Neale *et al. Genome Biology* 2014, **15**:R59
http://genomebiology.com/2014/15/3/R59

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]

Sequencing and Assembly of the 22-Gb Loblolly Pine Genome

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

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

Unique Features of the Loblolly Pine (*Pinus taeda* L.) Megagenome Revealed Through Sequence Annotation

Jill L. Wegrzyn,*[,1] John D. Liechty,* Kristian A. Stevens,[†] Le-Shin Wu,[‡] Carol A. Loopstra,[§] Hans A. Vasquez-Gross,* William M. Dougherty,[†] Brian Y. Lin,* Jacob J. Zieve,* Pedro J. Martínez-García,* Carson Holt, Mark Yandell,** Aleksey V. Zimin,[††] James A. Yorke,[††,‡‡] Marc W. Crepeau,[†] Daniela Puiu,[§§] Steven L. Salzberg,[§§] Pieter J. de Jong,*** Keithanne Mockaitis,[†††] Doreen Main,[§§§] Charles H. Langley,[†] and David B. Neale***

*Department of Plant Sciences, and [†]Department of Evolution and Ecology, University of California, Davis, California 95616, [‡]National Center for Genome Analysis Support, Indiana University, Bloomington, Indiana 47405, [§]Department of Ecosystem Science and Management, Texas A&M University, College Station, Texas 77843, **Department of Human Genetics, University of Utah, Salt Lake City, Utah 84112, [††]Institute for Physical Sciences and Technology, and [‡‡]Departments of Mathematics and Physics, University of Maryland, College Park, Maryland 20742, [§§]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University, Baltimore, Maryland 21205, ***Children's Hospital Oakland Research Institute, Oakland, California 94609, [†††]Department of Biology, Indiana University, Bloomington, Indiana 47405, and [§§§]Department of Horticulture, Washington State University, Pullman, Washington 99163

J Club 16.05.2014

Miks sekveneerida okaspuu genoomi?

- Fülogeneetiline kaardistamine
  - Seni väga puudulik
  - Paljasseemnetaimede suurtest harudest vanim (300 milj. aastat)
- Parasvöötmes laialt levivad
- Majanduslik olulisus (tõrvikumänni ümarpuit ~18% maailma toodangust)
- Abiks kõrgemate taimede geneetilise mitmekesisuse uurimisel

