Sequence-based assembly of chromosome 7A and comparison to diploid progenitors

Gabriel Keeble-Gagnere¹
J Nystrom-Persson¹, C Cavanagh², D Fleury³, H Webster¹, R Appels¹

¹ Veterinary and Life Sciences, Murdoch University, South Street, Perth, WA 6150
² CSIRO Plant Industry, Clunies Ross Street, Black Mountain ACT 2601
³ Australian Centre for Plant Functional Genomics, Hartley Grove, Urrbrae, Glen Osmond SA 5064
Talk outline

- Assembly of chromosome 7A
- Assembly results and next steps
- Post-genomics on chromosome 7A
Reference-level assembly of 7A

Mingcheng Luo, UC Davis
Reference-level assembly of 7A

BAC library fingerprinted

Physical assembly with LTC

Mingcheng Luo, UC Davis

Zeev Frenkel, Korol lab, Haifa University
Reference-level assembly of 7A

- **BAC library fingerprinted**
  - Mingcheng Luo, UC Davis

- **Physical assembly with LTC**
  - Zeev Frenkel, Korol lab, Haifa University

- MTP defines pools of BACs to sequence
Reference-level assembly of 7A

BAC library fingerprinted

Physical assembly with LTC

MTP defines pools of BACs to sequence

Illumina Hiseq sequencing

Mingcheng Luo, UC Davis

Zeev Frenkel, Korol lab, Haifa University

150bp reads, ~350bp paired-end library
Reference-level assembly of 7A

1. BAC library fingerprinted
   - Mingcheng Luo, UC Davis
2. Physical assembly with LTC
   - Zeev Frenkel, Korol lab, Haifa University
3. MTP defines pools of BACs to sequence
4. Illumina Hiseq sequencing
   - 150bp reads, ~350bp paired-end library
5. Assembly
   - Abyss
Reference-level assembly of 7A

- **BAC library fingerprinted**
  - Mingcheng Luo, UC Davis

- **Physical assembly with LTC**
  - Zeev Frenkel, Korol lab, Haifa University

- **MTP defines pools of BACs to sequence**

- **Illumina Hiseq sequencing**
  - 150bp reads, ~350bp paired-end library

- **Anchoring to genetic map**

- **Assembly**
  - Abyss
Reference-level assembly of 7A

BAC library fingerprinted

Physical assembly with LTC

MTP defines pools of BACs to sequence

Illumina Hiseq sequencing

150bp reads, ~350bp paired-end library

Assembly

Anchoring to genetic map

Integration of genetic and physical map

Mingcheng Luo, UC Davis

Zeev Frenkel, Korol lab, Haifa University

Abyss
Reference-level assembly of 7A

BAC library fingerprinted

Physical assembly with LTC

MTP defines pools of BACs to sequence

Illumina Hiseq sequencing

Assembly

Anchoring to genetic map

Integration of genetic and physical map

Mingcheng Luo, UC Davis

Zeev Frenkel, Korol lab, Haifa University

150bp reads, ~350bp paired-end library

Abyss
## Physical map assembly

<table>
<thead>
<tr>
<th>Arm</th>
<th># contigs &gt; 5 clones</th>
<th>Contig N50</th>
<th>Contig L50</th>
<th>Clones in MTP</th>
<th>Estimated total length</th>
</tr>
</thead>
<tbody>
<tr>
<td>7AS</td>
<td>380</td>
<td>1.38Mb</td>
<td>81/299</td>
<td>5,280</td>
<td>353Mb</td>
</tr>
<tr>
<td>7AL</td>
<td>352</td>
<td>1.70Mb</td>
<td>64/288</td>
<td>5,832</td>
<td>402Mb</td>
</tr>
</tbody>
</table>

- 11,012 BACs in MTP
- 732 physical contigs (BAC pools) to sequence
  - Barcoded 96 pools per lane
LTC vs. FPC

A single, large physical contig from FPC assembly revealed to be 4 separate contigs joined by single clones (likely contaminated wells)
Chromosome 7A sequence assembly summary

