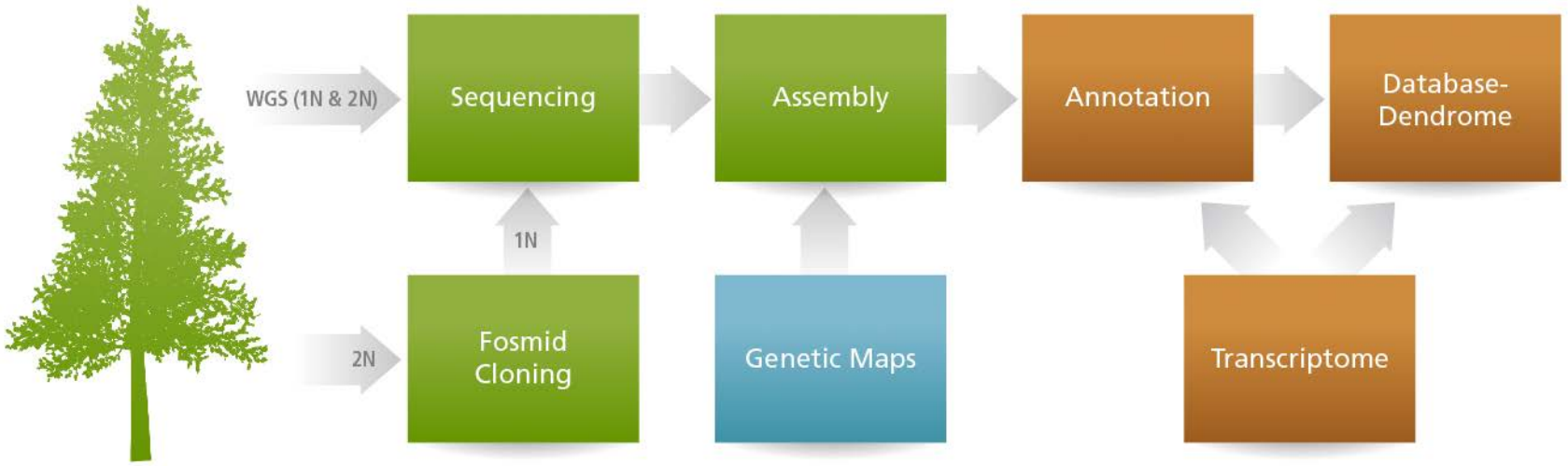
Annotating the sugar pine genome: transcriptome survey


Jill Wegrzyn

Department of Ecology and Evolutionary Biology
Institute of Systems Genomics
University of Connecticut

PineRefSeq

Elements of a Conifer Genome Sequencing Project



Loblolly pine transcriptome sequencing

Generate a comprehensive transcriptome reference:

- **Assemble coding regions for scaffolding 1.0 to 1.01 loblolly pine genome**
- **Integrate community data into assemblies (EST resources)**
- **Develop resources for the training of gene prediction tools (MAKER-P)**

Early Development
seeds
young seedlings

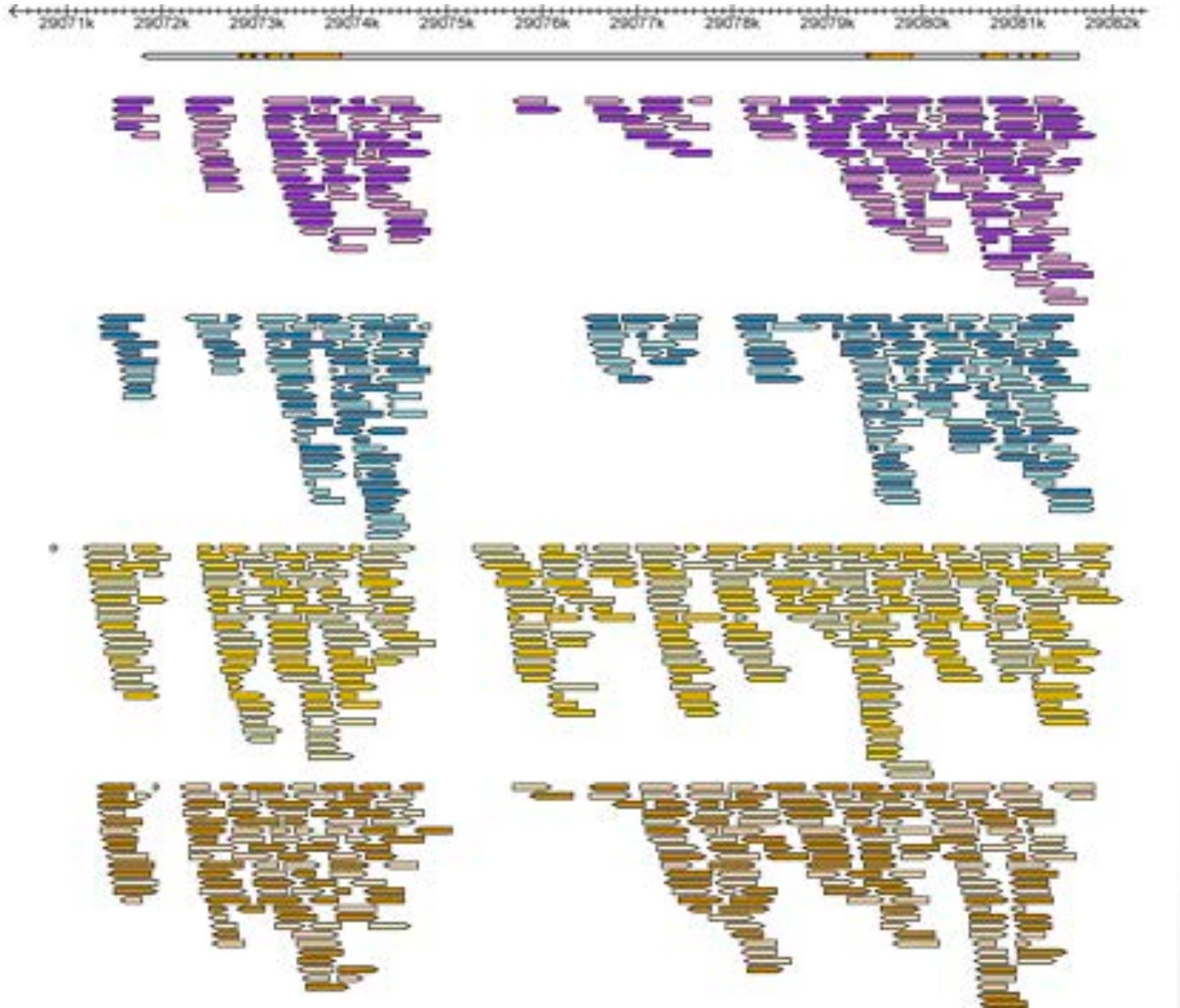
Reproductive Development
megastrobili
microstrobili

Vegetative Organs
vegetative buds
candles
stems
needles
roots

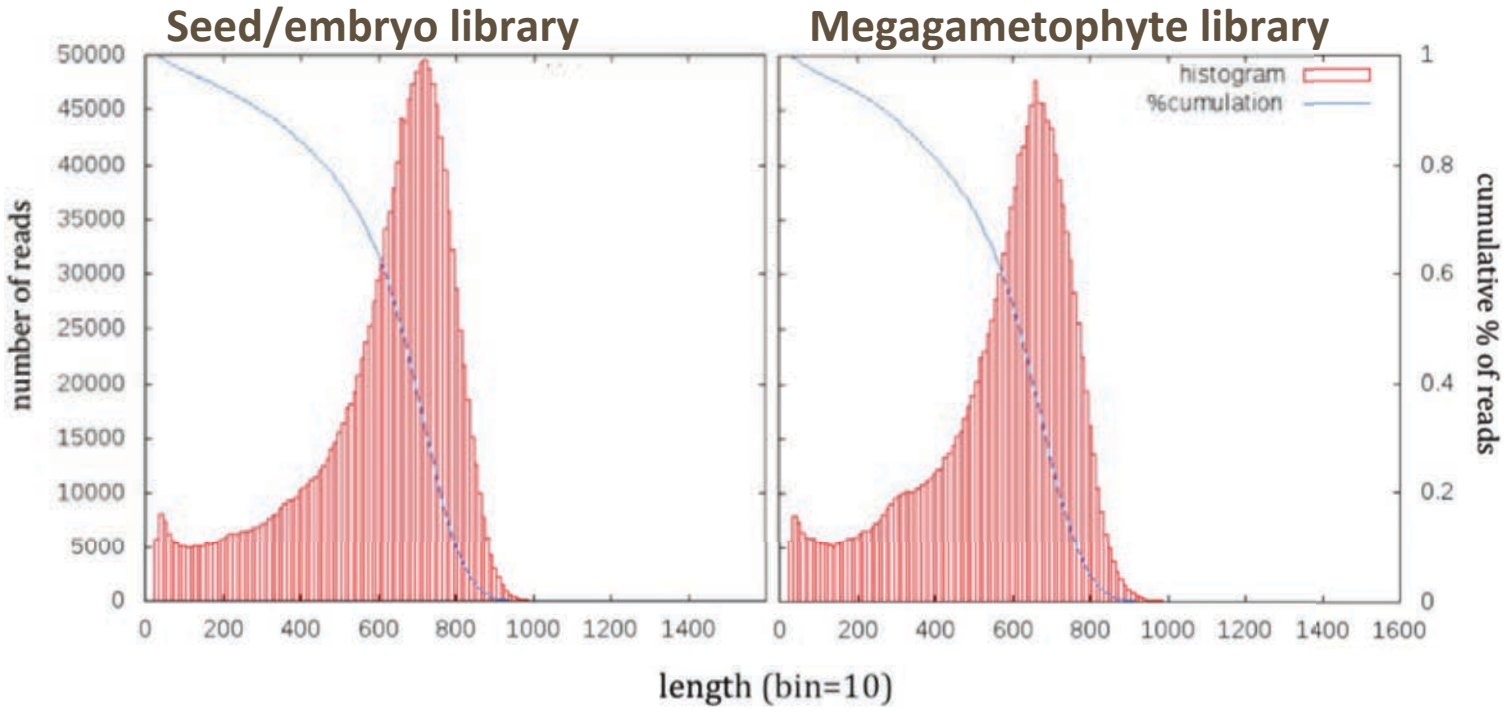
Early Stress Signaling Responses
cold
heat
elevated UV
compression

Carol Loopstra (Texas A&M University) and **Keithanne Mockaitis** (Indiana University)

Mapping Transcriptomes to the Genome



Loblolly pine transcriptome



Early Development (tissues from 20-1010 individual)

Seeds → Embryo

Seedlings → Young tissues/incremental stages

Conifer Reference Assemblies

***Pinus taeda* (loblolly pine):**

(3rd) Assembly v1.01 (Sept 2013)

