The Wheat Genome Sequence

Nils Stein, IPK Gatersleben

International Wheat Congress
8 December 2016
Frankfurt, Germany
Genome size / Genome sequencing

*Hordeum vulgare*
5 Gb (12 x rice)

*A.thaliana*
0.15 Gb

*Triticum aestivum*
17 Gb (40 x rice)
Wheat is a challenge for genomic studies & sequencing

Challenging Bread Wheat Genome

Average plant genome size ~ 6.000 Mb

~ 90% Repeat sequences

~ 16,000 Mb

Human ~ 3.000 Mb

~ 5,000 Mb

~ 2,300 Mb

50-80%

~ 355 Mb

~50%

~ 140 Mb

~10%

- A. thaliana (2x)
- Rice (2x)
- Maize (2x)
- Barley (2x)
- Bread wheat (6x)

120-fold A. thaliana genome
45-fold rice & Brachypodium genomes

Courtesy: Catherine Feuill
Wheat genome sequencing in 2005

- Too big genome for Sanger Sequencing
- Too big task for single group/lab
- Too costly
**A chromosome-based approach**

- 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)

*Doležel et al., Chromosome Res. 15: 51, 2007*
An integrated and ordered 3B reference sequence

MetaQTL analysis  
3B consensus map (5000 markers)  
3B Physical map  
3B pseudomolecule

Courtesy: Catherine Feuillet
The International Wheat Genome Sequencing Consortium

2016

Board of Directors
Role: Overall strategy and organization
Up to 10 directors including at least 1 Outside Director

Leadership Team
Role: Daily management
Members: Appels, Eversole, Feuillet, Keller, & Rogers

Coordinating Committee
Role: Scientific strategy
Members: Sponsors & Leaders of IWGSC Projects & Initiatives

Members
Role: Participation & Input
Members: Open to Anyone

COUNTRIES
59

SPONSORS
19

MEMBERS
1400

INSTITUTES/COMPANIES
434

www.wheatgenome.org
Vision and roadmap

**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

Chromosomal Survey Sequence

Chromosomal-based Physical Maps

MTP Sequencing

Chromosomal Survey Sequence

Whole Genome Shotgun Assembly

A REFERENCE SEQUENCE LINKED TO GENETIC AND PHENOTYPIC MAPS

www.wheatgenome.org
Roadmap to the Wheat Genome Sequence

Illumina sequencing of individual chromosomes

- Chromosome Survey Sequence v2 (2014)
- Whole genome mate pairs
- IWGSC CSS v3 (2016)

Physical maps of individual chromosomes

- MTP sequencing
- Pseudomolecule assembly

- 100%
- 62%

- Chromosome 3B (2014)
- 20 chromosomes (2016)

NRGene-Illumina WGS

- IWGSC Whole Genome Assembly (2016)

BioNanoGenomics, optical, RH, HiC maps

Genetic, LD maps

MTP sequence tags.....

Reference Genome Sequence (2017)

www.wheatgenome.org
Progress towards completion of Bread Wheat Reference Genome Sequence

*Flags represent countries where work is underway, as of December 2016

All physical maps available at IWGSC repository at URGI

All pseudo-molecules completed

www.wheatgenome.org
Wheat News from Hogwarts

Re: https://thescienceweb.wordpress.com/2016/01/10/assembly-problem-solved-by-wizardry/
The IWGSC CS WGA Project – timeline 2015

- Agreement Go!
- DNA Extraction
- Libraries Complete
- Sequencing Complete
- Assembly V0.1
- Assembly V0.2
- Complete QA

August | Sept | October | November | December | January | February | March

~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-reads sequencing data only (x200 coverage)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly size:</td>
<td>14.5 Gbp</td>
</tr>
<tr>
<td>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>L50:</td>
<td>7.1 Mbp</td>
</tr>
<tr>
<td>N50 (#sequences):</td>
<td>566</td>
</tr>
<tr>
<td>L90:</td>
<td>1.3 Mbp</td>
</tr>
<tr>
<td>N90 (#sequences):</td>
<td>2,363</td>
</tr>
<tr>
<td>MAX Scaffold:</td>
<td>45.8 Mbp</td>
</tr>
</tbody>
</table>
WGA Assembly Statistics

N50 = 7.4 Mb

size = 14.5 Gb

covered assembly length

10 Gb

1 Gb

100 Mb

100 Mb

10 Mb

1 Mb

100 kb

100 kb

minimal sequence length

10 Mb

1 Mb

100 kb

10 kb

1 kb

100 bp

IWGSC WGA
TGAC WGS
IWGSC CSS
JGI WGS
Concerted integration of resources: RefSeq

- IWGSC WGA
- Chromosome-specific MTP assemblies
- Physical maps
- Bionano optical maps
- WGP™ tags of all MTPs
- Genetic maps
- RH maps
- POPSeq
- HiC

