A comprehensive study of the sugar pine (*Pinus lambertiana*) transcriptome implemented through diverse next-generation sequencing approaches

Pedro J. Martínez-García, Daniel Gonzalez-Ibeas, Randi A. Famula, Annette Delfino-Mix, Kristian A. Stevens, Jeffrey D. Puryear, Carol A. Loopstra, Charles H. Langley, David B. Neale, Jill L. Wegrzyn

Conifer Reference Assemblies

Loblolly pine (*P. taeda*)
Assembly LP_v1.01
~ 65X coverage
Total Sequence: 23.2 Gbp
N50 Scaffold: 66.9Kbp

Sugar pine (*P. lambertiana*)
(just released)
Assembly SP_v1.0
~ 53X (PE)
17.3X (MP)
0.5-1X (DiTags)
Total Sequence: 31Gbp
N50 Scaffold: 246.5Kbp

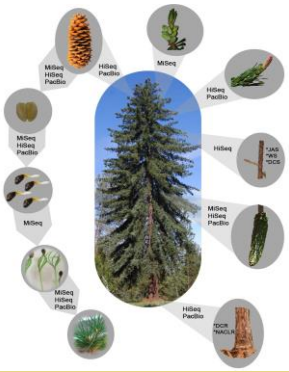
Douglas fir (*Pseudotsuga menziesii*)
Assembly Psme_v1.0
(just released)
~ 60X (PE)
11X (MP)
Total Sequence: 15.7Gbp
N50 Scaffold: 340.7Kbp

Previous Genomic resources for white pines

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

Generating new resources for sugar pine

Tissues collection



Stress Treatments

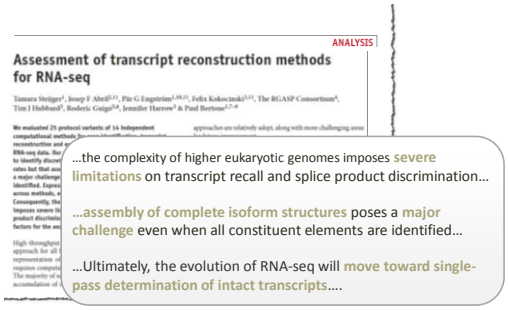
Needles

Stems

Roots

- Drought Stress
- Cold Stress
- Heat Shock
- Flooding
- Salt
- Wounding
- Jasmonic Acid
- Blister rust disease

Difficulties Resolving Transcriptomes with Short Reads


ANALYSIS

Assessment of transcript reconstruction methods for RNA-seq

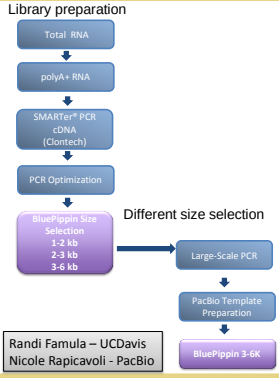
...the complexity of higher eukaryotic genomes imposes severe limitations on transcript recall and splice product discrimination...

...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....

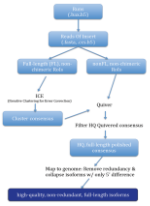
Full-length reads with Iso-Seq

Library preparation



Different size selection

Bioinformatics workflow



Clustering step ICE/Quiver:
Consensus isoforms (Pb1)
Low quality polished sequences (Pb2)
High quality polished sequences (Pb3)

Randi Famula – UCDavis
Nicole Rapicavoli - PacBio

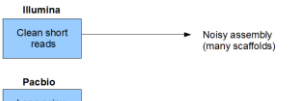
Transcriptome Sequencing Strategy

Hybrid Approach to Sequencing:

