# 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*

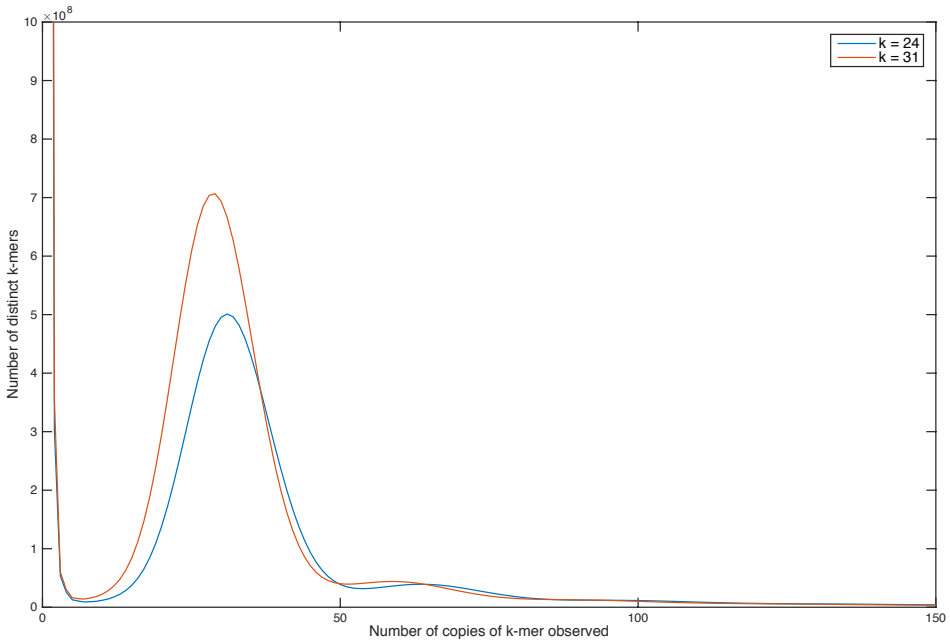


**Figure S1** The k-mer histogram of the error-corrected *P. lambertiana* paired-end reads shows a strong distinct peak at 1C depth consistent with haploid DNA. The peak at roughly twice the expected coverage (putative recent duplications) represents approximately 7% of the genome and appears more pronounced 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 |

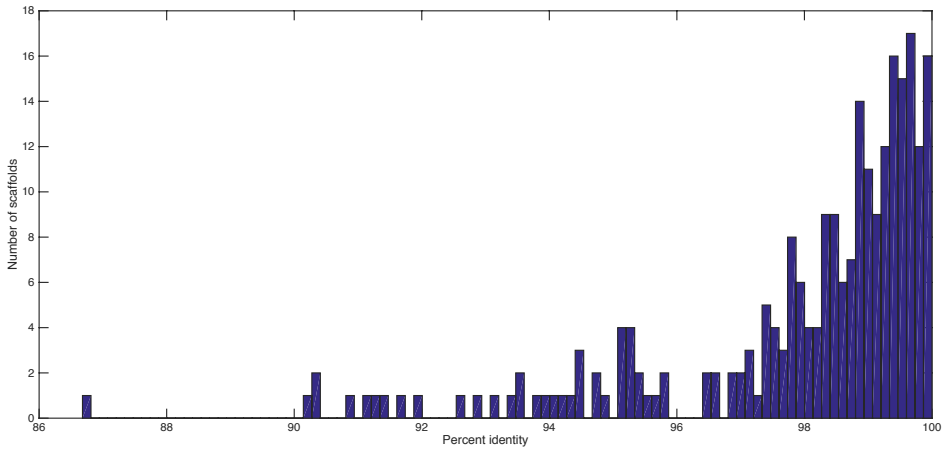


Figure S2 Histogram of %identity weighted across the nucmer alignment for each scaffold when comparing fosmid assemblies to the WGS assembly. The median %identity for an aligned scaffold was 98.82%.**Annotation (genes and transposable elements)**

*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).