IWGSC chromosome survey sequence
IWGSC RefSeq Project

- WGP tags (mostly from MTP BACs) for all chromosome except 3B
  - 693,697 BACs, 4,305,249 unique tags
- 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
- Physical maps for all chromosomes
  - 1,839,128 BACs, 47,810 contigs, 380,675 singletons
- Optical maps for 7BS, 7BL and 7DS
  - 1,335 Bionano contigs aligned to the WGA assembly
- GBS map of the SynOp RIL population
  - 179 RILs, 4074 markers
## IWGSC RefSeq Comparison

<table>
<thead>
<tr>
<th></th>
<th>v0.4</th>
<th>v1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>number/size of scaffolds/contigs</td>
<td>138,607/14.5 Gb</td>
<td>138,665/14.5 Gb</td>
</tr>
<tr>
<td>number/size of scaffolds/contigs &gt;= 100 kb</td>
<td>4,442/14.2 Gb</td>
<td>4,443/14.2 Gb</td>
</tr>
<tr>
<td>N50 of scaffolds / chromosomal superscaffolds</td>
<td>7.0 Mb</td>
<td>22.8 Mb</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 chr. &gt;= 100 kb</td>
<td>14.1 Gb (99.1 %)</td>
<td>14.1 Gb (99.1 %)</td>
</tr>
<tr>
<td># scaffolds/superscaffolds on chromosomes</td>
<td>3,975</td>
<td>1,601</td>
</tr>
<tr>
<td># 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 (95.0 %)</td>
</tr>
<tr>
<td>oriented sequence &gt;= 100 kb</td>
<td>13.1 Gb (92.4 %)</td>
<td>13.8 Gb (97.3 %)</td>
</tr>
</tbody>
</table>

~75 scaffolds / chromosome
Data provided by Etienne Paux
IWGSC RefSeq Annotation – General

RNA-seq → Isoform sequencing → Reference proteins

Merging of predicted exon/intron structures

Prediction of coding sequences

Confidence classification

High confidence

Low confidence
IWGSC RefSeq 1.0 Gene / Genome

IWGSC RefSeq v1.0 pseudomolecules

Triannot pipeline

Repeat-masked scaffolds

Repeat database ClariTeRep

Agreed Gene Model Evidence: RNASEq, ISOSeq, fLcDNA, ...

Repeat database ReCAT

gene models

Combined gene models

Gene model evaluation, Earlham Institute

Functional annotation, GO, InterPRO

Classification, pseudogene identification

IWGSC RefSeq v1.0 pseudomolecules for genome-wide analysis

MIPS pipeline

gene calls

Continued manual anno and curation of annotations

Future Annotated RefSeqv2.0

manual anno of gene families: NBS-LRR, AA-transporter, PPR...

Repeat database

ClariTeRep

Agreed Gene Model Evidence: RNASEq, ISOSeq, fLcDNA, ...

ReCAT

gene models

IWGSC RefSeq v1.0 pseudomolecules for publication
IWGSC RefSeq Data Access & Availability

https://wheat-urgi.versailles.inra.fr/
IWGSC accomplished its goal of generating a reference sequence for bread wheat almost within 10 years after its initiation.

This reference sequence is built of highly diverse community resources, a.o. physical maps of chromosomes, CSS assemblies, BAC-based MTP chromosome sequences, high quality whole genome shotgun assembly, millions genetic markers, Hi-C scaffolding.

draft pseudomolecules were made pre-publication accessible in summer 2016.

The reference sequence will be published in 2017.

Future improvements on sequence and annotation of the CS RefSeq will be coordinated by IWGSC.

The reference sequence is expected to:
- reduce time and improve success of cloning genes and QTL
- unlimited access to DNA markers for MAS and GS, perfect markers
- facilitate exploration of diversity in genetic resources for pre-breeding
Wheat genome sequencing – work in progress

Reference quality assemblies in progress:

- Triticum monococcum (2n)
- Aegilops tauschii (2n)
- Wild emmer (4n)
- Durum wheat (4n)
- CS (6n)

