The Reference Sequence of Wheat and IWGSC Phase II:

Ensuring a Full Genomic Toolbox for Wheat

Kellye Eversole
IWGSC Executive Director
Hélène Bergès
Member of the IWGSC Coordinating Committee

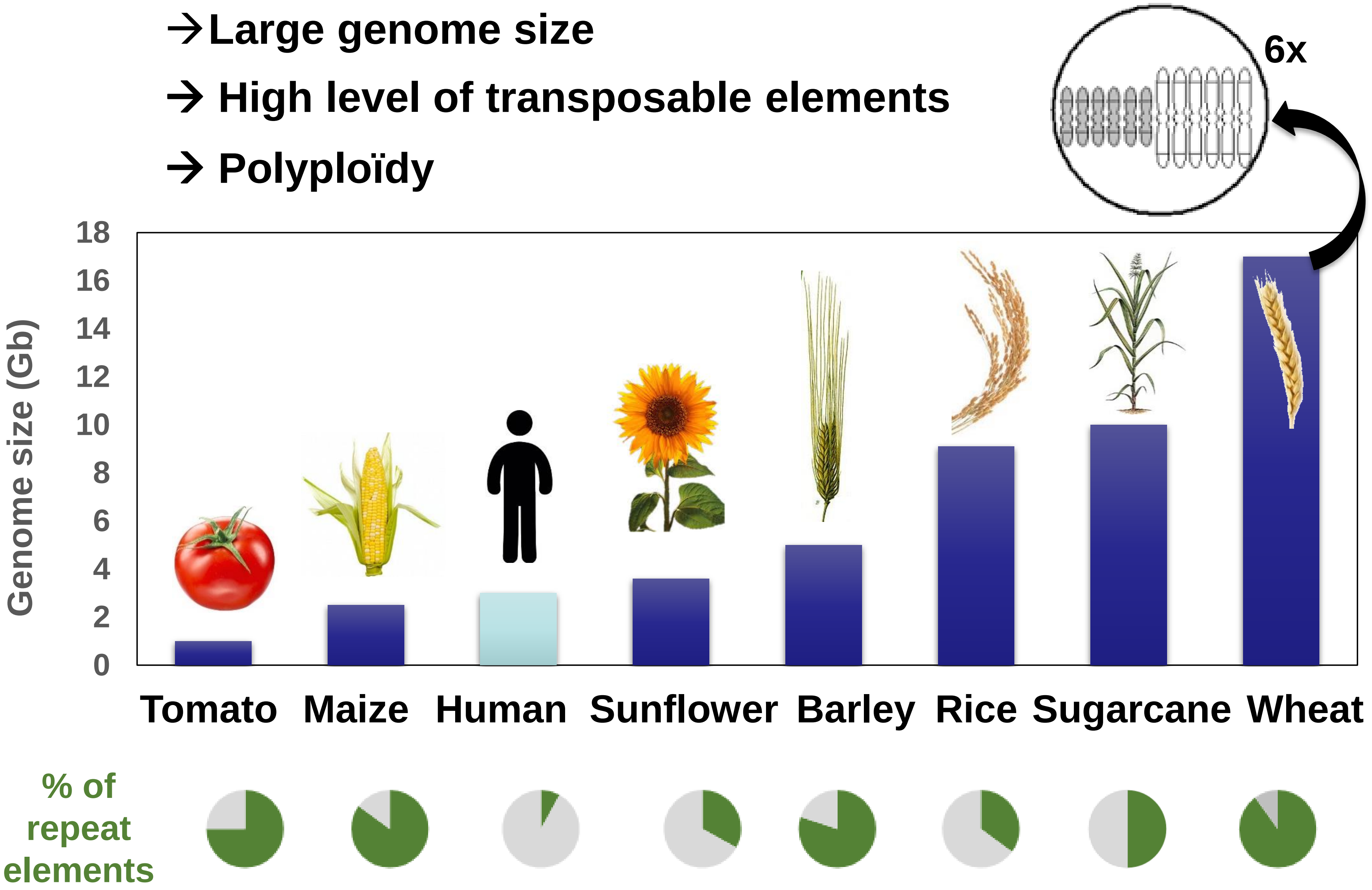
WHEATS&WOMEN INTERNATIONAL CONFERENCE
Roma – June 2018

