Homoeolog-Specific Expression in Soybean Through Illumina RNA-Sequencing

National Center for Genome Resources

Cornell University

Mizzou University of Missouri

Purdue University

Jimmy Woodward
jew@ncgr.org
Whole Transcriptome Shotgun Sequencing of Soybean High vs Low Protein Lines

- Four seed developmental timepoints
- From each of two soybean NILs
- Each exhibiting either high or low protein content

<table>
<thead>
<tr>
<th>Stage</th>
<th>WW% dry matter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
</tr>
<tr>
<td>4</td>
<td>46</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
</tr>
</tbody>
</table>

Seed Protein

A
B
Area Under the Linkage Group I QTL

The Seed Protein, Oil, and Yield QTL on Soybean Linkage Group I

J. Chunga, H. L. Babkab, G. L. Graefc, P. E. Staswickc, D. J. Leec,
P. B. Cregan, R. C. Shoemaker and J. E. Specht: *Crop Science*
# Data Summary and Alignment to the JGI 7x Soybean Genome

<table>
<thead>
<tr>
<th>Sample</th>
<th>Reads</th>
<th>Reads Aligned</th>
<th>Reads Uniquely Aligned</th>
<th>Total SNPs</th>
<th>In/Dels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gm_A1</td>
<td>7,835,063</td>
<td>5,808,687 (74%)</td>
<td>2,909,296 (37%)</td>
<td>1,374,602</td>
<td>4,429</td>
</tr>
<tr>
<td>Gm_A2</td>
<td>9,673,118</td>
<td>7,490,621 (77%)</td>
<td>2,662,020 (28%)</td>
<td>1,115,564</td>
<td>6,081</td>
</tr>
<tr>
<td>Gm_A3</td>
<td>9,102,649</td>
<td>6,700,125 (74%)</td>
<td>4,788,833 (53%)</td>
<td>1,502,187</td>
<td>4,418</td>
</tr>
<tr>
<td>Gm_A4</td>
<td>7,052,993</td>
<td>5,339,938 (76%)</td>
<td>3,519,076 (50%)</td>
<td>915,889</td>
<td>3,457</td>
</tr>
<tr>
<td>Gm_B1</td>
<td>16,988,687</td>
<td>13,394,686 (79%)</td>
<td>7,783,311 (46%)</td>
<td>2,610,326</td>
<td>8,545</td>
</tr>
<tr>
<td>Gm_B2</td>
<td>7,950,528</td>
<td>6,296,363 (79%)</td>
<td>2,181,665 (27%)</td>
<td>885,383</td>
<td>5,128</td>
</tr>
<tr>
<td>Gm_B3</td>
<td>9,201,789</td>
<td>7,400,210 (80%)</td>
<td>4,106,899 (45%)</td>
<td>1,014,933</td>
<td>5,677</td>
</tr>
<tr>
<td>Gm_B4</td>
<td>8,909,676</td>
<td>6,831,318 (77%)</td>
<td>4,161,016 (47%)</td>
<td>1,011,387</td>
<td>5,316</td>
</tr>
</tbody>
</table>
PCA of Loess Normalized, Log Transformed
Unique Read Counts
Alignment of Solexa Reads Against JGI 7x Soybean Genome

Alignment of reads against reference. [Note: Green (low) and red (high) read coverage]
No gene annotated, but all libraries suggest transcribed region.
No gene annotated, Expression only in Line “A”
**Gene expression within the linkage group I QTL region**

- Four differentially expressed genes
- 124 novel “exons” (An additional 13,300 outside of QTL region).

