Whole-exome targeted sequencing of the uncharacterized pine genome

Leandro G. Neves¹, John M. Davis¹,²,³, William B. Barbazuk¹,³,⁴,* and Matias Kirst¹,²,³,*

¹Graduate Program in Plant Molecular and Cellular Biology, University of Florida, Gainesville, FL 32611, USA,
²School of Forest Resources and Conservation, University of Florida, Gainesville, FL 32611, USA,
³University of Florida Genetics Institute, University of Florida, Gainesville, FL 32611, USA, and
⁴Department of Biology, University of Florida, Gainesville, FL 32611, USA

Received 1 October 2012; revised 26 April 2013; accepted 26 March 2013; published online 4 April 2013.
*For correspondence (e-mails bbarbazuk@ufl.edu and mkirst@ufl.edu).

SUMMARY

The large genome size of many species hinders the development and application of genomic tools to study them. For instance, loblolly pine (Pinus taeda L.), an ecologically and economically important conifer, has a large and yet uncharacterized genome of 21.7 Gbp. To characterize the pine genome, we performed exome capture and sequencing of 14,729 genes derived from an assembly of expressed sequence tags. Efficiency of sequence capture was evaluated and shown to be similar across samples with increasing levels of complexity, including haploid cDNA, haploid genomic DNA and diploid genomic DNA. However, this efficiency was severely reduced for probes that overlapped multiple exons, presumably because intron sequences hindered probe:exon hybridizations. Such regions could not be entirely avoided during probe design, because of the lack of a reference sequence. To improve the throughput and reduce the cost of sequence capture, a method to multiplex the analysis of up to eight samples was developed. Sequence data showed that multiplexed capture was reproducible among 24 haploid samples, and can be applied for high-throughput analysis of targeted genes in large populations. Captured sequences were de novo assembled, resulting in 11,396 expanded and annotated gene models, significantly improving the knowledge about the pine gene space. Interspecific capture was also evaluated with over 98% of all probes designed from P. taeda that were efficient in sequence capture, were also suitable for analysis of the related species Pinus elliottii Engelm.

Keywords: conifers, Pinus taeda, Pinus elliottii, sequence capture, target enrichment, genome complexity reduction, genome de novo assembly, technical advance.

INTRODUCTION

Sequencing the large genome of gymnosperms remains a challenge despite the rapid increase in DNA sequencing throughput and cost reductions (Schatz et al., 2010; Baker, 2012). At 15,480 Mbp, the modal genome size of gymnosperms is an order of magnitude larger than that of most sequenced angiosperms, and 6.7 times larger than the human genome (Leitch et al., 2001). The large genome of gymnosperms, and of pines in particular, does not appear to be the consequence of recent whole-genome duplications, but rather derived from retrotransposon activity (Morse et al., 2009). The transcriptome of conifers has also not expanded significantly compared with other vascular plants such as Arabidopsis (Kirst et al., 2003). Thus, coding and regulatory sequence is likely to represent only a minute fraction of a conifer genome – much smaller than that of other species sequenced to date. An alternative to complete sequencing of a large and complex genome is to focus on its coding and regulatory regions, which are of primary interest for population and quantitative genetics studies, and comparative genomics.

Reducing genome complexity for sequencing single-copy coding and regulatory regions can be achieved with methylation filtration (Rabinowicz et al., 1999) and C₀t
filtration (Peterson et al., 2002), and the utility of these methods has been well demonstrated in plants such as maize (Whitelaw et al., 2003; Barbazuk et al., 2005). Both methods rely on specific properties of repetitive DNA to exclude it from a genome sample. However, these approaches are labor intensive and relatively inconsistent when common regions of the genome are targeted for sequencing in multiple samples. Alternative methods of genome complexity reduction that can retrieve targeted sequences, including coding regions, were introduced recently. Initially, genomic DNA was hybridized to microarrays that contained oligonucleotides complementary to the target sequences, followed by elution and sequencing of the hybridized DNA (Albert et al., 2007). This approach has been further modified to use biotin-labeled oligonucleotides that can be retrieved by streptavidin beads, thus enabling solution-based hybridization (Gnirke et al., 2009). These methods are referred to hereafter as ‘sequence capture’. Sequence capture has been used extensively in human genome studies to detect nucleotide and structural variants (Gnirke et al., 2009; Walsh et al., 2010) and its recent use in maize demonstrated its utility in plants (Fu et al., 2010). A targeted sequence capture approach applied to uncharacterized large genomes, such as those of conifers, holds the promise of expanding the knowledge of coding and regulatory sequences, and support the identification of sequence variation.