- **Assembly + Trans Scaffolding**
- Approximately **65X** coverage
- Total Sequence: **22.1 Gbp**
- **N50 Scaffold:**
  - **66.9Kbp** (14.4 m)

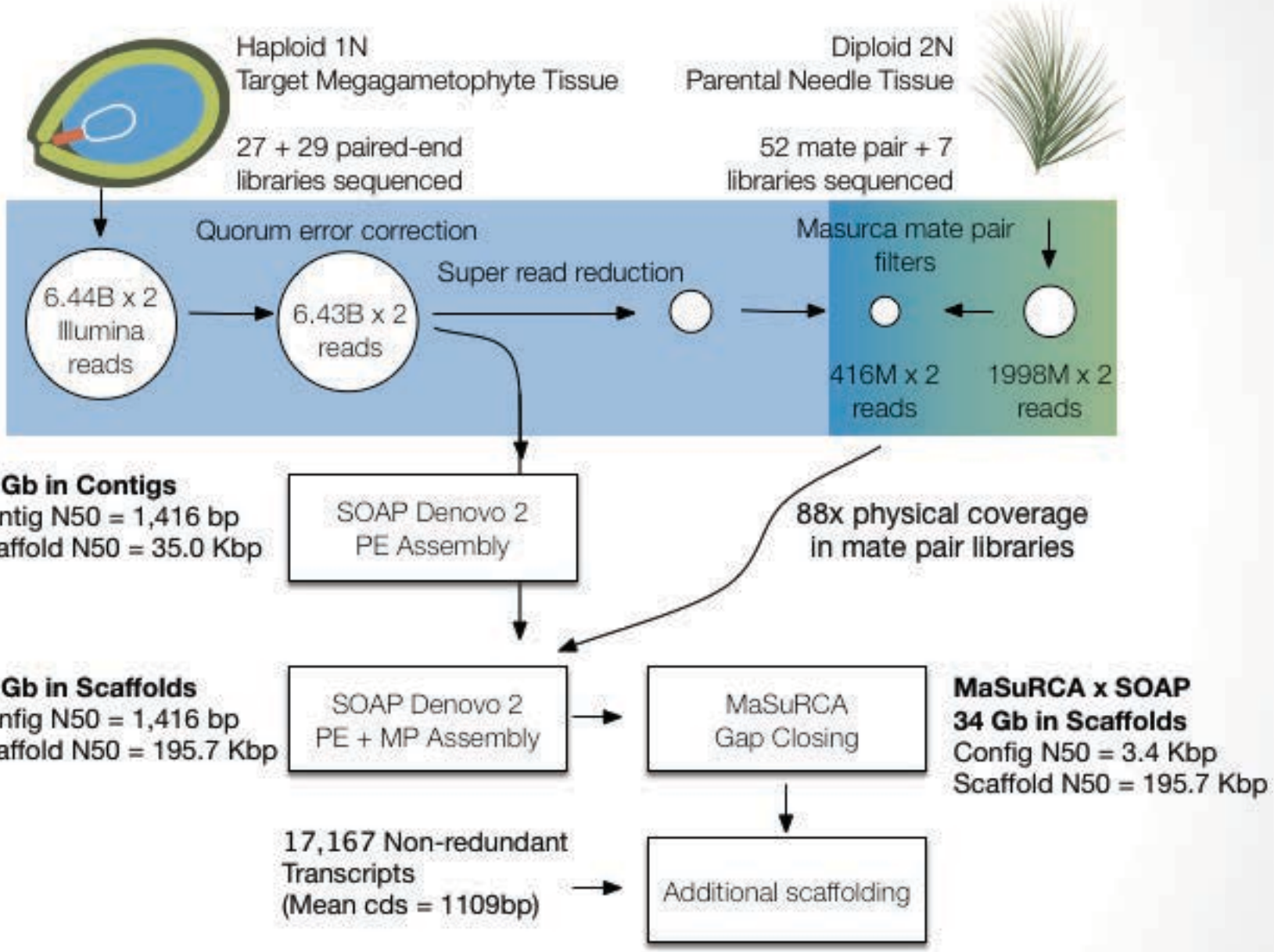
***Pinus lambertiana* (sugar pine):**

(1st) Assembly v0.5 (Aug 2014)

- **First pass assembly only**
- Approximately **62x** coverage
- Total Sequence: **33 Gbp**
- **N50 Scaffold:**
  - **34.9 Kbp**

Pinus taeda version 1.0 to 1.01 reflects (16 to 14 mil)

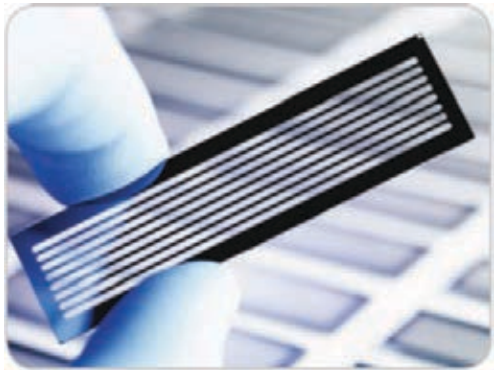
- SOAPdenovo scaffolding
- Independent scaffolding with transcriptome (nucmer)
  - Over **75,000 unique full-length genes**


Haploid 1N
Target Megagametophyte Tissue
27 + 29 paired-end
libraries sequenced
Diploid 2N
Parental Needle Tissue
52 mate pair + 7
libraries sequenced
Quorum error correction
6.44B x 2
Illumina
reads
6.43B x 2
reads
Super read reduction
Masurca mate pair
filters
416M x 2
reads
1998M x 2
reads
33 Gb in Contigs
Contig N50 = 1,416 bp
Scaffold N50 = 35.0 Kbp
SOAP Denovo 2
PE Assembly
88x physical coverage
in mate pair libraries
34 Gb in Scaffolds
Config N50 = 1,416 bp
Scaffold N50 = 195.7 Kbp
SOAP Denovo 2
PE + MP Assembly
MaSuRCA
Gap Closing
MaSuRCA x SOAP
34 Gb in Scaffolds
Config N50 = 3.4 Kbp
Scaffold N50 = 195.7 Kbp
17,167 Non-redundant
Transcripts
(Mean cds = 1109bp)
Additional scaffolding

Generating resources for sugar pine

Summary of Genomic Resources

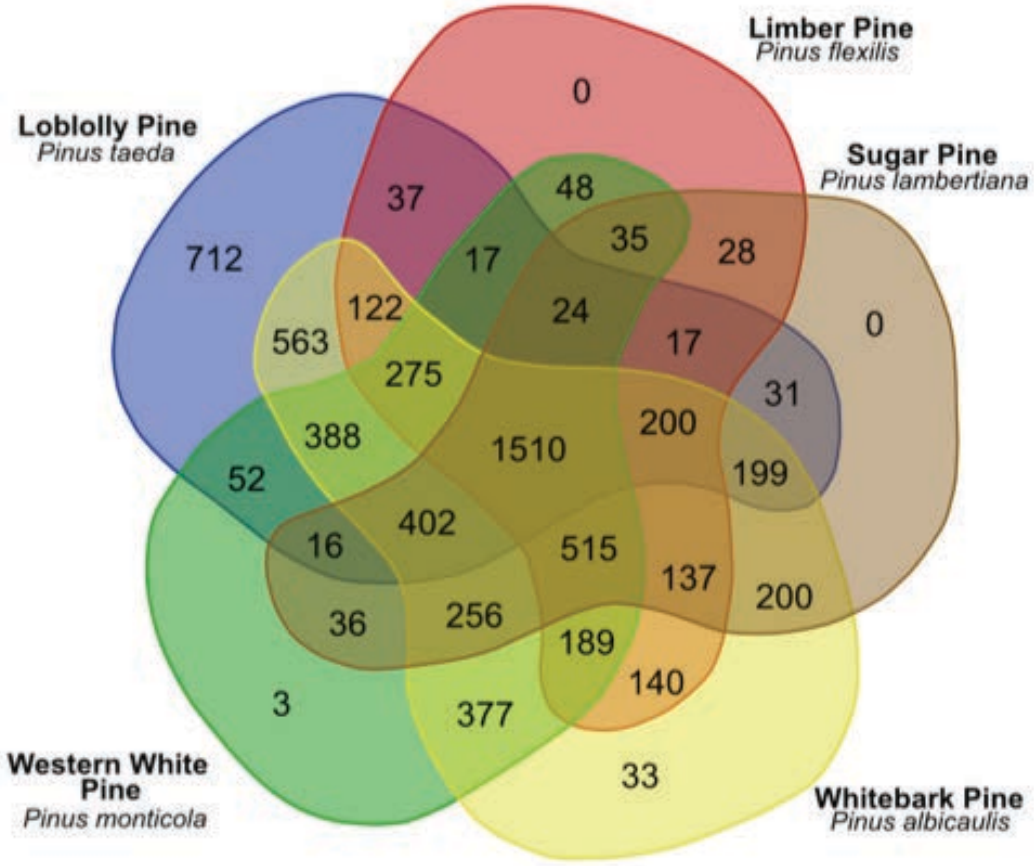
Species	Technology	Reads	Tissue	Reads after QC
Sugar pine Jessica Wright	Illumina GA IIx	SE, 80bp (3 lanes)	needle	66,894,169
Sugar pine (Lorenz et al. 2012)	Roche 454	SE, 350 bp (avg)	stem, needle	952,310
Limber pine Jeff Mitton	Illumina HiSeq	PE, 100bp (2 lanes)	needle	374,191,816
Whitebark pine Patricia Maloney	Illumina HiSeq	PE, 100bp (3 lanes)	needle	839,389,034
Western white pine (J-J. Liu et al 2013)	Illumina GA IIx	PE, 76bp	needle	208,059,003



White pine transcriptomes assembled

Species	Annotation rate	Informative Hits	Number of full-length genes	Avg contig length/N50	Annotation rate of full-length	Contaminants (%)
Sugar pine	61.78%	58.52%	**10,798**	1,319/1,506	93.84%	.36%
Limber pine	74.00%	67.71%	**15,090**	1,303/1,491	92.31%	0.23%
Whitebark pine	38.60%	34.73%	**25,780**	1,572/1,806	93.10%	0.40%
Western white pine	62.00%	57.27%	**24,082**	1,455/1,638	93.21%	0.29%

Needle Transcriptomes Compared



- TRIBE-MCL analysis
- Examination of Orthologous Groups (Proteomes)
- Identifying Taxonomically Restricted Sequences