Plant's genome exhibits high levels of complexity

→Large genome size

→ High level of transposable elements

→ Polyploïdy



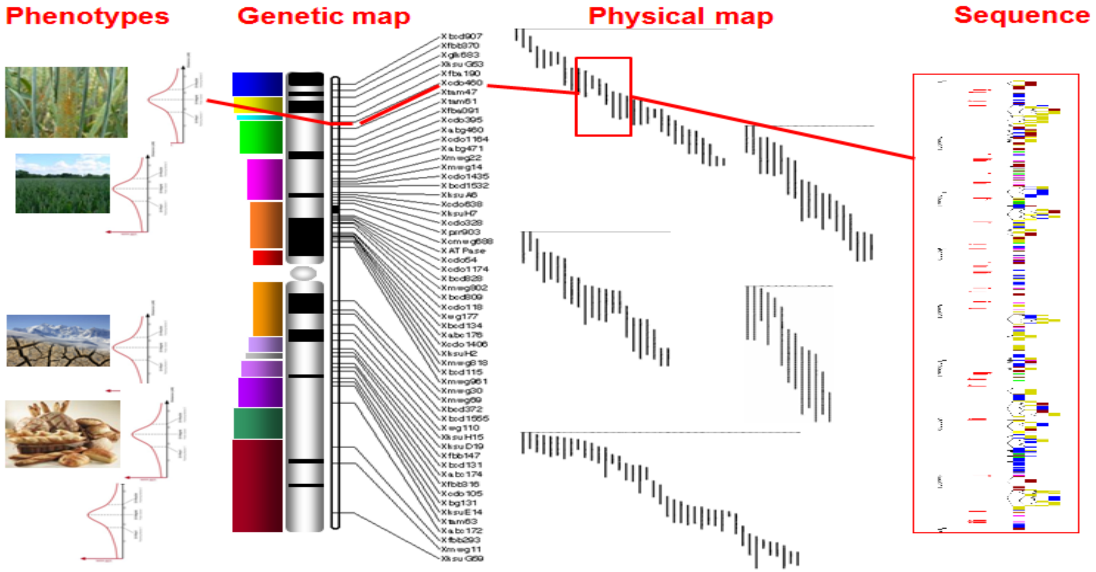
2005 – IWGSC Goal & Vision

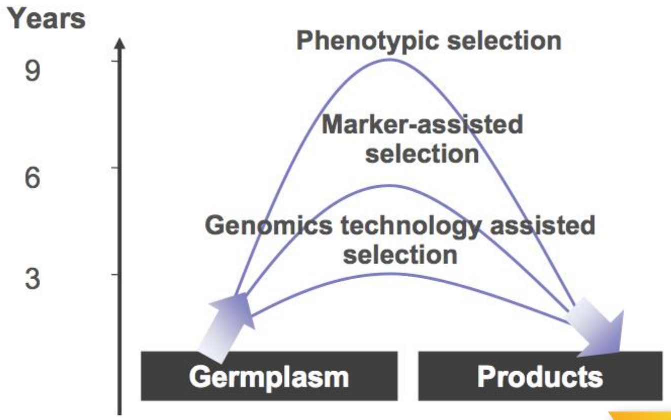
Goal

- 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



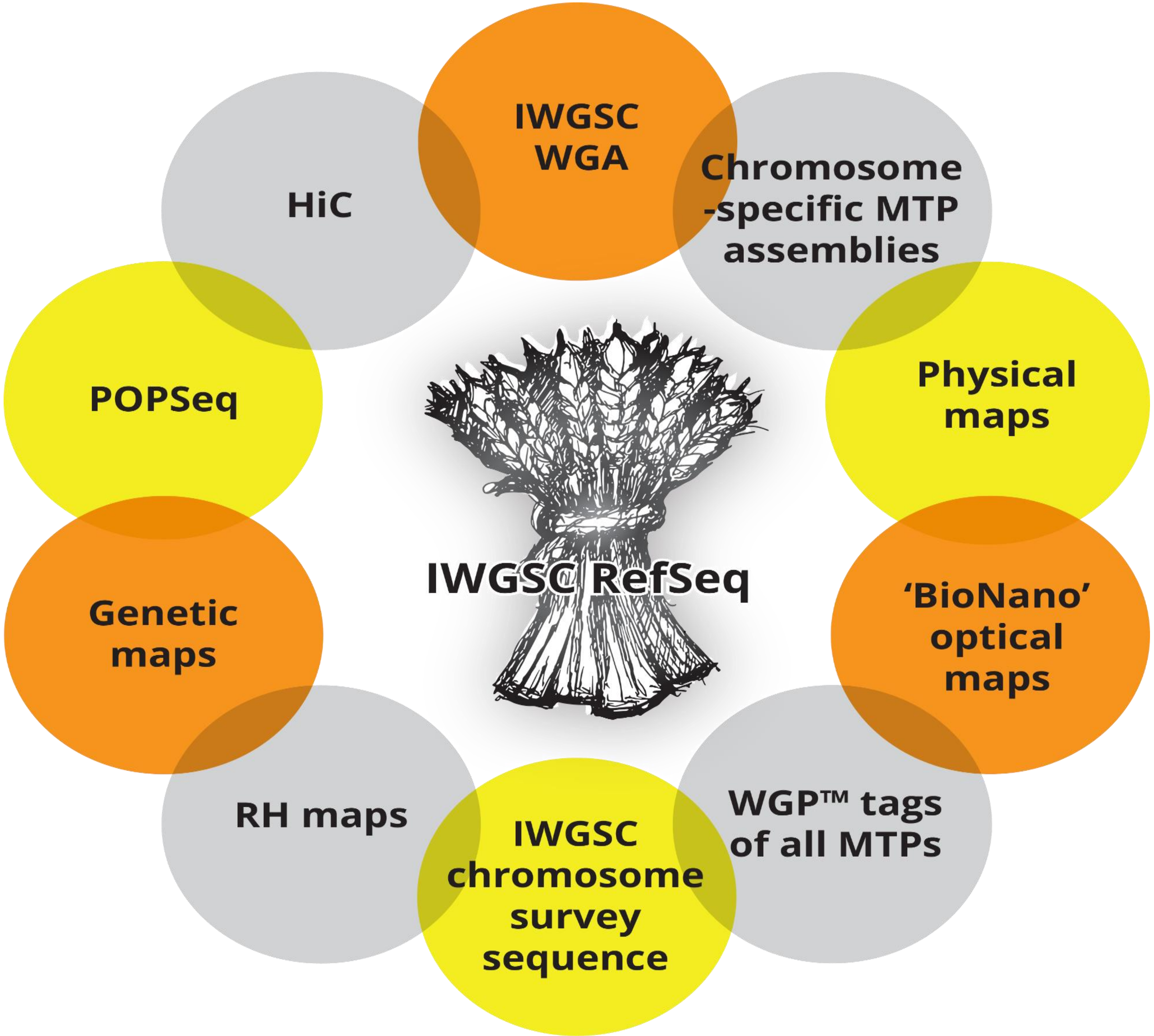


IWGSC RefSeq

The Reference Sequence for the Bread Wheat Genome

Frédéric Choulet
GDEC, INRA, UCA, Clermont-Ferrand, France

Concerted integration of resources: RefSeq v1.0



IWGSC WGA

Chromosome-specific MTP assemblies

Physical maps

'BioNano' optical maps

WGP™ tags of all MTPs

IWGSC chromosome survey sequence

RH maps

Genetic maps

POPSeq

HiC

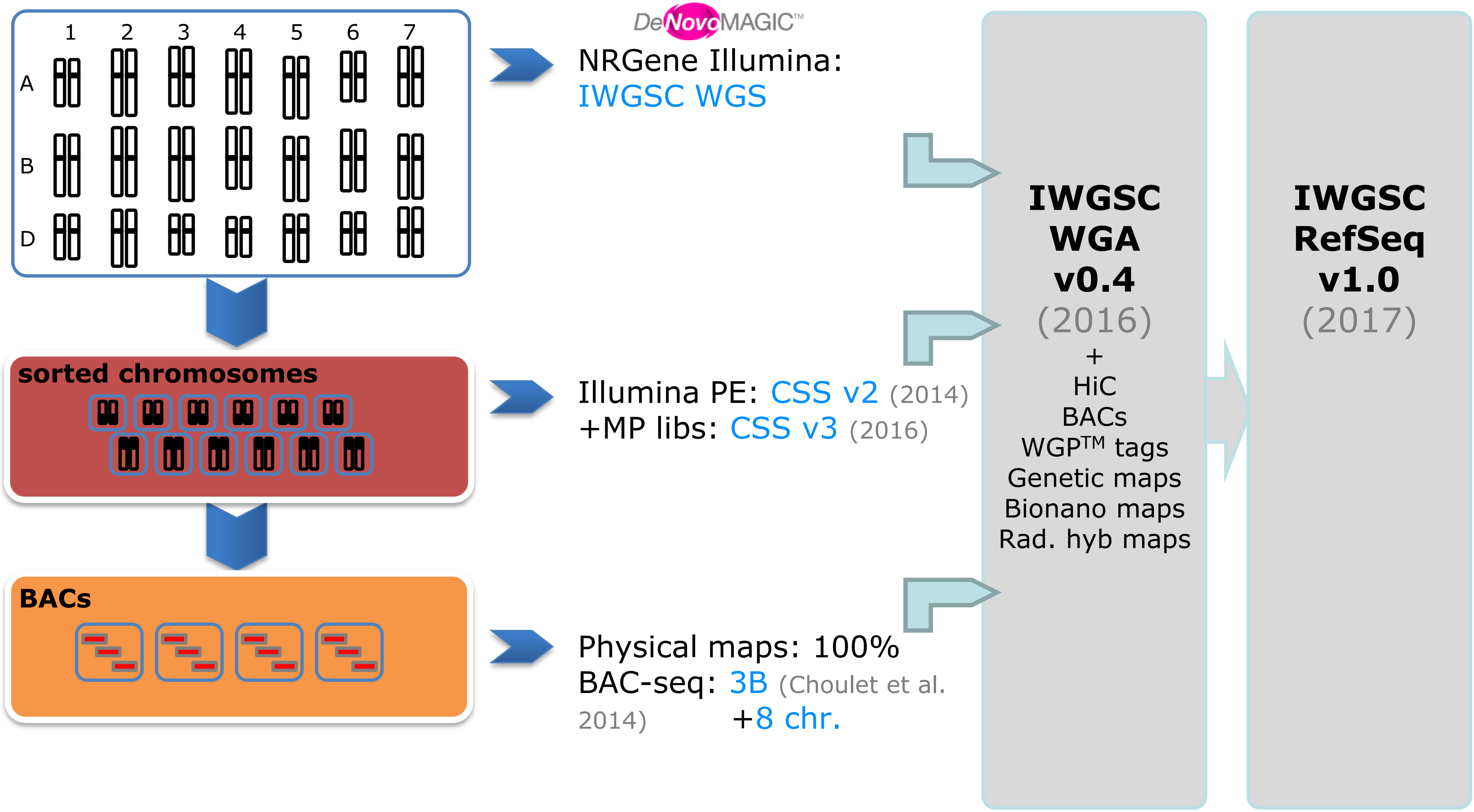
IWGSC RefSeq

IWGSC

IWGSC RefSeq

RefSeq v1.0

A fully annotated and anchored reference sequence assembly of the wheat genome



sorted chromosomes

BACs

DeNovoMAGIC™
NRGene Illumina:
IWGSC WGS

Illumina PE: CSS v2 (2014)
+MP libs: CSS v3 (2016)

Physical maps: 100%
BAC-seq: 3B (Choulet et al. 2014) +8 chr.