Applying sequence capture to conifers presents challenges that have not been encountered previously. First, it is unclear how efficient sequence capture will be in recovering a much smaller fraction of the genome, compared with previous studies (Gnirke et al., 2009; Zhou and Holliday, 2012). Second, unknown introns may impact hybridization efficiency and capture success, when positioned within a probe designed from transcript data. Here we report the development and synthesis of probes for 14 729 loblolly pine (Pinus taeda L.) UniGenes for sequence capture from genomic sequence. We describe the efficiency of capture, and examine success in light of the fact that much of our probe design was done in the absence of known gene structure. We create a de novo reference underlying the genomic region within which the probes reside, which expands the initial reference and annotates the gene space. We also compare the capture efficiency from cDNA and gDNA from both haploid and diploid tissue; hence, we examine the effect of targets with different genetic complexities on capture efficiency. Additionally, to support processing large numbers of samples, we determine the effect of multiplexing capture reactions on capture efficiency. Finally, we demonstrate how a probe set developed for P. taeda can be efficiently used to capture the target genes in a related species, slash pine (Pinus elliottii Engelm).

RESULTS

Sequence capture probe design from an uncharacterized genome

The efficiency of sequence capture probes designed from transcriptome assemblies may be affected by their position relative to exon-exon boundaries. Expectedly, a probe designed from cDNA sequences that spans two exons will hybridize only partially to genomic DNA. In this study, probes were designed from 15 983 loblolly pine transcript sequences derived from the assembly of expressed sequence tags. For 4500 transcripts, partial genomic sequences were available. Thus, incomplete gene models could be predicted by aligning transcripts to the corresponding genomic sequence with GMAP (Wu and Watanabe, 2005). Boundaries between exons were identified for 3946 transcripts, and probes were designed to avoid these regions. For the remaining 12 037 transcripts, probes were designed to tile the entire transcript sequence, with no overlap between them. A custom pipeline was applied (see Experimental Procedures) to selected probes that minimize hybridization to multiple transcripts, resulting in a final set of 54 773 120-mer oligonucleotides (6.57 Mbp) representing 14 729 transcripts. The average length of each transcript was 660.2 bp. On average, 2.3 probes could be designed for transcripts in which exon-exon boundaries were predicted (3615) and 4.2 probes for the remainder (11 114). Avoiding exon-exon boundaries resulted in shorter sections for which 120-mer probes could be designed and, consequently, fewer probes.

Efficient sequence capture of the loblolly pine genome with different complexities

Sequence capture may be less efficient when applied to conifers such as loblolly pine, because of the low target to genome size ratio. To assess the effect of genome complexity on the efficiency of sequence capture, the 54 773 custom-designed probes were hybridized to: (i) complementary DNA (cDNA) synthesized from megagametophyte mRNA; as well as (ii) haploid genomic DNA from megagametophyte DNA (1N); and (iii) diploid genomic DNA obtained from the needles of loblolly pine (2N). Sanger sequencing of the captured fragments and alignment (BLASTN E-value = 1 x 10^-10) to the reference transcript sequences showed that efficiency was similar regardless of the genome complexity – 71, 87 and 83% of the fragments aligned uniquely to the reference for haploid cDNA, haploid DNA and diploid DNA, respectively (Table 1). To verify the capture efficiency in a larger sample, the haploid DNA sample was also sequenced in a Pacific Biosciences SMRT cell. In total, 25 912 circular consensus sequences (CCS) were generated, of which 65% aligned to the reference at the same threshold used for the Sanger sequences (Table 1).
Further analysis suggested that regions of the genome that contained paralogous genes or related pseudogenes (referred collectively hereafter as ‘paralogs’) were also captured from the haploid and diploid DNA samples. The median proportion of nucleotide difference from the Sanger data to the probes was 0, 6.2 and 8.5% for haploid cDNA, haploid DNA and diploid DNA, respectively (Table 1). Pseudogenes are not likely to be represented in cDNA libraries, but genomic DNA populations may contain these sequence relatives. Our results suggest that paralogs are also captured from the genomic DNA fraction.

