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)

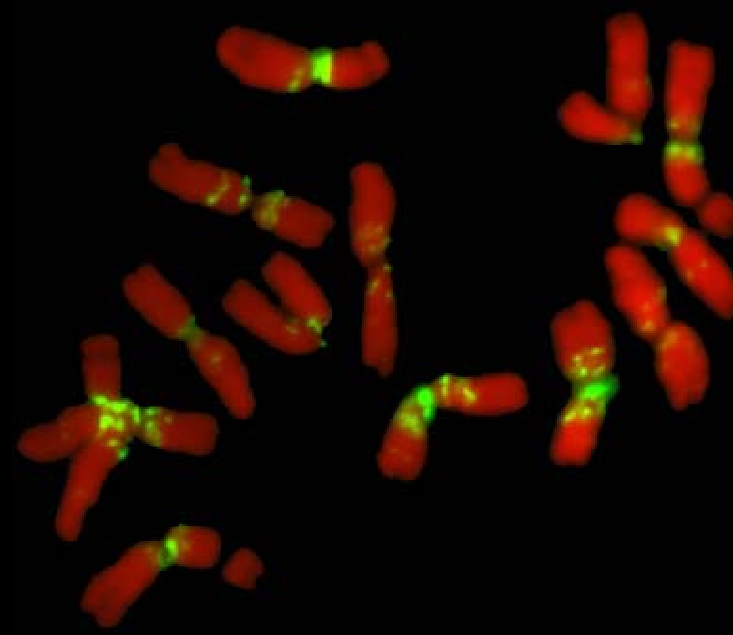
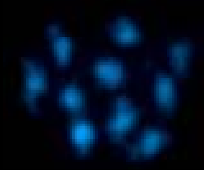


image: A Houben

A.thaliana



0.15 Gb

Triticum aestivum

17 Gb (40 x rice)

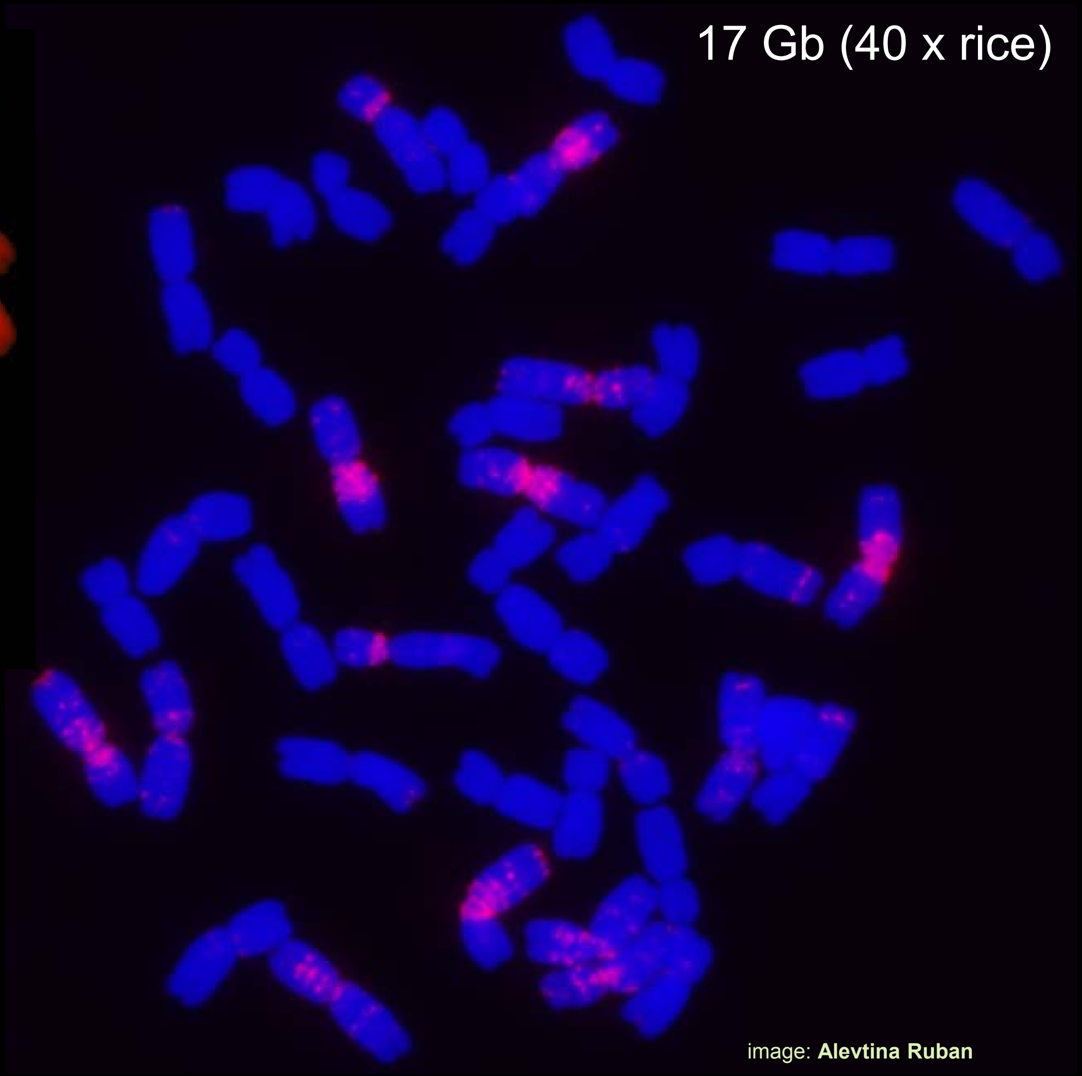
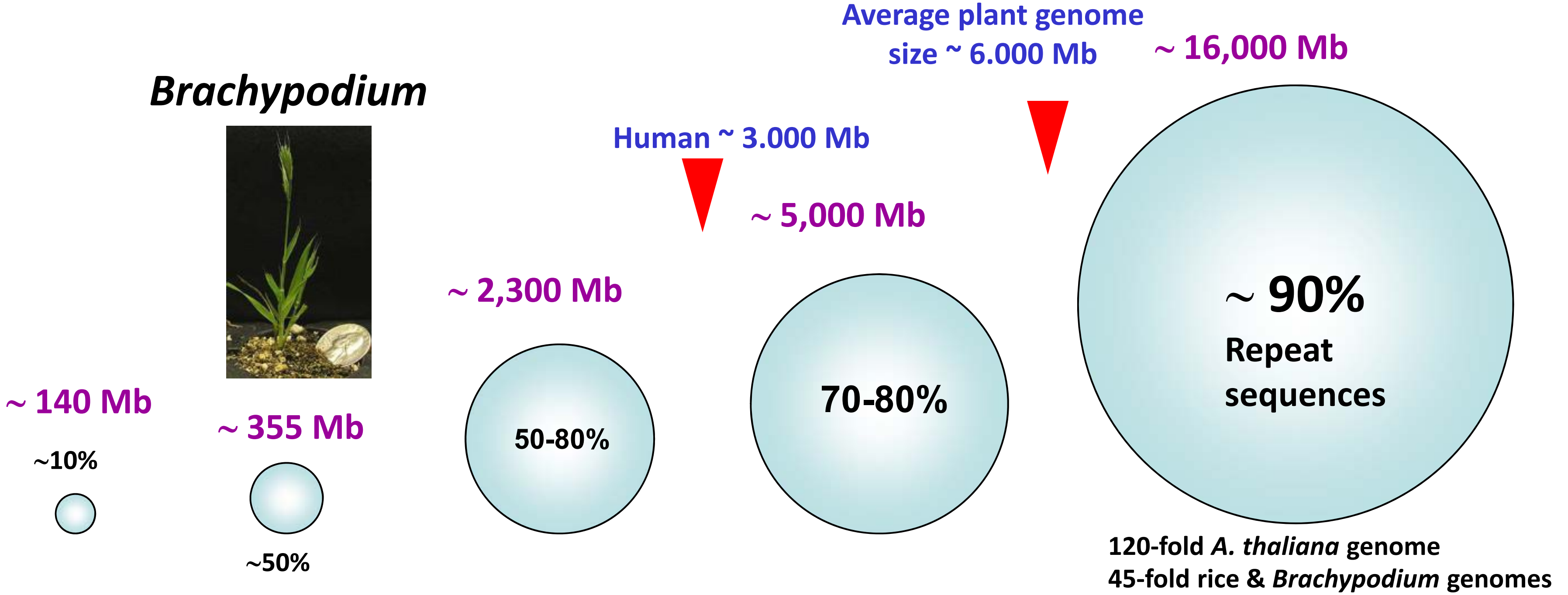


