Structure and Dynamics of the Hexaploid Wheat Genome

Frédéric CHOULET

Genetics Diversity Ecophysiology of Cereals
INRA – U. Clermont-Ferrand, France
Introduction to Wheat Genome Research

- Structure, evolution of wheat genome
- Recombination
- Paleogenomics
- Grain composition
- Response to abiotic stress
- Resistance to pathogens
- Diversity, selection

Objectives

- Resources
  - Sequencing the wheat genome
  - Markers
  - Bioinformatics

- Research
  - Structure / Expression / Evolution
    - TEs
    - Gene space
    - Duplications
    - Structural variations
    - Epigenome
Why wheat?

- Important crop
- Complex genome
17 Gb
AABBDD
85% TEs
- Launched in 2005
- Goal
  - Produce a high quality ref seq of the bread wheat genome
- Strategy
  - Reduce the complexity
WGS
- Brenchley et al. Nature 2012
- Ling et al. Nature 2013
- Jia et al. Nature 2013

Chr. Survey Seq (=CSS)
- IWGSC Science 2014

MTPseq
- Choulet et al. Science 2014
- Daron et al. Genom Biol 2015
- Pingault et al. Genom Biol 2015
- Glover et al. Genome Biol 2015
**Chr Survey Seq (2014)**

- **Resources**
  - 1 draft seq / chromosome arm
  - **10 Gb - 10 M** contigs (N50: 2.4 kb)
  - **99,000** genes
  - ~60% genes "zipped"

- **Main Results**
  - Gene loss--
  - SSD++
  - Dominance--
3B pseudomolecule (3BSEQ project)

- Protein coding genes (pseudogenes: 27%) 7,264
- Transposable elements 252,879 (86%)
Recombination
LD

centromeric retrotransposons

pericentromere

68 Mb 122 Mb 59 Mb

774 Mb
Gene density (/10 Mb)

TE density (%)

Expression (# conditions)

Choulet et al. Science 2014
Chromosome partitioning: 3B-specific??

Expression breadth

3A

3D

barley

maize
Accelerated evolution in the *Triticeae*

- Wheat 3B
  - 27% nonsyntenic genes

- Brachy-2
- Rice-1
  - ~10%
- Sorghum-3

- inter-chromosomal gene duplications
- intra-chromosomal

- More duplicated genes in the chr. extremities
- Enriched in adaptation functions
- Annotation challenge
- Impact on genome biology
New tools: CLARITE and ClariTeRep

3B → 252,000 TEs

Copia

CACTA

Gypsy
- silenced until polyploidization...
- ... but shared betw A-B-D
Evolutionary forces driving TE distribution

# copies

0-1 MYA  1-2 MYA  2-3 MYA  >3 MYA
The diagrams show gene distribution and synteny status. The top graph illustrates non-syntenic genes with a green line indicating a percentage distribution across kb positions. The bottom graph, marked with a purple outline, focuses on syntenic genes with a similar percentage distribution pattern. On the right side, a bar graph distinguishes between non-syntenic (red) and syntenic (blue) genes across -20 to +20 kb, with a green bar highlighting the region of interest.
Structural Variations
  o  CNVs
  o  PAVs
Structural Variations (SVs)

- Small indels
- Inversions, translocations
- **Duplications & Deletions --> CNVs & PAVs**

Main questions:

- Extent of SVs among the *Triticaceae*, hexaploid wheat acc.?
- Impact of polyploidization?
- Relationships betw SVs and chr. organization?
- Impact on phenotypes?
Structural Variations (SVs) – Detection

- Aligning orthologous sequenced loci

- Using resequencing data (short read-based)
  - not properly mapped paired-reads
  - split reads
  - depth of coverage

→ Limitations in polyploid TE-rich genomes
- **limitations**: homeologs and paralogs (repeated genes)
Resequencing data

- **advantage:** diploid context
- **limitations:** 3B DNA amplified before seq.
Resequencing data

45 accessions

- 22 hexaploids: *T. aestivum*
- 6 hexaploids: *T. macha, spelta*
- 17 tetraploids: *T. durum, dicoccoides, dicoccum, carthlicum*

- Illumina 2x100bp
- Depth: ~40x
- TEs (85%)
- Genes (2%)
SVs in genes – Methodology

