

# Annotation of the loblolly pine megagenome

CATTAGCTCTGGTCATCAAGTCATCCATGATTAGCT

**Jill Wegrzyn**

Department of Ecology and Evolutionary Biology

University of Connecticut



# Annotation of the Genome

## Overview

- Alignment of existing resources
- Alignment of *de novo* transcriptome assembly
- Gene space prediction
- Gene family analysis
- Repeat Sequence
  - Repeat Library
  - Interspersed
  - Tandem
- Database Resources
- Community Annotation



### Assembly v1.0 (March 2013)

- Approximately 65X coverage
- Total Sequence: 20.1 Gbp
- Total Contig: 2.3 Gbp
- N50 Contig: 8.2 Kbp (11.6 m)
- Total Scaffold: 17.8 Gbp
- N50 Scaffold: 30.7 Kbp (4.8 m)

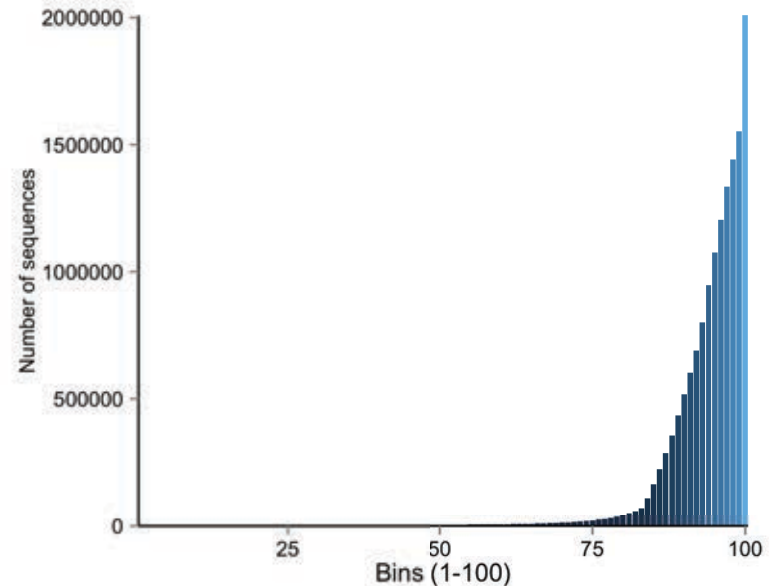
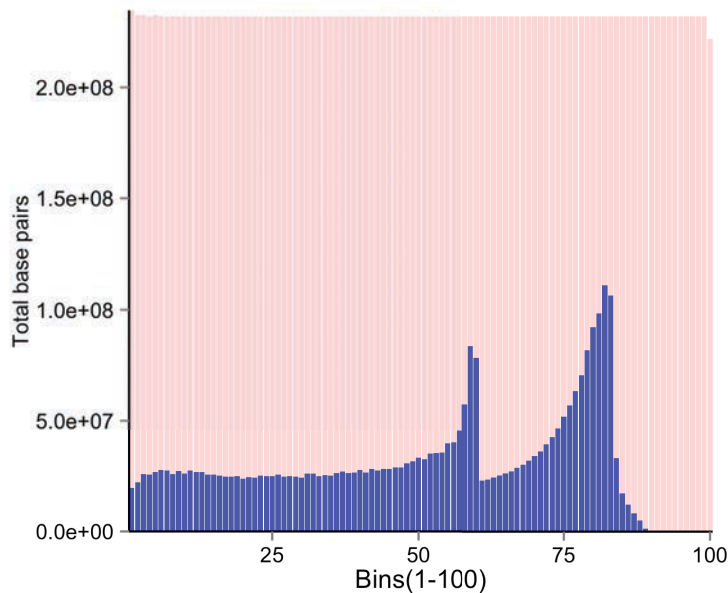
### Assembly v1.01

- Total Sequence: 20.1 Gbp
- N50 Contig 8.2 Kbp
- N50 Scaffold: 66.9 Kbp

# Annotation of the Genome

## Mapping Existing Datasets to the Genome

- Mapping resources initially on unmasked genome (1.0 and 1.01)
  - Divide the sequence into 100 bins of equal size
    - Descending order of scaffold length
      - Parallelize
      - Examine the effects of masking, fragmentation, repetitive sequence



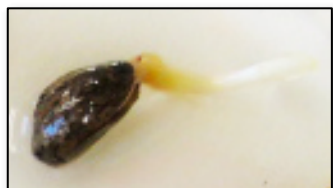
# Annotation of the Genome

## Mapping Existing Datasets to the Genome

Project	Total sequence	Identity	Coverage	Unique hits	Non-unique hits	Total percent mapped
Pinus taeda (reclustered ESTs)	45,085	98	98	26,700	712	60.8
Pinus taeda (reclustered ESTs)	45,085	98	95	29,676	1,845	69.91
Pinus taeda (reclustered ESTs)	45,085	95	95	31,324	2,074	74.01
Pinus palustris (454)	16,832	95	95	11,242	719	71.06
Pinus palustris (454)	16,832	95	50	11,181	1,949	78.06
Pinus lambertiana (454 + RNASeq)	40,619	95	95	13,134	317	33.11
Pinus lambertiana (454 + RNASeq)	40,619	95	50	23,376	3,792	66.88
Pinus banksiana (Treegenes clusters)	13,040	95	95	9,703	513	78.34
Pinus banksiana (Treegenes clusters)	13,040	95	50	9,470	1,473	83.92
Pinus contorta (Treegenes clusters)	13,570	95	95	9,575	396	73.48
Pinus contorta (Treegenes clusters)	13,570	95	50	9,534	1,083	78.24
Pinus pinaster (Treegenes clusters)	15,648	95	95	9,738	943	68.26
Pinus pinaster (Treegenes clusters)	15,648	95	50	10,221	2,491	81.24

# Progressive Transcript Profiling

First comprehensive transcriptome generated (27 unique RNA libraries)



**Early Development**  
seeds  
young seedlings



**Reproductive Development**  
megastrobili  
microstrobili



**Vegetative Organs**  
vegetative buds  
candles  
stems  
needles  
roots



**Early Stress Signaling Responses**  
cold  
heat  
elevated UV  
compression

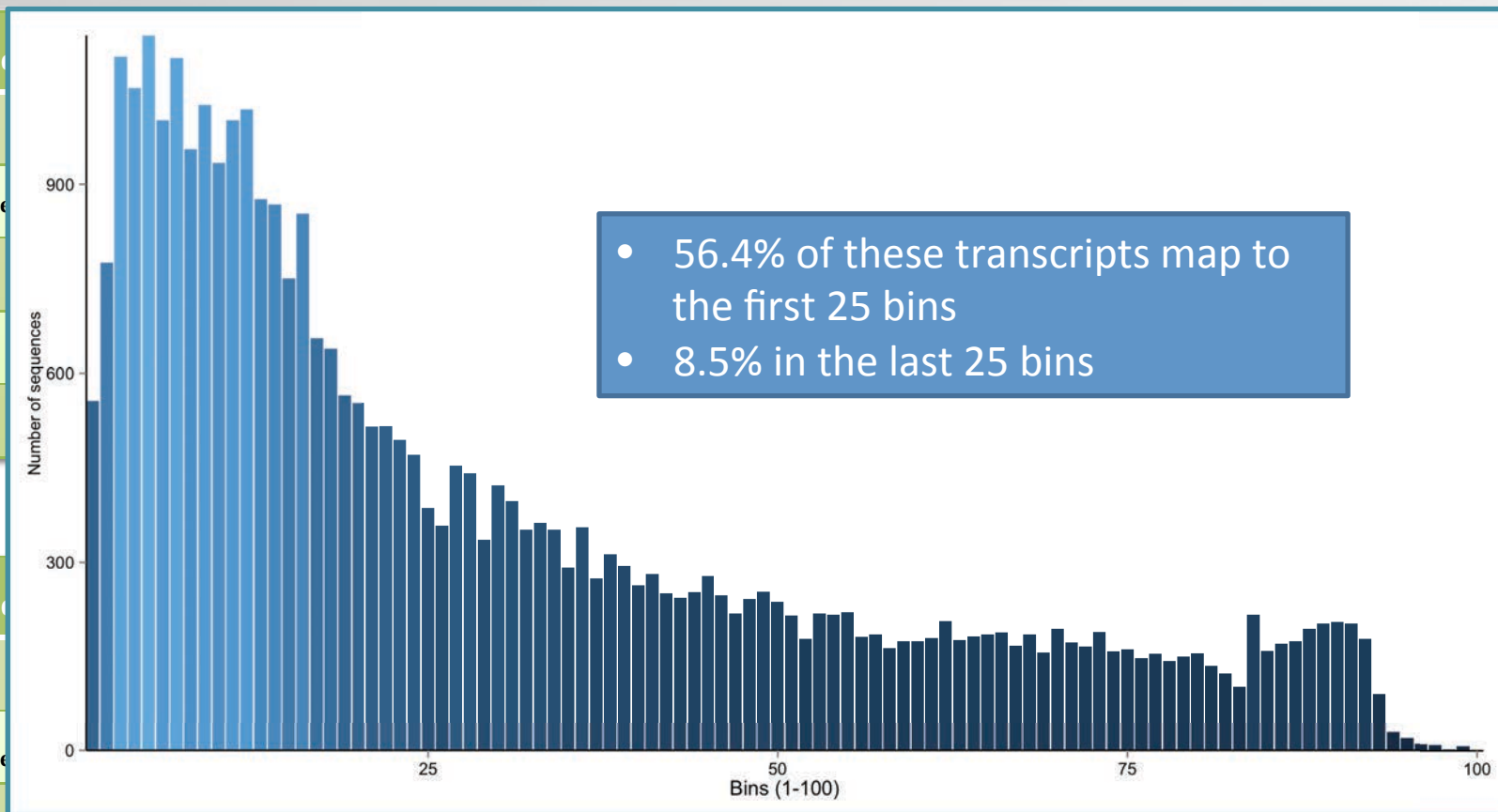


# Transcriptome Assembly

- Considerable variation in *de novo* transcriptome assemblies
  - Influenced by heterozygosity, sampling methods, normalization strategies
  - Used a **compare and compete** methodology to select the final transcripts
    - Independent, parallel assemblies from 2 Trinity versions and 6 Velvet/Oasis (different k-mer sizes)
  - **EvidentialSuite** – Full length, unique, protein coding (87,241)
  - 83,285 transcripts aligned at least partially to the genome (v1.01)