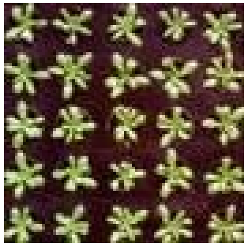
image: **Alevtina Ruban**

Challenging Bread Wheat Genome



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

A. thaliana **(2x)**



Rice(2x)

Maize (2x)

Barley (2x)

Bread wheat (6x)

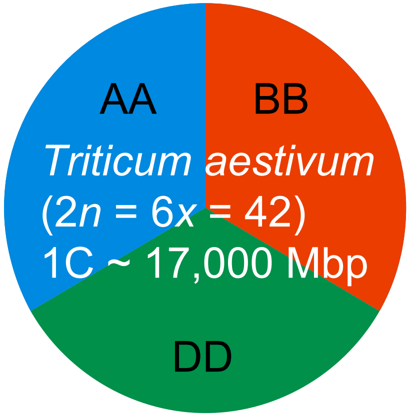
Wheat is a challenge for genomic studies & sequencing

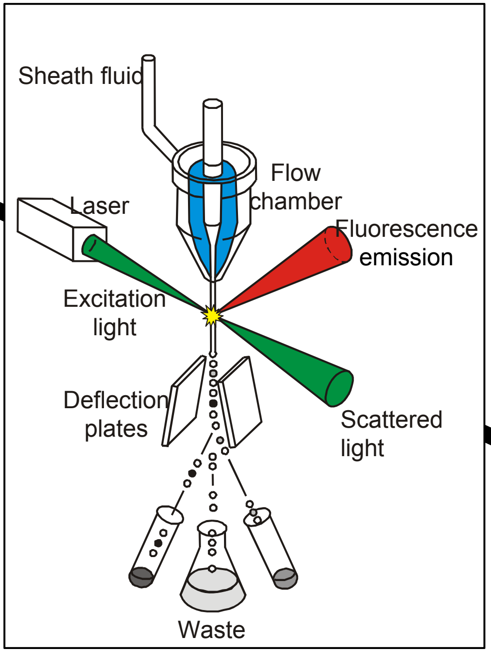
Courtesy: Catherine Feuill

Wheat genome sequencing in 2005

- Too big genome for Sanger Sequencing
- Too big task for single group/lab
- Too costly

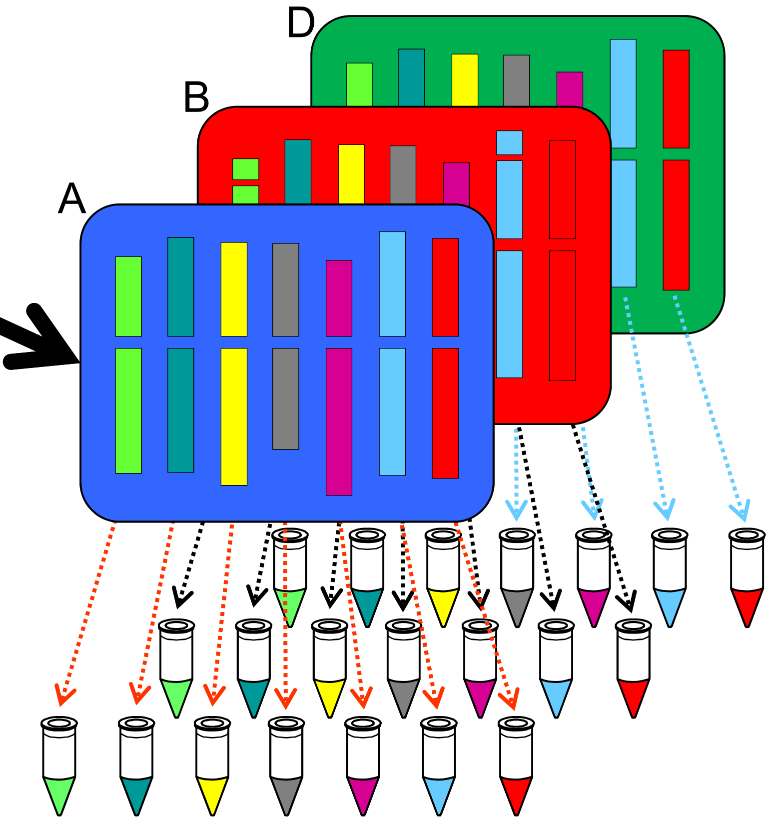
A chromosome-based approach





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

Dissection of the genome to single chromosomes (arms) representing individual (sub)genomes