Taimegenoomide suurused

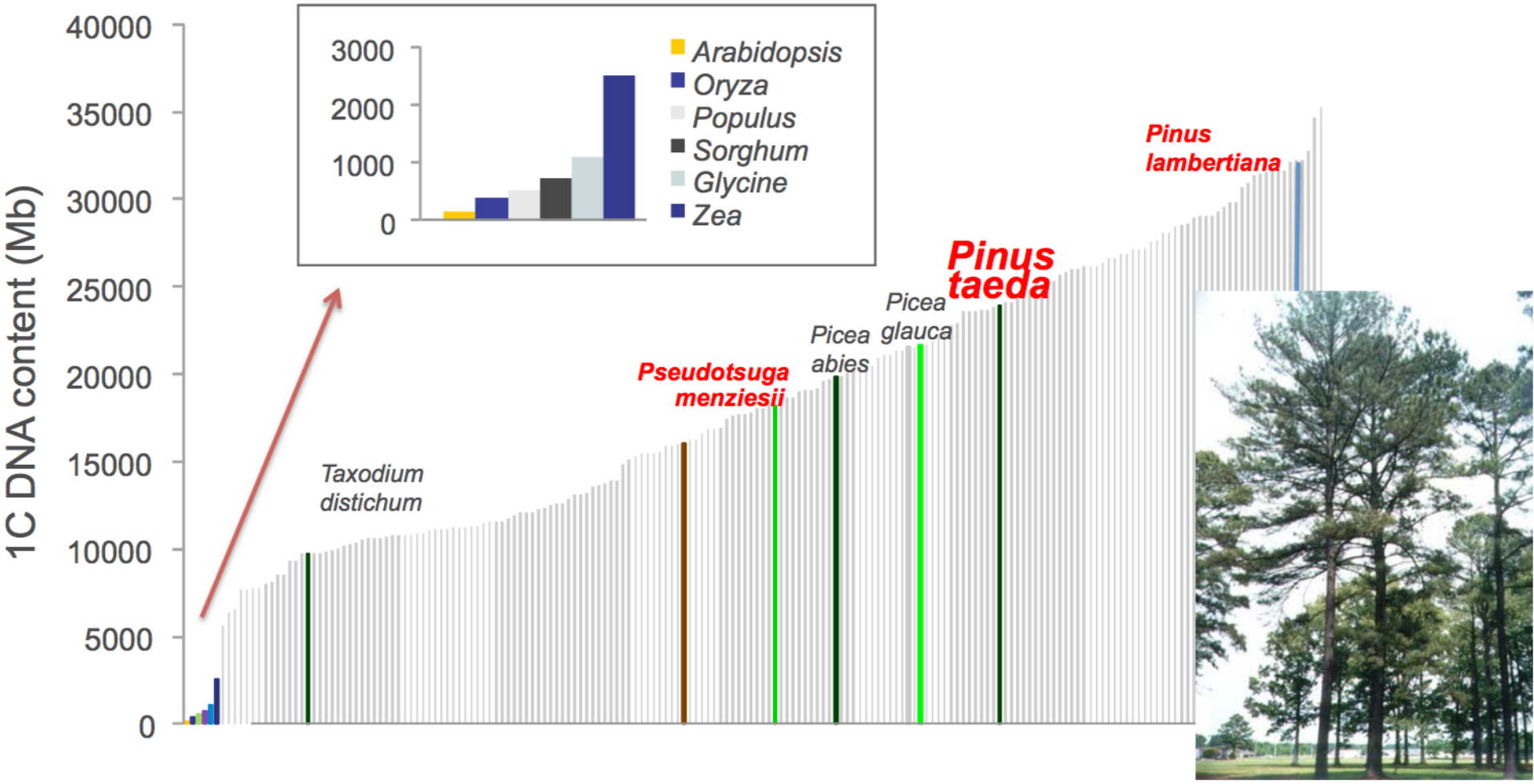


Image Credit: Modified from Daniel Peterson, Mississippi State University

DNA hankimine

Haploid

Haploid megagametophyte tissue
1N
Shotgun sequenced

Diploid

Diploid needle tissue
2N
40 Kb cloned fosmids, pooled and sequenced

Figure Credit: Nicholas Wheeler, University of California, Davis

DiTag raamatukogu loomine

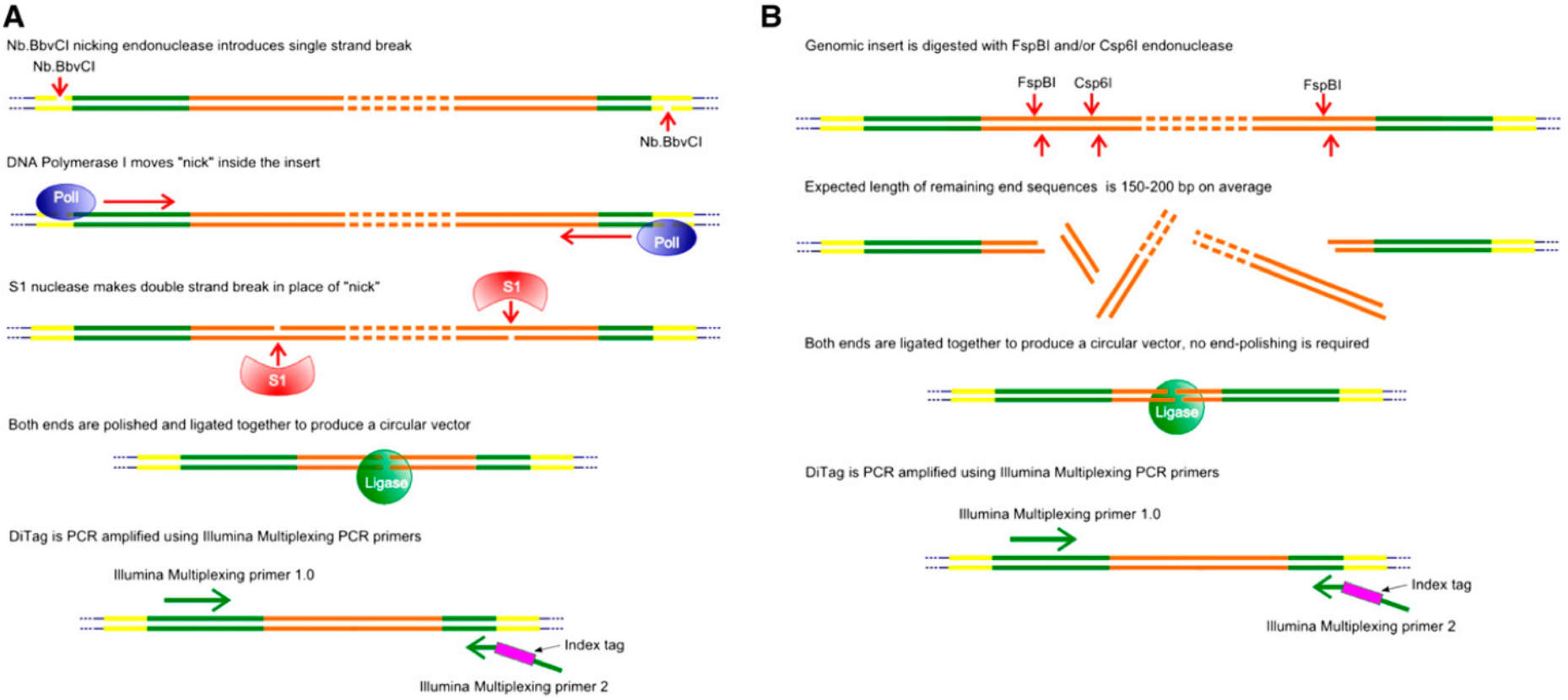


Figure 3 Schematic for our two methods for converting a fosmid library into an Illumina compatible DiTag library using the fosmid vector created for this project. (A) A nick translation approach, similar to the approach used in Williams *et al.* (2012), was implemented for approximately one-half of the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

Lugemite raamatukogud

Table 1 Overview of WGS sequence obtained for this project

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

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

- Lugemite arv: >16 miljardit
- Lugemite kogupikkus: >1.7 triljonit bp

MaSuRCA assembly

- MaSuRCA assemblerit jooksutati 64 tuumalisel 1 TB mäluga serveris
- Lugemite vigade parandamine (QuORUM): 800 GB/10 päeva
- Superlugemite loomine + mate pair filtreerimine: 400 GB/11 päeva
- Kontiigide ja scaffoldide loomine (CABOG): 450 GB / >60 päeva
- Aukude täitmine varieeruva pikkusega k-meridest loodud superlugemitega: 300 GB/8 päeva

Superlugemid (1)

- Jellyfishi abil loeti kõigi lugemite pealt kokku kõigi k-mer'ide sagedused (pikkusega näiteks 30)

AGCTGACTGACTGGTAACAA
AGCTGACTGA
GCTGACTGAC
……

- Kasutatkse ainult k-mer'e, mille arv > lävend T (näit. T=3)

