The Australian contribution to the R570 sugarcane genome sequence
Karen Aitken

Australian Sugar Industry

- 6000 cane growers (4,500 farming business)
- 35 million tonnes of sugarcane
- 4.75 million tonnes of raw sugar
- 80% exported
  - Major export customers include Japan, Korea, Malaysia, Taiwan, Saudi Arabia, New Zealand, Canada and USA
- $1.75 billion to the Australian economy

Sugarcane growing regions

Sugarcane genetics

- Cultivated sugarcane is a polyploid
- A hybrid between S. officinarum and S. spontaneum
- Both of which are polyploid but have different basic chromosome numbers
  - S. officinarum is octoploid, x = 10
  - S. spontaneum ploidy level varies from 4 to 12, x = 8, most common is octoploid

Sugarcane Genome

- Sugarcane is highly heterozygous and outcrosses
- Cultivars have from 90-115 chromosomes
- Autopolyploid/Allopolyploid
  - 80% from S. officinarum
  - 10-15% from S. spontaneum
  - 5-10% are recombinant chromosomes

Challenges of polyploidy

- Dramatic size increase
  - Much more data required
- Higher sequence similarity in functional regions
  - Hampers molecular approaches
  - Increased computational complexity of data analysis
- More functional complexity
  - Potential exponential increase in gene-gene interactions

10000 Mbp 1600 Mbp
**Generation of sugarcane genomic sequence - BAC sequencing**

- Generation of a reference sequence by sequencing the minimum tilling path of BAC clones aligned to the sorghum genome
- Selected BAC clones that were linked to traits of agronomic importance by screening the BAC library using both the qPCR 3d pooling strategy and the macroarray hybridisation method using markers with highly significant and robust associations to traits

**BAC sequencing and assembly**

- Sequenced 452 BAC clones using the Illumina HiSeq2000 platform
  - Generated 500 to 10,000 fold coverage of each BAC
  - Assembled reads using a custom pipeline developed in collaboration with the University of Queensland
- Sequenced and assembled an additional 696 BAC clones through KeyGene using the Pacific Biosciences long-read platform
  - This platform has been shown to reduce the computational cost of the assembly process

**Number of contiguous sequences assembled per BAC for Illumina and Key Gene sequence methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Sequences Assembled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina HiSeq</td>
<td>68%</td>
</tr>
<tr>
<td>Pacific Bioscience long-read</td>
<td>90%</td>
</tr>
</tbody>
</table>

**BAC sequence**

- In total we have assembled 987 BAC clones
- Total assembled sequence length is 121 Mbp
- Approximately 17% of the monoploid sugarcane genome
- Aligned to sorghum

**Whole genome shotgun sequencing strategy**

- Parallel approach to BAC by BAC sequencing method using same variety R570
- Will be combined with BAC data in sequencing consortium
- Short-read Illumina sequences from a range of overlapping fragments
- Advanced computational methods to assemble sequence into scaffolds

**Whole genome shotgun assembly**

- Illumina short reads were generated as paired-end data spanning 180bp to 600bp
- Enabled construction of high quality localised contiguous sequences
- Illumina long mate-pair libraries 2kbp to 32kbp
- Scaffolding of these localised sequences into longer fragments
- Ability to bridge repetitive DNA sequence
- Pacific Bioscience long read technology enables scaffolding to occur with reads up to 20kbp
  - Has the capacity to close gaps of scaffold assemblies
- BioNano provides physical mapping over very large distances up to 150kbp
  - Enables long distance scaffolding in sequence approaching chromosome pseudo-molecule status
Whole genome shotgun data sets