<table>
<thead>
<tr>
<th>Arm</th>
<th># Scaffolds (bp)</th>
<th>Mean (bp)</th>
<th>N50 (bp)</th>
<th>Max scaffold (bp)</th>
<th>Total assembly length (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7AS</td>
<td>33,541</td>
<td>11,704</td>
<td>26,896</td>
<td>264,183</td>
<td>392</td>
</tr>
<tr>
<td>7AL</td>
<td>38,731</td>
<td>11,674</td>
<td>27,032</td>
<td>274,707</td>
<td>452</td>
</tr>
<tr>
<td>Both</td>
<td>72,272</td>
<td>11,688</td>
<td>26,953</td>
<td>274,707</td>
<td>844</td>
</tr>
</tbody>
</table>

- All 732 BAC pools (physical contigs) have now been sequenced
  - One Illumina run only pooled half the BACs for each pool - has been re-sequenced; analysis is underway
MAGIC

Multi-parent Advanced Generation Inter-Cross, designed to get around limitations of double-haploid populations for mapping traits.

Applied to crops for the first time by Colin Cavanagh at CSIRO.

- 8-way cross
  - Baxter, Yitpi and Westonia (Australia)
  - AC Barrie (Canada), Alsen (US), Pastor (Mexico), Xiaoyan 54 (China), Volcani (Israel)
- 5,000 lines
- GBS sequencing of 4,800 markers in population in 980 lines (Matt Hayden, Victoria DEPI)
7A chromosome structure based on MAGIC

Approx. 75% of physical contigs included in preliminary ordering by MAGIC
Approx. 75% of physical contigs included in preliminary ordering by MAGIC

GWAS data (Xia Xianchun, CAAS, unpublished) indicates major grain hardness association on 7AL
Gene density higher at telomeric ends
7A chromosome structure based on MAGIC

Gene density lower in centromere and high occurrence of characteristic repeats.
CS x Renan population

The CS x Renan population (developed at INRA, Clermont-Ferrand) has 276 lines mapped with over 5,000 markers on chromosome 7A:

- This is being used to build a pseudo-molecule for 7A to use as a reference because the CS x Renan population was also used to build a reference pseudo-molecule for chromosome 3B (Choulet et al, Science, under submission)
- The 330 CS x Renan bins from this 7A pseudo-molecule are currently the targets for refinement using the MAGIC 7A mapping data
- The BAC pools (= MTP contigs) in the CS x Renan bins are also being re-examined in light of the sequence data using LTC (Zeev Frenkel, Korol lab, Haifa University)
Gene annotation of genome assembly to date

A collaborative effort to annotate the 7A genome assembly has been established:

- Philippe Leroy and INRA group (TriAnnot) at Clermont-Ferrand (France)
- Francisco Camara group (geneID), CRG (Spain)
- Angela Juhasz (Martonvásár, Hungary)
- Colleagues in Adelaide (ACPFG, Delphine Fleury, Diane Mather, Ute Baumann), Canberra (CSIRO, Jen Taylor) and Perth (Murdoch University, Michael Francki, Shahid Islam)
Using *T. urartu* sequence to guide scaffolding

Whole-genome shotgun assembly of *T. urartu* used 8 library sizes (200bp-20kb) for scaffolding (Ling et al, 2013).

We found we can use the scaffolds from *T. urartu* to order our scaffolds from 7A:

- First, align our scaffolds to *T. urartu* scaffolds
- Use alignment to determine potential ordering of our own scaffolds
- Check our own mate-pair data for evidence supporting the join
Using *T. urartu* sequence to guide scaffolding