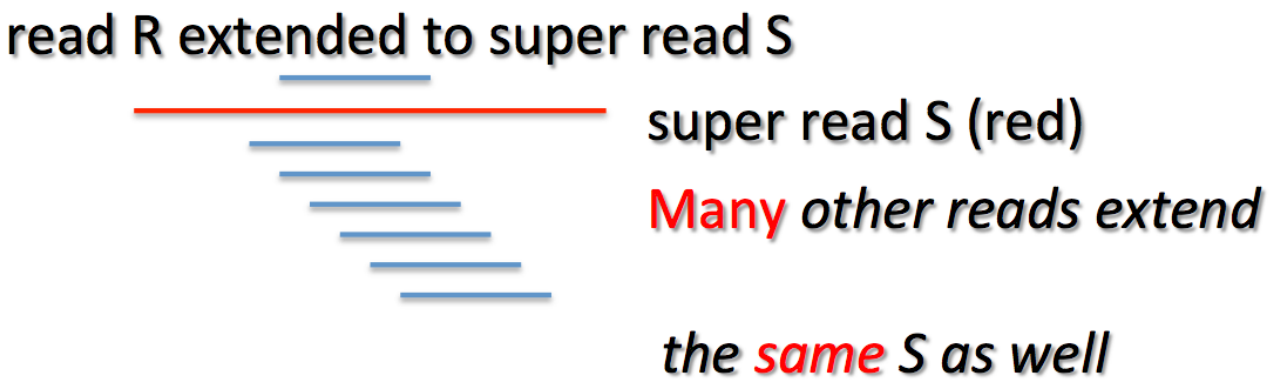
- Idee on teha lugemid pikemaks, mitte tükeldada need k-mer'ideks

Superlugemid (2)

- Lugemi pikendamine:

```
            CGACTGACCAGATGACCATGACAGATACATGGT
extend 5     GACTGACCAG                    ATACATGGTA 10  stop
extend 3    CGACTGACCA                     ATACATGGTC 2   stop
```

- Illumina sekveneerimisprojekti tulemuseks >50x katvus. Seega 100bp lugemite korral algab keskmiselt igas positsioonis uus lugem



Andmete vähendamine

- Alguses 16 miljardit lugemit keskmise pikkusega 120 bp
- Pärast filtreerimist ja superlugemite loomist: 150 miljonit lugemit (>100 korra vähem)
- Superlugemid sisaldasid 52 Gbp
- 50% järjestustest olid 500 bp pikad või pikemad

Assembly statistika

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

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

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

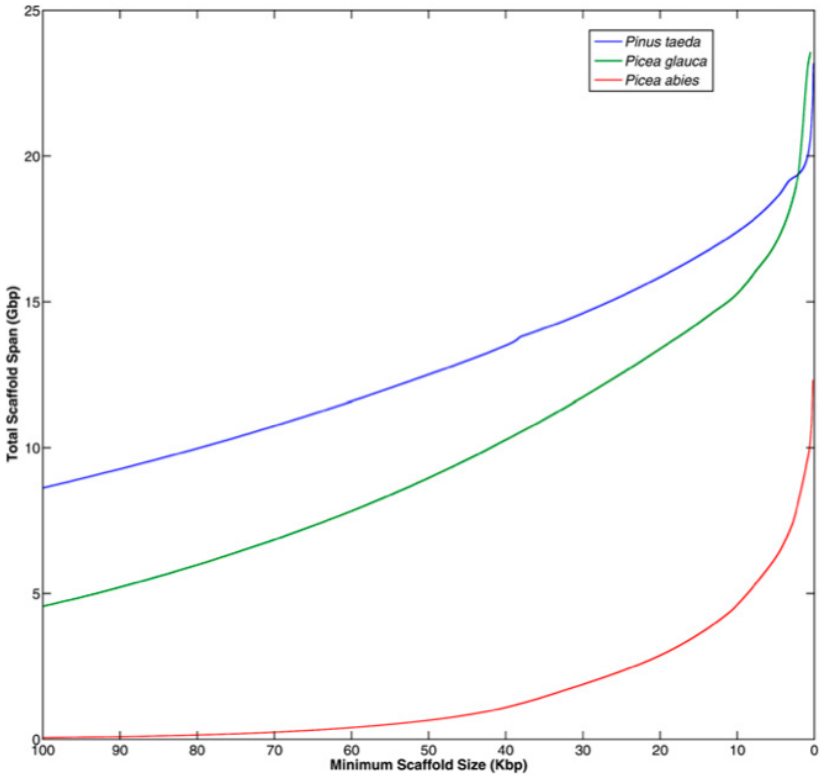


Figure C1 The contiguity of the loblolly pine v1.01 assembly is compared to contemporary draft conifer assemblies. Total scaffold span is plotted against a minimum scaffold size threshold. Loblolly pine is relatively more complete when considering large-gene-sized (>10 kbp) scaffolds. This is reflected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