- BWA & samtools

44 accessions  3B

sorted 3B reads

Chinese Spring 3B

PAVs

<10% of the gene length covered by reads

CNVs

depth of coverage-based approach

Normalization:
- GC%
- sequencing depth
- gene length
- \( \log_2(\text{cov}_{\text{acc}}/\text{cov}_{\text{Chinese Sp}}) \)
SVs in genes – Results

**PAVs**

XX% genes deleted in 1+ accessions [XX..XX]

**CNVs**

down\textsuperscript{CNVs}: X% of the 3B genes (on average per acc.)

up\textsuperscript{CNVs}: X%

XXXX (XX%) genes with no variation among 45 accessions
SVs in genes – Results

T. aestivum (AABBDD)  
Nanking_NO25

T. dicoccoides (AABB)  
acc_26676
SVs in genes – Methodology

=> develop a fine-tuned strategy?... in progress
SVs in TEs – Methodology

TE junction-based approach (**PAVs** only)

+/-75 bp

101 bp reads
SVs in TE–Results

3B → ~500k loci to study TE-PAVs along chr3B

- Polymorphic loci among 45 accessions: XX%
- Average per accession: XX% [XX% .. XX%]
SVs in TEs – Results

- *T. dicoccoides* (AABB): 3
- *T. dicoccum* (AABB): 4
- *T. carthlicum* (AABB): 4
- *T. petropavlovskyi* (AABBDD)
- *T. polonicum* (AABB)
- *T. durum* (AABB): 6
- *T. turgidum* (AABB)
- *T. macha* (AABBDD): 3
- *T. spelta* (AABBDD): 3
- *T. sphaerococcum* (AABBDD)
- *T. aestivum* (AABBDD): 22
- *T. compactum* (AABBDD)
- *T. vavilovii* (AABBDD)

Graphs showing RNA-seq coverage for acc-33757 and Nanking_NO25.
Conclusions

- Gene PAVs limited (on 3B)
- CNVs: **XX%** of the 3B genes (versus ~10\% in barley
  
  [Munoz-Amatriain et al. 2013] 14 wild+cultiv. genotypes)
- High level of **TE**-related SVs
- Importance of **chr. extremities** in the diversity of **Triticeae**
  (also observed in barley)

Next

- Validate the approach/thresholds used for CNV calling
- SVs at the whole genome level
  - use 3B-vs-3B results as QC
  - increase sample size through exon capture
- GO term enrichment
- Relationships betw **TE** SVs / gene expression
- Unmapped reads (pangenome)
Wheat genome seq initiatives in 2016:

- MTPseq 1A, 1B, 6B, 7A, 7B, 7D
- **IWGSC Whole Genome Assembly (NRGene)**
- TGAC WGS (several varieties)
- U Maryland WGS Pacbio+Ill
- UC Davis Ae. *tauschii* WGS+MTPseq
- BGI *T. urartu* WGS+MTPseq
- Wild Emmer Wheat (NRGene)
chr-by-chr approach, status in 2015...
**IWGSC-Whole Genome Assembly**

**Partners:**
- N. Stein
- C. Pozniak
- J. Poland
- A. Diestelfeld
- A. Scharpe
- G. Ronen
- M. Thompson
- K. Eversole, J. Rogers
- F. Choulet

**Strategy:**
WGS Illumina 180x - 3 MP lib
*DeNovoMAGIC™ 2.0*

**Timeline:**
- Aug 2015 -> start
- Sept
- Oct
- Nov -> Sequencing done
- Dec -> Assembly v0.1
  - QC *(M. Mascher, F. Choulet)*
- Jan. 2016 -> Assembly v0.2
### IWGSC-Whole Genome Assembly

<table>
<thead>
<tr>
<th>Year</th>
<th>Organism</th>
<th>Institute</th>
<th>Assembly Code</th>
<th>N50</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>T.ur</td>
<td>BGI</td>
<td>T.ura-BGI</td>
<td>64 kb /  19000 scaff</td>
</tr>
<tr>
<td>2013</td>
<td>Ae.t</td>
<td>BGI</td>
<td>Ae.ta-BGI</td>
<td>58 kb /  19000 scaff</td>
</tr>
<tr>
<td>2014</td>
<td>CS</td>
<td>IWGSC</td>
<td>CSS</td>
<td>2 kb /    &gt;1M scaff</td>
</tr>
<tr>
<td>2014</td>
<td>CS(3B)</td>
<td>GDEC/CNS</td>
<td>3B-pseudo</td>
<td>892 kb /  296 scaff</td>
</tr>
<tr>
<td>2015</td>
<td>Synth</td>
<td>IPK/JGI</td>
<td>Syn-JGI</td>
<td>21 kb / 1200000 scaff</td>
</tr>
<tr>
<td>2015-Jul</td>
<td>WEW</td>
<td>NRGene</td>
<td>WEW-NRGene</td>
<td>7000 kb / 414 scaff</td>
</tr>
<tr>
<td>2015-Dec</td>
<td>CS</td>
<td>IWGSC</td>
<td>IWGSC-WGA</td>
<td>7394 kb /  547 scaff</td>
</tr>
</tbody>
</table>
**IWGSC-Whole Genome Assembly**