<table>
<thead>
<tr>
<th>LIBRARY</th>
<th>CURRENT SEQUENCE COVERAGE (x)</th>
<th>CURRENT PHYSICAL COVERAGE (x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina 180bp</td>
<td>12.14</td>
<td>10.93</td>
</tr>
<tr>
<td>Illumina 250bp</td>
<td>3.74</td>
<td>3.52</td>
</tr>
<tr>
<td>Illumina 300bp</td>
<td>11.08</td>
<td>15.59</td>
</tr>
<tr>
<td>Illumina 600bp</td>
<td>8.70</td>
<td>26.10</td>
</tr>
<tr>
<td>Illumina 2,000bp</td>
<td>2.81</td>
<td>31.44</td>
</tr>
<tr>
<td>Illumina 2,500bp</td>
<td>18.65</td>
<td>41.94</td>
</tr>
<tr>
<td>Illumina 4,500bp</td>
<td>18.66</td>
<td>76.25</td>
</tr>
<tr>
<td>Illumina 5,000bp</td>
<td>2.94</td>
<td>78.89</td>
</tr>
<tr>
<td>Illumina 7,500bp</td>
<td>1.21</td>
<td>45.38</td>
</tr>
<tr>
<td>Illumina 12,000bp</td>
<td>0.37</td>
<td>58.61</td>
</tr>
<tr>
<td>PacBio 20,000bp</td>
<td>3.20</td>
<td>320.39</td>
</tr>
</tbody>
</table>

Assembly of whole genome shotgun data

- 4 High-Performance Computing facilities tested (up to 6 Tb)
- 3 Assembly algorithms compared
  - Velvet (public), AllPaths-LG (public), Biokanga (CSIRO-developed)
- Cumulative >6000 days of compute time & max. 1.5 Tb RAM used

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number assembled scaffolds</td>
<td>367094</td>
</tr>
<tr>
<td>Total assembly size (bp)</td>
<td>7398491276</td>
</tr>
<tr>
<td>Total assembly N50 (bp)</td>
<td>3489</td>
</tr>
<tr>
<td>Overall longest contig (bp)</td>
<td>71417</td>
</tr>
<tr>
<td>Number contigs &gt; 1 kbp</td>
<td>19461.395 (48%)</td>
</tr>
<tr>
<td>Number contigs &gt; 10 kbp</td>
<td>30502</td>
</tr>
</tbody>
</table>

The first draft assembly represents ~74% of the complete polyploid R570 genome sequence.

Assembly validation... (?)

CEGMA (Core Eukaryotic Genes Mapping Approach)

- Partial hits to 224 proteins (90% complete), 2.8 copies on average
- Complete hits 176 proteins (69% complete), 3.9 copies on average

Comparisons (BLASTN, evalue 1e-20)
- Saccharum Gene Index
  - 116,078 (95.7%)
- Sugarcane Transcribed Unigenes (Cardosa-Silva, 2014)
  - 72,269 (97.1%)

Comparison to Sorghum transcripts

Biokanga BLITZ
- BLAT-like alignment, developed for transcript/nucleome alignment
- Extracts intron-exon structures

Sorghum genome v2.1 annotated CDS
- Out of 33,032 transcripts, 29,249 (88.5%) were represented in assembly
- Non-overlapping alignments totaled ~152 Mbp
- ~400% of Sorghum transcriptome size

Can we observe allelic diversity in this assembly?

Distribution of allelic diversity across the sorghum genome

- All genes that were longer than 300bp across the genome
- At least 5 alleles
- Analyzed to determine whether the gene copies originated from *S. officinarum* or *S. spontaneum*
Comparison to Sorghum transcripts

The results indicate that it could be possible to determine the contributions of alleles from parental species.

Gbrowse

Summary

- The present assembly includes the PacBio long read data but was of limited help due to the large number of short contigs and limited long read sequence.
- To help improve the assembly a physical map of R570 has been generated using BioNano Genomics optical mapping technologies.
- The R570 mitochondrial genome sequence has been assembled.
- 2 large contigs.
- The current preliminary assembly of R570 has been used to extract genomic context information for sucrose accumulation genes and flowering genes. We see this as a major resource for further sugarcane research.
- The current assembly has also been used to validate all alleles of genes in the sequenced BAC clones.
- A JGI CSP project ‘Understanding polyploidy through the generation of the first sugarcane genome sequence’ has started this year to improve the current assembly.

Acknowledgements

Team members
- Paul Berkman
- Jiri Stiller
- Jai Perroux
- Jingchuan Li
- Rosanne Casu
- Anne Rae