# Aligning the loblolly pine transcriptome

Results against 1.0 and 1.01



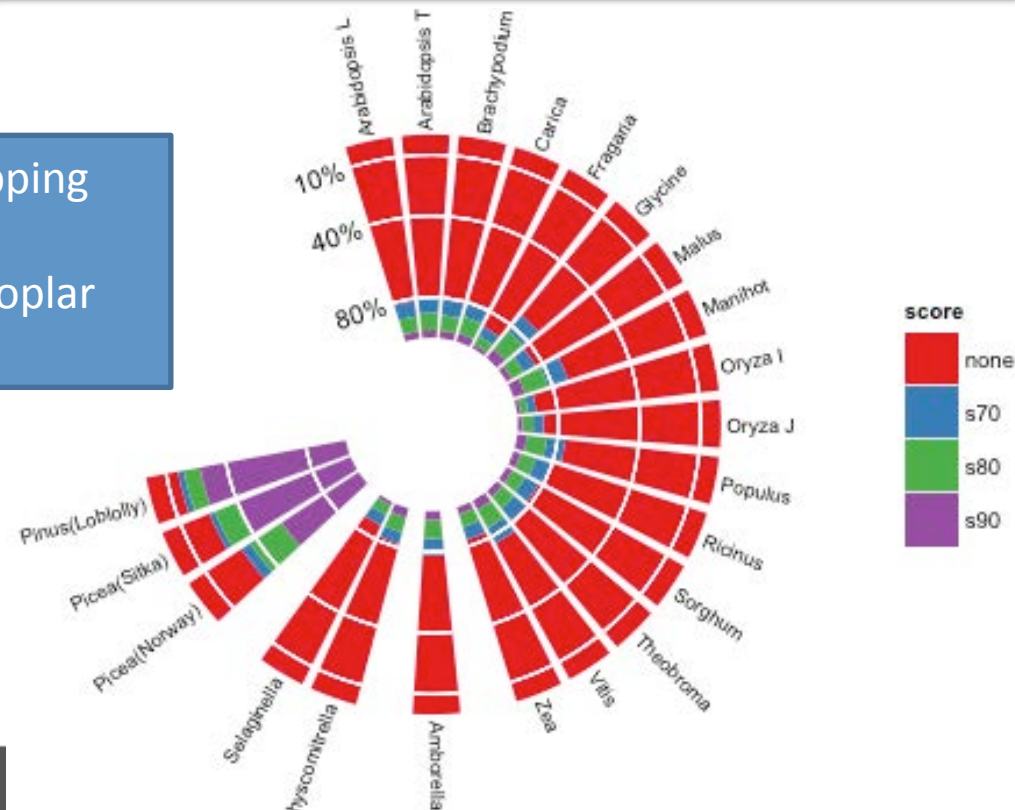
83,285 uni						
Method						t mapped
Blat/Exonerate						
GMAP						9
						5
NUCmer						4
83,285 uni						
Method						ent mapped
Blat/Exonerate						21
	83,285	98	95	42,822	5,130	57.58
	83,285	95	95	43,972	5,409	59.29
	83,285	95	50	44,469	28,116	87.15

# Annotation of the Genome

## Mapping Existing Datasets to the Genome

Full-length proteins	Total sequence	Unique hits	Non-unique hits	Total percent mapped	Data source
Picea abies	22,070	11,580	3,638	68.95	Nystedt et al. 2013
Picea sitchensis	10,793	6,516	1,574	74.95	Genbank
Pinus taeda	83,285	45,656	24,427	84.15	Current assembly
PLAZA (24 species)	653,613	90,149	19,492	16.77	Van Bel et al. 2012

Marginally better mapping rates for soybean and cassava followed by poplar and grape.





# MAKER-P

## Identification of the Gene Space

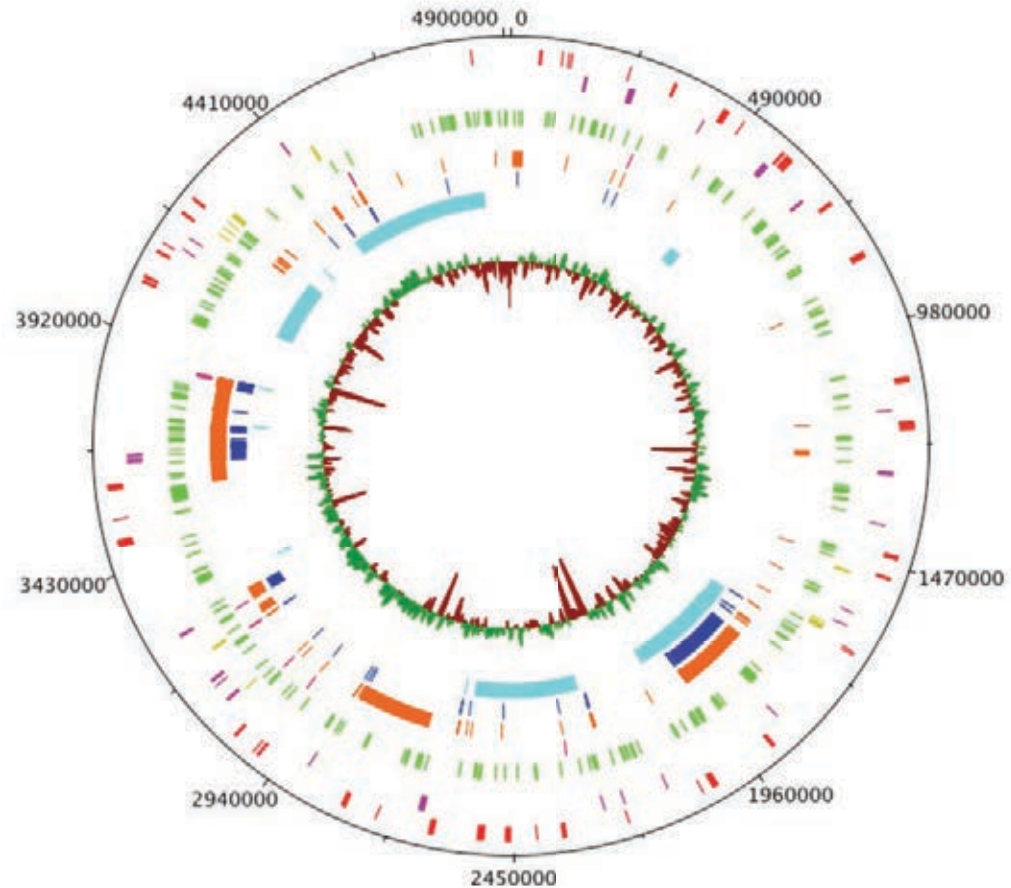


Provided *ab initio* and reference guided annotations

Texas Advanced Computing Center (TACC)  
8,640 CPUs (216 CPUs for each of 40 jobs) 17 hours total run time

Produced over 90,000 gene models that were further filtered to reduce the false positives associated with Pseudogenes

- Multi-exonic
- Recognizable protein domain
- One of more forms of supporting Evidence

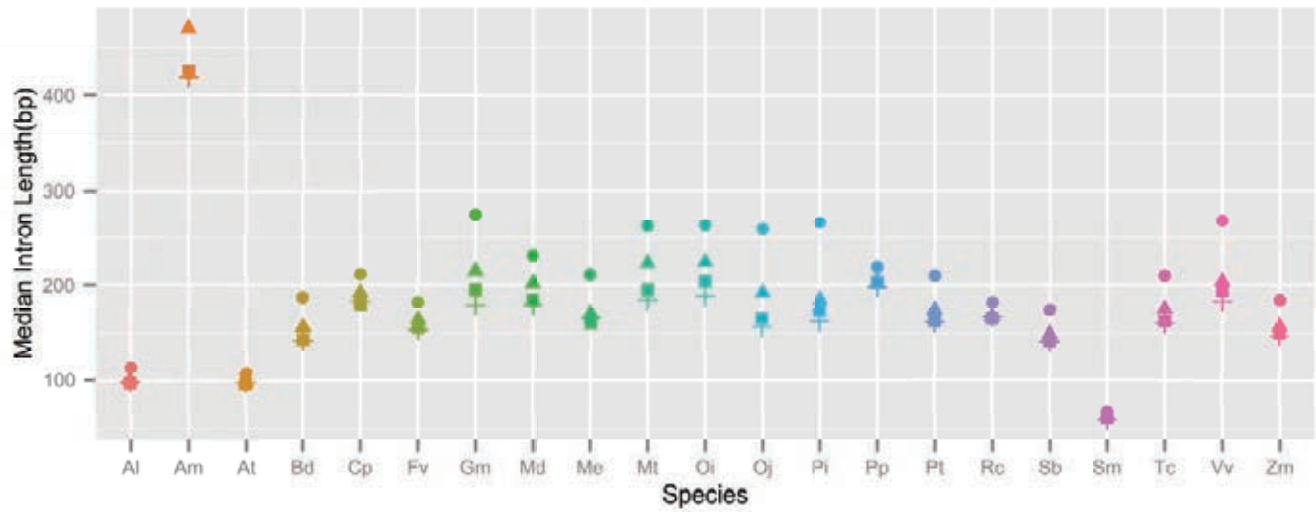
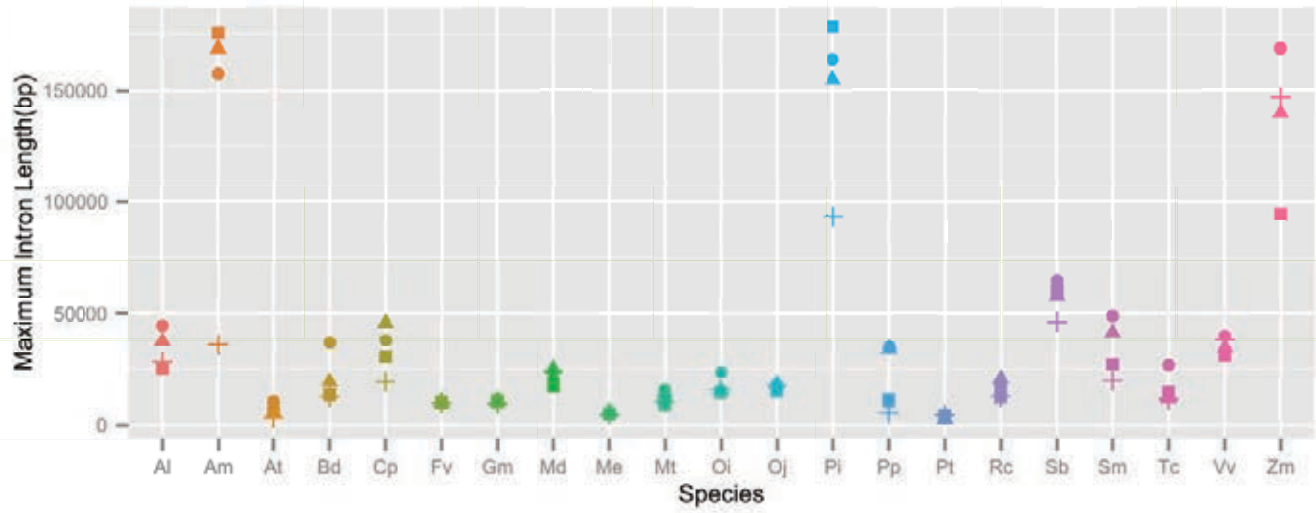


