Genome Biology

RESEARCH Open Access

Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies

David B Neale[1*], Jill L Wegrzyn[1], Kristian A Stevens[2], Aleksey V Zimin[3], Daniela Puiu[4], Marc W Crepeau[2], Charis Cardeno[2], Maxim Koriabine[5], Ann E Holtz-Morris[5], John D Liechty[1], Pedro J Martínez-García[1], Hans A Vasquez-Gross[1], Brian Y Lin[1], Jacob J Zieve[1], William M Dougherty[2], Sara Fuentes-Soriano[6], Le-Shin Wu[7], Don Gilbert[6], Guillaume Marçais[3], Michael Roberts[3], Carson Holt[8], Mark Yandell[8], John M Davis[9], Katherine E Smith[10], Jeffrey FD Dean[11], W Walter Lorenz[11], Ross W Whetten[12], Ronald Sederoff[12], Nicholas Wheeler[1], Patrick E McGuire[1], Doreen Main[13], Carol A Loopstra[14], Keithanne Mockaitis[6], Pieter J deJong[5], James A Yorke[3], Steven L Salzberg[4] and Charles H Langley[2]

Abstract

Background: The size and complexity of conifer genomes has, until now, prevented full genome sequencing and assembly. The large research community and economic importance of loblolly pine, *Pinus taeda* L., made it an early candidate for reference sequence determination.

Results: We develop a novel strategy to sequence the genome of loblolly pine that combines unique aspects of pine reproductive biology and genome assembly methodology. We use a whole genome shotgun approach relying primarily on next generation sequence generated from a single haploid seed megagametophyte from a loblolly pine tree, 20-1010, that has been used in industrial forest tree breeding. The resulting sequence and assembly was used to generate a draft genome spanning 23.2 Gbp and containing 20.1 Gbp with an N50 scaffold size of 66.9 kbp, making it a significant improvement over available conifer genomes. The long scaffold lengths allow the annotation of 50,172 gene models with intron lengths averaging over 2.7 kbp and sometimes exceeding 100 kbp in length. Analysis of orthologous gene sets identifies gene families that may be unique to conifers. We further characterize and expand the existing repeat library based on the *de novo* analysis of the repetitive content, estimated to encompass 82% of the genome.

Conclusions: In addition to its value as a resource for researchers and breeders, the loblolly pine genome sequence and assembly reported here demonstrates a novel approach to sequencing the large and complex genomes of this important group of plants that can now be widely applied.

Background

Advances in sequencing and assembly technologies have made it possible to obtain reference genome sequences for organisms once thought intractable, including the leviathan genomes (20 to 40 Gb) of conifers. Gymnosperms, represented principally by a diverse and majestic array of conifer species (approximately 630 species, distributed across eight families and 70 genera [1]), are one of the oldest of the major plant clades, having arisen from ancestral seed plants some 300 million years ago. Conifers will likely provide many genome-level insights on the origins of genetic diversity in higher plants.

Though today's conifers may be considered relics of a once much-larger set of taxa that thrived throughout the age of the dinosaurs (250 to 65 millions of years ago) [2,3], they remain the dominant life forms in many of the temperate and boreal ecosystems in the Northern Hemisphere and extend into subtropical regions and the Southern Hemisphere.

We chose to investigate the loblolly pine (*Pinus taeda* L.) genome because of its well-developed scientific resources. Over 1.5 billion seedlings are planted annually, approximately 80% of which are genetically improved,

* Correspondence: dbneale@ucdavis.edu
[1]Department of Plant Sciences, University of California, Davis, CA, USA
Full list of author information is available at the end of the article

driving its selection as the reference conifer genome. Among conifers, its genetic resources are unsurpassed in that three tree improvement cooperatives have been breeding loblolly pine for more than 60 years and manage millions of trees in genetic trials. The current consensus reference genetic map for loblolly pine is made up of 2,308 genetic markers [4]. Extensive QTL and association mapping studies in loblolly pine have revealed a great deal about the genetic basis of complex traits such as physical and chemical wood properties, disease and insect resistance, growth, and adaptation to changing environments. Current research focuses on the potential of genomic selection for continued genetic improvement [5].

The tree selected for sequencing, '20-1010', is a member of the North Carolina State University-Industry Cooperative Tree Improvement Program and the property of the Commonwealth of Virginia Department of Forestry, which released this germplasm into the public domain. In accordance with open access policies [6], we released the first draft genome of loblolly pine in June 2012, which made it the first draft assembly available for any gymnosperm. The draft described here represents a significant advance over available gymnosperm reference sequences [7,8].

Results and discussion

Sequencing and assembly

The loblolly pine genome [9] joins the two other conifer reference sequences produced recently [7,8]. With an estimated 22 billion base pairs [10], it is the largest genome sequenced and assembled to date. Our experimental design leveraged a unique feature of the conifer life cycle and new computational approaches to reduce the assembly problem to a tractable scale [9,11]. From the first whole genome shotgun (WGS) assembly of the 1.8 million base pair *Haemophilus influenzae* genome in 1995 to the orders-of-magnitude larger three-billion-base-pair mammalian genomes that followed years later [12], the WGS protocol has been an efficient and effective method of producing high quality reference genomes. This was in part made possible by the overlap layout consensus (OLC) assembly paradigm championed by Myers [13] and ubiquitously implemented in first-generation WGS assemblers. When next-generation sequencing disruptively ushered in a new era of WGS sequencing, the extremely large numbers of reads exceeded the capabilities of existing OLC assemblers. To circumvent this, new assemblers were developed, using short k-mer based methods first described by Pevzner [14]. The giant panda [15] was the first mammalian species to have its genome produced using strictly NGS reads. For loblolly pine, we utilized a hybrid assembly method that incorporates both k-mer based and OLC assembly methods.

Figure 1A illustrates the two sources of DNA that comprised the sequencing strategy. As outlined below (see [9] for details), the majority of the WGS sequence data in Table 1 was generated from a single pine seed megagametophyte. The small quantity of genomic DNA obtained from the haploid megagametophyte tissue was used to construct a series of 11 Illumina paired end libraries with sufficient complexity to form the basis of a high quality WGS assembly. The use of haploid DNA greatly simplifies assembly, but the limited quantity of haploid DNA was insufficient for the entire project. Diploid needle tissue served as an abundant source of parental DNA for the construction of long-insert linking libraries. This included 48 libraries ranging from 1 to 5.5 kilobase pairs (Kb) and nine fosmid DiTag libraries spanning 35 to 40 Kb.

An overview of the assembly process is presented in Figure 1B. The combined 63× coverage from megagametophyte libraries (approximately 15 billion reads) was used for error correction and for the construction of a database of 79-mers appearing in the haploid genome. This database was used to filter highly divergent haplotypes from the diploid sequence data. The super-read reduction implemented in the MaSuRCA assembler [11] condensed most of the haploid paired-end reads into a set of approximately 150 M longer 'super-reads'. Each super-read is a single contiguous haploid sequence that contains both ends of one or more paired-end reads. The construction process ensured that no super-read was contained in another super-read. Critically, the number of megagametophyte-derived reads was reduced by a factor of 100. The combined dataset was 27-fold smaller than the original, and was sufficiently reduced in size to make overlap-based assembly using CABOG [16] possible. The output of the MaSuRCA assembly pipeline became assembly 1.0. Additional scaffolding methods were implemented to improve the assembly by taking advantage of the deeply sampled transcriptome data [17], ultimately producing assembly v1.01. Finally, to further assess completeness, a scan for the 248 conserved core genes in the CEGMA database [18] was performed on all conifer assemblies (Figure 1B). The resulting annotations are classified as full length and partial. The loblolly pine v1.01 assembly has the largest number of total annotations (203) of the three conifers as well as the largest fraction of full length annotations (91%).

