Sequence of the Sugar Pine Megagenome

Kristian A. Stevens,*,[1,2] Jill L. Wegrzyn,[†,1] Aleksey Zimin,[‡] Daniela Puiu,[§] Marc Crepeau,* Charis Cardeno,* Robin Paul,[†] Daniel Gonzalez-Ibeas,[†] Maxim Koriabine, Ann E. Holtz-Morris,** Pedro J. Martínez-García,[††] Uzay U. Sezen,[†] Guillaume Marçais,[‡] Kathy Jermstad,[‡‡] Patrick E. McGuire,[††] Carol A. Loopstra,[§§] John M. Davis,*** Andrew Eckert,[†††] Pieter de Jong,** James A. Yorke,[‡] Steven L. Salzberg,[§,‡‡‡] David B. Neale,[††] and Charles H. Langley*,[2]**

*Department of Evolution and Ecology and [††]Department of Plant Sciences, University of California at Davis, California, 95616 [†]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut, 06269 [‡]Institute for Physical Sciences and Technology (IPST), University of Maryland, College Park, Maryland, 20742 [§]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine and [‡‡‡]Departments of Biomedical Engineering, Computer Science, and Biostatistics, Johns Hopkins University, Baltimore, Maryland, 21205 **Children's Hospital Oakland Research Institute, Oakland, California, 94609 [‡‡]United States Department of Agriculture Forest Service, Pacific Southwest Research Station, Placerville, California, 95667 [§§]Department of Ecosystem Science and Management, Texas A&M University, College Station, Texas, 77843 ***School of Forest Resources and Conservation, University of Florida, Gainesville, Florida, 32603 and [†††]Department of Biology, Virginia Commonwealth University, Richmond, Virginia 23284

ABSTRACT Until very recently, complete characterization of the megagenomes of conifers has remained elusive. The diploid genome of sugar pine (*Pinus lambertiana* Dougl.) has a highly repetitive, 31 billion bp genome. It is the largest genome sequenced and assembled to date, and the first from the subgenus *Strobus*, or white pines, a group that is notable for having the largest genomes among the pines. The genome represents a unique opportunity to investigate genome "obesity" in conifers and white pines. Comparative analysis of *P. lambertiana* and *P. taeda* L. reveals new insights on the conservation, age, and diversity of the highly abundant transposable elements, the primary factor determining genome size. Like most North American white pines, the principal pathogen of *P. lambertiana* is white pine blister rust (*Cronartium ribicola* J.C. Fischer ex Raben.). Identification of candidate genes for resistance to this pathogen is of great ecological importance. The genome sequence afforded us the opportunity to make substantial progress on locating the major dominant gene for simple resistance hypersensitive response, *Cr1*. We describe new markers and gene annotation that are both tightly linked to *Cr1* in a mapping population, and associated with *Cr1* in unrelated sugar pine individuals sampled throughout the species' range, creating a solid foundation for future mapping. This genomic variation and annotated candidate genes characterized in our study of the *Cr1* region are resources for future marker-assisted breeding efforts as well as for investigations of fundamental mechanisms of invasive disease and evolutionary response.

KEYWORDS conifer genome; transposable elements; white pine blister rust

THE gymnosperm genus *Pinus* is diverse and ubiquitous in temperate zones (Critchfield and Little 1966; Farjon and Filer 2013). Pines are often the keystone trees of terrestrial ecosystems (Richardson and Rundel 1998; Keane *et al.* 2012, and citations therein). Typical of conifers, pines have megagenomes that vary greatly in size among species, yet their karyotype is highly conserved. *Pinus* is divided into two large, ancient monophyletic subgenera, *Strobus* and *Pinus*, "white pines" and "yellow pines," respectively (Critchfield and Little 1966; Gernandt *et al.* 2005). The first *Pinus* genome sequence (22 Gbp) was recently reported for *Pinus taeda* L. (Zimin *et al.* 2014), a yellow pine commonly known as loblolly pine. The genomes of white pines are larger and more variable in size (Tomback 1982). Fossils allied with *Strobus* are known from the early Tertiary and late Cretaceous (Millar 1998), consistent with molecular phylogenetic dating of the crown group *Strobus* at 45–85 MYA (Willyard *et al.* 2007; DeGiorgio

doi: 10.1534/genetics.116.193227
Manuscript received June 29, 2016; accepted for publication October 25, 2016; published Early Online October 28, 2016.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.193227/-/DC1.
[1]These authors contributed equally to this work.
[2]Corresponding authors: Department of Evolution and Ecology, University of California at Davis, One Shields Ave, Davis, CA 95616. E-mail: kastevens@ucdavis.edu; and chlangley@ucdavis.edu

et al. 2014). Populations of a number of the majestic white pines of North America, and their associated ecosystems, have been devastated over the last century by white pine blister rust, WPBR (Kinloch 1992) caused by a highly pathogenic and invasive fungus, *Cronartium ribicola* J.C. Fischer ex Raben. While major gene resistance to this disease has been discovered in several species, and loci have been placed on the genetic maps of *Pinus lambertiana* Dougl. (Harkins *et al.* 1998; Jermstad *et al.* 2011) and *P. monticola* Dougl. ex D.Don (Liu *et al.* 2006), the discovery of the underlying genes, and of markers serviceable for genetic improvement in reforestation, may be greatly accelerated by the genome sequence itself.

P. lambertiana, commonly known as sugar pine, is a white pine native to western North America that is distributed from northern Oregon to Baja California at a wide span of altitudes. It is currently the tallest pine species, with heights reaching 76 m. The female cones of sugar pine are also gigantic, often longer than 600 mm (Kinloch and Scheuner 1990; Van Pelt 2001; American Forests 2015). *P. lambertiana* trees may live $>$ 500 years, and the onset of the species' sexual reproduction is delayed compared to other pines, possibly due to the height and girth needed to support these massive strobili. Paralleling these oversized dimensions, the genome of *P. lambertiana* was estimated from cytometry to be 31 Gbp (see below), nearly 50% larger than that of *P. taeda* and ten times the size of the human genome. While *P. lambertiana* was historically a significant timber source, heavy harvesting, and the arrival of the devastating white pine blister rust to its range, has changed the management focus. Since this species plays important ecological roles in the maintenance of biodiversity, carbon sequestration, soil stabilization, and watershed protection (Maloney 2012), considerable effort and resources have been deployed both by the US Forest Service and the private sector to structure the genetics of reforestation to fit the ecological factors, especially WPBR (reviewed in Waring and Goodrich 2012). In particular, the screening by progeny testing of diverse seed sources for individual trees carrying the major gene for WPBR resistance, *Cr1* (Kinloch 1992), has been ongoing for more than a decade. These extra costs of collecting seeds from candidate trees throughout the species range, of progeny testing for WPBR resistance (requiring several years), and the deployment of resistant seedlings, are significant components of forest management. Genotyping by markers with strong associations to WPBR resistance has the potential to greatly reduce both the effort and time required by the ongoing approach, and could open new strategies. Here, we demonstrate that the sequencing, assembly, and annotation of the genome sequence of *P. lambertiana* greatly accelerates the discovery of such genetic tools.

Conifer evolution and genome size

All members of the genus *Pinus* have 12 chromosomes (Saylor 1960) and are considered to be karyotypically stable throughout their evolutionary history (Sax 1960; Saylor 1964). With the exception of a potential event preceding the radiation (Li *et al.* 2015), whole genome polyploidy is thought to be absent among the $\geq$100 species. However, the amount of nuclear DNA that comprises a single copy of a pine genome can vary widely between species. Flow cytometric estimates for the genus *Pinus* in the *C*-values database (Bennett and Leitch 2012) range from a low of 20 Gbp for *P. muricata* D. Don, to a high of 35 Gbp for *P. ayacahuite* Ehrenb. ex Schltdl. (Figure 1B). The correlates and causes of this variation in genome size, including in *Pinus*, are an open topic of speculation and investigation (Williams *et al.* 2002; Grotkopp *et al.* 2004; Ahuja and Neale 2005; Morse *et al.* 2009).

The two subgenera of *Pinus* diverged $\sim$45–85 MYA ago (Figure 1A) (see also Willyard *et al.* 2007). Members of *Strobus* have an average genome size 5.2 Gbp larger than the subgenus *Pinus* (Figure 1B) (Grotkopp *et al.* 2004). The majority of sequenced conifer megagenomes are composed of interspersed repetitive sequences, with estimates ranging from 69% for *Picea abies* (L.) H. Karst. (Nystedt *et al.* 2013) to 80% for *P. taeda* (Wegrzyn *et al.* 2014). The evolutionary dynamics of transposable elements (TEs) have long been suspected to shape genomic change, including overall genome size, in numerous species (Orgel and Crick 1980; Hawkins 2006; Piegu *et al.* 2006; Tenaillon *et al.* 2011), including conifers (Nystedt *et al.* 2013). In contrast to angiosperms, where genome duplication events and LTR retrotransposon bursts are frequent, and account for most of the genome size expansions, a continual accretion of repeats may provide a better explanation of genome size variation within the genus *Pinus* (Morse *et al.* 2009). The genome sequence of *P. lambertiana* presents a new opportunity to address elements of the hypothesis that TE dynamics are behind these significant changes in genome size.

White pine blister rust

WPBR, the non-native heteroecious fungus *Cronartium ribicola*, infects North American pines of the *Strobus* subgenus. An invasive species, *C. ribicola* has devastated populations of five-needle pines, including *P. strobus* L. (eastern white pine), *P. monticola* (western white pine), *P. lambertiana* (sugar pine), *P. flexilis* James (limber pine), and *P. albicaulis* Engelm. (whitebark pine), and foxtail pine, along with closely related bristlecone pines (subgenus *Pinus* subsection *Balfourianae*) since its introduction from Asia or Europe a century ago. Damage from *C. ribicola* is known to reduce reproduction and survival of the majority of white pine species (Kinloch 1970; Waring and Goodrich 2012). Exacerbated by recent outbreaks of the mountain pine beetle, decreasing pine populations have affected wildlife, biodiversity, watershed, and timber potential. Rare individuals among the white pines species exhibit innate and heritable resistance that forms the basis for various selective reforestation efforts (Kinloch 2003). A major "gene" of resistance (MGR) to WPBR was mapped in *P. lambertiana* over 40 years ago (Kinloch 1970). An apparently biallelic locus, $Cr1^R/Cr1^r$ locus has been mapped in several *P. lambertiana* families (Devey *et al.* 1995; Harkins *et al.* 1998; Jermstad *et al.* 2011). In this work, we leverage these markers and the

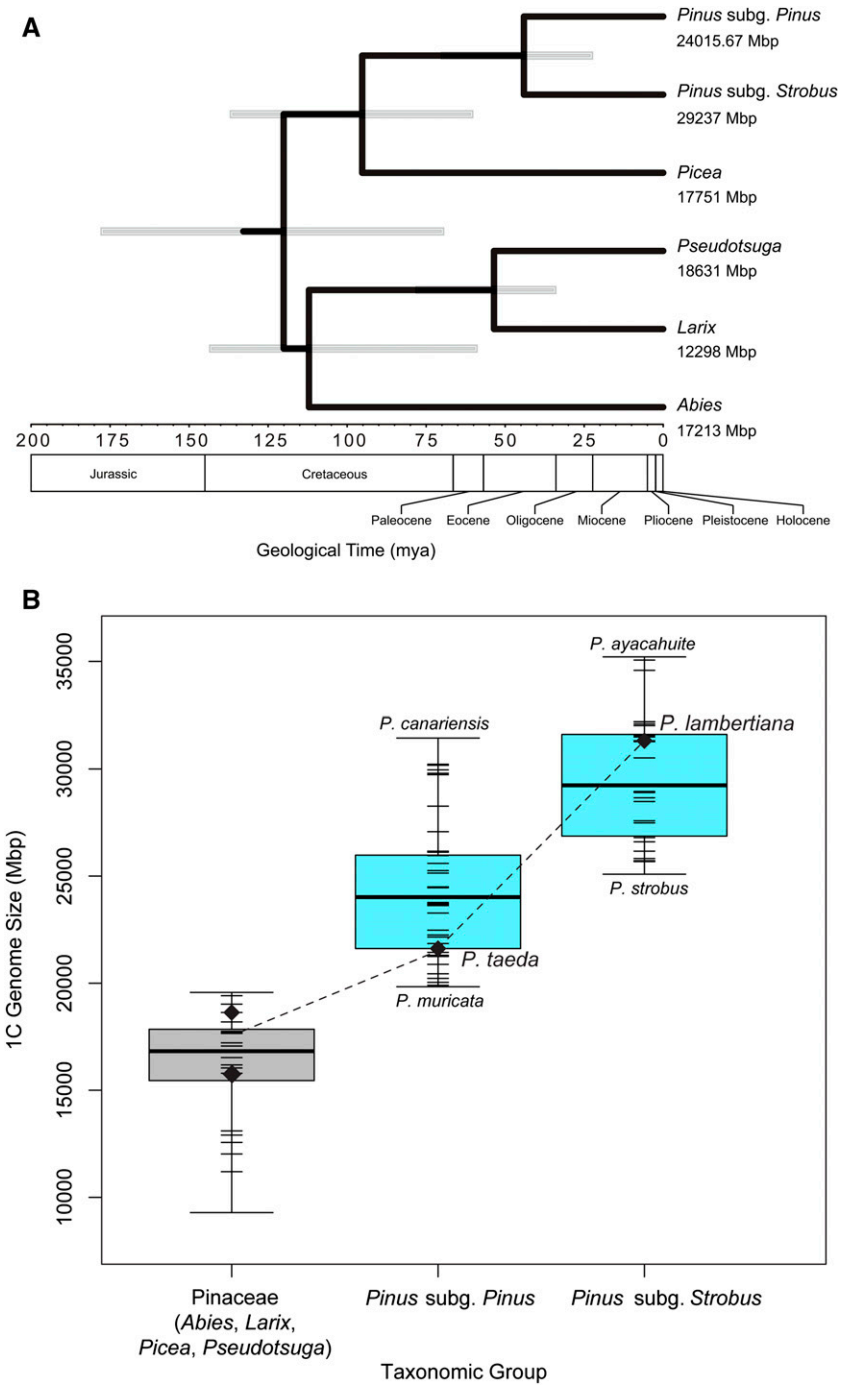

Figure 1 (A) The phylogeny of major genera within the Pinaceae along with genome size estimates. *P. lambertiana* falls in the *Strobus* subgenus. Inference was conducted using Bayesian analysis as implemented in BEAST ver. 2.2.0 (Bouckaert *et al.* 2014). Gray bars represent the 95% highest posterior density range for the age of the node. Data used for inference were 28 independent nuclear gene regions (see Eckert *et al.* 2013a,b), sequenced and assembled for representative taxa selected within each taxonomic group [*Pinus* subg. *Pinus*: *P. taeda*; *Pinus* subg. *Strobus*: *P. lambertiana*; *Picea*: *P. abies*; *Pseudotsuga*: *P. menziesii* (Mirb.) Franco; *Larix*: *L. decidua* Mill.; *Abies*: *A. alba* Mill.]. Details are presented in the Supplementary Methods in File S1 (B) Illustration of the genome size trends of major genera within Pinaceae. Genome sizes are from the *c*-values database (Bennett and Leitch 2012). Diamonds mark the estimates of genomes with a reference sequence. Point estimates in each category are shown as short horizontal lines. Species from other genera within the Pinaceae are shown in gray.

assembled *P. lambertiana* genome to identify large genomic scaffolds tightly linked to *Cr1* and SNPs in strong association with $Cr1^R$. We discuss possible *Cr1* candidates among the annotated genes.

