeated genes. Therefore, these regions were not removed from the assembly.

The comparison of the distribution lengths of the genes, exons and introns estimated in silver fir with the values found in the assemblies of other conifers showed some interesting results. First, the genes of silver fir were on average shorter than in the other conifer species, except for *P. glauca* (1,190 bp vs 1,330 bp; Warren et al., 2015) and *L. sibirica* (982 bp). However, this might be an effect of the sequencing strategy used and the presence of many

554 short scaffolds in the silver fir assembly, and it will require confirmation with future
555 improvements to the genome sequence. Second, the comparison of the silver fir exons in the
556 current study with those in the other conifers showed similar values for the number, mean length
557 and maximum length of exons, as well as the total amount of exonic sequence (63.7 Mb versus
558 the mean of 50.8 Mb for all compared annotations). This result confirmed that the number and
559 the length of exons are well conserved across species (Sena et al., 2014). The average number
560 of exons per gene was less conserved and the smallest in silver fir (1.92) compared to all other
561 conifers (2.26-8.80). The mean number of exons per gene averaged for all seven species was
562 4.08, which is very close to the value of 3.66 predicted for species such as conifers (Table 2 in
563 Koralewski & Krutovsky, 2011). Given that the average amount of exonic sequence in the
564 conifer genomes analyzed here is only 50.8 Mb, the differences in genome size among conifers
565 are presumably due in large part to the large fraction of repetitive sequences they contain
566 (Morse et al., 2009; Wegrzyn et al., 2013, 2014). Moreover, one of the major components of
567 plant genomes are the transposable elements, which may also affect the evolution of the intron
568 size (Kumar & Bennetzen, 1999).

569 Although intron size has been positively correlated with genome size across eukaryotes
570 (Vinogradov, 1999), this trend is not a rule for seed plants (Wan et al., 2018). Previous studies
571 have reported larger intron sizes in conifers than in angiosperms (Nystedt et al., 2013; Neale et
572 al., 2014; Guan et al., 2016; Sena et al., 2014). This difference is probably related to the high
573 percentage of repetitive sequences, which are the major component of all gymnosperm genomes
574 sequenced to date. Across gymnosperms, *Ginkgo biloba* has longer introns (Guan et al., 2016)
575 than *P. taeda*, but a smaller genome. When comparing the distribution of intron lengths across
576 genomes in several conifers, we found a similar distribution and average between silver fir and
577 *P. glauca* (311 bp vs 511 bp), with the genome size of the latter being almost double (33 Gb)

578 that of silver fir. In contrast, in *P. taeda* and *P. menziesii* the correlation between intron size
579 and genome size was supported by our results, since the intron size was bigger in *P. taeda*
580 (3,004 bp vs 2,301 bp) and also its genome is bigger (20 Mb vs 16 Mb). Moreover, the highest
581 mean intron length across these six species was measured in *P. lambertiana* (10,164 bp) that
582 had a genome size similar to that in *P. glauca* (31 Mb and 32 Mb, respectively), and the smallest
583 both mean and maximum intron lengths were observed in *A. alba* and *L. sibirica* that have also
584 the smallest genome sizes, 16.19 Gb (Roth et al., 1997) and 12.03 Gb (Ohri & Khoshoo, 1986),
585 respectively.

586 Another aspect related to intron length is the suggestion that the expansion of introns
587 occurred early in conifer evolution (Nystedt et al., 2013). This hypothesis was confirmed by the
588 comparison between orthologous introns of *P. taeda* and *G. biloba* that showed a high content
589 of repeats in long introns in both species (Wan et al., 2018). In addition, our analysis showed
590 that the maximum intron length corresponds to *P. taeda* and *P. lambertiana*, and their mean
591 intron length was higher than in other conifer species. The geological timescale calculated for
592 the Pinaceae showed that *Pinus* is the older genus across the Pinaceae, since its presence was
593 confirmed starting from the Early Cretaceous (Wang et al., 2000). The genus *Abies* should be
594 closer to *Pseudotsuga* than to *Picea* and *Pinus* (Wang et al., 2000). Nevertheless, likely due to
595 the high fragmentation of the silver fir genome sequence reported here, the estimated maximum
596 intron length in *A. alba* was only half of that estimated for *P. menziesii*.