For validation purposes, we used a large pool of approximately 4,600 fosmid clones to approximate a random sample of the genome [9]. The sequenced and assembled pool contained 3,798 contigs longer than 20,000 bp, each putatively representing more than half of a fosmid insert, with a total span of 109 Mbp. When aligned to the genome 98.63% of the total length of these contigs was covered by the WGS assembly. A total of

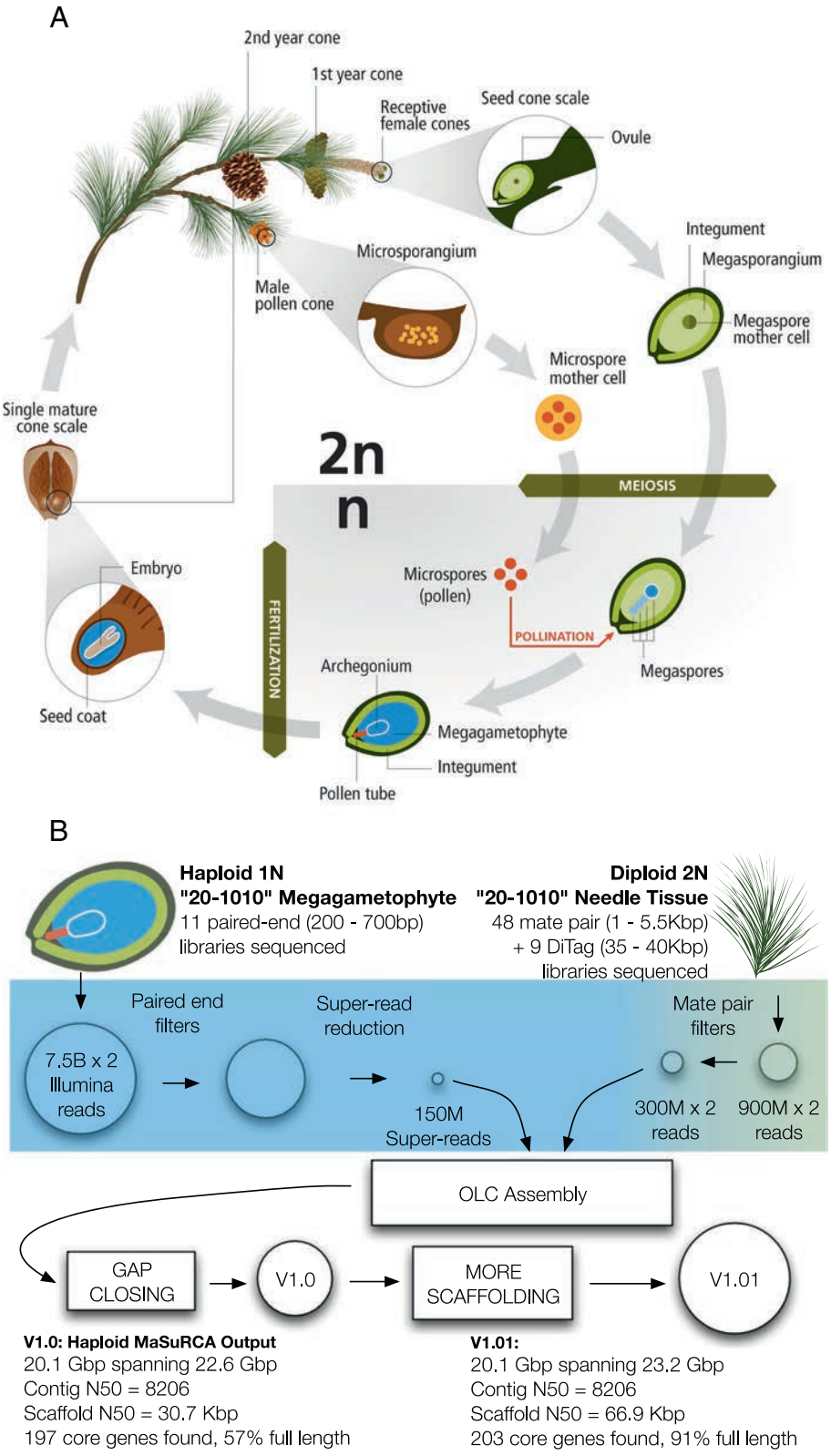

Figure 1 (See legend on next page.)

(See figure on previous page.)
Figure 1 (A) The sources of haploid and diploid genomic DNA. The reproductive cycle of a conifer showing the unique sources of haploid and diploid genomic DNA sequenced. Both the ova pronucleus and the megagametophyte are derived by mitotic divisions from a single one of the four haploid meiotic segregant megaspores. The tissue from a single megagametophyte formed the basis for all of our shorter insert paired end Illumina libraries (Table 1). To construct longer insert libraries (Illumina mate pair and Fosmid DiTag) requiring greater amounts of starting DNA, needles from the parental genotype (20-1010) were used. **(B)** Sequencing and assembly schematic. An overlap layout consensus assembly, made possible by MaSuRCA's critical reduction phase, was followed by additional scaffolding, incorporating transcript assemblies, to improve contiguity and completeness [9,11].

2,120 of the aligned contigs had 99.5% or higher similarity, implying a combined error rate of less than 0.5%.

Annotation

A *de novo* transcriptome assembly of 83,285 unique, full-length contigs from several tissue types and existing nucleotide resources (ESTs and conifer transcriptomes) supported a set of 50,172 unique gene models, derived from the MAKER-P annotation pipeline (Table 2) [19,20]. From the *de novo* transcriptome assembly, 42,822 aligned uniquely (98% identity and 95% coverage) to the genome. Of the 45,085 re-clustered loblolly pine EST sequences, 27,412 aligned (98% identity and 98% coverage). The frequent occurrence of pseudogenes (gene-like fragments representing 2.9% of the genome), required the use of conservative filters to define the final gene space [20]. The selected models represent coding-sequence lengths between 120 bp and 12 Kbp. Gene and exon lengths were comparable with angiosperm species; however, the number of full-length genes identified, even in a more fragmented genome, was greater than in other species (Figure 2A). Introns numbered 144,579 with an average length of 2.7 Kbp and a maximum length of 318 Kbp. A total of 6,267 (4.4%) of the introns were greater than 20 Kbp in length. This distribution far exceeds the intron lengths reported in other plant species and is, on average, longer than estimates in *Picea abies* [8,20]. The final gene models were identified on 31,284 scaffolds that were at least 10 Kbp in length. A total of 3,835 scaffolds contained three or more genes. Given the fragmentation of the genome and long intron lengths observed, it is likely that the genome contains additional genes, but also that some of the 50,172 models defined here may later be merged together.

Table 1 Characteristics of the loblolly pine v1.01 draft assembly

Estimated 1 N genome size	22 Gbp [10]
Number of chromosomes	12
G + C%	38.2%
Sequence in contigs >64 bp	20,148,103,497 bp
Total span of scaffolds	23,180,477,227 bp
Contig N50	8,206 bp
Scaffold N50	66,920 bp
Haploid paired end libraries 200-600 bp	11 libraries
	7.5x billion x 2 reads (GA2x + HiSeq + MiSeq)
	1.4 trillion bp total read length
	63x sequence coverage
	150 million maximal super-reads
	52 billion total bp
	2.4x sequence coverage
Diploid mate pair libraries 1,000-5,500 bp	48 libraries
	863 million x 2 reads (GA2x)
	273 billion total read length
	270 million x 2 reads after filtering
	37x physical coverage
DiTag libraries 35-40 Kbp	9 libraries
	46 million x 2 reads (GA2x)
	4.5 million reads x 2 after filtering
	7.5x physical coverage

We clustered the protein sequences in order to identify orthologous groups of genes [23]. Comparisons with 14 species, ignoring transposable elements, yielded 20,646 gene families with two or more members and 1,476 gene families present in all species (Figure 2A). Of the full set, 1,554 were specific to conifers and 159 of those were specific to loblolly pine [20].

The majority of characterized plant resistance proteins (R proteins) are members of the NB-ARC and NB-LRR families, and are associated with disease resistance [24]. Several independent families were identified containing one or both of these domains. The largest contained 43 loblolly pine members and 14 spruce members. Several other smaller families contained members exclusively from loblolly pine ranging from two to five members each. Other gene families with roles in disease resistance were also identified, including Chalone synthases (CHS) (three in loblolly pine and one spruce member). Increased expression of CHS is associated with the salicylic acid defense pathway [25].

Response to environmental stress, such as salinity and drought, has been investigated at length in conifers. Three different sets of Dehydrin (DHN) domains were noted, the largest with 10 loblolly pine members and 12

Table 2 Comparison of gene metrics among sequenced plant genomes

	Pinus taeda	*Picea abies* [8]	*Arabidopsis thaliana* [21]	*Populus trichocarpa* [21]	*Vitis vinifera* [21]	*Amborella trichopoda* [22]
Genome size (assembled) (Mbp)	20,148	12,019[a]	135	423	487	706
Chromosomes	12	12	5	19	19	13
G + C content (%)	38.2	37.9	35.0	33.3	36.2	35.5
TE content (%)	79	70	15.3	42	41.4	N/A
Number of genes[b]	50,172	58,587[c]	27,160	36,393	25,663	25,347
Average CDS length (bps)	965	723	1102	1143	1095	969
Average intron length (bps)	2,741	1,020	182	366	933	1,538
Maximum intron length (bps)	318,524	68,269	10,234	4,698	38,166	175,748

[a]Estimated genome size is 19.6 Gbp.
[b]Number of full-length genes >150 bp in length and validated through current annotations.
[c]High and medium confidence genes from the Congenie project [8].

spruce members. The first genetic evidence of dehydrins playing a role in cellular protection during osmotic shock was in 2005, in *Physcomitrella patens* [26]. Subsequently, it was noted that transcription levels of a DHN increased in *Pinus pinaster* when exposed to drought conditions [27].