Sequencing and assembly

The sequencing and assembly approach used here for *P. lambertiana* is an adaptation of the approach successfully used for *P. taeda* (Neale *et al.* 2014; Zimin *et al.* 2014). We have found that the haploid DNA obtainable from a single megagametophyte from the target genotype is sufficient to form the basis of a high quality whole genome shotgun assembly. For additional contiguity, haploid megagametophyte coverage is supplemented with longer linking mate pair libraries using DNA isolated from abundantly available diploid needle tissue of the maternal parent. For additional contiguity of the gene space, we performed transcriptome-based scaffolding using deep coverage RNA-Seq data. The nearly 50% larger size of the *P. lambertiana* genome required changes to the previous software methods to make assembly tractable. The resulting draft genome sequence described here has an N50 scaffold size of 246.6 kbp and a total estimated genome size

of 31 Gbp, making it the largest genome sequenced and assembled to date.

Materials and Methods

Plant material

Our target tree for reference genome sequencing was *P. lambertiana* genotype 5038 in the collection of the United Stated Department of Agriculture (USDA) Forest Service, which is in the public domain. Haploid megagametophyte tissue was sourced from wind-pollinated seeds from grafted ramets collected in 1994 from the USDA Forest Service Badger Hill site. *P. lambertiana* needle tissue was collected in August 2011 from a ramet at the same location.

DNA isolation

As described by Zimin *et al.* (2014), our sole source of haploid DNA was a single megagametophyte. Prior to DNA extraction, seeds were immersed in water for 4 days, after which individual haploid megagametophytes were dissected from each seed. DNA was subsequently extracted from dissected megagametophytes as described in Zimin *et al.* (2014). For diploid DNA, large-scale extractions were prepared from *P. lambertiana* needles. For long insert mate pair libraries, nuclei were isolated and DNA was extracted and quantified at University of California (UC) Davis using the methods previously reported (Zimin *et al.* 2014). The resulting DNA was treated with 0.33 μl PreCR Repair Mix (New England Biolabs) per microgram DNA prior to use in library construction. DNA for the *P. lambertiana* fosmid pools was isolated and quantified at CHORI using the slightly modified method previously reported (Zimin *et al.* 2014). Further details can be found in the Supplemental Material, File S1 (Supplementary Methods).

Error correction

Paired end reads were error corrected using QuorUM (Marçais *et al.* 2013), as packaged in the MaSuRCA 2.3.0 assembly pipeline (Zimin *et al.* 2013). Only *k*-mers from the haploid sequences were used in constructing the error correction database. Detailed error correction results are given in Tables S1–S3 in File S1.

Super-read construction

Error-corrected data were used to construct super-reads (Zimin *et al.* 2013), which are longer, nonredundant, and overall much more compact than the original read data. For the *P. lambertiana* paired end sequence data, the super-reads procedure reduced the 6.36 billion error-corrected read pairs to 148 million super reads (Figure 2). The average length of the super-reads was 502 bp with a total length of 75 Gbp. By comparison, the average super-read length for *P. taeda* was 362 bp (Zimin *et al.* 2014).

Mate pair cleaning and filtering

Mate pairs from diploid libraries were cleaned and filtered as follows. (1) Mate pair sequence were error corrected by QuorUM, using a k-mer database from the haploid data. This step had the secondary effect of enriching for our target haplotype. (2) Nonjunction fragments, "short innies," were detected and removed using a procedure that attempted to connect pairs by *k*-mer extensions (again using *k*-mers from the haploid data) off the "wrong" ends. (3) Reads <100 bp were extended via unique *k*-mers to a length of 64–100 bp. If both reads in a pair could not be extended to at least 64 bp, the pair was discarded.

Initial assembly

The preprocessed reads from both the haploid and diploid libraries were then assembled with SOAPdenovo2 (Luo *et al.* 2012) using a *k*-mer size of 99. Paired end libraries (Table S2 in File S1) were divided into three progressively less reliable fragment sizes: <200, 200–400, and >400 bp. Mate-pair libraries (Table S4 in File S1) were divided into two groups: <10 and >10 kbp.

Gap closing

To increase contiguity, gap closing was performed on the output of the SOAPdenovo assembler using the MaSuRCA gap closer, plus the super-read sequences to "patch" gaps in the SOAP assembly.

Transcriptome scaffolding

Additional scaffolding steps used a set of transcript sequences assembled from Pacific Biosciences (PacBio) and Illumina RNA-seq data (Table S7) from Gonzalez-Ibeas *et al.* (2016). We aligned transcript sequences to the whole genome shotgun (WGS) scaffolds using both *nucmer* (-maxmatch -nosimplify -l 45 -c 4) (Kurtz *et al.* 2004) and *bwa-mem* (-k 45 -O 60 -E 10) (Li 2013). We then merged alignments that were adjacent on both the transcript and the corresponding scaffold. For pairs of scaffolds that were aligned adjacently to the same transcript, we subsequently created a link. We sorted the links in descending order according to intron size. Next, we built a graph by visiting links in order. Each link corresponds to a potential edge in the graph between vertices corresponding to scaffolds. We added a link/edge to the graph if it did not create a cycle or a vertex degree >2. Upon completion, the graph consisted only of paths, which we converted to superscaffolds that contained one or more of the original assembly scaffolds.

Transcriptome assembly

Thirty-one tissue-specific samples, including needle, root, stem, pollen, cone, strobili, and embryonic tissues, were used for the construction of cDNA libraries. A variety of treatments were applied to seedlings before RNA extraction, including: cold/heat shock, flood/drought stress, wounding, and salicylic or methyl jasmonate exposure. RNA sequencing was done using Illumina MiSeq and HiSeq to generate short (100–300 bp) reads, and PacBio Iso-Seq reads, which range from 1000 to over 6000 bp. Seven Miseq libraries, nine Hiseq, and 18 PacBio libraries were created, and a total of 40 SMRT cells (1–4 SMRT cells per library) was sequenced.

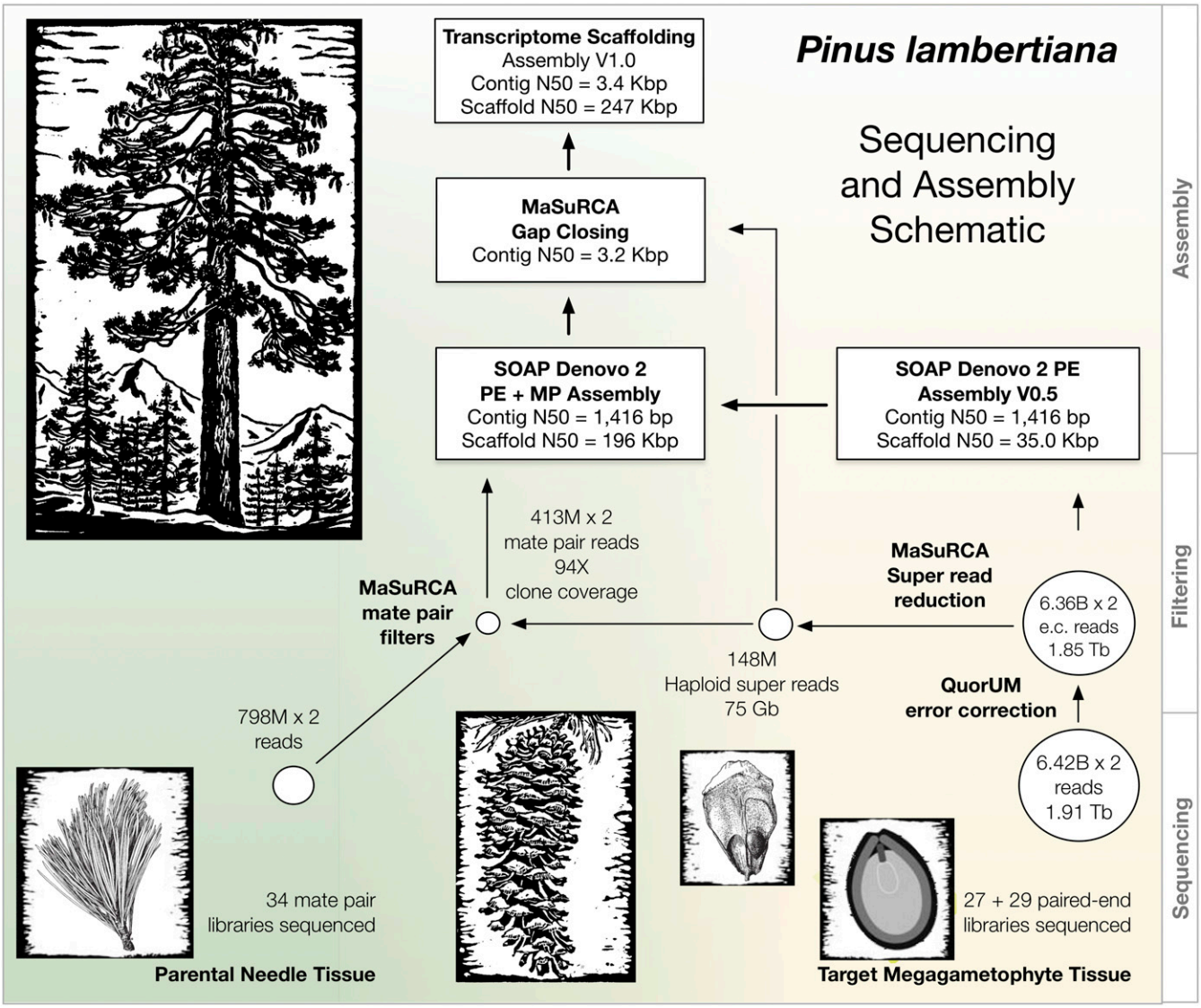

Figure 2 Overview of sequencing and assembly strategy for *P. lambertiana*. Woodcut images used with permission from "The trees of Yosemite; a popular account." Library of Congress call number QK484.C2 T7 1932.

Quality trimmed reads were used for assembly with Trinity (Haas *et al.* 2013), and protein coding sequences (CDS) were identified with Transdecoder (Haas *et al.* 2013). All CDSs were clustered at 95% sequence identity with Uclust (Edgar 2010) (usearch v8.1.1861) to generate a nonredundant set of transcripts.

Identification of genomic scaffolds and mapping in the Cr1 region

Jermstad *et al.* (2011) reported the sequences from cloned RAPD bands OP_G16 and BC_432 that were linked to *Cr1*. To identify these genomic loci, the representative consensus sequence for each RAPD band was aligned to the *P. lambertiana* genome assembly using *gmap* (Wu and Watanabe 2005). In both cases, a unique top hit (path1) was observed identifying target scaffolds, which we used to develop new markers.

Target scaffolds were masked for annotated simple and interspersed repeats (see Supplementary Methods in File S1). We designed pairs of nested PCR primers using PRIMER3 (Rosen and Skaletsky 1999) for unique regions in these two target scaffolds. All of the PCR assays used standard PCR reaction conditions: 2.0 mM $MgCl_2$, 0.2 mM each of dNTPs, 0.5 mM each of forward and reverse primers, 1 U of *Taq* polymerase, and 50 ng of DNA. For validation purposes, we used the available primer sequences of PCR amplicon, UMN_3258_01 (http://treegenesdb.org/ftp/CRSP/) to develop a new marker, *cr1lC*.

Gene annotation

Annotations were generated using the automated genome annotation pipeline MAKER-P (Campbell *et al.* 2014). Inputs and training sets for MAKER-P included the *P. lambertiana* genome assembly, a *P. lambertiana* transcriptome assembly

(see Supplementary Methods in File S1), ESTs from spruce and pine (1,027,297 downloaded from GenBank), protein sequence data from *Vitis vinifera* L. (25,665), *Amborella trichopoda* Baill. (25,354), *Populus trichocarpa* Torr. and A.Gray ex Hook (38,655), *Picea abies* (22,721), *Picea sitchensis* (Bong.) Carrière (17,841), *Pinus taeda* (34,059), and RNA-seq data from *P. lambertiana*. Default MAKER-P mapping parameters were used (80% coverage and 85% identity threshold for EST-genome alignments, and 50% coverage and 40% identity for protein-genome alignments). More details can be found in the Supplementary Methods in File S1.

Interspersed repeat annotation

To find interspersed repeat elements, we used both similarity and *de novo* based approaches (Figure S3 in File S1). RepeatModeler combines two complementary *de novo* repeat element prediction algorithms: RECON (Bao *et al.* 2002) and RepeatScout (Price *et al.* 2005). To make the RepeatModeler computation tractable, we used only the Illumina sequenced fosmid pools (above), along with the longest 2.5% of genomic scaffolds. We also used a combination of TEclass (Abrusán *et al.* 2009), CENSOR (Kohany *et al.* 2006), and manual characterization to identify the uncharacterized elements from the repeat library produced by RepeatModeler. We used this library, along with the plant Repbase library (Jurka *et al.* 2005) (plant component only, v19.01) as the reference database for RepeatMasker (Tarailo-Graovac *et al.* 2009). Full-length elements were determined by applying a cut-off of 80-80-80 (80% sequence similarity, and 80 bp minimum length) (Wicker *et al.* 2007).

Data availability

The *P. lambertiana* assembly and annotation are available from GenBank as accession GCA_001447015.2 and BioProject 174450, and also from http://www.pinegenome.org/pinerefseq. Genomic DNA and RNA reads are also available under BioProject 174450.

Results

Sequencing

Our sequencing strategy for conifer genomes has taken advantage of the haploid tissue of the conifer megagametophyte (Neale *et al.* 2014; Zimin *et al.* 2014). Fortunately the observed correlation over the evolutionary diversity of gross seed weight with genome size (Wakamiya *et al.* 1993) (Grotkopp *et al.* 2004) in the genus *Pinus* worked to our advantage. Our collection of *P. lambertiana* megagametophytes had an average weight of 225 mg compared to only 23.5 mg for *P. taeda*. This translated into substantially larger yields of haploid genomic DNA from single seeds. From our target *P. lambertiana* megagametophyte, we were able to obtain 36.2 mg of DNA, from which we generated 1.91 trillion base pairs of sequence (Figure 2 and Table 1), representing ~62× coverage of the 31 Gbp haploid genome.

Table 1 Characteristics of the *P. lambertiana* sequence data and 1.0 assembly, compared to known cytometric and cytological properties

Cytometric Genome Size	31 Gbp
Chromosome number	12
Assembly V1.0	
Total size	
Scaffolds ≥ 200 bp	4,259,911 scaffolds 27.6 Gbp including gaps 25.5 Gbp without gaps
Scaffolds ≥ 500 bp	1,089,992 scaffolds 26.9 Gbp including gaps 24.7 Gbp without gaps 54,147,744 contigs
Contigs < 200 bp ("chaff")	6.5 Gbp
N50 scaffold size (31 Gb)	246.6 kbp
N50 contig size (31 Gb)	4.25 kbp
Sequence data	
Number of paired-end libraries	56
Paired end sequencing depth	1,910 Gbp (61.5×)
By platform	
Hiseq 2000 (125 bp + 125 bp)	2.8×10^{11} bp (9.0×)
Hiseq 2500 (150 bp + 150 bp)	1.4×10^{12} bp (45.1×)
GAIIx (160 bp + 156 bp)	1.8×10^{11} bp (5.8×)
MiSeq (255 bp + 255 bp)	4.7×10^{10} bp (1.5×)
By fragment size	
[200 bp, 400 bp]	9.6×10^{11} bp (31.0×)
[400 bp, 600 bp]	4.6×10^{11} bp (15.0×)
[600 bp, 900 bp]	4.8×10^{11} bp (15.6×)
Long fragment libraries (1.5–25 kbp)	34
Long fragment coverage	
Illumina Trueseq	22.5× physical coverage
Nextera mate pair	71.2× physical coverage

N50 statistics were calculated using an estimated genome size of 31 Gbp. Paired end sequencing depth represents the raw output prior to error correction. Physical coverage estimated by MaSuRCA (including the inferred DNA fragement) is reported here for all libraries by chemistry (see Supplementary Methods in File S1).

