Sequencing and Assembly of the 22-Gb Loblolly Pine Genome

Aleksey Zimin,*,[1] Kristian A. Stevens,[†,1,2] Marc W. Crepeau,[†] Ann Holtz-Morris,[‡] Maxim Koriabine,[‡] Guillaume Marçais,* Daniela Puiu,[§] Michael Roberts,* Jill L. Wegrzyn, Pieter J. de Jong,[‡] David B. Neale,** Steven L. Salzberg,[§] James A. Yorke,*,[††] and Charles H. Langley[†]**

*Institute for Physical Sciences and Technology and [††]Departments of Mathematics and Physics, University of Maryland, College Park, Maryland 20742, [†]Department of Evolution and Ecology and **Department of Plant Sciences, University of California, Davis, California 95616, [‡]Children's Hospital Oakland Research Institute, Oakland, California 94609, [§]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University, Baltimore, Maryland 21205

ABSTRACT Conifers are the predominant gymnosperm. The size and complexity of their genomes has presented formidable technical challenges for whole-genome shotgun sequencing and assembly. We employed novel strategies that allowed us to determine the loblolly pine (*Pinus taeda*) reference genome sequence, the largest genome assembled to date. Most of the sequence data were derived from whole-genome shotgun sequencing of a single megagametophyte, the haploid tissue of a single pine seed. Although that constrained the quantity of available DNA, the resulting haploid sequence data were well-suited for assembly. The haploid sequence was augmented with multiple linking long-fragment mate pair libraries from the parental diploid DNA. For the longest fragments, we used novel fosmid DiTag libraries. Sequences from the linking libraries that did not match the megagametophyte were identified and removed. Assembly of the sequence data were aided by condensing the enormous number of paired-end reads into a much smaller set of longer "super-reads," rendering subsequent assembly with an overlap-based assembly algorithm computationally feasible. To further improve the contiguity and biological utility of the genome sequence, additional scaffolding methods utilizing independent genome and transcriptome assemblies were implemented. The combination of these strategies resulted in a draft genome sequence of 20.15 billion bases, with an N50 scaffold size of 66.9 kbp.

GYMNOSPERMS diverged from the lineage leading to angiosperms >300 million years ago (Bierhorst 1971; Magallón and Sanderson 2005). In comparison to their more recently diversified sister taxa, the flowering plants, gymnosperms possess dramatically larger genomes (Bowe *et al.* 2000; Peterson *et al.* 2002; Morse *et al.* 2009; Zonneveld 2012). While rapid progress has been made in characterizing the genomes of angiosperms, the same is not true for gymnosperms, in part due to their size and complexity. Comparative studies indicate that repetitive sequences, transposable elements, and gene duplication have proliferated in gymnosperms (Ahuja and Neale 2005; Morse *et al.* 2009; Kovach *et al.* 2010; Mackay *et al.* 2012).

doi: 10.1534/genetics.113.159715
Manuscript received November 13, 2013; accepted for publication December 10, 2013
Available freely online through the author-supported open access option.
Data can be accessed at http://www.pinegenome.org/pinerefseq and at National Center for Biotechnology Information BioProject 174450.
[1]Joint first authors.
[2]Corresponding author: Department of Evolution and Ecology, University of California, Davis, CA 95616. E-mail: kastevens@ucdavis.edu

The gymnosperms diversified early after the divergence from angiosperms and are represented today by strikingly diverse ancient lineages (Magallón and Sanderson 2005). Less than 100 MYA, in the Cretaceous period, the Pinaceae lineage diversified to become the dominant terrestrial plant in temperate and subarctic climatic zones. Conifers are by far the most abundant gymnosperm, representing >80% of the Earth's biomass (Neale and Kremer 2011).

The first target for genome sequencing from the genus *Pinus* was chosen for its economic (agricultural) as well as biological (phylogenetic and ecological) importance. *Pinus taeda* (loblolly pine) is a classic representative of the largest genus in the order Conifereae. The genus *Pinus* consists of >100 species (Mirov 1967). Loblolly pine is a native forest tree species in the southeastern United States that is harvested for commercial lumber, pulp, and paper markets (Frederick *et al.* 2008). Because of its economic value, loblolly pine has been at the center of a number of long-term breeding programs (Schutz 1997; McKeand *et al.* 2003) as well as more fundamental genetics investigations and resource

development (*e.g.*, genetic maps). The genotype selected for sequencing, designated "20-1010," is the property of the Virginia State Department of Forestry, which released this germplasm into the public domain for use without restriction. Standard flow cytometry methods applied to *P. taeda* placed the size of the genome at ~21.6 Gb (O'Brien *et al.* 1996), more than seven times larger than the human genome. It is the largest genome sequenced and assembled to date using any technology. Achieving a successful outcome required leveraging unique aspects of conifer biology and employing novel computational methodologies to reduce the assembly problem to a tractable scale.

Among the many distinguishing characteristics of gymnosperms, the most important for our project is the haploid female gametophyte (or megagametophyte) in each seed. The genome of the megagametophytic cells is derived by mitosis from the same meiotic product (megaspore) as the egg nucleus in the archagonium. The haploid megagametophyte serves as a nutrient source for the developing embryo.

A single haploid megagametophyte offers a major reduction in sequence complexity although at the expense of constraining the amount of DNA available. Haploid sequence is much easier to assemble than outbred diploid sequence. As described below, the majority of our raw sequence data came from a single pine seed, but the source was insufficient to provide DNA for our long-range "linking" libraries. Paired reads from longer DNA fragments require considerably more DNA, and for these libraries we used diploid DNA obtained from needles of the maternal tree (see *Materials and Methods*). The large size of the *P. taeda* genome necessitated many individual long-insert libraries to achieve sufficient physical coverage. To link the assembly together over large distances, we created two types of linking libraries: (1) mate pair "jumping" libraries in which paired reads were ~1–6 kbp apart and (2) fosmid ends, or DiTags, in which paired reads were from the ends of a 35- to 40-kbp fosmid clone. Fosmid clones are less expensive to generate than BACs and are well known to have less cloning bias and a more narrowly controlled insert size (Kim *et al.* 1992).

Our assembly strategy was built around the MaSuRCA genome assembler (Zimin *et al.* 2013). Several of its innovations were developed specifically to handle the demands of this very large project. The key idea in MaSuRCA is to reduce high-coverage paired-end reads to a much smaller and more concise set of "super-reads."

Applied to our data, the MaSuRCA assembler can be conceptually divided into four major phases. The first phase corrects errors in the Illumina reads, using the QuORUM error corrector (Marcais *et al.* 2013). The second phase reduces the short and highly redundant Illumina paired-end reads to a concise set of super-reads. To do this, MaSuRCA utilizes an exhaustive list of distinct sequences of length k, or k-mers, derived from the paired-end reads as a summary of the haploid genome. A "super-read" is constructed from a paired-end read when the ends can be extended uniquely (as judged from the list of k-mers) to form a single uninterrupted sequence containing both reads. For a deep-coverage data set, many read pairs will extend into the same super-read, which results in significant compression of the data. The assembly uses only maximal super-reads, *i.e.*, those super-reads that are not properly contained in other super-reads. The third phase of assembly uses an adapted overlap-based assembler, CABOG (Miller *et al.* 2008), to assemble the super-reads together with filtered read pairs from the longer diploid libraries. The final phase further increases contiguity by using a local assembly algorithm to fill gaps in the scaffolds.

Following assembly by MaSuRCA, we implemented additional custom scaffolding methods, described below, using both an independent SOAPdenovo2 (Luo *et al.* 2012) assembly and high-quality transcriptome data to further improve the assembly's contiguity.

Two other conifer genome sequences, of the Norway spruce and white spruce, have recently been reported (Birol *et al.* 2013; Nystedt *et al.* 2013). Our initial draft assembly of loblolly pine, release 0.6, was released on June 11, 2012 (http://loblolly.ucdavis.edu/bipod/ftp/Genome_Data/genome/pinerefseq/Pita/v0.6/). It contained 18.5 Gbp of sequence with an N50 contig size of 800 bp.

The much-improved genome sequence of Loblolly pine (Neale *et al.* 2014), whose sequencing and assembly are described here, contains 20.1 Gbp with an N50 contig size of 8.2 kbp and an N50 scaffold size of 66.9 kbp. By many measures, it represents the most contiguous and complete conifer (gymnosperm) genome sequence available (*Appendix C*).

Materials and Methods

Plant material

Wind-pollinated seeds (the source of a megagametophyte DNA) were collected from a grafted ramet of *P. taeda* genotype 20-1010 in a Virginia Department of Forestry seed orchard near Providence Forge, Virginia. Needles were collected from grafted ramets of the *P. taeda* genotype 20-1010 growing at the Erambert Genetic Resource Management Area near Brooklyn, Mississippi, and the Harrison Experimental Forest near Saucier, Mississippi.

DNA isolation

Megagametophytes were individually dissected from the seeds. Each megagametophyte was frozen in liquid nitrogen and ground to a fine powder with a mortar and pestle, and DNA was extracted using the NucleoSpin Plant II kit with buffer PL2. DNA yield was quantified using a PicoGreen fluorometric assay. Nuclei were isolated from the needles using the method in Peterson *et al.* (2000) with the exception that two layers of MiraCloth were substituted for cheesecloth in the second filtration step and the 37.5% Percoll gradient was replaced by a layered gradient of 30, 60, and 90% Percoll atop a final layer of 85% sucrose (intact nuclei accumulated just above the sucrose layer). A total of 500 ml of extraction buffer nuclear isolation buffer (MEB) was used for every 50 g of needles. (Note that during preparation of MEB the pH must first be stabilized between pH 3 and 4

with concentrated HCl before being brought back to a final pH of 6.0 with NaOH.) In some instances, the above method was modified by doubling the concentrations of anti-oxidants and nuclease inhibitors in the extraction buffer and extracting the crude homogenate with chloroform just prior to the first centrifugation step.