597 The assembly of the silver fir chloroplast genome resulted in a single scaffold of 120,908
598 bp that comprised 11 contigs. Each chloroplast has its own genome (cpDNA) that for most
599 plants is formed by four parts: two large inverted repeats, one large single-copy and one small
600 single-copy region. Pinaceae chloroplast genomes lack the inverted repeats. Moreover, the
601 chloroplast genomes in Pinaceae are characterized by the presence of many small repeats and

602 are known to vary in organization (Hipkins, Krutovskii, & Strauss, 1994). The cpDNA
603 organization in Pinaceae was investigated using the *Cedrus* cpDNA as reference, showing the
604 presence of at least three organization types: one similar to *Cedrus* and also found in *Picea*,
605 another similar to *Pseudotsuga*, and another similar to *Larix* (Wu et al., 2011). In addition to
606 *Cedrus*/*Picea*, *Pseudotsuga* and *Larix* organizations, another form of organization was
607 recognized in *Abies* (Tsumura, Suyama, & Yoshimura, 2000). In the current study, we only
608 showed that the chloroplast sequence of silver fir is highly similar and collinear to two other
609 *Abies* species. In addition, the length of the silver fir chloroplast genome is also similar to the
610 other *Abies* chloroplast genome assemblies (Semerikova & Semerikov, 2007; Yi et al., 2015)
611 as well as to that of the *P. abies* chloroplast genome.

612

613 ## 5. CONCLUSION AND PERSPECTIVES

614 Here, we present a draft version of the silver fir genome, which represents a first step towards
615 the full deciphering of this giga-genome in its full complexity. This research is part of the Silver
616 Fir Genome Project, which is a community effort within the Alpine Forest Genomics Network
617 (AForGeN, IUFRO WP 2.04.11; Neale et al., 2013a). The genome sequencing was financed by
618 a bottom-up approach among partners, and the first result is the draft genome sequence
619 presented here (ABAL 1.1)—possibly a profitable strategy for many (plant) genome
620 sequencing initiatives in the future (Twyford, 2018). Long-read sequencing and other
621 approaches for improving the scaffolding are the next steps to be undertaken. Recent advances
622 in genome research have shown that very large and complex genomes may be described in high
623 detail (i.e. Nowoshilow et al., 2018; International Wheat Genome Sequencing Consortium,
624 2018). Therefore, we foresee to improve the genome assembly through additional sequencing
625 approaches complementary to the available Illumina PE and MP reads, such as Bionano optical

626 mapping and PacBio or Oxford Nanopore long-read sequencing, to overcome stretches of
627 repetitive sequences during assembly. Further development of this study could include
628 comparative genomic research exploring phylogenies and evolution in conifer species.
629 Moreover, future research projects could utilize the draft silver fir genome as a reference to re-
630 sequence a diverse panel of trees from contrasting environments and to develop a genotyping
631 array with thousands of single-nucleotide polymorphisms (SNP). Such SNP resources will be
632 useful in many types of demographic studies and, along with the gene annotation presented
633 here, will enable genomic studies and experiments aimed at discovering those genes that are
634 relevant for particular traits (e.g. related to growth) and adaptive responses (e.g. drought
635 tolerance).

ACKNOWLEDGEMENT

The authors thank Berta Fusté from the CNAG-CRG, Centre for Genomic Regulation for her help in managing this project. We would also thank Aleksey Zimin, Daniela Puiu and Michael Schatz for their comments and advice about genome and organelle assembly. This work was in part supported by grants of the National Bioinformatics Institute (INB), PRB2-ISCIII (PT13/0001/0044 to JG). Authors would like to thank "ELIXIR-ITA HPC@CINECA" for providing the computing resources to complete some bioinformatic tasks within this project.

AUTHOR CONTRIBUTIONS

Project conception: D. Neale, B. Heinze, M. Höhn, and C. Sperisen.

Project design: D. Neale, M. Troggio, T. Alioto, F. Gugerli, B. Heinze and M. Höhn.

Financial support provided through: E. Bazin, B. Fady, M. Fladung, B. Fussi, D. Gömöry, S. C. González-Martínez, D. Grivet, F. Gugerli, O.K. Hansen, M. Höhn, B. Heinze, K.V. Krutovsky, G.G. Vendramin, Z. Zaya, B. Ziegenhagen and M. Westergren.