Tissues and Technologies

Tissue	Description	Technology
pollen	pollen	MiSeq
female conelets	early female conelets before pollination	MiSeq
pollen cones	pollen cones	MiSeq
cone	9-inch cone	MiSeq
seed	germinating sugar pine seed	MiSeq/HiSeq/PacBio
root, stem and needles	"basket stage" seedling	MiSeq
root, stem and needles	"primary needle stage" seedling	MiSeq
root	seedling slowly drought-stressed	MiSeq
needle	seedling slowly drought-stressed	MiSeq
stem	seedling slowly drought-stressed	MiSeq
root	Flood roots after treatment	MiSeq
needle	Flood needles after treatment	MiSeq
stem	Blister Susceptible stem (LCO2-15)	MiSeq
stem	Blister Resistant stem (LCO2-03)	MiSeq
needle	seedling cold-shocked	HiSeq
needle	seedling heat-shocked	HiSeq
stem	stem after wounding	HiSeq
root	root after NaCl treatment	HiSeq
root	Salicylic Acid treatment for 5 hrs	HiSeq
needle	Salicylic Acid treatment for 5 hrs	HiSeq
stem	Salicylic Acid treatment for 5 hrs	HiSeq
root	Methyl jasmonate treatment for 5 hrs	HiSeq
needle	Methyl jasmonate treatment for 5 hrs	HiSeq
stem	Methyl jasmonate treatment for 5 hrs	HiSeq
female cones	2 cm female cones	HiSeq/PacBio
female strobili	female strobili near pollination	HiSeq/PacBio
root	seedling was not drought-stressed	HiSeq/PacBio
needle	seedling was not drought-stressed	HiSeq/PacBio
stem	seedling was not drought-stressed	HiSeq/PacBio
needle	Blister Susceptible needles (LCO2-15)	HiSeq/PacBio
needle	Blister Resistant needles (LCO2-03)	HiSeq/PacBio

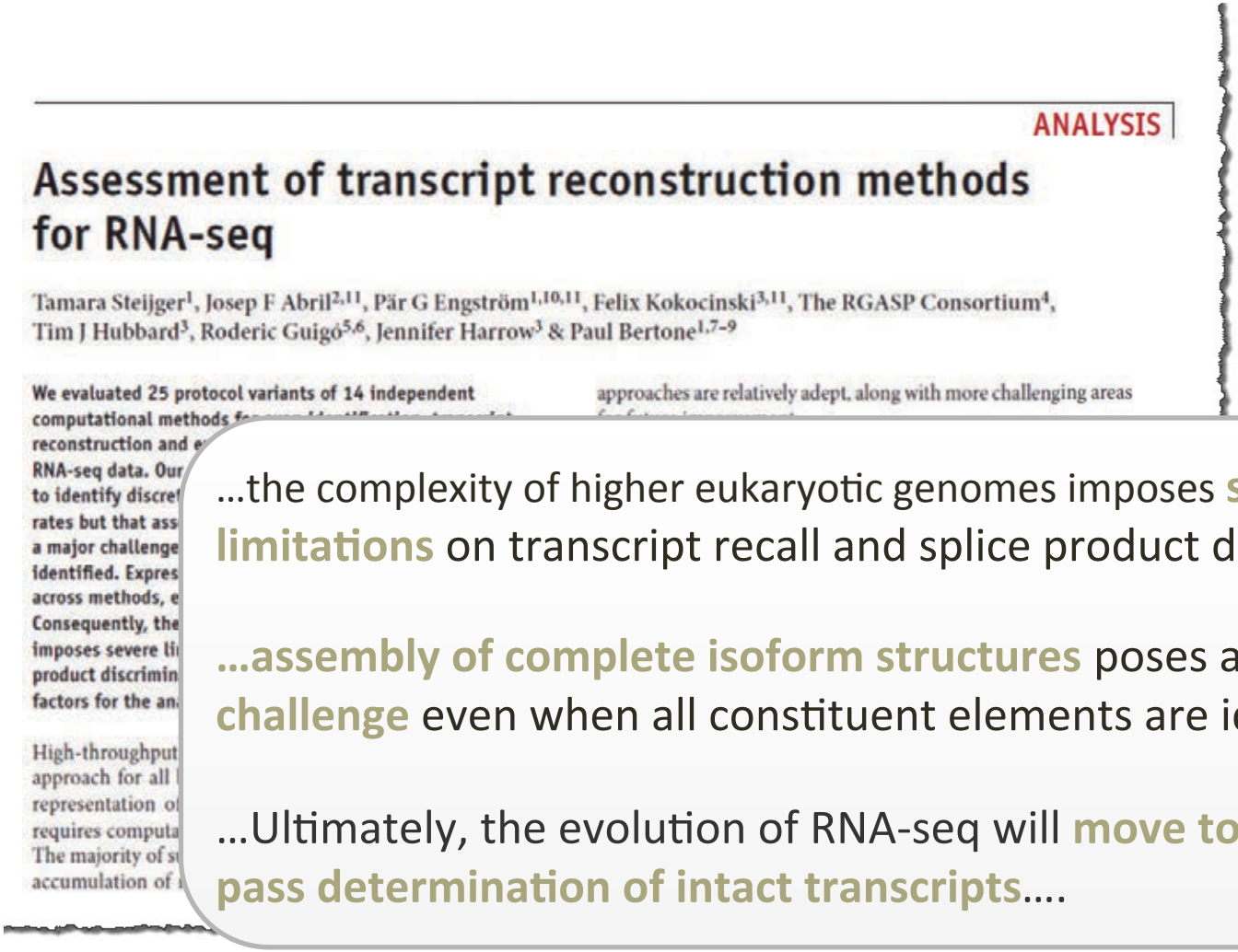
Two individuals for sequencing:

Susceptible and Resistant to white pine blister rust (WPBR)

Transcriptome Assembly:

- Yield a set of transcripts for scaffolding the genome
- Develop resources for full annotation of the genome
- Identify candidate genes for resistance (white pine blister rust)

Difficulties Resolving Transcriptomes with Short Reads



ANALYSIS

Assessment of transcript reconstruction methods for RNA-seq

Tamara Steijger[1], Josep F Abril[2,11], Pär G Engström[1,10,11], Felix Kokocinski[3,11], The RGASP Consortium[4], Tim J Hubbard[3], Roderic Guigó[5,6], Jennifer Harrow[3] & Paul Bertone[1,7–9]

We evaluated 25 protocol variants of 14 independent computational methods for exon identification and transcript reconstruction and e
RNA-seq data. Our
to identify discre
rates but that ass
a major challenge
identified. Expres
across methods, e
Consequently, the
imposes severe li
product discrimin
factors for the an

High-throughput
approach for all
representation o
requires computa
The majority of s
accumulation of

approaches are relatively adept, along with more challenging areas

…the complexity of higher eukaryotic genomes imposes **severe limitations** on transcript recall and splice product discrimination…

…assembly of complete isoform structures poses a **major challenge** even when all constituent elements are identified…

…Ultimately, the evolution of RNA-seq will **move toward single-pass determination of intact transcripts**….

Transcriptome Sequencing Strategy

Hybrid Approach to Sequencing:

Illumina

Clean short reads → Noisy assembly (many scaffolds)

Pacbio

Long noisy reads → High quality assembly (few scaffolds)

HiSeq
Average length=(100x2)
180 million reads/lane
Accuracy: 99%

MiSeq
Average length=(300x2)
25 million reads/lane
Accuracy: 99.6%

PacBio SMRT II Iso-Seq
Size selected lengths (5-6Kb, 10% over 10Kb)
40,000 reads/SMRT cell (run)
Accuracy: 86%

Workflow Options for Full-Length Transcripts with double BluePippin Size Selection

Total RNA

↓

polyA+ RNA

↓

SMARTer® PCR cDNA (Clontech)