Estimating genome size

We analyzed the *k*-mer distribution of the paired reads to derive an independent estimate of the haploid size of the genome for coverage estimates. Using the jellyfish program (Marçais and Kingsford 2011), we computed distributions of *k*-mer depth for $k = 24$ and $k = 36$ for all the paired sequences derived from our megagametophyte. We estimated genome size from the *k*-mer distribution as described previously (Zimin *et al.* 2014), using both the mean and the mode of the distributions for $k = 24$ and $k = 31$. As shown in Table 2, all four estimates of the genome size are in close agreement, ranging from 30.9 to 31.9 Gbp.

Our haploid library based estimates were in the range of previous experimental estimates in the literature. The Gymnosperm DNA *C*-values Database release 6.0 (Bennett and Leitch 2001) contains three flow cytometry-based estimates for the genome size of *P. lambertiana*: 33.4 Gbp (Grotkopp *et al.* 2004); 31.1 Gbp (Williams *et al.* 2002); and 29.4 Gbp (Wakamiya *et al.* 1993). The authors of the 33.4 Gbp estimate noted that their genome size estimates of various species were consistently higher than values already in the literature. The mean of these experimental estimates, 31 Gbp, is in close

Table 2 Estimates of the genome size of *P. lambertiana* based on the distribution of *k*-mers in the paired read data

	k = 24	*k* = 31
Total *k*-mers	1.56×10^{12}	1.47×10^{12}
Erroneous *k*-mers	1.20×10^{10}	2.20×10^{10}
Total correct *k*-mers	1.55×10^{12}	1.45×10^{12}
E(unique *k*-mer depth) mode	49.72	46.77
Estimated genome size	31.1 Gbp	30.9 Gbp
E(unique *k*-mer depth) mean	48.53	46.02
Estimated genome size	31.9 Gbp	31.4 Gbp

Erroneous *k*-mers refer to *k*-mers that were identified as likely to contain errors, and these were removed from the calculation.

Table 3 Assemblies of the four fosmid pools sequenced with PacBio technology

Pool	Contigs	Minimum	Mean	Maximum	Length	N50
SPPB1	61	979	31983	45177	1950994	34685
SPPB2	54	586	33949	44946	1833274	35595
SPPB3	58	586	29525	43039	1712462	35375
SPPB4	73	551	27960	43934	2041131	35324

Each pool contained 48 fosmids.

agreement with our sequenced-based estimates, and therefore we chose this value as the estimated total size of the genome.

Assembly

Super-reads (Zimin *et al.* 2013) played a fundamental role in the assembly of *P. lambertiana*, where they allowed us to dramatically reduce the size of the input to subsequent assembly steps (Figure 2). Nevertheless the CABOG assembler (Miller *et al.* 2008) used for the 22 Gbp genome of *P. taeda* could not process the larger *P. lambertiana* genome, so we instead used the *de Bruijn* graph-based SOAPdenovo 2 assembler (Luo *et al.* 2012) for initial contig and scaffold construction. Following this step, we reassembled the contigs, with SOAPdenovo 2, adding the 93× coverage from long-fragment libraries, yielding scaffolds with an N50 size of 196 kbp. We then ran a separate gap-closing procedure to reduce the number of intrascaffold gaps, which closed 12.6 million out of 26.2 million gaps in the assembly. This reduced the total gap length by ∼780 Mbp, and increased the N50 contig size to 3.4 kbp.

Finally, we used transcript sequences to improve contiguity in the vicinity of genes. We aligned a set of 17,167 assembled transcripts (see *Materials and Methods*) to the scaffolds. We joined scaffolds together if the links created were consistent with a colinear transcript alignment. In total, 32,619 scaffolds were merged during this step. The resulting assembly (version 1.0) has an N50 scaffold length of 246.6 kbp. The combined length of the assembly, including all scaffolds and contigs >200 bp, is 27.6 Gbp (Table 1). The assembly contains another 6.48 Gbp in contigs, and scaffolds ≤200 bp that were not considered for most analyses.

Validation

As an independent assessment of assembly quality, we sequenced four pools of 48 fosmids each using the PacBio RS II platform (see Supplementary Methods in File S1). We collected deep coverage (>250×) of each pool. The vector-trimmed HGAP3-assembled pools are reported in Table 3. Most of the assembled contigs appeared to span the full length of a fosmid, ∼40 kbp (Table 3, and Table S6 in File S1). Overall, the PacBio fosmid assemblies were 98.8% identical to the WGS assembly, which covered >95% of their total length. Because the haploid fosmids were constructed from diploid needle tissue, at most half were expected to match exactly. Thus, the 1.2% divergence represents an upper bound on alignment and assembly errors, or, alternatively, half the heterozygosity rate.

As a measure of the correctness of the WGS assembly, we looked for large insertions, deletions, or rearrangements between the PacBio and WGS assemblies. The comparison yielded only one noncolinear alignment, and one WGS scaffold with a large 7.6 kbp deletion, for which we could not rule out haplotype differences. A second scaffold with a 5.3 kbp deletion was clearly a heterozygous insertion of an LTR element in the assembled fosmid.

For further validation, we examined the alignment of the WGS scaffolds just prior to transcriptome scaffolding to our collection of 12,533 PacBio and 4634 Illumina assembled transcripts; >99% of these alignments were consistent. When examining the 1% that were not colinear, we found that these were dominated by Illumina-based transcripts, leading to the conclusion the most of these represented errors in the transcript assembly rather than the WGS assembly.

Gene content

Annotation yielded 13,936 high-quality gene models and 71,117 low-quality models, the presence of direct RNA evidence being the primary distinction between the two classes (Supplementary Methods in File S1). A total of 11,769 scaffolds were annotated with at least one high-quality gene model, ranging from one to eight models per scaffold (1.2 models/scaffold on average). Only 33 scaffolds were annotated with five or more models. Completeness of the gene space evaluated with BUSCO (Simão *et al.* 2015) was 53% when using the high-quality models, and 58% when the low-quality models were included. Alternatively, DOGMA (Dohmen *et al.* 2016) estimated a coverage of 94% for their Conserved Domain Arrangements, For comparison, when run on the complete set of *P. taeda* gene models, BUSCO estimated 50% completeness and DOGMA estimated 61% (Table S8 in File S1)

In total, 11,595 of the 13,936 gene models were functionally annotated with a characterized plant protein sequence. A total of 2041 were classified as uninformative (protein alignment with no functional assignment), and 300 showed no homology to characterized proteins. As expected, *Vitis vinifera*, *Arabidopsis thaliana* (L.) Heynh., and *Ricinus communis* L. were the species that contributed the most to the functional annotations. The largest *P. lambertiana* intron, at 578 kbp, is the second largest (after one in *P. taeda*) found in

Table 4 Comparison of gene metrics among sequenced conifer genomes and select angiosperms

	Pinus taeda	*Picea abies*	*Pinus lambertiana*	*Picea glauca*	*Arabidopsis thaliana*	*Populus trichocarpa*	*Vitis vinifera*	*Amborella trichopoda*
Genome size (Mbp)	20,148	19,600	31,000	20,000	135	423	487	706
Chromosomes	12	12	12	12	5	19	19	13
Gc+cC content (%)	38.2	37.9	35.1	31.1	35	33.3	36.2	35.5
TE content (%)	74	70	79	N/A	15.3	42	41.4	N/A
Number of genes	9,024	26,359[a]	13,936	14,462	27,160	36,393	25,663	25,347
Average CDS length (bp)	1,562	931	1,330	1,421	1,102	1,143	1,095	969
Average intron length (bp)	12,875	1,020	8,039	603	182	366	933	1,538
Maximum intron length (bps)	8,91,919	68,269	5,78,081	1,19,319	10,234	4,698	38,166	1,75,748

[a] High confidence genes from the Congenie project.

a plant genome to date (Table 4), although the draft state of the genome means that larger introns are highly likely to be scattered among multiple scaffolds.

Transposable elements

TE sequences constitute 79% of the *P. lambertiana* genome, higher than the 74% found in *P. taeda* (see Supplementary Methods in File S1). Of these, 67% of the transposable sequences in *P. lambertiana* are LTR retrotransposons. The distribution of transposable element families is very similar in the two species (see Figure 3). The most substantial difference in repeat content observed between the genomes is a 35% greater proportion of Gypsy elements in *P. lambertiana*. The distributions of estimated insertion times among LTR retrotransposons are congruent with those reported for spruce in Nystedt *et al.* (2013) (Figure S5). The median LTR insertion time for *P. lambertiana* (16.0 MYA) is younger than that of *P. taeda* (17.4 MYA). As a class, *P. lambertiana* Gypsy elements are significantly younger (14.5 MYA; $P < 1.5 \times 10^{-12}$), consistent with their increased numbers and a lineage-specific expansion. These observations are consistent with the hypothesis that TEs make up the bulk of the enlarged genomes of subgenus *Strobus*, with much of the expansion in *P. lambertiana* attributable to Gypsy.

The similarity in TEs among the sequenced conifer genomes supports the hypothesis that conifers have experienced massive expansion of TEs throughout their history (Neale *et al.* 2014), likely including the period prior to the radiation of *Pinus*, yielding their large and varied sizes. The bulk of TE sequences are ancient and diverged. Consistent with this, we observed that partial elements are far more abundant than full-length sequences in *P. lambertiana*, representing 67.3% of the genome, and 87% of the total repetitive content. And while the vast majority of LTRs were ancient and inactive, we did find evidence of recent transposition in the form of a recently inserted heterozygous TE. We observed a complete heterozygous insertion of a PARTC element in a genomic segment captured in an assembled fosmid clone. Heterozygosity is inferred from the insertion of the element, and the presence of a target site duplication in the alignment to the alternate haplotype (Figure S6). Previous analysis of the many copies of the PARTC subfamily suggested that it was dead (Zuccolo *et al.* 2015). However, this copy has identical LTR sequences, and apparently functional proteins.

Identification of genomic scaffolds and mapping in the Cr1 region

Using the draft WGS assembly, we succeeded in anchoring the cloned RAPD sequences, *scar*OPG16_950 and *scar*BC432_1110, which had previously been mapped near the *Cr1* locus (Jermstad *et al.* 2011), to two distinct scaffolds (Table S13). No longer limited to designing PCR primers within those cloned sequences, we utilized the entire repeat masked scaffolds as a resource, and were able to identify many clear single nucleotide polymorphisms (SNPs) in each flanking amplicon, including that adjacent to scarBC432_1110, which had previously yielded no scorable SNPs (Jermstad *et al.* 2011).

PCR primers were designed to amplify two small genomic loci, one in scaffold 223,058 and the other in 370,413 (Table S10). The amplicons of successful primer pairs were sequenced and tested for segregation in a small sample of both $Cr1^R$ and $Cr1^r$ segregant megagametophytes from maternal tree 5701 ($Cr1^R/Cr1^r$), for which the rescued embryos were genotyped for *Cr1*. Note, the pollen parent of the rescued embryos was assumed to be $Cr1^r/Cr1^r$ because the frequency of $Cr1^R$ is assumed to be rare. Alternative haplotype sequences were found for both amplicons that segregated (see Supplementary Methods in File S1 for the Fasta sequence), and appeared to be linked to one another and to the *Cr1* locus.

A large sample of megagametophytes was efficiently genotyped using Cleaved Amplified Polymorphic Sequence (CAPS) assays (Konieczny and Ausubel 1993). We developed two new CAPS markers, *cr1lA* and *cr1lB*, based on the sequence variation in these two amplicons physically linked via the assembly to the previously reported RAPD markers (see Table S13 in File S1). Genotyping of *cr1lA* and *cr1lB* on a sample of 245 megagametophytes from maternal tree 5701 yielded two apparent single crossovers between markers *cr1lA* and *cr1lB* (both $cr1lA^R$ - $cr1lB^r$), and 225 non-recombinants (Table S12 in File S1). We were not able to confirm the Harkins *et al.* (1998) gene order BC_432_1110 – *Cr1* – OPG_16_950. For the RAPD markers BC_432_1110 and OPG_16_950, Harkins *et al.* (1998) reported recombination fractions of 3%, and genetic map distances of 1.2 cM between both markers and *Cr1* for maternal tree 5701. For our data, Harkins *et al.* (1998) gene order results in

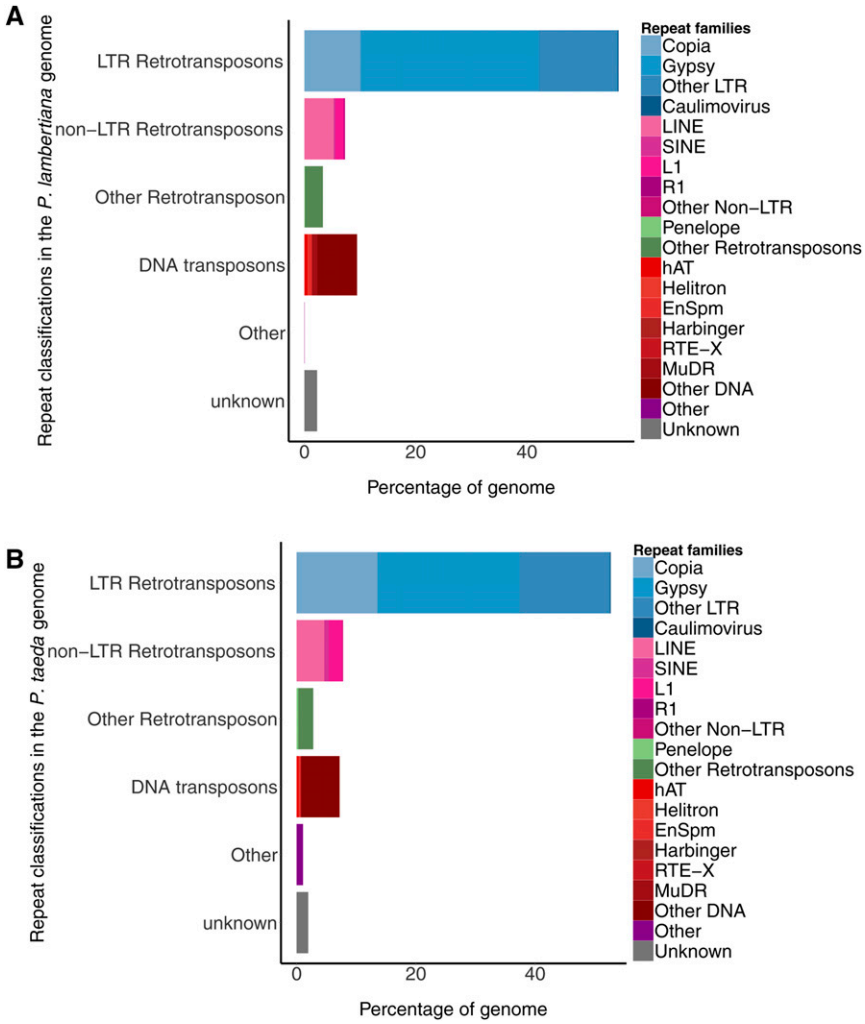

Figure 3 Comparison of repetitive content between transposable element repeat families in *P. lambertiana* (top) and *P. taeda* (bottom).

12 putative double recombinants, which can alternatively be interpreted as *Cr1* genotyping error.

Our two crossovers between *cr1lA* and *cr1lB* indicate that *Cr1* is closer to *cr1lB*. To validate this result, and place it in a slightly broader mapping context, we added a new marker *cr1lC* to the genetic map, derived from a previously characterized PCR amplicon (Jermstad *et al.* 2011; UMN_3258_01; http://treegenesdb.org/ftp/CRSP/) Genotyping with SNPs derived from this amplicon placed it closest to scar-OPG_16_959 on the side away from *Cr1* in the Jermstad *et al.* (2011) map for maternal tree 5701 at a distance of ~12 cM. We genotyped *cr1lC* in the two *cr1lA* - *cr1lB* recombinant megagametophytes from 5701, in three randomly selected $Cr1^r$ nonrecombinants, and four randomly selected $Cr1^R$ nonrecombinants to further refine marker order. Two distinct *cr1lC* haplotypes were determined among these progeny. None were recombinant between *cr1lA* and *cr1lC*, thus placing *Cr1* outside of these loci (Figure 4, left), consistent with the gene order (*cr1lC* – *cr1lA*) – (*Cr1* – *cr1lB*).