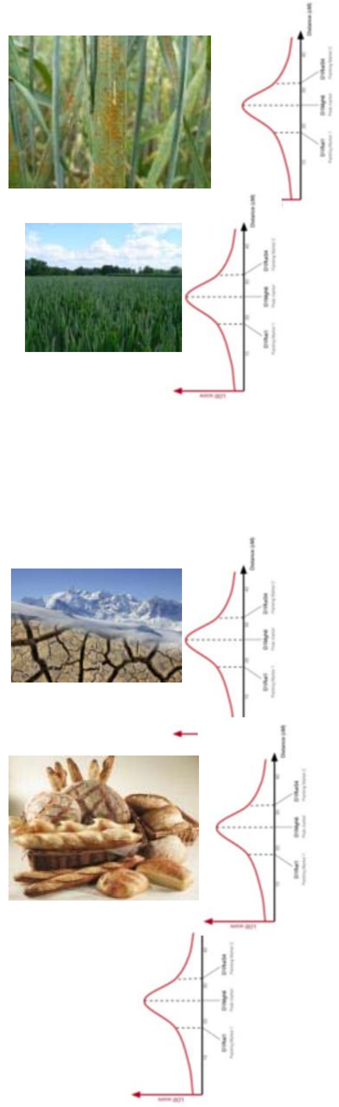


- Chromosomes: 605 - 995 Mbp (3.6 – 5.9% of the genome)
- Chromosome arms: 225 - 585 Mbp (1.3 – 3.4% of the genome)

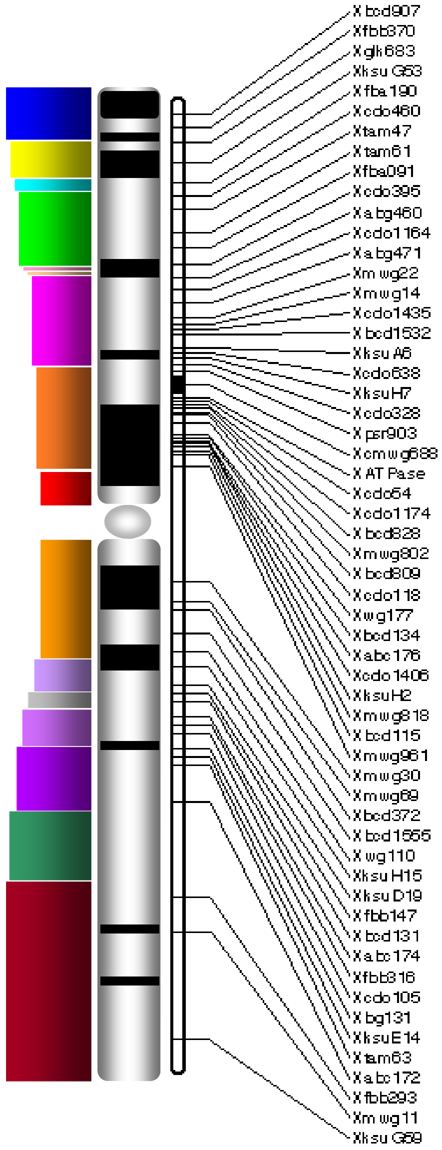
- **Chromosome specific BAC libraries (June 2012)**
- **Amplified DNA for chromosome survey (Nov 2011)**

An integrated and ordered 3B reference sequence

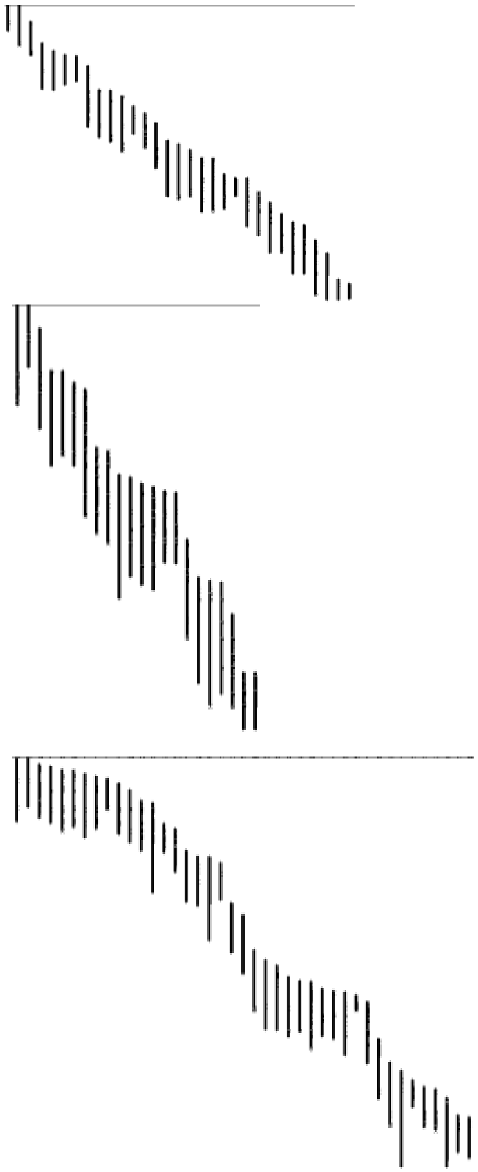
MetaQTL analysis



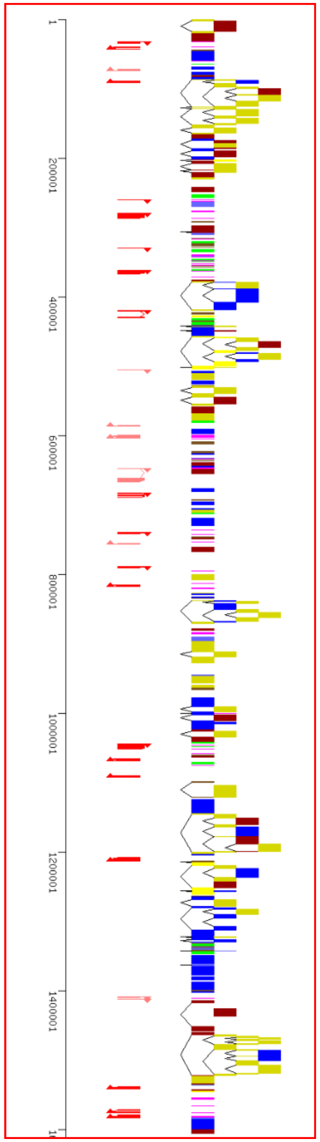
3B consensus map (5000 markers)