Isolated nuclei for fosmid cloning were embedded in agarose plugs, and nuclear DNA was further purified by pulsed-field gel electrophoresis (*Appendix A*). Nuclei for jumping library construction were resuspended in 4 ml of MPDB (2-methyl-2,4-pentanediol buffer), and then aqueous SDS was added to a final concentration of 2% (w/v). Contents were mixed gently to lyse the nuclei and placed on ice for 20 min. A 1/100 volume of 10 mg/ml Proteinase K was added, and the mixture was incubated for 30 min at 37°. A second 1/100 volume aliquot of Proteinase K was added, followed by another 30-min incubation at 37°, and then 7.565 g of finely ground $CsCl_2$ was added and the mixture was incubated at room temperature for 30 min with gentle rocking. Volume was brought to 8.5 ml with MPDB, and the mixture was centrifuged at 90,000 × *g* for ~72 hr. Fractions were collected from the resulting gradient, and the fractions containing DNA were dialyzed for several days against multiple 4-liter volumes of 1× TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). DNA yield was quantified using a PicoGreen fluorometric assay. To reduce the presence of single-strand nicks (which increase the fraction of nonjunction mate pairs), DNA for jumping library construction was in some instances treated with 10 units of S1 nuclease per microgram of DNA, thus converting single-strand nicks into double-strand breaks.

Megagametophyte selection

Care was taken to select a single haploid megagametophyte for sequencing from recently collected 20-1010 seeds. Illumina library bias and complexity is well known to be affected by the quality and quantity of the input DNA (Ross *et al.* 2013). The amount of DNA recovered in a preliminary experiment from the haploid tissue of a loblolly pine megagametophyte varied considerably (mean: 1.35 μg; standard deviation: 0.50 μg) and was always below the ideal for the Illumina paired-end library protocol. From a set of 17 megagametophyte DNA samples selected for preliminary library construction and sequencing, one with a high DNA yield (2.1 μg; in the upper 10%) was selected for construction of a series of short-insert libraries to be deeply sequenced. Coincidentally, these selected libraries ranked among the lowest in the amount of contaminating plastid DNA as well as in a measure of bias using homology to repeats in existing BAC sequences (Kovach *et al.* 2010).

Sequence data

Second-generation genome sequencing projects can benefit greatly by sequencing a diversity of libraries, which can reduce bias (Ross *et al.* 2013) and provide linking information at different scales (Gnerre *et al.* 2011; Birol *et al.* 2013; Wang *et al.* 2013). We sequenced 11 haploid paired-end Illumina libraries on three Illumina platforms. The HiSeq 2000 was the least expensive per base pair, but read length (100 bp) was the shortest. We obtained reads up to 160 bp from the GAIIx, but with lower throughput and increased cost, and 250-bp paired-end reads from a MiSeq instrument, which has the lowest throughput and the highest cost per base. In total, >1.4 trillion base pairs (Tbp) of short-insert, paired-end, high-quality sequence was obtained (Table 1), which corresponds to 64× coverage of a 22-Gbp genome.

Table 1 Overview of WGS sequence obtained for this project

Library type	Instrument	Fragment size (bp)	Read length (bp)	Coverage
Illumina Paired-end	GAIIx	200–657	156–160	22×
Illumina Paired-end	HiSeq	200–657	100–128	42×
Illumina Paired-end	MiSeq	350–657	255	<1×
Illumina Mate Pair	GAIIx	1300–5500	156–160	13×
Fosmid DiTag	GAIIx	35,000–40,000	156–160	<1×

The final column reports the nonredundant depth of coverage that was obtained for each library and instrument type based on a genome size of 22 Gbp.

A total of 48 Illumina long-insert mate pair libraries were sequenced on the Illumina GAIIx (*Appendix B*, Table B2). As expected from the limited efficiency of the library construction protocol, the complexity of each library (defined as the numbers of unique molecules sequenced) was much lower than for the short-insert libraries. The total raw sequence coverage obtained from long-insert libraries was ~13×.

Construction of paired-end libraries from megagametophytes

The entire DNA amount extracted, ~2.1 μg for the target megagametophyte, was fragmented by sonication using a Bioruptor UCD-200 (Diagenode) at high power for 15 cycles of 30 sec on/30 sec off. Universal Illumina paired-end adapter was ligated to the end-repaired, A-tailed fragments. To create highly size-specific libraries, but preserve the maximum amount of DNA for library construction, the adapter-ligation product was size selected on a 2% agarose gel by collecting a series of ~1-mm gel slices along the length of the gel lane, with mean insert sizes ranging from 209 to 638 bp (see Figure 1). DNA was extracted from the gel slices using a QIAGEN MinElute kit, and each DNA aliquot was used in its entirety as template for a 10-cycle enrichment PCR with Phusion HF DNA polymerase and barcoded primers. Libraries were quantified on an Agilent Bioanalyzer 2100 and sequenced on the GAIIx and HiSequation 2000 platforms.

Subsequent alignment using Illumina's CASAVA pipeline (version 1.8.2) to the *P. taeda* chloroplast genome and available genomic sequence was used for determining insert-size statistics. To estimate the empirical distributions shown in Figure 1, ~100,000 correctly oriented paired-end read alignments were sampled from each library to construct an insert-size histogram.

Long-insert mate pair (jumping) libraries

Jumping libraries were constructed using standard Illumina methods with either Illumina Mate Pair Library v2 kits or

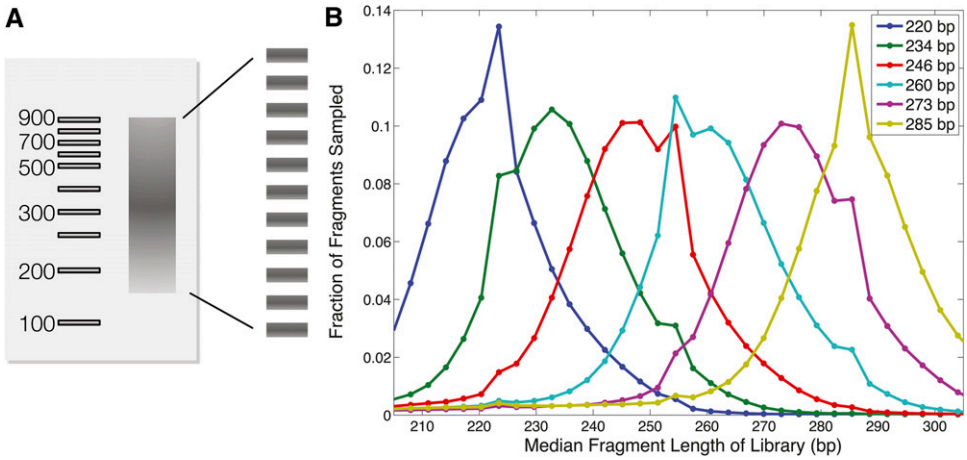

Figure 1 Partition library construction. (A) Scheme to partition the single megagametophyte DNA sample into multiple libraries. Megagametophyte DNA was sonicated and then run out on an agarose gel; the target size range was excised and partitioned into equal spaced slices. The goal was to set the number and sizes of slices such that the coefficient of variation on the insert-size distribution within each fraction was small enough to support high quality *de novo* assembly. (B) The empirical fragment size distributions of partitioned libraries made from our target megagametophyte.

comparable reagents from other vendors. Briefly, at least 10 μg of DNA was fragmented to the desired size range using a HydroShear Plus (Digilabs). DNA was end-repaired for 15 min with unmodified dNTPs and then for an additional 15 min with biotinylated dNTPs. Where the Illumina kit was not used, we found that a mixture of 0.75 mM biotin-16-AA-2′-dUTP (TriLink Biotechnologies) and 0.75 mM biotin-16-AA-2′-dCTP (TriLink Biotechnologies) could be substituted for the biotin-dNTP supplied in the kits. Biotinylated DNA was size-selected on a 0.6% MegaBaseagarose (BioRad) gel run overnight at ~1 V/cm. The gel was cut to collect DNA fractions with defined size ranges, and DNA was extracted from each gel slice using the QIAEX II kit (Qiagen). Circularization reactions containing up to 600 ng (2- to 5-kb libraries) or 1200 ng (>5-kb libraries) of DNA were run overnight at 30°. Where non-kit reagents were used, we found that T4 DNA ligase (New England Biolabs) at a final concentration of 90 units/μl achieved efficient circularization. Noncircularized fragments were removed by digestion with exonuclease, and the remaining circularized molecules were fragmented by sonication using a Bioruptor UCD-200 (Diagenode). Biotinylated (junction) fragments were pulled down using Dynabeads M-280 streptavidin magnetic beads (Invitrogen), and end-repair, A-tailing, and adapter-ligation reactions were performed on the biotinylated fragments bound to the beads. The adapter-ligated fragments were then used as template for 18 cycles of enrichment PCR. All jumping libraries were made using multiplex paired-end adapters and oligos as described for fragment libraries above. We ultimately preferred to use KAPA HiFiHotStartReadyMix (KAPA Biosystems) over Phusion polymerase for the enrichment PCR, taking care to follow the manufacturer's recommendations regarding denaturation times and temperatures.

Finally, the libraries were quantified on an Agilent Bioanalyzer and sequenced on a GAIIx with a paired-end read length of 160 × 156 bp, a run regime chosen for compatibility with short-insert libraries sequenced simultaneously on adjacent lanes of the flow cell. The sequenced reads were aligned to available high-quality *P. taeda* reference sequence to confirm that the insert sizes and numbers of nonjunction fragments were within acceptable ranges.