RNA-seq raamatukogud eri kudedest

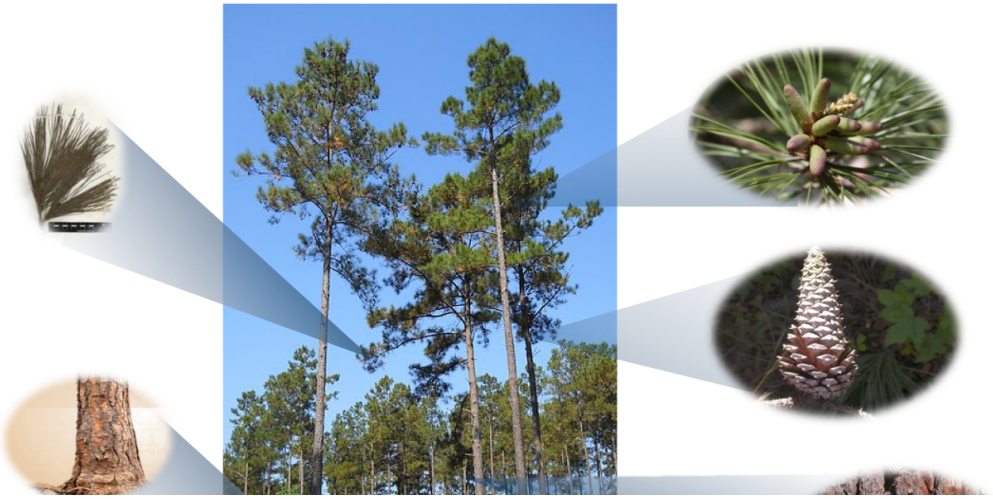


Table 1 Mapping EST/transcriptome resources against *P. taeda* version 1.01 genome

Project	Total sequence	Identity	Coverage	Unique hits	Non-unique hits	Total % mapped
P. taeda (reclustered ESTs)	45,085	98	98	26,700	712	60.8
P. taeda (reclustered ESTs)	45,085	98	95	29,676	1,845	69.91
P. taeda (reclustered ESTs)	45,085	95	95	31,324	2,074	74.01
P. taeda (reclustered ESTs)	45,085	95	50	29,744	5,486	78.14
P. taeda (*de novo*)	83,285	98	98	29,262	1,731	35.21
P. taeda (*de novo*)	83,285	98	95	42,822	5,130	57.58
P. taeda (*de novo*)	83,285	95	95	43,972	5,409	59.29
P. taeda (*de novo*)	83,285	95	50	44,469	28,116	87.15
P. palustris (454)	16,832	95	95	11,242	719	71.06
P. palustris (454)	16,832	95	50	11,181	1,949	78.06
P. lambertiana (454 + RNASeq)	40,619	95	95	13,134	317	33.11
P. lambertiana (454 + RNASeq)	40,619	95	50	23,376	3,792	66.88
P. banksiana (TreeGenes clusters)	13,040	95	95	9,703	513	78.34
P. banksiana (TreeGenes clusters)	13,040	95	50	9,470	1,473	83.92
P. contorta (TreeGenes clusters)	13,570	95	95	9,575	396	73.48
P. contorta (TreeGenes clusters)	13,570	95	50	9,534	1,083	78.24
P. pinaster (TreeGenes clusters)	15,648	95	95	9,738	943	68.26
P. pinaster (TreeGenes clusters)	15,648	95	50	10,221	2,491	81.24

