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

Pedro J. Martínez-Garcia, Daniel Gonzalez-Ibeas, Randi A. Famula, Annette Delfino-Mix, Charles H. Langley, David B. Neele, Jill L. Wegrzyn

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

Sugar lime (P. lambertiana) (just released)
Assembly SP_v1.0
- 53X (PE)
17.3X (MP)
0.5-1X (DI Tags)
Total Sequence: 310gb
N50 Scaffold: 246.5kb

Douglas fir (Pseudotsuga menziesii)
Assembly Poma_v1.0
- 60X (PE)
11X (MP)
Total Sequence: 15.70gb
N50 Scaffold: 340.7kb

Previous Genomic resources for white pines

Species | Technology | Reads | Tissue | Reads after QC |
--- | --- | --- | --- | --- |
Sugar pine | Illumina GA llx | SE, 80bp (3 lanes) | needle | 66,894,169 |
Sugar pine (Lorenz et al. 2012) | Roche 454 | SE, 350 bp (avg) | stem | 952,310 |
Limber pine | Illumina HiSeq | PE, 100bp (2 lanes) | needle | 374,191,816 |
Whitebark pine | Illumina HiSeq | PE, 100bp (3 lanes) | needle | 839,389,034 |
Western white pine | Illumina GA llx | PE, 76bp | needle | 208,059,005 |

Generating new resources for sugar pine

Tissues collection

Stress Treatments

- Drought Stress
- Cold Stress
- Heat Shock
- Flooding
- Salt
- Wounding
- Jasmonic Acid
- Blister rust disease
**Difficulties Resolving Transcriptomes with Short Reads**

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-base determination of intact transcript...

---

**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 (5-6 Kb, 10% over 10 Kb)
- 40,000 reads/SMRT cell (run)
- Accuracy: 86%

---

**PacBio**

<table>
<thead>
<tr>
<th>Description</th>
<th>Tissue</th>
<th>Methodology</th>
<th>Insert size (kb)</th>
<th>Full-length reads QC</th>
<th>Polished high quality reads QC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual female cones</td>
<td>Stem</td>
<td>TruSeq RNASeq</td>
<td>2</td>
<td>224,866</td>
<td>2,444,368</td>
</tr>
<tr>
<td>Residual female cones</td>
<td>Stem</td>
<td>SMARTer-V3</td>
<td>2</td>
<td>224,866</td>
<td>2,444,368</td>
</tr>
<tr>
<td>Residual female cones</td>
<td>Stem</td>
<td>Illumina</td>
<td>2</td>
<td>224,866</td>
<td>2,444,368</td>
</tr>
</tbody>
</table>

---

**Eukaryote Non-Model Transcriptome Annotation Pipeline**

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

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

---

**Full-length reads with Iso-Seq**

**Library preparation**

- Total RNA
- Poly-A RNA
- SMARTer V3 (DNA) library
- SMARTer V3
- POLY-ADENYLATION

**Bioinformatics workflow**

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

2.5 billion reads

---

**1.6 million reads**
Transcriptome statistics

<table>
<thead>
<tr>
<th>Assembled transcripts (number of sequences)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total transcripts</td>
</tr>
<tr>
<td>HiSeq</td>
</tr>
<tr>
<td>MiSeq</td>
</tr>
<tr>
<td>PacBio</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set of unique transcripts/full length transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number (scaffolding)</td>
</tr>
<tr>
<td>Average length</td>
</tr>
<tr>
<td>Shortest transcript</td>
</tr>
<tr>
<td>Largest transcript</td>
</tr>
<tr>
<td>N50 Statistic</td>
</tr>
</tbody>
</table>

Functional annotation

| Annotated | 30639 |
| Informative | 26568 |
| Uninformative | 3923 |
| Unannotated | 1243 |
| Contaminants | 1399 |

Comparison of Sequencing Technologies

Embryo

| PacBio | Illumina |

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 an 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