- 50,172 gene models mapped to 31,284 scaffolds at least 10 Kbp in length
- 3,835 scaffolds contained three or more genes

# MAKER-P

## Final Gene Models

Total Sequences	Total	Introns	Longest	Average	Number of	Number of	Number of	Number of	Number of	Number of
50,172										FIRST introns >100Kbp
15,653										63
										13



- Intron expression
- Word frequency
- the first intron
- 40%
- 38%
- 8%
- Transcription peptide

ces to

steine

# Comparing Plant Genomes

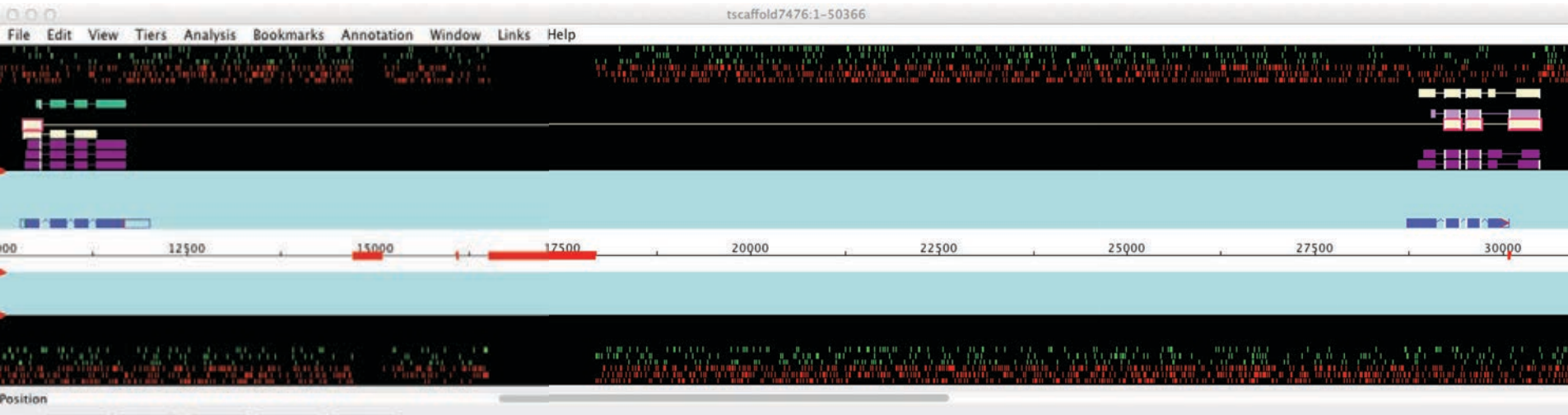
## Gene Space

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

- **How many genes are in the pine genome?**
  - Fragmentation, Pseudogenes, Repeats still present issues in constructing complete models
  - Bias for known protein domains and intron-less genes (up to 20%)
  - Previous estimates in literature range from 30K to over 90K

# Apollo

## Manual Annotation Needed

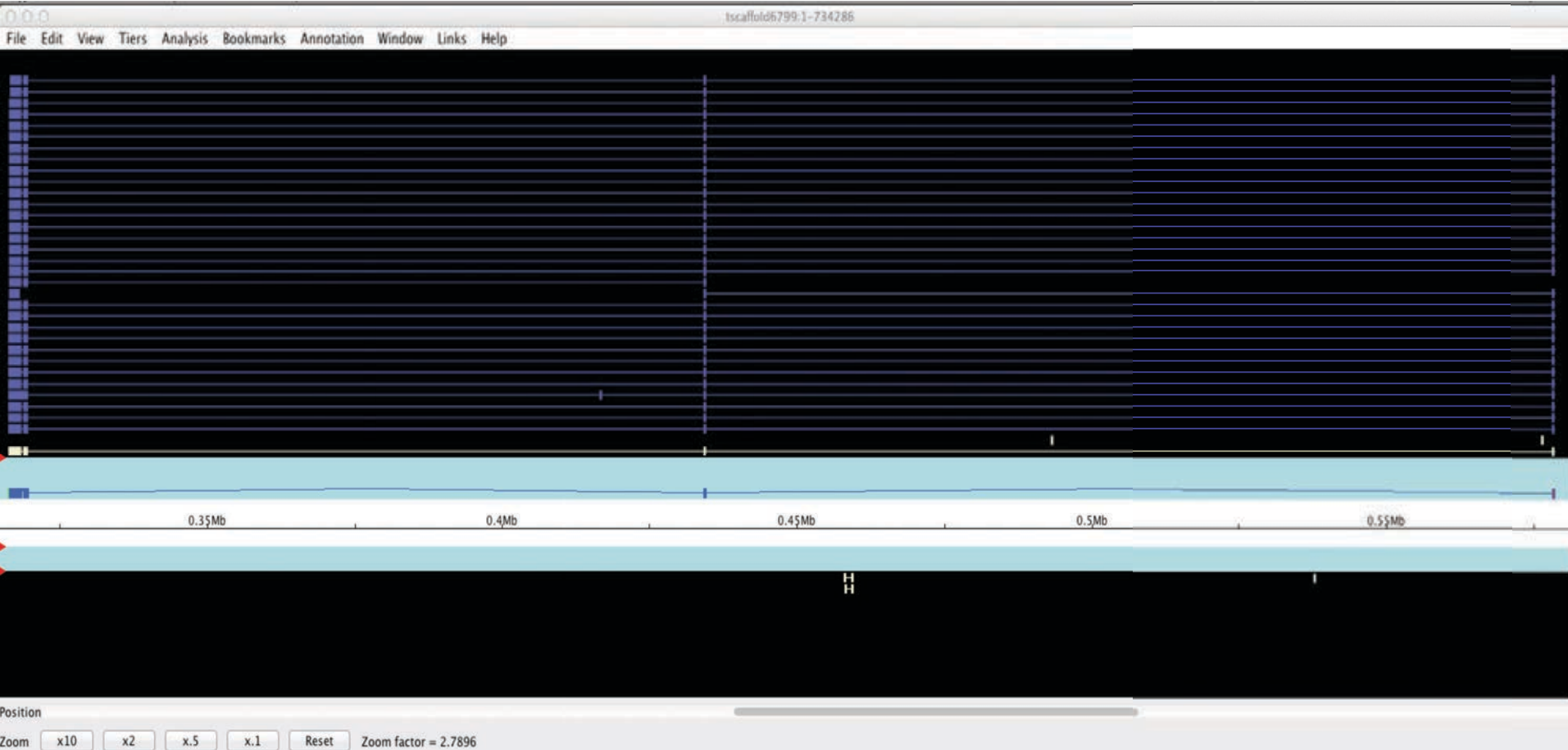


Yellow = Loblolly pine transcriptome  
Green = High quality Norway spruce protein  
Purple = Low Quality PLAZA proteins (70/70)  
Blue-Violet = MAKER annotation

# Long introns, over 100k

De novo transcript evidence and orthologous proteins

Pinus sylvestris phytochrome (NCBI BLASTP)



# Orthologous Proteins

## Examining Gene Families

- 50K gene models to the 352,151 proteins curated from 13 plant species resulted in **20,646 unique gene families**
- 11 PLAZA species + Amborella + Norway spruce + Sitka spruce
- Grouped into dicots, basal, mosses, monocots, and conifers

### Dicots

Arabidopsis thaliana: 26304 / 24766  
Glycine max: 36271 / 35969  
Populus trichocarpa: 35516 / 33358  
Ricinus communis: 30314 / 24039  
Theobroma cacao: 28222 / 27154  
Vitis vinifera: 24479 / 21795

### Basal

Amborella trichopoda: 24611 / 21191

### Mosses

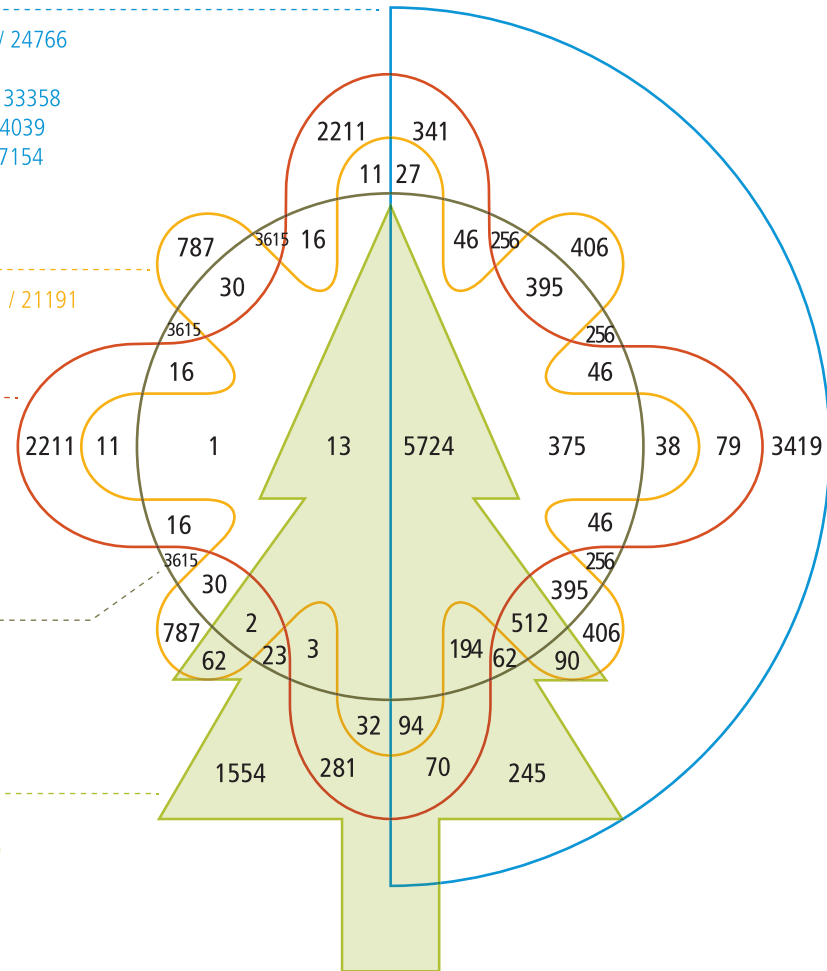
Selaginella moellendorffii:  
16832 / 15909  
Physcomitrella patens:  
25938 / 19359