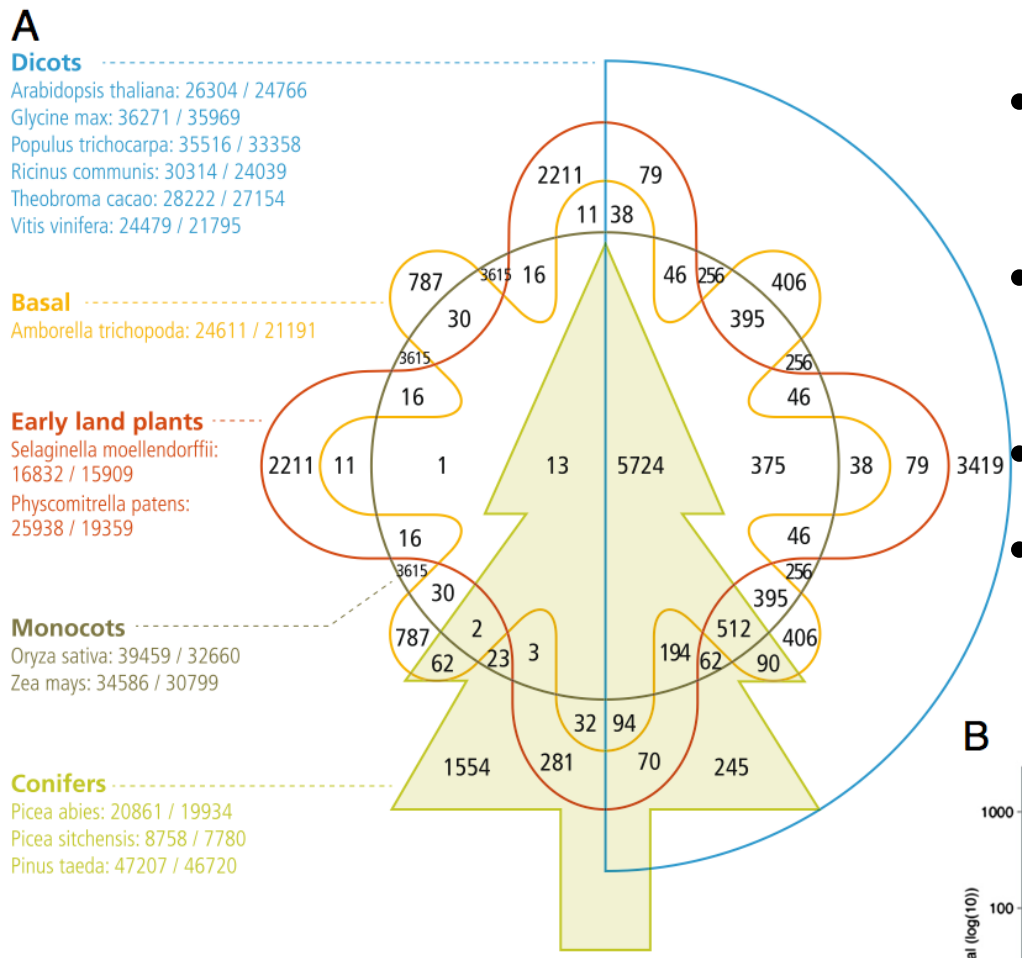
Tõrvikumänni geenid

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].