Lab work: C. Rellstab, S. Brodbeck, C. Sperisen, M. Gut.

Bioinformatic analysis: T. Alioto, L. Bianco, F. Cruz, M. Gut, J. Gómez Garrido, K. Ulrich, E. Mosca

Manuscript preparation: E. Mosca, D. Neale, L. Bianco, M. Troggio, F. Cruz, J. Gómez Garrido, T. Alioto , F. Gugerli, C. Rellstab, S. Brodbeck, K. Csilléry, and K. Ullrich.

All authors approved the manuscript.

DATA ACCESSIBILITY

The silver fir genome assembly ABAL 1.1 is available in the TreeGenes Database with the following link: https://treegenesdb.org/FTP/Genomes/Abal/

DISCLOSURE DECLARATION

The authors declare no competing interest.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Molecular Biology Journal, 215,* 403-410.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pevzner, P. A. (2002). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology, 19,* 455-477.

Bennett, M. D., & Leitch, I. J. (2011). Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Annals of Botany, 107,* 467-590.

Birol, I., Raymond, A., Jackman, S.D., Pleasance, S., Coope, R., Taylor, G. A., … Jones, S. J. M. (2014). Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics, 29,* 1492-1497.

Boetzer, M., & Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome Biology, 13,* R56.

Bondar, E. I., Putintseva, Y. A., Oreshkova, N. V., Krutovsky, K. V. (2018). Siberian larch (*Larix sibirica* Ledeb.) chloroplast genome and development of polymorphic chloroplast markers. *BMC Bioinformatics* (accepted, in press)

Büntgen, U., Tegel, W., Kaplan, J. O., Schaub, M., Hagedorn, F., Bürgi, M., … Liebhold, A. (2014). Placing unprecedented recent fir growth in a European-wide and Holocene-long context. *Frontiers in Ecology and the Environment, 12,* 100-106.

Cailleret, M., Nourtier, M., Amm, A., Durand-Gillmann, M., Davi, H. (2014). Drought-induced decline and mortality of silver fir differ among three sites in Southern France. *Annals of Forest Science, 71,* 1-15

Cailleret, M., Davi, H. (2012). Effects of climate on diameter growth of co-occurring *Fagus sylvatica* and *Abies alba* along an altitudinal gradient. *Trees, 25,* 265–276.

Crepeau, M. W., Langley, C. H., & Stevens, K. A. (2017). From Pine Cones to Read Clouds: Rescaffolding the Megagenome of Sugar Pine (*Pinus lambertiana*). *G3* (Bethesda, Md.), *7(5),* 1563-1568.

Csilléry, K., Sperisen, C., Ovaskainen, O., Widmer, A., Gugerli, F. (2018). Adaptation to local climate in size, growth and phenology across 19 silver fir (*Abies alba* Mill.) populations from Switzerland. bioRxiv 292540, doi: 10.1101/292540.

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics, 21,* 3674-3676.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., …Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29,* 15-21.

Ellenberg, H. 2009. Coniferous woodland and mixed woods dominated by conifers. Pages 191–242. Vegetation ecology of Central Europe. Cambridge University Press, Cambridge, UK.

Elling, W., Dittmar, C., Pfaffelmoser, K., Rötzer, T. (2009). Dendroecological assessment of the complex causes of decline and recovery of the growth of silver fir (*Abies alba* Mill.) in Southern Germany. *Forest Ecology and Management, 257,* 1175-1187.

Genomic Services Lab of HustonAlpha (2010). BAM2FASTQ version1.1.0. https://gsl.hudsonalpha.org/information/software/bam2fastq (18 August 2010).

George J.P., Schueler, S., Karanitsch-Ackerl, S., Mayer, K., Klumpp, R. T., Grabner M. (2015). Inter- and intra-specific variation in drought sensitivity in *Abies* spec. and its relation to wood density and growth traits. *Agricultural and Forest Meteorology,* 214-215, 430-443.

Grattapaglia, D., Orzenil, B. S.-J., Tassinari Resende, R., Cappa, E. P., Salomão de Faria Müller, B., Tan B., … El-Kassaby, Y.A. (2018). Quantitative genetics and genomics converge to accelerate forest tree breeding. *Frontiers in Plant Science, 9,* 1693

Grotkopp, E., Rejmánek, M., Sanderson, M. J., & Rost, T. L. (2004). Evolution of genome size in pines (*Pinus*) and its life history correlates: supertree analyses. *Evolution, 58,* 1705-1729.