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

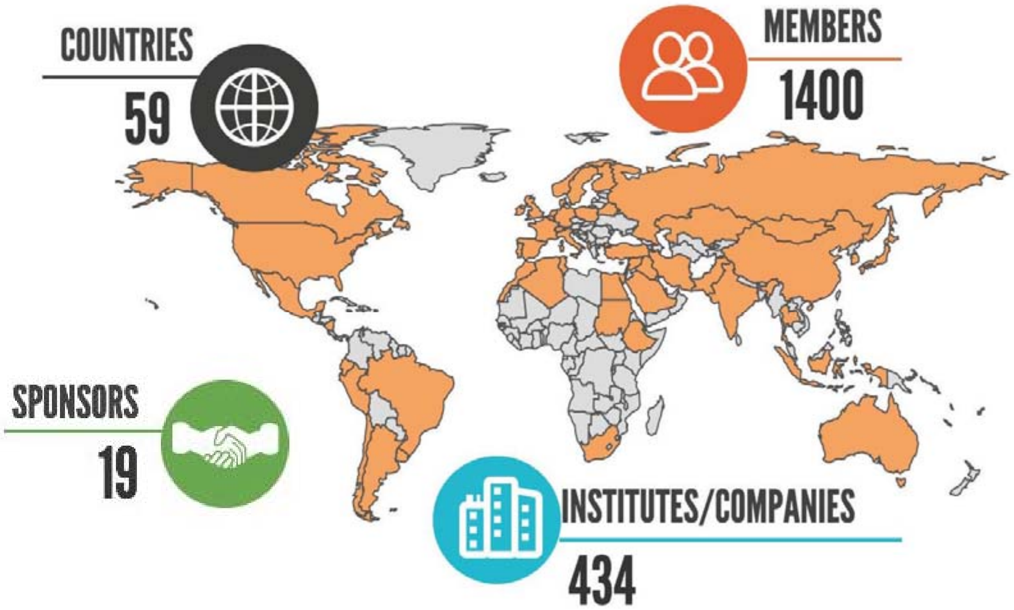
Role: Scientific strategy

Members: Sponsors & Leaders of IWGSC Projects & Initiatives

Members

Role: Participation & Input

Members: Open to Anyone



Arcadia Biosciences
ACPFG
biogemma
Bayer CropScience
CSIRO
CIMMYT
Céréales Vallée
DUPONT The miracles of science
FLORIMOND DESPREZ
INRA SCIENCE & IMPACT
Institute of Experimental Botany of the AS CR, v. v. i.
KWS
KeyGene
Limagrain from earth to life
Together, we are KANSAS WHEAT
MONSANTO
UNIVERSITY OF SASKATCHEWAN Crop Development Centre
RAGT
syngenta
CTAG²