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

Mi-Seq
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%



Illumina

Description	Tissue	Technology	Number of reads after QC	Total trinity 'genes'	Total trinity transcripts	libs/lane
Blister Resistant needles (LCO2-03)	needle	HiSeq	264207835	73581	112529	1
Seedling was not drought-stressed	Root	HiSeq	232870127	100782	150944	1
Seedling was not drought-stressed	Stem	HiSeq	272808440	64239	104148	1
Germinating sugar pine seed	Embryo	HiSeq	232349072	50340	79393	1
Methyl jasmonate treatment for 5 hrs	Stem	HiSeq	255543502	58239	92768	1
NaCl treatment	Root	HiSeq	237568138	68961	101178	1
Female cones (2cm)	Female cones	HiSeq	241875834	51363	79350	1
Female strobili near pollination	Female strobili	HiSeq	271381487	71438	110173	1
Wounding treatment	Stem	HiSeq	268217058	71226	110544	1
Blister Resistant stem (LCO2-03)	Stem	MiSeq	33374491	64531	94955	1
Seedling slowly drought-stressed	Needle	MiSeq	32947205	69877	95471	1
Pollen	Pollen	MiSeq	33131689	69359	102314	1
"Basket stage" seedling	Root, stem and needles	MiSeq	33725577	52998	80065	1
Germinating sugar pine seed	Embryo	MiSeq	28838541	48942	73271	1
Early female conelets before pollination	female conelets	MiSeq	37166765	115809	155024	1
Pollen cones	Pollen cones	MiSeq	27597667	71414	102389	1
		Total	**2503603428**	**1103099**	**1644516**	**16**

2.5 billion reads

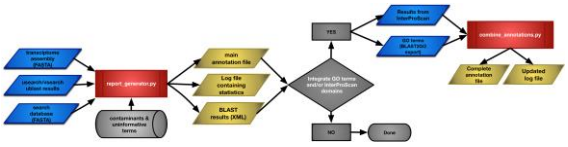
PacBio

Description	Tissue	size selection	Number SMRT cells	Reads of Insert after QC	Number of full-length non-chimeric reads	Number of consensus isoforms	Number of polished low-quality isoforms	Number of polished high-quality isoforms
Seedling was not drought-stressed	stem	1 Kb	4	201744	102092	77603	2564	6574
Seedling was not drought-stressed	stem	2 Kb	4	259868	103022	81386	2398	5646
Germinating sugar pine seed	Embryo	1 Kb	1	60646	14343	10890	497	835
Germinating sugar pine seed	Embryo	2 Kb	4	256631	125057	88710	2881	9854
Germinating sugar pine seed	Embryo	3-6 Kb	1	44815	19413	14100	1090	539
2 cm female cones	female cones	1 Kb	3	162898	61611	46531	2062	3780
2 cm female cones	female cones	2 Kb	4	224284	84540	64185	2563	5487
2 cm female cones	female cones	3-6 Kb	3	249945	85231	60218	41383	18834
Female strobili near pollination	female strobili	1 Kb	4	228401	111972	67126	3158	9830
Female strobili near pollination	female strobili	2 Kb	4	199885	83411	55594	3415	5117
Female strobili near pollination	female strobili	3-6 Kb	3	238685	82948	53428	34317	19111

1.6 million reads

Eukaryote Non-Model Transcriptome Annotation Pipeline

QC-sickle -> Trinity_denovo_asembly -> Transdecoder -> USEARCH/UCLUST + enTAP



(enTAP, https://github.com/SamGinzburg/WegrzynLab)

Transcriptome statistics

Assembled transcripts (number of sequences)

Total transcripts	278812
HiSeq	75175
MiSeq	45524
PacBio	158113

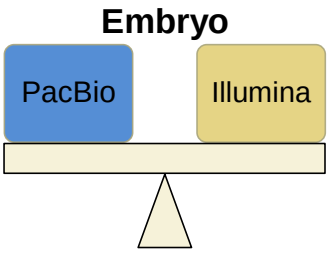
Set of unique transcripts/full length transcripts

Number (scaffolding)	33113
Average length	1144
Shortest transcript	300
Largest transcript	13236
N50 Statistic	1386

Functional annotation

Annotated	30839
Informative	26568
Uninformative	3923
Unannotated	1243
Contaminants	1399

Comparison of Sequencing Technologies



TRANSCRIPTS
Length - Completeness - Mapping rates

TRANSCRIPTOME
Coverage - Diversity

Comparison – Transcripts

- PacBio highest number of complete coding regions (8940 PacBio > 8782 HiSeq > 7892 MiSeq)
- PacBio longer lengths before TS, BUT after TS lengths were similar between technologies
- Less 4% of transcripts were full-length (70%-70%) with either technology before TS, with a increment up to 21% of full-length after TS
- Less than 50% of PacBio transcripts mapped (Illumina >70%) (after TS: PacBio ~60%, Illumina ~90%)