Cupins are members of a large, diverse family belonging to the Germin and Germin-like superfamily (GLP) [28]. In this analysis, several families, including one large family contained one or more domains related to Cupin 1. The largest family contained 23 loblolly pine members and five spruce members. These genes, similar to other GLPs, are expressed during somatic embryogenesis in conifers [28]. Cupins have therefore been associated with plant growth, and more recently associated with disease resistance in rice [29].

The *COPI C* family (58 members) was the largest exclusively identified in loblolly pine. Vesicle coat protein complexes containing *COPI* family members mediate transport between the ER and golgi, and interact with Ras-related transmembrane proteins, p23 and p24 [30]. Members of the Ras superfamily, *Arf* and *ArfGap*, also identified in loblolly pine, are involved in *COPI* vesicle formation [31]. These proteins were assigned to the small molecule binding GO category, which is enriched in pine and other conifers as compared with angiosperms. The other notable GO assignments include nucleic acid binding, protein binding, ion binding, and transferase activity which are consistent with the most populated categories for the other species included in the comparison (Figure 2B).

Repetitive DNA content

Previous examination of the loblolly pine BAC and fosmid sequences led to the development of the Pine Interspersed Element Resource (PIER), a custom repeat library [32]. *De novo* analysis of 1% of the genome yielded 8,155 repeats, bringing PIER's total to 19,194 [20]. The plethora of novel repeat content may be explained in part by the highly diverged nature of the repeat sequences, which prevents accurate identification from a reference library due to obscured similarities. Homology analysis demonstrated that retrotransposons dominated, representing 62% of the genome (Figure 3A). Seventy percent of these were long terminal repeat (LTR) retrotransposons. *PtConagree* [32] covered the largest portion of the genome, followed by *TPE1* [33], *PtRLC_3*, *PtRLX_3423* [20], *PtOuachita*, and *IFG7* [34]. Among introns, the estimated repetitive content was 60%. Introns were relatively rich in DNA transposons, at 3.31% (Figure 3A). Overall, the combined similarity and *de novo* approaches estimate that 82% of the pine genome is repetitive in nature (Figure 3A).

Though the genome is inundated with interspersed repetitive content, analysis revealed that only 2.86% is composed of tandem repeats, the majority of which are comprised of millions of retrotransposon LTRs. This estimate is comparable to the frequencies observed in other members of the Pineaceae (2.71% in *Picea glauca* (v1.0) and 2.40% in *Picea abies* (v1.0)) [20]. The number of tandem repeats may be dependent on sequencing and assembly methodologies. As shown previously, loblolly pine ranges from 2.57% in the WGS assembly [20] to 3.3% in Sanger-derived BACs [32]. Similar to most species, the relative frequencies of the repeating units of tandem loci are heavily weighted towards minisatellites (between 9 and 100 bp). This attribute is ubiquitous across plants but the smaller volume of microsatellites (1 to 9 bp repeating units) in conifers when compared to angiosperms and the increased contribution from heptanucleotides is significant (Figure 3B). A substantial number of loblolly pine's tandem repeats are

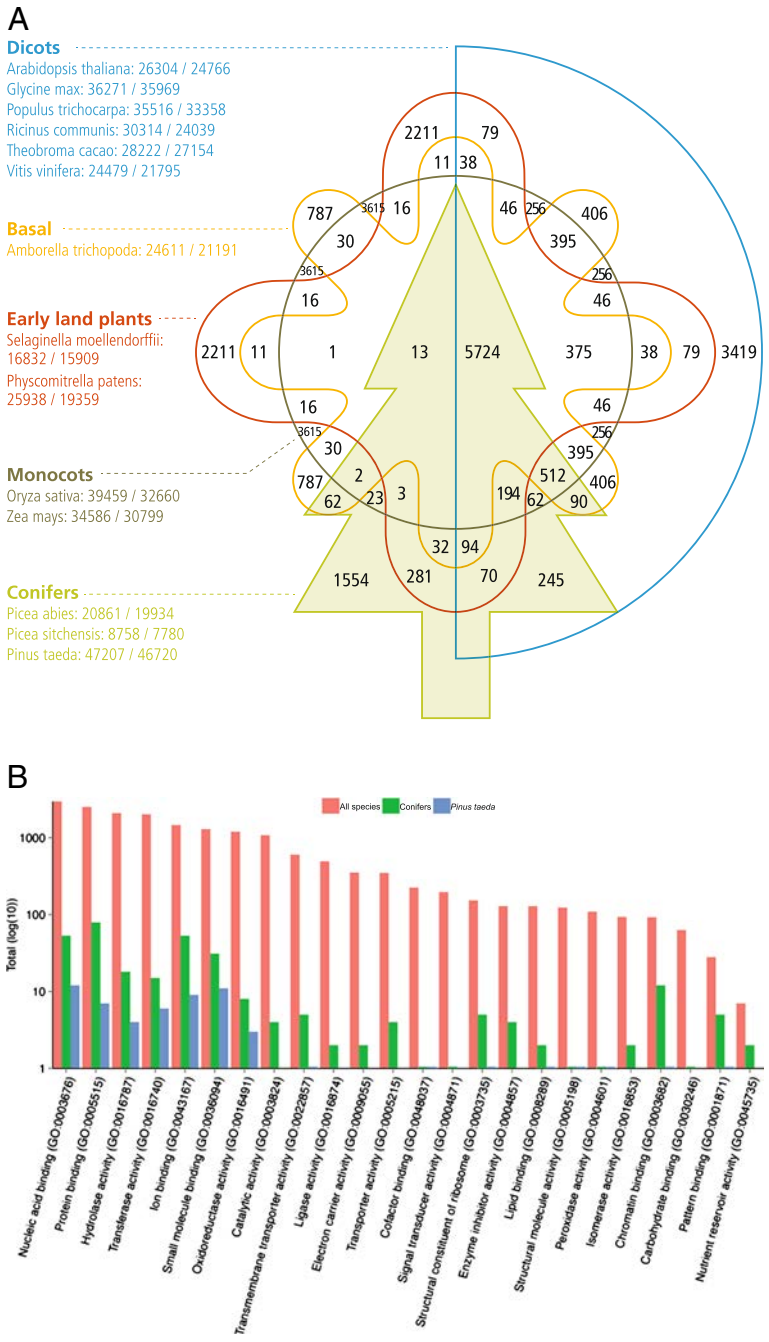

Figure 2 Unique gene families and Gene Ontology term assignments. (A) Identification of orthologous groups of genes for 14 species split into five categories: conifers (*Picea abies*, *Picea sitchensis*, and *Pinus taeda*), monocots (*Oryza sativa* and *Zea mays*), dicots (*Arabidopsis thaliana*, *Glycine max, Populus trichocarpa*, *Ricinus communis*, *Theobroma cacao*, and *Vitis vinifera*), early land plants (*Selaginella moellendorffii* and *Physcomitrella patens*), and a basal angiosperm (*Amborella trichopoda*). Here, we depict the number of clusters in common between the biological categories in the intersections. The total number of sequences for each species is provided under the name (total number of sequences/total number of clustered sequences). **(B)** Gene ontology molecular function term assignments by family for all species (red), conifers (green), and *Pinus taeda* exclusively (blue).

telomeric sequences (TTTAGGG)n (approximately 23,926 loci) and candidate centromeric sequences (TGGAAACCC-CAAATTTTGGGCGCCGGG)n (5,183 loci, 1.8 Mbp). Originally identified in *Arabidopsis thaliana* [35], the telomeric sequences were found interstitially as well as at the end of the chromosomes in loblolly pine and other conifers [36]. The interstitial presence of the heptanucleotide repeat may explain the increased observation of this microsatellite in conifers. Pines have especially long telomeres, reaching up to 57 Kbp as found in *Pinus longaeva* [37,38].