↓

PCR Optimization

↓

BluePippin Size Selection
1-2 kb
2-3 kb
3-6 kb

→ Large-Scale PCR

↓

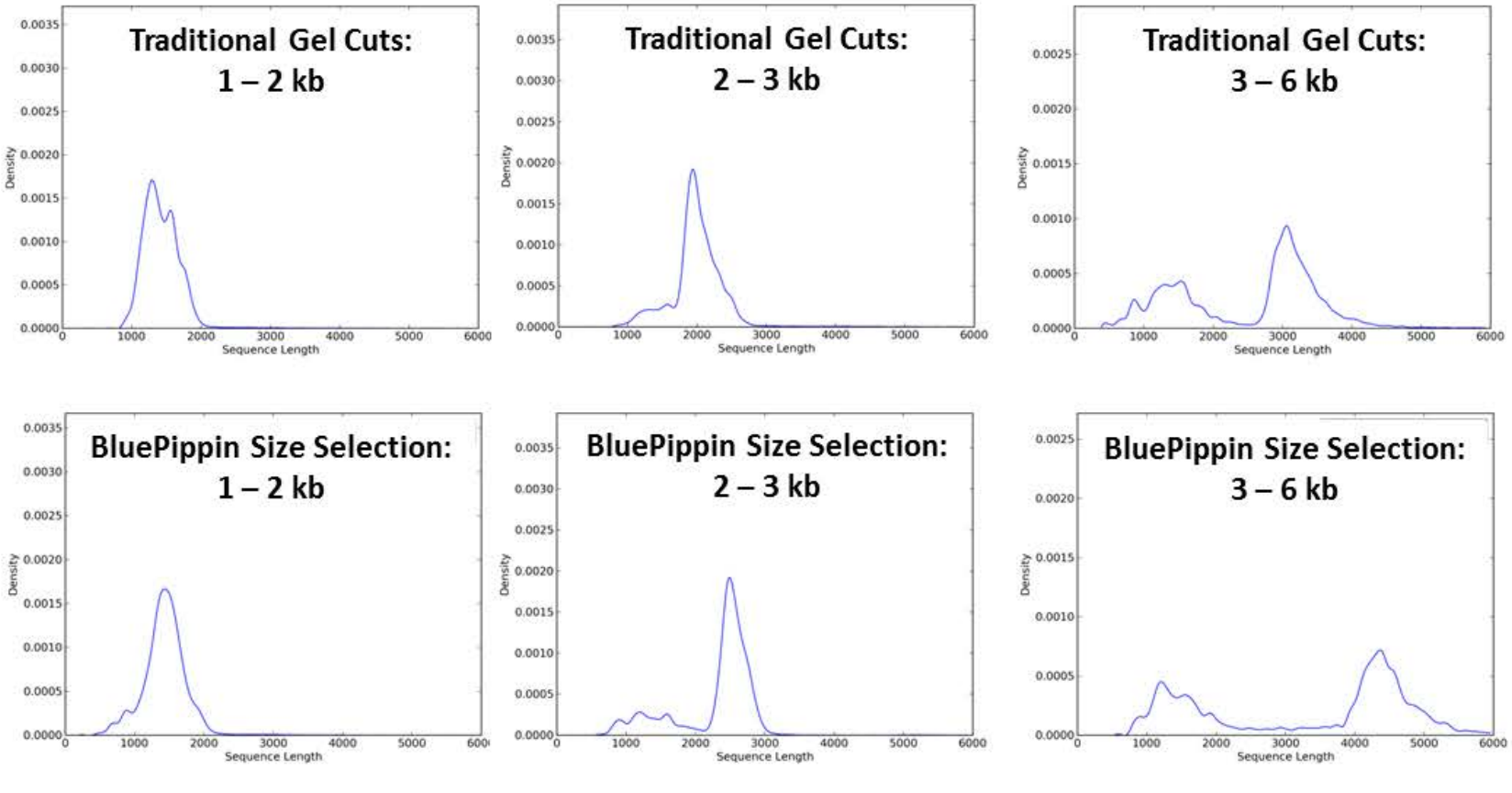
PacBio Template Preparation

↓

BluePippin 3-6K

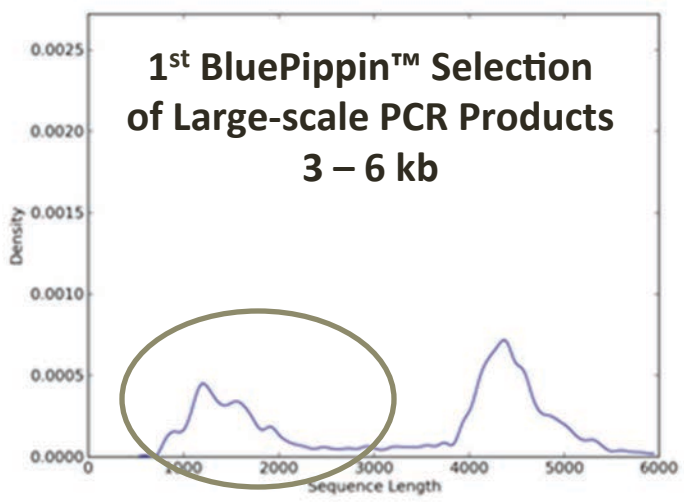
Randi Famula – UCDavis
Nicole Rapicavoli - PacBio

BluePippin as an alternative to gel cutting

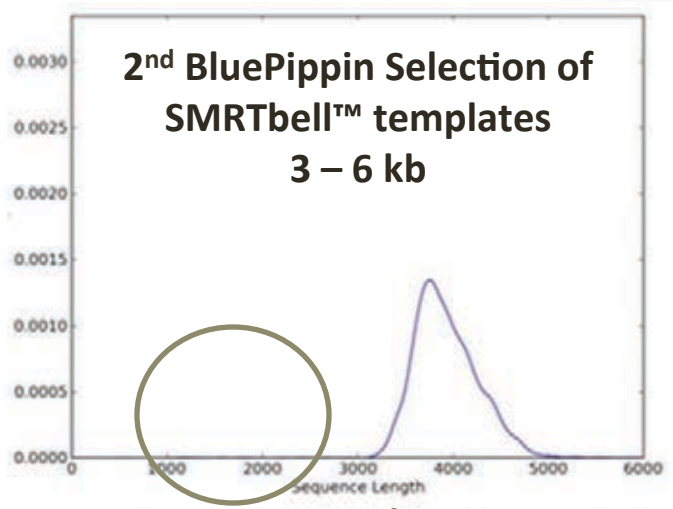


- BluePippin size selected samples tend to give longer transcripts within the target range
- 3-6K fraction from gel cuts have transcripts up to 4.5 kb long where as the fraction from the BluePippin sample have transcripts up to 6 kb

Improving the Detection of 3-6 kb Transcripts



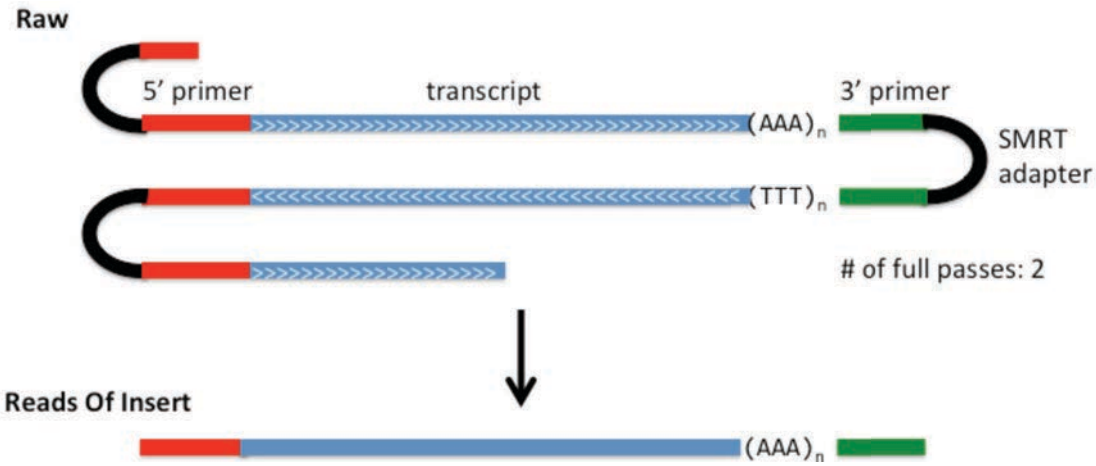
First size selection to select 3-6 kb fraction

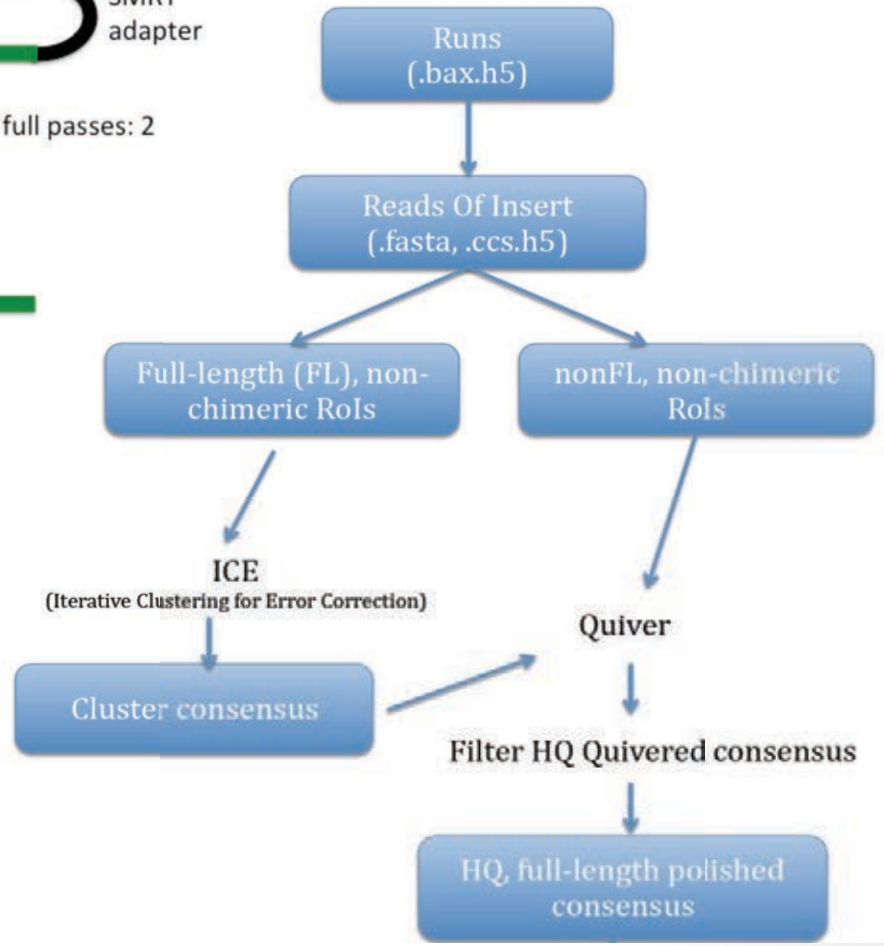


Final size selection of 3-6 kb SMRTbell library

- Shorter SMRTbell templates will impact the loading of 3-6 Kb templates
- Removal of short templates can be accomplished by a second round of BluePippin size selection on the library

Full-length reads with Iso-Seq





Sequencing Statistics

Samples selected for technology comparison

Transcript selection (CDS identification with Transdecoder + clustering at 90 % identitywith UCLUST)