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

Comparison - Transcriptome – Completeness

- Lower coverage and higher variation by PacBio

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

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

**Differential Expression – GO categories**

<table>
<thead>
<tr>
<th>GO-ID</th>
<th>Term</th>
<th>Category</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0016784</td>
<td>epidermal cell differentiation</td>
<td>Biological_process</td>
<td>0.011213</td>
</tr>
<tr>
<td>GO:0015632</td>
<td>ATP binding</td>
<td>Molecular_function</td>
<td>0.096198</td>
</tr>
<tr>
<td>GO:0005824</td>
<td>ATP binding</td>
<td>Molecular_function</td>
<td>0.016897</td>
</tr>
<tr>
<td>GO:0048767</td>
<td>receptor protein tyrosine kinase signaling pathway</td>
<td>Regulation</td>
<td>0.021576</td>
</tr>
<tr>
<td>GO:0009809</td>
<td>lignin biosynthetic process</td>
<td>Molecular_function</td>
<td>0.033588</td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th>Sample Type</th>
<th>Mapped to Genome</th>
<th>% of Sample Mapped to Genome</th>
<th>Total Trinity Transcripts</th>
<th>Total Trinity &quot;genes&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female strobili (MiSeq)</td>
<td>32260898</td>
<td>218646</td>
<td>180581</td>
<td>89203</td>
</tr>
<tr>
<td>Female cones (MiSeq)</td>
<td>31376822</td>
<td>128302</td>
<td>103083</td>
<td>59841</td>
</tr>
<tr>
<td>Female conelets (PM, Hiseq)</td>
<td>27460906</td>
<td>61700</td>
<td>44981</td>
<td>36784</td>
</tr>
<tr>
<td>&quot;Basket stage&quot; seedling (PM, Hiseq)</td>
<td>26048750</td>
<td>55835</td>
<td>40070</td>
<td>34728</td>
</tr>
<tr>
<td>Pollen (PM, Hiseq)</td>
<td>29461236</td>
<td>56963</td>
<td>42176</td>
<td>32970</td>
</tr>
<tr>
<td>Pollen cones (PM, Hiseq)</td>
<td>25211652</td>
<td>67374</td>
<td>49795</td>
<td>8081</td>
</tr>
<tr>
<td>Cold Root (Tree 1, Hiseq)</td>
<td>24537808</td>
<td>77775</td>
<td>57633</td>
<td>9942</td>
</tr>
<tr>
<td>Primary needle stage (is this tree #1, Hiseq)</td>
<td>24592668</td>
<td>61195</td>
<td>43902</td>
<td>37262</td>
</tr>
<tr>
<td>Susceptible WPBR (Stem) #2, Hiseq</td>
<td>154080770</td>
<td>130597</td>
<td>72125</td>
<td>29509</td>
</tr>
<tr>
<td>Susceptible WPBR (Root) #2, Hiseq</td>
<td>142288544</td>
<td>128879</td>
<td>88061</td>
<td>22143</td>
</tr>
<tr>
<td>Susceptible WPBR (Needle) #2, Hiseq</td>
<td>114994106</td>
<td>146059</td>
<td>97675</td>
<td>28718</td>
</tr>
<tr>
<td>Susceptible WPBR (Stem) #1, Hiseq</td>
<td>99080770</td>
<td>112587</td>
<td>73700</td>
<td>28605</td>
</tr>
<tr>
<td>Susceptible WPBR (Root) #1, Hiseq</td>
<td>102288544</td>
<td>128879</td>
<td>88061</td>
<td>22143</td>
</tr>
<tr>
<td>Susceptible WPBR (Needle) #1, Hiseq</td>
<td>114994106</td>
<td>146059</td>
<td>97675</td>
<td>28718</td>
</tr>
<tr>
<td>Resistance WPBR (Stem) #1, Hiseq</td>
<td>92196090</td>
<td>111708</td>
<td>78176</td>
<td>16581</td>
</tr>
<tr>
<td>Resistance WPBR (Root) #1, Hiseq</td>
<td>187244006</td>
<td>195158</td>
<td>138647</td>
<td>27903</td>
</tr>
<tr>
<td>Resistance WPBR (Needle) #1, Hiseq</td>
<td>109706506</td>
<td>136890</td>
<td>96206</td>
<td>19621</td>
</tr>
<tr>
<td>Resistance WPBR (Stem) #2, Hiseq</td>
<td>109114558</td>
<td>124300</td>
<td>85202</td>
<td>23869</td>
</tr>
<tr>
<td>Resistance WPBR (Root) #2, Hiseq</td>
<td>100142518</td>
<td>134295</td>
<td>94199</td>
<td>21987</td>
</tr>
<tr>
<td>Resistance WPBR (Needle) #2, Hiseq</td>
<td>157203340</td>
<td>147026</td>
<td>99169</td>
<td>23038</td>
</tr>
</tbody>
</table>

Total: 1729543178 reads, 2443193 mapped reads, 1736953 "genes".

---

### Acknowledgements

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

University of California, Davis
- Randi Famula
- Hans Vasquez-Giens
- 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