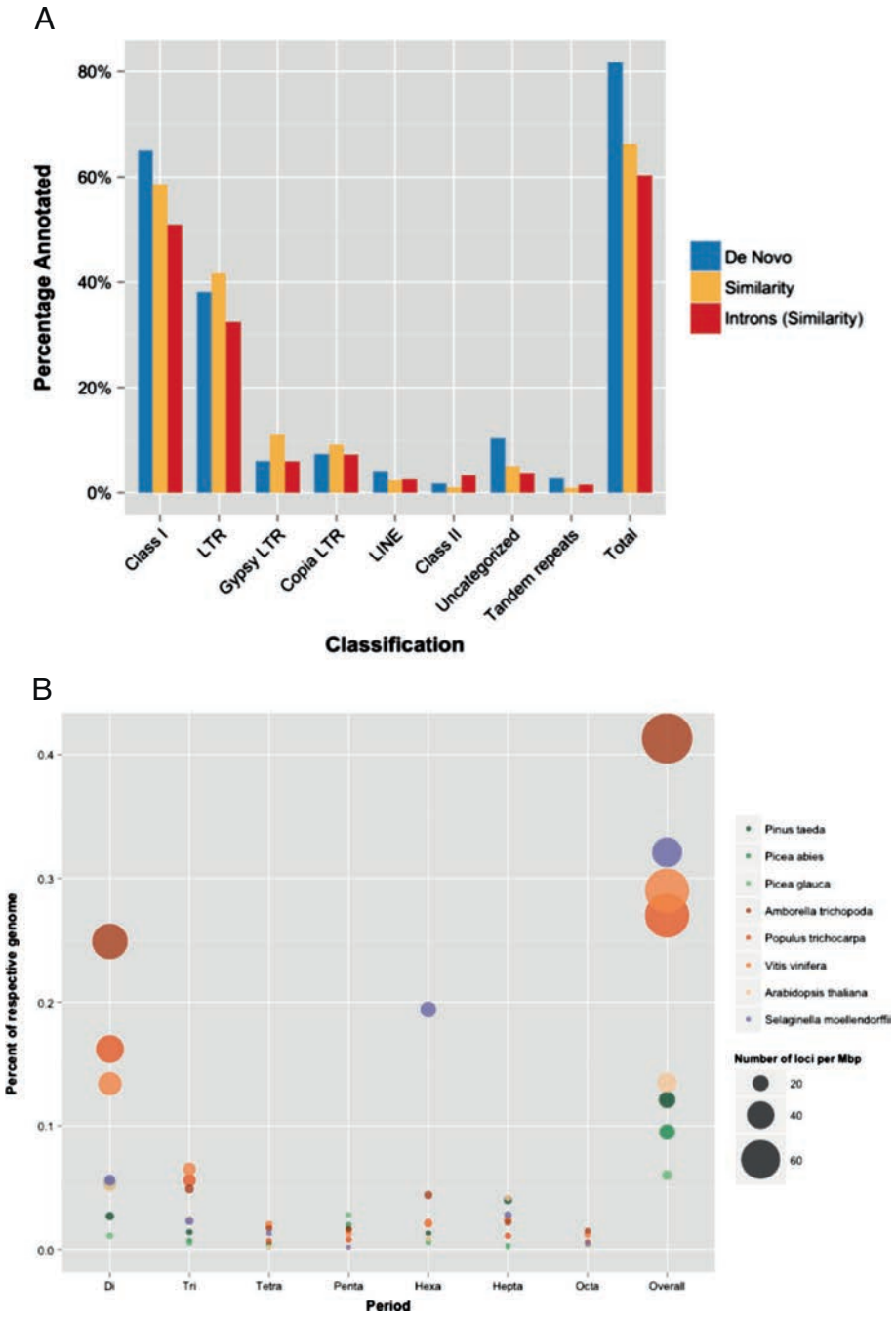

Figure 3 Interspersed and tandem repetitive content. (A) Overview of repetitive content in the *Pinus taeda* genome for similarity (blue) and *de novo* (yellow) approaches. Introns are evaluated with similarity methods against PIER 2.0 [32]. **(B)** Overview of microsatellite content across species with exclusion of mononucleotide repeats. Orange, green, and purple points represent angiosperm, gymnosperm, and lycophyte species, respectively. Each point displays both the density (point size) and length (y-axis) of di-, tri-, tetra-, penta-, hexa-, hepta-, and octanucleotide tandem repeats (x-axis). The *Overall* category is an accumulation of the previous seven categories.

Organelle genomes

The mitochondrial genome was identified and assembled separately, taking advantage of its deeper coverage and distinctive GC content. The assembly was built primarily from 28.5 million high-quality 255-bp MiSeq reads generated during WGS sequencing. The read were first assembled with SOAPdenovo2 [39], and the resulting 7,559 scaffolds were aligned to the loblolly pine chloroplast genome [40] and to 557 complete and partial plant mitochondrial genomes. Twenty-seven scaffolds aligned

to the chloroplast and 90 aligned to mitochondria. The mitochondrial scaffolds were distinguished by their coverage, which averaged >14x, while the coverage depth of chloroplast scaffolds was far deeper, more than 100x. The original reads represented just 0.3x coverage of the nuclear DNA. The mitochondrial GC-content was 44% *versus* 38.2% for the genome and 39.5% for the chloroplast. Based on these results, we identified 33 unaligned scaffolds longer than 1 Kb likely to be mitochondrial, with > = 44% GC content and coverage between 8x and 50x. These plus the 90 previously aligned scaffolds were reassembled, using additional reads from two WGS jumping libraries (lengths 3,800 bp and 5,200 bp), extracting only those pairs that matched the mitochondrial contigs. The resulting mitochondrial genome assembly has 35 scaffolds containing 40 contigs, with a total contig length of 1,253,551 bp and a maximum contig size of 256,879 bp.

New insights in conifer functional biology

The draft genome sequence and transcriptome assemblies have enabled discovery of genes that underlie ecologically and evolutionarily important traits, illuminated larger-scale genomic organization of gene families, and revealed missing genes that evolved in angiosperms and not gymnosperms.

Disease resistance

The genome revealed that a partial EST containing a SNP was actually a candidate gene for rust resistance in loblolly pine. We mapped the SNP genetically, associated it with rust resistance then determined it was a toll-interleukin receptor/nucleotide binding/leucine-rich repeat (TNL) gene [41] containing signature domains only present in the new transcript and genome assemblies. Rust pathosystems can provide useful insights into host-pathogen co-evolution, because host resistance genes interact genetically with pathogen avirulence genes [42,43]. Analysis of fusiform rust pathogen *Cronartium quercuum* (Berk.) Miyabe ex Shirai f.sp. *fusiforme* (*Cqf*) genetic interactions with *Pinus* hosts [44,45] led to mapping of *F*usiform *r*ust resistance *1* (*Fr1*; [46]) to LG2 [47,48]. *P. lambertiana Cr1* for white pine blister rust resistance [49] was also mapped to the same linkage group using syntenic markers [50].

Two large mapping populations were used to assign genetic map positions to 2,308 SNP markers that were mapped to genomic scaffolds [4], whose SNPs were then tested for association with rust resistance [48] in a family-based, clonal population [51]. The top-ranked SNP for rust resistance in a parent segregating for *Fr1* mapped to LG2 (31.3 cM; Figure 4A) and occurred in a transcript model encoded by a TNL-type gene located on a genomic scaffold (Figure 4B). The TNL-type gene is related to *N* from *Nicotiana* [52] that belongs to a large class of genes for resistance to biotrophic pathogen-induced diseases [41,53]. Prior to this work the loblolly pine gene product appeared to lack TIR and NB domains because the EST was truncated. Based on OrthoMCL analysis of the full-length proteins [20], the gene belongs to a class of TNLs that have expanded in conifers (N = 780 in loblolly pine; N = 180 in *P. abies*) but not *Arabidopsis* (N = 3). By contrast, most TNL genes in *Arabidopsis* belong to a large class (N = 138) not found in loblolly pine or *P. abies*.

The genome sequence has revealed that distinct classes of TNLs have expanded in conifers and angiosperms, making it feasible to test conifer candidate genes for co-segregation with disease resistance, instead of using markers derived from incomplete ESTs or other species. The transcript of the TNL gene is detected in young stems, reaction wood, in hymenial layers obtained from fusiform rust galls, and is a candidate for *Fr1. Avr1*, the avirulence gene that specifically interacts with *Fr1* [54], has been genetically mapped on LGIII of *Cqf* and the genome sequence is now available [55]. These genome-based discoveries open the door to understanding the effects of host *Fr* genes on allelic diversity and frequency in their corresponding *Avr* genes, and vice versa, at large geographic scales. The practical outcomes for *Fr* gene durability are significant given the widespread planting of >500 million loblolly pine seedlings each year that harbor one or more fusiform rust resistance loci as a consequence of parental selection, selective breeding, and screening for fusiform rust resistance [56]. Comparisons of *Fr1* and *Cr1* loci should generate new insights into evolution of resistance genes [57] within a genus that arose 102-190 million years ago [58].

Stress response

Conifers dominate a variety of biomes by virtue of their capacity to survive and thrive in the face of extreme abiotic stresses. For example, pronounced resistance to water stress, particularly in mature trees, has enabled conifers to spread across deserts and alpine areas, well beyond the range of most competing woody angiosperms. At the same time, water stress is a major cause of mortality for conifer seedlings [59], and predictions hold that differential susceptibility of conifer species to water stress will have profound consequences for forest and ecosystem dynamics under future climate change scenarios [60]. Variation in drought resistance in conifers has long been recognized to have a genetic basis [61,62], and substantial effort has previously been devoted to attempts at using molecular and genomic tools to uncover the responsible genetic determinants [63-65].