Guan, R., Zhao, Y., Zhang, H., Fan, G., Liu, X., Zhou, W., … Chen, W.B. (2016). Draft genome of the living fossil *Ginkgo biloba*. *GigaScience, 21,* 49.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr., R. K., Hannick, L. I., … White, O. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research, 31,* 5654-5666.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., …Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments**.** *Genome Biology, 9,* R7.

Hipkins, V. D., Krutovskii, K. V., & Strauss, S. H. (1994). Organelle genomes in conifers: structure, evolution, and diversity. *Forest Genetics, 1,* 179–189.

Heer, K., Behringer, D., Piermattei, A., Bässler, C., Brandl, R., Fady, B., … Opgenoorth, L. (2018). Linking dendroecology and association genetics in natural populations: Stress responses archived in tree rings associate with SNP genotypes in silver fir (*Abies alba* Mill.). *Molecular Ecology, 27,* 1428-1438.

International Wheat Genome Sequencing Consortium (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science, 361,* 661-673.

Kandler, O. & Innes, J. L. (1995). Air pollution and forest decline in Central Europe. *Environmental Pollution, 90,* 171-180.

Kolmogorov, M., Raney, B., Paten, B., & Pham, S. (2014). Ragout—a reference-assisted assembly tool for bacterial genomes. *Bioinformatics, 30,* i302-i309.

Koralewski, T. E., Krutovsky, K. V. (2011). Evolution of exon-intron structure and alternative splicing. *PLoS ONE, 6,* e18055.

Kumar, A., & Bennetzen, J. L. (1999). Plant retrotransposons. *Annual Review of Genetics*, 33, 479-532.

Kurtz, S., Phillippy, A., Delcher, A., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S. (2004). Versatile and open software for comparing large genomes. *Genome Biology, 5,* R12

Kuzmin, D. A., Feranchuk, S. I., Sharov, V. V., Cybin, A. N., Makolov, S. V., Putintseva, Y. A., Oreshkova, N. V., & Krutovsky, K. V. (2018) Stepwise large genome assembly approach: A case of Siberian larch (*Larix sibirica* Ledeb.). *BMC Bioinformatics* (accepted, in press)

Lebourgeois, F., Rathgeber, C. B. K., & Ulrich, E. (2010). Sensitivity of French temperate coniferous forests to climate variability and extreme events (Abies alba, Picea abies and Pinus sylvestris). Journal of Vegetation Science, 21, 364–376.)

Leitch, I. J., Soltis, D.E., Soltis, P.S., Bennett, M. D. (2005). Evolution of DNA amounts across land plants (Embryophyta). *Annals of Botany, 95,* 207-217.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34,* 3094-3100.

Li, H. & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics, 26,* 589-595.

Lomsadze, A., Burns, P.D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm**.** *Nucleic Acids Research*, *42*, e119.

Majoros, W.H., Pertea, M., & Salzberg, S.L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, *20*, 2878-2879.

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., & Clavijo, B. J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies, *Bioinformatics*, *33*, 574-576.

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*, 764-770.

Martin, M. (2011). Cutadapt removes adapter sequences from high-through put sequencing reads. *EMBnet.journal*, *1*, 17.

Morse, A.M., Peterson, D.G., Islam-Faridi, M.N., Smith, K.E., Magbanua, Z., Garcia, S.A., …Davis, J.M. (2009). Evolution of genome size and complexity in *Pinus*. *PLoS ONE*, *4*, e4332.

Mosca, E., Eckert, A.J., Di Pierro, E.A., Rocchini, D., La Porta, N., Belletti, P., Neale, D. B. (2012). The geographical and environmental determinants of genetic diversity for four alpine conifers of the European Alps. *Molecular Ecology*, *21*, 5530-5545.

Mosca, E., Gonzáles-Martínez, S. C., Neale D. B. (2014). Environmental versus geographical molecular adaptation in two subalpine conifers. *New Phytologist*, *201*, 180-192.

Neale, D.B., Mosca, E., & Di Pierro, E. A. (2013a). Alpine forest genomics network (AForGeN): a report of the first annual meeting. *Tree Genetics & Genomes*, *9*, 879-881.