### Monocots

Oryza sativa: 39459 / 32660  
Zea mays: 34586 / 30799

### Conifers

Picea abies: 20861 / 19934  
Picea sitchensis: 8758 / 7780  
Pinus taeda: 47207 / 46720





# Orthologous Proteins

## Examining Gene Families

- 20,646 unique gene families contained 90% gene set (361,433 proteins) with an average of 17 genes/family
- 1,554 conifer specific families
- Of these, 152 were unique to loblolly pine (32 with 5 more members)

### Dicots

Arabidopsis thaliana: 26304 / 24766  
Glycine max: 36271 / 35969  
Populus trichocarpa: 35516 / 33358  
Ricinus communis: 30314 / 24039  
Theobroma cacao: 28222 / 27154  
Vitis vinifera: 24479 / 21795

### Basal

Amborella trichopoda: 24611 / 21191

### Mosses

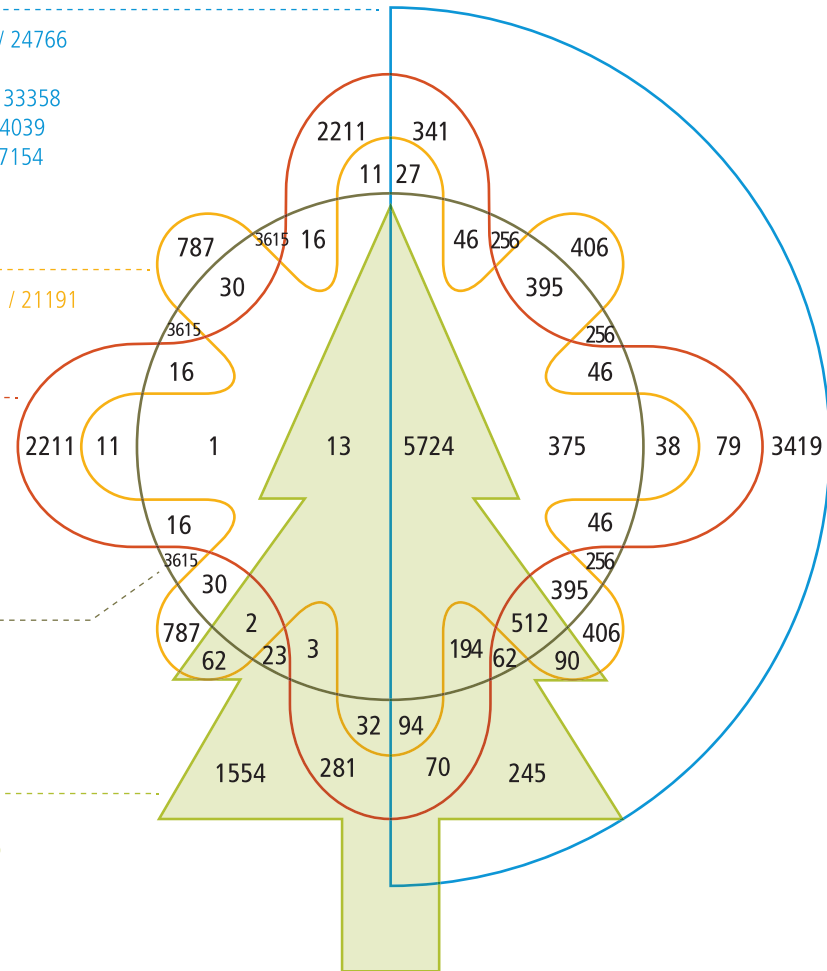
Selaginella moellendorffii:  
16832 / 15909  
Physcomitrella patens:  
25938 / 19359

### Monocots

Oryza sativa: 39459 / 32660  
Zea mays: 34586 / 30799

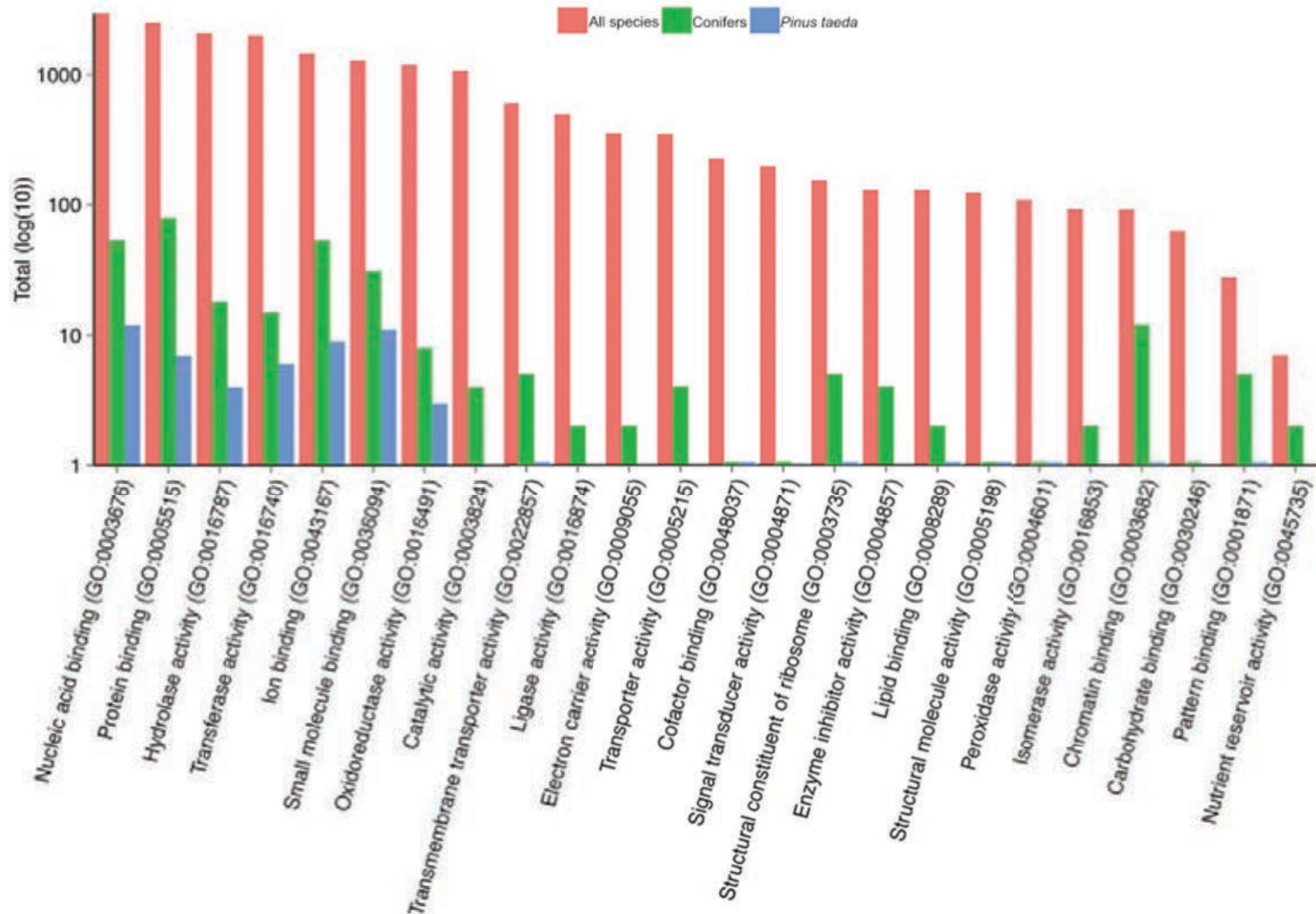
### Conifers

Picea abies: 20861 / 19934  
Picea sitchensis: 8758 / 7780  
Pinus taeda: 47207 / 46720



# Orthologous Proteins

## Examining Gene Families

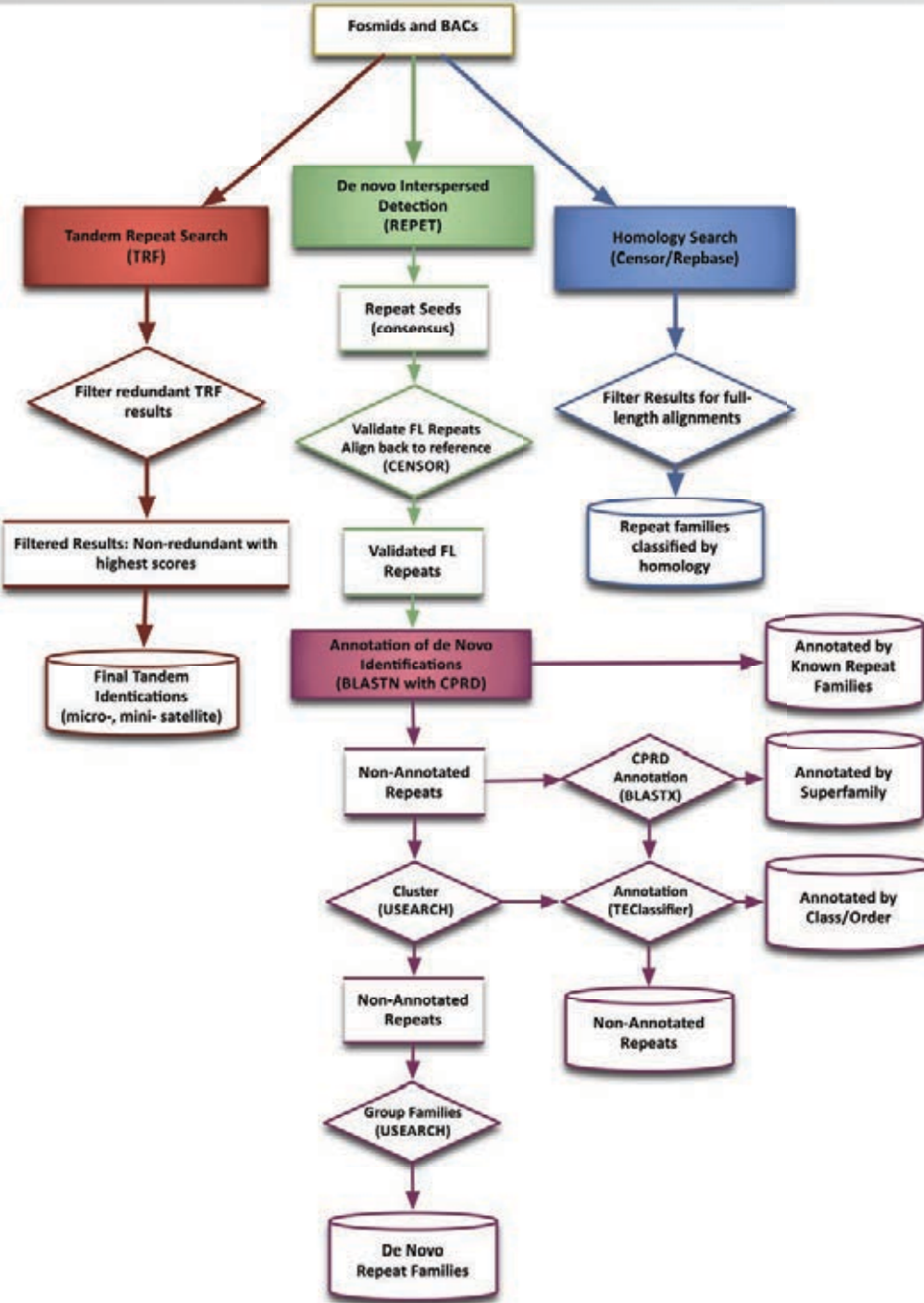


The *COPI C* family (58 members) was the largest exclusively identified in loblolly pine. Vesicle coat protein complexes containing *COPI* family members mediate transport between the ER and golgi, and interact with Ras-related transmembrane proteins, p23 and p24