One of the first drought-responsive conifer genes to be cloned and characterized was *lp3* [66], which was shown to share homology with a small family of nuclear-

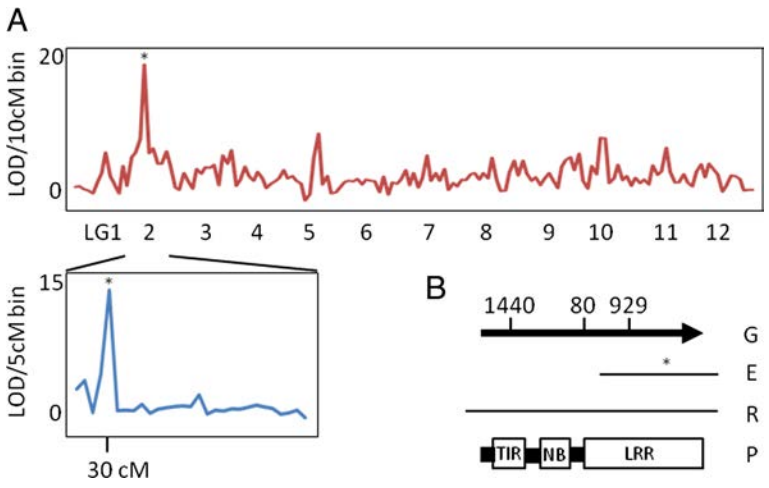

Figure 4 Identification of TNL candidate gene for Fr1. (A) Genome survey of rust resistance in segregating progeny of Fr1/fr1 *Pinus taeda* among clonally propagated half-siblings (upper) and full-siblings (lower). Bins with highest LOD scores contained. SNP 2_5345_01 (*). **(B)** Translated gene model (G) on genome scaffold jcf7180063178873 is interrupted by three introns with sizes given in bp, previously available EST (E) containing SNP 2_5345_01 (*), fully assembled transcript Evg1_1A_all_VO_L_3760_240252 from RNAseq (R) and the domain structure of the protein model (P).

localized, ABA-inducible genes (termed *ASR* for ABA-, Stress- and Ripening) initially identified in tomato [67,68]. Subsequent work has shown that *ASR* genes are broadly distributed in higher plants and adaptive alleles of these genes are determinants of drought resistance in wild relatives of various domesticated crops [69-71]. Transcriptomic studies have detected differential expression of *lp3* gene family members in drought-stressed pine [72,73], while other studies have linked expression of *lp3* gene family members to aspects of wood formation, that is, xylem development [74,75] and cold tolerance [76]. Genetic studies indicate that *lp3* alleles are under selection in pine and likely confer adaptive resistance to drought [77,78].

Protein sequences for the four distinct loblolly pine *lp3* gene family members in GenBank (AAB07493, AAB02692, AAB96829, AAB03388) aligned optimally to four of the high-confidence gene models. The *ASR* genes in tomato are physically clustered and have been held out as examples of tandemly arrayed genes that are important for adaptation [70]. In the v1.01 assembly, two of the pine *lp3* genes (AAB07493, AAB96829) were found to reside on the same scaffold (Figure 5). Very little is known about the physical clustering of gene family members in conifers, but the availability of the loblolly pine genome sequence now provides the opportunity to study such relationships and their contribution to adaptation in conifers.

Wood formation

The genome assembly and annotation provide new information on the roles of specific genes involved in wood formation. Pine secondary xylem contains large numbers of tracheids with abundant bordered pits for both mechanical support and water transport; by contrast, the secondary xylem of woody dicots typically has specialized vessel elements for water conduction, and fiber cells for mechanical support [79,80]. The chemical composition of gymnosperm xylem is characterized by a guaiacyl-rich (G type) lignin and the absence of syringyl (S type) subunits [81]. The lignins in the xylem of woody dicots, gnetales, and Selaginella (a lycopod) are characterized by a mixed polymer of S and G subunits (S/G lignin) [82]. The hemicelluloses of pines are a mixture of heteromannans, while dicot hemicelluloses are typically xylan-rich [83]. The functional and evolutionary differences in lignin composition, hemicellulose composition, and presence of vessel elements between gymnosperms and angiosperms are informed by the pine genome sequence. Expressed homologs of all but one of the known genes for lignin precursor (monolignol) biosynthesis [84] have been identified in the pine genome assembly. The exception is the gene encoding ferulate 5-hydroxylase (also called coniferaldehyde 5-hydroxylase), the key enzyme for the formation of sinapyl alcohol, the precursor for S subunits in S/G lignin [85]. The absence of a 5-hydroxylase homolog is significant because of the depth of pine sequencing and the quality of the annotation [20]. A putative homolog of a gene only recently implicated in monolignol biosynthesis, encoding caffeoyl shikimate esterase [86], has also been identified in the pine annotation. Some monolignol gene variants are associated with quantitative variation in growth and wood properties such as wood density or microfibril angle [87-89]. The draft pine genome assembly contains putative homologs of six cellulose synthase subunits (CesA1,

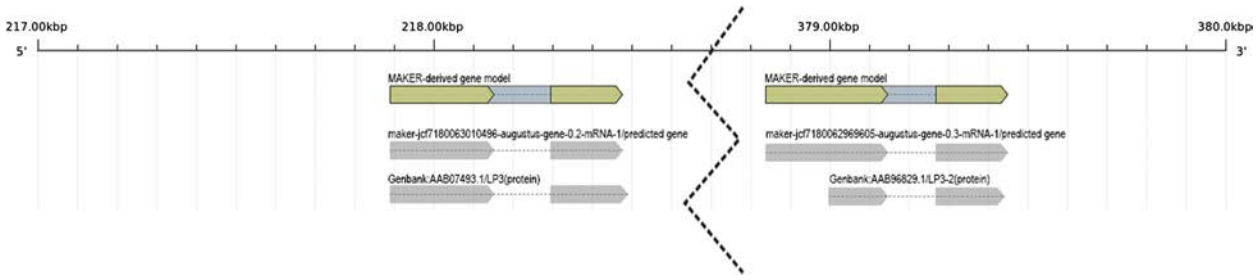

Figure 5 *Pinus taeda* lp3 sequences from Genbank (AAB07493, AAB96829) were aligned to the same scaffold in v1.01 and supported by two distinct MAKER-derived gene model.

3, 4, 5, 7, and 8), and two putative gene models for glucomannan 4-beta-mannosyltransferases and two for xyloglucan glycosyltransferases, consistent with the hemicellulose composition of pine. The pine genome assembly also contains putative homologs for many genes that encode transcription factors that regulate wood cell types or the perennial growth habit [90,91]. This information is useful to guide the genetic improvement of wood properties as resources for biomaterials and bioenergy.

Conclusions

The loblolly pine reference genome joins the recent genomes of Norway spruce and white spruce forming a foundation for conifer genomics. To tackle the problem of reconstructing reference sequences for these leviathan genomes, the three projects each used different approaches. The whole genome shotgun approach has been historically favored because it gives a rapid result. An alternative has been the expensive and time-consuming application of cloning to reduce to complexity of the problem for tractability or to obtain a better result [8]. Our combined strategy resulted in the most complete and contiguous conifer (gymnosperm) genome sequenced and assembled to date [9] with an assembled reference sequence consisting of 20.1 billion base pairs contained in scaffolds spanning 22.18 billion base pairs.

Our efforts to improve the quality of the loblolly pine reference genome sequence for conifers are continuing. The importance of a high quality and complete reference sequence for major taxonomic groups is well chronicled [92]. The loblolly pine reference genome was obtained from a single tree, 20-1010, for which significant and continuing open-access genome resources are freely available through the Dendrome Project and TreeGenes Database [93].

Materials and methods

Reference genotype tissue and DNA

All source material was obtained from grafted ramets of our reference *Pinus taeda* genotype 20-1010. Our haploid target megagametophyte was dissected from a wind-pollinated pine seed collected from a tree in a Virginia Department of Forestry seed orchard near Providence Forge, Virginia. Diploid tissue was obtained from needles collected from trees at the Erambert Genetic Resource Management Area near Brooklyn, Mississippi and the Harrison Experimental Forest near Saucier, Mississippi. A detailed description of the preparation and QC of DNA from these tissue samples is contained in [9].

Sequencing, assembly, and validation

A detailed description of the whole genome shotgun sequencing, assembly, and validation of the V1.0 and V1.01 loblolly pine genomes is contained in [9].

To compare the contiguity of our V1.01 whole genome shotgun assembly to contemporary conifer genome assemblies the scaffold sequences for white spruce genome [7] and Norway spruce [8] were obtained from Genbank.

CEGMA analysis of the core gene set [18] performed on the V1.0 and V1.01 loblolly pine genomes was obtained as described in [9]. Similarly, a Norway spruce analysis was performed with results consistent with those reported in [8]. The results for the white spruce assembly were taken directly from [7].