Fosmid DiTag libraries

We constructed fosmid DiTag libraries as follows. DNA from diploid needle tissue was cloned using the same method as described below for fosmid pool construction to create a particle library of ~20 million clones in vector pFosDT5.4. The vector pFosDT5.4 incorporates features allowing two methods for the creation of fosmid DiTags (Figure 2). The first, a nick-translation method (Williams *et al.* 2012), utilizes two Nb.BbvCI recognition sites located on opposite vector strands flanking the insert (Figure 3A). Digestion with enzyme Nb.BbvCI introduced a nick in each strand. This was followed by nick translation with *Escherichia coli* DNA polymerase I that moved nicks toward each other and inside the insert. The distance between insert ends and nicks is controlled by the polymerase concentration. Nicks were then converted to double-strand breaks with S1 nuclease, repaired to blunt ends, and the resulting internally deleted fosmid vectors with truncated insert sequences attached were recirculated at low DNA concentration to prevent formation of chimeras by intermolecular ligation.

Alternatively, the nick translation approach was replaced with an endonuclease digestion method (Figure 3B). This method was facilitated by engineering the pFosDT5.4 vector to contain no FspBI or Csp6I recognition sites. Fosmid molecules were digested with each enzyme either singly or in combination (the two enzymes generate compatible ends) to release a large internal fragment of the genomic insert while leaving behind ends of indeterminate length. This was followed by recircularization of the internally deleted fosmids.

Following recircularization, the remaining insert ends were amplified by PCR using the Illumina primers built into the vector, and the resulting library was size-selected on an agarose gel to collect library molecules with insert sizes ranging from 150 to 450 bp. As an internal control, we supplemented the libraries at a 1% ratio with library molecules made from *Drosophila melanogaster* genomic DNA using the same vector. The resulting DiTag libraries were sequenced on an Illumina GAIIx (*Appendix B*, Table B3).

Estimating the haploid genome size from haploid k-mer data

Recall that a k-mer is a contiguous sequence of length k; thus, a read of length L contains $L - k + 1$ k-mers. We computed k-mer histograms of haploid sequence data using the program jellyfish (Marcais and Kingsford 2011). To reduce the effect of low-quality data, we used the QuORUM error-corrected reads (below). To avoid quantization error, the total count for every distinct k-mer was computed independently of the histogram.

The genome size was estimated as the total number of k-mers divided by the expected depth or multiplicity of distinct k-mers in the data. The approach used to estimate the latter quantity is to focus on the range of the distribution in which most k-mers are unique and substitute an estimate of the expected depth of a unique k-mer instead. Upon inspecting our empirical data, we determined that k-mers occurring 15 or fewer times were mostly errors. We also considered k-mers observed >120 times (approximately twice the most common multiplicities of unique k-mers) to be repeats. We then determined the expected value of the k-mer depth distribution from a cubic interpolating spline fit to the data, using MATLAB fit, after outliers had been removed. This gave a slightly smaller estimate of the expected k-mer depth than the more typical approach of using the maximum (mode).

Whole-genome shotgun assembly with MaSuRCA:

Sequence preprocessing and error correction: The MaSuRCA assembly pipeline first corrected errors in the Illumina reads using QuORUM (Marcais *et al.* 2013), a method that uses k-mer frequencies and quality scores to correct errors. A notable aspect of this was that the list of "good" 24-mers used for error correction was compiled only from the haploid paired-end data. During the process of error correction SNPs between the haplotypes are indistinguishable from error. A fraction of SNPs in the reads from the diploid mate pair and DiTag libraries was changed to match our target megagametophyte haplotype. When larger differences were present, such as insertions or deletions, the error corrector deleted these reads. In addition, the set of 24-mers belonging to Illumina adapters, chloroplast and mitochondrial DNA were identified and added to a special QuORUM "contaminant" list. QuORUM truncated the read if it encountered a contaminant 24-mer. The trimming of mate pair and DiTag reads to the junction site was also accomplished during this phase.

Optimized super-read construction: For the super-read reduction, the MaSuRCA assembler utilized k-mers from the error corrected paired-end reads as a summary of the underlying haploid genome. The k-mer length was optimized to maximize the number of distinct k-mers utilized by the super-read reduction (see Zimin *et al.* 2013 for details). If k is too small, the number of distinct k-mers will decrease due to short repeats. Conversely, if k is too large, distinct regions of the genome may be missing due to low coverage. A grid search to find the optimal value was performed between $k = 70$ and $k = 85$. From this we inferred an optimal value for the number of utilized k-mers at $k = 79$. This became the chosen k-mer length.

We then applied the super-reads reduction using a database of 79-mers to the deep-coverage paired-end data from the megagametophyte. This produced a set of super-reads based strictly on the haploid DNA. We discuss in detail the outcome of this phase in the *Results*.

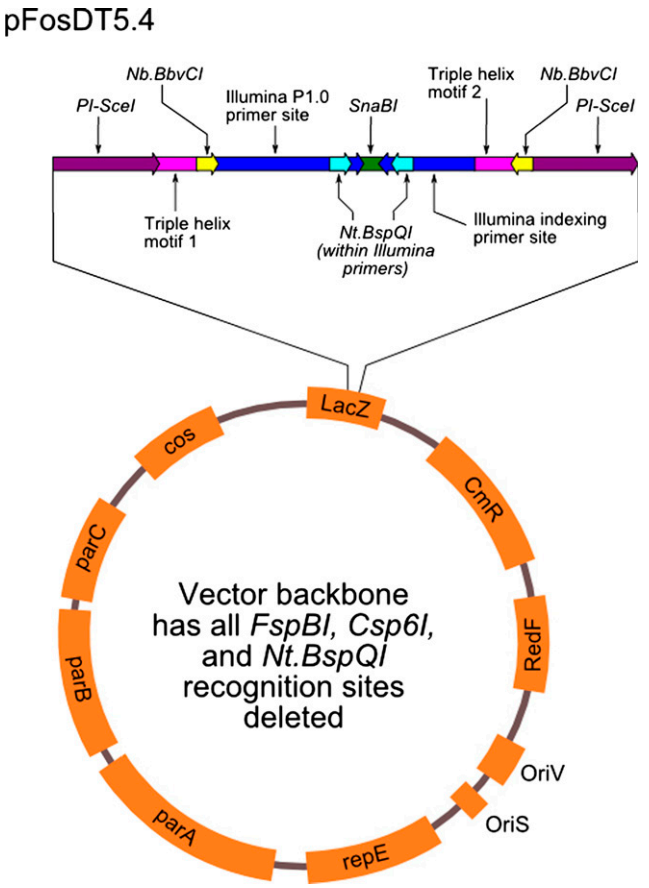

Figure 2 An overview of the pFosDT5.4 fosmid cloning vector used for DiTag creation. The cloning site (where the genomic DNA is inserted) is flanked by forward and reverse Illumina primers to enable end sequencing. Two Nb.BbvCI nicking endonuclease sites are located 5′ of the Illumina primers on both strands to allow for creation of DiTags by a nick translation method. The vector backbone has had all FspBI and Csp6I sites removed to allow for creation of DiTags by endonuclease digestion.

Assembly with diploid reads: Because the long-insert libraries were constructed from diploid pine needles, half of these fragments, on average, derived from a different parental genome than the haploid megagametophyte. Many of these pairs contained differences that distinguished them from the haploid DNA. Although most of the reads with haplotype differences were corrected or trimmed, some remained after error correction. To reduce the number of scaffolding conflicts caused by heterozygosity, we required that all 57-mers in both reads in a linking pair were present in the haploid paired-end data;

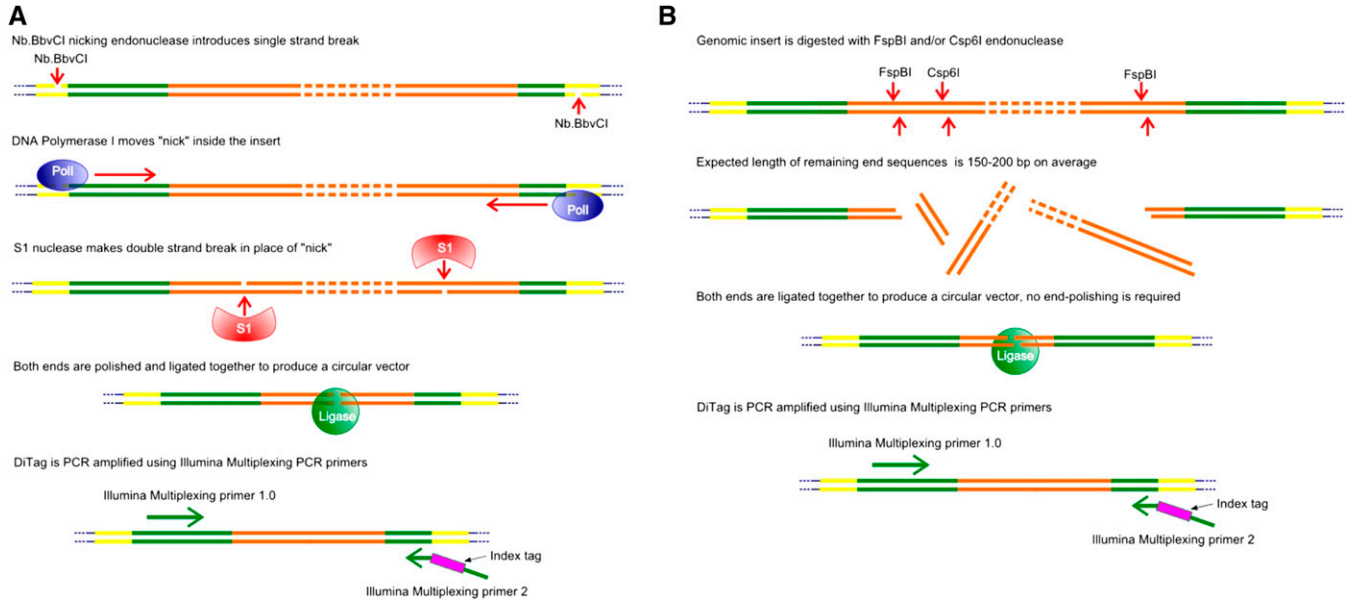

Figure 3 Schematic for our two methods for converting a fosmid library into an Illumina compatible DiTag library using the fosmid vector created for this project. (A) A nick translation approach, similar to the approach used in Williams *et al.* (2012), was implemented for approximately one-half of the libraries. (B) An endonuclease digestion protocol was also used for approximately one-half of the libraries. Although the location of the junction sites is more constrained, in practice, we obtained higher yields from this method.

otherwise, the read pair was rejected. This step should have retained all pairs from the same haplotype.