Samples can be multiplexed and efficiently captured

Sequence capture is costly, laborious and time-consuming. These obstacles limit its use for quantitative and population genetics studies, and other applications that require resequencing large numbers of individuals. To address these limitations, library preparation was customized to allow multiplexing of up to eight haploid samples in each hybridization. Illumina’s universal adapters were modified by the addition of custom barcodes, and samples with ligated adapters were pooled in equimolar concentrations prior to hybridization to probes. This simple approach reduces cost and increases throughput by almost one order of magnitude relative to the standard single-sample hybridization procedure. The products of a multiplexed reaction were sequenced in a Pacific Biosciences SMRT cell, and on average 3000 captured fragments were characterized from each individual (24 002 total CCS reads). Capture was reduced in multiplexed, compared with non-multiplexed, reactions from 65 to 50% (Table 1), suggesting a trade-off between increased throughput and efficiency. However, this trade-off is compensated by the reduction in cost and increase in throughput of sequence capture.

Sequence capture is reproducible among unrelated samples

Loblolly pine has very high levels of intra-specific genetic diversity (Brown et al., 2004). Such high diversity often hampers the development of molecular assays that require sequence conservation, and could result in variable capture efficiency among unrelated individuals. Here, the

Table 1 Validation of the capture efficiency in pine across different genomic complexity. A sample of the captured fragments (N) was sequenced using Sanger and/or Pacific Bioscience methods

<table>
<thead>
<tr>
<th>Genomic complexity</th>
<th>Sanger</th>
<th>PacBio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Aligned</td>
</tr>
<tr>
<td>Haploid cDNA</td>
<td>75</td>
<td>53 (71%)</td>
</tr>
<tr>
<td>Haploid gDNA</td>
<td>77</td>
<td>67 (87%)</td>
</tr>
<tr>
<td>Diploid gDNA</td>
<td>83</td>
<td>69 (83%)</td>
</tr>
<tr>
<td>Haploid gDNA multiplex</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

*a*WU-BLASTN alignment against UniGene reference at e^-10 stringency.

*b*Sequence capture reaction performed on a pool of eight individually barcoded samples. Aligned results are an average of the eight samples.

Figure 1. Efficiency of multiplexed capture in *Pinus taeda*. Shown is the percentage of the Illumina sequenced reads that align to the predicted targets (Y-axis) when increasing values of mismatch between the reads and the reference is accepted (X-axis). Average and standard deviation are shown for 24 haploid samples.
reproducibility of multiplex sequence capture was assessed by sequencing libraries from haploid DNA of 24 unrelated loblolly pine samples (Figures 1 and 2). A reproducible capture should result in the gene being sequenced at similar depth, across all samples of a population, for the probes targeting the gene. On average, the coefficient of variation for sequencing depth across samples was calculated (X-axis) for the subset of the genes with at least 80% of the samples that show positive depth.

![Variation of sequencing depth between genotypes at the gene level](image)

**Figure 2.** Reproducibility of the multiplexed capture in *Pinus taeda*. The average sequencing depth for target genes was calculated for each haploid samples. Coefficient of variation for sequencing depth across samples was calculated (X-axis) for the subset of the genes with at least 80% of the samples that show positive depth.

<table>
<thead>
<tr>
<th>Target</th>
<th>Percentage of samples with target sequenced</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Probe level</td>
<td>26341</td>
</tr>
<tr>
<td>Gene level</td>
<td>9997</td>
</tr>
</tbody>
</table>

**Table 2** Reproducibility of the capture on 24 unrelated haploid samples of *P. taeda*. Average sequencing depth was calculated for each individual and used to generate the results shown.

A 100% level means that all 24 samples had at least one read aligning to the target, whereas 0% level indicate the opposite, no reads aligning to the target.

Probe design that avoids boundaries between exons leads to higher capture efficiency

Initially, sequence capture probes could be designed to avoid boundaries between exons for 3946 transcripts for which partial genomic sequences were available. After the completion of probe designs and hybridizations, a draft sequence of the *P. taeda* genome became available (v. 0.8). As a consequence, the position of boundaries between exons could be predicted for 2749 transcript sequences used for probe design based on the draft genome assembly. For probes designed in a region of boundary between exons, the median sequencing depth was 3.6× across the 24 haploid samples. For the same genes, probes that did not overlap two exons had a median depth of 11.2× (t-test *P*-value 5.9 × 10⁻¹³) (Figure 3).