# Prior Studies in Gymnosperms

Sequence Type	Species	Paper	Size of Resource	Repetitive Content Estimate	Elements Identified
BACs	<i>Picea glauca</i>	Hamberger et al (2009)	150 Kb	40%	
BACs	<i>Pinus taeda</i>	Morse et al (2009)	1,612 Kb		Gymny
BACs	<i>Pinus taeda</i>	Kovach et al (2010)	887 Kb	24-80%	PtIFG7
BACs	<i>Taxodium disitchum</i>	Liu et al (2011)	580 Kb	90%	
Fosmids	<i>Taxus mairei</i>	Hao et al (2010)	1,923 Kb	20.8%	
Southern Hybridization	<i>Pinus pinaster</i>	Rocheta et al (2006)			PpRT1
Southern Hybridization	<i>Pinus elliottii</i>	Kamm et al (1996)			TPE1

# Similarity and De Novo Repeat Identification



## Tandem Repeat Finder (TRF)

## Homology (Censor against RepBase)

Summary of Repbase v17.07

- Number of entries: 28,155
- Number of species represented: 715
- Number of repeat families: 280
  - Angiosperm entries: 131
  - **Gymnosperm entries (conifer):13**

## De Novo (REPET/TEannot)

- Self-alignment (all vs all) with BLAST to find HSPs is followed by clustering with **Grouper**, **Recon**, and **Piler**
- 3 sets of clusters are aligned with a MSA (**MAP**) to derive a consensus sequence
- Structural search runs simultaneously (**LTR Harvest**) to detect highly diverged LTRs
- Final **Blastclust** to cluster potential sequences

# Genomic Sequence

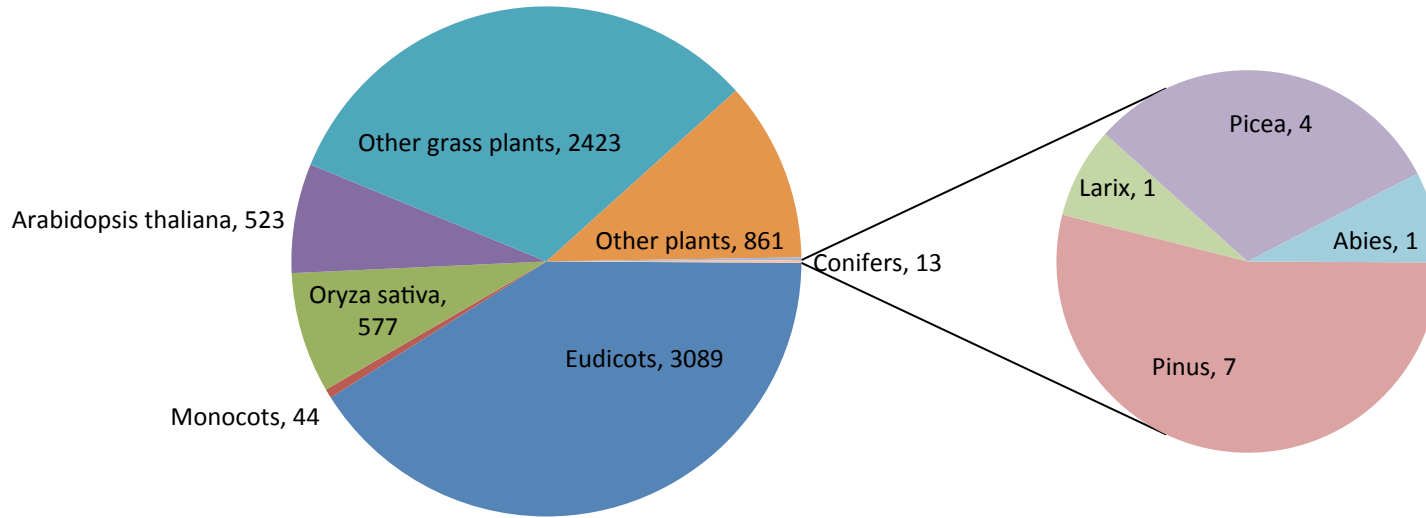
*Pinus taeda* BACs and Fosmids

	<i>Pinus taeda</i> BACs	<i>Pinus taeda</i> Fosmids
<b>Total number of sequences</b>	103	90,973
<b>Average sequence length</b>	115,130	2,918
<b>Median sequence length</b>	118,782	475
<b>N50 sequence length (bp)</b>	127,167	16,204
<b>Shortest sequence length</b>	1,392	201
<b>Longest sequence length</b>	235,088	75,791
<b>Total length (bp)</b>	11,858,447	265,511,345
<b>GC %</b>	37.98%	38.09%
<b>A : C : T : G%</b>	31.27 : 18.79 : 31.32 : 18.62	30.94:19.07:30.97:19.03

Combined sequence resource represents roughly 1% of the estimated genome

# Similarity Search Analysis

RepBase



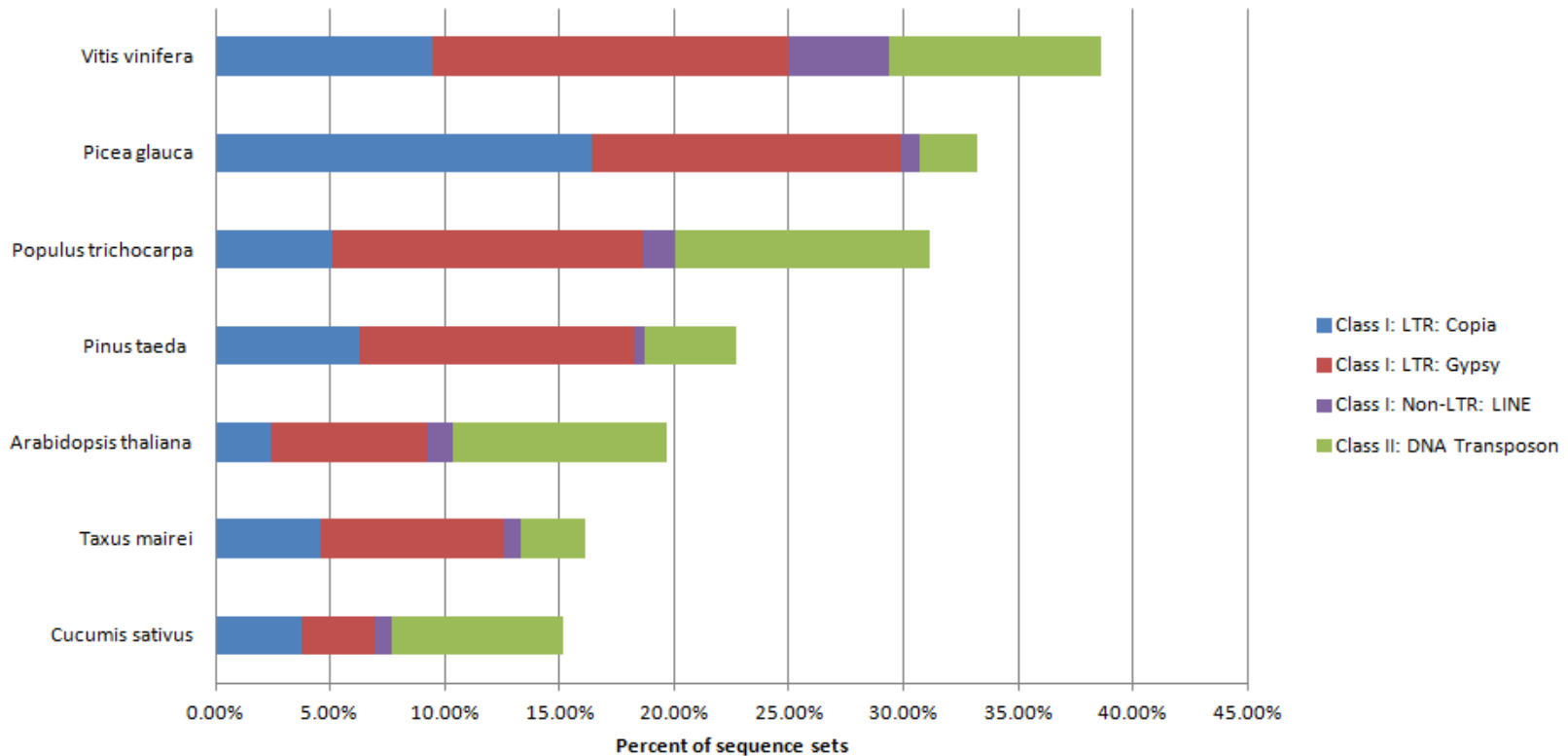
Pinus elliotti	Pinus thunbergii	Pinus radiata	Picea abies	Picea glauca	Abies veitchii	Larix
TY1_PE (Copia-LTR)	PLN1_PT (LINE) PLN2_PT (LINE) PLN3_PT (LINE) RT_PT (Copia-LTR)	IFG7_I (Gypsy-LTR) IFG7_LTR (Gypsy-LTR)	AlISEI_LTR(Gypsy-LTR) AlISEI_I(Gypsy-LTR)	PGGYPSYX1 (Gypsy-LTR) PCOPIAX1 (Copia-LTR)	ROMANIAV1 (Gypsy-LTR)	SAT1_LM (Satellite)



# Homology Search Results

Censor (BLAST-style) comparisons against Repbase

Partial and Full-length Interspersed Alignments (compared across species)