**QC :: Completeness?**

- **exons:** 98.7% match with 100%id-100%ov
- **ISBPs (4.2M):** 97.7% match with 100%id-100%ov
- **WGPtags-4B (0.9M):** 96.2% match with 100%id-100%ov

➡ Completeness+++
- **IWGSC-Whole Genome Assembly**

  QC :: Chimeras?

  ![Diagram showing QC results](image)

  - alignment to CSS genes and ISBPs
  - alignment to MTPseq (3B, 1B)
  - alignment to genetic map (POPseq, CsRe)
  - alignment to physical map (WGPtags)
  - alignment to HiC map

  ➤ **Chimeric scaffolds found (<200) - corrected in v0.2/v0.3**
- **IWGSC-Whole Genome Assembly**

  **IWGSC-WGA v0.2 metrics:**
  - #Scaff $\geq 2$kb: 37,872
  - Size: 14.532 Gb
  - Gaps: 1.8%
  - L50: 7.058 Mb / 566 scaff
  - L90: 1.261 Mb / 2,363 scaff
  - max: 45.794 Mb

  21 pseudomolecules constructed with HiC data (IPK)
  - $\rightarrow$ 14.0 Gb (96%) IWGSC-WGA ordered
IWGSC roadmap update

![Diagram showing steps: sorted chromosomes, Physical maps, IWGSC-WGA, CSS, MTPseq, and Gold Standard labels.]

- Sorted chromosomes
- Physical maps
- IWGSC-WGA
- CSS
- MTPseq
- Gold Standard labels
Integration WGA ⇔ MTPseq in progress...

Strategy based on comparing ISBPs (speed++ specificity++)
207 WGA-scaff joined by 1B-MTP-scaff representing 593 Mb
The diagram illustrates the relationship between read length and log(N50) for different wheat genotypes and assemblies. The x-axis represents different wheat genomes (AA, BB, DD, AABB, AABBD, AABBDD, others, Chinese Spring). The y-axis represents read length in kilobases (100 kb, 1 Mb, 10 Mb, 100 kb). The different genotypes and assemblies are represented by different markers and colors:

- T. urartu (WGS and BAC+WGS)
- Ae. tauschii (WGS and BAC+WGS)
- WEW-WGS
- IWGSC-WGA
- 3B
- Ill-Pacbio
- Cadenza
- Synthetic
- Kronos
- CSS
- TGACv1
- 5x-bbsrc
- Chinese Spring

The chart shows how the read length and log(N50) vary across different assemblies and genotypes, indicating the quality and coverage of the sequencing data.
Thanks

**INRA GDEC**
- Etienne Paux
- Hélène Rimbert
- Ambre-A. Josselin
- Romain De Oliveira
- Jonathan Kitt
- Benoit Darrier
- Nicolas Guilhot
- Philippe Leroy
- Pierre Sourdille
- François Balfourier
- Charles Poncet
- Josquin Daron
- Lise Pingault
- Natasha Glover
- Sébastien Theil
- Aurélie Evrard
- Emeric Dynomant
- Aurélien Bernard

**CEA-Génoscope**
- P. Wincker, V. Barbe et al.

**INRA CNRGV**
- H. Bergès et al.

**INRA URGI**
- H. Quesneville, M. Alaux et al.

**IEB**
- J. Dolezel et al.

**IWGSC**
- K. Eversole, J. Rogers

**IWGSC-WGA working team**
- IPK, U Sask, KSU, U TelAviv, GIFS, Illumina, NRGene, ...
Frédéric CHOULET

GDEC
Genetics Diversity Ecophysiology of Cereals