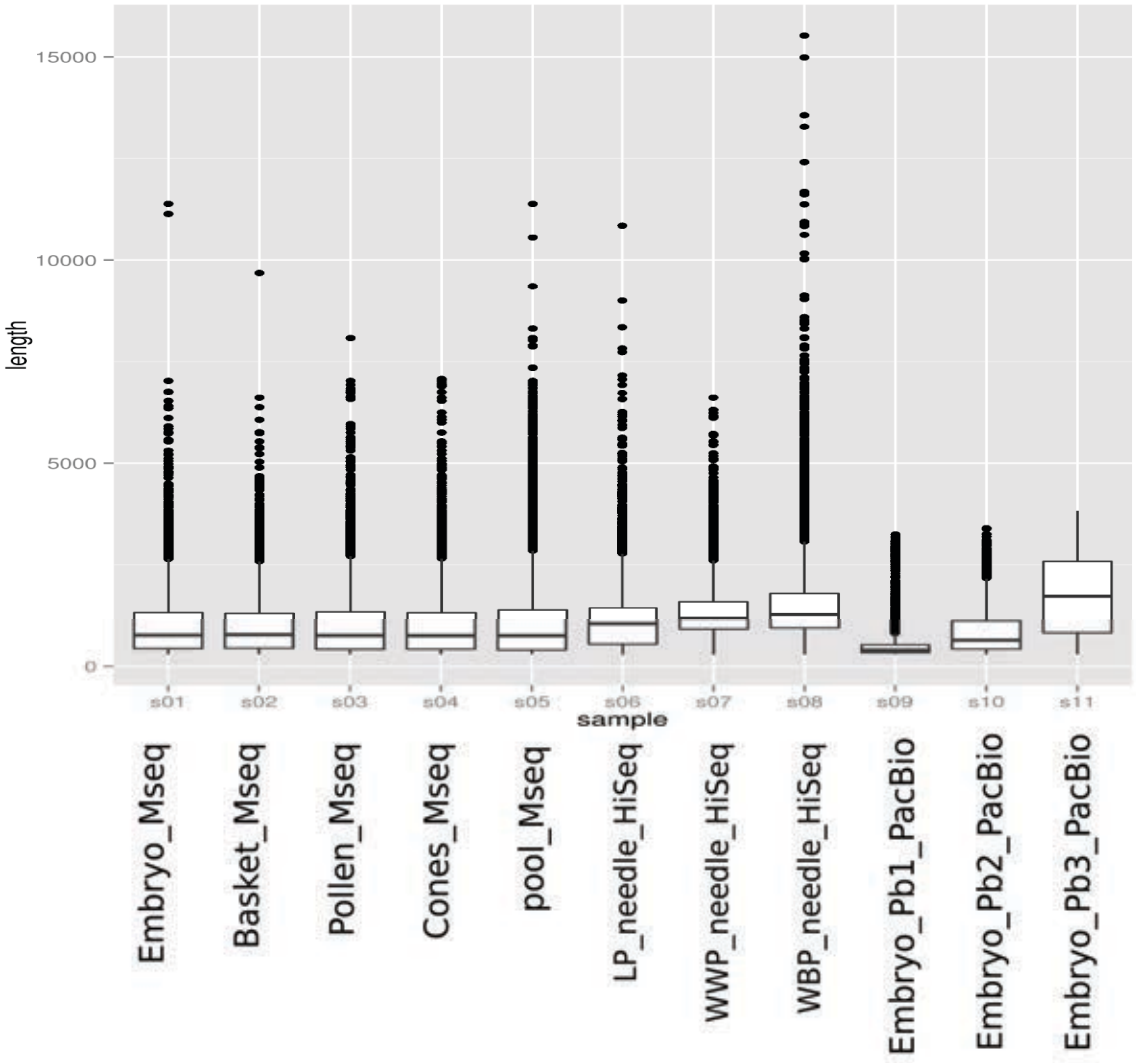
A. Illumina libraries

Library name	Total trinity 'genes'	Total trinity transcripts	Total assembled bases	Percent GC	Contig N50	Average contig length	selection (CDS identification + clustering)
Basket_Mseq	52998	80065	75802148	40.98	1137	946.76	8268
Pollen_Mseq	115809	155024	104420360	40.48	891	673.58	9008
Cones_Mseq	69359	102314	94035094	40.65	1086	919.08	9382
Embryo_Mseq	48942	**73271**	72587376	40.71	1209	**990.67**	**7892**
pool_Mseq	116155	163590	143195342	39.97	1012	875.33	12679

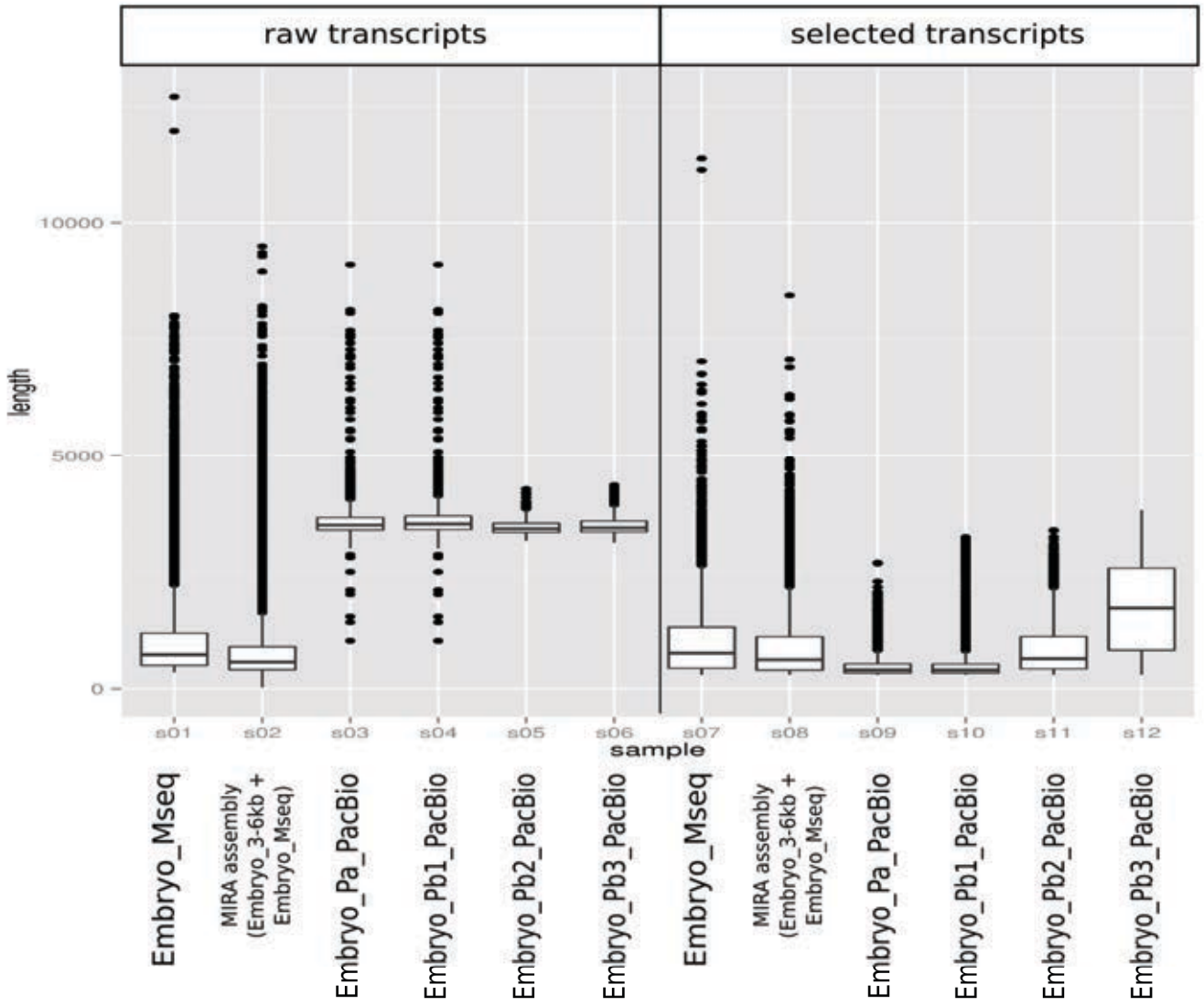
B. SMRT cells from PacBio libraries

Sample name	size selection	Number SMRT cells	Reads of Insert	Read Quality of Insert	Mean Number of Passes	Number of full-length non-chimeric reads (Pa)	Average full-length non-chimeric read length	Number of consensus isoforms (Pb1)	Average consensus isoforms read length	Number of polished low-quality isoforms (Pb2)	Number of polished high-quality isoforms (Pb3)	transcript selection (CDS identification + clustering)
DCS_Stem	1 Kb	4	201744	0,92	7,75	102092	1540,50	77603	1547,75	2564	6574	--
DCS_Stem	2 Kb	4	259868	0,92	8,00	103022	1869,75	81386	1922,00	2398	5646	--
Embryo	1 Kb	1	60646	0,93	9,00	14343	2074,00	10890	2101,00	497	835	--
Embryo	2 Kb	4	256631	0,92	7,50	125057	1586,25	88710	1678,75	2881	9854	--
Embryo	3-6 Kb	1	44815	0,88	3,00	**19413**	**3570,00**	14100	**3595,00**	1090	539	**8940**
ConesB	1 Kb	3	162898	0,90	6,33	61611	2072,33	46531	2101,00	2062	3780	--
ConesB	2 Kb	4	224284	0,91	7,00	84540	1938,25	64185	1990,50	2563	5487	--
Strobilli	1 Kb	4	228401	0,92	7,00	111972	1595,00	67126	1591,75	3158	9830	--
Strobilli	2 Kb	4	199885	0,89	5,00	83411	2275,50	55594	2243,75	3415	5117	--

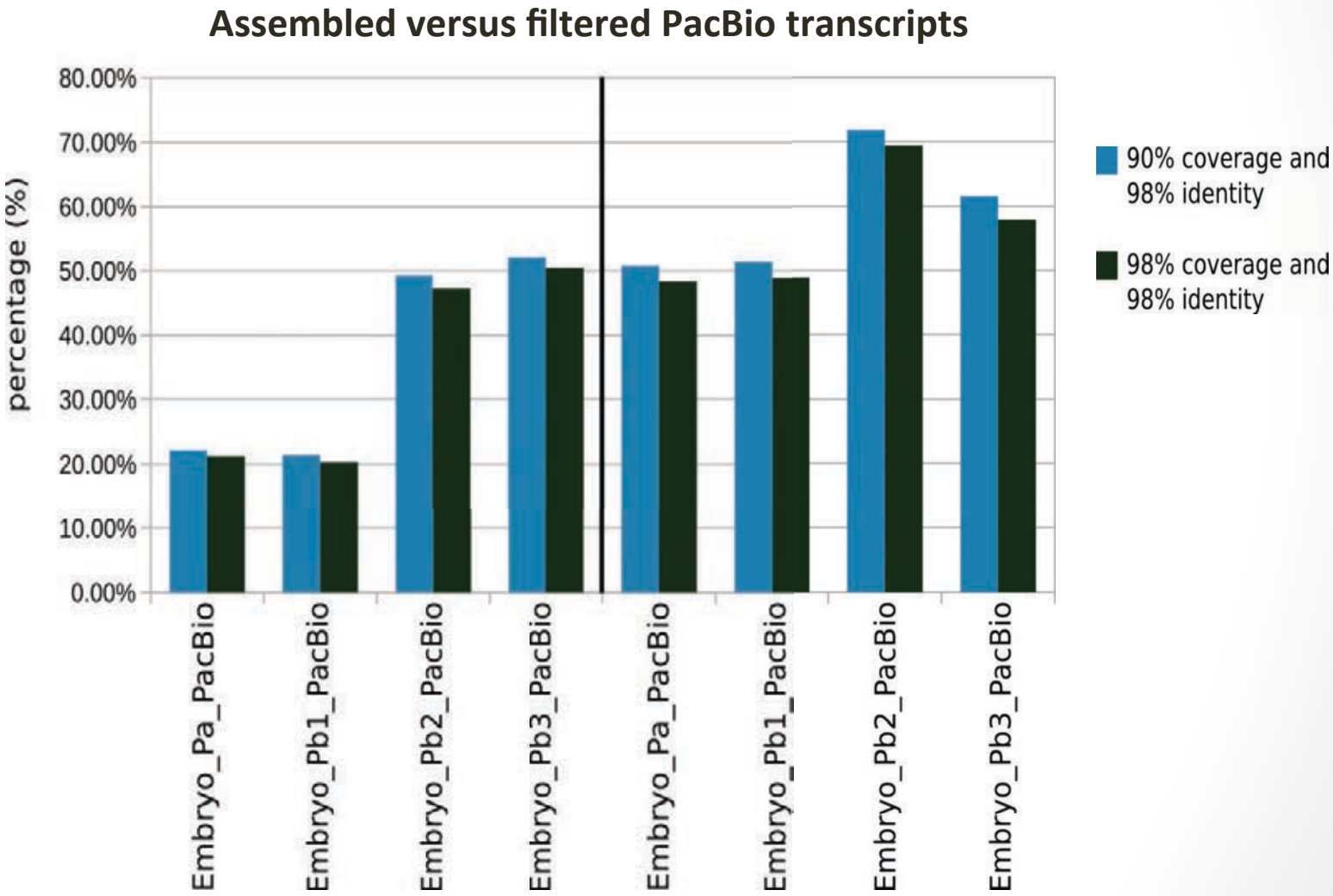
Assembled Transcript Lengths Technology Evaluation