The group of probes with introns predicted to occur within their sequences (*n* = 4937) was used to further investigate whether sequencing depth decayed as a result of the
position of the intron relative to the probe. The 120-nt long probe was assumed symmetric for this analysis and the 60 positions were evaluated, 1 being the outside of the probe and 60 being the middle of the probe. Figure 4 shows that there is a negative correlation ($r^2 = -0.19$) between median sequencing depth and position of the intron in the probe, as expected from the notion that an intron in the middle of the probe will hinder hybridization the most. The large variation observed in the data is probably a result of other factors not accounted for in this analysis, such as GC content, size of the intron, among others.

**Sequence capture expands gene models**

Sequence capture has been broadly used to characterize genetic variation in exon regions. However, captured sequences typically expand beyond the target region because hybridized captured fragments are longer than probes. For species with an uncharacterized genome, this offers an opportunity to expand gene models beyond the available coding sequence, to also include introns, and untranslated or putative regulatory regions. To develop such expanded models for loblolly pine, a bioinformatics pipeline was developed to utilize sequencing reads that span the interval between two capture probes, and extend beyond it. Initially all reads were aligned to the reference sequences used for probe design, and those with similarity at both or single ends (referred hereafter as orphaned reads) were retained to support the de novo assembly of each gene. The longest contigs were then annotated and a new reference was created. To avoid the mis-assembly of paralogs and pseudogenes, the alignment step was performed at relatively high stringency (3% mismatch accepted). Furthermore, the data for all the 24 sequenced haploid samples were combined for the de novo assembly. In total, 11 396 expanded gene models were created, generated from 13 705 contigs. For the majority (9288) of the expanded gene models, reads were assembled into single contigs. However, for 1907 and 201 expanded gene models, two and three contigs were generated, respectively. The average total contig size was 883 bp, ranging from 74 to 2418 bp (Table 3 and Figure 5).

The goal of the de novo assembly was to expand the gene models in pine. Four scenarios were considered during the annotation: (i) expansion of the reference beyond its 5’ end (five_prime_expansion); (ii) complete assembly of an intron (complete_non_reference_fragment); (iii) partial assembly of an intron (incomplete_non_reference_fragment); and (iv) expansion of the reference beyond its 3’ end (three_prime_expansion) (Figure 6). On average, we were able to expand 141 bases 5’ to the reference for 7437 genes and 140 bases 3’ to the reference for 6727 genes (Table 3). A set of 3517 introns in 2746 genes was sequenced completely. These have an average length of 122 bp, and a range of 9–425 bp (Table 3 and Figure S2). Short candidate introns may represent coding sequence insertions absent in the reference. Finally, there were 8609 instances in which introns were partially assembled (average length = 152 bp; Table 3) within 5327 genes. These instances are likely to represent larger introns, where the sequence captured by the nearest probe was not sufficiently long to span the entire intron. For 967 out of 5327 genes that contained partially assembled introns, we were able to connect adjacent incomplete intron fragments based on the following empirical rules: (i) two incomplete introns for the gene are contained within two contigs; (ii) the first contig ends with an incomplete region and the second starts with an incomplete region; and (iii) the two contigs are adjacent to each other relative to the reference (Figure 6b). When these connections between incomplete introns were detected, they were...
Probes developed for loblolly pine are efficient for sequence capture of a related species.

Because the capture between unrelated samples of *P. taeda* was highly reproducible, we tested the suitability of the probes in capturing homologous regions from the genome of *P. elliottii*. Sequence capture was carried out as described for *P. taeda* – i.e. from haploid megagametophyte DNA in three multiplex reactions with eight unrelated samples each, totaling 24 samples. Overall, the metrics of the capture were almost identical to that of *P. taeda* and, 98.9% of the probes that captured the expected sequence efficiently in *P. taeda* were also efficient in the capture of the same regions in *P. elliottii* (data not shown). Capture was also reproducible across the 24 haploid samples sequenced, with a sequencing depth coefficient of variation across samples of 45.3% at the gene level. In total, 11 104 genes were consistently captured in at least 20 of the samples and, similarly to *P. taeda*, 3138 genes were not captured in any of the haploid samples of *P. elliottii*. A de novo reference was also created using the captured data and 13 566 contigs were formed for 11 160 genes. These results validate the flexibility of hybridization-based target enrichment as a way to characterize the genome of closely related species using a common set of probes.