Neale, D.B., Langley, C.H., Salzberg, S. L., & Wegrzyn, J. L. (2013b). Open access to tree genomes: the path to a better forest. *Genome Biology*, *14*, 120.

Neale, D.B., Wegrzyn, J.L., Stevens, K. A., Zimin, A. V., Puiu, D., Crepeau, M. W., … Langley, C. H. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, *15*, R59.

Neale, D.B., McGuire, P.E., Wheeler, N.C., Stevens, K.A., Crepeau, M.W., Cardeno, C., … Wegrzyn, J.L. (2017). The Douglas-fir genome sequence reveals specialization of the photosynthetic apparatus in Pinaceae. *G3: Genes, Genomes, Genetics*, *9*, 3157-3167.

Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pahl, A. W. C., Pippel, M., … Meyers, E. W. (2018). The axolotl genome and the evolution of key tissue formation regulators. *Nature, 554*, 50-55.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., … Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature, 497*, 579-584.

Ohri, D., Khoshoo, T. N. (1986). Genome size in gymnosperms. *Plant Systematics and Evolution*, *153*, 119-132.

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, *23*, 1061-1067.

Parra, G., Blanco, E., & Guigo, R. (2000). GeneID in Drosophila. *Genome Resource*, *10*, 511-515.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*, 290-295.

Piotti, A., Leonarduzzi, C., Postolache, D., Bagnoli, F., Spanu, I., Brousseau, L., Vendramin, G. G. (2017). Unexpected scenarios from Mediterranean refugial areas: disentangling

complex demographic dynamics along the Apennine distribution of silver fir. *Journal of Biogeography*, *44*, 1547-1558.
Roth, R., Ebert, I., Schmidt, J. (1997). Trisomy associated with loss of maturation capacity in a long-term embryogenic culture of *Abies alba*. *Theoretical and Applied Genetics*, *95*, 353-358.
Roschanski, A. M., Csilléry, K., Liepelt, S., Oddou-Muratorio, S., Ziegenhagen, B., Huard, F., …Fady, B. (2015). Data from: Evidence of divergent selection for drought and cold tolerance at landscape and local scales in *Abies alba* Mill. in the French Mediterranean Alps. Dryad Digital Repository. https://doi.org/10.5061/dryad.t671s.
Roschanski, A. M., Csilléry, K., Liepelt, S., Oddou-Muratorio, S., Ziegenhagen, B., Huard, F. … Fady, B. (2016). Evidence of divergent selection for drought and cold tolerance at landscape and local scales in *Abies alba* Mill. in the French Mediterranean Alps. *Molecular Ecology*, *25*, 776-794.
Roschanski, A. M., Fady, B., Ziegenhagen, B., & Liepelt, S. (2013). Annotation and re-sequencing of genes from de novo transcriptome assembly of *Abies alba* (Pinaceae). *Applications in Plant Sciences*, *1*, 1-8.
Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J., & Arvestad, L. (2014). BESST - Efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics*, *15*, 281.
Semerikova, S. A., & Semerikov, V. L. (2007). The diversity of chloroplast microsatellite loci in Siberian fir (*Abies sibirica* Ledeb.) and two Far East fir species *A. nephrolepis* (Trautv.) Maxim. and *A. sachalinensis* Fr. Schmidt. *Genetika* , *43*, 1637-1646.
Sena, J. S., Giguère, I., Boyle, B., Rigault, P., Birol, I., Zuccolo, A., … Mackay, J. (2014). Evolution of gene structure in the conifer *Picea glauca*: a comparative analysis of the impact of intron size. *BMC Plant Biology*, *14*, 95
Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*, 3210-3212.
Simpson, J. T. (2014). Exploring genome characteristics and sequence quality without a reference. *Bioinformatics*, *30*, 1228-1235.
Simpson, J. T., & Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, *22*, 549-556.
Slater, G. S., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*, 31.
Stanke, M., Schoffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, *7*, 62.
Stevens, K. A., Wegrzyn, J. L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., … Langley, C.H. (2016). Sequence of the sugar pine megagenome. *Genetics*, *204*, 1613-1626.
Tinner, W., Colombaroli, D., Heiri, O., Henne, P. D., Steinacher, M., Untenecker, J., … Valsecchi, V. (2013) The past ecology of *Abies alba* provides new perspectives on future responses of silver fir forests to global warming. *Ecological Monographs*, *83*, 419-439.
Tsumura, Y., Suyama, Y., & Yoshimura, K. (2000). Chloroplast DNA inversion polymorphism in populations of *Abies* and *Tsuga*. *Molecular Biology and Evolution*, *17*, 1302-1312.
Twyford, A. D. (2018). The road to 10,000 plant genomes. *Nature Plants*, *4*, 312-313.
Vinogradov, A. E. (1999). Intron genome size relationship on a large evolutionary scale. *Journal of Molecular Evolution*, *49*, 376-384.