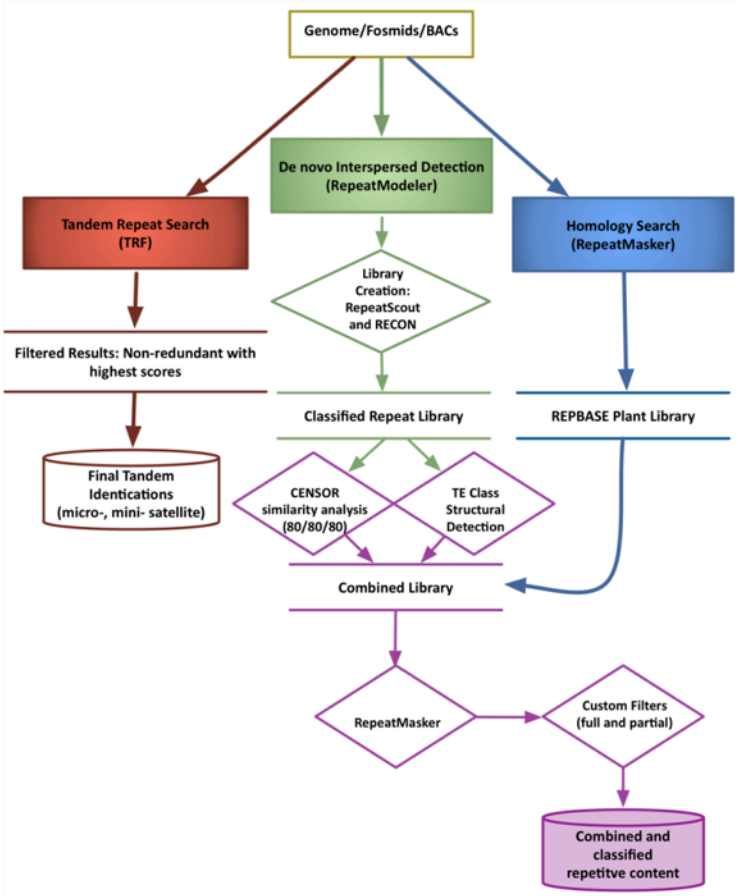


**Figure S3.** Methodology for identification of repeat elements in the *Pinus lambertiana* and *P. taeda* genomes. Both *de novo* repeat methodology algorithms such as RECON and RepeatScout as well as similarity search using RepeatMasker were used. Full-length repeat

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).

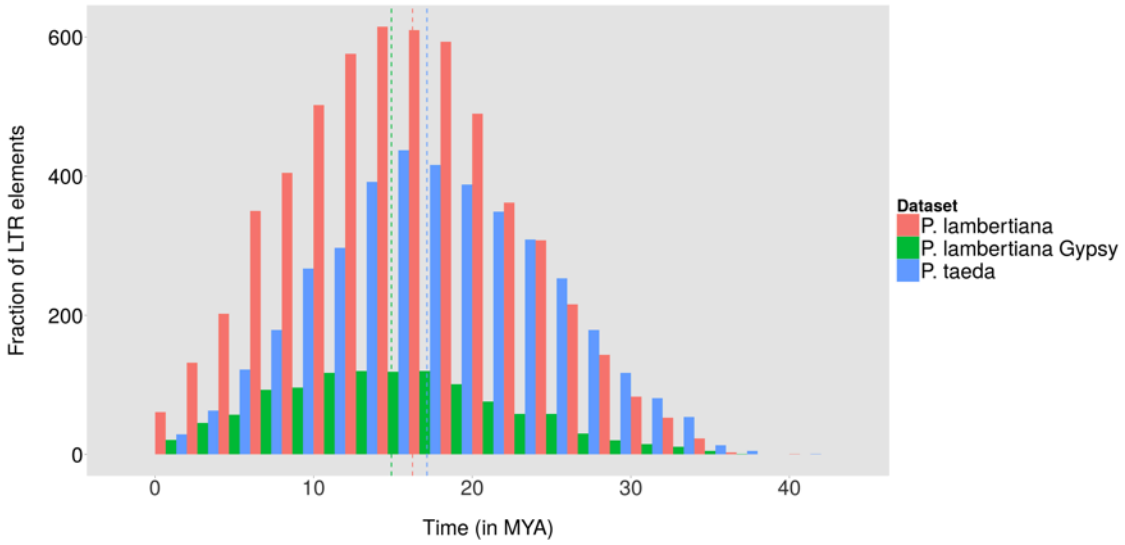


**Figure S4.** Histogram depicting insertion times of various retrotransposons in the combined fosmid dataset of *P. lambertiana* and *P. taeda*. The dotted lines represent the average insertion time of the respective datasets. Histograms have been created using substitution rates of 2.2 × 10-9 mutations 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*

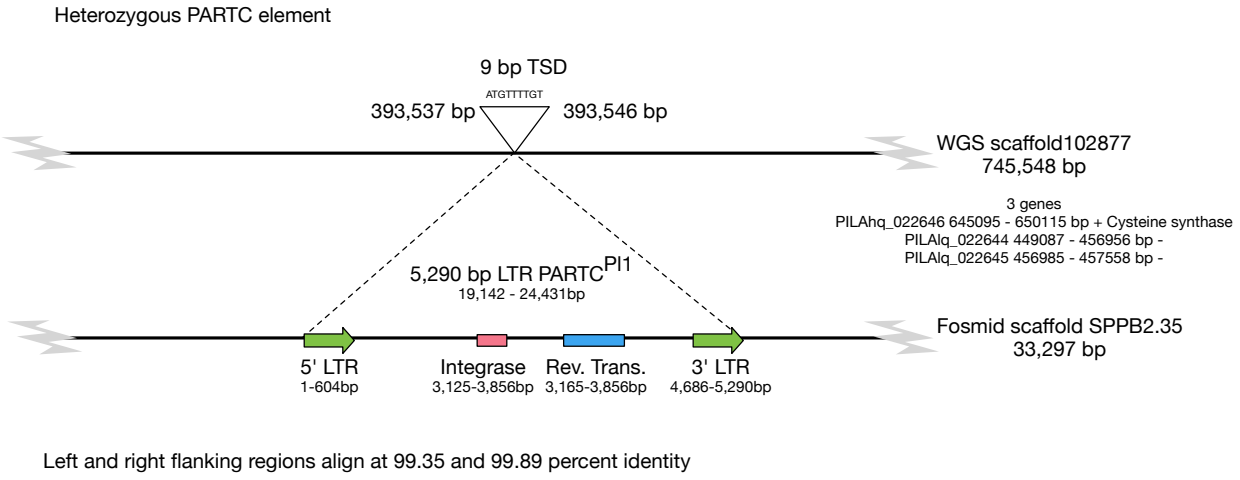


**Figure S5.** Evidence of a heterozygous and active PARTC element: PARTC$^{Pl1}$ The alignment of genomic scaffold102877 to fosmid scaffold SPPB2.35 reveals a single large structural difference, the insertion of a PARTC into the fosmid scaffold. The 5' and 3' LTR sequences are identical. The coding 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.