**DISCUSSION**

Here we characterized a reduced representation of the complex genome of the conifers loblolly and slash pine, using in solution target enrichment followed by next-generation sequencing. We focused on a set of 14 729 genes for which transcript sequences were available for probe design. Considering an estimated haploid genome size of 21.7 Gbp for *P. taeda* (Wakamiya et al., 1993) and a probe space of 6.57 Mb, the target region represents only 0.03% of the genome. Sequence capture was evaluated on different genome complexities; including haploid cDNA, haploid genomic DNA and diploid genomic DNA. Overall, more than 70% of captured and sequenced fragments had sequence similarity to the targets, regardless of the sample complexity. A similar success was reported by Saintenac et al. (2011) in the capture of 0.035% of the wheat genome, in which 60% of the reads aligned to targets.

Despite the early success in plant sequence capture (Cronn et al., 2012), the procedure is laborious, limiting its use when the characterization of a large numbers of individuals is required. Here, multiplexing the reaction to simultaneously capture eight samples minimized this limitation.
While this approach seemed to result in some loss in capture efficiency (from 65 to 50%), the significant gains in throughput and cost savings compensated for this drawback. These results are comparable with those reported in human genome studies, in which multiplexed capture was also performed (Cummings et al., 2010; Bansal et al., 2011). Despite the complexity of the multiplex capture, there was good agreement between the fractions of the genome obtained from the samples analyzed. The variation across genes, however, was high. While the mean sequencing depth of the experiment was 108×, the median value was only 10× due to the presence of a few genes sequenced at extremely high depth. While attempts were made during probe design to minimize cross-hybridization, the high copy number of certain genes in conifer genome cannot be overlooked as an obstacle for suitable probe design. This limitation was also reported for wheat (Saintenac et al., 2011), but little action can be done to avoid it when a reference genome is not available. During the UniGene annotation step we observed that the 100 most captured genes were enrichment for those annotated as being located in the chloroplast or as retrotransposon-related proteins (Data S8), and such genes should be avoided in future studies. Therefore, it is advised that new applications of sequence capture in such complex genomes begin with the analysis of few individuals and, if necessary, probes should then be redesigned to limit the occurrence of repetitive regions inferred from the depth of coverage.

Lack of a reference genome sequence also clearly contributed to lower capture efficiency among probes designed from sequences in which the boundaries between exons were unknown. We observed that probes spanning these
boundaries had lower sequencing depth (Figure 3) and that depth was lower the more central those boundaries were relative to the probe (Figure 4). For example, the sequence of gene 0_10007 (Figure 6a) was captured by three probes. An intron detected near the middle of the second probe (relative position 54), most probably contributed significantly to a reduction in median sequencing depth to 4.1 ×, compared with 10.5 × and 14.5 × in the first and third probes, respectively. Although other factors probably contribute to poor hybridization of probes to their targets, such as GC content (Gnirke et al., 2009; Bundock et al., 2012; Zhou and Holliday, 2012), we show that those spanning an intron junction are likely to have less favorable hybridization kinetics when hybridized to genomic DNA.

It was noted by Gnirke et al. (2009) that captured fragments can often be larger than the probes designed to recover them and, thus, regions beyond the probe targets will be captured and sequenced. In this study, approximately 35% of aligned Illumina paired-end data were represented by ‘orphaned’ mates – i.e. only one of two mates aligned to the UniGene reference used for probe design. Consequently, we tested if sequence capture data could be used to expand gene models. For this, the preferred approach was to align the reads to the UniGene reference, identify and recruit the orphaned mates to the aligned ones, and to perform an independent de novo assembly for each UniGene reference. A bioinformatics pipeline (Recriuter) was developed, that implements this approach in an automated fashion. A similar approach was reported recently by Lemmon et al. (2012) on a comparatively simple genome and augmented with manual intervention to annotate the best contigs. With the fully automated Recruiter pipeline we reconstituted 13 705 high-quality contigs for 11 396 genes with no manual intervention. Newly assembled contigs were annotated in GFF (General Feature Format file) files that contained information about introns and 5’/3’ expansions relative to the original expressed sequence tag (EST)-based reference (see Figure 6 for an example of the GFF file for two genes).