IWGSC
WGA
v0.4
(2016)
\+
HiC
BACs
WGP™ tags
Genetic maps
Bionano maps
Rad. hyb maps

IWGSC
RefSeq
v1.0
(2017)

❑ **IWGSC RefSeq v1.0** – *Community resources*

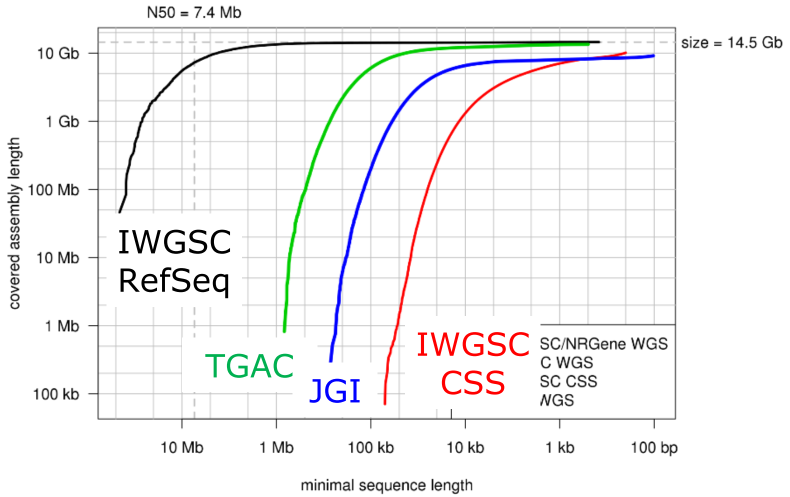
- Physical maps for all chromosomes
  - 1,839,128 BACs
- WGP tags (mostly from MTP BACs) for all chromosomes except 3B
  - 4,305,249 unique tags for 693,697 BACs
- BAC sequence assemblies
  - 8 chr: 1A, 1B, 3B, 3D, 6B, 7A, 7B, 7D + 2 chr arms: 4AL, 5BS
  - 52,890 BACs (9.7 Gb), N50=68 kb
- BioNano maps for 7A, 7B and 7DS
  - 1,335 BioNano contigs
- 4.7 M molecular markers

❑ **IWGSC RefSeq v1.0** – *Metrics*

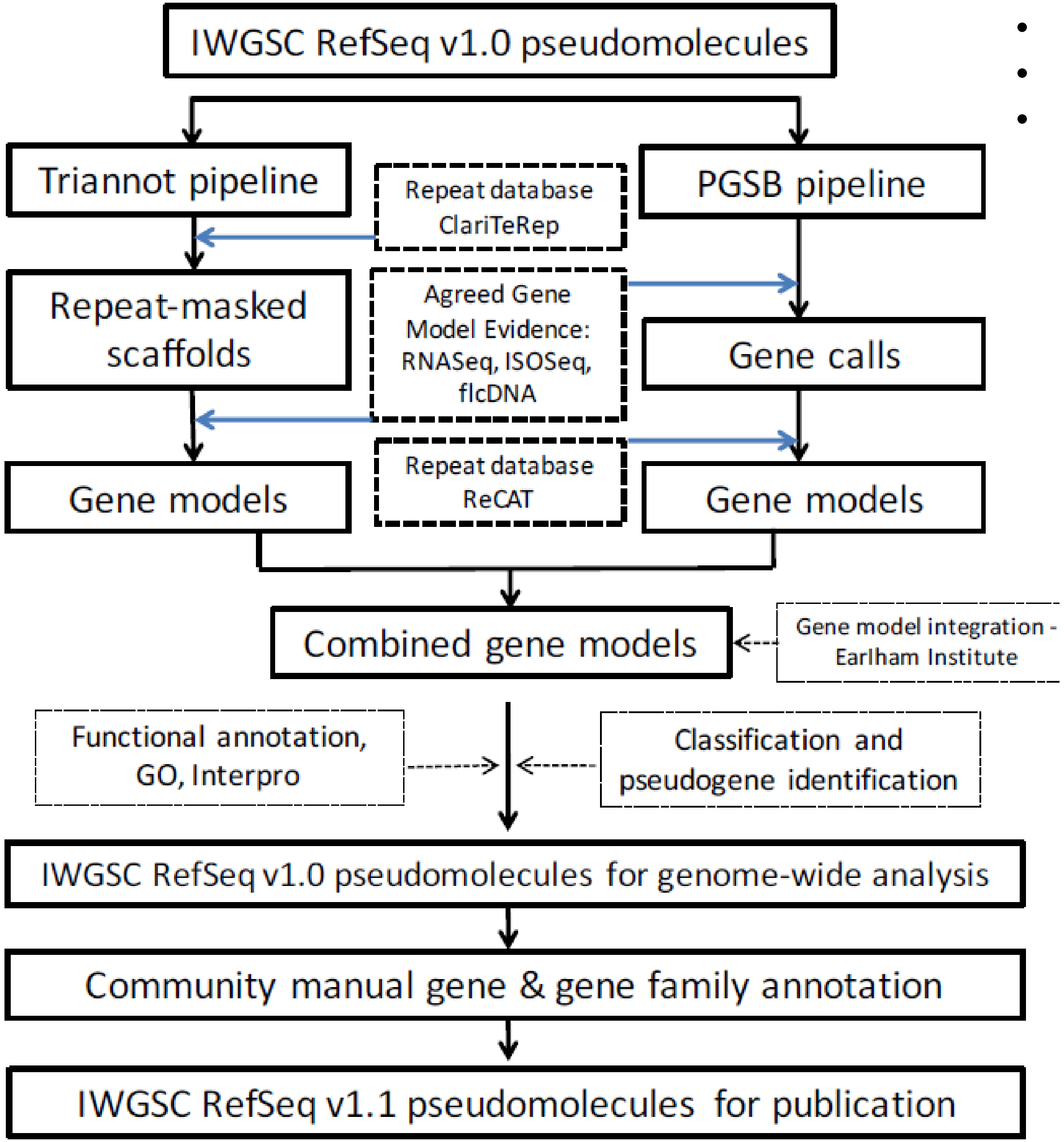
total size	14.5 Gb (genome size ~15.5 -> 94%)
completeness	97-99%
contig N50	52 kb
scaffold N50	7 Mb
superscaffold N50	23 Mb

21 pseudomolecules:

total size	14.1 Gb (97%)
superscaffolds	1601 (avg 76 per chr.)