Vitali, V., Büntgen, U., & Bauhus, J. (2017) Silver fir and Douglas fir are more tolerant to extreme droughts than Norway spruce in south-western Germany. *Global Change Biology, 23,* 5108-5119.

Vrška, T., Adam, D., Hort, L., Kolář, T., & Janík, D. (2009). European beech (*Fagus sylvatica* L.) and silver fir (*Abies alba* Mill.) rotation in the Carpathians – a developmental cycle or a linear trend induced by man? *Forest Ecology and Management, 258,* 347-356.

Wan, T., Liu, Z. M., Li, L. F., Leitch, A. R., Leitch, I. J., Lohaus, R., Liu, Z. J., … Wang, X. M. (2018). A genome for gnetophytes and early evolution of seed plants. *Nature Plants*, *4*, 82-89.

Wang, X.-Q., Tank, D. C., & Sang, T. (2000). Phylogeny and divergence times in Pinaceae: evidence from three genomes. *Molecular Biology and Evolution, 17,* 773-781.

Warren, R. L., Keeling, C. I., Yuen, M. M., Raymond, A. , Taylor, G. A., Vandervalk, B. P., … Bohlmann, J. (2015), Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *Plant Journal, 83,* 189-212.

Wegrzyn, J. L., Liechty, J. D., Stevens K. A., Wu, L. S., Loopstra, C. A., … Neale D. B. (2014). Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *, 196,* 891-909.

Wegrzyn, J. L., Lin, B. Y., Zieve, J. J., Dougherty, W. M., Martínez-García, P. J., Koriabine, M., … Stevens K. A. (2013). Insights into the loblolly pine genome: characterization of BAC and fosmid sequences. *PLoS ONE, 8,* e72439.

Wegrzyn, J. L., Main, D., Figueroa, B., Choi, M., Yu, J., Neale, D. B., Jung, S., … , Abbott, A. G. (2012). Uniform standards for genome databases in forest and fruit trees. *Tree Genetics & Genomes*, *8*, 549-557.

Wyman, S. K., Jansen, R. K., & Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics, 20,* 3252-3255.

Wolf, H. (2003). EUFORGEN Technical Guidelines for genetic conservation and use for silver fir (*Abies alba*). Rome: International Plant Genetic Resources Institute; p. 6.

Wu, C.-S., Lin, C.-P., Hsu, C.-Y., Wang, R.-J., & Chaw, S.-M. (2011). Comparative chloroplast genomes of Pinaceae: insights into the mechanism of diversified genomic organizations. *Genome Biology and Evolution, 3*, 309-319.

Wu, T. D., Reeder, J., Lawrence, M., Becker, G., & Brauer, M. J. (2016). GMAP and GSNAP for Genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods in Molecular Biology, 1418,* 283-334.

Yi, D.-K., Yang, J.C., So, S. K., Joo, M., Kim, D. K., Shi, C. H., Lee, Y. M., Choi, K. (2015). The complete plastid genome sequence of *Abies koreana* (Pinaceae: Abietoideae). *Mitochondrial DNA, 27,* 2351-2353.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics, 29,* 2669-2677.

Zhi-Liang, H., Bao, J., & Reecy, J. M. (2008). CateGOrizer: A web-based program to batch analyze Gene Ontology classification categories. *Online Journal of Bioinformatics*, *9*, 108-112.

Zonneveld, B. J. M. (2012). Genome sizes of 172 species, covering 64 out of the 67 genera, range from 8 to 72 picogram. *Nordic Journal of Botany, 30,* 490-502.