We also implemented a preprocessing step to remove apparently duplicate molecules from the long-insert libraries. This step is critical for long-insert libraries because of their reduced library complexity. By filtering these out, we ensure that each link inferred from long-insert reads is independently ascertained.

We discuss details of coverage and filtering results for both types of long-insert linking libraries in the *Results*. The filtered linking reads were combined with super-reads and assembled using an adapted version of CABOG.

The MaSuRCA assembler was run on a 64-core IA64 computer with 1 terabyte of RAM. Running the assembly pipeline took 3 months. The maximum memory usage during the assembly reached 800 GB during the super-reads step. The output of MaSuRCA was the 1.0 version of the assembly (see Table 3).

Results and Discussion

Haploid megagametophyte library complexity

The complexity of a library is typically measured as the number of distinct molecules in the library (Daley and Smith 2013). For the library construction protocols described here, this is less than the total number of molecules in the library due to a PCR amplification step, as well as *in vivo* amplification for fosmid clone-based libraries.

Library complexity curves, which plot the number of distinct molecules against the number of molecules sequenced, were used to characterize the rate of diminishing returns because a library of finite complexity is sequenced more deeply. We constructed these curves on QuORUM error-corrected data as follows. We represented each molecule by a concatenation of a short prefix and suffix of length k, giving a sequence, or tag, of length $2k$. For strand consistency, the prefix was obtained from the first read and the suffix was obtained from the reverse complemented second read. We chose the parameter k to be large enough (here, $k = 24$) so that the tags were typically unique. Two tags were identified as the same molecule if their sequences matched exactly in the same or opposite orientations. From the tags we constructed a histogram of molecular frequency in the sequenced sample, which we sampled without replacement to create the complexity curve.

Figure 4 plots the number of distinct read pairs against the total number of sequenced pairs. The shape of the curve indicates how fast the library's complexity is being depleted, *i.e.*, how the likelihood of a duplicate increases as sequencing continues. For our most deeply sequenced library, Figure 4 shows the limits of paired-end library complexity when constructed from a megagametophyte. For our deeply sequenced library, the curve shows that additional sequencing would sample new molecules at a rate of $<40\%$. It also indicates that, for most of our libraries, deeper sequencing would yield additional useful data. When considered with the partition scheme described previously, we see that our megagametophyte WGS protocol is a valuable method for obtaining highly broad and deep paired-end sequence data types for *de novo* assembly.

Haploid k-mers and an estimate of the 1N genome size

Using short k-mers from all the reads, we estimated the genome size to provide an assessment of the overall paired-end library

Table 2 Haploid (1N) genome size estimates

	k-mer length	
	31	24
Total k-mers	1.08×10^{12}	1.16×10^{12}
E(distinct k-mer depth)	53.76 occurrences	56.99 occurrences
Genome size (Gbp)	20.12	20.42

bias and complexity and to appraise previous estimates based on flow cytometry.

Figure 5 shows a histogram of k-mer counts in our haploid error-corrected reads, computed using the program jellyfish (Marçais and Kingsford 2011) for $k = 24$ and $k = 31$. Small values were chosen for tractability as well as good separation between the unique haploid k-mers and uncorrected errors. Each point in the plot corresponds to the number of k-mers with a particular count; for example, the number of 31-mers that occur 50 times is plotted in red at $x = 50$. Note that we expect a k-mer that occurs just once in the genome (*i.e.*, is nonrepetitive) to occur C times on average, where C is depth of coverage. For haploid data, this plot would ideally take the form of a single-mode Poisson distribution with mean equal to C (Lander and Waterman 1988). Deviations are attributable to other properties of the genome and to systematic errors in library construction and sequencing. Each sequencing error may create up to k single-copy k-mers, producing a peak around $x = 1$. The number of these k-mers have been greatly reduced in Figure 5 by the QuORUM error-correction algorithm. The long tail in the plot contains conserved high-copy-number repeats.

We estimated the genome size as the ratio of the total number of k-mers divided by the mean number of occurrences of the unique k-mer counts (an estimate of the expected depth of distinct sequences of length k), as shown in Figure 5. Table 2 gives these quantities as well as the genome size estimates for $k = 24$ and $k = 31$. Both estimates are slightly smaller than the value determined by flow cytometry (21.6 Gbp) (O'Brien *et al.* 1996) and are consistent with our final total contig length.

Haploid sequence data reduction: from reads to super- reads

The goal of the super-read algorithm is to reduce the quantity of sequence data presented to the overlap-based assembler. Each maximal super-read included in the downstream assembly passed two critical tests: successful filling of the gap between a pair of reads and determination that the extended super-read was not properly contained within another super-read (maximal).

Paired-end gap fill: The extension of a read by the super-read algorithm is done conservatively with the philosophy of relegating the important assembly decisions to the overlap-based assembly. Hence, read extension will halt when it cannot be done unambiguously. For a pair of reads to result in a super-read, the algorithm needs to successfully fill in the sequence of the gap (see Zimin *et al.* 2013). Two properties of the sequenced genome play a large role in the success rate of gap filling. The first is the number and size of repetitive sequences in the genome. An unresolved repetitive sequence will prevent filling the gap. More repeats can be resolved by using a longer k-mer size; as explained above, we chose $k = 79$ for *P. taeda*. We avoided the second property of concern, divergence between the two parental chromosomes in a diploid genome, by constructing all short-insert libraries from DNA of a single haploid megagametophyte.

As the gap size increases, so does the likelihood of encountering ambiguity, making it more difficult to fill gaps for longer fragments. For our haploid read pairs, the rate of successful gap filling was as high as 96% for overlapping GAIIx reads from our shortest library. Over the range of gap sizes in our data, we observed an approximately linear 6% reduction in the rate of successful gap filling for every 100-bp increase in the expected gap size (*Appendix B*, Table B1). This is consistent with a relatively low per-base probability ($\sim 6 \times 10^{-4}$) that a problem is encountered during gap filling over the range of gap sizes observed.

Maximality: Only the subset of "maximal" super-reads is passed to the overlap-based assembler because fully contained reads would not provide additional contiguity information. There is a trade-off in producing maximal super-reads. Assuming that reads have the same length, it is easier to fill the gap between pairs that come from shorter fragments. But for libraries of longer fragments, although successful gap filling is less frequent, more of the gap-filled pairs generate maximal super-reads. We observed that the percentage of maximal super-reads uniformly increased with library fragment length, despite reduced gap-filling effectiveness (*Appendix B*, Table B1). Thus, the lower gap-filling rate for longer fragments did not detract from their utility.

We can use the rate of maximal super-read creation to estimate the marginal impact of additional paired-end sequencing. To improve the assembly, additional paired-end reads must be able to generate new maximal super-reads. In our analysis, we found that ~24% of the read pairs in longer (500 bp) libraries produced additional maximal super-reads *vs.* only 8–10% of the pairs in libraries shorter than 300 bp (*Appendix B*, Table B1). This suggests that the most effective way to improve the assembly further would be to prioritize additional paired-end data from 500-bp or longer fragment libraries.

Overall paired-end data reduction: MaSuRCA's error-correction and super-reads processes reduced the 1.4 Tbp of raw paired-end read data (in ~15 billion reads) to 52 Gbp in super-reads, a 27-fold reduction. The reduced data contained ~150 million maximal super-reads with an average length of 362 bp, equivalent to 2.4× coverage of the genome. This 100-fold reduction in the number of paired-end reads is critical for the next step of assembly, which computes pairwise overlaps between these reads along with the linking libraries.

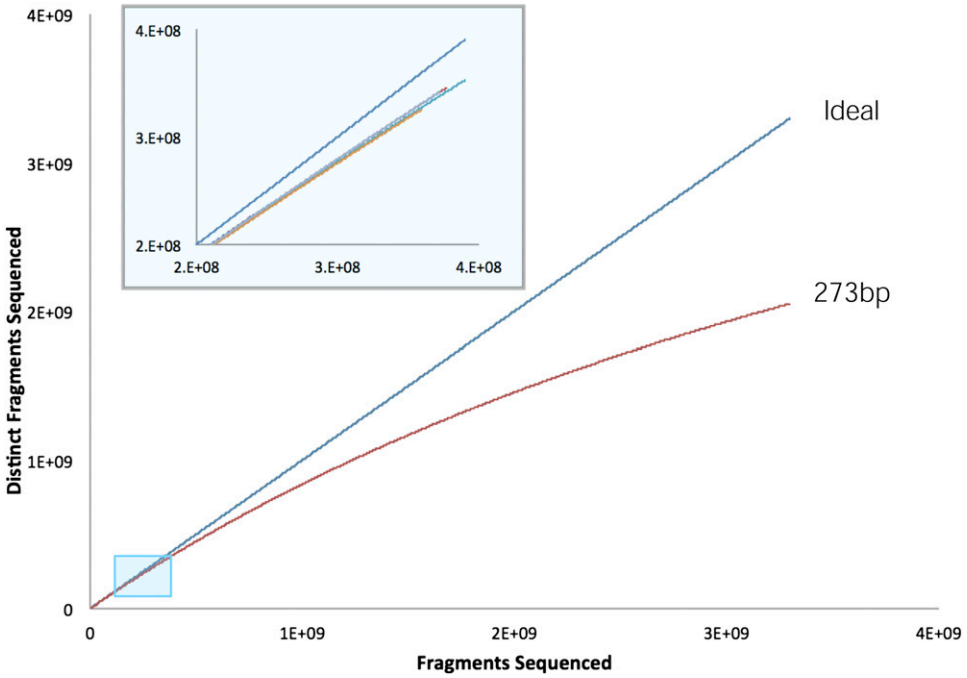

Figure 4 Library complexity curves quantify library complexity and the diminishing returns as sequencing progresses. Library complexity is an estimate of the number of distinct molecules in the library (Daley and Smith 2013). The convexly shaped complexity curve plots the number of distinct molecules observed against the number of sequenced reads. Only our deeply sequenced library exhibited diminishing returns. Libraries sequenced to lower depth are shown in the inset.