❑ **IWGSC RefSeq v1.0** – *Genome annotation pipeline*



IWGSC RefSeq v1.0 pseudomolecules

Triannot pipeline

Repeat database ClariTeRep

PGSB pipeline

Repeat-masked scaffolds

Agreed Gene Model Evidence: RNASeq, ISOSeq, flcDNA

Gene calls

Gene models

Repeat database ReCAT

Gene models

Combined gene models

Gene model integration - Earlham Institute

Functional annotation, GO, Interpro

Classification and pseudogene identification

IWGSC RefSeq v1.0 pseudomolecules for genome-wide analysis

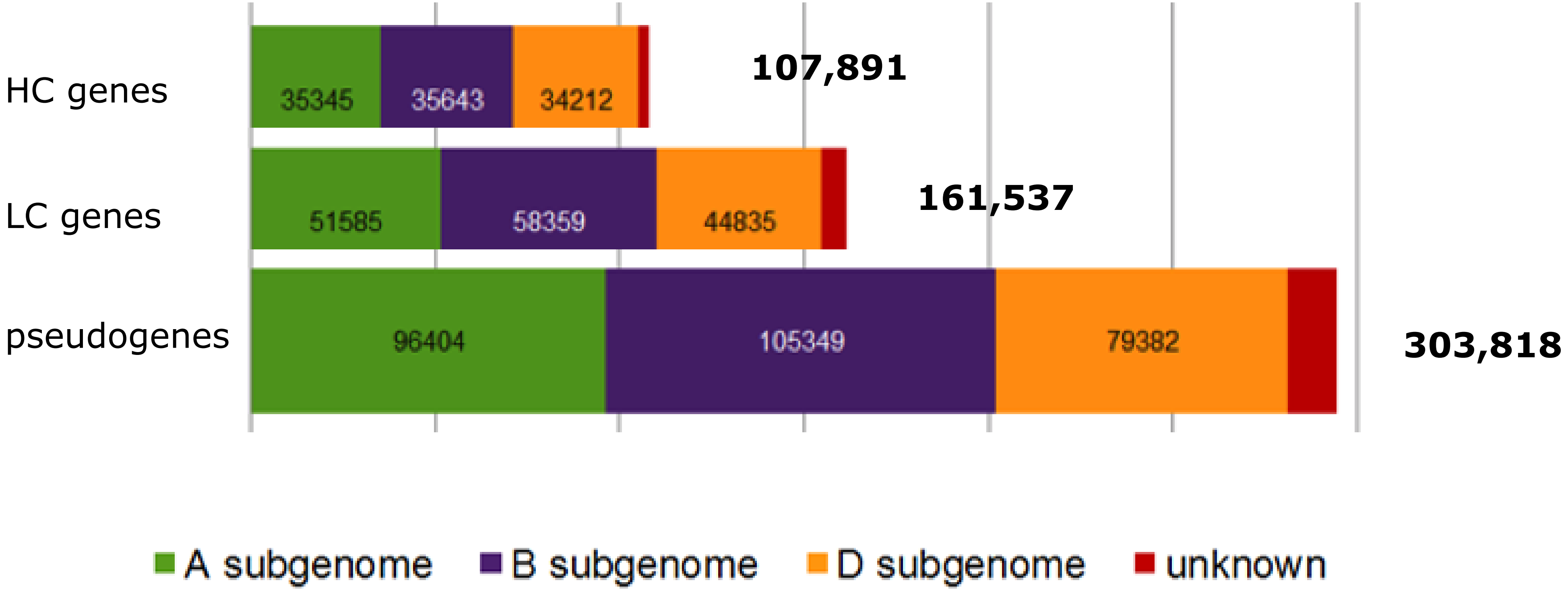
Community manual gene & gene family annotation

IWGSC RefSeq v1.1 pseudomolecules for publication

- **GDEC INRA** Clermont, France – *Rimbert Leroy Choulet et al.*
- **PGSB** Munich, Germany – *Spannagl Twardziok et al.*
- **Earlham Inst.** Norwich, UK – *Swarbreck Venturini et al.*

❑ Annotation – gene models

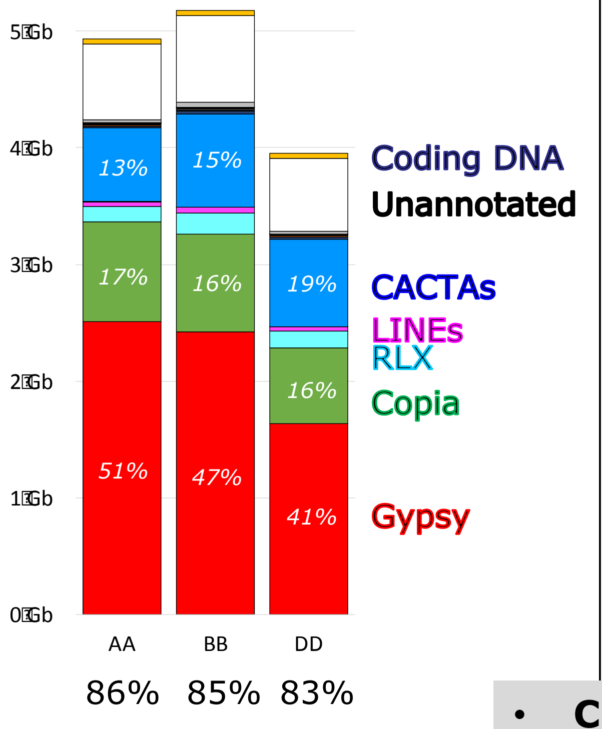
v1.0: automated
v1.1: incorporate 3685 manually annotated genes (available upon publication)

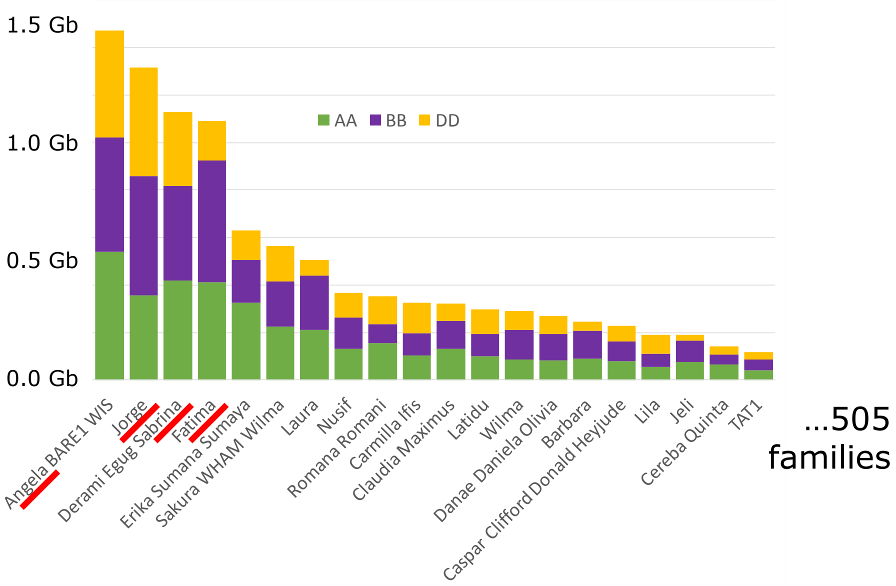


- *T. Wicker (UZH)*
- *H. Gundlach (PGSB)*
- *F. Choulet (GDEC)*

❑ Annotation – transposable elements

- CLARITE, TREP, *denovo* fl-LTR-RTs detection…
- 3,968,974 elements belonging to 505 families (84% <=> 11.9 / 14.1 Gb)





- **Conserved TE composition of A-B-D (no family specific to 1 subgenome)**
- **Complete reshuffling of TEs between homeologous loci**
- **TE activity after polyploidization very limited**