Increasing the Cr1 genomic region

To expand our annotated intervals linked to *Cr1*, we walked outward from the two marker-anchored scaffolds using physical linkage inferred from one or more aligned fosmid DiTag reads not included in the assembly. Using this approach, an additional gene-containing scaffold was physically linked to one of our anchored scaffolds by two fosmid DiTags (Figure 4).

The genome assembly allowed a more targeted identification of potential gene candidates for *Cr1*. Figure 4 shows a total of 14 gene annotations on the two scaffolds genetically linked to *Cr1*, and a third scaffold that was physically linked by fosmid DiTags. Of the 14 linked genes, PILA_lg017786 stands out as a candidate because it contains both the NB-ARC and LRR domains that are common elements of disease-resistance genes. We looked for direct evidence of expression in transcriptome assemblies and found only one transcript (TR43508|c1_g1_i2|m.82078; see Supplementary Methods in File S1) assembled from a library constructed from a WPBR resistant tree. The transcript overlaps two exons of the candidate gene (red bar above the gene in Figure 4). The most similar known gene is in *P. monticola* (Western white pine), a TIR-NBS-LRR protein (GI:321530320). The closest well-annotated gene appears to be the disease resistance protein RGA2 in the grass *Aegilops tauschii* Coss. (GI:475615320).

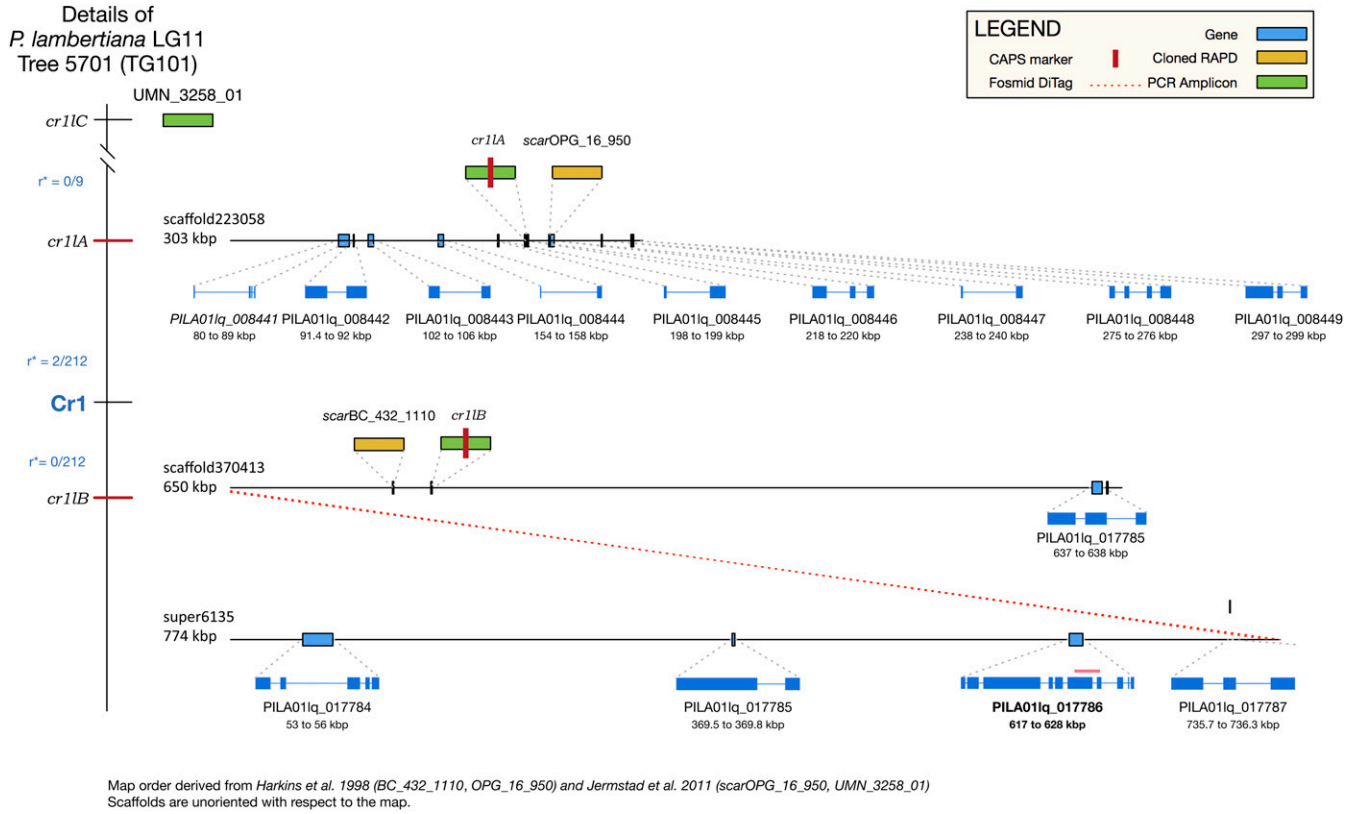

Figure 4 Annotated scaffolds and elements linked to *Cr1*. On the left is a tentative map of the *Cr1* region of chromosome 11 showing the positions of identified markers. The gene order shown was derived from Harkins *et al.* (1998) (BC_432_1110 labeled *cr1lB*, *Cr1*, OPG_16_950 labeled *cr1lA*) and Jermstad *et al.* (2011) (*Cr1*, scarOPG_16_950, UMN_3258 labeled *cr1lC*). To the right are five scaffolds and 14 gene annotations that are linked to the *Cr1* gene. The evidence of expression of PILA_lg017786 was a single transcript (red bar) assembled from a library constructed from a resistant tree (Supplementary Methods in File S1). Scaffold super6135 is physically linked to scaffold 370413 that harbors *cr1lB* by two fosmid DiTags.

Cr1 association

To both confirm the tight linkage of *cr1lB* to *Cr1* and to provide a potential resource for marker-assisted selection, small, representative samples of *Cr1*-genotyped trees were genotyped by sequencing the amplicon from a single megametophyte. A total of six $Cr1^R/Cr1^R$, 12 $Cr1^R/Cr1^r$, and 22 $Cr1^r/Cr1^r$ genotyped sugar pine seed trees from the center of the species' range were assayed (Table S15 in File S1). Genotyping of the diploid parent for *Cr1* was done by the Forest Service at the El Dorado National Forest, Placerville Nursery, using their standard protocol of germinating, and scoring at least 56 exposed seed trees for WPBR resistance. These $Cr1^R/Cr1^R$ and $Cr1^R/Cr1^r$ trees were previously reported in Vangestel *et al.* (2016).

We selected one megagametophyte each from a maternal parent that had been genotyped for resistance. The *cr1lB* primers appeared to work well outside of their original context (maternal genotype 5701) and the haploid nature of the DNA afforded additional confirmation of the sequencing results. Evaluating only sequences associated with known *Cr1* alleles (*i.e.*, transmitted from $Cr1^r/Cr1^r$, $Cr1^R/Cr1^R$, or phenotyped progeny of 5701) we identified a five-site motif that predicted the *Cr1* allele nearly completely (see Figure 5). All seven of our $Cr1^R$ associated haplotypes (six transmitted from $Cr1^R/Cr1^R$ and the $Cr1^R$ linked haplotype of 5701) had the motif "TTACT." Furthermore, 23 out of 24 of our $Cr1^r/Cr1^r$ transmitted haplotypes had the alternate $Cr1^r$ linked motif "GCGGC." The association is almost complete; the differences in the frequencies of the two haplotypes transmitted with known *Cr1* genotypes is statistically significant, $P < 10^{-5}$ by χ^2 with 2 d.f. Both motifs segregated in the progeny of one $Cr1^r/Cr1^r$ parent. The observation of this single heterozygous tree is consistent with a low frequency of "recombinant" haplotypes. Still the association of *Cr1* with SNPs in the *cr1lB* amplicon on scaffold 370,314 is strong.

Discussion

A key step in the sequencing strategy for *P. lambertiana* was the generation of deep sequencing coverage of the haploid genome. Even so, the unprecedented amount of data, two trillion bases, required an alternative strategy in order to assemble the genome in a reasonable time frame. The contiguity of the *P. lambertiana* assembly, as measured by the N50 scaffold size, is higher than previous conifer genome assemblies (Birol *et al.* 2013; Nystedt *et al.* 2013; Neale *et al.* 2014;

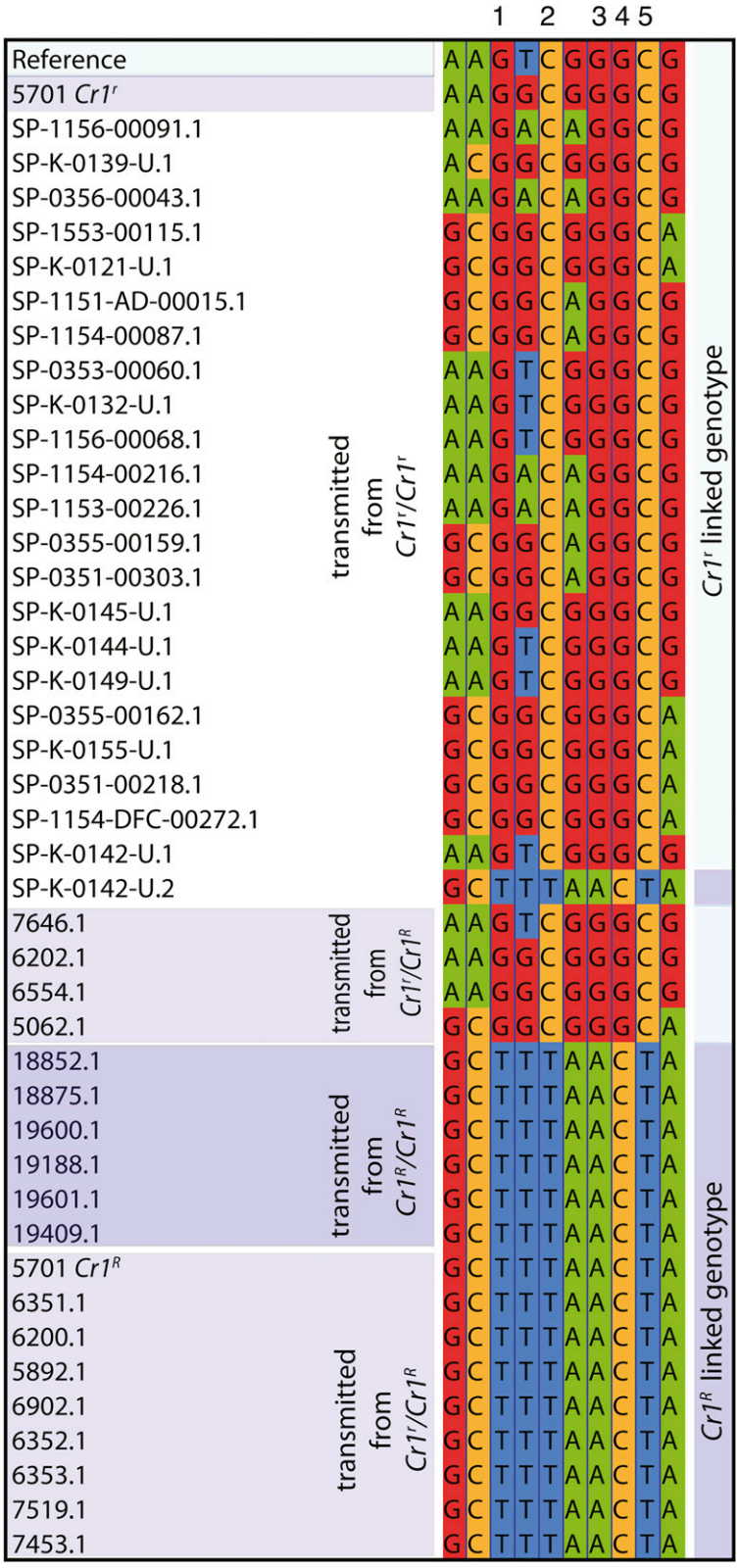

Figure 5 Multiple alignment of association samples showing the most variable sites, 40% or more consensus differences. The numbered five site *Cr1* linked motif is seen as two haplotypes, the *Cr1*^{r} linked GCGGC and the *Cr1*^{R} linked TTACT. One haplotype (SP-K-0142-U.2) transmitted from a *Cr1*^{r}/*Cr1*^{r} parent genotyped as a putative *Cr1*^{R} linked "TTACT" recombinant.

Warren *et al.* 2015). A combination of factors, including deeper sequence coverage, more physical coverage from new linking mate pair library chemistries, and better computational methods, all likely contributed to the advance. Like other conifers, a critical biological aspect of the *P. lambertiana* genome that allows it to be assembled, is the accumulated divergence among the ancient repeats comprising the majority of the genome. This increased contiguity of the *P. lambertiana* assembly

clearly suggests that the contiguity of conifer genome assemblies will continue to increase as scalable, long-range linking methods become available.

The characterization of *P. lambertiana* transposable element sequences supports the hypothesis advanced by Nystedt *et al.* (2013) that an ancient accretion of mostly inactive TEs at a rate faster than they are removed, explains the majority of the increased genome size observed in the *Pinus* subgenus *Strobus*. Given the huge genome sizes, the time scale involved, and the still sparse sampling of genome sequences of conifer species, recent TE dynamics (if such exist) are difficult to detect. Nevertheless, we made two observations relevant to the hypothesis. First, sequences of gypsy families are more abundant in the *P. lambertiana* genome lineage, and this likely contributed to the increase in genome size. This hypothesis is supported by Gypsy families having increased fractions of repeat sequences with younger age. Second, we detected what appears to be an actively transposing *Part-C* element, based on its fully intact coding genome, and its heterozygous insertion state. These observations are consistent with the simplest hypothesis that the many transposon families remain an active but small cohort, and that their sequences accumulate over millions of years because their replicative transposition rate exceeds their removal rate. So far, there is no evidence for any very recent huge expansion of specific families. We did detect the signature of recent duplication in the *P. lambertiana* genome in the *k*-mer distribution, perhaps evidence of nonhomologous crossover. However such duplications were not abundant enough to explain the difference in genome sizes. While ancient genome duplication (Li *et al.* 2015) may also have played a role, the hypothesized event predates the radiation of *Pinus*.

The immense size and repetitive nature of the conifer genome, especially that of *P. lambertiana*, has been, and remains, a daunting barrier to genetic analyses, especially the investigation of pathogen resistance. And this challenge, compounded with those inherent to the long generation time, as well as resource requirements, have translated into strenuous efforts to achieve modest advances in understanding and impacts on the genetics of reforestation. This reference genome brings new powerful tools to genetics/genomic research in *P. lambertiana*. We sought to apply the new reference genome sequence to the characterization of the genetics of resistance to WPBR, building on the rich previous research, and indeed the availability of genomic samples from now classic efforts to genetically map a major disease resistance gene. Also (as discussed above) strong ecological and economic considerations motivate the pursuit of both new knowledge, and effective practical tools that can be applied to forest management (Waring and Goodrich 2012). Large scaffolds in the assembly of *P. lambertiana* bearing short sequences previously linked to *Cr1* (Harkins *et al.* 1998; Jermstad *et al.* 2011) were identified, validated as linked to *Cr1*, and annotated as containing a promising candidate gene. Of substantial immediate practical relevance is the strong association between SNPs anchored in one of these scaffolds and *Cr1* in natural populations. Genotyping with such SNPs is a long-sought-after tool that will increase the efficiency of ongoing and future WPBR-resistant reforestation. The present expensive and time consuming process of identifying candidate trees, collecting seed (during a narrow period), and waiting 2 years for infection bioassay results, does ultimately identify trees heterozygous (or rarely homozygous) for $Cr1^R$ that can then be harvested for seeds to go into reforestation. But the efficiency is low, and the cost to identify a single such tree is thousands of dollars [see the estimated replacement costs in a 2013 supplement to a US Forest Service Handbook (page 5), available at http://www.fs.fed.us/im/directives/field/r5/fsh/2409.18/r5U2409U18U50U2013U1.doc]; furthermore, the supply is not always adequate or ecologically optimal. Ongoing efforts to develop these and other SNPs as practical tools for sugar pine forest management have great promise, and may lead the way to similar tools for other white pines.