<table>
<thead>
<tr>
<th>Query</th>
<th>QS</th>
<th>QE</th>
<th>SS</th>
<th>SE</th>
<th>Comm</th>
</tr>
</thead>
<tbody>
<tr>
<td>7AS-11582.draft080314.scaffold0227</td>
<td>1282</td>
<td>3591</td>
<td>3052</td>
<td>5360</td>
<td>MPS: 2</td>
</tr>
<tr>
<td>7AS-11582.draft080314.scaffold0062</td>
<td>430</td>
<td>3361</td>
<td>6270</td>
<td>9575</td>
<td>MPS: 2</td>
</tr>
<tr>
<td>7AS-11582.draft080314.scaffold0019</td>
<td>1286</td>
<td>10152</td>
<td>22464</td>
<td>MPS: 3</td>
<td></td>
</tr>
<tr>
<td>7AS-11582.draft080314.scaffold0012</td>
<td>15479</td>
<td>22854</td>
<td>36495</td>
<td>MPS: 6</td>
<td></td>
</tr>
<tr>
<td>7AS-11582.draft080314.scaffold0248</td>
<td>10364</td>
<td>37602</td>
<td>48668</td>
<td>MPS: 0</td>
<td></td>
</tr>
</tbody>
</table>

*T. urartu* scaffold

Wheat 7A scaffolds
Using *T. urartu* sequence to guide scaffolding

Mate-pairs from wheat required confirmed join
Post-genomics on chromosome 7A

From genome to metabolome:

- **Transcriptome (RNA-seq)**
  - Hollie Webster PhD. thesis, ACCWI
- **Proteome (8-plex iTRAQ)**
  - Shahid Islam, ACCWI
- **Metabolomics**
  - Camilla Hill, Ute Roessner (Melbourne University)
RNA-seq study (Hollie Webster)

Investigating the effects of drought on the developing head, in two double haploid lines, D02-105 (drought intolerant) and D08-299 (drought tolerant).

- Two years: 2011, 2012
- Two varieties: D02-105, D08-299 (plus Westonia, Kauz in 2011)
- Four timepoints: AR05, AR10, AR15, FHE
- Two experimental factors: Control, Drought
- Between 3 and 6 biological replicates per sample, and up to 3 technical replicates per biological replicate
  - After filtering, around 200 technical replicates in total (for 2012 experiment)
A total of 230 anther-specific rice genes identified by Deveshwar et al. (2011) could be identified in our RNA-Seq data from the developing spike of wheat.

24 of these genes were differentially expressed in response to water deficit early in spike development.

4 of these 24 genes were located in a small region on 5BL that also defines a QTL for the timing for the start of head development.

These genes are currently under further investigation.
iTRAQ proteomics (Shahid Islam)

18, 8-plex, iTRAQ experiments which replicates the experimental design of Hollie Webster’s RNA-seq study, plus two extra time points (7- and 30-days post-anthesis).

- Largest shotgun proteomics study in wheat
- Approximately 1,500 peptides per experiment
Metabolomics work carried out by Camilla Hill in Ute Roessner lab, Melbourne University

GC-MS study of drought stress in Excalibur/Kukri double haploid population published in 2013 showed highest number of metabolic QTLs (mQTLs) mapped to chromosome 7A

LC-MS study manuscript in preparation
- High number of mQTLs on chromosome 7A
- First mapping of mQTL to genomic sequence in wheat
**Friedrich**

Friedrich is a framework for bioinformatics application development in Scala. It is especially well suited for heavy data processing in a flexible, experimental setting. A basic genome assembler, the first application built on Friedrich, is included.

A paper was presented at PRIB 2012:

- An Open Framework for Extensible Multi-Stage Bioinformatics Software

Developed in collaboration between:

- Australia-China Centre for Wheat Improvement, Murdoch University, Perth, Australia (ACCWI)
- National Institute for Biomedical Innovation, Osaka, Japan (NIBIO)

(Previously: Centre for Comparative Genomics, Murdoch University and NIBIO)

[https://bitbucket.org/jtnystrom/friedrich/](https://bitbucket.org/jtnystrom/friedrich/)

Open source, under intensive development...
Thanks

Funding:
- Bioplatforms Australia (BPA)
- Grain Research Development Corporation (GRDC)

Collaborators:
- Australian Genome Research Facility (AGRF)
- Department of Environment and Primary Industries, Melbourne (DEPI)
  - Matt Hayden, Josquin Tibbits
- CSIRO
- INRA

ACCWI software developer: Johan Nystrom-Persson