Comparison - Transcriptome – Coverage and Diversity

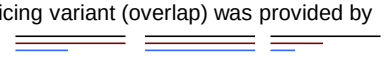
Embryo transcriptome: 17505 unique embryo transcripts mapped against SP_v1.0 (100%coverage/100%identity)

- 9249 transcripts map uniquely
- 74% covered by PacBio
- 1615 unique hits covered by all three technologies

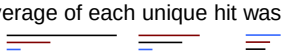
Comparison - Transcriptome – Coverage and Diversity

1615 covered by all three technologies

- Longest splicing variant (overlap) was provided by HiSeq



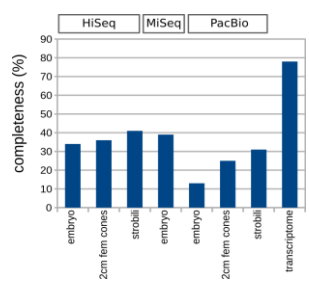
- A single technology producing the longest splice variants (in number) was HiSeq
- Contribution to coverage of each unique hit was better by HiSeq



- Highest number of non-redundant splice variants was provided by PacBio

Comparison - Transcriptome - Completeness

- Lower coverage and higher variation by PacBio



Benchmarking Universal Single-Copy Orthologs **BUSCO** (Simão et al 2015)

Take-home message…

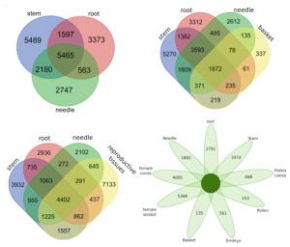
PacBio: better in splice variant detection and highly productive for full-length (not necessarily the longest) final transcripts

Illumina: better in term of transcripts coverage/length, longest splice variant and high depth for expression studies

Differential Expression – without replicates

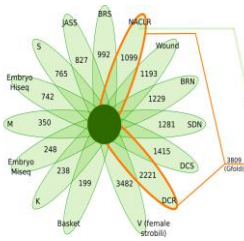
Gfold
(Feng et al. 2012)
-without replicates-

- Stem tissue more similar to needle tissue
- Low number in Basket
- Reproductive tissues more similar to stem
- High number provided by female reproductive tissue



Differential Expression – without replicates

- More than 10000 transcripts shared by all libraries
- Overall 5958 transcripts were differential expressed with fold change >2.0
- NACL-roots 1099 transcripts library specific, 233 differently expressive



GO-ID	Term	Category	FDR
NACLR library-specific transcripts			
GO:0046034	ATP metabolic process	P	6.55E-002
GO:0016469	proton-transporting two-sector ATPase complex	C	7.34E-002
NACLR vs. untreated control differentially expressed transcripts (Gfold)			
GO:0042555	MCM complex	C	5E-02
GO:0043168	anion binding	F	8E-18
GO:0005524	ATP binding	F	2E-09
GO:0016887	ATPase activity	F	5E-03
GO:0005034	osmosensor activity	F	5E-02
GO:0010817	regulation of hormone levels	P	6E-06
GO:0048765	root hair cell differentiation	P	2E-02
GO:0048767	root hair elongation	P	7E-03

Differential Expression – GO categories

GO-ID	Term	Category	FDR
NACLR treatment			
GO:0042555	MCM complex - mostly helicases	C	0,049359
GO:0043168	anion binding	F	7,84E−18
GO:0005524	ATP binding	F	2,14E−09
GO:0016887	ATPase activity	F	0,004976
GO:0005034	osmosensor activity - three different histidine kinases	F	0,046233
GO:0010817	regulation of hormone levels	P	6,04E−06
GO:0048765	root hair cell differentiation	P	0,018912
GO:0048767	root hair elongation	P	0,006801
GO:0009809	lignin biosynthetic process	P	0,023398
JASS treatment			
GO:0010583	response to cyclopentenone - most topoisomerases	P	0,007481
GO:0009605	response to external stimulus	P	0,018085
GO:0043207	response to external biotic stimulus	P	0,021576
GO:0051707	response to other organism	P	0,021576
GO:0016458	gene silencing	P	0,023677
GO:0051567	histone H3-K9 methylation	P	7,68E−09
GO:0042742	defense response to bacterium	P	0,028652
GO:0010476	gibberellin mediated signaling pathway	P	0,011213
GO:0042221	response to chemical	P	1,90E−07