IWGSC

www.wheatgenome.org

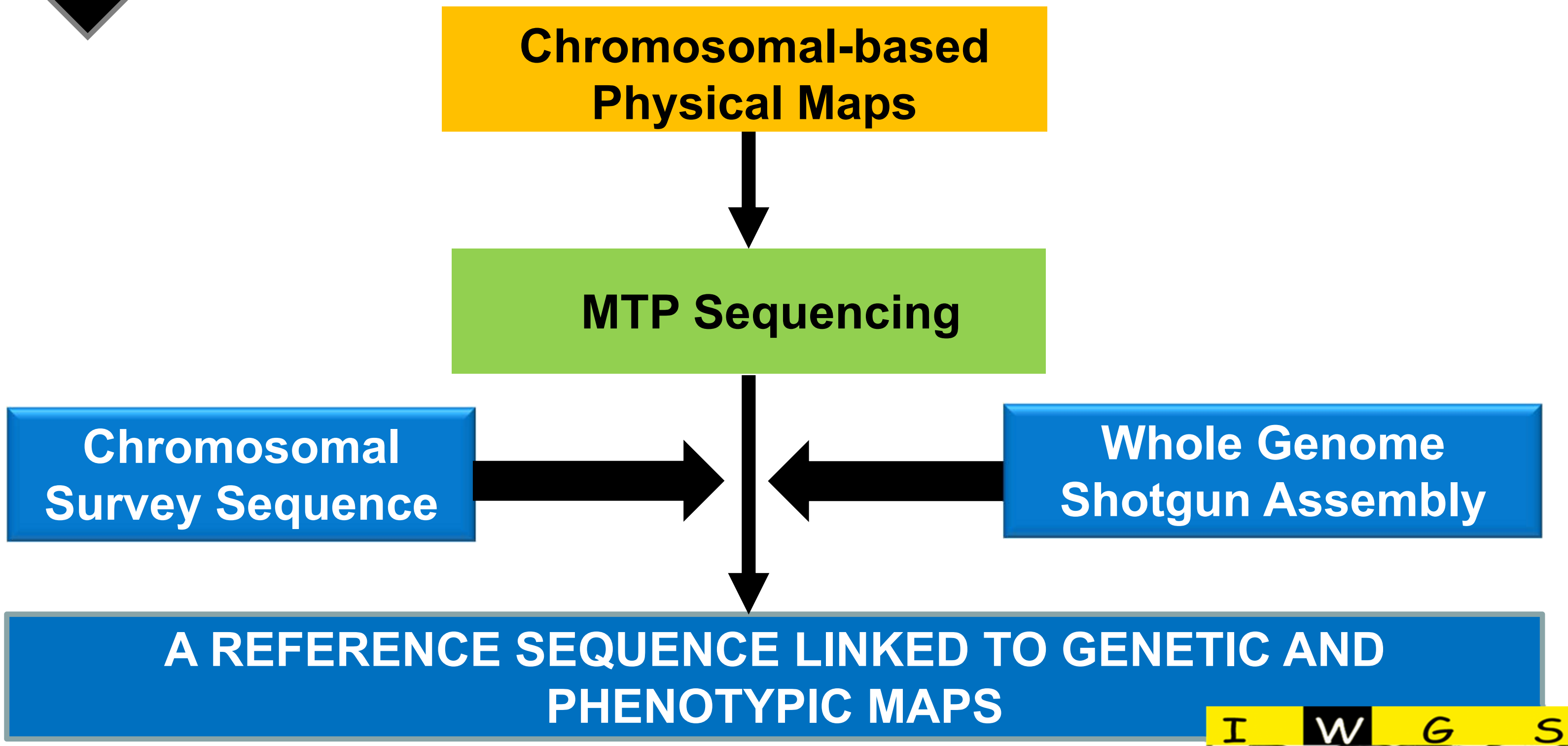
Vision and roadmap

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-based Physical Maps

MTP Sequencing

Chromosomal Survey Sequence

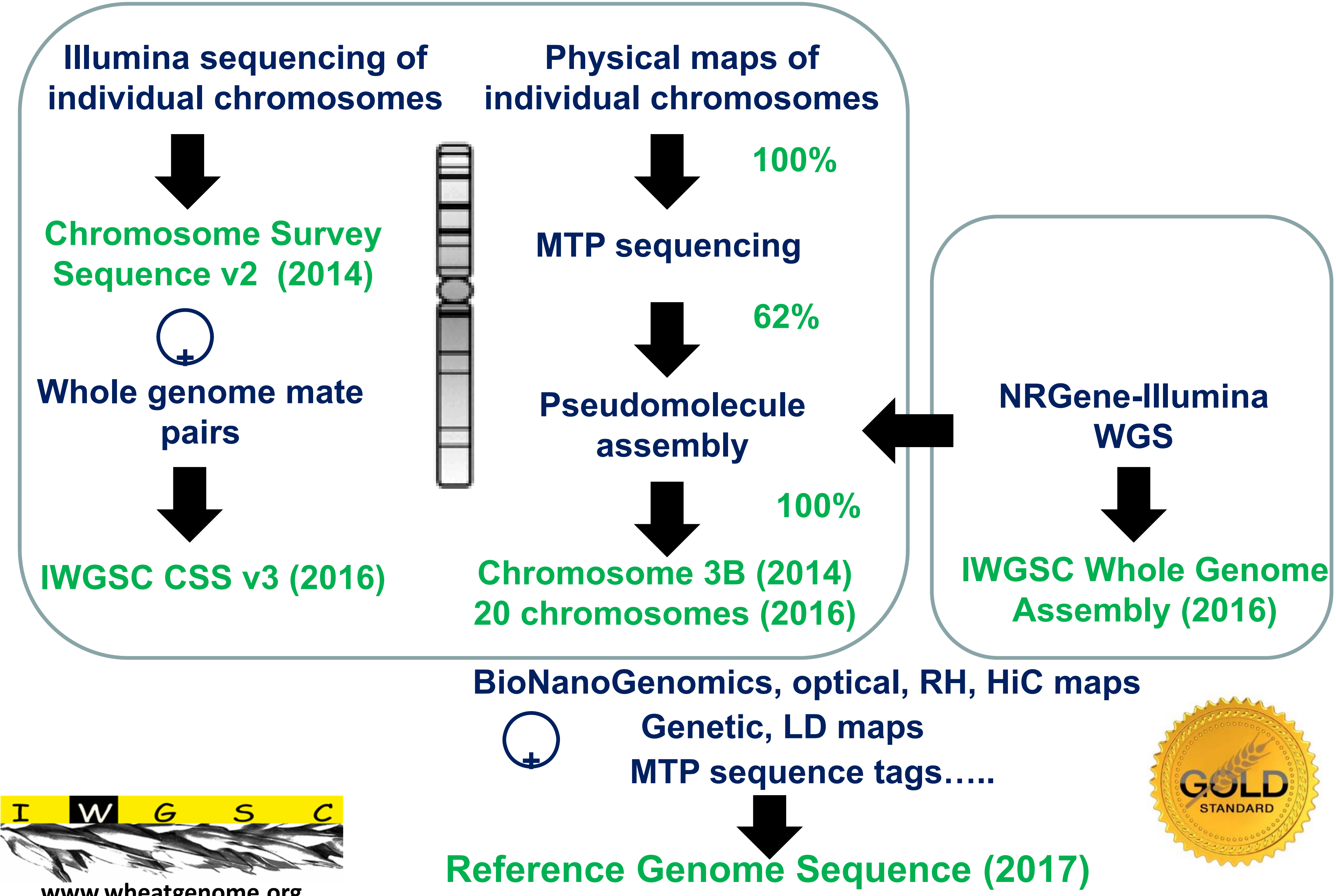
Whole Genome Shotgun Assembly

A REFERENCE SEQUENCE LINKED TO GENETIC AND PHENOTYPIC MAPS

GOLD STANDARD

IWGSC

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

100%

MTP sequencing

62%

Pseudomolecule assembly

100%

Chromosome 3B (2014)
20 chromosomes (2016)

NRGene-Illumina WGS

IWGSC Whole Genome Assembly (2016)

BioNanoGenomics, optical, RH, HiC maps

⊕

Genetic, LD maps

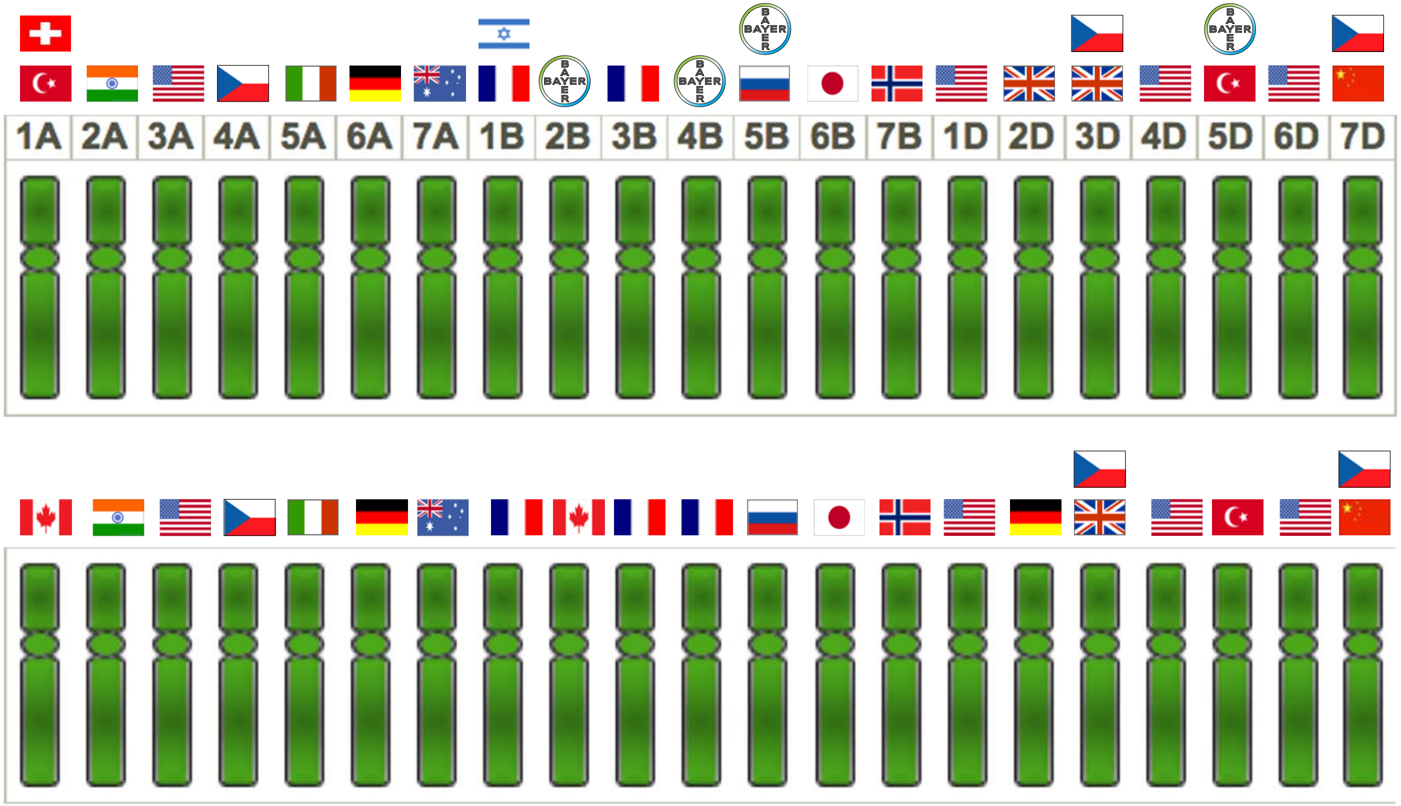
MTP sequence tags…..