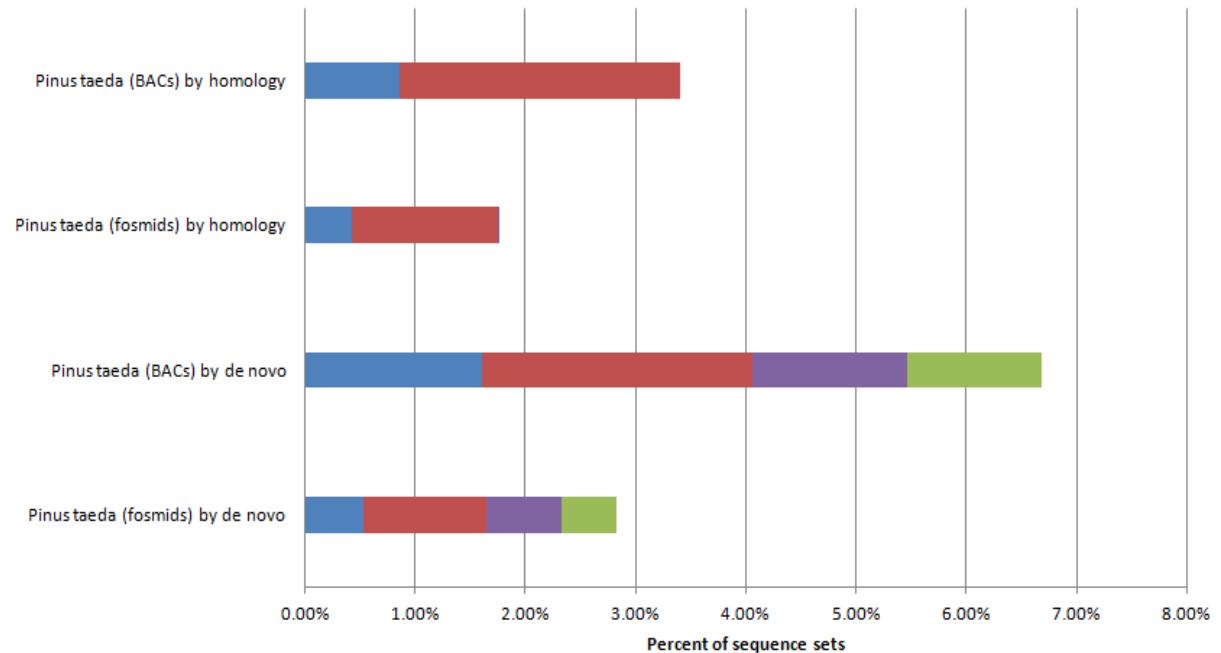
# Summary of Combined Homology and De Novo Approach

	P + F (Homology)	Filtered (Homology)	P + F (de novo)	Filtered (de novo)
<b>Class I</b>	20.41%	1.39%	73.39%	21.82%
<b>Class II</b>	4.03%	0%	1.52%	0.53%
<b>Other (Tandem)</b>	3.06%	2.6%	12.97%	6.22%
<b>Total</b>	<b>27.50%</b>	<b>3.99%</b>	<b>87.89%</b>	<b>28.58%</b>

## Full Length Sequences

80-80-80 Rule (Wicker et al. 2007)

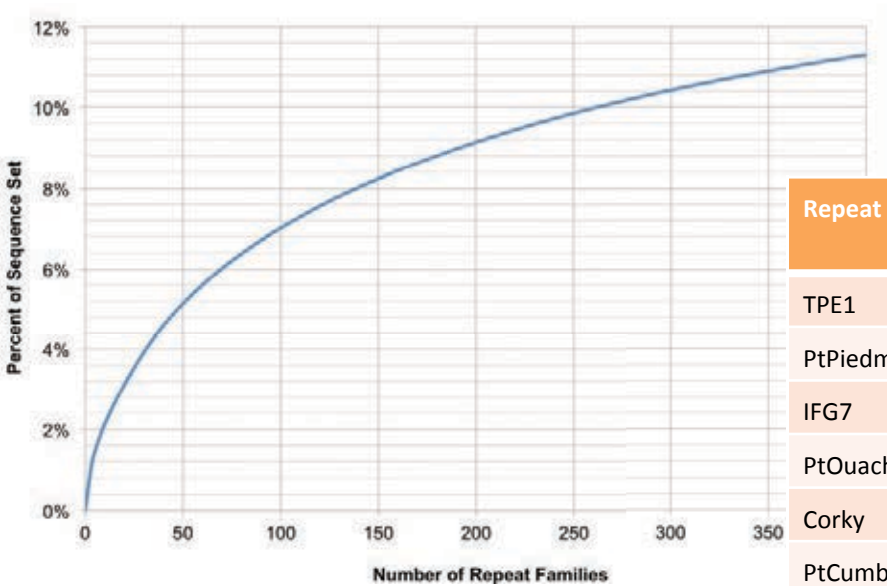
- 80 bp in length
- 80% identity
- 80% coverage



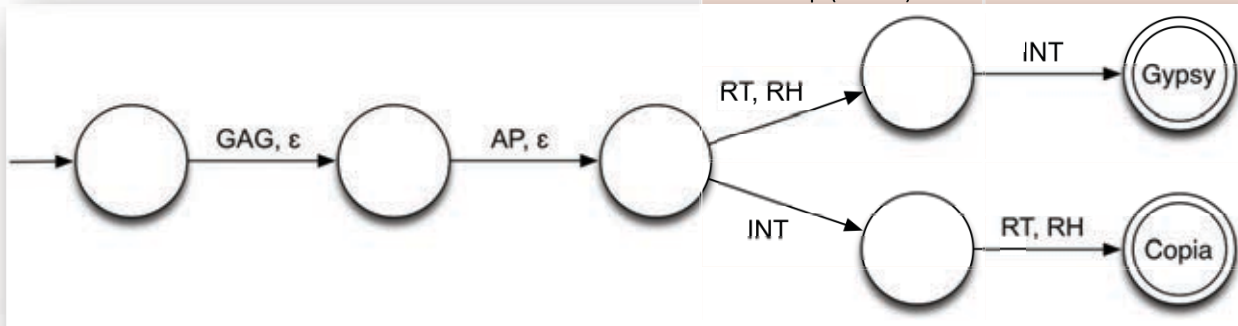
# Novel Repeat Elements

Diverged LTRs are annotated as 6,270 novel families

Top 400 elements only cover 12% of the combined sequence sets

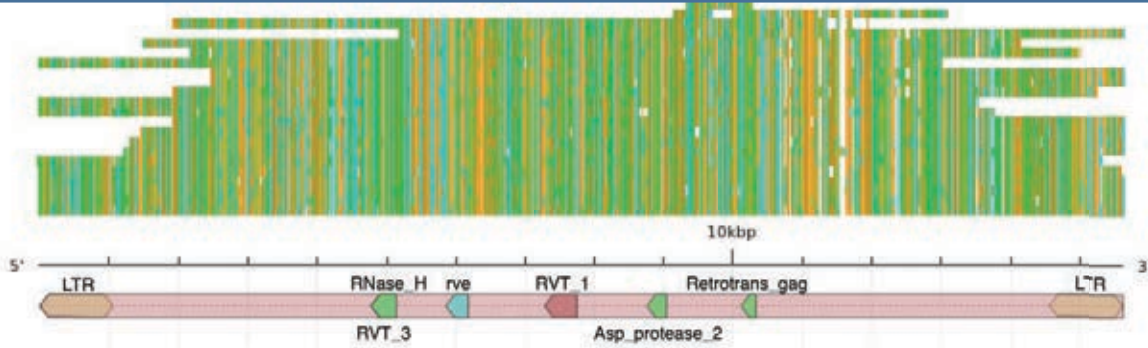


Repeat family	Full-Length Copies	Length (bp)	Percent of Sequence Set
TPE1	159	1,077,598	0.39%
PtPiedmont (93122)	133	969,109	0.35%
IFG7	162	956,018	0.34%
PtOuachita (B4244)	47	576,871	0.21%
Corky	78	469,286	0.17%
PtCumberland (B4704)	67	431,492	0.16%
PtBastrop (82005)	38	378,631	0.14%
		378,020	0.14%
		367,653	0.13%
		322,632	0.12%
		309,248	0.11%
		291,479	0.11%
PtConagree (B3341)	50	285,850	0.10%
PtTalladega (215311)	33	274,826	0.10%
Total	982	7,088,713	2.56%

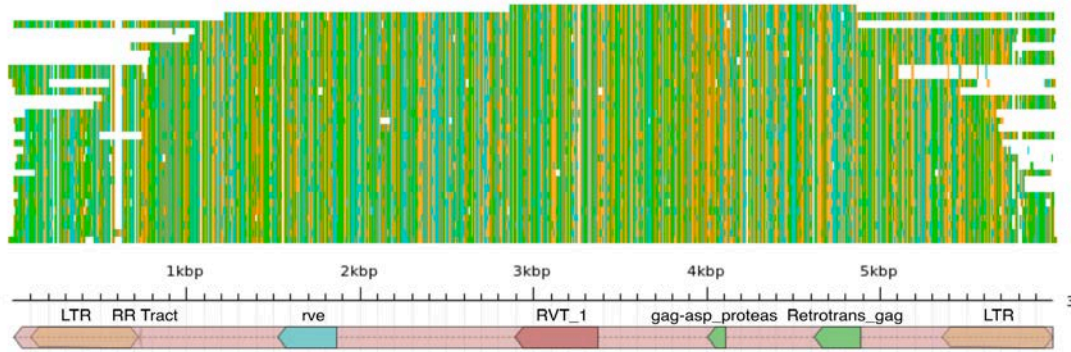


# Novel Repeat Elements

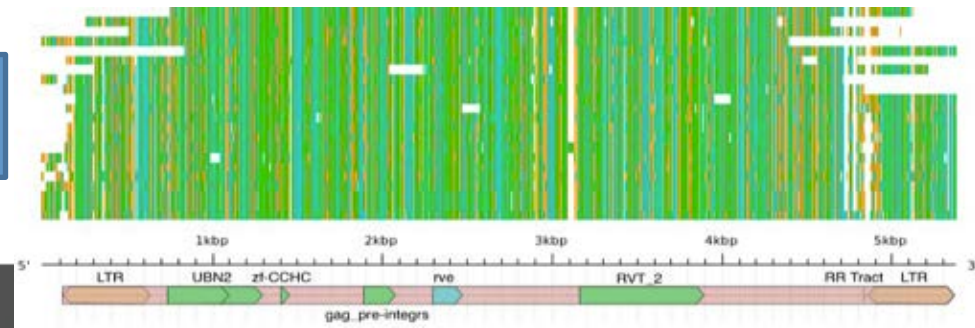
Multiple sequence alignment with annotations of the novel Gypsy LTR - PtOuachita



Multiple sequence alignment with annotations of the novel Gypsy LTR - PtAppalachian