Differential Expression – GO categories

GO-ID	Term	Category	FDR
Wounding			
GO:0006950	response to stress	P	8,39E−09
GO:0006952	defense response	P	2,54E−06
GO:0005911	cell-cell junction	C	0,000132
GO:0030855	epithelial cell differentiation	P	0,010485
GO:0009913	epidermal cell differentiation	P	0,010485
GO:0008544	epidermis development	P	0,013693
GO:0060429	epithelium development	P	0,017064
GO:0042545	cell wall modification	P	0,035399
Reproductive tissue			
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	P	2,86E−07
	oxidoreductase activity, acting on diphenols and related substances		
GO:0016682	as donors, oxygen as acceptor	F	0,002152
GO:0009751	response to salicylic acid	P	0,002236
GO:0010333	terpene synthase activity	F	6,87E−14
GO:0009740	gibberellic acid mediated signaling pathway	P	0,000720
GO:0048506	regulation of timing of meristematic phase transition	P	0,000373
GO:0007389	pattern specification process	P	0,002009
GO:0009955	adaxial/abaxial pattern specification	P	0,008520
GO:0007165	signal transduction	P	1,66E−17
GO:0050793	regulation of developmental process	P	2,67E−05
GO:0010476	gibberellin mediated signaling pathway	P	0,000174
GO:0009686	gibberellin biosynthetic process	P	0,000641

Genome expansion in conifers

- Expansion consequence of transposable element (TE) proliferation (80%) rather than genome duplications
- Unique small RNA profile (24-nt) in conifers associated with epigenetic processes and control of repetitive element proliferation
- Potential lineage-specific Dicer-like (DCL) (key in sRNA biogenesis) proteins were identified (Dolgosheina et al. 2008) in conifers.
- 12 transcripts in reproductive tissue with sequence similarity and domain topology matching DCL features expanding their characterization in sugar pine.
- 6 were supported by gene models (genome v1.0).

Additional 1.7 billion Illumina reads

Samples	Technology	Reads of Insert after QC	Total trinity transcripts	Total trinity 'genes'	line
Female strobili	MiSeq	32260898	218646	180581	89203
Female cones	MiSeq	31376822	128302	103083	59841
Female conelets (PM)	Hiseq	27460906	61700	44981	36784
"Basket stage" seedling	Hiseq	26048750	55835	40070	34728
pollen (PM)	Hiseq	29461236	56963	42176	32970
pollen cones (PM)	Hiseq	25211652	67374	49795	8081
Cold Root (Tree 1)	Hiseq	24537808	77775	57633	9942
Primary needle stage (is this tree #1)	Hiseq	24592668	61195	43902	37262
Susceptible WPBR (Stem) #2	Hiseq	99080770	112587	73700	28605
Susceptible WPBR (Root) #2	Hiseq	102288544	128879	88061	22143
Susceptible WPBR (Needle) #2	Hiseq	114994106	146059	97675	28718
Susceptible WPBR (Stem) #1	Hiseq	76938486	105459	69517	28291
Susceptible WPBR (Root) #1	Hiseq	217641798	195948	137558	26214
Susceptible WPBR (Needle) #1	Hiseq	142041716	177094	116622	31923
Resistance WPBR (stem) #1	Hiseq	92196090	111708	78176	16551
Resistance WPBR (root) #1	Hiseq	187244006	195158	138647	27903
Resistance WPBR (needle) #1	Hiseq	109706506	136890	96206	19621
Resistance WPBR (stem) #2	Hiseq	109114558	124300	85202	23869
Resistance WPBR (root) #2	Hiseq	100142518	134295	94199	21987
Resistance WPBR (needle) #2	Hiseq	157203340	147026	99169	23038
	Total	**1729543178**	**2443193**	**1736953**	**607674**

Acknowledgements

University of Connecticut
- Jill Wegrzyn
- Daniel Gonzalez-Ibeas
- Ethan Baker
- Robin Paul

University of California, Davis
- Randi Famula
- Hans Vasquez-Gross
- David Neale
- Kristian Stevens
- Charles Langley

Texas A&M University
- Carol Loopstra
- Jeff Puryear

USDA Forest Service
- Detlev Volger
- Annette Delfino-Mix

Indiana University
- Keithanne Mockaitis