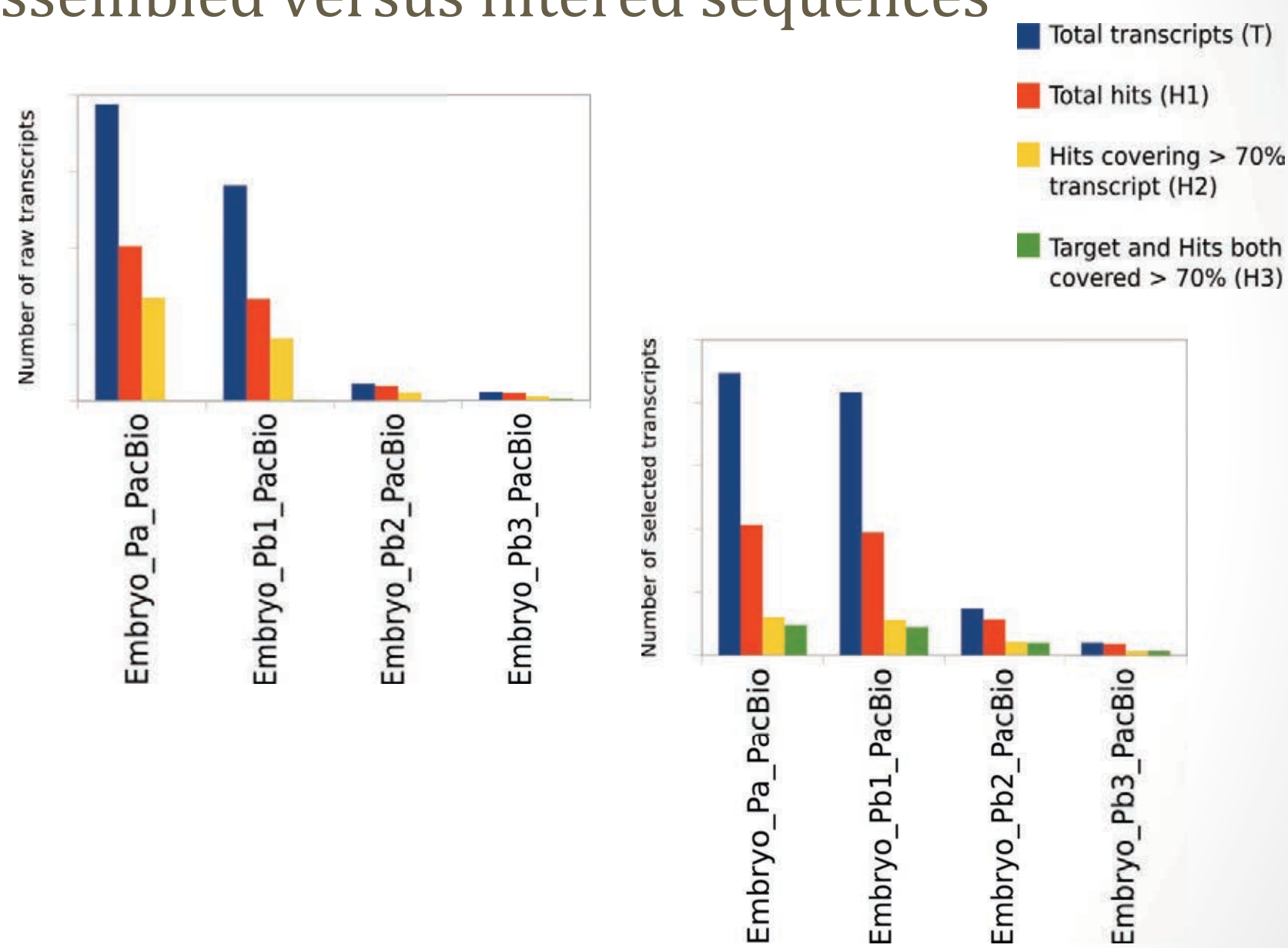
Assembled versus filtered transcripts



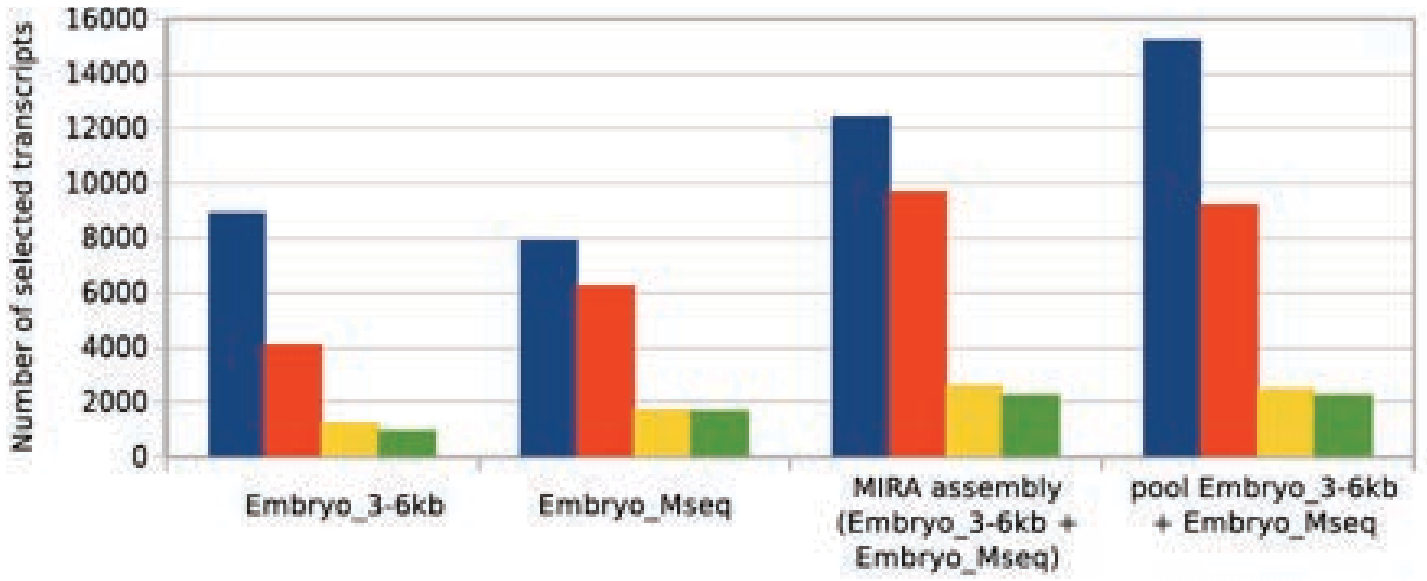
Alignment to the sugar pine genome Assembled versus filtered transcripts



Comparison of annotation rates Assembled versus filtered sequences

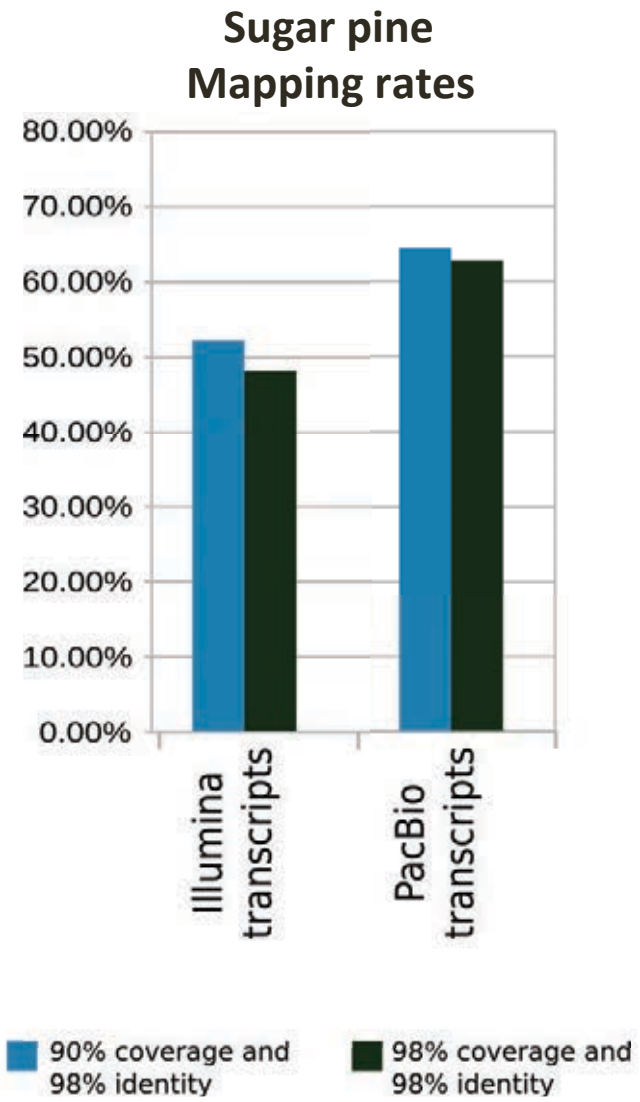


Comparison of annotation rates among assembly approaches



- MIRA performs a hybrid assembly with MiSeq reads and error corrected PacBio reads
- Pooled method clusters independent assemblies
  - SMRT assembly of Embryo PacBio
  - Trinity assembly of Embryo MiSeq

Aligning to the sugar pine genome (v0.5)