Our sequencing strategy used the lowest-cost and most accurate data—the reads from the HiSeq platform—to provide the bulk of the k-mers required for error correction. We then generated super-reads from all error-corrected reads, including the longer reads from the GAIIx and MiSeq instruments. As expected, the GAIIx and MiSeq reads contributed a greater proportion of maximal super-reads. The optimal mix of read lengths is difficult to determine and likely varies for every genome.

Incorporating diploid mate pair sequences

We obtained 1726 million paired reads from 48 mate pair libraries with insert sizes of ~1300–5500 bp (Table 2 and *Appendix A*, Table A1). After filtering to remove reads that failed to match the haploid DNA, 1156 million reads (69%) remained. Another 70 million paired reads that failed to contain a biotin junction and 171 million paired reads identified as duplicates were removed. Finally, we eliminated pairs in which either read was shorter than 64 bp, the minimum read length for the CABOG assembler. The final set of jumping reads contained 540 million reads (270 million pairs), which is ~37-fold physical coverage of the genome (*Appendix B*, Table B2). Notably, there was little variation in the rate at which reads were filtered due to error correction and haplotype differences as the library length increased.

Fosmid DiTags

Particularly useful for establishing long-range links in an assembly are paired reads from the ends of fosmid (Kim *et al.* 1992) or BAC clones (Shizuya *et al.* 1992), which span tens of kilobases. Long-range paired-end data are even more critical for second-generation sequencing projects (Schatz *et al.* 2010), where read lengths are significantly shorter than they are for first-generation sequencing projects. Recent reports demonstrate that mammalian genome assemblies containing high-quality fosmid DiTags can attain contiguity nearly as well as traditional Sanger-based assemblies (Gnerre *et al.* 2011).

We generated 46 million pairs of DiTag reads (*Appendix B*, Table B3). Read pairs from DiTag libraries are subject to the same artifacts as those from mate pair libraries, namely duplicate molecules and nonjumping pairs. Application of the aforementioned MaSuRCA filtering procedures for jumping libraries was applied and yielded ~4.5 million distinct mapped read pairs for downstream overlap layout consensus assembly. The largest reduction (47% of reads removed) was due to the removal of duplicate read pairs, a result of the limited library complexity (*Appendix B*, Table B3). The filtering statistics also indicated that the nick translation protocol was the most efficient at generating distinct DiTag read pairs. To obtain an unbiased estimate of the length distribution of these libraries, we incorporated a set of *D. melanogaster* fosmids into select pools at 1% concentration prior to DiTag construction. The median insert size, measured after removal of nonjumping and chimeric outliers, was 35.7 kbp. Using a genome-size estimate of 22 Gbp for loblolly pine, this represents 7.3× physical coverage of the haploid genome.

While the inclusion of fosmid DiTag libraries was helpful, only approximately one-third of the V1.0 MaSuRCA assembled genome scaffolds (by length) contained the two or more DiTags required to make a long-range link. This leads us to conclude that deeper fosmid DiTag coverage could significantly boost scaffold contiguity in future assemblies.

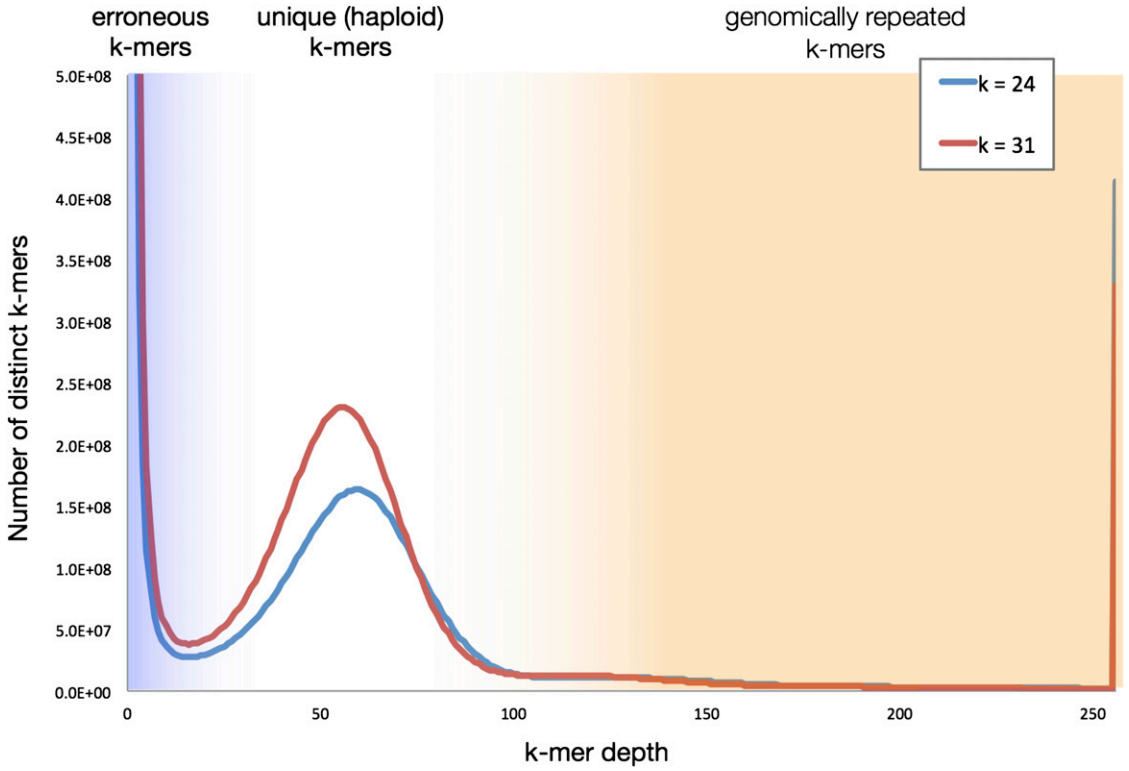

Figure 5 Plot of k-mer counts in the haploid paired-end WGS sequence. After counting all 24- and 31-mers in the QuORUM error-corrected reads, we plotted the number of distinct k-mers (*y*-axis) that occur exactly X times (*x*-axis). For our haploid data, each plot has a single primary peak at the X-value corresponding to the depth of coverage of the 1N genome.

Additional scaffolding procedures for final assembly

To produce the final assembly (version 1.01) from the output of MaSuRCA (version 1.0), we employed two additional strategies that improved the length and accuracy of some of the scaffolds. First, we ran an independent WGS assembly using SOAPdenovo2, including its gap-closing module (Luo *et al.* 2012). We used *all* haploid paired-end data for the contig assembly and added the mate pair and DiTag libraries for the scaffolding step only. Following sequencing and quality control, the adapter sequence was removed from reads using cutadapt (Martin 2011), and the input data were filtered to remove sequences similar to the chloroplast reference sequence (Parks *et al.* 2009). After error correction, 11.44 billion reads were used for contig assembly and 11.98 billion reads for scaffolding. We used a k-mer length of 79 for contig assembly.

Although the SOAPdenovo2 contigs are much smaller than in the 1.0 assembly, with an N50 size of just 687 bp, the scaffolds are larger, with an N50 size of 54.7 kbp. Therefore, we ran the SOAPdenovo2 scaffolder again, using the MaSuRCA scaffolds as input "contigs" and using a conservatively filtered set of mate pairs as linking information. This produced a new set of 8.34 million scaffolds with an N50 size of 86.8 kbp, substantially larger than the original value.

As discussed above, the DiTag sequences prior to MaSURCA processing contained a moderate number of nonjunction read pairs (*Appendix B*, Table B3). To improve the assembly fidelity during this step, we broke all newly created intrascaffold links for which all supporting mates originated from a DiTag library. The resulting modified assembly had 14.4 million scaffolds, with an N50 size of 64.6 kbp, still over twice as large as the original scaffold N50 size.

We used transcripts assembled *de novo* from RNA-seq reads and predicted to encode full-length proteins (see National Center for Biotechnology Information BioProject 174450) to identify scaffolds that could be joined or rearranged to be congruent with transcripts. Because transcripts can be assembled erroneously, we followed a conservative strategy that primarily linked together scaffolds that contained a single gene.

We aligned the 87,241 assembled transcripts to all scaffolds in the modified assembly described above, using the *nucmer* program in the MUMmer package (Kurtz *et al.* 2004), and collecting only those transcripts with at least two exact matches of 40 bp or longer. This yielded 68,497 transcripts mapping to 47,492 different scaffolds.

Alignments between transcripts and the WGS assembly were checked for consistency, looking for instances where the scaffold would require an inversion or a translocation to make it align. We found 1348 inconsistencies, 2% of the total. These could represent errors in either the *de novo* transcriptome assembly or the WGS assembly, and they will be investigated and corrected if necessary in future assembly releases.

