Physical map of the wheat chromosome arm 3DS

Jan Bartoš

Centre of Region Haná for Biotechnological and Agricultural Research
Institute of Experimental Botany
Šlechtitelů 31
783 71 Olomouc - Holice
Wheat chromosome arm 3DS

Chromosome arm 3DS characteristics

- Estimated size 321 Mbp
- Less than 2% of wheat genome
- Low level of polymorphism in D genome
- Important genes
  - Ph2 locus (pairing homologs)
  - Yr49 (yellow rust resistance)
3DS physical map

BAC library and fingerprinting
• 36,864 clones
• 11x chromosome coverage
• 27,880 useful fingerprints

Automated assembly
• FPC based
• Following IWGSC rules
• Cut-off: 1e-75 => 1e-45

Automated assembly
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cut-off</td>
<td>1e-45</td>
</tr>
<tr>
<td>Contigs</td>
<td>1,360</td>
</tr>
<tr>
<td>Q-clones</td>
<td>282</td>
</tr>
<tr>
<td>Assembly length (Mb)</td>
<td>310 (97%)</td>
</tr>
<tr>
<td>Longest contig (kb)</td>
<td>1,092</td>
</tr>
<tr>
<td>N50 contig length (kb)</td>
<td>244</td>
</tr>
<tr>
<td>MTP (clones)</td>
<td>3,823</td>
</tr>
</tbody>
</table>

http://olomouc.ueb.cas.cz/
### 3DS physical map

**Manual assembly**
- FPC based
- Cut-off: $1e^{-45}$ => $1e^{-15}$
- Correction using LTC

**Automated assembly**
- Cutoff: $1e^{-45}$
- Contigs: 1,360
- Q-clones: 282
- Assembly length (Mb): 310 (97%)
- Longest contig (kb): 1,092
- N50 contig length (kb): 244
- MTP (clones): 3,823

**Manual assembly**
- Cutoff: $1e^{-15}$
- Contigs: 918
- Q-clones: 499
- Assembly length (Mb): 278 (87%)
- Longest contig (kb): 1,870
- N50 contig length (kb): 412
- MTP (clones): ---
**In silico anchoring workflow**

1. MTP definition (FPC)
2. Markers e.g. IWGSC survey sequence
3. MTP 3-D pool sequencing
4. Read mapping to marker sequences (bwa)
5. Resolving unique reads
6. Positive pool detection (for each seq)
7. BAC address deconvolution
8. Quality control

http://olomouc.ueb.cas.cz/
MTP pool sequencing

- MTP 3,823 clones
- Fifty 3D MTP pools (10 plates, 16 rows, 24 columns)
- Pools of each dimensions sequenced as indexed libraries on Illumina HiSEQ
- 367,907,030 reads (2 x 100 bp)
- Unequal pool coverage
- 6 – 166x (mean 35x; median 23.5x)

http://olomouc.ueb.cas.cz/
Read mapping to marker sequences

Reference sequence
- IWGSC 3DS survey sequence
- 314,944 sequences
- Total length 145,374,274 bp (45% of chromosome arm)

Read alignment
- Using Burrows-Wheeler aligner
- Reads of each pool renamed to track their origin
- Maximal coverage 30x/pool
Positive pool identification

• Only reads mapped to unique position with no mismatch used

• **Positive pools identified individually for each sequence**

• Aligned reads counted for each pool

• Number of aligned reads normalized by pool coverage

• Pool positive if normalized read number \( \geq 20\% \) of average for pools with at least one aligned read

• At least 1 plate, 1 row and 1 column pool for 258,146 seqs

http://olomouc.ueb.cas.cz/
BAC address deconvolution

1) One positive pool in each dimension (1 – 1 – 1)
   • Direct BAC clone identification
     Plate07 – RowC – Column18 --> TaaCsp3DShA_0055B07

2) Multiple positive pools in at least one dimension (e.g. 2 – 2 – 2)
   • Identification of all candidate BAC clones
     a) Check contig information for all clones
     b) Check possible overlap in case of end clones

3) Sequence not anchored if:
   a) Positive pool is missing for plate, row or column.
   b) Five or more positive pools in at least one dimension.
   c) No positive clone was found in step 2).

http://olomouc.ueb.cas.cz/
Anchoring results

- Anchored 184,880 sequences
  - 58.7% survey sequences
- 96,784,747 bp anchored
  - 66.6% of survey sequence length
  - 30.2% estimated arm length
- 878 contigs with at least one sequence
- 1 – 2,514 sequences per contig

Anchored sequences

- type_1: 3%
- type_2a: 63%
- type_2b: 34%
Analysis of anchored sequences

184,880 anchored sequences

**DArT**
- 194 DArTs identified in 182 sequences
- 125 contigs anchored by DArT markers (1 – 6 markers/contig)

**Gene fragments**
- 1,906 gene models/fragments identified in 3DS survey sequences
- 1,408 (73.9%) genes/fragment anchored (by 1,372 sequences)
- 377 contigs contain at least one gene (1 – 24/contig)
- 793 organized using GenomeZipper approach
- 291 contigs anchored by GenomeZipper (1-17 gene fragments/contig)

319 contigs anchored - 53.4% of physical map length
Analysis of anchored sequences

Repeat junctions
- IsbpFinder used to identify repeat junctions (potential ISBP markers)
- 24,517 TE insertions with preserved ends were found in 3DS survey sequences
- 17,684 (72.1%) ISBPs anchored to contigs (in 13,870 sequences)
- Up to 232 ISBPs in one contig
- 652 contigs (85.6% of physical map length) have at least one insertion site
Quality control

**DArT**
- 40 contigs with more than one DArT
- 74% same or close position on DArT map

**GenomeZipper**
- 192 contigs with more than one gene fragment
- 70% neighbour positions on GenomeZipper

**Hmmm… Anchoring error rate is overestimated**

Additional sources of error
- BAC contig miss-assembly
- Genetic mapping of DArT markers
- Incorrect position of gene fragment in GenomeZipper

http://olomouc.ueb.cas.cz/
Contig miss-assembly is significant source of error
**Physical localization of gene fragments at identical GenomeZipper position**

- **178 GenomeZipper positions with multiple gene fragments**
- For 161 (90.5%) fragments have identical position
Additional assembly improvement

6,362 sequences of anchoring type 2b) could be used to merge contigs

- Sequences anchored to clones in different contigs
- Match of the clones at e-10
Conclusion

• We developed protocol for high-throughput contig anchoring
• 66% of survey sequence (97 Mbp) anchored to physical map
• 74% genes identified in survey sequences localized in BAC clones
• 53% of the physical map organized through anchoring to DArT genetic map and 3DS GenomeZipper

Future perspective

• Additional validation of results (including wet lab)
• Cleaning and integration of ISBP markers, polymorphism identification within CS x Renan population
• Sequencing of 3,823 clones of MTP

http://olomouc.ueb.cas.cz/
Acknowledgement

Jaroslav Doležel
Kateřina Cvikova
Jan Šafář
Hana Šimková

Federica Cattonaro

Michael Alaux

Andrzej Killian