Sequencing the hexaploid wheat genome in 42 simple steps

David Edwards
University of Queensland
Dave.Edwards@uq.edu.au
Outline

• Wheat genome overview

• Summary of work
  • 7DS
  • 7BS/4AL translocation
  • Current group 7 assemblies

• Future work
Wheat genomics

- AABBDD

- Highly repetitive >80%
  - Multiple genomes
  - Transposon activity


Wheat genomics

Wheat–rice genome relationships.

Chromosome arm sequencing

• DNA from Jaroslav Dolezel (Czech Rep.)
  • Flow sorted chromosomes
  • Cytogenetic stock with arm deletion

• Benefits
  • Better resolution of smaller “genome size”
  • Reduces repetitive sequence
  • Simpler assembly
  • Sequencing 42 rice genomes
Second-generation sequencing (2GS)

- Illumina GAIIx and HiSeq2000
  - ↑↑↑↑ sequence
  - ↓ read-length
  - ↓ money
  - ↓ time
  - ↑ computation
Illumina paired reads

Insert size

- Illumina GAIIx and HiSeq
- Read length (35 bp – 150 bp)
  - ~ Normal distribution
  - Standard deviation ~ 10% mean
Mapping 7DS reads to reference genomes

Brachypodium

rice
Mapping reads to reference genomes
7DS assembly

- Velvet assembly from 17.6x coverage
  - Total assembly 153,653,984 bp (40% of 7DS)
  - Remainder of arm present as collapsed repeats
  - Longest contig 32,648 bp

- Syntenic build
  - Total length 7,814,423 bp
  - 1,735 genes
  - 1,072 syntenic to *B. distachyon*
Have we assembled all the 7DS genes?

- Compare sequences of cDNAs which had been bin mapped to the 7DS assembly
  - 315 of the 354 cDNAs (88.5%) are represented in the assembly
Have we assembled all the 7DS genes?

• Compare sequences of cDNAs which had been bin mapped to the 7DS assembly
  – 315 of the 354 cDNAs (88.5%) are represented in the assembly

• However:
  – None of the missing 39 cDNAs match the syntenic region of Brachypodium
Comparison with genetic map

- 65 Aegilops tauschii 7S markers (Luo et al., 2009)
- 60 matched our assembly
- 5 had no significant sequence identity
Comparison with genetic map

- 65 *Aegilops tauschii* 7S markers (Luo et al., 2009)
- 60 matched our assembly
  - these are all SNP markers
- 5 had no significant sequence identity
  - these are all RFLPs
7DS syntenic build

www.wheatgenome.info

7DS syntenic build

- Missing region?
- Corresponds to 7BS/4AL translocation
7BS/4AL translocation

- Velvet assembly from 21.0x coverage
  - Total assembly 176,154,889 bp (49% of 7BS)
  - Longest contig 29,196 bp

- Syntenic build
  - Total length 6,508,016 bp
  - 1,632 genes included
  - 967 syntenic to *B. distachyon*
7BS/4AL translocation

7DS and 7BL sequence similarity with Brachypodium
7BS/4AL translocation

- Bradi1g49550 to Bradi1g52510 missing in 7BS assembly
- Bradi1g49470 to Bradi1g52330 found in 4AL 454 data
- Translocation between Bradi1g49500 and Bradi1g49550
  - Intervening 4 genes missing from all assemblies
- ~13% genes moved from 7BS to 4AL
  - 13 genes moved from 4AL to 7BS
7BS/4AL translocation

Genes in SynBuilds

Genes NOT in SynBuilds

7DS assembly (purple)
7BS assembly (grey)
7BS/4AL translocation

- Genes conserved from *B. distachyon* syntenic region:
  - 38.6% in 7DS
  - 39.1% in 7BS

- Genes conserved in homoeolog syntenic regions:
  - 84.6% between 7BS/7DS
  - 83.7% between 4AL/7DS

- Conservation consistent between homoeologs and *B. distachyon*

- Gene count extrapolates ~77,000 genes in wheat
  - 45,000 – 50,000 in syntenic blocks to *Brachypodium*
### Group 7 chromosomes