We scanned the alignments looking for transcripts that spanned two distinct scaffolds. For each transcript, we sorted

Table 3 Comparison of *P. taeda* whole-genome assemblies before and after additional scaffolding

	P. taeda 1.0 (before)	*P. taeda* 1.01 (after)
Total sequence in contigs (bp)	20,148,103,497	20,148,103,497
Total span of scaffolds (bp)	22,564,679,219	23,180,477,227
N50 contig size (bp)	8,206	8,206
N50 scaffold size (bp)	30,681	66,920
No. of contigs >500 bp	4,047,642	4,047,642
No. of scaffolds >500 bp	2,319,749	2,158,326

The contigs are the same for both assemblies. N50 statistics use a genome size of 22×10^9.

the scaffolds aligned to it and created a directed link between each pair of scaffolds. The links were weighted using the sum of the alignment lengths. If multiple transcripts aligned to the same pair of scaffolds, the link between them was given the sum of all the weights. (This can happen when, for example, two transcripts represent alternative splice variants of the same gene.) We rebuilt the scaffolds using these new links in order of priority from strongest (greatest weight) to weakest, checking for circular links at each step and discarding them.

This transcript-based scaffolding step linked together 31,231 scaffolds into 9170 larger scaffolds, slightly increasing the N50 size and substantially improving the number of transcripts that align to a single scaffold. The final assembly, version 1.01, has 14.4 million scaffolds spanning 23.2 Gbp. Excluding gaps, the total size of all contigs is 20.15 Gbp. The largest scaffold is 8,891,046 bp. The scaffold N50 size is 66,920 bp, based on a 22 Gbp total genome size (Table 3).

Assessing assembly contiguity and completeness

We used the CEGMA pipeline (Parra *et al.* 2007) to assess the contiguity of both the V1.0 and V1.01 assemblies with respect to a set of highly conserved eukaryotic genes. CEGMA uses a set of conserved protein families from the Clusters of Orthologous Groups database (Tatusov *et al.* 2003) to annotate a genome by constructing DNA-protein alignments between the proteins and the assembled scaffolds. Of 248 highly conserved protein families, CEGMA found full-length alignments to the V1.0 assembly for 113 (45%). Including partial alignments brings this total to 197 (79%). In the V1.01 assembly, CEGMA found 185 (75%) full-length alignments and 203 (82%) full-length and partial alignments. There was little change in the total number of alignments between V1.0 and V1.01. However, the fraction of all alignments to the assembly that were full length increased from 57% in the V1.0 assembly to 91% in the V1.01 assembly, further quantifying the biological utility of the additional scaffolding phase. There are a number of possible explanations for the discrepancy between the coverage and the observed rates of conserved gene annotations, principally assembly fragmentation and potential ascertainment bias in the determination of the gene set.

Comparison of fosmid and whole-genome assemblies

For validation purposes we constructed, sequenced, and assembled a large pool of ~4600 fosmid clones (*Appendix A*). The assembly contained 3798 contigs longer than 20,000 bp (>50% of the fosmid insert in each case), which we used to check the completeness and quality of the WGS assembly. Because the fosmids were selected at random, the amount of sequence covered by the WGS assembly should provide an estimate of how much of the entire genome has been captured in that assembly. The 3798 contigs contain 109 Mbp, or ~0.5% of the entire genome.

We aligned contigs to the WGS assembly (V1.0/V1.01 contigs) using MUMmer (Kurtz *et al.* 2004) and found that all but 4 of the 3798 contigs aligned across nearly their entire length at an average identity of 99%. Nearly all (98.63%) of the combined length of the fosmid pool contigs is covered by alignments to the WGS assembly. The unaligned fraction consists of the 4 contigs that failed to align (117,681 bp) plus the unaligned portions of the remaining contigs. Assuming that the fosmids represent a random sample, this suggests that the WGS assembly covers >98% of the whole genome (with the reservation that hard-to-sequence regions might be missed by both the WGS and fosmid sequencing approaches, and thus not counted in this evaluation).

Fosmid haplotype comparisons

We expect that half of the fosmid contigs would correspond to the same haplotype as the aligned WGS contig. Thus while many contigs should be close to 100% identical, contigs from the alternative haplotype (untransmitted) would show a divergence rate corresponding to the differences between haplotypes, estimated to be 1–2%. We considered the possibility that the fosmid contigs would thus fall into two distinct bins: near-identical and 1–2% divergent. However, relatively few contigs showed this degree of divergence: 2120 contigs (56%) were 99.5% or more identical, and the others matched the WGS contigs at a range of identities, with the vast majority >98% identical (*Appendix A*, Figure A1). Only 189 contigs (5%) were <95% identical, with the lowest identity at 91%. This suggests that divergence between the two parental haplotypes is <1% on average, with a small number of highly divergent regions.

Conclusions

The sequencing and assembly strategy described here, of the largest genome to date, resulted in a haploid assembly composed of 20.15 billion base pairs. By many measures, it is the most contiguous and complete draft assembly of a conifer genome (*Appendix C*). This project required a close

coupling between sequencing and assembly strategy and the development of new computational methods to address the challenges inherent to the assembly of mega-genomes.

Conifer genomes are filled with massive amounts of repetitive DNA, mostly transposable elements (Nystedt *et al.* 2013; Wegrzyn *et al.* 2013, 2014), which might have presented a major barrier to successful genome assembly. Fortunately, most of these repeats are relatively ancient, so their accumulated sequence differences allowed the assembly algorithms to distinguish individual copies.

A major concern in assembly of diploid genomes is the degree of divergence between the two parental copies of each chromosome. Even small differences between the chromosomes can cause an assembler to construct two distinct contigs, which can be indistinguishable from two-copy repeat sequences. Conifers offer an elegant biological solution to this problem by producing megagametophytes with a single copy of each chromosome, from which sufficient DNA can be extracted to cover the entire genome.

While the current assembly represents an important landmark, and will serve as a powerful resource for biological analyses, it remains incomplete. To further increase sequence contiguity, we plan to generate additional long-range paired reads, particularly DiTag pairs, that should substantially increase scaffold lengths. Noting the success of the smaller fosmid pool assemblies presented here, inexpensive sequencing allows for the possibility of sequencing many more of these pools and merging the long contigs from these assemblies into the WGS assembly (Zhang *et al.* 2012; Nystedt *et al.* 2013). In parallel with efforts to improve the next *P. taeda* reference sequence, we are gathering a large set of genetically mapped SNPs that will allow placement of scaffolds onto the 12 chromosomes, providing additional concrete physical anchors to the assembly. Finally, we note that ongoing efforts to expand the diversity of conifer genome reference sequences will provide a solid foundation for comparative genomics, thus improving the assemblies and their annotations.

Acknowledgments

This work was supported in part by U.S. Department of Agriculture/National Institute of Food and Agriculture (2011-67009-30030). We wish to thank C. Dana Nelson at the USDA Forest Service Southern Research Station for providing and verifying the genotype of target tree material.

Note added in proof: See Wegrzyn *et al.* 2014 (pp. 891–909) in this issue for a related work.

Literature Cited

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

Bai, C., W. S. Alverson, A. Follansbee, and D. M. Waller, 2012 New reports of nuclear DNA content for 407 vascular plant taxa from the United States. Ann. Bot. (Lond.) 110: 1623–1629.

Bierhorst, D. W., 1971 *Morphology of Vascular Plants*. Macmillan, New York.

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.

Bowe, L. M., G. Coat, and C. W. Depamphilis, 2000 Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Proc. Natl. Acad. Sci. USA 97: 4092–4097.

Daley, T., and A. D. Smith, 2013 Predicting the molecular complexity of sequencing libraries. Nat. Methods 10: 325–327.

Frederick, W. J., S. J. Lien, C. E. Courchene, N. A. DeMartini, A. J. Ragauskas *et al.*, 2008 Production of ethanol from carbohydrates from loblolly pine: a technical and economic assessment. Bioresour. Technol. 99: 5051–5057.

Fuchs, J., G. Jovtchev, and I. Schubert, 2008 The chromosomal distribution of histone methylation marks in gymnosperms differs from that of angiosperms. Chromosome Res. 16: 891–898.

Gnerre, S., I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton *et al.*, 2011 High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA 108(4): 1513–1518.

Kim, U.-J., H. Shizuya, P. J. de Jong, B. Barren, and M. I. Simon, 1992 Stable propagation of cosmid-sized human DNA inserts in an F factor based vector. Nucleic Acids Res. 20: 1083–1085.

Kovach, A., J. L. Wegrzyn, G. Parra, C. Holt, G. E. Bruening *et al.*, 2010 The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics 11(1): 420.

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.

Lander, E. S., and M. S. Waterman, 1988 Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2(3): 231–239.

Li, R., H. Zhu, J. Ruan, W. Qian, X. Fang *et al.*, 2010 De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 20(2): 265–272.

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 DOI:10.1186/2047-217X-1-18.

Mackay, J., J. F. Dean, C. Plomion, D. G. Peterson, F. M. Cánovas *et al.*, 2012 Towards decoding the conifer giga-genome. Plant Mol. Biol. 80: 555–569.

Magallon, S., and M. J. Sanderson, 2005 Angiosperm divergence times: the effects of genes, codon positions, and time constraints. Evolution 59: 1653–1670.

Marçais, G., and C. Kingsford, 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27(6): 764–770.

Marcais, G., J. Yorke, and A. Zimin, 2013 QuorUM: an error corrector for Illumina reads, arXiv:1307.3515 [q-bio.GN].

Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17(1): 10.

McKeand, S., T. Mullin, T. Byram, and T. White, 2003 Deployment of genetically improved loblolly and slash pines in the south. J. For. 101(3): 32–37.

Miller, J. R., A. L. Delcher, S. Koren, E. Venter, B. P. Walenz *et al.*, 2008 Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 24(24): 2818–2824.