TE companion paper…

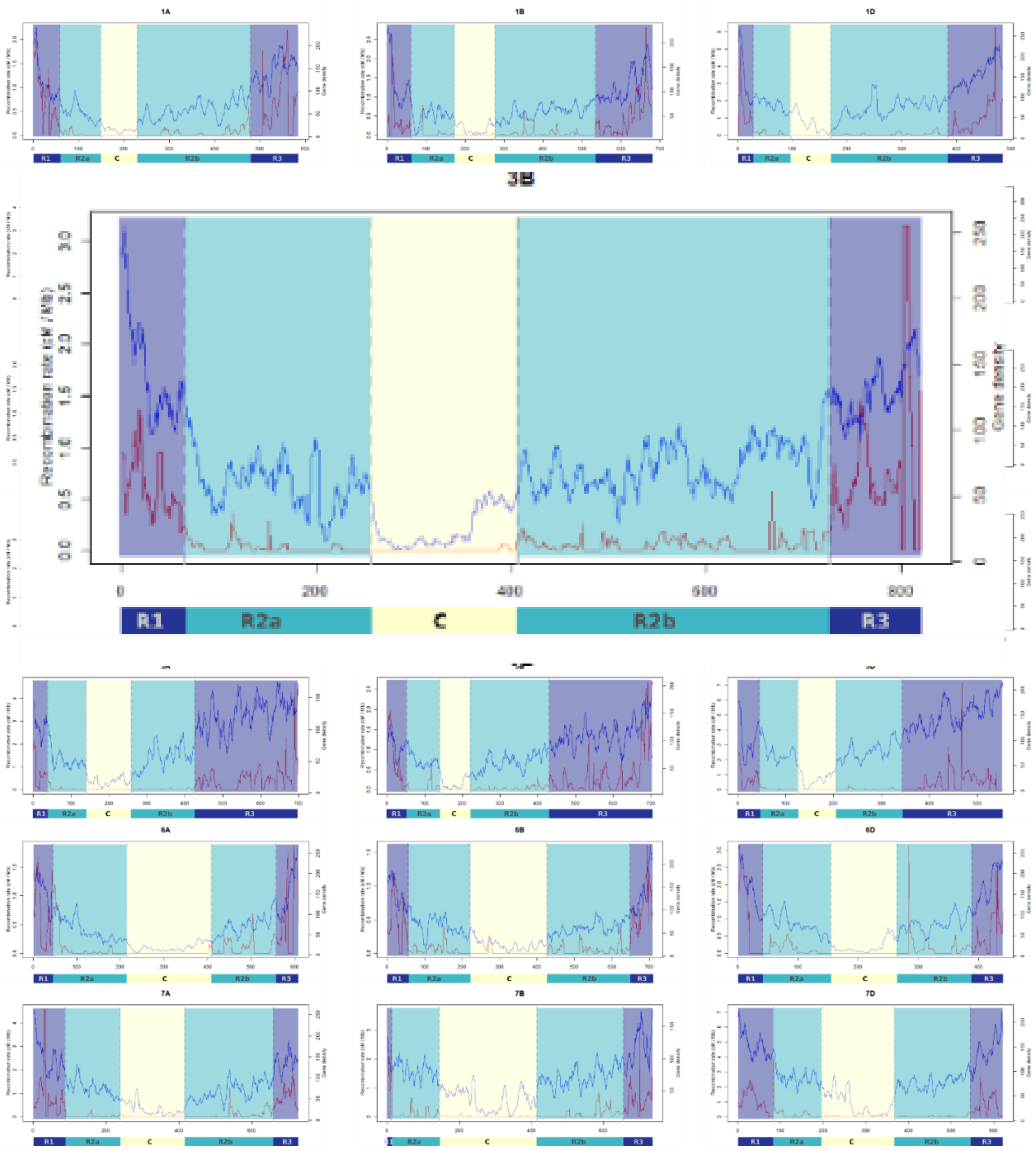
- *Etienne Paux (INRA GDEC)*

❑ Genome partitioning

R1/3=distal

C=proximal

R2a/b=interstitial



R1/3= high gene / low TE

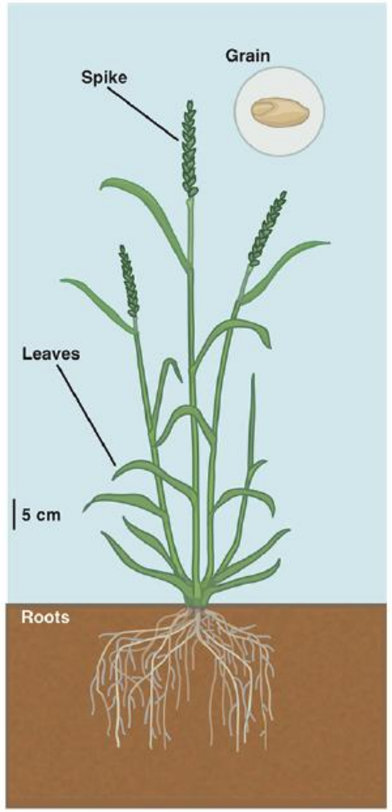
- *Daniel Lang (PGSB)*
- *Manuel Spannagl (PGSB)*

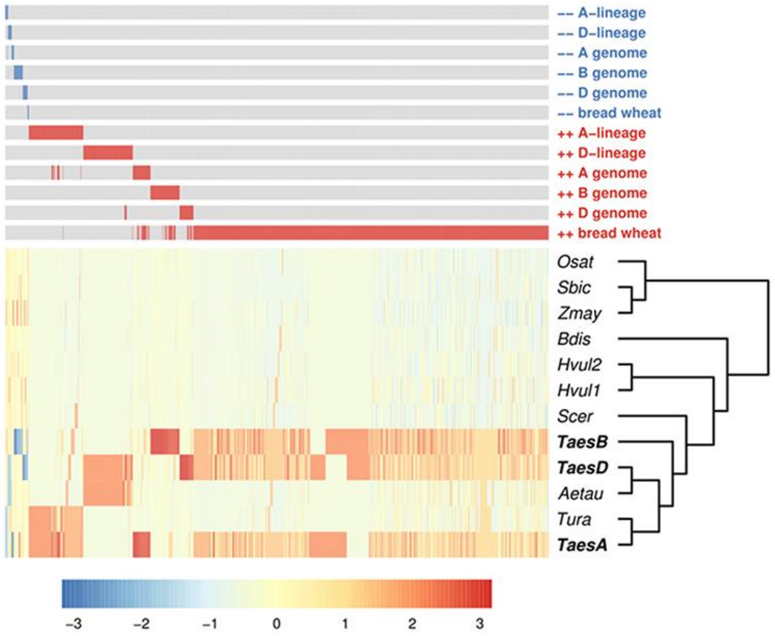
❑ A-B-D comparative analyses

Analysis of gene families of wheat

Gene family expansion/contraction associated with traits targeted by breeder selection

- 26080 gene families
- 33% expanded in the wheat lineage / 0.1% contracted families
  72% expanded in A+B+D





GO/PO/TO term enrichment and expanded gene families:

- **morphological traits** related to leaf, spike and root development, leaf size, tillering, vegetative growth and development, seed, endosperm, and embryo-development and morphology
- **grain yield** and **quality** (seed maturation, dormancy and germination)
- **abiotic stress** tolerance (salt stress, cold stress)