**Known Genes**

**Novel Exons**
Homoeologous Soybean BACs
Lines indicate homoeolog pairs.
Blue lines indicate inter-homoeolog differential gene expression.
## Soybean Illumina Data Overview

<table>
<thead>
<tr>
<th>Sample</th>
<th>Reads</th>
<th>Reads Aligned</th>
<th>Reads Uniquely Aligned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apical Meristem</td>
<td>6,477,456</td>
<td>5,669,285 (88%)</td>
<td>3,885,504 (60%)</td>
</tr>
<tr>
<td>Flower</td>
<td>5,176,140</td>
<td>4,275,733 (83%)</td>
<td>3,322,166 (64%)</td>
</tr>
<tr>
<td>Green Pods</td>
<td>4,183,024</td>
<td>3,694,647 (88%)</td>
<td>1,442,311 (34%)</td>
</tr>
<tr>
<td>Leaves</td>
<td>5,290,196</td>
<td>4,740,401 (90%)</td>
<td>2,772,753 (52%)</td>
</tr>
<tr>
<td>Nodule</td>
<td>6,335,851</td>
<td>5,265,596 (83%)</td>
<td>3,386,046 (53%)</td>
</tr>
<tr>
<td>Root</td>
<td>6,107,312</td>
<td>5,154,176 (84%)</td>
<td>3,740,765 (61%)</td>
</tr>
<tr>
<td>Root tip</td>
<td>5,154,499</td>
<td>4,212,283 (82%)</td>
<td>3,187,659 (62%)</td>
</tr>
</tbody>
</table>

Read length: 36  
Average read quality: 37-39  
Number of gene matches: 43,455-49,992 in JGI 8X Genome
Inter-Homoeolog Differential Expression

<table>
<thead>
<tr>
<th>Tissue</th>
<th>p=0.05</th>
<th>p=1e-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root</td>
<td>7,340</td>
<td>2,168</td>
</tr>
<tr>
<td>Root Tip</td>
<td>6,570</td>
<td>1,729</td>
</tr>
<tr>
<td>Nodule</td>
<td>6,675</td>
<td>2,031</td>
</tr>
<tr>
<td>Apical Meristem</td>
<td>7,462</td>
<td>2,175</td>
</tr>
<tr>
<td>Leaves</td>
<td>6,600</td>
<td>1,741</td>
</tr>
<tr>
<td>Flower</td>
<td>7,245</td>
<td>2,266</td>
</tr>
<tr>
<td>Green Pods</td>
<td>5,017</td>
<td>878</td>
</tr>
</tbody>
</table>

Of approximately 14,000 homoeolog pairs
No Significant Inter-Homoeolog Differences

**NUCLEAR TRANSCRIPTION FACTOR, X-BOX BINDING 1 (NFX1)**

- Unique Reads/Million

**CHAPERONE BINDING PROTEIN**

- Unique Reads/Million

**MULTI-COPPER OXIDASE**

- Unique Reads/Million
One Homoeolog Preferentially Expressed
Subfunctionalized Homoeologs?

### SERINE-THREONINE PROTEIN KINASE, PLANT-TYPE

- **Glyma13g42930.1**
- **Glyma15g02450.1**

### PHOSPHOGLYCERATE KINASE

- **Glyma08g17600.1**
- **Glyma15g41540.1**

### CYTOCHROME P450

- **Glyma08g10950.1**
- **Glyma11g37110.1**
- **Glyma11g37110.1**
Annual and perennial *Glycine* diverged around 5 million years ago

Recent (ca. 50,000 yr.) allopolydiploidy in subgenus *Glycine*

(From $2n = 38, 40$ to $2n = 78, 80$)

Jeff Doyle
Daniel C. Ilut
Thomas Owens

Perennial *Glycine* species occur throughout Australia

What physiological changes might have promoted polyploid diversification?
Photoprotection

Suite of mechanisms for dissipating excess light energy before it causes damage:

Primary pathways function in Photosystem II

Also called “nonphotochemical quenching” (NPQ)

Experimental Design

- **Species**
  - Diploids: D3 x D4
  - Tetraploid: T2

- **Daytime Light intensity:**
  - Limiting light: 125 µmol m⁻²s⁻¹
  - Excess light: 800 µmol m⁻²s⁻¹

- Whole transcriptome shotgun sequencing (mRNA Seq)
Solexa results

Summary Statistics:

<table>
<thead>
<tr>
<th>Sample</th>
<th>Channels</th>
<th>Reads</th>
<th>Reads Aligned</th>
<th>Percent Aligned</th>
<th>Reads Uniquely Aligned</th>
<th>Percent Uniquely Aligned</th>
<th>Genes sequenced</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3E</td>
<td>1</td>
<td>5065610</td>
<td>3957656</td>
<td>78.13%</td>
<td>2000660</td>
<td>39.49%</td>
<td>43520</td>
</tr>
<tr>
<td>D3L</td>
<td>1</td>
<td>5148542</td>
<td>3928079</td>
<td>76.29%</td>
<td>1642540</td>
<td>31.90%</td>
<td>41311</td>
</tr>
<tr>
<td>D4E</td>
<td>2</td>
<td>732016</td>
<td>402852</td>
<td>55.03%</td>
<td>258535</td>
<td>35.32%</td>
<td>31898</td>
</tr>
<tr>
<td>D4L</td>
<td>3</td>
<td>11544369</td>
<td>8517865</td>
<td>73.78%</td>
<td>4611424</td>
<td>39.95%</td>
<td>47357</td>
</tr>
<tr>
<td>T2E</td>
<td>2</td>
<td>8383794</td>
<td>6057267</td>
<td>72.25%</td>
<td>3947004</td>
<td>47.08%</td>
<td>48294</td>
</tr>
<tr>
<td>T2L</td>
<td>3</td>
<td>13707519</td>
<td>10345139</td>
<td>75.47%</td>
<td>6831894</td>
<td>49.84%</td>
<td>49291</td>
</tr>
</tbody>
</table>

- >50.8 million reads = >1.8 Gb
- 55-78% of reads aligned
- 32-50% uniquely aligned
- Soybean genome draft 7X: 58,556 genes
Detection of transcription over a wide range of expression levels

43,520 genes ranked by expression
Contributions of homoeologues to total expression in limiting light for 47 genes encoding subunits of photosystem II.
Reciprocal homoeologue expression biases within gene families

Are homoeologue expression biases simply carried over from diploids?
Homoeologue expression is not simply carried over from diploids

Diploid expression: LL

Homoeologue expression in T2: LL

Homoeologue expression is not simply carried over from diploids.
Does the overall pattern of gene expression in the tetraploid look more like one diploid progenitor than the other?

Yes, T2 is more like D3 under both high and low light conditions.

Heat map of pairwise Pearson correlation coefficients, using unique RPM and all genes.
Are similar suites of genes up- or down-regulated in diploids and the polyploid under excess light conditions?

What fraction of transcriptomes show >2x response on shift from low to excess light?
Over 20% of diploid transcriptomes are light responsive

D3
- 79% no change
- 16% 2-fold increase
- 5% 2-fold decrease
- <2-fold change

D4
- 76% no change
- 12% 2-fold increase
- 12% 2-fold decrease
- <2-fold change
Only 12% of the polyploid transcriptome is light responsive
Very different suites of genes are up-regulated under excess light in the two diploids.
Tetraploid has a unique response, distinct from either diploid.
Tetraploid response shares more genes with D3 than with D4

Tetraploid combines D3 and D4 light-responsive genes
What is the biology of the genes showing different expression patterns?

Fig. 2. (A) Genes up-regulated under excess light conditions in D3, D4, and T2, showing overlapping sets of genes. (B,C, and D) GO Biological Process (B), Cellular Component (C), and Molecular Function (D) category classifications for each of the gene sets from the Venn diagram (e.g., the 808 genes up-regulated only in T2 are broken down in the histograms labeled T2). The histograms represent about 25% of the genes from the Venn diagram; around 50% are unclassified, and an additional approximately 25% are classified in the two most abundant classes for each category, which show similar levels in all seven comparisons, and have been omitted to facilitate visualization of the remaining classes.
mRNA Sequencing Provides...

- Global gene expression profiling.
  - Expression level comparisons
  - Identification of structural variants (alternative splicing)
  - Gene copy (duplicate gene) - specific expression

- The ability to identify genetic variants.
  - Population level genetic diversity of complex genomes
  - Understand impact of genetic variants on gene expression