Acknowledgments

We thank Carson Holt and Mark Yandell for their modifications to their MAKER-P pipeline to support conifer genomes. Funding for this project was provided through a United States Department of Agriculture/ National Institute of Food and Agriculture (USDA/NIFA) (2011-67009-30030) award to D.B.N. at University of California, Davis.

Note added in proof: See Gonzalez-Ibeas *et al.* 2016 (pp. 3787–3802) in *G3: Genes, Genomes, Genetics* for a related work.

Literature Cited

Abrusán, G., N. Grundmann, L. DeMester, and W. Makalowski, 2009 TEclass—a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics 25(10): 1329–1330.

Ahuja, M. R., and D. B. Neale, 2005 Evolution of genome size in conifers. Silvae Genet. 54(3): 126–137.

American Forests, 2015 This Is It! The Quest for a New Champion Sugar Pine. Available at: http://www.americanforests.org/blog/quest-for-a-new-champion-sugar-pine/.

Bao, Z., and S. R. Eddy, 2002 Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 12(8): 1269–1276.

Bennett, M. D., and I. J. Leitch, 2012 Plant DNA C-values database, release 6.0, Dec. 2012. Available at: http://data.kew.org/cvalues/.

Birol, I., A. Raymond, S. D. Jackman, S. Pleasance, R. Coope *et al.*, 2013 Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics 29: 1492–1497.

Campbell, M. S., M. Law, C. Holt, J. C. Stein, G. D. Moghe *et al.*, 2014 MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 164(2): 513–524.

Critchfield, W. B. and E. L. Little, Jr. 1966. Geographic distribution of the pines of the world. USDA Forest Service Miscellaneous Publication 991. US Department of Agriculture, Washington DC.

DeGiorgio, M., J. Syring, A. J. Eckert, A. Liston, R. Cronn *et al.*, 2014 An empirical evaluation of two-stage species tree inference strategies using a multilocus dataset from North American pines. BMC Evol. Biol. 14(1): 67.

Devey, M. E., A. Delfino-Mix, B. B. Kinloch, and D. B. Neale, 1995 Random amplified polymorphic DNA markers tightly linked to a gene for resistance to white pine blister rust in P. lambertiana. Proc. Natl. Acad. Sci. USA 92(6): 2066–2070.

Dohmen, E., L. P. Kremer, E. Bornberg-Bauer, and C. Kemena, 2016 DOGMA: domain-based transcriptome and proteome quality assessment. Bioinformatics 32: 2577–2581.

Eckert, A. J., J. L. Wegrzyn, J. D. Liechty, J. M. Lee, W. P. Cumbie *et al.*, 2013a The evolutionary genetics of the genes underlying phenotypic associations for loblolly pine (Pinus taeda, Pinaceae). Genetics 195: 1353–1372.

Eckert, A. J., A. D. Bower, K. D. Jermstad, J. L. Wegrzyn, B. J. Knauss *et al.*, 2013b Multilocus analyses reveal little evidence for lineage wide adaptive evolution within major clades of soft pines (Pinus subgenus Strobus). Mol. Ecol. 22: 5635–5650.

Edgar, R. C., 2010 Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19): 2460–2461.

Farjon, A., and D. Filer, 2013 *An Atlas of the World's Conifers*, Brill Publishing, Leiden, The Netherlands.

Fattash, I., R. Rooke, A. Wong, C. Hui, T. Luu *et al.*, 2013 Miniature inverted-repeat transposable elements: discovery, distribution, and activity 1. Genome 56(9): 475–486.

Gernandt, D. S., G. G. López, S. O. García, and A. Liston, 2005 Phylogeny and classification of Pinus. Taxon 54(1): 29–42.

Gonzalez-Ibeas, D., P. J. Martínez-García, R. A. Famula, A. Delfino-Mix, K. A. Stevens *et al.*, 2016 Assessing the gene content of the megagenome: sugar pine (*Pinus lambertiana*) G3 (Bethesda) 6: 3787–3802.

Grotkopp, E., M. Rejmánek, M. J. Sanderson, and T. L. Rost, 2004 Evolution of genome size in pines (Pinus) and its life-history correlates: supertree analysis. Evolution 58(8): 1705–1729.

Harkins, D. M., P. A. Skaggs, A. D. Mix, G. E. Dupper, M. E. Devey *et al.*, 1998 Saturation mapping of a major gene for resistance to white pine blister rust in P. lambertiana. Theor. Appl. Genet. 97(8): 1355–1360.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood *et al.*, 2013 De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Prot. 8(8): 1494–1512.

Hawkins, J. S., H. Kim, J. D. Nason, R. A. Wing, and J. F. Wendel, 2006 Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res. 16(10): 1252–1261.

Jermstad, K. D., A. J. Eckert, J. L. Wegrzyn, A. Delfino-Mix, D. A. Davis *et al.*, 2011 Comparative mapping in Pinus: P. lambertiana (Pinus lambertiana Dougl.) and P. taeda (Pinus taeda L.). Tree Genet. Genomes 7(3): 457–468.

Keane, R. E., D. F. Tomback, C. A. Aubry, A. D. Bower, E. M. Campbell *et al.*, 2012 A range-wide restoration strategy for whitebark pine (Pinus albicaulis). Gen. Tech. Rep. RMRS-GTR-279. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. p. 108.

Kinloch, Jr., B. B., 1992 Distribution and frequency of a gene for resistance to white pine blister rust in natural populations of P. lambertiana. Can. J. Bot. 70(7): 1319–1323.

Kinloch, Jr., B. B., 2003 White pine blister rust in North America: past and prognosis. Phytopathology 93(8): 1044–1047.

Kinloch, Jr., B. B., and W. H. Scheuner, 1990 Pinus lambertiana Dougl., P. lambertiana. Agric. Handb 654: 370–378.

Kinloch, Jr., B. B., G. K. Parks, and C. W. Fowler, 1970 White pine blister rust: simply inherited resistance in sugar pine. Science 167(3915): 193–195.

Kohany, O., A. J. Gentles, L. Hankus, and J. Jurka, 2006 Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7(1): 474.

Konieczny, A., and F. M. Ausubel, 1993 A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. Plant J. 4(2): 403–410.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway *et al.*, 2004 Versatile and open software for comparing large genomes. Genome Biol. 5(2): R12.

Li, Z., A. E. Baniaga, E. B. Sessa, M. Scascitelli, S. W. Graham *et al.*, 2015 Early genome duplications in conifers and other seed plants. Science Advances 1(10): e1501084.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997.

Liu, J. J., A. K. Ekramoddoullah, R. S. Hunt, and A. Zamani, 2006 Identification and characterization of random amplified polymorphic DNA markers linked to a major gene (Cr2) for resistance to Cronartium ribicola in Pinus monticola. Phytopathology 96(4): 395–399.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang *et al.*, 2012 SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1(1): 18.

Maloney, P. E., D. R. Vogler, A. J. Eckert, C. E. Jensen, and D. B. Neale, 2011 Population biology of sugar pine (*Pinus lambertiana Dougl.*) with reference to historical disturbances in the Lake Tahoe Basin: implications for restoration. Forest Ecology and Management 262: 770–779.

Marçais, G., J. A. Yorke, and A. Zimin, 2013 QuorUM: an error corrector for Illumina reads. arXiv preprint arXiv:1307.3515.

Millar, C. I., 1998 Early evolution of pines, pp. 69–94 in *Ecology and Biogeography of Pinus*, edited by D. M. Richardson. Cambridge University Press, Cambridge, UK.

Morse, A. M., D. G. Peterson, M. N. Islam-Faridi, K. E. Smith, Z. Magbanua *et al.*, 2009 Evolution of genome size and complexity in Pinus. PLoS One 4(2): e4332.

Neale, D. B., J. L. Wegrzyn, K. A. Stevens, A. V. Zimin, D. Puiu *et al.*, 2014 Decoding the massive genome of P. taeda using haploid DNA and novel assembly strategies. Genome Biol. 15(3): R59.

Nystedt, B., N. R. Street, A. Wetterbom, A. Zuccolo, Y.-C. Lin *et al.*, 2013 The Norway spruce genome sequence and conifer genome evolution. Nature 497(7451): 579–584.

Orgel, L. E., and F. H. Crick, 1980 Selfish DNA: the ultimate parasite. Nature 284(5757): 604.

Piegu, B., R. Guyot, N. Picault, A. Roulin, A. Saniyal *et al.*, 2006 Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res. 16(10): 1262–1269.

Richardson, D. M., and P. W. Rundel, 1998 Ecology and biogeography of Pinus: an introduction. pp. 3–48. in *Ecology and Biogeography of Pinus*, edited by D. M. Richardson. Cambridge University Press, Cambridge, UK.

Rosen, S., and H. J. Skaletsky, 1999 Primer3 on the WWW for general users and for biologist programmers, pp. 365–386. in *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, edited by S. Krawetz, and S. Misener. Humana Press: Totowa.

Sax, K., 1960 Meiosis in intraspecific pine hybrids. For. Sci. 6: 135–138.

Saylor, L. C., 1961 A karyotypic analysis of selected species of Pinus. Master's Thesis, North Carolina State University. Genetica 10: 77–84.
Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.
Tenaillon, M. I., M. B. Hufford, B. S. Gaut, and J. Ross-Ibarra, 2011 Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians. Genome Biol. Evol. 3: 219–229.
Tomback, D. F., 1982 Dispersal of whitebark pine seeds by Clark's nutcracker: a mutualism hypothesis. J. Anim. Ecol. 51: 451–467.
Van Pelt, R., 2001 Forest giants of the Pacific coast. University of Washington Press, Seattle.
Vangestel, C., A. Vázquez-Lobo, P. J. Martínez-García, I. Calic, J. L. Wegrzyn *et al.*, 2016 Patterns of neutral and adaptive genetic diversity across the natural range of sugar pine (Pinus lambertiana Dougl.). Tree Genet. Genomes 12(3): 1–10.
Wakamiya, I., R. J. Newton, J. S. Johnston, and H. J. Price, 1993 Genome size and environmental factors in the genus Pinus. Am. J. Bot. 80: 1235–1241.
Waring, K. M., and B. A. Goodrich, 2012 Artificial regeneration of five-needle pines of western North America: a survey of current practices and future needs. Tree Planters Notes 55: 55–71.
Warren, R. L., C. I. Keeling, M. M. Yuen, A. Raymond, G. A. Taylor *et al.*, 2015 Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. Plant J. 83(2): 189–212.
Wegrzyn, J. L., B. Y. Lin, J. J. Zieve, W. M. Dougherty, P. J. Martínez-García *et al.*, 2013 Insights into the P. taeda genome: characterization of BAC and fosmid sequences. PLoS One 8(9): e72439.
Wegrzyn, J. L., J. D. Liechty, K. A. Stevens, L. S. Wu, C. A. Loopstra *et al.*, 2014 Unique features of the P. taeda (Pinus taeda L.) megagenome revealed through sequence annotation. Genetics 196(3): 891–909.
Williams, C. G., K. L. Joyner, L. D. Auckland, S. Johnston, and H. J. Price, 2002 Genomic consequences of interspecific Pinus spp. hybridization. Biol. J. Linn. Soc. Lond. 75(4): 503–508.
Willyard, A., J. Syring, D. S. Gernandt, A. Liston, and R. Cronn, 2007 Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for Pinus. Mol. Biol. Evol. 24(1): 90–101.
Wu, T. D., and C. K. Watanabe, 2005 GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21(9): 1859–1875.
Zimin, A. V., G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg *et al.*, 2013 The MaSuRCA genome assembler. Bioinformatics 29(21): 2669–2677.
Zimin, A., K. A. Stevens, M. W. Crepeau, A. Holtz-Morris, M. Koriabine *et al.*, 2014 Sequencing and assembly of the 22-Gbp P. taeda genome. Genetics 196(3): 875–890.
Zuccolo, A., D. G. Scofield, E. De Paoli, and M. Morgante, 2015 The Ty1-copia LTR retroelement family PARTC is highly conserved in conifers over 200MY of evolution. Gene 568(1): 89–99.

Communicating editor: S. C. Gonzalez-Martinez

Supplementary Methods, Tables, and Figures

Sequencing and Assembly

Paired-end libraries from megagametophytes

Paired-end libraries were constructed as described in Zimin *et al.* (2014). Briefly: approximately 5 μg of DNA from our target megagametophyte was fragmented by sonication, end-repaired, and A-tailed. Universal Illumina paired-end adapters were ligated to the fragments and agarose gel size selection was used to collect a series of ligation-product fractions with mean insert sizes ranging from 180 to 880 bp. Ten ng of each fraction was used as template for a 10-cycle enrichment PCR with barcoded primers. Libraries were quantified on an Agilent Bioanalyzer 2100 and sequenced on the GAIIx and HiSeq 2500 platforms.

Two enrichment PCR chemistries were used: the Illumina-recommended Phusion HF master mix (New England Biolabs) and KAPA HiFi HotStart master mix (Kapa Biosystems). In a side-by-side comparison of k-mer depth distributions the Kapa Biosytems chemistry demonstrated a lower variance in coverage and it was therefore used for all remaining library construction.

Paired-end sequencing

Table S1 Paired end sequencing results by platform. The majority of paired end sequence data came from the HiSeq 2500 platform which replaced the GAIIx as a high throughput longer-read solution achieving an average error-corrected read length just 3 bp shorter than the GAIIx. ('C. len' is corrected length in bp).

Platform	Read length	Reads sequenced	Reads after E.C.	%	Bases sequenced	Bases after E.C. >=31bp	%	C. len	%
MiSeq	255+255	191329972	190012005	99.3	47165405920	44250142585	93.8	234	91.9
HiSeq 2500	150+150	3704633253	3670172611	99.1	5.55695E+11	5.4229E+11	97.6	148	98.5
HiSeq 2500	151+151	5577432158	5518035319	98.9	8.42192E+11	8.20401E+11	97.4	149	98.5
HiSeq 2000	125+125	2250040534	2220695615	98.7	2.81255E+11	2.71663E+11	96.6	122	97.9
GAIIx	160+156	1134732636	1127425204	99.4	1.81557E+11	1.71896E+11	94.7	152	96.5

Table S2 Paired end sequencing results by insert size. We observed a slight reduction in the efficiency of error correction for the longer insert libraries.

Insert size	Libraries	Reads	Reads after E.C.	%	Bases sequenced	Bases after EC >=31bp	%
[200bp, 400bp)	32	6686446005	6634318205	99.2	9.59584E+11	9.32758E+11	97.2
[400bp, 600bp)	12	2961998624	2936847083	99.2	4.64692E+11	4.52177E+11	97.3
[600bp, 900bp)	12	3209723924	3155175466	98.3	4.83589E+11	4.65565E+11	96.3

Paired-end super-reads

The *k*-mer size for the construction of paired-end super reads was optimized to maximize the number of distinct *k*-mers in the error-corrected paired-end data. The value of 89 was chosen using a grid search, implemented by repeatedly running super-read construction, and identifying a local maximum (Table S3).