Reference Genome Sequence (2017)

I W G S C

www.wheatgenome.org

Progress towards completion of Bread Wheat Reference Genome Sequence



Physical maps

IWGSC Reference Pseudo-molecule*

All physical maps available at IWGSC repository at

All pseudo-molecules completed

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

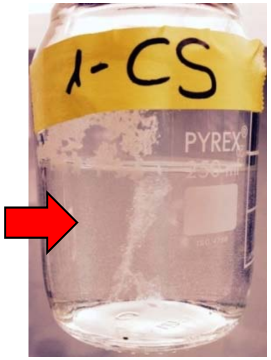
www.wheatgenome.org

IWGSC

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

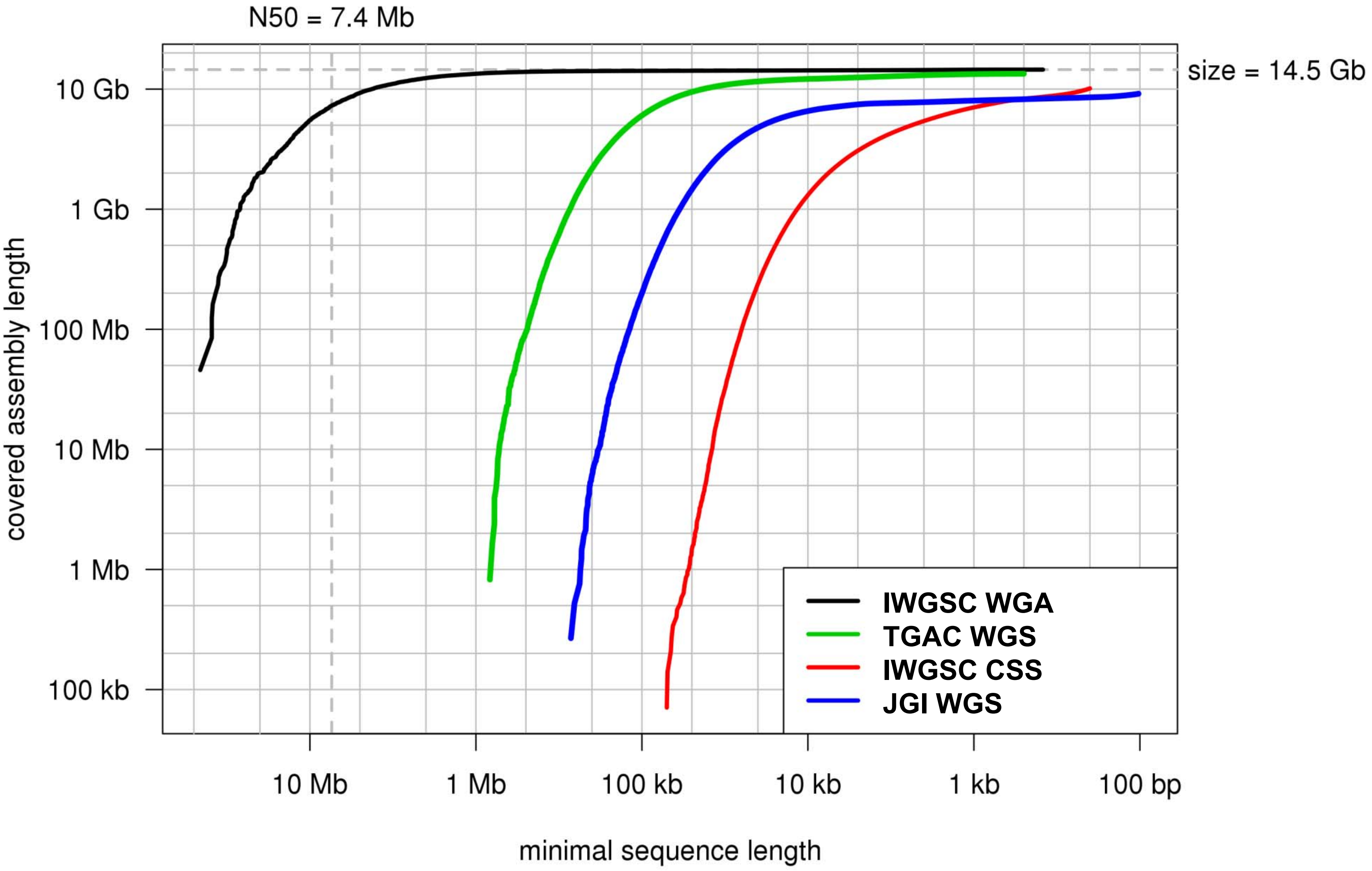
IWGSC Whole Genome Assembly Project

De novo assembly:

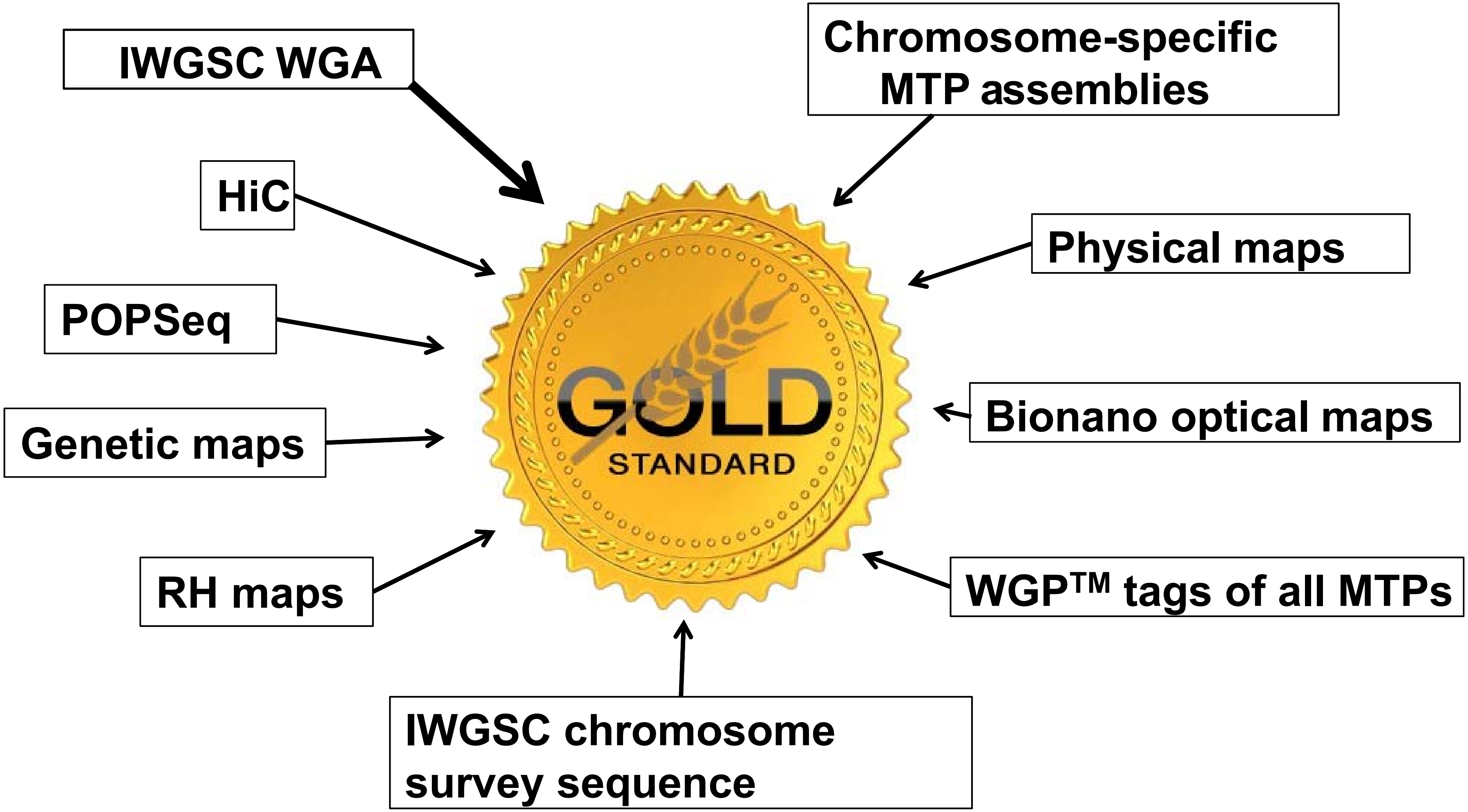
- NRGene's DeNovoMagic-2 platform, total run time < 3 weeks, 1Tb RAM computer
- illumina short-reads sequencing data only (x200 coverage)

Assembly size:	14.5 Gbp
Gaps size:	262 Mbp
Gaps %:	1.80
Total # scaffolds:	138,484
L50:	7.1Mbp
N50 (#sequences):	566
L90:	1.3 Mbp
N90 (#sequences):	2,363
MAX Scaffold:	45.8 Mbp