To assemble the mitochondrial genome, a subset of the WGS sequence consisting of 255 bp paired end MiSeq reads from four Illumina paired end libraries (median insert sizes: 325, 441, 565, and 637) were selected for an independent organelle assembly. The 28.5 Mbp of sequence, representing less than 0.3× nuclear genomic coverage, was assembled using SOAPdenovo2 (K = 127). The resulting contigs were aligned using nucmer to a database containing the loblolly pine chloroplast, sequencing vector, 102 BACs, and 50 complete plant mitochondria. Contigs were identified and labeled as mitochondrial if they aligned exclusively to existing mitochondrial sequence and had high coverage (> = 8×) and G + C% (> = 44%). The contigs were then combined with additional linking libraries, the LPMP_23 mate pair library and all DiTag libraries, and assembled a second time with SOAPdenovo2. Subsequently intra-scaffold gaps were closed using and GapCloser (v1.12). The

assembled sequences were iteratively scaffolded and gaps were closed until no assembly improvements could be made.

Annotation

The assembled genome was annotated with the MAKER-P pipeline [19] as described in [20]. Prior to gene prediction, the sequence was masked with similarity searches against RepBase and the Pine Interspersed Element Resource (PIER) [32]. Following the annotation, the TRIBE-MCL pipeline [94], was used to cluster the 399,358 protein sequences from 14 species into orthologous groups as described in [20].

Repetitive DNA content

Interspersed repeat detection was carried out in two stages, homology-based and *de novo* as described in [20]. For homology-based identification, RepeatMasker 3.3.0 [95] was run against the PIER 2.0 repeat library [32] for both the full genome and introns. REPET 2.0 [96] was implemented with the pipeline described in [32] for *de novo* repeat discovery. Only the 63 longest scaffolds were used in the all-vs-all alignment (approximately 1% of the genome). In addition, PIER 2.0, the spruce repeat database, and publicly available transcripts from *Pinus taeda* and *Pinus elliottii* were utilized as input for known repeat and host gene recognition.

To identify tandem repeats, Tandem Repeat Finder (v4.0.7b) [97] was run on both the genome and transcriptome as described in [20]. Filtering of multimeric repeats and overlaps with interspersed repeats, helped assess total tandem coverage and relative frequencies of specific satellites.

Data availability

Primary sequence data may be obtained from NCBI and is indexed under BioProject PRJNA174450. The whole genome shotgun sequence obtained for this assembly is available from the sequence read archive (SRA: SRP034079). The V1.0 and V1.01 genome sequences are available at [98]. Access to gene models, annotations, and Genome Browsers [99,100] are available through the TreeGenes database [93,101].

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DBN, JLW, KAS, AZ, DM, CAL, KM, PJDJ, JAY, SLS, and CHL designed the research. DBN, JLW, KAS, AZ, DP, MC, CC, MK, AEH, JDL, PJMG, HAVG, BYL, JJZ, WMD, SFS, LW, DG, GM, MR, CH, MY, JMD, KS, JFDD, WWL, RWW, RS, DM, CAL, KM, PJDJ, JAY, SLS, and CHL performed research and analyzed data. DBN, JLW, KAS, JMD, JFDD, RWW, RS, NW, PEM, CAL, SLS, and CHL wrote the article. All authors read and approved the final manuscript.

Acknowledgements

Funding for this project was made available through the USDA/NIFA (2011-67009-30030) award to DBN at University of California, Davis. We gratefully acknowledge the assistance of C. Dana Nelson at the USDA Forest Service Southern Research Station for providing and verifying the genotype of target tree material. We also wish to thank the management and staff of the DNA Technologies core facility at the UC Davis Genome Center for providing expert and timely HiSeq sequencing services to this project.

Author details

[1]Department of Plant Sciences, University of California, Davis, CA, USA. [2]Department of Evolution and Ecology, University of California, Davis, CA, USA. [3]Institute for Physical Sciences and Technology (IPST), University of Maryland, College Park, MD, USA. [4]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA. [5]Children's Hospital Oakland Research Institute, Oakland, CA, USA. [6]Department of Biology, Indiana University, Bloomington, IN, USA. [7]National Center for Genome Analysis Support, Indiana University, Bloomington, IN, USA. [8]Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT, USA. [9]School of Forest Resources and Conservation, Genetics Institute, University of Florida, Gainesville, FL, USA. [10]Southern Institute of Forest Genetics, USDA Forest Service, Southern Research Station, Saucier, MS, USA. [11]Warnell School of Forestry and Natural Resources, University of Georgia, Athens, GA, USA. [12]Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC, USA. [13]Department of Horticulture, Washington State University, Pullman, WA, USA. [14]Department of Ecosystem Science and Management, Texas A&M University, College Station, TX, USA.

Received: 11 February 2014 Accepted: 4 March 2014
Published: 4 March 2014

References

1. Farjon A: *World Checklist and Bibliography of Conifers.* 2nd edition. Richmond: Kew Publishing; 2001.
2. Farjon A: *The Natural History of Conifers.* Portland, OR: Timber Press; 2008.
3. Leslie AB, Beaulieu JM, Rai HS, Crane PR, Donoghue MJ, Mathews S: **Hemisphere-scale differences in conifer evolutionary dynamics.** *Proc Natl Acad Sci U S A* 2012, **109:**16217–16221.
4. Martínez-García PJ, Stevens K, Wegrzyn J, Liechty J, Crepeau M, Langley C, Neale D: **Combination of multipoint maximum likelihood (MML) and regression mapping algorithms to construct a high-density genetic linkage map for loblolly pine (*Pinus taeda* L.).** *Tree Genetics & Genomes* 2013, **9:**1529.
5. Zapata-Valenzuela J, Isik F, Maltecca C, Wegrzyn J, Neale D, McKeand S, Whetten R: **SNP markers trace familial linkages in a cloned population of *Pinus taeda*-prospects for genomic selection.** *Tree Genetics & Genomes* 2012, **8:**1307–1318.
6. Neale D, Langley C, Salzberg S, Wegrzyn J: **Open access to tree genomes: the path to a better forest.** *Genome Biol* 2013, **14:**120.
7. Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Saint Yuen MM, Keeling CI, Brand D, Vandervalk BP, Kirk H, Pandoh P, Moore RA, Zhao YJ, Mungall AJ, Jaquish B, Yanchuk A, Ritland C, Boyle B, Bousquet J, Ritland K, MacKay J, Bohlmann J, Jones SJM: **Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data.** *Bioinformatics* 2013, **29:**1492–1497.
8. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hallman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Kaller M, Luthman J, Lysholm F, Niittyla T, Olson A, Rilakovic N, Ritland C, Rossello JA, Sena J, *et al*: **The Norway spruce genome sequence and conifer genome evolution.** *Nature* 2013, **497:**579–584.
9. Zimin A, Stevens K, Crepeau M, Holtz-Morris A, Korabline M, Marçais G, Puiu D, Roberts M, Wegrzyn J, de Jong P, Neale D, Salzberg S, Yorke J, Langley C: **Sequencing and assembly of the 22-Gb loblolly pine genome.** *Genetics* 2014, **196:**875–890.
10. O'Brien IEW, Smith DR, Gardner RC, Murray BG: **Flow cytometric determination of genome size in *Pinus*.** *Plant Sci* 1996, **115:**91–99.
11. Zimin A, Marais G, Puiu D, Roberts M, Salzberg S, Yorke J: **The MaSuRCA genome assembler.** *Bioinformatics* 2013, **29:**2669–2677.

12. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XQH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang JH, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau C, McKusick VA, Zinder N: **The sequence of the human genome.** *Science* 2001, **291:**1304–1351.
13. Myers E: **Toward simplifying and accurately formulating fragment assembly.** *J Comput Biol* 1995, **2:**275–290.
14. Pevzner PA: **1-Tuple DNA sequencing: computer analysis.** *J Biomol Struct Dyn* 1989, **7:**63–73.
15. Li RQ, Fan W, Tian G, Zhu HM, He L, Cai J, Huang QF, Cai QL, Li B, Bai YQ, Zhang ZH, Zhang YP, Wang W, Li J, Wei FW, Li H, Jian M, Li JW, Zhang ZL, Nielsen R, Li DW, Gu WJ, Yang ZT, Xuan ZL, Ryder OA, Leung FCC, Zhou Y, Cao JJ, Sun X, Fu YG: **The sequence and *de novo* assembly of the giant panda genome.** *Nature* 2010, **463:**311–317.
16. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G: **Aggressive assembly of pyrosequencing reads with mates.** *Bioinformatics* 2008, **24:**2818–2824.
17. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40:**D290–D301.
18. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics* 2007, **23:**1061–1067.
19. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, Lei J, Achawanantakun R, Jiao D, Lawrence CJ, Ware D, Shiu SH, Childs KL, Sun Y, Jiang N, Yandell M: **MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations.** *Plant Physiol* 2014, **164:**513–524.
20. Wegrzyn JL, Liechty JD, Stevens KA, Wu L-S, Loopstra CA, Vasquez-Gross HA, Dougherty WM, Lin BY, Zieve JJ, Martinez-Garcia PJ, Holt C, Yandell M, Zimin AV, Yorke JA, Crepeau MW, Puiu D, Salzberg SL, de Jong PJ, Mockaitis K, Main D, Langley CH, Neale DB: **Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation.** *Genetics* 2014, **196:**891–909.
21. Goodstein DM, Shu SQ, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS: **Phytozome: a comparative platform for green plant genomics.** *Nucleic Acids Res* 2012, **40:**D1178–D1186.
22. Amborella genome project: **The *Amborella* genome and the evolution of flowering plants.** *Science* 2013, **342:**1467.
23. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res* 2002, **30:**1575–1584.
24. Glowacki S, Macioszek VK, Kononowicz AK: **R proteins as fundamentals of plant innate Immunity.** *Cell Mol Biol Lett* 2011, **16:**1–24.
25. Dao TTH, Linthorst HJM, Verpoorte R: **Chalcone synthase and its functions in plant resistance.** *Phytochem Rev* 2011, **10:**397–412.
26. Saavedra L, Svensson J, Carballo V, Izmendi D, Welin B, Vidal S: **A dehydrin gene in *Physcomitrella patens* is required for salt and osmotic stress tolerance.** *Plant J* 2006, **45:**237–249.
27. Velasco-Conde T, Yakovlev I, Majada JP, Aranda I, Johnsen O: **Dehydrins in maritime pine (*Pinus pinaster*) and their expression related to drought stress response.** *Tree Genetics & Genomes* 2012, **8:**957–973.
28. Mathieu M, Lelu-Walter MA, Blervacq AS, David H, Hawkins S, Neutelings G: **Germin-like genes are expressed during somatic embryogenesis and early development of conifers.** *Plant Mol Biol* 2006, **61:**615–627.
29. Carrillo MGC, Goodwin PH, Leach JE, Leung H, Cruz CMV: **Phylogenomic relationships of rice oxalate oxidases to the cupin superfamily and their association with disease resistance QTL.** *Rice* 2009, **2:**67–79.
30. Faini M, Prinz S, Beck R, Schorb M, Riches JD, Bacia K, Brugger B, Wieland FT, Briggs JAG: **The Structures of COPI-coated vesicles reveal alternate coatomer conformations and interactions.** *Science* 2012, **336:**1451–1454.
31. Pucadyil TJ, Schmid SL: **Conserved runctions of membrane active GTPases in coated vesicle formation.** *Science* 2009, **325:**1217–1220.
32. Wegrzyn J, Lin B, Zieve J, Dougherty W, Martínez-García P, Koriabine M, Holtz-Morris A, de Jong P, Crepeau M, Langley C, Puiu D, Salzberg S, Neale D, Stevens K: **Insights into the loblolly pine genome: characterization of BAC and fosmid sequences.** *PLoS ONE* 2013, **8:**e72439.
33. Kamm A, Doudrick RL, HeslopHarrison JS, Schmidt T: **The genomic and physical organization of Ty1-copia-like sequences as a component of large genomes in *Pinus elliottii* var *elliottii* and other gymnosperms.** *Proc Natl Acad Sci U S A* 1996, **93:**2708–2713.
34. Kossack DS, Kinlaw CS: **IFG, a gypsy-like retrotransposon in *Pinus* (Pinaceae), has an extensive history in pines.** *Plant Mol Biol* 1999, **39:**417–426.
35. Richards EJ, Ausubel FM: **Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*.** *Cell* 1988, **53:**127–136.
36. Leitch AR, Leitch IJ: **Ecological and genetic factors linked to contrasting genome dynamics in seed plants.** *New Phytol* 2012, **194:**629–646.
37. Aronen T, Ryynanen L: **Variation in telomeric repeats of Scots pine (*Pinus sylvestris* L.).** *Tree Genetics & Genomes* 2012, **8:**267–275.
38. Flanary BE, Kletetschka G: **Analysis of telomere length and telomerase activity in tree species of various life-spans, and with age in the bristlecone pine *Pinus longaeva*.** *Biogerontology* 2005, **6:**101–111.
39. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung D, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J: **SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler.** *GigaScience* 2012, **1:**18.
40. Parks M, Cronn R, Liston A: **Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes.** *BMC Biol* 2009, **7:**84.
41. Dangl JL, Horvath DM, Staskawicz BJ: **Pivoting the plant immune system from dissection to deployment.** *Science* 2013, **341:**746–751.
42. Flor HH: **Inheritance of pathogenicity in *Melampsora lini*.** *Phytopathology* 1942, **32:**653–669.
43. Flor HH: **Current status of the gene-for-gene concept.** *Annu Rev Phytopathol* 1971, **9:**275–296.
44. Griggs MM, Walkinshaw CH: **Diallel analysis of genetic-resistance to *Cronartium quercuum* f. sp. *fusiforme* in slash pine.** *Phytopathology* 1982, **72:**816–818.
45. Powers HR: **Pathogenic variation among single-aeciospore isolates of *Cronartium quercuum* f. sp. *fusiforme*.** *For Sci* 1980, **26:**280–282.
46. Wilcox PL, Amerson HV, Kuhlman EG, Liu BH, OMalley DM, Sederoff RR: **Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping.** *Proc Natl Acad Sci U S A* 1996, **93:**3859–3864.
47. Amerson HV, Nelson CD, Kubisiak TL: **R gene detection and mapping in fusiform rust disease.** *Forests* 2013. in review.
48. Quesada T, Resende MFRJ, Munoz P, Wegrzyn JL, Neale DB, Kirst M, Peter GF, Gezan SA, Nelson CD, Davis JM: **Mapping fusiform rust resistance genes within a complex mating design of loblolly pine.** *Forests* 2013, **5:**347–362.
49. Kinloch BB, Parks GK, Fowler CW: **White pine blister rust. Simply inherited resistance in sugar pine.** *Science* 1970, **167:**193–195.
50. Jermstad KD, Eckert AJ, Wegrzyn JL, Delfino-Mix A, Davis DA, Burton DC, Neale DB: **Comparative mapping in *Pinus*: sugar pine (*Pinus lambertiana* Dougl.) and loblolly pine (*Pinus taeda* L.).** *Tree Genetics & Genomes* 2011, **7:**457–468.
51. Kayihan GC, Huber DA, Morse AM, White TL, Davis JM: **Genetic dissection of fusiform rust and pitch canker disease traits in loblolly pine.** *Theor Appl Genet* 2005, **110:**948–958.
52. Whitham S, Dineshkumar SP, Choi D, Hehl R, Corr C, Baker B: **The product of the tobacco mosaic-virus resistance gene N: Similarity to toll and the interleukin-1 receptor.** *Cell* 1994, **78:**1101–1115.
53. Meyers BC, Morgante M, Michelmore RW: **TIR-X and TIR-NBS proteins: two new families related to disease resistance TIR-NBS-LRR proteins encoded in *Arabidopsis* and other plant genomes.** *Plant J* 2002, **32:**77–92.
54. Kubisiak TL, Amerson HV, Nelson CD: **Genetic interaction of the fusiform rust fungus with resistance gene *Fr1* in loblolly pine.** *Phytopathology* 2005, **95:**376–380.
55. Kubisiak TL, Anderson CL, Amerson HV, Smith JA, Davis JM, Nelson CD: **A genomic map enriched for markers linked to *Avr1* in *Cronartium quercuum* f.sp. *fusiforme*.** *Fungal Genet Biol* 2011, **48:**266–274.
56. McKeand S, Mullin T, Byram T, White T: **Deployment of genetically improved loblolly and slash pines in the south.** *J For* 2003, **101:**32–37.
57. Jermstad KD, Sheppard LA, Kinloch BB, Delfino-Mix A, Ersoz ES, Krutovsky KV, Neale DB: **Isolation of a full-length CC–NBS–LRR resistance gene analog candidate from sugar pine showing low nucleotide diversity.** *Tree Genetics & Genomes* 2006, **2:**76–85.
58. Willyard A, Syring J, Gernandt DS, Liston A, Cronn R: **Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*.** *Mol Biol Evol* 2007, **24:**90–101.

