Odyssey of the IWGSC Reference Genome Sequence: 12 years 1 month 28 days 11 hours 10 minutes and 14 seconds.

Kellye Eversole
IWGSC Executive Director

Plant Genomics and Gene Editing Congress
Amsterdam, The Netherlands
16 March 2017
The odyssey begins…  2005
2005 - Genome sequencing – the ‘wheat’ challenge

**Wheat**
*Triticum aestivum*
(16 Gb)

**Barley**
*Hordeum vulgare*
(5 Gb)

**Corn**
*Zea mays*
(2.5 Gb)

**Soybean**
*Glycine max*
(1.1 Gb)

**Rice**
*Oryza sativa*
(0.45 Gb)

**Arabidopsis**
*Arabidopsis thaliana*
(0.15 Gb)

Dal-Hoe Koo, Wheat Genetics Resource Center, Kansas State University

Gb = 1,000,000,000 DNA base pairs
How to produce a useful sequence?
Goal

- Lay a foundation to accelerate wheat improvement
- Increase profitability throughout the industry

Vision

- High quality annotated genome sequence, comparable to rice
- Physical map-based, integrated and ordered sequence
Dissection of the genome to single chromosomes (arms) representing individual (sub)genomes

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 (2006 - 2012)
- Amplified DNA for chromosome survey (2010 - 2011)
An integrated and ordered 3B reference sequence

<table>
<thead>
<tr>
<th>MetaQTL analysis</th>
<th>3B consensus map (5000 markers)</th>
<th>3B Physical map</th>
<th>3B pseudomolecule</th>
</tr>
</thead>
</table>

Feuillet, Paux, and Choulet, Science 2008 and Science 2014
Roadmap to the Wheat Genome Sequence

Illumina sequencing of individual chromosomes
- IWGSC CSS v2 (2014)
  - Whole genome mate pairs
  - IWGSC CSS v3 (2016)

Physical maps of individual chromosomes
- MTP sequencing
  - Pseudomolecule assembly
  - Chromosome 3B (2014)
  - 20 chromosomes (2016)

NRGene-Illumina WGS
- IWGSC Whole Genome Assembly v0.4 (2016)

Radiation Hybrid, Hi-C, genetic, LD maps
- BioNanoGenomics optical maps
- MTP sequence tags.....

Reference Genome Sequence (2017)
The IWGSC CS WGA Project – timeline 2015

~2 months from data accumulation to completion of first assembly
**IWGSC Whole Genome Assembly Project**

De novo assembly:
- NRGene's DeNovoMagic-2 platform, total run time < 3 weeks, 1Tb RAM computer
- Illumina short-read sequencing data only (200 x coverage)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly size</td>
<td>14.5 Gbp</td>
</tr>
<tr>
<td>Est. gaps size</td>
<td>262 Mbp</td>
</tr>
<tr>
<td>Gaps %</td>
<td>1.80</td>
</tr>
<tr>
<td>Total # scaffolds</td>
<td>138,484</td>
</tr>
<tr>
<td>N50</td>
<td>7.1Mbp</td>
</tr>
<tr>
<td>L50 (#sequences)</td>
<td>566</td>
</tr>
<tr>
<td>N90</td>
<td>1.3 Mbp</td>
</tr>
<tr>
<td>L90 (#sequences)</td>
<td>2,363</td>
</tr>
<tr>
<td>MAX Scaffold</td>
<td>45.8 Mbp</td>
</tr>
</tbody>
</table>
Concerted integration of resources: RefSeq v1.0
IWGSC RefSeq Project

- Physical maps for all chromosomes
  - 1,839,128 BACs, 47,810 contigs, 380,675 singletons
- WGP tags (mostly from MTP BACs) for all chromosomes except 3B
  - 4,305,249 unique tags, 693,697 BACs
- BAC sequence assemblies for 8 chromosomes (1A, 1B, 3B, 3D, 6B, 7A, 7B, 7D) and partial MTP data for two arms (4AL, 5BS)
  - 52,890 BACs (9.7 Gb), N50 - 68 kb
- Optical maps for 7A, 7B and 7DS
  - 1,335 BioNanoGenomics contigs aligned to the WGA assembly
- GBS map of the SynOp RIL population
  - 179 RILs, 4074 markers
## Comparison of IWGSC Assembly Releases

