The international effort to sequence the 17Gb wheat genome:

Yes, Wheat can!

Catherine Feuillet

INRA Clermont-Ferrand, France

Genetics, Diversity & Ecophysiology of Cereals
The International Wheat Genome Sequencing Consortium

Launched in 2005 on the initiative of Kansas Growers

IWGSC Sponsors

Co-Chairs
- R. Appels
  - Australia
- J. Dvorak
  - USA
- C. Feuillet
  - France
- B. Gill
  - USA
- B. Keller
  - Switzerland
- Y. Ogihara
  - Japan

Executive Director:
K. Eversole

21 Sponsors

Funding or Scientific Contributors

Coordinating Committee

General Members

64 members, 20 countries

> 228 members
28 countries

www.wheatgenome.org
An integrated and ordered wheat genome sequence

- Access to regulatory sequences and all kind of structural polymorphisms
- Perfect markers (causal) for precision breeding
- Genetic resources mining
- GM wheat
The Breadwheat genome is……

1. Big: 17Gb (5 x human genome, 40 x rice…)

2. Polyploid: 2n= 42 = 6x

3. Full of TEs (>90%)
Dissection of the genome to single chromosomes (arms) representing individual (sub)genomes

- **Triticum aestivum**
  - $(2n = 6x = 42)$
  - $1C \sim 17,000$ Mbp

- Sheath fluid
- Laser
- Excitation light
- Flow chamber
- Deflection plates
- Scattered light

Doležel et al., Chromosome Res. 15: 51, 2007

- Chromosomes: $605 - 995$ Mbp
  - (3.6 – 5.9% of the genome)
- Chromosome arms: $225 - 585$ Mbp
  - (1.3 – 3.4% of the genome)

- Chromosome specific BAC libraries (June 2012)
- Amplified DNA for chromosome survey (Nov 2011)
Combined strategies to establish a wheat reference genome sequence

Physical mapping of individual chromosomes

Survey sequencing of individual chromosomes
- Gene catalog
- Virtual order
- Markers
- Short term

MTP sequencing
- Anchored and ordered sequence
- Intergenic regions
- Markers
- Long term

A reference sequence anchored to the genetic and phenotypic maps
An international effort

1A  2A  3A  4A  5A  6A  7A

1B  2B  3B  4B  5B  6B  7B

1D  2D  3D  4D  5D  6D  7D

T. aestivum

cv Chinese Spring
Physical map of the 1GB chromosome 3B

- 131,792 fingerprinted BACs by SNaPshot
- 1,283 contigs (average size = 749 kb) with FPC
- 961 Mb coverage (97% chromosome)
- 4,367 molecular markers (SSRs, ISBPs, unigenes…)
- 919 anchored contigs (740 Mb) in 16 deletion bins
- 19.2 X coverage
- MTP (8,448 clones)

A Physical Map of the 1-Gigabase Bread Wheat Chromosome 3B

Paux et al, Science 2008; Rustenholz et al, Plant Physiol 2011

Improving physical maps

✓ Pilot sequencing showed that:

10% of BACs in contigs are mis-assembled, some chimerical contigs (0.6/10Mb) and an average contig size smaller than in small genomes

Whole Genome Profiling

Sequence-based physical mapping of complex genomes by whole genome profiling

Jan van Oeveren,1 Mario de Ruiter,1 Taco Jesse,1 Hein van der Poel,1 Jifeng Tang,1 Feyruz Yalcin,1 Antoine Janssen,1 Hanne Volpin,1 Keith E. Stormo,2 Robert Bogden,2 Michiel J.T. van Eijk,1 and Marcel Prins1,3

1Keygene N.V., Wageningen, The Netherlands; 2Ambicon Express Inc., Pullman, Washington 99163, USA

Linear Topography Contig

LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes

Zeev Frenkel,1 Etienne Pauw,1 David Mester1, Catherine Feillet2, Abraham Korol1

University of Haifa
Whole Genome Profiling (WGP): a new sequence-based physical mapping technology.