Mirov, N. T., 1967 *The Genus Pinus*. Ronald Press, New York.

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., and A. Kremer, 2011 Forest tree genomics: growing resources and applications. Nat. Rev. Genet. 12(2): 111–122.

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

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

O'Brien, I. E. W., D. R. Smith, R. C. Gardner, and B. G. Murray, 1996 Flow cytometric determination of genome size in Pinus. Plant Sci. 115: 91–99.

Parks, M., R. Cronn, and A. Liston, 2009 Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 7(1): 84.

Parra, G., K. Bradnam, and I. Korf, 2007 CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23(9): 1061–1067.

Peterson, D. G., S. R. Schulze, E. B. Sciara, S. A. Lee, J. E. Bowers *et al.*, 2002 Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res 12: 795–807.

Peterson, D. G., J. P. Tomkins, D. A. Frisch, R. A. Wing, and A. H. Paterson, 2000 Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide. J. Agric. Genomics 5: 1–100.

Ross, M. G., C. Russ, M. Costello, A. Hollinger, N. J. Lennon *et al.*, 2013 Characterizing and measuring bias in sequence data. Genome Biol. 14(5): R51.

Schatz, M. C., A. L. Delcher, and S. L. Salzberg, 2010 Assembly of large genomes using second-generation sequencing. Genome Res. 20(9): 1165–1173.

Schultz, R. P., 1997 Loblolly pine: the ecology and culture of loblolly pine (*Pinus taeda L.*), in *Agriculture Handbook, Washington, (713)*. U.S. Department of Agriculture, Forest Service. Washington, D.C.

Shizuya, H., B. Birren, U.-J. Kim, V. Mancino, T. Slepak *et al.*, 1992 Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc. Natl. Acad. Sci. USA 89(18): 8794–8797.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin *et al.*, 2003 The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4(1): 41.

Wang, N., M. Thomson, W. J. A. Bodles, R. M. M. Crawford, H. V. Hunt *et al.*, 2013 Genome sequence of dwarf birch (Betula nana) and cross-species RAD markers. Mol. Ecol. 22: 3098–3111.

Wegrzyn, J., B. Y. Lin, J. J. Zieve, W. M. Dougherty, P. J. Martínez-García *et al.*, 2013 Insights into the loblolly pine genome: characterization of BAC and fosmid sequences. PLoS ONE 8: e72439.

Wegrzyn, J. L., J. D. Liechty, K. A. Stevens, L. Wu, C. A. Loopstra *et al.*, 2014 Unique Features of the Loblolly Pine (*Pinus taeda* L.) Megagenome Revealed Through Sequence Annotation. Genetics 196: 891–909.

Williams, L. J. S., D. G. Tabbaa, N. Li, A. M. Berlin, T. P. Shea *et al.*, 2012 Paired-end sequencing of Fosmid libraries by Illumina. Genome Res. 22(11): 2241–2249.

Zhang, G., X. Fang, X. Guo, L. Li, R. Luo *et al.*, 2012 The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490(7418): 49–54.

Zimin, A. V., G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg *et al.*, 2013 The MaSuRCA genome assembler. Bioinformatics 29: 2669–2677.

Zonneveld, B. J. M., 2012 Genome sizes of 172 species, covering 64 out of the 67 genera, range from 8 to 72 picogram. Nord. J. Bot. 30: 490–502.

Communicating editor: M. Johnston

Appendix A: Fosmid Pool Sequencing and Assembly for Validation

Fosmid Pool Construction

Independently of the whole-genome shotgun libraries, a library of fosmid particles was constructed in the vector pFosTH. pFosTH is derived from pFosDT5.4 by replacing the cloning site region, including the Illumina primers and Nb.BbvCl sites as present on a small PI-*Sce*I fragment, with a synthetic DNA fragment consisting of the triple-helix motif sites 1 and 2 flanking the unique *Eco*47-III cloning site. Extracted genomic DNA from diploid needle tissue was sheared to an average size of ~40 kbp, converted to blunt ends, and size-purified by pulsed-field electrophoresis. This was followed by ligation to excess vector (*Eco*47-III digested, dephosphorylated ends) and λ-phage packaging (extracts from mcrA, -B, -C strains) to create a particle library (see Figure 2). The particle library was then titered, portioned, and transduced into *E. coli*. After 1 hr of recovery, the cells were frozen and stored as glycerol stocks directly from the liquid media to minimize duplication. This procedure resulted in a primary library of ~7 million *E. coli* colony forming units, each containing a single type of fosmid. Smaller subpools of fosmids were also created. After titration of the frozen stock, specific aliquots of the main pool were streaked on LB chloramphenicol agarose plates, incubated to form primary colonies, then recovered by scraping off the agarose, suspended in media, pooled, and used to inoculate liquid cultures for DNA amplification and purification. The resulting defined (low complexity) DNA samples were sequenced and assembled using independent software for the purposes of validating the WGS assembly.

Libraries and Sequencing

From the primary fosmid library described above, 11 smaller subpools of ~580 ± 10% *E. coli* colonies were created. The fosmid DNA of the 11 pools of harvested bacterial colonies was then amplified in vivo. Fosmid DNA was subsequently purified and digested with the homing endonuclease PI-*Sce*I, which has a 35-bp recognition site. With the isolated DNAs quantified by PicoGreen, the purified insert DNAs were portioned out to create an equal molar super-pool of all 11 component fosmid pools, as well as a set of 3 nested equal molar super-pools of 8, 4, and 2 small fosmid pools.

The fosmid DNA was then subsequently processed into Illumina libraries using the reagents and protocols as previously described. The largest super pool was converted to an Illumina long-insert mate pair library while the smaller nested super pools were converted into short-insert paired-end libraries. Paired-end libraries were subsequently sequenced on an Illumina GAIIx (SCS version 2.9.35, RTA version 1.9.35). The two smallest fosmid pools were primarily used for calibration purposes and are therefore not considered further.

Subsequent ungapped single-seed alignment using Illumina's CASAVA pipeline (version 1.7.0) to the *E. coli* and fosmid vector genomes was used to determine insert-size statistics. The paired-end library had a mean fragment length of 261 bp while the mate pair library had a median fragment length of 3345 bp (Table A1). The reported rates of the noncanonical paired-end alignment orientations were also consistent with high-quality libraries.

Table A1 Libraries and read statistics for sequences contributing to our largest fosmid pool assembly

Library	Pool size (fosmids)	Read lengths (bp)	Mean fragment length (bp)	No. of pairs	No. of pairs after cleaning
Paired-end	4600 ± 10%	160, 156	261	90,392,267	73,325,963
Mate pair	6400 ± 10%	160, 156	3345 ± 151	11,978,560	8,390,844

Fosmid Assembly

We cleaned the raw read data by identifying and filtering out contaminating sequence from reads that contained *E. coli* DNA (1.0% in the short library, 1.76% in the jumping library), fosmid vector (17.9% in the short library, 29.3% in the jumping library), or adapter sequence (0.4% in the short library, 1.2% in the jumping library). After experiments with different assembly parameters and multiple assemblers, we concluded that slightly better assemblies resulted when all reads were

Table A2 Assembly statistics for the largest pool of fosmids

	No.	Sum of lengths	Vector on either end
Contigs >20,000 bp	3798	109,412,316	623
Contigs >30,000 bp	1651	56,742,427	585
Scaffolds >20,000 bp	5323	171,852,787	2001
Scaffolds >30,000 bp	3719	131,752,852	1911

The final column shows the number of contigs/scaffolds for which one or both ends contained fosmid vector sequence, indicating that the contig extended all the way to the end of the insert.

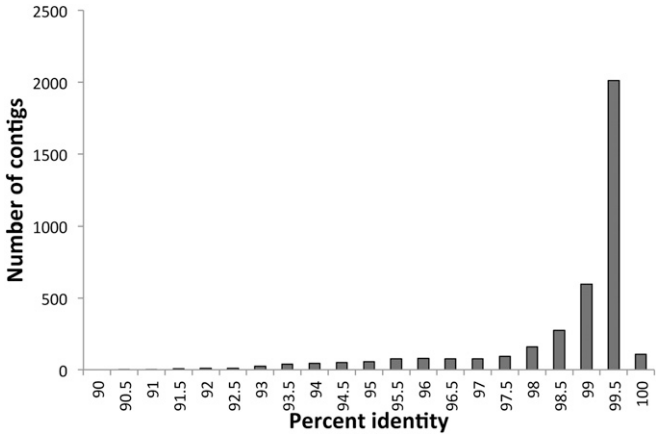

Figure A1 The percentage of identity of long contigs (>20,000 bp) from the fosmid pool assembly when aligned to the WGS assembly. A total of 109 contigs were 100% identical, and another 2011 were between 99.5 and 100% identical.

trimmed to 125 bp, thereby removing low-quality bases from the 3′ ends of the reads. We assembled the fosmid pool with *SOAPdenovo* (release 1) (Li *et al.* 2010) using a k-mer size of 63.

Our largest pool contained ∼4600 fosmids. If the expected size of the loblolly DNA in the fosmids was 36 kbp (see *Results*), then the total pool size was 166 Mbp, or <1% of the total genome size for loblolly pine. Thus, we would expect ∼1% of the randomly chosen fosmids to overlap another fosmid in the same pool.

The total amount of sequence from this pool was 81.7 million read pairs (Table A1), which provided ∼120× coverage, on average, for each fosmid. The resulting pooled assembly showed remarkably high contiguity, with 3719 scaffolds spanning >30 kbp (∼80% of the maximum length for a fosmid) and 5323 scaffolds longer than 20 kbp, as shown in Table A2.

The high level of contiguity for the fosmid pool assembly is likely a result of the haploid nature of the fosmids combined with the use of a jumping library containing 3.5-kbp paired reads.