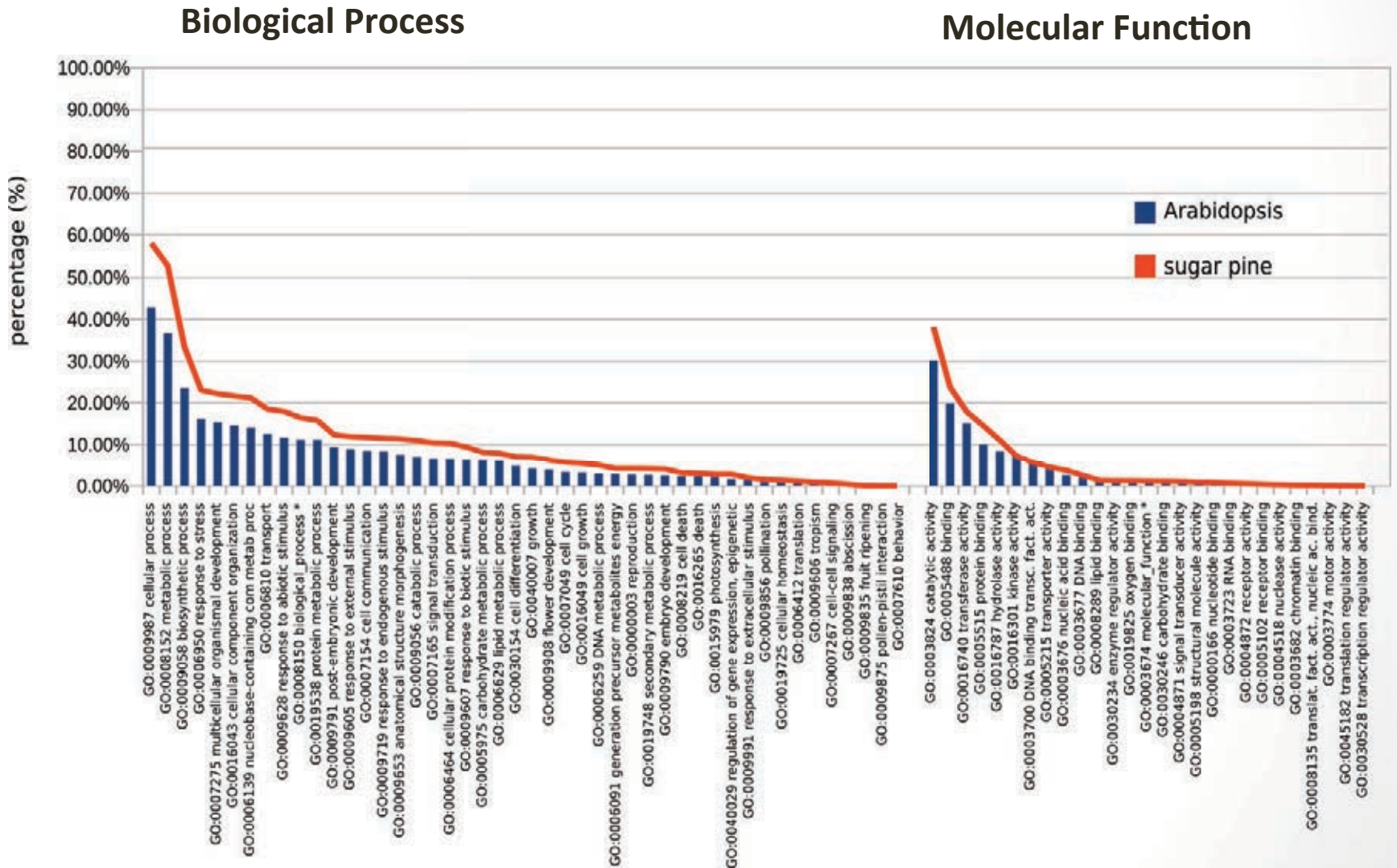


Final Scaffolding Sets

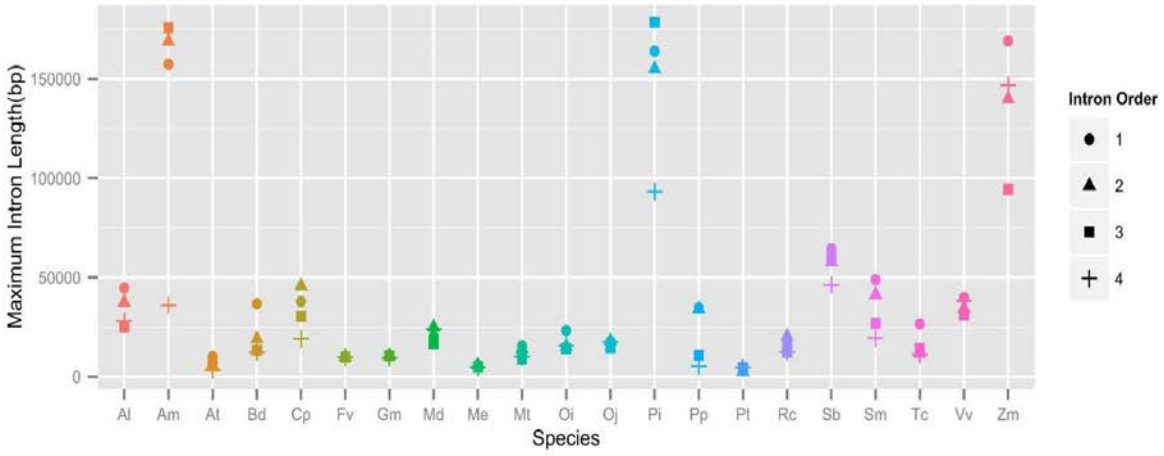
MIRA assembly and pooled assembly did not yield significant differences in annotation or genome mapping rate

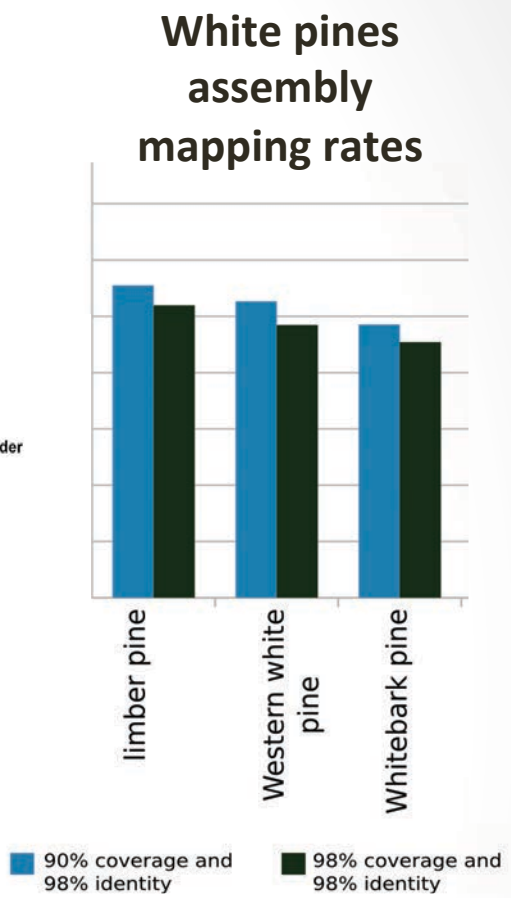
Total: 66,132
High quality: 17,167

Evaluating Completeness of Scaffolding Set



Intron Analysis

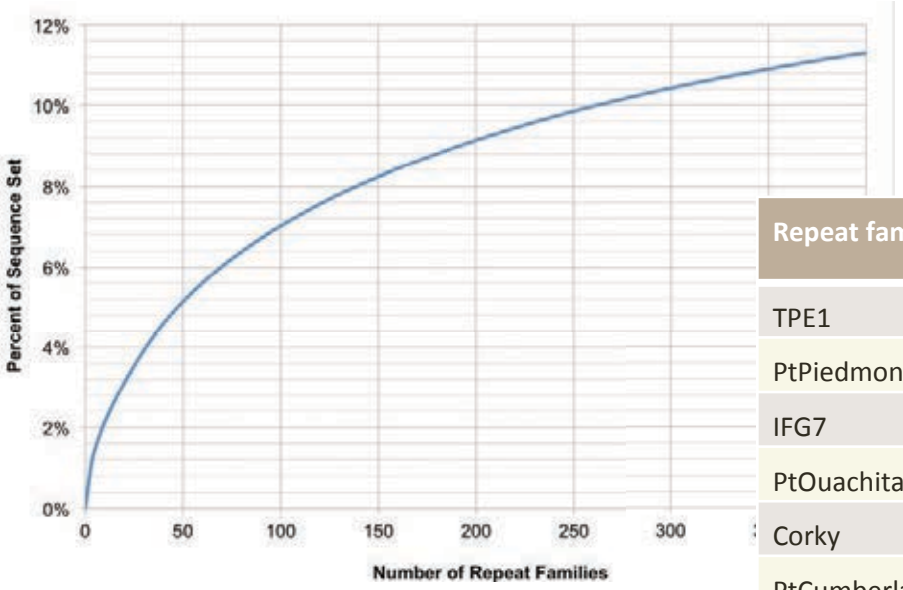




Species	Avg. Intron Length	Max Intron Length (Kbp)	Avg. number of exons
limber pine	3273	146.6	4.8
western white pine	3155	146.6	4.9
sugar pine	6255	273.4	5.9

Novel Repeat Elements
Diverged LTRs are annotated as 6,270 novel families

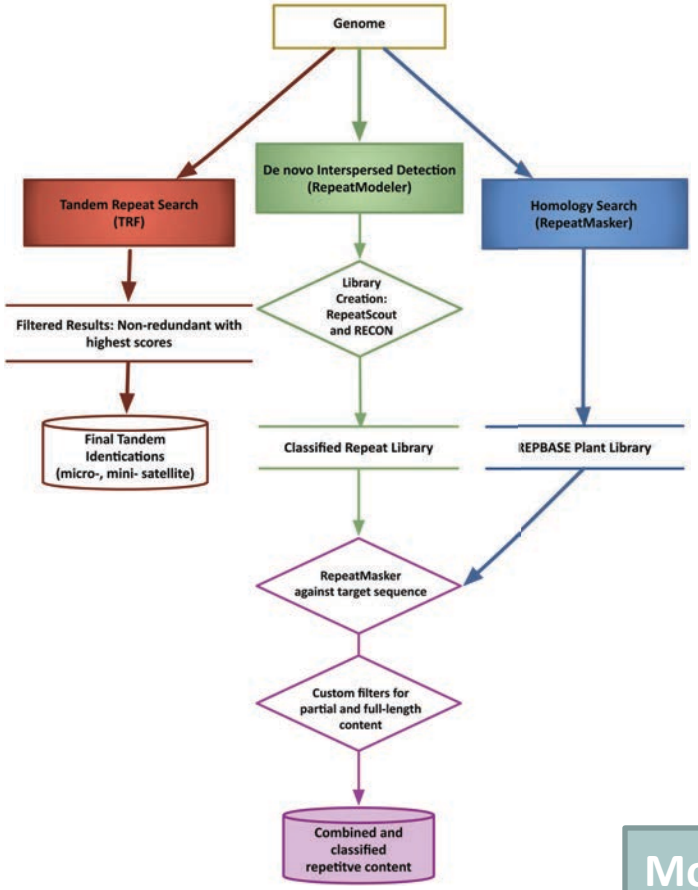
Top 400 elements only cover 12% of the combined sequence sets