59. Burns RM, Honkala BH: **Silvics of North America.** In *Agriculture Handbook 654*. 2nd edition. Washington, DC: U.S: Department of Agriculture, Forest Service; 1990:877.
60. Mueller RC, Scudder CM, Porter ME, Trotter RT, Gehring CA, Whitham TG: **Differential tree mortality in response to severe drought: evidence for long-term vegetation shifts.** *J Ecol* 2005, **93:**1085–1093.
61. Brix H: **Determination of viability of loblolly pine seedlings after wilting.** *Bot Gaz* 1960, **121:**220–223.
62. Ferrell WK, Woodard ES: **Effects of seed origin on drought resistance of douglas-fir (*Pseudotsuga menziesii*) (Mirb) Franco.** *Ecology* 1966, **47:**499.
63. Hamanishi ET, Campbell MM: **Genome-wide responses to drought in forest trees.** *Forestry* 2011, **84:**273–283.
64. Newton R, Padmanabhan V, Loopstra C, Dias M: **Molecular responses to water-deficit stress in woody plants.** In *Handbook of Plant and Crop Stress*. Boca Raton, FL: M. Pessarakl, CRC Press; 1999:641–657.
65. Newton RJ, Funkhouser EA, Fong F, Tauer CG: **Molecular and physiological genetics of drought tolerance in forest species.** *For Ecol Manage* 1991, **43:**225–250.
66. Chang SJ, Puryear JD, Dias MADL, Funkhouser EA, Newton RJ, Cairney J: **Gene expression under water deficit in loblolly pine (*Pinus taeda*): isolation and characterization of cDNA clones.** *Physiol Plant* 1996, **97:**139–148.
67. Frankel N, Carrari F, Hasson E, Iusem ND: **Evolutionary history of the *Asr* gene family.** *Gene* 2006, **378:**74–83.
68. Iusem ND, Bartholomew DM, Hitz WD, Scolnik PA: **Tomato (*Lycopersicon esculentum*) transcript induced by water-deficit and ripening.** *Plant Physiol* 1993, **102:**1353–1354.
69. Cortes AJ, Chavarro MC, Madrinan S, This D, Blair MW: **Molecular ecology and selection in the drought-related Asr gene polymorphisms in wild and cultivated common bean (*Phaseolus vulgaris* L.).** *BMC Genet* 2012, **13:**58.
70. Fischer I, Camus-Kulandaivelu L, Allal F, Stephan W: **Adaptation to drought in two wild tomato species: the evolution of the *Asr* gene family.** *New Phytol* 2011, **190:**1032–1044.
71. Shen G, Pang YZ, Wu WS, Deng ZX, Liu XF, Lin J, Zhao LX, Sun XF, Tang KX: **Molecular cloning, characterization and expression of a novel *Asr* gene from *Ginkgo biloba*.** *Plant Physiol Biochem* 2005, **43:**836–843.
72. Lorenz WW, Alba R, Yu YS, Bordeaux JM, Simoes M, Dean JFD: **Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (*P. taeda* L.).** *BMC Genomics* 2011, **12:**264.
73. Lorenz WW, Sun F, Liang C, Kolychev D, Wang HM, Zhao X, Cordonnier-Pratt MM, Pratt LH, Dean JFD: **Water stress-responsive genes in loblolly pine (*Pinus taeda*) roots identified by analyses of expressed sequence tag libraries.** *Tree Physiol* 2006, **26:**1–16.
74. Gonzalez-Martinez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB: **Association genetics in *Pinus taeda* L. I. Wood property traits.** *Genetics* 2007, **175:**399–409.
75. Paiva JAP, Garces M, Alves A, Garnier-Gere P, Rodrigues JC, Lalanne C, Porcon S, Le Provost G, Perez DD, Brach J, Frigerio JM, Claverol S, Barre A, Fevereiro P, Plomion C: **Molecular and phenotypic profiling from the base to the crown in maritime pine wood-forming tissue.** *New Phytol* 2008, **178:**283–301.
76. Joosen RVL, Lammers M, Balk PA, Bronnum P, Konings MCJM, Perks M, Stattin E, Van Wordragen MF, van der Geest AHM: **Correlating gene expression to physiological parameters and environmental conditions during cold acclimation of Pinus sylvestris, identification of molecular markers using cDNA microarrays.** *Tree Physiol* 2006, **26:**1297–1313.
77. Eveno E, Collada C, Guevara MA, Leger V, Soto A, Diaz L, Leger P, Gonzalez-Martinez SC, Cervera MT, Plomion C, Garnier-Gere PH: **Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses.** *Mol Biol Evol* 2008, **25:**417–437.
78. Grivet D, Sebastiani F, González-Martínez SC, Vendramin GG: **Patterns of polymorphism resulting from long-range colonization in the Mediterranean conifer Aleppo pine.** *New Phytol* 2009, **184:**1016–1028.
79. Donaldson LA: **Lignification and lignin topochemistry - an ultrastructural view.** *Phytochemistry* 2001, **57:**859–873.
80. Sperry JS, Hacke UG, Pittermann J: **Size and function in conifer tracheids and angiosperm vessels.** *Am J Bot* 2006, **93:**1490–1500.
81. Sarkanen KV, Ludwing CH: *Lignins: Occurrence, Formation, Structure and Reactions*. Wiley Interscience: New York, NY; 1971.
82. Weng JK, Akiyama T, Bonawitz ND, Li X, Ralph J, Chapple C: **Convergent evolution of syringyl lignin biosynthesis via distinct pathways in the lycophyte *Selaginella* and flowering plants.** *Plant Cell* 2010, **22:**1033–1045.
83. Scheller HV, Ulvskov P: **Hemicelluloses.** *Annu Rev Plant Biol* 2010, **61:**263–289.
84. Shi R, Sun YH, Li QZ, Heber S, Sederoff R, Chiang VL: **Towards a systems approach for lignin biosynthesis in *Populus trichocarpa*: transcript abundance and specificity of the monolignol biosynthetic genes.** *Plant Cell Physiol* 2010, **51:**144–163.
85. Osakabe K, Tsao CC, Li LG, Popko JL, Umezawa T, Carraway DT, Smeltzer RH, Joshi CP, Chiang VL: **Coniferyl aldehyde 5-hydroxylation and methylation direct syringyl lignin biosynthesis in angiosperms.** *Proc Natl Acad Sci U S A* 1999, **96:**8955–8960.
86. Vanholme R, Cesarino I, Rataj K, Xiao YG, Sundin L, Goeminne G, Kim H, Cross J, Morreel K, Araujo P, Welsh L, Haustraete J, McClellan C, Vanholme B, Ralph J, Simpson GG, Halpin C, Boerjan W: **Caffeoyl shikimate esterase (CSE) is an enzyme in the lignin biosynthetic pathway in *Arabidopsis*.** *Science* 2013, **341:**1103–1106.
87. Gion JM, Carouche A, Deweer S, Bedon F, Pichavant F, Charpentier JP, Bailleres H, Rozenberg P, Carocha V, Ognouabi N, Verhaegen D, Grima-Pettenati J, Vigneron P, Plomion C: **Comprehensive genetic dissection of wood properties in a widely-grown tropical tree: *eucalyptus*.** *BMC Genomics* 2011, **12:**301.
88. Thumma BR, Nolan MR, Evans R, Moran GF: **Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in *Eucalyptus* spp.** *Genetics* 2005, **171:**1257–1265.
89. Yu Q, Li B, Nelson CD, McKeand SE, Batista VB, Mullin TJ: **Association of the *cad-n1* allele with increased stem growth and wood density in full-sib families of loblolly pine.** *Tree Genetics & Genomes* 2006, **2:**98–108.
90. Demura T, Fukuda H: **Transcriptional regulation in wood formation.** *Trends Plant Sci* 2007, **12:**64–70.
91. Melzer S, Lens F, Gennen J, Vanneste S, Rohde A, Beeckman T: **Flowering-time genes modulate meristem determinacy and growth form in *Arabidopsis thaliana*.** *Nat Genet* 2008, **40:**1489–1492.
92. Fraser CM, Eisen JA, Nelson KE, Paulsen IT, Salzberg SL: **The value of complete microbial genome Sequencing (you get what you pay for).** *J Bacteriol* 2002, **184:**6403–6405.
93. Wegrzyn JL, Lee JM, Tearse BR, Neale DB: **TreeGenes: a forest tree genome database.** *Int J Plant Genom* 2008, **2008:**412875.
94. van Dongen S, Abreu-Goodger C: **Using MCL to extract clusters from networks.** *Bacterial Molecular Networks* 2012, **804:**281–295.
95. **RepeatMasker.** [http://www.repeatmasker.org/]
96. Flutre T, Duprat E, Feuillet C, Quesneville H: **Considering transposable element diversification in *de novo* annotation approaches.** *PLoS ONE* 2011, **6:**e16526.
97. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573–580.
98. **Pine reference sequences.** [http://www.pinegenome.org/pinerefseq]
99. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE: **Web Apollo: a web-based genomic annotation editing platform.** *Genome Biol* 2013, **14:**R93.
100. Stein LD, Mungall C, Shu SQ, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12:**1599–1610.
101. **TreeGenes. A forest tree genome database.** [http://dendrome.ucdavis.edu/treegenes]

doi:10.1186/gb-2014-15-3-r59
Cite this article as: Neale *et al.*: Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology* 2014 **15**:R59.