Table S3 Selecting a value of *k* for the MaSuRCA assembler.

k	Distinct *k*-mer count	*k*-unitig count	Average *k*-unitig length
79	26,999,996,380	585,474,925	125.12
89	27,134,295,936	458,204,603	148.22
99	27,105,725,056	350,678,093	176.30

K-mer histogram for error-corrected paired-end reads

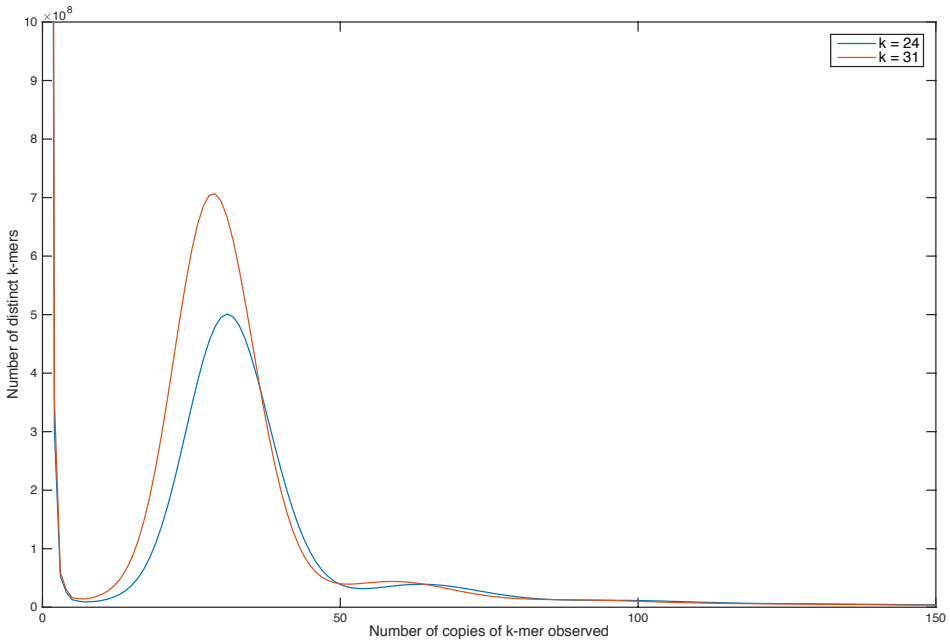

Figure S1 The k-mer histogram of the error-corrected *P. lambertiana* paired-end reads shows a strong distinct peak at 1C depth consistent with haploid DNA. The peak at roughly twice the expected coverage (putative recent duplications) represents approximately 7% of the genome and appears more pronounced than in *P. taeda* (Zimin *et al*. 2014). We observed that 39.4% of the 31-mers were in the more highly repeated tail, to the far right. In *P. taeda* a slightly smaller fraction (34.1%) of 31-mers were in this tail.

Mate pair libraries

For all mate pair libraries input DNA was first treated with 0.33 µl PreCR Repair Mix (New England Biolabs) per microgram of DNA following the manufacturer's guidelines. Jumping libraries were constructed using two methods. Initially libraries were constructed as in (Zimin *et al.* 2014) using the Illumina Mate Pair Library v2 protocol. Later we switched to Illumina's Nextera Mate Pair kit because it gave superior results, particularly for longer-range linkage. Nextera Mate Pair libraries were constructed following the "gel-plus" method in the kit instructions but with the following modifications: input DNA amounts and reagent/reaction volumes for steps up to agarose gel size-selection were tripled in order to achieve increased yields. For longer-range libraries (i.e. > 10 Kbp) the amount of tagmentation enzyme was reduced to 1 µl per microgram input DNA, which shifted the fragment-length distribution to higher molecular weights. Bst polymerase (8 U/µl; New England Biolabs) was sometimes substituted for Strand Displacement Polymerase when kit volumes ran short. PCRClean DX beads (Aline Biosciences) were substituted for Ampure XP beads throughout. 0.6% MegaBase agarose gels were run overnight using a Bio-Rad FIGE Mapper. Shearing of circularized molecules was performed using a Diagenode Bioruptor NGS at high power for 8 cycles of 15 seconds on/90 seconds off. Fifteen cycles of enrichment PCR were performed.

Diploid mate pair sequencing and pre-processing

Deep fragment coverage from long-range paired reads is essential for constructing large scaffolds (Gnerre *et al.* 2011; Ross *et al.* 2013; Zimin *et al.* 2014). Fragment or "clone" coverage refers to the coverage of the genome represented by the entire DNA fragment. Thus if a pair of 100-bp reads is sequenced from both ends of a 5000-bp fragment, the fragment coverage will be 25 times deeper than the actual read coverage. In total, 20 modified Illumina Trueseq and 14 Illumina Nextera mate pair libraries were constructed from diploid maternal genomic DNA. We monitored library complexity during the sequencing process as described in Zimin *et al.* (2014). An initial investigation determined that our modified Illumina Trueseq libraries would be impractical for obtaining deep coverage on the larger genome, particularly for longer fragment sizes. After an evaluation of Illumina's Nextera mate pair libraries, in which we observed deeper per-library coverage, we chose these libraries for the bulk (76%) of our long-fragment sequencing.

Raw sequence from mate pair libraries was processed through a special module of MaSuRCA (Zimin *et al.* 2013) to make the reads match the target haplotype. We used a database of haploid 24-mers to correct errors and single-nucleotide polymorphisms in the diploid read pairs. This correction procedure yielded over 93X fragment coverage in paired reads where both reads had been corrected to match the haploid data (Table S4). This represents more than twice the fragment coverage obtained for *P. taeda* (Zimin *et al.* 2014).

Table S4 Mate pair libraries, MaSuRCA-processed reads, and estimated physical coverage by insert size.

Insert Size Range	Count	Processed reads	Physical coverage
[1Kbp, 5Kbp)	14	358,618,948	18.8X

[5Kbp, 10Kbp)	10	268,825,892	30.3X
[10Kbp, 15Kbp)	7	157,651,636	32.1X
[15Kbp, 25Kbp)	3	41,269,998	12.6X

Illumina sequenced fosmid pool

For use in repeat-library construction, a pool of approximately 5000 *P. lambertiana* fosmid clones (0.5% of the genome) was prepared and sequenced following our previous method (Wegrzyn *et al.* 2013; Zimin *et al.* 2014). Paired-end and Illumina mate pair libraries were prepared as described above. Both libraries were sequenced in a single HiSeq 2500 lane in high-throughput mode (Table S5). Data were processed with RTA 1.17.21.3 and CASAVA 1.8.2. Sequence was subsequently filtered and assembled with SOAPdenovo2 using the method reported in Wegrzyn *et al.* (2013) yielding a 159 Mbp assembly containing 4963 scaffolds greater than 20 Kbp (a fosmid may generate only one of these).

Table S5 Illumina sequencing of fosmid pools.

Library type	Insert size	Number of paired 150 bp reads (Millions)	Number of bases (Mbp)	Estimated coverage
paired-end	400 bp	46.4	13,928	67X
mate pair	3 Kbp	22.5	6,750	32X physical coverage

PacBio sequenced fosmid pools

Four identical fosmid pools of 48 fosmids each were constructed from the larger pool above. These were prepared and sequenced using PacBio RS II for validation purposes. Additional details on the sequencing depth and alignment assembled pools to the WGS assembly are given here.

Table S6 PacBio sequencing of fosmid pools.

Fosmid Pool	Number of reads	Mean read length	N50 read length	Number of bases	Estimated coverage
SPPB1	82,563	7,421	10,863	612,739,994	255X
SPPB2	91,904	6,974	9,815	640,943,357	266X
SPPB3	106,393	6,333	8,969	673,810,507	280X
SPPB4	92,381	6,312	9,023	583,153,465	242X

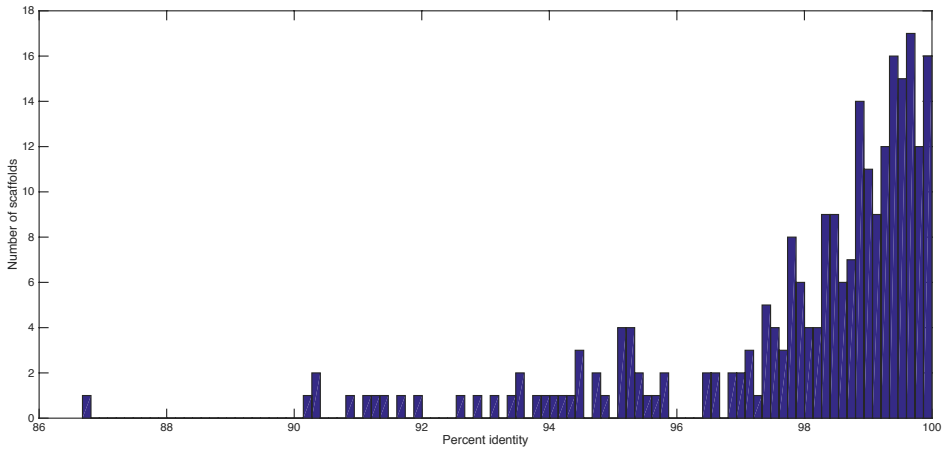

Figure S2 Histogram of %identity weighted across the nucmer alignment for each scaffold when comparing fosmid assemblies to the WGS assembly. The median %identity for an aligned scaffold was 98.82%.**Annotation (genes and transposable elements)**

Libraries used for gene annotation and transcriptome scaffolding

A subset of the libraries and sequence described in Gonzalez-Ibeas *et al.* (2016) were used to construct the transcripts used for scaffolding and annotating the genome. Additional information about those libraries is available here.

Table S7. *P. lambertiana* RNA libraries used in this paper. More details are available in Gonzalez-Ibeas *et al.* (2016). Sequence data is available at GenBank under the NCBI Bioproject 174450.

Library ID and Description	Library type	Sequencing	Transcriptome Scaffolding	Gene Annotation
K, from pollen	RNA-seq	MiSeq	**X**	**X**
M, from early female cones (2 weeks before pollination)	RNA-seq	MiSeq	**X**	**X**
Embryo, from germinating sugar pine seed	RNA-seq	HiSeq, MiSeq, PacBio	**X**	**X**
Basket, from "basket stage" seedling (root, stem, and needles)	RNA-seq	MiSeq	**X**	**X**
S, from 2-cm female cones	RNA-seq	HiSeq, PacBio	**X**	**X**
V, from female cones at the time of pollination	RNA-seq	HiSeq, PacBio	**X**	**X**

DCS, from stem of control plants (no treatment)	RNA-seq	HiSeq, PacBio	**X**	**X**
BRN, from Blister Resistant needles (LCO2-03)	RNA-seq	HiSeq	**O**	**X**
DCR, from root of control plants (no treatment)	RNA-seq	HiSeq	**O**	**X**
JASS, from stem after Methyl jasmonate treatment	RNA-seq	HiSeq	**O**	**X**
NACLR, from root after NaCl treatment	RNA-seq	HiSeq	**O**	**X**
WS, from stem after wounding	RNA-seq	HiSeq	**O**	**X**
BRS, Blister Resistant stem (LCO2-03)	RNA-seq	MiSeq	**O**	**X**
SDN, from needles of seedling slowly drought-stressed	RNA-seq	MiSeq	**O**	**X**
P, from pollen cones	RNA-seq	MiSeq	**O**	**X**

Gene model identification and annotation

Annotation of the *P. lambertiana* genome was performed with MAKER-P. Models that did not contain at least one protein domain as defined in Pfam/Panther via InterProscan were removed. For the high quality set, due to the potential high content of pseudogenes, only multi-exonic models supported by RNAseq data were considered, and remaining models were moved to the low quality set. Manual inspection of gene coordinates of the high quality set and comparison with transcriptome data revealed that the genes could have been split during the identification process (that is, the gene is fragmented in several parts which are counted as independent consecutive gene models sorted on the same genomic area). The problem of genes fragmented into >1 *loci* within the same scaffold during gene prediction has been also reported for other conifers (Nystedt *et al.* 2013). We followed a merging strategy by combining MAKER gene predictions that were mapped under the same transcript source (that is, after mapping the transcript on the genome, it overlapped with split consecutive models). This way, 5,133 original MAKER models were collapsed, resulting in 1,454 merged models. Additionally, we rescued 807 mono-exonic MAKER models by using more stringent criteria (they were full-length, with a recognizable protein domain, supported by RNA-seq data and protein evidence from species relatives and whose *Arabidopsis* counterpart is also mono-exonic (TAIR10 database, e-value cut-off 1e-09)) to be added to the high quality set. Transcripts that were not used by MAKER were aligned to the genome using GMAP and included (1,745 models). In total, 13,936 gene models were

considered the final high-quality set (combined categories) for downstream analysis, and 71,117 were flagged as low quality (Table S8). Categories of the high-quality set included 1) original MAKER predictions (being 9,930 non-merged multi-exonic and 807 mono-exonic, both with RNA-seq support but different selection criteria), 2) 1,454 merged MAKER models, and 3) 1,745 models built from RNA-seq data.

Gene models were subsequently functionally annotated with a characterized plant protein sequence via our in-house annotation pipeline, enTAP (https://github.com/SamGinzburg/WegrzynLab)

Table S8. *P. lambertiana* gene models

Category	**Gene models**	
	Pinus lambertiana	***Pinus taeda***
1) MAKER models with RNA support	**10737**	**5877**
2) Models added from RNAseq data	**1745**	**1466**
3) Total merged models	**1454**	**1681**
Total high quality gene models	**13936**	**9024**
BUSCO gene space completeness (%)	53	30
Models without RNA support (low quality)	71117	75528
Total gene models	85053	84552
BUSCO gene space completeness (%)	58	50
DOGMA gene space completeness (%)	94	61

Tandem repeat identification

P. lambertiana genome v1.0 scaffolds greater than 400 bp were used for tandem repeat analysis. A total of 1,184,160 scaffolds were present in the resulting dataset. Tandem repeat finder (Benson 1999) was used to detect simple repeats across the full genome. Tandem repeats that overlapped interspersed repeats were removed. Tandem repeats

were categorized as microsatellites (2-8bp), minisatellites (9-100bp), or satellites (>100bp). Mononucleotide repeats were excluded as less reliable.

Interspersed repeat identification

To find interspersed repeat elements, we used both similarity and *de novo* based approaches (Supplementary Figure S3). RepeatModeler combines two complementary *de novo* repeat element prediction algorithms: RECON (Bao *et al.* 2002) and RepeatScout (Price *et al.* 2005). To make the RepeatModeler computation tractable, we used only the Illumina sequenced fosmid pools (above) along with the longest 2.5% of genomic scaffolds. We also used a combination of TEclass (Abrusán *et al.* 2009), CENSOR (Kohany *et al.* 2006), and manual characterization to identify the uncharacterized elements from the repeat library produced by RepeatModeler. We used this library along with the plant Repbase library (Jurka *et al.* 2005) (plant component only, v19.01) as the reference database for RepeatMasker (Tarailo-Graovac *et al.* 2009). Full-length elements were determined by applying a cut-off of 80-80-80 (80% sequence similarity and 80 bp minimum length) (Wicker *et al.* 2007).

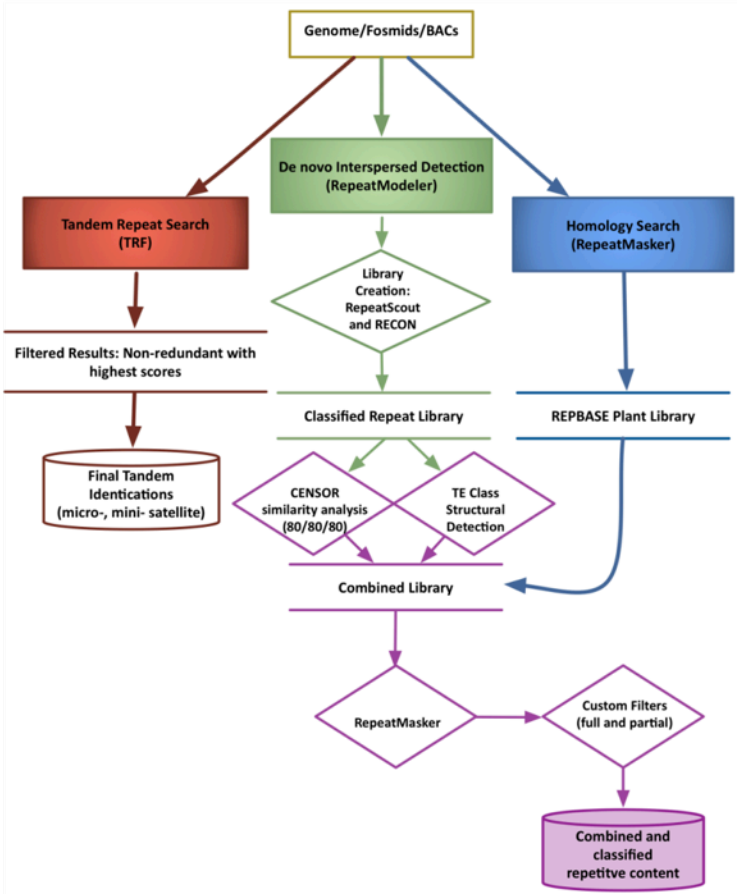

Figure S3. Methodology for identification of repeat elements in the *Pinus lambertiana* and *P. taeda* genomes. Both *de novo* repeat methodology algorithms such as RECON and RepeatScout as well as similarity search using RepeatMasker were used. Full-length repeat

datasets were obtained by using a cut-off of 80% sequence similarity and a minimum of 80bp alignment length (Wicker *et al.* 2007).