B sub-genome expanded families enriched for genes related to plant and leaf size and development, tillering and vegetative growth time:

predominant impact of the B genome on vegetative plant growth and development

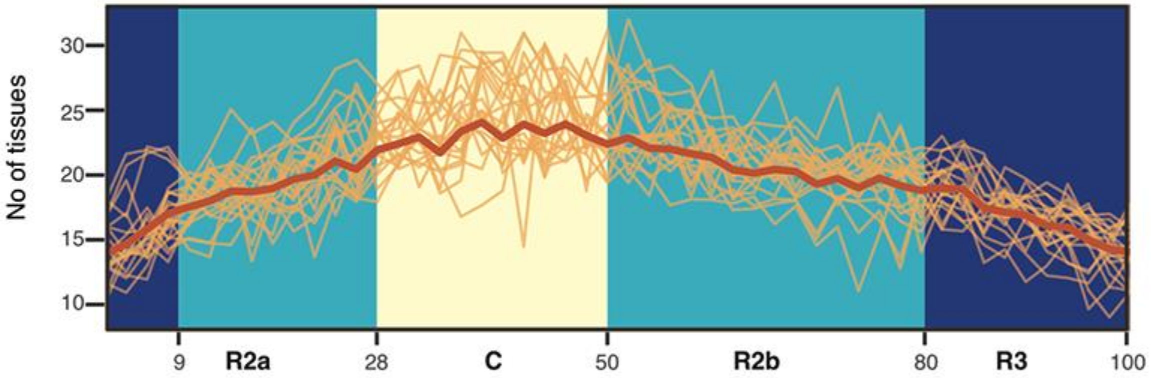
- *C. Uauy (JIC)*
- *P. Borrill*
- *R. Ramirez-Gonzalez*

❑ Transcriptome atlas

Atlas of transcription reveals trait associated gene co-regulation networks

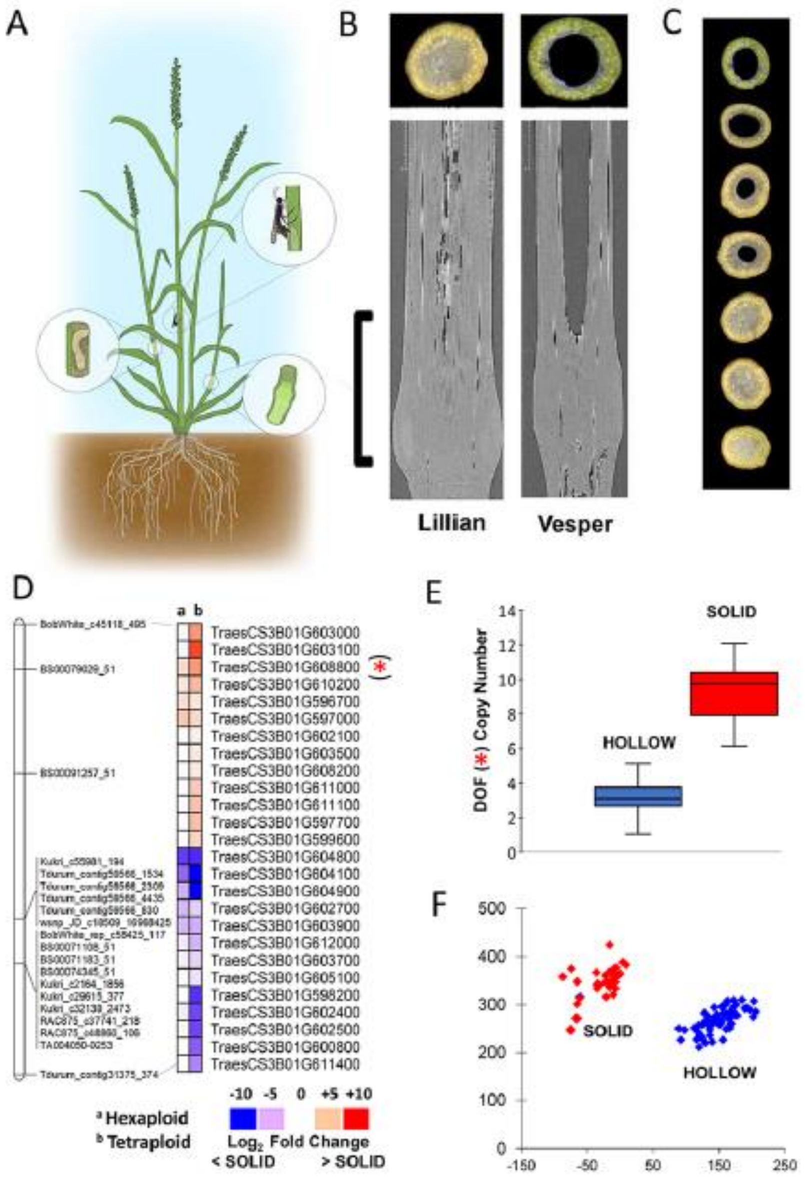
- 850 RNASeq samples – 32 tissues
  - 85% HC genes expressed
  - 49% LC genes expressed

- Chromosome partitioning
  - Distal region genes expressed in fewer tissues and enriched for responses to stress



Genes located in the distal R1 and R3 regions : lower expression breadth than those in the proximal regions

❑ Reference-guided dissection of an insect/abiotic stress resistance QTL for improved marker assisted selection



- traits of agronomic importance are inherited as QTL

- informative DNA markers in strong linkage disequilibrium with QTL are essential
- IWGSC RefSeq v1.0 provides template for assigning molecular markers to coordinates in linear DNA molecules anchored to high density molecular maps
- Example described : dissecting a QTL associated with insect/abiotic stress resistance
  - ➢ **Diagnosis marker**

- ➢ **Combining this knowledge with major loci selected for breeding programs provides a new framework for breeders to tackle the challenges of the new agriculture (balance the selection processes for adaptation to biotic and abiotic stress, end-use quality, and yield improvement)**

- *M. Alaux (INRA URGI)*

❑ IWGSC RefSeq Data Access & Availability

https://wheat-urgi.versailles.inra.fr/Seq-Repository/

Xwmc430

Examples: Xwmc430, QTL, TaeCsp3B

Pre-publication data access under Toronto Agreement:

- IWGSC **WGA v0.4**: June 2016
- IWGSC **RefSeq v1.0**: January 2017
- IWGSC **Annotation v1.0**: May 2017
- IWGSC **Annotation v1.1**: upon publication
- Manuscript under review

http://www.wheatgenome.org/Tools-and-Resources

IWGSC Phase II:

What’s Next for the IWGSC

WHEATS&WOMEN INTERNATIONAL CONFERENCE
Roma – June 2018

Measuring the Impact of IWGSC RefSeq v1.0

476,595 BLAST searches
22,935 downloads

2017

> 100 papers

Genome structure (9)

Genome expression (4)