For validation purposes, the contigs >20,000 bp in Table A2 were aligned against the WGS assembly using MUMmer. Detailed results from these alignments are presented above. Figure A1 gives additional detail as to the distribution of the percentage of identity of each fosmid contig aligned to the WGS assembly.

Appendix B: Additional Library Statistics

Haploid Paired-End Libraries

We sequenced 11 paired-end libraries derived from the DNA of a single megagametophyte, using three Illumina platforms (HiSeq2000, GAIIx, and MiSeq). Table B1 shows that we obtained >1.4 Tbp of short-insert, paired-end, high-quality sequence, corresponding to nearly 64-fold coverage of the loblolly pine genome. As expected from the narrow and consistent distributions in Figure 2, the empirical fragment size coefficient of variation for the 11 haploid paired-end libraries was uniformly small, between 4 and 6%.

Mate Pair Libraries

We sequenced 48 mate pair libraries, with insert sizes ranging from ∼1300 to 5500 bp. Table B2 groups these by insert size and summarizes the outcome of the MaSuRCA jumping library filters on these as well as the estimated clone coverage presented to the downstream overlap assembler.

DiTag Libraries

We sequenced nine DiTag libraries in 40 fractional GAIIx lanes, yielding 46 million read pairs, summarized in Table B3.

Table B1 Selected statistics for haploid paired-end sequence data by platform and fragment size

Platform	Median fragment size (bp)	Sequenced Gbp	Error-corrected Gbp	Read 1 expected error-corrected length (bp)	Read 2 expected error-corrected length (bp)	Expected gap size (bp)	Gaps filled (%)	Maximal super-reads (%)
GAIIx 160+156	209	2.3	2.2	157.0	151.3	−99.3	96	6
	220	2.9	2.8	157.1	151.3	−88.4	96	7
	234	3.3	3.1	157.6	149.7	−73.3	95	8
	246	2.3	2.2	157.8	151.7	−63.5	94	9
	260	2.4	2.3	157.8	150.9	−48.7	94	9
	273	263	252	157.5	150.0	−34.5	92	12
	285	2.4	2.3	157.8	151.4	−24.2	92	12
	325	2.2	2.1	157.5	149.3	18.3	91	15
	441	1.9	1.8	157.1	141.7	142.2	84	18
	565	1.5	1.4	156.9	140.5	267.6	76	22
	637	24.0	1.1	156.6	134.8	345.6	71	23
HiSeq 2x125	273	326.0	315.4	123.8	122.0	27.1	83	9
HiSeq 2x128	220	96.6	92.9	126.5	124.1	−30.6	90	6
	234	59.5	57.7	127.2	125.4	−18.6	87	6
	246	61.6	59.6	126.6	125.3	−5.9	87	7
	260	100.2	97.1	127.1	125.1	−2.3	84	9
	285	95.6	91.7	126.4	123.2	35.4	86	10
	325	92.3	88.9	126.2	124.3	74.5	80	12
	441	35.8	34.4	126.2	124.0	190.8	73	15
	565	43.4	41.4	125.7	122.2	317.1	68	19
MiSeq 2x255	325	2.5	2.38	248.2	238.8	−162.0	95	14
	441	2.0	1.85	247.5	232.2	−38.6	89	18
	565	1.6	1.44	246.9	246.9	96.0	83	23
	637	1.2	1.03	246.6	246.6	177.9	79	24

The percentage of reads that are gap filled is from the total number of reads entering gap fill, and the percentage of reads becoming maximal super-reads is from those passing error correction. As fragment size increases, so does the expected gap size. Both the error-corrected read length and the rate at which the gaps are successfully filled are decreasing functions of the fragment size. Nevertheless, the overall trend is that participation in the formation of super-reads increases with insert size. Only the paired-end reads that are successfully transformed into maximal super-reads are passed to subsequent assembly steps.

Table B2 Summary of outcomes from preprocessing stages for long fragment libraries, with lengths ranging from 1000 to 5500 bp

Estimated jump size (bp)	Library count	Reads sequenced	After error correction and mapping[a]	Nonjunction pairs	Redundant reads	Final reads with both >63 bp	Estimated clone coverage
1000–1999	5	127.3	85.3 (67%)	5.6 (7%)	9.7 (12%)	42.1	1.5×
2000–2999	16	651.9	430.0 (66%)	26.6 (6%)	43.0 (11%)	207.4	11.9×
3000–3999	18	705.4	474.4 (67%)	26.4 (6%)	88.1 (20%)	213.2	16.2×
4000–4999	6	186.6	127.8 (69%)	6.6 (5%)	13.6 (11%)	63.8	6.3×
5000–5500	3	55.3	38.2 (69%)	5.7 (15%)	19.8 (61%)	57.6	2.1×

Read counts are given in millions.
[a] Reads that successfully passed error correction and mapped to the haplotype of the target megagametophyte.

Table B3 Summary of DiTag sequencing

Library	Lanes sequenced	Median reads sequenced per lane (millions)	Median reads after error correction and mapping (millions)	Median nonjunction reads (millions)	Median redundant reads (millions)	Median final reads with both >63 bp (millions)	Median clone coverage per lane (sum)
N1	6	2.5	1.5 (57%)	0.2 (16%)	0.5 (37%)	0.3 (12%)	0.3× (1.8×)
N2	6	1.8	1.2 (68%)	0.03 (3%)	0.6 (53%)	0.3 (15%)	0.2× (1.2×)
N3	6	0.9	0.6 (59%)	0.05 (8%)	0.2 (46%)	0.1 (13%)	0.1× (0.6×)
N4	3	1.5	1.0 (66%)	0.1 (5%)	0.3 (37%)	0.3 (22%)	0.3× (0.9×)
R1	2	4.1	1.8 (44%)	0.7 (39%)	0.9 (85%)	0.2 (4%)	0.1× (0.2×)
R2	2	2.9	1.1 (39%)	0.4 (33%)	0.5 (63%)	0.1 (5%)	0.1× (0.2×)
R3	2	5.7	2.2 (39%)	1.0 (45%)	1.2 (97%)	0.2 (3%)	0.1× (0.2×)
All lanes	40	93	39 (42%)	7 (18%)	15 (47%)	9 (10%)	(7.5×)

Four nick translation libraries (N1–N4) and three endonuclease digestion libraries (R1–R3) are detailed. Each library was sequenced in replicate in a number of lanes. All values reported other than the totals are medians. To allow for a more direct comparison between libraries, median values per lane are used.

Appendix C: Conifer Assembly Statistics

The loblolly pine genome (Neale *et al.* 2014), whose sequencing and assembly we described here, joins two other conifer genome sequences, Norway spruce (Nystedt *et al.* 2013) and white spruce (Birol *et al.* 2013), as a foundation of conifer genomics. We report selected assembly statistics from these three genomes for comparison (Table C1 and Figure C1).

Table C1 Loblolly pine V1.01 assembly compared to contemporary draft conifer genomes

Species	Loblolly pine (*Pinus taeda*)	Norway spruce (*Picea abies*)	White spruce (*Picea glauca*)
Cytometrically estimated genome size (Gbp)	21.6[a]	19.6[b]	15.8[c]
Total scaffold span (Gbp)	22.6	12.3	23.6
Total contig span[d](Gbp)	20.1	12.0	20.8
Referenced genome-size estimate (Gbp)	22	18	20
N50 contig size (kbp)	8.2	0.6	5.4
N50 scaffold size (kbp)	66.9	0.72	22.9
No. of scaffolds	14,412,985	10,253,693	7,084,659
Annotation of 248 conserved CEGMA genes (Parra *et al*. 2007)	185 (74%) complete 203 (82%) complete + partial, 91% annotated full length	124 (50%) complete 189 (76%) complete + partial, 66% annotated full length	95 (38%) complete 184 (74%) complete + partial, 52% annotated full length

N50 contig and scaffold sizes are based on the estimated genome size listed in the table.
[a] O'Brien *et al.* (1996).
[b] Fuchs *et al.* (2008).
[c] Bai *et al.* (2012).
[d] Determined as the number of non-N characters in the published reference sequence.

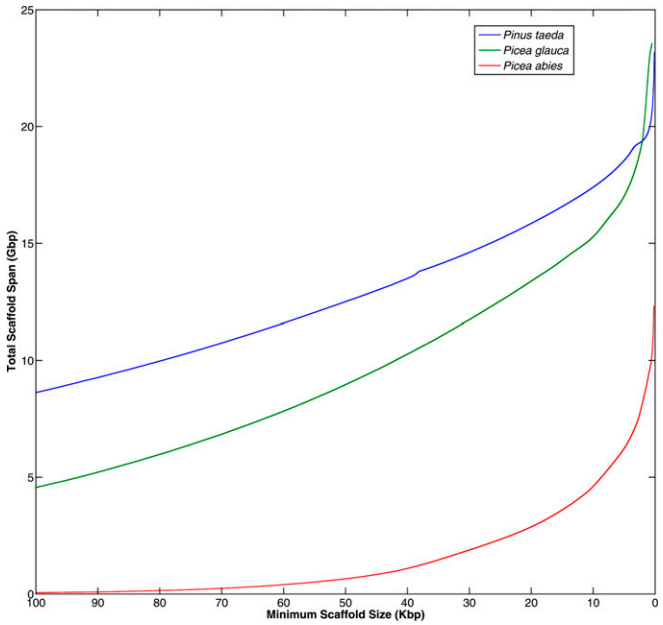

Figure C1 The contiguity of the loblolly pine v1.01 assembly is compared to contemporary draft conifer assemblies. Total scaffold span is plotted against a minimum scaffold size threshold. Loblolly pine is relatively more complete when considering large-gene-sized (> 10 kbp) scaffolds. This is reflected in the CEGMA results (Table C1).