The expansion of gene models into non-coding regions generated here provides a comprehensive assessment of the gene structure of conifers. Although the approach is biased towards assembling smaller introns, it suggests that loblolly pine appears to have most of its intron structures similar to other plants and different than humans (Hong et al., 2006), with an average length of 122 and 329 bp, depending on whether they were completely or partially assembled, respectively (Table 3). Unexpectedly, a large portion of the introns was only assembled partially (8609 events for 5327 genes), a finding that suggests the presence of large introns in the genome of pine. In addition to introns, expanded 5’ and 3’ ends of the genes were generated, potentially capturing portions of the regulatory region of some genes. The summary results presented in Table 3 show that 5’ and 3’ ends were expanded on average 141 and 140 bp, respectively; and that complete introns and incomplete but continuous introns were 122 and 329 bp on average, respectively. Thus, relative to the average original reference size (660 bp), reference sequences were expanded by 21% beyond the 5’ and 3’ ends; 18% for complete introns; and 50% for incomplete but continuous introns, a significant contribution for future assembly and characterization of the genome of conifers.

Despite the magnitude and complexity of the conifer genomes, we describe how sequence capture was used to efficiently enrich and sequence the genic portions of the genome of *P. taeda*. We present a series of variables that affect capture efficiency and modified the approach to allow for higher throughput multiplexing the reaction, efficiently capturing 11 695 genes from a genomic background of eight haploid genomes (about 173.6 Gbp). Finally, using the probes designed for *P. taeda*, we show that the interspecific capture of *P. elliottii* worked equally well in retrieving the target genes for this closely related species. Coupled with our de novo assembly and capture multiplexing strategies, this approach provides a proof-of-concept for how sequence capture can be used to quickly expand genome research in uncharacterized and complex genomes. As further improvements to the method are developed, and novel cost reducing modifications are implemented, such as the method proposed by Querfurth et al. (2012), sequence capture presents a powerful method for characterizing genetic diversity in thousands of selected regions of complex plant genomes.

**EXPERIMENTAL PROCEDURES**

**Genotypes, mRNA and DNA extraction**

For sequence capture of haploid cDNA and haploid genomic DNA, megagametophytes were collected from seeds of loblolly pine genotype 17 (Kayihan et al., 2005) and from 24 unrelated trees (for genomic DNA) sampled from the natural distribution of each *P. taeda* and *P. elliottii*. Seeds were stratified at 4 °C. Prior to mRNA extraction, seed germination was stimulated by placing them in a Petri dish with moist filter paper for 4 days, after which seeds were dissected to separate the haploid megagametophyte from the diploid embryo. Whole lyophilized megagametophytes were ground individually in a tissue homogenizer. Messenger RNA and DNA were extracted from the haploid megagametophyte tissue using Dynabeads mRNA DIRECT kit (Invitrogen, http://www.invitrogen.com) and DNeasy Plant Mini Kit (Qiagen, http://www.qiagen.com/), respectively. For sequence capture of diploid genomic DNA, needles were collected from genotype 352 (Quezada et al., 2010). The needle tissue sample was ground manually in liquid nitrogen and DNA was extracted as described above.

**cDNA synthesis**

Messenger RNA was fragmented to an average length of 200 nucleotides with RNA Fragmentation Reagents (Ambion) in a 10 μl reaction, followed by incubation for 3 min and 50 sec at 70°C. Fragmented mRNA was purified using RNA Clean and Concentrator-5 (Zymo Research). First-strand cDNA was synthesized using
SuperScript II Reverse Transcriptase (Invitrogen) with the following modifications: 3 μg of Random Primers (Invitrogen) and 20 μl of RNasin Ribonuclease Inhibitor (Promega) were used in each reaction. Second-strand cDNA synthesis was done using 30 μl dNTP mix (Promega), 2 μl of RNase H (Invitrogen) and 50 μl of DNA polymerase I (Invitrogen) in a 100 μl reaction volume (buffer: 500 mM Tris-HCl pH 7.8, 50 mM MgCl₂, 10 mM DTT). Reactions were incubated at 16°C for 2.5 h and the product was purified with MinElute Reaction Cleanup (Qiagen).