<table>
<thead>
<tr>
<th></th>
<th>IWGSCv0.4</th>
<th>RefSeqv1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number / coverage of scaffolds/contigs</td>
<td>138,607 / 14.5 Gb</td>
<td>138,665 / 14.5 Gb</td>
</tr>
<tr>
<td>Number / coverage of scaffolds/contigs &gt;=100kb</td>
<td>4,442 / 14.2 Gb</td>
<td>4,443 / 14.2 Gb</td>
</tr>
<tr>
<td>N50 scaffolds / superscaffolds</td>
<td>7.0 Mb</td>
<td>22.8 Mb</td>
</tr>
<tr>
<td>L50 (no. sequences (\rightarrow) N50)</td>
<td>566</td>
<td>166</td>
</tr>
<tr>
<td>N90 scaffolds / superscaffolds</td>
<td>1.3 Mb</td>
<td>4.1 Mb</td>
</tr>
<tr>
<td>L90 (no. sequences (\rightarrow) N50)</td>
<td>2363</td>
<td>718</td>
</tr>
<tr>
<td>Gaps filled with BAC sequences</td>
<td></td>
<td>183 (1.7 Mb)</td>
</tr>
<tr>
<td>Average size of inserted BAC sequences</td>
<td></td>
<td>9.5 kb</td>
</tr>
<tr>
<td>Sequence assigned to chromosomes</td>
<td>14.1 Gb (96.8%)</td>
<td>14.1 Gb (96.8%)</td>
</tr>
<tr>
<td>Sequence assigned to chromosomes (&gt;=100kb)</td>
<td>14.1 Gb (99.1%)</td>
<td>14.1 Gb (99.1%)</td>
</tr>
<tr>
<td>No. scaffolds / superscaffolds on chromosomes</td>
<td>3,975</td>
<td>1,601</td>
</tr>
<tr>
<td>No. oriented scaffolds / superscaffolds</td>
<td>2,464</td>
<td>1,243</td>
</tr>
<tr>
<td>Oriented sequence</td>
<td>13.1 Gb (90.2%)</td>
<td>13.8 Gb (95%)</td>
</tr>
<tr>
<td>Oriented sequence &gt;=100kb</td>
<td>13.1 Gb (92.4%)</td>
<td>13.8 Gb (97.3%)</td>
</tr>
</tbody>
</table>

*RefSeq v1.0 contains ~75 scaffolds per chrom.*
IWGSC RefSeq v1.0 Annotation

- **Triannot pipeline**
  - Repeat database: ClariTeRep
  - Agreed Gene Model Evidence: RNASeq, ISOSeq, flicDNA
  - Repeat database: ReCAT

- **PGSB pipeline**
  - Gene calls

- **Gene Models**
  - Combined gene models
  - Functional annotation: GO, InterPRO
  - Classification: pseudogene identification

- **IWGSC RefSeq v1.0 pseudomolecules**
  - For genome-wide analysis
  - For publication

- **Manual annotation of gene families**: NB-LRR, AA-transporter, PPR, ...

- **Continued manual annotation and curation of annotations**

- **Future Annotated IWGSC RefSeq v2.0**
IWGSC RefSeq Data Access & Availability

Pre-publication data access:
IWGSC WGA v0.4: June 13, 2016
IWGSC RefSeq v1.0: January 14, 2017

Gene models completed: March 2017
Final analyses completed: April/May 2017
Manuscript submission: Summer 2017

https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies
ARRIVAL IN PORT OF HISTORIC IWGSC CRUISE
1ST REFERENCE SEQUENCE OF BREAD WHEAT ACHieved
IWGSC 2.0

• Manual and functional annotation of the sequence to empower gene discovery and gene cloning to understand the molecular basis of traits

• Coordination of re-sequencing activities for diversity panels that represent the breadth of geographic distribution of germplasm for breeders

• Support the development of user-friendly, integrated databases
Lessons learned

• At least one high quality, manually annotated reference sequence
• Physical map-based for adaptability to any technology
• Maintain flexibility for new technologies without losing sight of quality
• Stay on the course towards your vision
Acknowledgments

IWGSC Leadership: Rudi Appels, Kellye Eversole, Catherine Feuillet, Beat Keller, Jane Rogers

IWGSC Chromosome Leaders:

Etienne Paux, Frédéric Choulet
Jaroslav Dolezel, Hana Simkova, Miroslav Valarik, Jan Bartos
Hirokazu Handa
Luigi Cattivelli
Elena Salina
Nikolai Ravin

Bikram Gill
Catherine Feuillet
John Jacobs
Beat Keller
Abraham Korol
Odd-Arne Olsen

Rudi Appels
Hikmet Budak
Nils Stein
Thorsten Schnurbusch
Curtis Pozniak
Andrew Sharpe
Kuldeep Singh
Song Weining
Matt Clark
All physical maps and pseudo-molecule sequences available at IWGSC repository: https://wheat-urgi.versailles.inra.fr
Thank you for your attention!