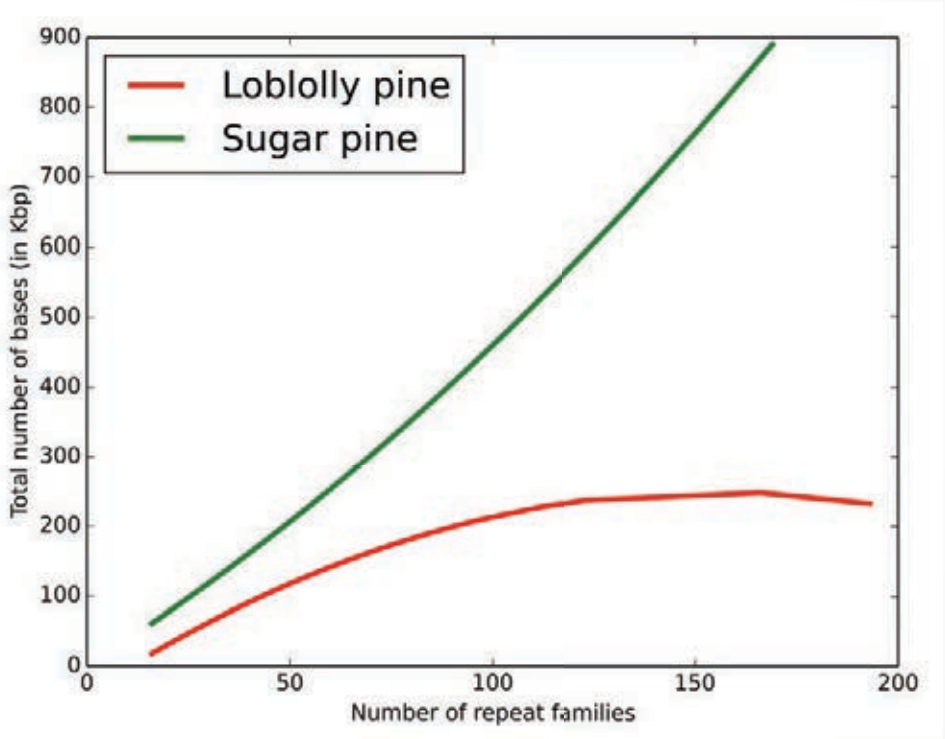


Repeat family	Full-length Copies	Length (bp)	Sequence Set
TPE1	159	1,077,598	0.39%
PtPiedmont	133	969,109	0.35%
IFG7	162	956,018	0.34%
PtOuachita	47	576,871	0.21%
Corky	78	469,286	0.17%
PtCumberland	67	431,492	0.16%
PtBastrop	38	378,631	0.14%
PtOzark	32	378,020	0.14%
PtAppalachian	67	367,653	0.13%
PtPineywoods	68	322,632	0.12%
PtAngelina	24	309,248	0.11%
Gymny	24	291,479	0.11%
PtConagree	50	285,850	0.10%
PtTalladega	33	274,826	0.10%
Total	982	7,088,713	2.56%

Repeat Sequence Detection

Developing strategy and resources in fosmids

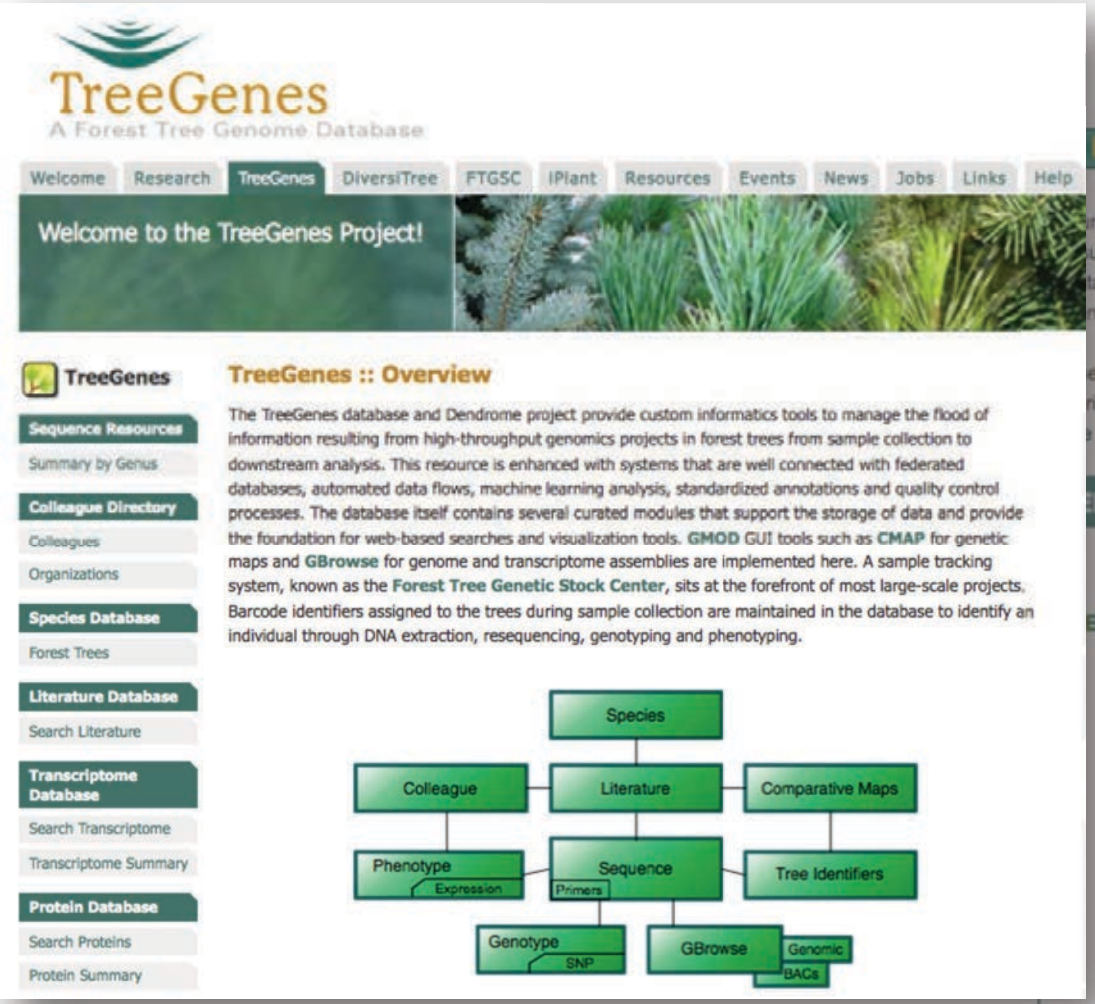


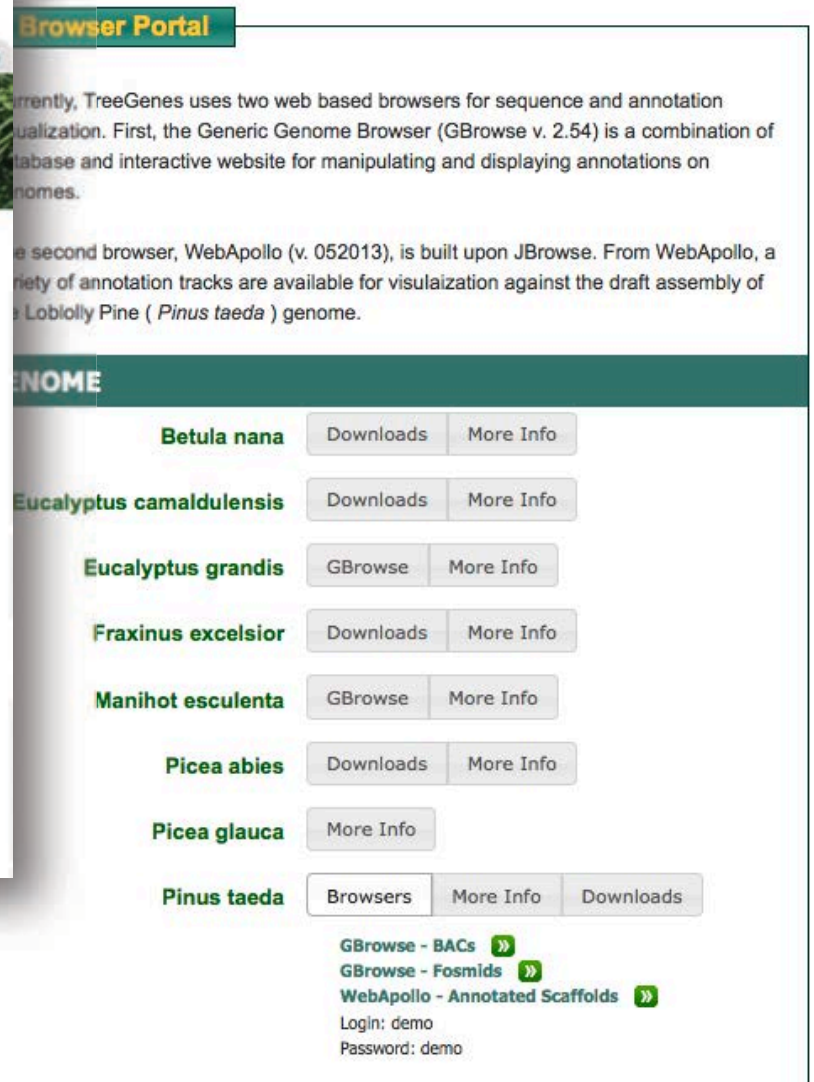


Monday at 1:20pm P0988 - Repeat Sequence Characterization in Sugar Pine (*Pinus lambertiana*) and Loblolly Pine (*Pinus taeda*) (**Robin Paul**)

Dendrome Project

TreeGenes Database to Distribute Transcriptome and Genome





Sunday at 1:30pm – Forest Tree Workshop
Tuesday at 1:50pm – TreeGenes Computer Demo

Acknowledgements

University of Connecticut

- Daniel Gonzalez-Ibeas
- Ethan Baker
- Sam Ginzburg
- Robin Paul

University of California, Davis

- Pedro J. Martinez-Garcia
- Kristian Stevens
- John L. Liechty
- Patricia Maloney
- Randi Famula
- Hans Vasquez-Gross
- Emily Grau
- Charles Langley
- David Neale

University of Colorado

- Jeffrey Mitton

Texas A&M University

- Carol Loopstra
- Jeff Puryear

USDA Forest Service

- Detlev Volger
- Camille Jensen
- Annette Delfino-Mix
- Jessica Wright

Indiana University

- Keithanne Mockaitis

Pacific Biosciences

- Nicole Rapicavoli

More Information on the sugar pine transcriptome:

Monday at 11:40am P0987
Daniel Gonzalez-Ibeas

PineRefSeq Genome Team
University of Maryland
Johns Hopkins University
CHORI