Figure captions

FIGURE 1. Distribution of 17-mers in the whole-genome sequence of *Abies alba* using raw paired-end (PE) 2 × 151 bp reads generated from the PE300 library with 300 bp long fragment inserts and estimated with Jellyfish 2.2.0 (Marçais & Kingsford, 2011). The high peak at very low depths is caused by sequencing errors.

FIGURE 2. Spectra Copy Number in the *Abies alba* genome ABAL 1.1. Comparison between the *k*-mer (*k*=27) spectra of paired-end (PE) 300 2 x 151 bp reads generated from the PE300 library with 300 bp long fragment inserts and the ABAL 1.1 assembly. This stacked histogram was produced with KAT (Mapleson et al., 2016) that shows the spectra copy number classes along the assembly.

FIGURE 3. Violin plot of the distribution length of the genes, transcripts, exons and introns across the *Abies alba* (Abies_al) high-quality genes and full-length genes (indicated as "full"; **A**). The length was log10 transformed. Violin plot of the distribution lengths of genes (**B**), exons (C) and introns (D) across the *Abies alba* (A_alba) high-quality genes and full-length genes, *Pseudotsuga menziesii* (Ps_menz), *Picea abies* (P_abies), *Picea glauca* (P_glauca), *Pinus taeda* (P_taeda), *Pinus lambertiana* (P_lamb).

List of supplementary material

TABLE S1. Estimation of DNA concentration, 260/280 and 260/230 ratios and DNA integrity in the two sample types (megagametophyte and needle) used for DNA extraction in *A. alba*.

TABLE S2. Gene ontology (GO) term categories used to count the GO terms of *A. alba*. GO_slim2 is an option in CateGOrize software and myclass2 accounts for 50 additional categories.

TABLE S3. *A. alba* genome annotation statistics considering two types of gene models (protein coding genes and full-length genes).

TABLE S4. Intron and exon statistics for silver fir (*A. alba*) and Douglas-fir (*Pseudotsuga menziesii*) gene models.

TABLE S5. Count and percentage (fraction) of the GO terms assigned in each category using the two classification lists (**A**: slim2 and **B**: myclass2) to be complemented.

FIGURE S1. Distribution map of *A. alba* natural stand, compiled by the EUFORGEN Network members (EUFORGEN 2009).

FIGURE S2. (**A**) Location of the 19 sampled Swiss populations and tree AA_WSL01. Modified after Csilléry et al. (2018). (**B**) The log-likelihood from Structure runs with K = 2 to K=10. (**C**) Ancestry proportions of AA_WSL01 and the 19 genotyped Swiss populations for K=3 and K=4.

FIGURE S3. Plot produced with DNAdiff for the comparison between *A. alba* and *A. sibirica* chloroplasts (**A**) and *A. alba* and *A. koreana* chloroplasts (**B**).

FIGURE S4. Boxplots of the distribution lengths of the genes, transcripts, exons and introns across the *A. alba* high-quality genes and full-length genes (indicated as "full"). The distribution is log10 transformed.

FIGURE S5. Boxplots of the distribution lengths of the genes (**A**), exons (**B**), and introns (**C**) across the *Abies alba* (A_alba) high-quality genes and full-length genes (indicated as "full"), *Pseudotsuga menziesii* (Ps_menz), *Picea abies* (P_abies), *Picea glauca* (P_glauca), *Pinus taeda* (P_taeda), *Pinus lambertiana* (P_lamb).

FIGURE S6. Distribution of the most abundant Gene Ontology (GO) terms assigned to the *A. alba* genome using slim2 categories (**A**) and myclass2 categories (**B**). The percentage (fraction) of the term assigned in each category is represented only for values > 0.2%. All categories are given in Table S2, all count and percentages in Table S5.

TABLE 1 Summary of the raw data for Illumina paired-end (PE) and mate-pair (MP) libraries for whole-genome sequencing of *Abies alba*.