Multiple sequence alignment with annotations of the novel Copia LTR - PtPineywoods



# Summary of Combined Homology and De Novo Approach

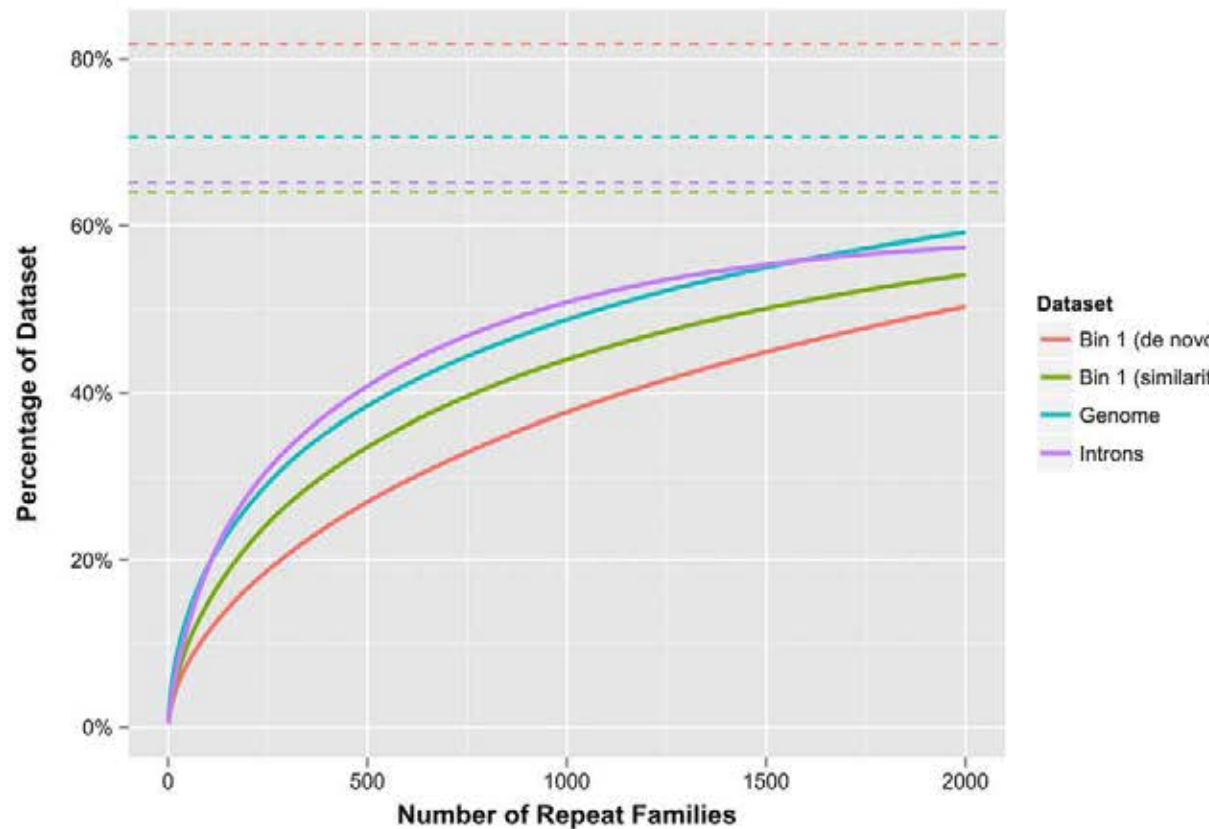
- 88% repetitive (partial and full-length)
- 29% repetitive (full-length only defined by 80-80-80)
  - 87% of the full-length content is characterized as LTR retrotransposon (Gypsy and Copia)
- Repeats are highly diverged
  - Only 23% identified by homology for full and partial elements
  - Repbase contains just 15 (+5) gymnosperm elements
  - 6,270 novel families discovered with no homology
    - 5,155 are single copy
- Nested repeats common in LTR retrotransposons

# PIER library

## Pine Interspersed Element Repeat Library (PIER 2.0)

5,280 elements belonging to six families (plnrep, grasrep, mcotrep, dcotrep, oryrep, and athrep) from Repbase (17.07), 5 additional elements previously characterized, and 9,415 elements characterized in loblolly pine BAC/fosmids. Total = 14,700

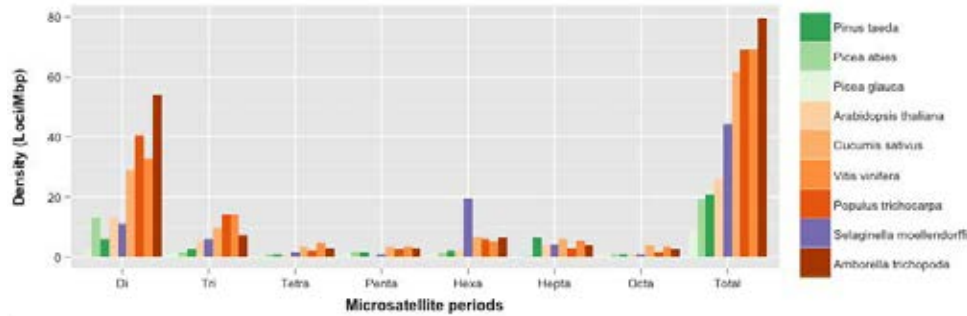
- Similarity analysis alone revealed just over 60% of the genome to be repetitive
  - Much less full length content identified than in BACs/fosmids
- De novo analysis applied to bin1 estimates 82% repetitive
- Repeat element families are very diverged
- High copy and highest coverage elements different than those identified in BACs/fosmids originally
  - First 100 elements account for less than 10% of the genome
  - TPE1 and PtConagree



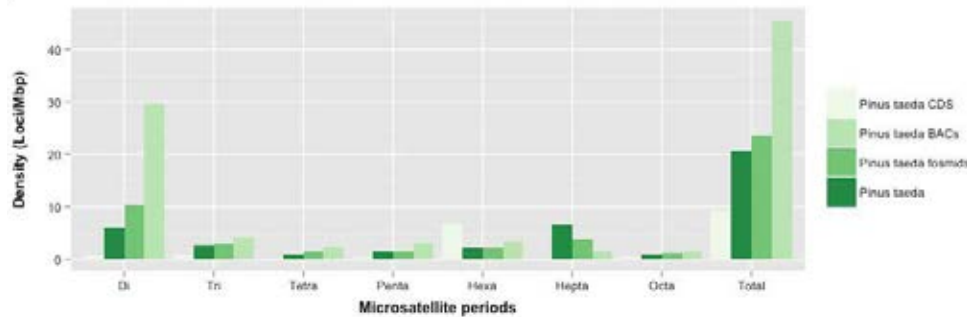


# Tandem Repeats

Comparison across sequenced angiosperms and other gymnosperms



Total tandem content for loblolly pine:  
 3.31% of BACs  
 2.59% of fosmids  
 2.86% of genome



	<i>Pinus taeda</i> v1.01			<i>Picea glauca</i> v1.0			<i>Picea abies</i> v1.0		
	Micro	Mini	Sat	Micro	Mini	Sat	Micro	Mini	Sat
Most frequent period size	7	21	123	2	21	102	2	50	101
Cumulative length (Mbp)	8.93	20.09	8.66	2.26	8.70	4.89	10.38	29.00	5.25
Num. of loci	145,992	361,356	27,422	62,592	179,716	23,256	255,380	285,648	24,895
Total cumulative length (Mbp)	27.33	396.42	221.01	7.48	357.90	198.83	18.63	299.25	151.81
Total %	0.12%	1.76%	0.98%	0.04%	1.72%	0.96%	0.10%	1.53%	0.77%
Total overall content	<b>2.86%</b>			<b>2.71%</b>			<b>2.40%</b>		

# Genome Database (TreeGenes)

- Genome/Transcriptome Delivery
- Community-level annotation
- TreeGenes resources