Library preparation

Sequence-ready libraries from haploid DNA and cDNA isolated from megagametophyte, and diploid DNA from needles, were prepared according to Bentley et al. (2008) with modifications. DNA was mechanically sheared to a mean fragment length of 200 base pairs using a Covaris E220 machine, according to manufacturer’s recommendations. Fragments were end-repaired with the End-It DNA End-Repair kit (Epicenter Biotechnologies) and purified with a MinElute Reaction Cleanup (Qiagen). A single qAMP was then added to the 3′ end of the fragments with 15 U of 3′ to 5′ exonuclease Klenow fragment (New England Biolabs) and 10 μl dATP (Promega). When samples were multiplexed in individual hybridization reactions, custom adapters that included 5′ barcode sequences were added and purified as described in the previous step. Adapter ligated fragments of 250-500 bp were excised from an agarose gel and purified with MinElute Gel Extraction (Qiagen). Enrichment PCR (polymerase chain reaction) was performed (up to 18 cycles) using Phusion Polymerase (New England Biolabs) and standard Illumina paired-end primers followed by purification with a MinElute PCR cleanup kit (Qiagen). Each library was quantified with a Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen). A detailed library preparation protocol including the sequence of the barcoded adapters is available in the Supplemental Methods (Methods S1).

Probe design for target enrichment

For probe design a custom pipeline was developed and 15 983 UniGenes assembled from Sanger-sequenced loblolly pine ESTs were used as input (http://denisrime.ucdavis.edu/treegen/). For each UniGene, adjacent non-overlapping 120-nt long regions were extracted as putative probes from the inner part of the UniGene sequence (or predicted exon). This resulted in a total of 64 032 putative probes that were filtered by comparing each probe against the UniGene database at decreasing stringency (from e-value 1 × 10⁻²⁰ to 1 × 10⁻⁵) using WU-Blast BLASTN (Gish, 1996–2004). At each stringency level, probes matching multiple regions were discarded. The probe enrichment was stopped when there were fewer unique probes than the probe set space would accommodate (55 000 probes). The final stringency filter (e-value 1 × 10⁻⁵) resulted in a unique sequence capture set of 54 773 probes representing 14 729 UniGenes, and this probe set was used to prepare a custom SureSelect assay by Agilent Technologies (Data S2).

Sequence capture hybridization reaction and sequencing

Initially, one sample from each genome complexity (haploid cDNA, haploid genomic DNA and diploid genomic DNA) was processed individually and hybridized in individual reactions. The captured fragments for these three libraries were cloned into a vector using the Zero Blunt TOPO PCR cloning kit (Invitrogen) and 96 clones from each library were Sanger sequenced. Upon verification of the feasibility of sequence capture in loblolly pine, eight barcoded libraries from haploid DNA, each representing a single individual, were pooled in equimolar concentrations prior to hybridization. Each pool was hybridized to capture probes following the Agilent SureSelect protocol, re-quantified using Quant-IT PicoGreen dsDNA Assay Kit and diluted to 10 nm. Each pool of eight haploid samples was sequenced on one lane of HiSeq 2000 (Illumina) following the manufacturer’s recommendations, in paired-end mode for 100 cycles. The haploid genomic DNA sample and one multiplexed reaction of eight individual diploid genomic samples were sequenced separately in two Pacific Biosciences SMRT Cells in CCS mode.

Sanger and Pacific Biosciences data processing and analysis

Data filtering. Raw fastq files had the Illumina systematic read identifier renamed to include project specific information (e.g. sequencing center, project). Based on the nucleotide barcode present in the 5′ end of every sequence, reads from each individual were collected into separate files using the barcode splitter function of FASTX-Toolkit. The read identifiers were changed again to include the individual they represent and the barcode sequence was removed from the reads. Low-quality bases (PHRED score <20) on the 3′ end of every read were removed using the fastq quality trimmer function of FASTX-Toolkit and only processed reads longer than 50 nucleotides were retained.

Alignment. Because these are paired-end reads, the quality-filtered files from both ends were re-synchronized to keep the same order of reads in both files. Reads that lost a mate during the data filtering step were treated as single-end reads in the alignment. The reference used for the alignment was composed of the sequences used for probe design. Reads were aligned to the references for these three loci (http://bioinformatics.bsc.es/marthlab/Mosaik) with the following parameters: maximum percentage of mismatches (-mmp 0.05), unique alignment mode turned on (-m unique), most accurate hashing strategy (-a all) and hash size (-h 15). Multiple alignments for a individual were merged using BamTools (Barnett et al., 2011). Raw sequence data for each
genotype was submitted to the Short Read Achieve under study accession SRP018726.