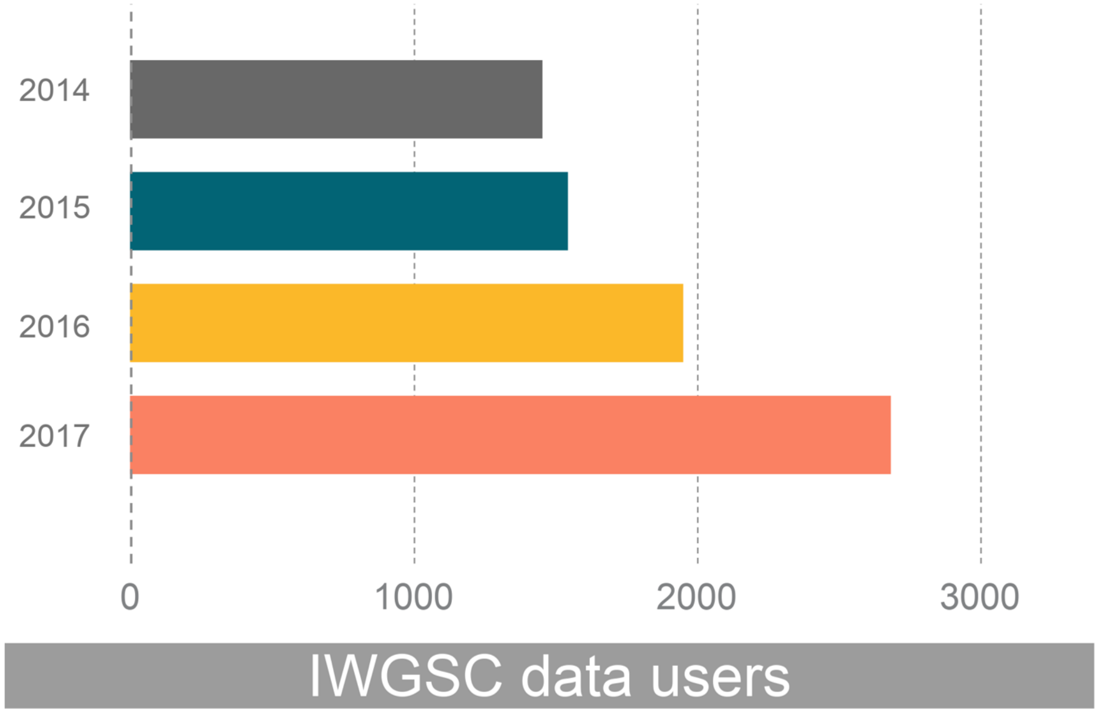
Genetic diversity (18)

Trait genomics (67)

Genome enabled breeding (9)

# of papers referencing use of IWGSC RefSeq related resources since January 2017

Measuring the Impact of IWGSC RefSeq v1.0



Progress

2005

Vision

- High quality annotated genome sequence, comparable to rice
- Physical map-based, integrated and ordered sequence

MISSION ACCOMPLISHED

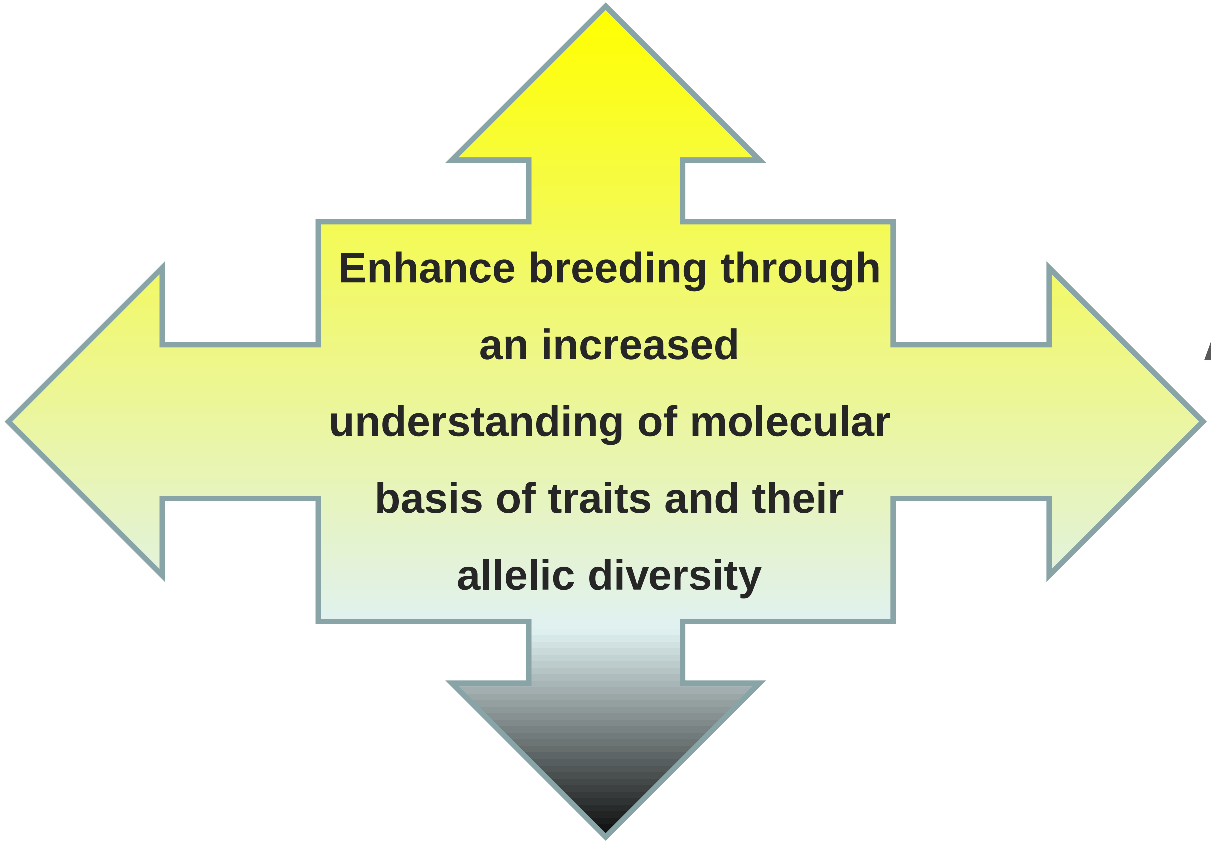
Looking into the Future

- **Achieve Gold Standard Reference**
- **Reach beyond reference to ensure a full genome-sequence based toolbox for wheat**

Goal: Accelerate wheat improvement

Vision and strategy for the IWGSC 2.0

Reference sequences with high quality functional annotations



Public, user-friendly, integrated databases and tools for all users

Enhance breeding through an increased understanding of molecular basis of traits and their allelic diversity

A wheat Pan-genome representing the worldwide diversity

An international, well connected and coordinated community

IWGSC

From RefSeq v1.0 to IWGSC Gold Standard

IWGSC RefSeq + **Manual & Functional Annotation** = IWGSC Gold Standard

- Pipeline for community generated manual annotation with future annotation releases
- Functional annotation
  - Gene family leaders
  - Applying machine learning techniques
- Closing gaps by incorporating new resources to build a IWGSC RefSeq v2.0
- IWGSC leaders: Rudi Appels, Fred Choulet…..

IWGSC 2.0 Activities: Pan-Genome

- Develop a wheat pan-genome that represents the breadth of worldwide wheat diversity
- Reference quality for 8-12 landraces
- Diversity panels at draft quality
- Haplotype database
- Skim-sequencing and high throughput genotyping of many lines
- IWGSC team: Etienne Paux, Sébastien Praud,…..

IWGSC 2.0 Activities: Database

- Develop user-friendly, integrated databases and tools

- IWGSC Pan-genome Database Task Force
  - Fred van Ex, Bayer CropScience
  - Magalie Leveugle, Biogemma
  - Matthieu Conte, Syngenta
  - Michael Alaux, URGI-INRA
  - …….

- What tools do you want?

IWGSC 2.0 Activities: IWGSC Exome Array

- Wheat exome capture array based on IWGSC RefSeq annotation v1.1.
- Arbor Biosciences will offer enrichment reagents and an enrichment service
- Exome Team: Burkhard Steuernagel, Sreya Ghosh, Sébastien Praud, Hikmet Budak, Etienne Paux, Ute Bauhman
- Discussion: What modules would you like to see?