- Unikaalsed transkriptid (*de novo*): 83,285 (42,822 genoomis – iden 98%, cov 95%)
- MAKER-P geenimudelid kontiigidelt: 50,172 (20,412 kattusid transkriptidega – iden 98%, cov 98%)



- 20,646 geeniperekonda kahes või enamas liigis
- 1,476 perekonda kõigis liikides
- 1,554 okaspuude spetsiifilised
- 159 tõrvikumännile spetsiifilised

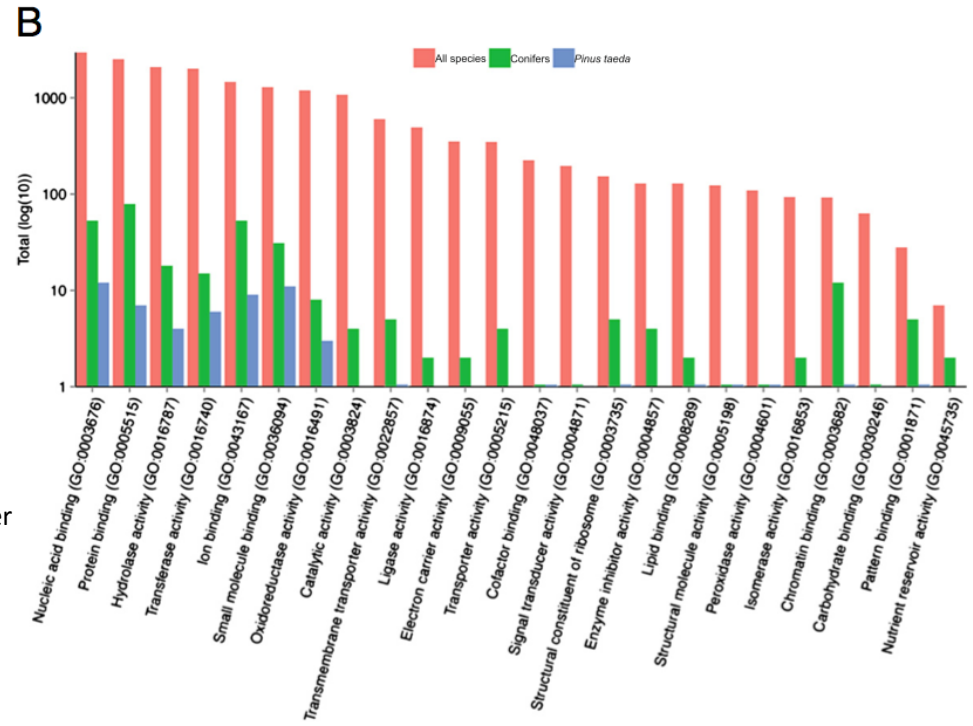
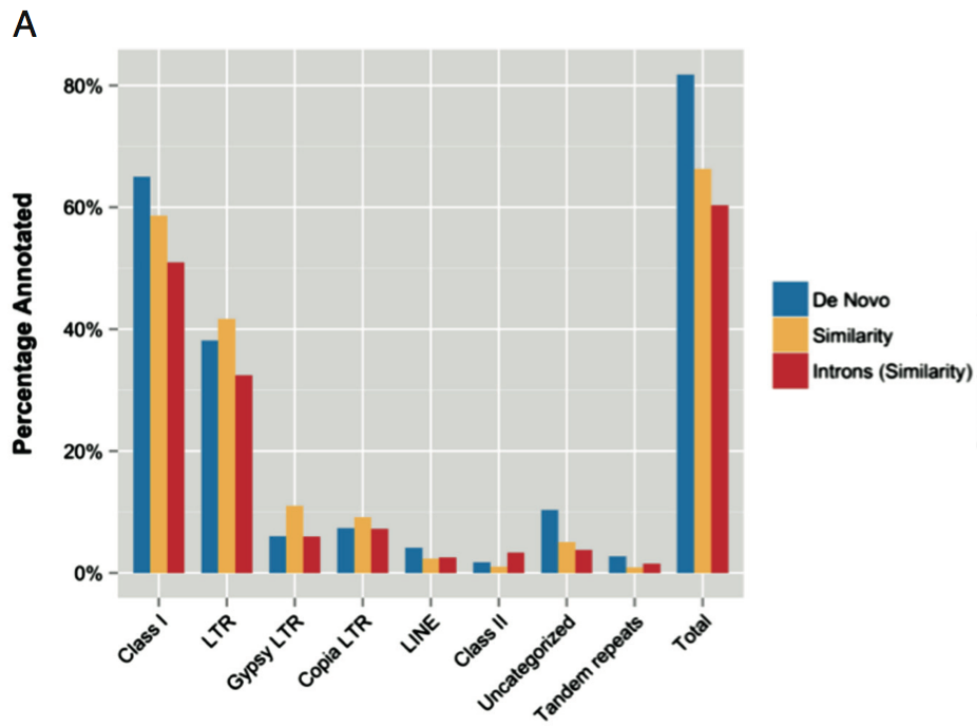