- **Current assembly status**

<table>
<thead>
<tr>
<th>Chromosome Arm</th>
<th>Data (Gbp)</th>
<th>Coverage</th>
<th>Current assembly (bp)</th>
<th>Longest contig (bp)</th>
<th>N50 (bp)</th>
<th>SynBuild Version</th>
<th>SynBuild Length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7DS</td>
<td>27.7</td>
<td>72.6x</td>
<td>153,680,095</td>
<td>37,458</td>
<td>1,159</td>
<td>1</td>
<td>7,814,423</td>
</tr>
<tr>
<td>7DL</td>
<td>26.5</td>
<td>76.7x</td>
<td>101,872,746</td>
<td>34,934</td>
<td>1,309</td>
<td>0.1</td>
<td>7,124,835</td>
</tr>
<tr>
<td>7AS</td>
<td>17.7</td>
<td>35.7x</td>
<td>148,258,996</td>
<td>11,044</td>
<td>148</td>
<td>0.1</td>
<td>2,735,305</td>
</tr>
<tr>
<td>7AL</td>
<td>20.6</td>
<td>50.5x</td>
<td>168,471,449</td>
<td>25,876</td>
<td>437</td>
<td>0.1</td>
<td>4,084,856</td>
</tr>
<tr>
<td>7BS</td>
<td>23.2</td>
<td>54.7x</td>
<td>176,154,889</td>
<td>29,196</td>
<td>472</td>
<td>1</td>
<td>6,508,016</td>
</tr>
<tr>
<td>7BL</td>
<td>15.1</td>
<td>28.0x</td>
<td>163,374,698</td>
<td>12,928</td>
<td>351</td>
<td>0.1</td>
<td>2,237,378</td>
</tr>
</tbody>
</table>

http://wheatgenome.info
Group 7 chromosomes

Welcome to WheatGenome.info

Wheat is probably the most important crop in the world, yet it has one of the most challenging genomes. Bread wheat is a hexaploid, with three complete genomes termed A, B and D in the nucleus of each cell. Each of these genomes is almost twice the human genome and consists of around 5,000 million letters. Several groups around the world are working towards sequencing wheat. Details of individual efforts can be found on the wiki below.

Genome sequencing projects can be generally divided into whole genome shotgun (WGS) methods or BAC by BAC methods.

WGS attempts to sequence the genome in one go, by generating a large amount of sequence data and then assembling this to produce a representation of the string of letters which make up the genome. WGS has the benefit in that it is quick and relatively inexpensive, but it is often confounded by the inability to stitch the individual sequence reads together, resulting in a poor quality assembly. This is particularly problematic for polyploids, where more than one genome is present in each cell, or where there is a substantial quantity of repetitive sequences. Wheat is a polyploid with 3 genomes, each of which is 80% repetitive, making WGS unattractive.

The alternative BAC by BAC approach requires breaking the genome down to relatively small pieces (c. 120 kbp), ordering these as a minimal tilling path, then sequencing each of the BACs in the tilling path. While sequence assembly or repetitive regions remains problematic, this approach offers the potential to produce the best quality finished genome. However, BAC by BAC sequencing of wheat is hugely expensive, time consuming and is still not guaranteed to produce a complete genome due to some regions being underrepresented in BAC libraries.

Individual chromosomes arms are being assembled and annotated. They are made publically available as they are produced using the genome viewer GBrowse2.

Links to available chromosome arms are below.

All chromosome arm specific sequence data is available at TARGdb.

All raw and assembled sequence data is freely available on request.

• Draft Wheat 7AS
• Draft Wheat 7AL
• Draft Wheat 7DL
• Draft Wheat 7BS
• Draft Wheat 7RL
• Draft Wheat 7DL
• Draft Wheat 7BS
• Wheat 7AS
• Wheat 7DL in preparation

Wheat genome assembly BLAST portal www.wheat genome.info/gbrowseblast

There is a large and increasing range of wheat genome resources available on the web. If you cannot find what you are looking for here, you may want to try some of the sites below.
Future work

• Complete syntenic builds for group 7 arms
  • New assemblies with all data

• Homeolog analysis
  • Identify genes preferentially lost/retained
  • Extract gene function/ontology
  • Investigate contributing factors to gene movement/loss

• Align gene expression data
• Distinguish homoeologous/varietal SNPs
• 3rd generation sequencing
Summary

• Shotgun assembly of 7A, 7B and 7D
• Model for identification of all wheat genes
• Framework for complete genome sequencing
• ~13% of 7BS genes translocated to 4AL
• Gene movement is consistent between arms
• We estimate ~77,000 genes in wheat
• Full comparison of homoeologs underway...
Acknowledgements

Paul Berkman
Jiri Stiller
Sahana Manoli
Emma Campbell
Michał Lorenc
Kaitao Lai
Michael Imelfort
Chris Duran
Terry Clark
Edmund Ling
Kenneth Chan
Hong Ching Lee

Adam Skarszewski
Daniel Marshall
Megan McKenzie
Lars Smits
Jacqueline Batley
Delphine Fleury
Bao-Lam Huynh
Jaroslav Doležel
Marie Kubaláková
Hana Šimková
Pilar Hernandez

Contact:
Dave.Edwards@uq.edu.au