❑ Conclusion

- IWGSC goal for high quality reference achieved
- IWGSC RefSeq v1.0 incorporates highly diverse community resources (WGS assembly + Hi-C links + physical maps + CSS + BAC sequences + millions of markers)
- Draft v0.4 + pseudomolecules v1.0 + annotation v1.0 available for download prior to publication
- IWGSC Phase II
  - Continued improvements from manual and functional annotation
  - IWGSC Exome Array based on the sequence
  - Pan-genome that covers the breadth of wheat diversity (landraces and elite var.)
  - Pan-genome database, visualization, and tools for breeders

❑ Acknowledgments: Leaders & Major Contributors

***IWGSC RefSeq PIs*:** **Rudi Appels, Kellye Eversole, Nils Stein, Jane Rogers, Catherine Feuillet, Beat Keller**

IWGSC Whole Genome Assembly PIs: **Curtis Pozniak, Nils Stein**, Frédéric Choulet, Assaf Distelfeld, Kellye Eversole, Jesse Poland, Jane Rogers, Gil Ronen and Andrew G. Sharpe

Whole Genome Sequencing and Assembly: **Curtis Pozniak, Gil Ronen, Nils Stein**, Omer Barad, Kobi Baruch, Frédéric Choulet, Gabriel Keeble-Gagnère, Martin Mascher, Andrew G. Sharpe and Gil Ben-Zvi

Hi-C Data Based Scaffolding: **Nils Stein** and **Martin Mascher**

Whole Genome Assembly QC & Analyses: **Frédéric Choulet, Gabriel Keeble-Gagnère, Martin Mascher** and **Jane Rogers**

Pseudomolecule Assembly: **Frédéric Choulet, Gabriel Keeble-Gagnère** and **Martin Mascher**

RefSeq Genome Structure and Gene Analyses: **Manuel Spannagl**, **Frédéric Choulet** and Daniel Lang

Automated Annotation: **Frédéric Choulet, Manuel Spannagl, David Swarbreck** and Hélène Rimbert

Manual Gene Curation: **Rudi Appels** and Hélène Rimbert

Sub-Genome Comparative Analyses: **Frédéric Choulet, Manuel Spannagl** and Daniel Lang

Transposable Elements: **Frédéric Choulet**, Thomas Wicker and Heidrun Gundlach

Phylogenomic Analyses: **Daniel Lang** and **Manuel Spannagl**

Transcriptome Analyses & RNASeq Data: **Cristobal Uauy**, Philippa Borrill and Ricardo H. Ramirez-Gonzalez

Whole Genome Methylome: **Curtis Pozniak, Stephen Robinson** and **Andrew G. Sharpe**

Histone Mark Analyses: **Moussa Benhamed and Etienne Paux**

BAC Chromosome MTP IWGSC-Bayer Whole Genome Profiling(WGP^{TM}) Tags: **Jane Rogers and John Jacobs**

Chromosome LTC Mapping & Physical Mapping Quality Control: **Abraham Korol**, Zeev Frenkel and Tzion Fahima

RH Mapping: **Vijay Tiwari**

Optical Mapping: **Jaroslav Doležel**

Recombination Analyses: **Pierre Sourdille**

Overall Gene Family Analyses: **Rudi Appels, Manuel Spannagl** and Daniel Lang

CBF gene family: **Delfina Barabaschi**

Dehydrin gene family: **Pilar Hernandez** and **Sergio Galvez**

❑ Acknowledgments: Leaders & Major Contributors

NLR gene family: **Burkhard Steuernagel**

PPR gene family: **Ian Small** and Joanna Melonek

Prolamin gene family: **Angéla Juhász** and Tatiana Belova

WAK gene family: **Kostya Kanyuka** and Robert King

Genome Size Analysis: **Jaroslav Doležel**

MicroRNA and tRNA annotation: **Hikmet Budak**

Sst Team: **Kirby Nilsen, Sean Walkowiak** and **Curtis Pozniak**

Genetic Maps and Mapping: Gary Muehlbauer and Jesse Poland

BAC libraries and Chromosome Sorting: **Jaroslav Doležel** and Hana Šimková

BAC Pooling, BAC library Repository, and Access: **Hélène Berges**

IWGSC Sequence & Data Repository and Access: Michael Alaux

1A BAC Sequencing & Assembly: Curtis Pozniak, Andrew G. Sharpe, and **Sean Walkowiak**

1B BAC Sequencing & Assembly: Frédéric Choulet and Etienne Paux

1D, 4D, 6D Physical Mapping: Bikram Gill

2AL Physical Mapping: Kuldeep Singh

2AS Physical Mapping: Nagendra K. Singh and **Jitendra P. Khurana**

2B, 2D, 4B, 5BL & 5DL IWGSC-Bayer Whole Genome Profiling(WGP™) Physical Maps: Jane Rogers and John Jacobs

3AL Physical Mapping: Bikram Gill

3DS Physical Mapping & BAC Sequencing & Assembly: Jan Bartoš

4A Physical Mapping, BAC Sequencing, Assembly, & Annotation: Miroslav Valárik

5BS BAC Sequencing, & Assembly: Elena Salina, Nikolai Ravin and **Konstantin Skryabin**

6B BAC Sequencing & Assembly: Hirokazu Handa

7A Physical Mapping & BAC Sequencing: Rudi Appels

7B Physical Mapping, BAC Sequencing, & Assembly: Odd-Arne Olsen and **Tatiana Belova**

7DS BAC Sequencing & Assembly: Hana Šimková

7DL Physical Mapping & BAC Sequencing: Song Weining

❑ Acknowledgments: Funders

ANR

DFG Deutsche Forschungsgemeinschaft

Australian Government
Grains Research and Development Corporation

GRDC
GRAINS RESEARCH & DEVELOPMENT CORPORATION

NMBU
Norwegian University of Life Sciences

Agriculture and Agri-Food Canada
Agriculture et Agroalimentaire Canada

DEPARTMENT OF BIOTECHNOLOGY
Ministry of Science & Technology

illumina

NRC·CNRC

European Commission

INRA
SCIENCE & IMPACT

Alberta Wheat COMMISSION
GROWTH.
INNOVATION.
RESULTS.

The Research Council of Norway

USDA
United States Department of Agriculture

National Institute of Food and Agriculture

FranceAgriMer

FSOV

ISRAEL SCIENCE FOUNDATION

CTAG²
Canadian Triticum Applied Genomics

GENTYANE
PLATE-FORME de GENOTYPAGE et SÉQUENÇAGE en AUVERGNE

NSF

Australian Government
Department of Industry, Innovation and Science

JUNTA DE ANDALUCIA

GenomeCanada

AUSTRALIA-CHINA
SCIENCE AND RESEARCH FUND

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

RSF - DFG

Government of Saskatchewan

BAYER

BBSRC
bioscience for the future

GenomePrairie

GAČR
CZECH SCIENCE FOUNDATION

Federal Ministry of Food and Agriculture

MAFF
Ministry of Agriculture, Forestry and Fisheries
農林水産省

日清製粉グループ
Nisshin Seifun Group

Sask Wheat
DEVELOPMENT COMMISSION

Computing infrastructure for Science

crea
Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria

Federal Ministry of Education and Research

MONTANA STATE UNIVERSITY

Western Grains Research Foundation

IWGSC RefSeq

IWGSC

IWGSC Sponsors

biogemma

BAYER Bayer CropScience

CIMMYT
International Maize and Wheat Improvement Center

DOW

DUPONT

FLORIMOND DESPREZ

INRA
SCIENCE & IMPACT

Institute of Experimental Botany of the AS CR, v. v. i.

Limagrain
from earth to life

KANSAS WHEAT
Rediscover Wheat

RAGT

syngenta

THE UNIVERSITY of ADELAIDE

I W G S C