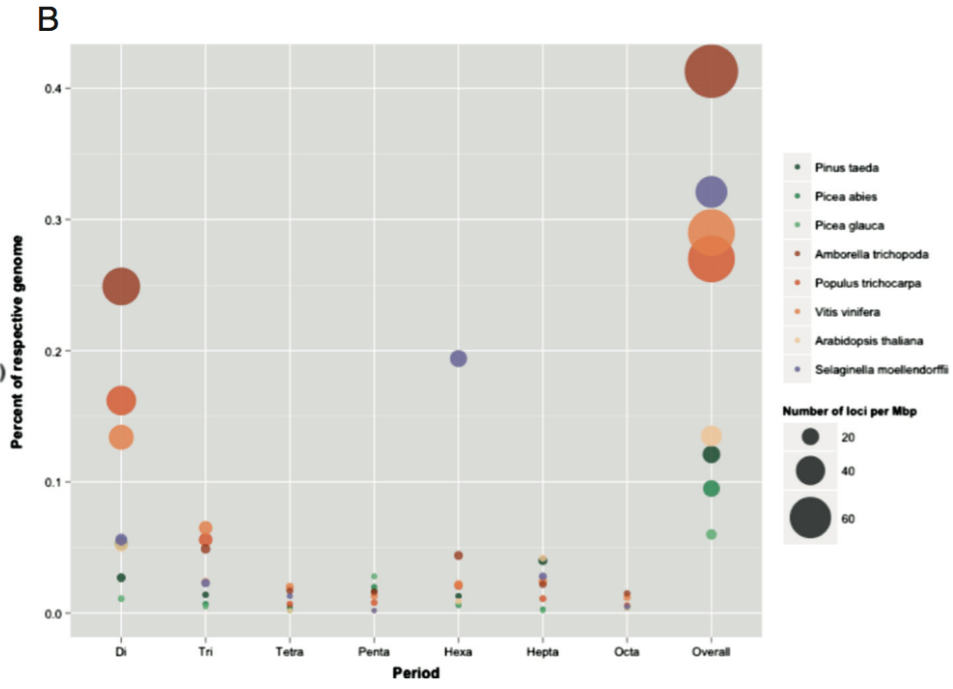
Figure 2 Unique gene families and Gene Ontology term assignments. (A) Identification of orthologous groups of genes for 14 species split into five categories: conifers (Picea abies, Picea sitchensis, and Pinus taeda), monocots (Oryza sativa and Zea mays), dicots (Arabidopsis thaliana, Glycine max, Populus 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).

Tõrvikumänni geenid (3)

- Okaspuude spetsiifilised geeniperekonnad sisaldavad muu hulgas:
  - transkriptsioonifaktoreid (Myb, WRKY ha HLH),
  - oksüoreduktaase (tsütokroom p450),
  - haigusvastaseid valke (NB-ARC, TNL),
  - stressiga seotud (lp3) ja puidukoe moodustamisega seotud geenide homoloogid

Kordused





- 82% tõrvikumänni genoomist on kordused (*de novo* + sarnased)
- Erinevalt õistaimedest on ainult 2.86% on tandeemsed, enamik retrotransposoonid (kuuskedel ~2.5%)
- Intronites ~60 % kortdustest
- Mändidel pikad telomeerid (kuni 57 Kbp) – heptanukleotiidsete mikrosatelliitide suhteliselt suurem hulk