Library	Read length (bp)	Insert size (kb)	Mean fragment size (bp)	Read Pairs (million)	Yield (Mb)	Coverage	Avg. Phix Error R1 (%)	Avg. Phix Error R2 (%)
PE300-1	2 x 151	-	304	3,274	989,029	57.103	0.646	0.908
PE300-2	2 x 151	-	307	1,886	569,617	32.888	0.883	1.126
PE300-3	2 x 151	-	312	1,066	322,181	18.602	0.768	1.081
MP1500	2 x 101	1.5	-	1,255	253,529	14.638	0.214	0.32
MP3000	2 x 101	3	-	1,277	257,985	14.895	0.214	0.32
MP8000	2 x 101	8	-	1,255	253,590	14.641	0.214	0.32
Total PE				6,226	1,880,827	108.593		
Total MP				3,787	765,104	44.175		

974 **TABLE 2** Summary statistics for the *Abies alba* whole-genome assembly version 1.1 (ABAL
975 1.1) and chloroplast assembly.
976

Genome	Feature	
Nuclear	Number of contigs	45,280,944
	Number of scaffolds	37,192,295
	Mean GC%	39.34
	Total length (Mb)	18,167
	Minimum scaffold length (bp)	106
	Maximum scaffold length (bp)	297,427
	Mean scaffold length (bp)	488.50
	Median scaffold length (bp)	115
	Contig N50 (bp)	2,477
	Scaffold N50 (bp)	14,051
Chloroplast	Total length (bp)	120,908
	Number of contigs	11
	Number of scaffolds	1
	Contig N50 (bp)	15,758

977

TABLE 3 Comparison of genome summary metrics from *A. alba* and other sequenced conifer genomes (version numbers in parentheses).

Genome summary metric	***Abies alba (1.0)***	***Pseudotsuga menziesii (1.5)***	***Pinus taeda (2.0)***	***Pinus lambertiana (1.5)***	***Picea glauca (3.0)***	***Picea abies (1.0)***	***Larix sibirica (1.0)****
Total length (Mb)	18,167	15,700	20,613	31,000	32,795	19,600	12,340
N50 scaffold (Kb)	14.05	372.39	2,108.3	2,509.9	110.56 34.40[§]	5.21	6.44
N of genes	94,205	54,830	47,602	71,117[¶]	102,915	70,968	49,521
N of full-length genes	50,757	20,616	NA	13,936[¶]	16,386[§]	28,354[ø]	32,482
N of exons	181,168	181,475	166,465	153,111	232,182	178,049	151,838
N of introns	64,728	145,595	108,809	121,858	124,951	107,313	101,675
Mean gene length (bp)	1,190	10,510	9,066	40,820	1,330	2,427	982
Mean exon length (bp)	352	231	320	241	320	312	324
Mean intron length (bp)	311	2,301	3,004	10,164	511	1,017	353
Maximum exon length (bp)	6,300	8,037	4,946	8,003	9,568	6,068	10,268
Maximum intron length (bp)	36,015	182,831	408,800	805,500	44,116	68,269	10,154
Exons per gene	1.92	8.80	3.50	5.25	2.26	3.78	3.03
Total exonic length	$6.4x10^6$	$4.2x10^6$	$5.3x10^6$	$1.8x10^6$	$7.4x10^6$	$5.6x10^6$	$4.9x10^6$

For the gene annotation and the definition of the "full-length genes" different approches were used across species. The scaffold N50 (scfN50) was calculated on the unshuffled assemblies and discarding scaffolds shorter than 200 bp.

*Kuzmin et al., 2018; K.V. Krutovsky, personal communication

[§] high confidence set (Warren et al., 2015; PG29 v3) and scaffold N50 calculated using sequences >= 500 bp: N50 is 71.5 kb if considering both clones (WS77111)

[¶] low-quality and high quality gene models from *Pinus lambertiana* v.1 (Stevens et al., 2016), the other were calculated on *Pinus lambertiana* v1.5 (Crepeau et al., 2017),

[ø] high confidence (Nystedt et al., 2013)

992 **TABLE 4** Genome annotation statistics for *A. alba* considering two types of gene models
993 (protein coding genes and full-length genes). All statistics are given in Table S3.
994

Features	**Protein-coding genes**	**Full-length genes**
Number of genes	94,205	50,757
Median gene length (bp)	558	804
Number of transcripts	98,227	53,487
Median transcript length (bp)	445	597
Number of exons	187,740	181,168
Coding GC content	46.4%	45.15%
Median exon length (bp)	224	237
Number of introns	89,618	64,728
Median intron length (bp)	146	145
Exons/transcript	2.00	2.32
Transcripts/gene	1.04	1.05

995