<table>
<thead>
<tr>
<th></th>
<th>Wild emmer</th>
<th>Durum</th>
<th>CS v0.4</th>
<th>CS v1.0</th>
<th>CDC Landmark</th>
<th>CDC Stanley</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly size:</td>
<td>10.5 Gbp</td>
<td>10.45 Gbp</td>
<td>14.5 Gbp</td>
<td>14.1 Gbp</td>
<td>14.5 Gb</td>
<td>14.5 Gb</td>
</tr>
<tr>
<td>Total # scaffolds:</td>
<td>151,912</td>
<td>129,464</td>
<td>138,484</td>
<td>1,601 [4,161]</td>
<td>146,603</td>
<td>166,832</td>
</tr>
<tr>
<td>L50:</td>
<td>6.9 Mbp</td>
<td>6.0 Mbp</td>
<td>7.1 Mbp</td>
<td>22.8 Mbp</td>
<td>6.1 Mbp</td>
<td>5.9 Mbp</td>
</tr>
<tr>
<td>N50 (#sequences):</td>
<td>414</td>
<td>493</td>
<td>566</td>
<td>166</td>
<td>727</td>
<td>684</td>
</tr>
<tr>
<td>L90:</td>
<td>1.1 Mbp</td>
<td>1.1 Mbp</td>
<td>1.3 Mbp</td>
<td>4.1 Mbp</td>
<td>1.1 Mbp</td>
<td>1.2 Mb</td>
</tr>
<tr>
<td>N90 (#sequences):</td>
<td>1,827</td>
<td>2,019</td>
<td>2,363</td>
<td>718</td>
<td>2,687</td>
<td>2,676</td>
</tr>
</tbody>
</table>
More wheat genome sequences – towards wheat pan-genome analysis

10 wheat genomes (and more) project:

- 2 Canadian varieties (CDC Landmark, CDC Stanley)
- 1(+2) US varieties (Jagger)
- 1 German winter wheat variety (Julius)
- 1 Swiss winter wheat variety (Arina)
- (2) Australian varieties (tbd)
- 1(+1) Chinese variety (Zhongmai175)
- CS, Cadenza, Paragon, Kronos, Robigus, Claire, Alchemy, Brompton, Hereward, Rialto, Soissons, Xi19

NRGene refseq quality assemblies
Earlham Institute assemblies

VARIETY 1
PAN GENOME
CORE GENOME
DISPENSIBLE GENOME
VARIETY 2
Sequencing of a german winter wheat: „Julius“

• Sequencing of 800 nt PE and MP libraries done
• Sequencing of 450 nt PE libraries ongoing

• Assembly, End of January 2017

• Hi-C/10x genomics based scaffolding, March 2017
Wheat genome sequencing – the future

- Chinese Spring “Gold Standard”
- Core and Pan Genomes
- Structural/PAV
- Low coverage WGS
- Haplotypes
- Exome Sequencing
  RNA Seq
  Reduced representation
- SNPs/Alleles
Summary and Conclusion

- Wheat CS RefSeqv1.0 will be publicly available in 2017
- similar resources are in progress for wheat species at all ploidy levels
- Additional reference quality sequence resources will become available for a minimum of 10 more haplotypes within the next 12-18 months initiating the era of wheat pan-genomics
- The Wheat Initiative and IWGSC will stay committed to communicate progress in wheat research to the community and inform about work in progress and data accessibility
- Adademia and Industry must be aware of the developments and get their Bioinformatics data storage and analysis infrastructure in place to efficiently accommodate multigenome information for breeding and research
Thanks to funders of IWGSC WGA Project
IWGSC Chromosome Leaders

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

Etienne Paux, Frédéric Choulet

Bikram Gill

Catherine Feuillet

Jaroslav Dolezel, Hana Simkova, Miroslav Valarik, Jan Bartos

John Jacobs

Rudi Appels

Beat Keller

Hirokazu Handa

Abraham Korol

Luigi Cattivelli

Kuldeep Singh

Hikmet Budak

Odd-Arne Olsen

Curtis Pozniak

Nils Stein

Thorsten Schnurbusch

Elena Salina

Kuldeep Singh

Song Weining

Matt Clark
Thanks to IWGSC Sponsors!
An international initiative to sequence the genome of the durum wheat cultivar Svevo

Luigi Cattivelli
Nicola Pecchioni

Aldo Ceriotti
Luciano Milanesi
Gabriella Sonnante

University of Bologna
M. Maccaferri, S.Salvi,
Roberto Tuberosa

Nils Stein
Martin Mascher

Hikmet Budak

University of Tel Avi
Assaf Distelfeld

Klaus Mayer

HelmholtzZentrum münchen
German Research Center for Environmental Health

University of Saskatchewan

Curtis Pozniak
Acknowledgements - WHEATSEQ

IPK GGR / GED
Sudharsan Padmarasu
Axel Himmelbach
Susanne König
Ines Walde

IPK DG
Martin Mascher

IPK BIT
Sebastian Beier
Uwe Scholz

PGSB/HMGU
Manuel Spannagl
Heidrun Gundlach
Sven Twardziok
Klaus Mayer