Table S9. Full-length and partial repeat elements in *P. lambertiana*

Repeat classification	Percentage of full-length repeat elements	Percentage of partial-length repeat elements
LTR/Gypsy	4.740	27.390
LTR/Copia	1.480	8.570
other LTR	2.070	12.010
Caulimovirus	0.025	0.150
LINE/L1	0.220	1.290
LINE/R1	0.020	0.118
other LINE	0.770	4.490
other SINE	0.045	0.260
other Non-LTR	0.009	0.049
Penelope	0.013	0.081
other Retrotransposon	0.480	2.734
hAT	0.079	0.462
EnSpm	0.084	0.489
Helitron	0.036	0.206
MuDR	0.147	0.852
other DNA	1.054	6.041
other repeat elements	0.006	0.035

LTR insertion time estimation

We used LTR Harvest (Ellinghaus *et al.* 2008) to identify long terminal repeats (LTRs) in the Illumina datasets of *P. lambertiana* and *P. taeda*. Full-length repeats were identified and probed for their respective LTR regions by searching for LTR harvest hits that were subsets of the full-length hits from RepeatModeler (or vice-versa). LTR Harvest alignments that fulfilled this criteria were aligned with MUSCLE (Edgar 2004) and percent identity between the LTR regions at two ends of the retro-transposon was computed. Divergence was calculated from the percent identity using the Jukes-Cantor formula (Chor *et al.* 2006). The insertion time was calculated from the divergence values as described by SanMiguel *et al.* (1998). The nucleotide substitution rates were used as described in the case of *Picea abies* (Nystedt *et al.* 2013).

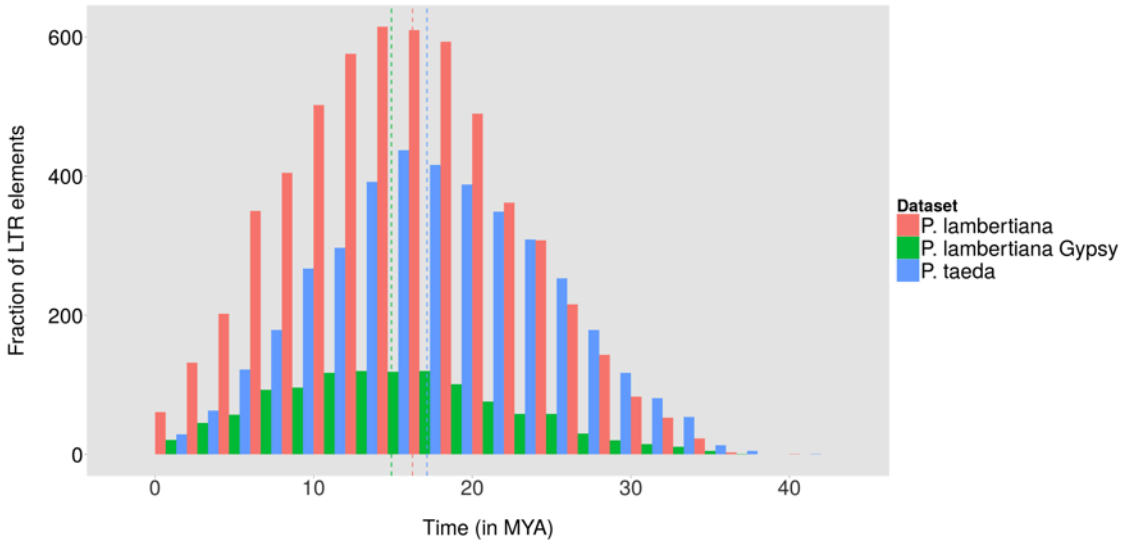

Figure S4. Histogram depicting insertion times of various retrotransposons in the combined fosmid dataset of *P. lambertiana* and *P. taeda*. The dotted lines represent the average insertion time of the respective datasets. Histograms have been created using substitution rates of 2.2 × 10-9 mutations per year from Nystedt *et al.* (2013). Dotted lines represent the average insertion time of their respective datasets in the histogram

Evidence of a recent LTR insertion

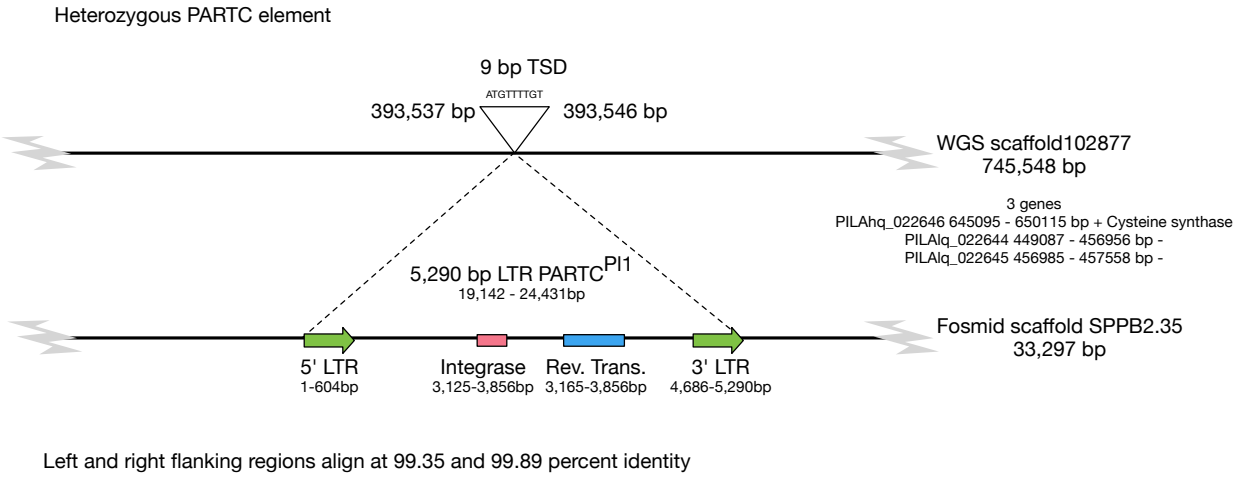

Figure S5. Evidence of a heterozygous and active PARTC element: PARTC^{Pl1} The alignment of genomic scaffold102877 to fosmid scaffold SPPB2.35 reveals a single large structural difference, the insertion of a PARTC into the fosmid scaffold. The 5' and 3' LTR sequences are identical. The coding regions of integrase and reverse transcriptase appear to be functionally conserved (no frameshift or stop codon mutations). At the site of insertion, there are 6-bp duplications at each end of the PARTC^{Pl1} element.

Genomics of the *C. ribicola* Resistance Gene *Cr1*

Megagametophyte DNA Prep

Prior to DNA extraction, megagametophytes were stored at -80℃. Approximately 1/6 of each megagametophyte was ground with two glass beads on a Mini BeadBeater 8 (BioSpec) at maximum speed for 2 minutes. DNA was extracted following a Qiagen DNeasy mini prep kit (Qiagen) with the addition of proteinase K. Quality and quantity were measured using Picogreen dye on a Qubit fluorometer (Invitrogen).

Identification of genomic loci of interest

Jermstad *et al.* (2011) reported the sequences from cloned RAPD bands OP_G16 and BC_432 that were linked to *Cr1*. To identify these genomic loci, the representative consensus sequences for each RAPD band were aligned to the V1.0 draft *P. lambertiana* genome using *gmap* (Wu and Watanabe 2005). In both cases, a unique top hit (path1) was observed and reported.

Primer design and sequencing

We designed nested PCR primers using PRIMER3 (Rosen and Skaletsky 1999) on the reference genome repeat masked changing RepeatMasker annotated repeats to N (Figure S3). Table S9 gives a list of primers and conditions. All of the PCR assays used standard PCR reaction conditions: 2.0 mM $MgCl_2$, 0.2 mM each of dNTPs, 0.5 mM each of forward and reverse primers, 1U of Taq and 50 ng of DNA.

For validation purposes, we used the available primer sequences of PCR amplicon, UMN_3258_01 (ftp://dendrome.ucdavis.edu/ftp/CRSP/) to develop a new marker *cr1lC*. In our case, genotype was determined by sequencing UMN_3258_01 and subsequent *phred* and *phrap* analysis as described below.

SNP discovery

The DNA sequences for each PCR amplicon were processed and assembled with *phred* and *phrap* (Greene *et al.* 1992) with default parameters. The resulting contigs were subsequently inspected with *consed*. If a single contig was produced, SNPs and short indels were determined by inspection for high-quality discrepancies with the consensus sequence. Most segregating loci produced two scaffolds. SNPs and short indels were identified by alignment of the sequences with *muscle* (Edgar 2004).

Table S10. PCR primer details

Scaffold	Primer name	Primer sequence	[Mg^{2+}]	Annealing		Size (bp)	Com-ment
				Temp (C)	Time (sec)		
370413	cr1lB_F1	GATAGGGAGGTTACAGGCCC	2.00	57	30	1083	External
370413	cr1lB_R1	TAGTGGATAGGAACCGTGGC					primers
370413	cr1lB_nF1b	ACAAGAATCTTACCTGGGCC	1.50	56	30	482	Nested
370413	cr1lB_nR1b	GTCTATTTAAGCCACGCCCC					primers
223058	cr1lA_F2	ATTTTCACGCCTTCTACGCC	2.00	57	30	1064	External
223058	cr1lA_R2	TTGCTAAGGACCCAGATCCC					primers
223058	cr1lA_nF2a	AGCTTTGAATTGCGCTAGGG	1.50	58	30	577	Nested
223058	cr1lA_nR2a	CGCTGAGTACCCATATCCCC					primers
277631	277631_F1	GGGGAGGGGTGTCATTGTTA	2.00	57	30	932	External
277631	277631_R1	CCCCAATGTTTGTGACCCAG					primers
277631	277631_nF1a	CCACCCTAGCTCCAAAGTGA	1.00	57	30	420	Nested
277631	277631_nR1a	GCATCTCCATTTGTTGCGGA					primers

Cleaved amplified polymorphic sequence (CAPS) assays

The two distinct haplotypes per loci that were identified with consed were mapped for restrictions sites using RestictionMapper (http://www.restrictionmapper.org/). We identified specific restriction enzymes that detect polymorphic cutting sites producing readily discernable banding patterns. Conditions and size distributions are described in Table S10. A set of 99 megagametophytes from randomly selected open-pollinated seeds of parent 5701 ($Cr1^R/Cr1^r$) were initially genotyped for the CAPS markers. A second expanded search for recombinants was made by pre-screening a larger set of 1054 megagametophytes for RAPD markers BC_432_1110 and OPG_16_950 used in Harkins *et al.* (1998). This screen resulted in an enriched subset of 146 proposed recombinants.

We expect the assignment of *Cr1* genotypes to be susceptible to error (Harkins *et al.* 1998) and we did observe a small number of 'double crossovers' based on their proposed gene order, OP_G16 – *Cr1* - BC_432. (Table S12). These were removed from downstream analysis.

Table S11. Restriction digest markers

Marker	Restriction enzyme	Sequence	Reaction conditions	Inactivation conditions	Haplotype 1	Haplotype 2
cr1lB	MseI (10 µl)	TTAA	37°C for 15 min	65C for 20 min	322 bp	116 and 164 bp
cr1lA	RsaI (10 µl)	GTAC	37C for 15 min	65C for 20 min	~290 bp	204 and ~290bp

Table S12 Restriction digest genotyping results

cr1lB	***cr1lA***

*Mse*I (TTAA)	*Cr1*	*Rsa*I (GTAC)	Count
116,164	R	204,292	74
322	r	284,292	138
322	r	204,292	2
116,164	r	204,292	5
322	R	284,292	7

Table S13. Sequenced cloned RAPD markers anchored to the assembly (top), and the corresponding cloned amplified polymorphic sequence (CAPS) assays (bottom).

RAPD/SCAR	Scaffold ID	Scaffold length (bp)	Position begin (bp)	Position end (bp)	Coverage	Identity
***(scar)*OPG16_950**	223058	303,049	221,124	221,124	98.1%	97.2%
***(scar)*BC432_1110**	370413	655,271	119,205	119,205	99.3%	95.7%
CAPS marker	**Linked RAPD**	**Restriction enzyme**	**Cut site**	**Amplicon size**	**Haplotype 1**	**Haplotype 2**
cr1lA	OPG16_950	MseI	TTAA	577	322 bp	116bp, 164bp
cr1lB	BC432_1110	RsaI	GTAC	482	~290 bp	204bp, ~290bp

Linking in additional scaffolds

We used Fosmid DiTag linking libraries not included in the assembly to link in additional scaffolds. The libraries were constructed using a refinement of the approach used in *Zimin et al.* 2014, modified so that library inserts containing a junction motif could be enriched by hybridization[1]. The Fosmid DiTag libraries were aligned to the genome using *bwa mem* (Li and Durbin 2010). Alignments were kept if their mapping quality exceeded a minimum threshold of 40 and both sequences aligned within 40 kbp to the end of a scaffold with an implied distance of less than 55 Kbp. We had the highest confidence in the link between scaffold370413 and super6135 which was witnessed by two DiTag pairs (Table S14).

Table S14. Linking fosmid DiTags in the *Cr1* region.

DiTag pair	Target scaffold ID	Alignment start (bp)	Mapping quality	Scaffold length (bp)	Offset from beginning (end) (bp)
2	scaffold370413	15586	60	655271	15586
2	super6135	770368	43	772474	(2106)
3	scaffold370413	6392	60	655271	6392
3	super6135	760342	60	772474	(12132)

[1] http://www.idtdna.com/pages/docs/default-source/xgen-libraries/xgen-lockdown-protocols/hybridization-capture-protocol-xgen-lockdown-probes-and-reagents.pdf

Table S15. Megagametophytes sequenced for the population sample. With one exception one megagametophyte from each phenotyped seed tree was sequenced. All 8 available megagametophytes were sequenced from SP-K-0142-U.