**De novo assembly of captured fragments and gene annotation**

**De novo assembly.** The Recruiter pipeline was developed to assemble the captured reads on a per-genome basis. The data for all 24 haploid samples of *P. taeda* or *P. elliottii* were used to create a consensus assembly for each species. For each individual, the bam alignment file was used to identify reads that aligned to the reference, either with one or both ends aligned. The aligned reads, together with their possible unaligned mate, were recruited from the re-synchronized quality-trimmed fastq file and the recruited data for all samples were combined for each reference UniGene. **CAP3** (Huang and Madan, 1999) is used by Recruiter to assemble the reads, with an assembly task being independently performed for each genome. The recruited fastq file is converted to **CAP3** format, maintaining the paired-end and base quality information, and using the option -h 100 to allow for maximum expansion of the contigs.

**Contig structural annotation.** The contigs formed are annotated to exclude low-confidence contigs and create a GFF-like file with annotation of the contigs relative to the reference. Contigs were annotated when they passed the following empirical criteria: minimum absolute size of 150 bp, minimum relative size compared with the reference of 0.25, minimum absolute number of reads of 20 and minimum relative number of reads relative to the total of 0.15. These values were selected by analyzing the distribution of length and number of reads from the assembly output. For the few cases in which more than three contigs remained after this selection, the top three contigs selected based on coverage relative to the reference were maintained for annotation. The following features were present in the GFF annotation file and their meaning is explained in the text: five_prime_expansion, three_prime_expansion, reference like_fragment, complete_non_reference_fragment and incomplete_non_reference_fragment. The source code for Recruiter is available upon request and the de novo assembly files for *P. taeda* and *P. elliottii* are provided in Supplementary Data (Data S3, S4, S5 and S6).

**UniGene functional annotation.** The original UniGenes used for probe design were functionally annotated using Blast2GO auto-assembled pipeline on default parameters (Conesa et al., 2005), with the only difference that BLASTX results (e-value threshold $1 \times 10^{-10}$) against NCBI non-redundant database were generated locally and imported to Blast2GO. The annotated results for all UniGenes are reported in Supplementary Data S7. A Fisher exact test was performed to compare gene ontology (GO) enrichment between the top 100 most captured genes and the remainder using Blast2GO automated pipeline.

**Intron prediction and depth analysis**

The *P. taeda* draft assembly (v. 0.8) was used to predict introns for an unbiased comparison of the effect of introns in the capture efficiency. The large size of the genome is not handled by **GMAP**, thus the option -h 100 to allow for maximum expansion of the contigs. The large size of the genome is not handled by **GMAP**, thus the relative position of the intron in the probe was obtained adjusted to the center, thus varying from 1 (borders or positions 1 and 120) to 60 (center or positions 60 and 61) to increase the number of observations per class. The sequence depth for each probe was calculated as the average number of reads that aligned to each of the 60 innermost bases of the probe. The center of the probe was used in an attempt to avoid neighboring capture effect from adjacent probes. The final probe depth used for the analysis is the median depth taken across the 24 haploid samples of *P. taeda*. Supplementary Data S9 provides the median depth for all the 54 773 probes designed for *P. taeda*.

**SUPPORTING INFORMATION**

Additional Supporting Information may be found in the online version of this article.

**Figure S1.** Distribution of reads per multiplexing reaction. Eight samples were combined in a multiplexed capture reaction (pool).

**Figure S2.** Size distribution of completely assembled introns ($n = 3517$).

**Methods S1.** Sequence capture protocol for multiplexing capture.

**Data S1.** Fasta file with the original *Pinus taeda* unigenes used for probe design.

**Data S2.** List of probes used in the experiment relative to the uni-gene reference.

**Data S3.** Fasta file with the de novo reference created for *Pinus taeda*.

**Data S4.** General feature format file with structural annotation for the *Pinus taeda* de novo reference.

**Data S5.** Fasta file with the de novo reference created for *Pinus elliottii*.

**Data S6.** General feature format file with structural annotation for the *Pinus elliottii* de novo reference.

**Data S7.** Functional annotation of unigenes used for capture using Blast2GO automated pipeline.

**Data S8.** Categories of gene ontology that are over represented in the top 100 most captured genes compared to the whole gene set.

**Data S9.** Median sequencing depth for the 54 773 probes across all 24 *P. taeda* haploid samples sequenced. Depth was calculated for the 60 innermost bases of the probe.

**REFERENCES**


Genetics 186, 677–696.

Gnirke, A., Melnikov, A., Maguire, J.

Gish, W.

Gnirke, A., Melnikov, A., Maguire, J.

Gish, W.