CAACAAGTCATCCATGATT  
TCCGCATAGTAGCTCATA  
TCATAGTCTTCAATGCA  
CAACAAGTCATCCATGATC  
TCATAGTAGCTCATA

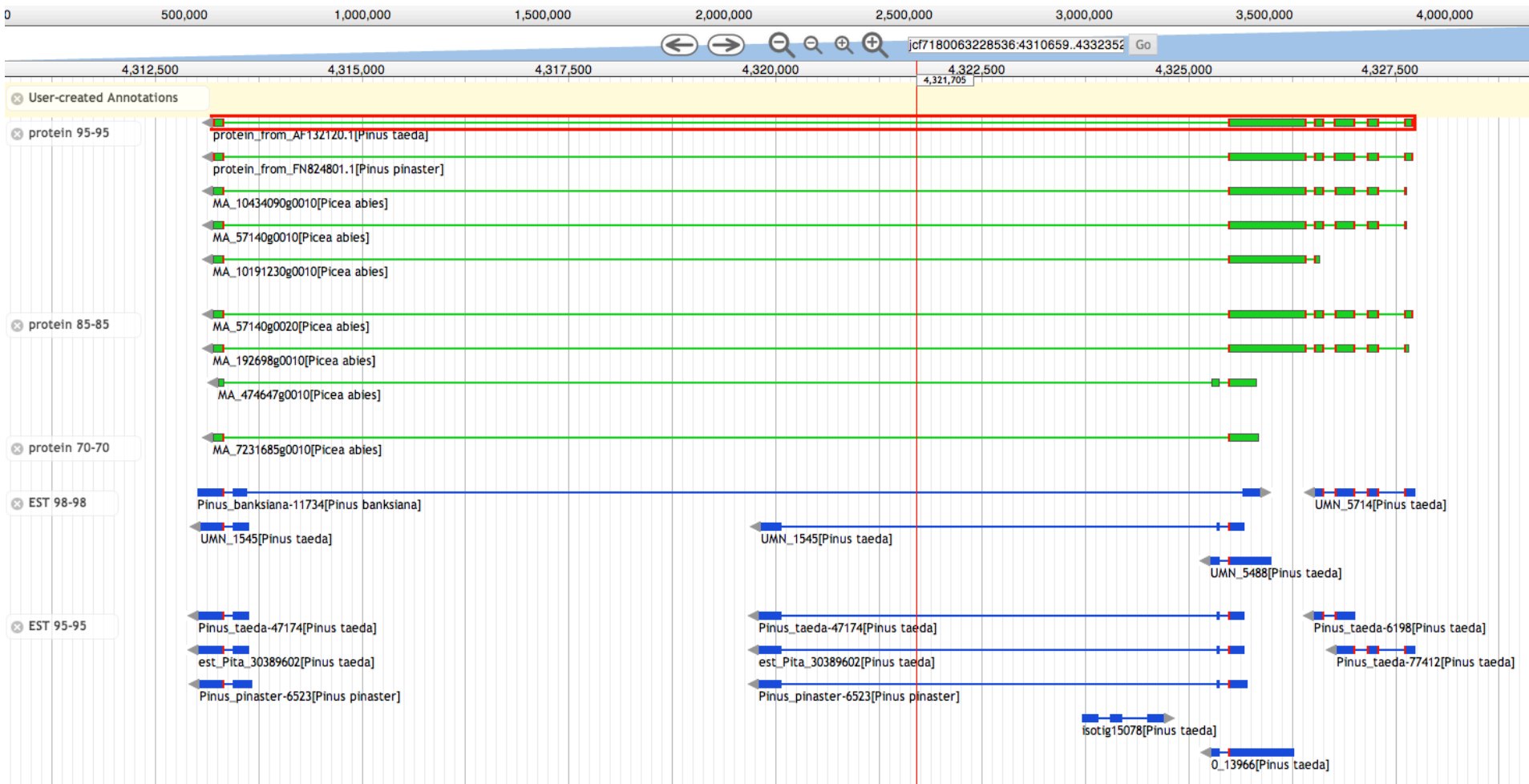
# Dendrome Project

## TreeGenes Database to Distribute Transcriptome and Genome

The screenshot displays the TreeGenes web interface. At the top, the 'TreeGenes' logo and 'A Forest Tree Genome Database' are visible. The main content area shows a genomic track for *P. taeda* (Loblolly pine) Genome v1.01, specifically a 4.05 kbp region from coordinates C31761558:476..4,525. The interface includes a search bar with the same coordinates, a 'Data Source' dropdown set to 'P.taeda (Loblolly pine) Genome v1.01', and navigation controls like 'Annotate Restriction Sites', 'Save Snapshot', and 'Load Snapshot'. The genomic track itself shows a scale from 0k to 10k, with a zoomed-in view of the 4.05 kbp region. Below the scale, various genomic features are annotated: a gene (maker-jcf7180059383434-snap-gene-0..2), its coding sequence (CDS), mRNA (maker-jcf7180059383434-snap-gene-0..2-mRNA-1), and both Five Prime UTR and Three Prime UTR regions. The interface also includes a 'Comparative Mapping Database' section at the bottom, listing other species like *Populus tremuloides*, *Populus trichocarpa*, and *Vitis vinifera* with links to 'Downloads' and 'More Info'.

# WebApollo on TreeGenes

## Conifer specific proteins



# GenSAS

## GENome Sequence Annotation Server

**GENome Sequence Annotation Server**

### Sequence Selection

Provide a group name for these sequences

Please paste sequence above

### Masking Tool Selection

Select a masking tool to identify repeats in your sequences. Tools added to the task in the Tool Selection box will be executed after all masking tools have finished, and they will use the masked sequence rather than the original non-masked sequence.

**Masking Tools**

**RepeatMasker**

### Tool Selection

*Intrinsic Gene Prediction*

Genscan

FGENESH

Augustus

*Extrinsic Gene Prediction*

Transcript BLAST

Protein BLAST

BLAT

*Other Features*

Microsatellite

GFF3 Importer

tRNAscan

getorf

### Task Queue

V. corymbosum 454 asse

Microsatellite\_all\_scaffok

MS\_all\_scaffolds

Loblolly Pine Genome v1

anthocyanin 5-aromatic a



# GenSAS

## GENome Sequence Annotation Server

# GenSAS

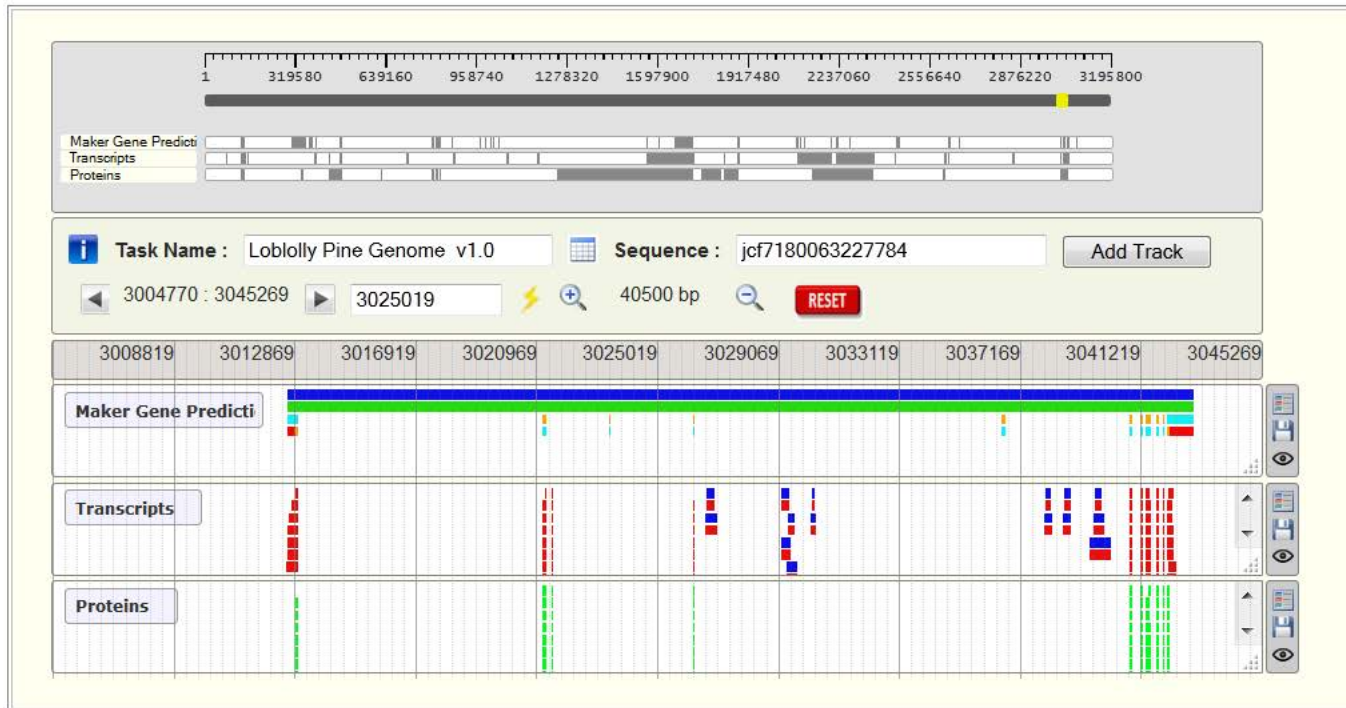
Genome Annotation and Curation

Enter the username to masquerade as.

Logged in as: ficklin | [Log Out](#)

[Home](#) [Use GenSAS](#) [Manage GenSAS](#) [GenSAS User Guide](#) [Contact](#) [Create content](#) [Administer](#) [Log out](#)





# GenSAS

## GENome Sequence Annotation Server

The screenshot displays the GenSAS web interface. At the top left, the logo "GenSAS" is shown with the subtitle "Genome Annotation and Curation". A navigation bar includes "Home", "Use GenSAS", and "Manage GenSAS". On the right, there is a search bar with a "Go" button and a login status "Logged in as: ficklin | Log Out".

The main content area features a genomic track for "Loblolly Pine" with coordinates from 3004770 to 3045269. The track includes "Maker Gene Predictions", "Transcripts", and "Proteins". A detailed view of a specific mRNA prediction is shown in a modal window:

Track	Maker Gene Predictions
Type	mRNA
Start	3012569
Stop	3042927
Strand	+
Source	maker
Score	-
Phase	-
ID	maker-jcf7180063227784-augustus-gene-14.22-mRNA-1
Parent	maker-jcf7180063227784-augustus-gene-14.22
Name	maker-jcf7180063227784-augustus-gene-14.22-mRNA-1
_AED	0.18
_eAED	0.18
_QI	267 0.63 0.75 0.83 0.45 0.5 12 772 362
Dbxref	<ul style="list-style-type: none"><li>Gene3D:G3DSA:3.60.21.10</li><li>InterPro:IPR004843</li><li>InterPro:IPR006186</li><li>PANTHER:PTHR11668</li></ul>
Ontology_term	GO:0016787

At the bottom of the modal window, there are options: "Add to Curation Track" and "No editable custom track".

# GenSAS

## GENome Sequence Annotation Server

The screenshot displays the GenSAS web interface. At the top left, the logo 'GenSAS' is shown with the tagline 'Genome Annotation and Curation'. On the top right, there is a search bar with a 'Go' button and a login status 'Logged in as: ficklin | Log Out'. A blue navigation bar contains 'Home', 'Use GenSAS', and 'Manage'.

The main content area is partially obscured by a 'Task Information' modal window. The modal window has a title bar with a close button and contains the following information:

**Task Information**

**GFF3 Importer**

Job Name: Maker Gene Predictions  
Status: completed

Feature Count

Type	Total
CDS	191,156
exon	197,597
five_prime_UTR	23,049
gene	50,172
mRNA	50,172
three_prime_UTR	16,450
<b>Total</b>	<b>528,596</b>

Tool Parameters: pita.genes.filtered.gff

Output Files

- Run Log
- Error Log
- Uploaded File

Please note, some files may be large and others may be empty dependent on how the tool works.

In the background, a 'Maker Gene Predictor' interface is visible, showing a genomic track with coordinates (1, 319) and a 'Task Name' field containing 'Loblo'. Below the track are buttons for 'Maker Gene Predictic', 'Transcripts', and 'Proteins'. To the right, another track is partially visible with coordinates (720, 800) and a 'Add Track' button.

## Acknowledgements

**Project Director:** David Neale (UCD)

**Annotation Team:**

John Liechty (UCD)

Kristian Stevens (UCD)

Hans Vasquez-Gross (UCD)

Pedro J. Martinez-Garcia (UCD)

Brian Lin (UCD)

Jacob Zieve (UCD)

William Dougherty (UCD)

**MAKER Team:**

Carson Holt (Utah)

Mark Yandell (Utah)

**Transcriptome Team:**

Le-Shin Wu (IU)

Keithanne Mockaitis (UI)

Carol Loopstra (TAMU)

**GenSAS Team:**

Stephen Ficklin (WSU)

Doreen Main (WSU)

**Sequence and Assembly Team:** Aleksey Zimin (UMD), Daniela Puiu (JHU), Steven Salzberg (JHU), Jim Yorke (UMD), Marc Crepeau (UCD), Daniela Puiu (JHU), Steven Salzberg (JHU), Maxim Koriabine (CHORI), Ann Holtz-Morris (CHORI), Pieter de Jong (CHORI)



# Thank you!



United States  
Department of  
Agriculture

National Institute  
of Food and  
Agriculture