Seed Tree ID	Resistance Phenotype	National forest	Ranger district	Elevation
19600	RR	Tahoe	Downieville	5500
19188	RR	Sierra	Minarets	5981
19409	RR	Stanislaus	Groveland	4500
19601	RR	Tahoe	Downieville	5500
18875	RR	n/a	n/a	n/a
18852	RR	n/a	n/a	n/a
6351	Rr	Shasta-Trinity	Mt. Shasta	5600
6200	Rr	Six Rivers	Lower Trinity	4900
5892	Rr	Klamath	Goosenest	6100
6902	Rr	Lassen	Hat Creek	5600
6352	Rr	Shasta-Trinity	Mt. Shasta	5800
5062	Rr	Klamath	Happy Camp	3700
7646	Rr	Sierra	Pine Ridge	5600
6353	Rr	Shasta-Trinity	Mt. Shasta	5900
7519	Rr	Eldorado	Georgetown	3000
6202	Rr	Six Rivers	Lower Trinity	4800
6554	Rr	Shasta-Trinity	Weaverville	5100
7453	Rr	Tahoe	Foresthill	4600
SP-1151-AD-00015	rr	Plumas	Beckwourth	7000
SP-0356-00043	rr	Eldorado	Placerville	7000
SP-0353-00060	rr	Eldorado	Georgetown	3500
SP-1156-00068	rr	Plumas	Quincy	6200
SP-1154-00087	rr	Plumas	Feather River	3000
SP-1156-00091	rr	Plumas	Quincy	7000
SP-1553-00115	rr	Sierra	Pine Ridge	6500
SP-K-0121-U	rr	Klamath	Ukonom	5500
SP-K-0132-U	rr	Klamath	Ukonom	1500
SP-K-0139-U	rr	Klamath	Ukonom	1020
SP-K-0142-U	rr	Klamath	Ukonom	2030
SP-K-0144-U	rr	Klamath	Ukonom	3070
SP-K-0145-U	rr	Klamath	Ukonom	1250
SP-K-0149-U	rr	Klamath	Ukonom	3601
SP-K-0155-U	rr	Klamath	Ukonom	4507
SP-0355-00159	rr	Eldorado	Pacific	6000
SP-0355-00162	rr	Eldorado	Pacific	5500
SP-1154-00216	rr	Plumas	Feather River	3500
SP-0351-00218	rr	Eldorado	Amador	4500
SP-1153-00226	rr	Plumas	Feather River	2400
SP-1154-DFC-00272	rr	Plumas	Feather River	4000
SP-0351-00303	rr	Eldorado	Amador	6500

Transcript evidence for linked and associated genes

Candidate transcripts were found by BLASTX search using the candidate genes. Transcripts were kept if the reciprocal best gmap alignment of the candidate transcript to the genome overlapped the candidate gene. The candidate transcript TR43508|c1_g1_i2|m.82078 was identified in a library constructed from needles of a resistant genotype inoculated with the fungus *C. ribicola*. The library was prepared, sequenced with the HiSeq platform, and analyzed by the same method described in Gonzalez-Ibeas *et al.* (2016). This library was not included neither in the scaffolding nor the annotation transcriptome sets.

Figure S6. Candidate transcript from a resistant library overlapping gene candidate PILA_017786.

```
>TR43508|c1_g1_i2|m.82078
AAACTCAGAAACCTTCAATACATCGATTTGGAAGGTGCTTCTAATTTGCAGATGCTTCCA
AATTCATTTGGGGATTTAACTCAACTCAAACATCTAATTTTGAAAAGGTGCTCTAATTTG
ACCATCTCCAGCGAAGCACTTGGAAATATTACCACGCTTAAAAGCTTAGATCTTTCATAT
TGTAACCAGGTGAAAGACGTGCCTCCCCAAGTCACACGTCAACTGTCCTTGCAAAACTTA
TATTTGAATGGATCAAAGTTAAAAGAATTGCCGAGCAATATTGGAGTCCTCTGCAATTTG
GAAGTTCTGCATTTAGGTAGCGATTTGTTGGAAGCGCTGCCAGATGGTCTTGGTGTCCTG
AATAGTTTGAAGAGATTATCACTCTCTTCTTCGCCGCAGTTGAAATCCTTGCCGGATTCC
ATTGGACTATTGACTCAGTTGAGAGTACTGGTCATAGAATCTTGCGGACTAGAATCCTTA
CCAAAAGAAATTTTCAAGATGAGTAATCTGAGAAGTTTAATGATACGGAATTGTCCGTTG
CGGGAACACCCATTTAGAAAGGAGTTTGAAGGAGTAAGAGAAACGCACTTATTATTGGAA
GGGGAAAGTGCGTTGAATAATTTGAACTCCTCCAATCACAGACGCATGTTTGGGCTCAAG
TGGTTAACCCTGTCAGGCACAGAAATAAGGGAGGTATTTTTTGATGAGGGCGTTTTCCCC
TGCGTTCAACAACTAAATGTTCTAGACTGCCCTGAGATACGTAAGTTGTCAGTGGAACAT
TTAACTTCTTTGGAGAATTTGGTTGTTCGGCAATGCAAGAATCTCCAGAGCATACTAGGG
TTGAGGCAGCTCACACAGCTTACAGAACTACATGTTTATGGATGCCCTGAGATACGAGAG
CTGCCAGGTGTGGAACAATTGGTTTCTTTGGAGATGTTGAAAATTGGGGAATGC
```

Table S16 Gene annotation for scaffolds linked to the *Cr1* locus.

Scaffold	Gene ID/Name	Annotation
scaffold370413	PILA_071809	Alias=uninformative, Interpro:IPR000757,PANTHER:PTHR31062,PANTHER:PTHR31062:SF18,Pfam:PF00722, note:partial
scaffold223058	PILA_008442	Alias=putative MYB DNA-binding domain superfamily protein,Interpro:IPR001005,PANTHER:PTHR10641,PANTHER:PTHR10641:SF460,Pfam:PF00249,note:partial
scaffold223058	PILA_008443	Alias=ATPUP11, putative,Interpro:IPR004853,Interpro:IPR012946, Interpro:IPR030182,PANTHER:PTHR31376,PANTHER:PTHR31376:SF2,Pfam:PF03151,Pfam:PF07983, note:complete
Scaffold223058	PILA_008444	Alias=adenosylhomocysteinase/s-adenosyl-l-homocysteine hydrolase,Interpro:IPR000043, Interpro:IPR015878,PANTHER:PTHR23420,Pfam:PF00670, note:complete
scaffold223058	PILA_008445	Alias=non-annotated model, Interpro:IPR000043,Interpro:IPR015878,PANTHER:PTHR23420,Pfam:PF00670, note:complete
scaffold223058	PILA_008446	Alias=PREDICTED: transcription factor MYB108-like, Interpro:IPR001005,PANTHER:PTHR10641,PANTHER:PTHR10641:S

		F484,Pfam:PF00249, note:complete
scaffold223058	PILA_008447	Alias=RAB GTPase homolog A4C,I nterpro:IPR001806,PANTHER:PTHR24073,PANTHER:PTHR24073:S F437,Pfam:PF00071, note:complete
scaffold223058	PILA_008448	Alias=PREDICTED: alpha-galactosidase-like isoform X1, Interpro:IPR000111,PANTHER:PTHR11452,PANTHER:PTHR11452:S F18,Pfam:PF02065, note:partial
scaffold223058	PILA_008449	Alias=R2R3-MYB transcription factor, Interpro:IPR001005,PANTHER:PTHR10641,PANTHER:PTHR10641:S F494,Pfam:PF00249, note:complete
super6135	PILA_017784	Alias=PREDICTED: probable xyloglucan endotransglucosylase/hydrolase protein 32-like, Interpro:IPR000757,PANTHER:PTHR31062,PANTHER:PTHR31062:S F18,Pfam:PF00722, note:complete
super6135	PILA_017785	Alias=putative DNAJ heat shock protein, Interpro:IPR002939,PANTHER:PTHR24077,Pfam:PF01556, note:complete
super6135	PILA_017786	Alias=uninformative,Interpro:IPR001611,Interpro:IPR002182, Interpro:IPR026906,PANTHER:PTHR23155,Pfam:PF00560,Pfam:PF 00931,Pfam:PF13306,Pfam:PF13504, note:complete
super6135	PILA_017787	Alias=uninformative,Interpro:IPR001452,Pfam:PF00018, note:complete
super6135	PILA_017787	Alias=uninformative, Interpro:IPR001452,Pfam:PF00018,note:complete

Pinaceae phylogenetic tree estimation

A multitude of studies has examined phylogenetic patterns within genera, as well as among genera. The vast majority of these studies, however, are based on chloroplast DNA (cpDNA; e.g. Eckert and Hall 2006; Gernandt et al. 2008, Parks et al. 2009; Hernandez-Leon et al. 2013) or handfuls of nuclear loci with or without inclusion of cpDNA (e.g., Wang et al. 2000; Syring et al. 2005; Willyard et al. 2007). Most studies have identified a broadly supported backbone for branching patterns for the phylogeny of the Pinaceae (Fig. 1). More contentious, however, is the estimation of divergence times, due not only to use of fossils in questionable placements in the phylogeny (Eckert and Hall, 2006; Willyard et al., 2007; Gernandt et al., 2008), but also to limited information about branch lengths across multiple, independent loci. Here, we utilize the resource provided in this paper to estimate a multilocus phylogeny for the Pinaceae based on 28 nuclear genes using the BEAST ver. 2.20 software (Bouckaert et al. 2014). Specifically, we explored estimates of divergence times in a six-taxon tree (*Pinus* subg. *Pinus, Pinus* subg. *Strobus, Picea, Larix, Pseudotsuga*, and *Abies*) representing approximately 55% of the genus-level diversity within the Pinaceae. Divergence times were estimated under two models of molecular evolution, each assuming an HKY+G substitution model - (1) a global, strict molecular clock and (2) a global, relaxed molecular clock parameterized with a lognormal distribution.

Parameters for both models were estimated using MCMC with 1.1 x 10^8 steps, a burn-in of 1.0 x 10^7, and a thinning interval of 1.0 x 10^4. Convergence was assessed for each model through comparisons of three independent runs of the MCMC routine, while mixing for each run was assessed using effective sample size (ESS) calculations based on the autocorrelation of parameter estimates along the Markov chains. Models were compared using Bayes factors (BFs) based on the marginal likelihoods for each model (Suchard et al. 2001). For comparison, we also report modified AIC values for each model (Baele et al. 2012). All post-MCMC analysis was conducted using Tracer ver. 1.6 (Rambaut et al. 2014).

Table S17. Summary of the 28 loci used for phylogenetic inference of divergence times within the Pinaceae. Putative homologs were identified via blastx analysis of the expressed sequence tag (EST) contig against the Reference Protein database housed at NCBI. More information about these loci is available in the DiversiTree database housed at the Dendrome website (https://dendrome.ucdavis.edu/DiversiTree/). Information about the assembly and sequencing of loci across the Pinaceae can be found in Eckert *et al.* (2013a, 2013b). Loci with NA in the E-value column did not have a putative homolog found in the Reference Protein database via blastx analysis of the EST contig listed in the second column of the table.

Locus id	EST contig id	Homolog	Gene Product	E-value
0_846_01	0_846	NM_129800	bZIP transcription factor	6.00E-14
0_5038_01	0_5038	XP_010248353	Phloem protein 2-Like A10-like protein	5.00E-21
0_6448_02	0_6448	NM_099986	ATP-dependent helicase (DCL1)	4.00E-130
0_8642_01	0_8642	XP_003635538	Elongation factor G-2, chlroplastic-like	6.00E-125
0_9383_01	0_9383	NM_106563	Ubiquitin thiolesterase	7.00E-53
0_10706_01	0_10706	NM_179945	Uncharacterized protein	3.00E-08
0_11772_01	0_11772	XP_003554743	Probable tRNA N6-adenosine threonylcarbamoyltransferase	5.00E-139
0_12745_01	0_12745	NM_122578	Kelch repeat-containing F-box family protein	5.00E-59
0_13240_01	0_13240	NM_121480	L-aspartate oxidase	6.00E-68
0_14122_02	0_14122	NM_113125	Uncharacterized protein	4.00E-75
0_15075_01	0_15075	NM_129383	CAX-interacting protein	6.00E-51

0_15762_01	0_15762	NA	NA	NA
2_1501_01	2_1501	XP_016463965	Uncharacterized protein	2.00E-42
2_1528_01	2_1528	XP_010497358	Mediator of RNA polymerase II transcription subunit 33B-like protein	4.00E-55
2_3742_03	2_3742	XP_010269982	LAG1 longevity assurance homolog 2-like protein	2.00E-35
2_8011_02	2_8011	NM_116232	Scarecrow-like transcription factor	2.00E-27
2_8443_01	2_8443	XP_016647698	Glycosyltransferase family protein 64 C5 isoform	7.00E-126
2_9456_01	2_9456	XP_006844460	E3 ubiquitin-protein ligase ORTHUS 2 isoform X1	5.00E-12
CL149Contig3_04	CL149Contig3	NM_112485	L-asparaginase	2.00E-70
CL516Contig1_07	CL516Contig1	XP_009410369	Pyrophosphate-energized vacuolar membrane proton pump-like protein	0.0
CL1064Contig1_02	CL1064Contig1	XP_006844510	Protein bicaudal C homolog 1	3.00E-25
CL2472Contig1_01	CL2472Contig1	XP_010919153	Lysine-specific histone demethylase 1 homolog 3	1.00E-32
CL3148Contig1_04	CL3148Contig1	XP_002318094	Leucine-rich repeat transmembrane protein kinase	3.00E-115
CL3770Contig1_01	CL3770Contig1	XP_008219652	Uncharacterized protein	3.00E-16
CL4354Contig1_01	CL4354Contig1	XP_002270378	Serine/threonine-protein phosphatase PP2A-2	3.00E-114
CL4481Contig1_04	CL4481Contig1	NP_564202	OB-fold nucleic acid binding domain-containing protein	3.00E-57
CL4511Contig1_02	CL4511Contig1	XP_013592268	Protein HHL1, chloroplastic-like isoform X2	2.00E-67
UMN_1023_01	UMN_1023	XP_00685278	F-box/LRR-repeat protein 14	3.00E-103

ADDITIONAL REFERENCES

Abrusán G., Grundmann N., DeMester L., Makalowski W. 2009. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics, 25(10):1329-1330.

Bao, W., M. G. Jurka, V. V. Kapitonov and J. Jurka 2009. New superfamilies of eukaryotic DNA transposons and their internal divisions. Molecular biology and evolution: msp013.

Bao Z. and Eddy S. R. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Research, 12(8):1269-1276.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research 27(2): 573.

Chor, B., M. D. Hendy and S. Snir 2006. Maximum likelihood Jukes-Cantor triplets: analytic solutions. Molecular biology and evolution 23(3): 626-632.

Eckert, A. J., J. L. Wegrzyn, J. D. Liechty, J. M. Lee, W. P. Cumbie, J. M. Davis, B. Goldfarb, C. A. Loopstra, S. R. Palle, T. Quesada, C. H. Langley, and D. B. Neale. 2013a. The evolutionary genetics of the genes underlying phenotypic associations for loblolly pine (Pinus taeda, Pinaceae). Genetics 195: 1353-1372.

Eckert, A. J., A. D. Bower, K. D. Jermstad, J. L. Wegrzyn, B. J. Knauss, J. V. Syring, and D. B. Neale. 2013b. Multilocus analyses reveal little evidence for lineage wide adaptive evolution within major clades of soft pines (Pinus subgenus Strobus). Molecular Ecology 22: 5635-5650.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32(5), pp.1792-1797.

Ellinghaus, D., S. Kurtz and U. Willhoeft 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC bioinformatics 9(1): 18.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences, 108(4), 1513-1518.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany and J. Walichiewicz 2005. "Repbase Update, a database of eukaryotic repetitive elements." Cytogenetic and genome research 110(1-4): 462-467.

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC bioinformatics, 7(1):474.

Price, A. L., N. C. Jones and P. A. Pevzner 2005. De novo identification of repeat families in large genomes. Bioinformatics 21(suppl 1): i351-i358.

Ross, M. G., C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum, and D. B. Jaffe. 2013 Characterizing and measuring bias in sequence data. Genome biology 14, no. 5 R51.

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

Tarailo-Graovac, M. and N. Chen 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics: 4.10. 11-14.10. 14.

Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante and O. Panaud 2007. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8(12): 973-982.

Zimin, A., Stevens, K. A., Crepeau, M. W., Holtz-Morris, A., Koriabine, M., Marçais, G., ... and Langley, C. H. 2014. Sequencing and assembly of the 22-Gb P. taeda genome. Genetics, 196(3), 875-890.