**SNaPshot**
- EcoRI, BamHI, XbaI, XhoI and HaeIII sites
- BAC clone
- Digested by 5 enzymes
- Labeled with 4 dyes
- Between 40 and 250 bands per BAC

**Tolerance of 0.4 bp between two bands to be considered as identical**

**WGP**
- EcoRI sites
- BAC clone
- EcoRI digestion -> adapter ligation
- Sequencing each side
- On average, 58 sequence tags of 36 bp per BAC (138kb, 1 EcoRI site / 5.1kb)

**Pilot project on 16,128 FP (9.6X) of 3B**
WGP improves physical mapping in wheat

Physical map assembly with FPC

<table>
<thead>
<tr>
<th></th>
<th>SNaPshot (1e-25)</th>
<th>WGP (1e-11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated coverage in length</td>
<td>236 Mb ± 65</td>
<td>199 Mb ± 42</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>631</td>
<td>434</td>
</tr>
<tr>
<td>Average contig size (Kb)</td>
<td>374</td>
<td>469</td>
</tr>
<tr>
<td>N50 (Kb)</td>
<td>455</td>
<td>567</td>
</tr>
</tbody>
</table>

Comparison to 12 sequenced contigs:

<table>
<thead>
<tr>
<th></th>
<th>SNaPshot (1e-25)</th>
<th>WGP (1e-11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage percentage in length</td>
<td>95.8%</td>
<td>94.9%</td>
</tr>
<tr>
<td>Number of chimerical contigs for 10 Mb</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>Percentage of mis-assembled BACs</td>
<td>9.5%</td>
<td>2.7%</td>
</tr>
</tbody>
</table>

- Equivalent coverage in length
- Less mis-assembled BACs in WGP
- Equivalent number of chimerical contigs

Philippe et al, BMC genomics 2012)
**FPC vs LTC**

### FPC

- **Based only on BACs similarity**
  - High initial stringency to limit chimerical contigs
  - Adding singletons to contigs extremity at each step
  - Merging contigs at each step

### LTC

1. **Cut-off $10^{-40}$**
2. **Cut-off $10^{-35}$**
3. **Cut-off $10^{-30}$**
4. **Cut-off $10^{-25}$**
5. **Cut-off $10^{-15}$**

- **Non-linear contigs**
- **Based on BAC similarity and contigs linearity**
  - Low initial stringency to limit the number of contigs
  - Checking contigs linearity

- **Splitting non linear contigs by increasing the stringency**

**Elimination of chimerical contigs**
LTC vs FPC (1BL physical map)

- 1BL estimated size = 535 Mb
- 65,413 useful fingerprints (SNaPshot)

LTC significantly improves physical mapping in wheat

IWGSC workshop
Tuesday 17th 1-6pm
IWGSC physical maps: January 2012

TOTAL: 2,253,312 BAC clones

~ 16 x, 118 kb

Completion by Summer 2012
Combined strategies to establish a wheat reference genome sequence

Physical mapping of individual chromosomes

Survey sequencing of individual chromosomes
- Gene catalog
- Virtual order
- Markers
- Short term

MTP sequencing
- Anchored and ordered sequence
- Intergenic regions
- Markers
- Long term

A reference sequence anchored to the genetic and phenotypic maps
Chromosome Survey Sequencing

Amplified DNA/sorted chromosomes

Illumina reads (2*108 bp/) PE 0.5 kb
Min 50 x

Assembly (ABySS)
K-mer 71
Contigs > 200bp N50 = 2.1 kb

http://urgi.versailles.inra.fr/Species/Wheat/Sequence-Repository
Combined strategies to establish a wheat reference genome sequence

Physical mapping of individual chromosomes

Survey sequencing of individual chromosomes

MTP sequencing

Gene catalog
Virtual order
Markers
Short term

Anchored and ordered sequence
Intergenic regions
Markers
Long term

A reference sequence anchored to the genetic and phenotypic maps
3B SEQuencing Project (1Gb)

Chr 3B physical map

1282 BAC-contigs
8441 BACs

Pool of 10 BACs
(Roche 454 GSFLX Titanium, 8 Kb MP)

922 pools

BAC pool

454 scaffolds

Illumina contigs

BAC-ends

Super-scaffolds

✓ Annotation (TriAnnot)
✓ Anchoring/orientation (ISBP SNPs)
✓ Resequencing and polymorphisms analyses
✓ Transcription map (15 RNASeq)
Wheat 3B sequence automated annotation

IWGSC workshop
Tuesday 17th 1-6pm

16,136 scaffolds
16,136 fasta files
Cluster 712 cores < 1 day

15,544 predicted genes

8,734 genes
BlastClust.pl
- Identity: 99%
- Overlap: 90%

TAAC.pm
- Filtering
  No homology: 7,282
  Similarity to Transposase: 44
- Automated validation
  Start/stop, splicing sites: 2,815 (34%)

9,351 genes selected

Leroy et al, Frontiers in Plant Science 2012)
IWGSC MTP sequencing

www.wheatgenome.org