WGA Assembly Statistics



Concerted integration of resources: RefSeq



IWGSC WGA

Chromosome-specific MTP assemblies

HiC

Physical maps

POPSeq

Bionano optical maps

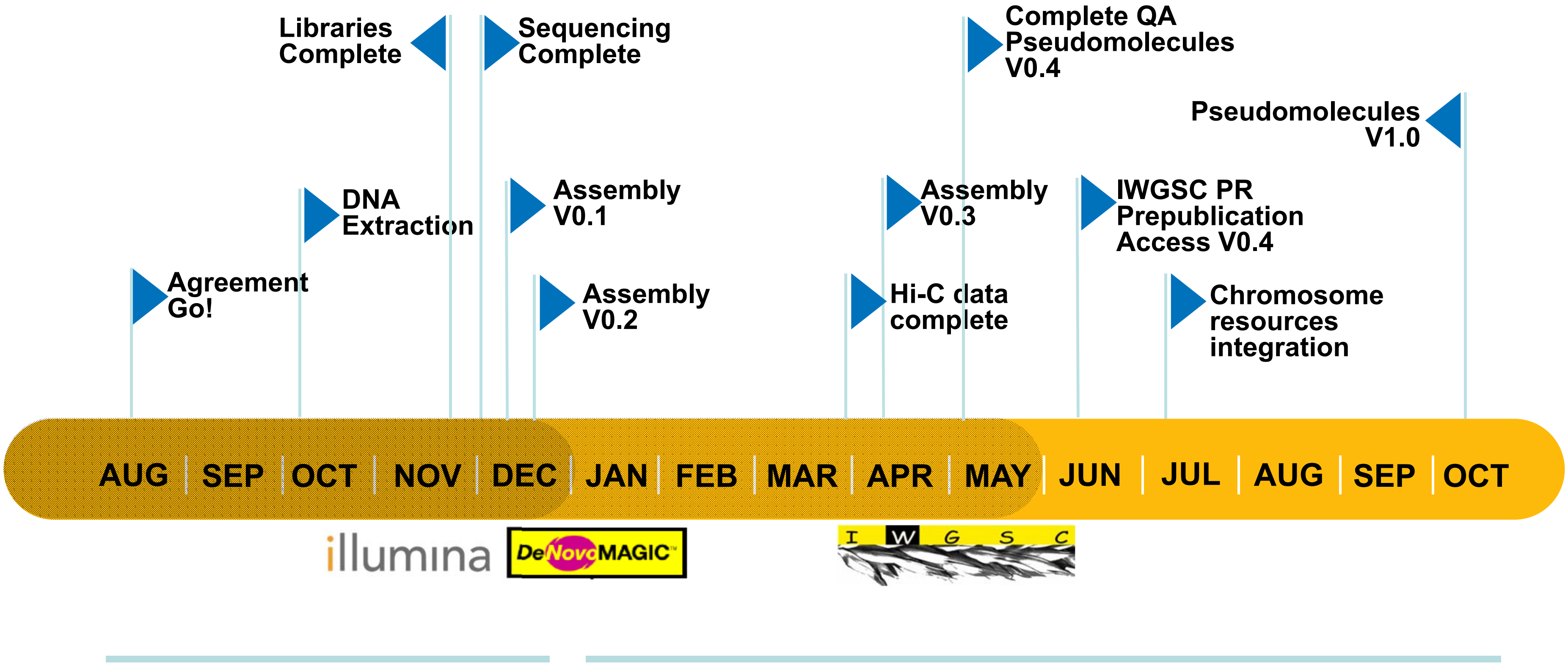
Genetic maps

RH maps

WGP™ tags of all MTPs

IWGSC chromosome survey sequence

IWGSC RefSeq v1.0 Project - Timeline



Agreement Go!

DNA Extraction

Libraries Complete

Sequencing Complete

Assembly V0.1

Assembly V0.2

Hi-C data complete

Assembly V0.3

Complete QA Pseudomolecules V0.4

IWGSC PR Prepublication Access V0.4

Chromosome resources integration

Pseudomolecules V1.0

AUG | SEP | OCT | NOV | DEC | JAN | FEB | MAR | APR | MAY | JUN | JUL | AUG | SEP | OCT

IWGSC RefSeq Project

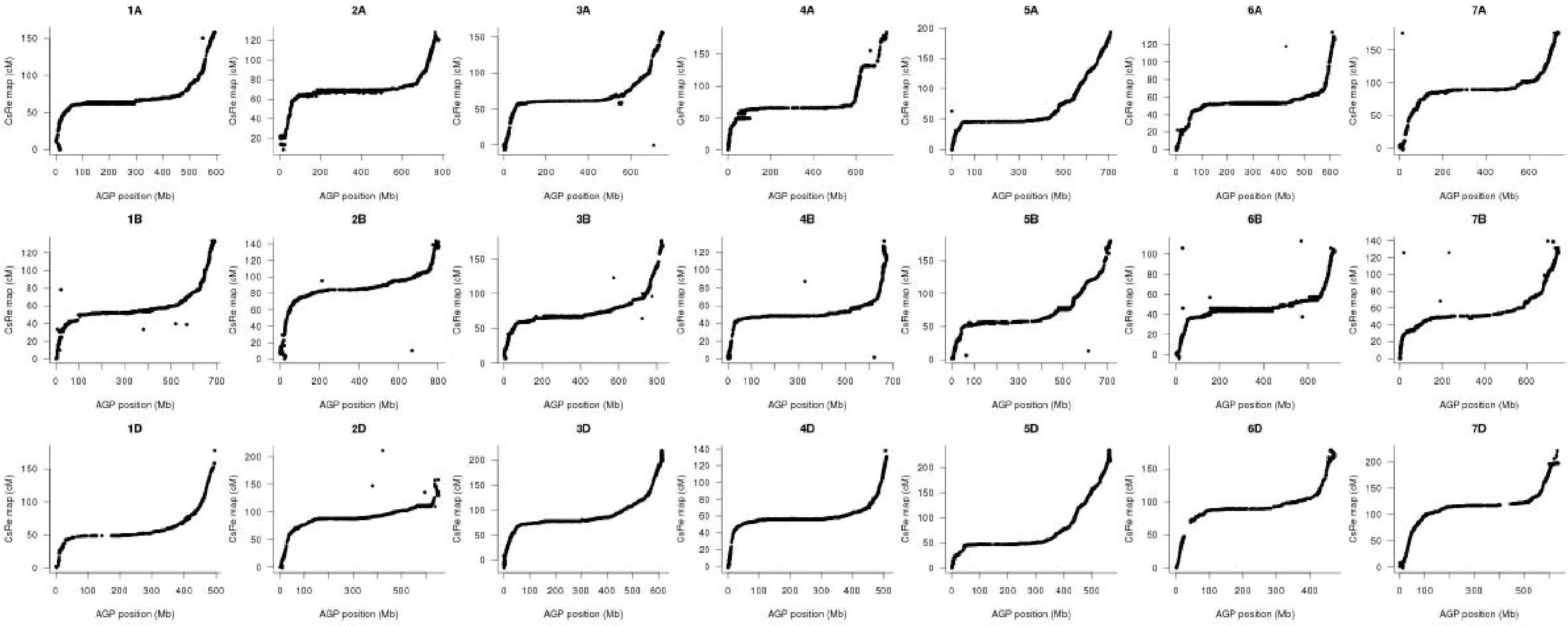
- WGP tags (mostly from MTP BACs) for all chromosome expect 3B
  - ▶ 693,697 BACs, 4,305,249 unique tags
- BAC sequence assemblies for 8 chromosomes (1A, 1B, 3B, 3D, 6B, 7A, 7B, 7D) and partial MTP data for two arms (4AL, 5BS)
  - ▶ 52,890 BACs (9.7 Gb), N50: 68 kb
- Physical maps for all chromosomes
  - ▶ 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

	v0.4	**v1.0**
number/size of scaffolds/contigs	138,607/14.5 Gb	138,665/14.5 Gb
number/size of scaffolds/contigs >= 100 kb	4,442/14.2 Gb	4,443/14.2 Gb
N50 of scaffolds / chromosomal superscaffolds	7.0 Mb	22.8 Mb
gaps filled with BAC sequences		183 (1.7 Mb)
average size of inserted BAC sequences		9.5 kb
sequence assigned to chromosomes	14.1 Gb (96.8 %)	14.1 Gb (96.8 %)
sequence assigned to chr. >= 100 kb	14.1 Gb (99.1 %)	14.1 Gb (99.1 %)
# scaffolds/superscaffolds on chromosomes	3,975	1,601
# oriented scaffolds/superscaffolds	2,464	1,243
oriented sequence	13.1 Gb (90.2 %)	13.8 Gb (95.0 %)
oriented sequence >= 100 kb	13.1 Gb (92.4 %)	13.8 Gb (97.3 %)

~75 scaffolds / chromosome

IWGSC RefSeq / Renan Map

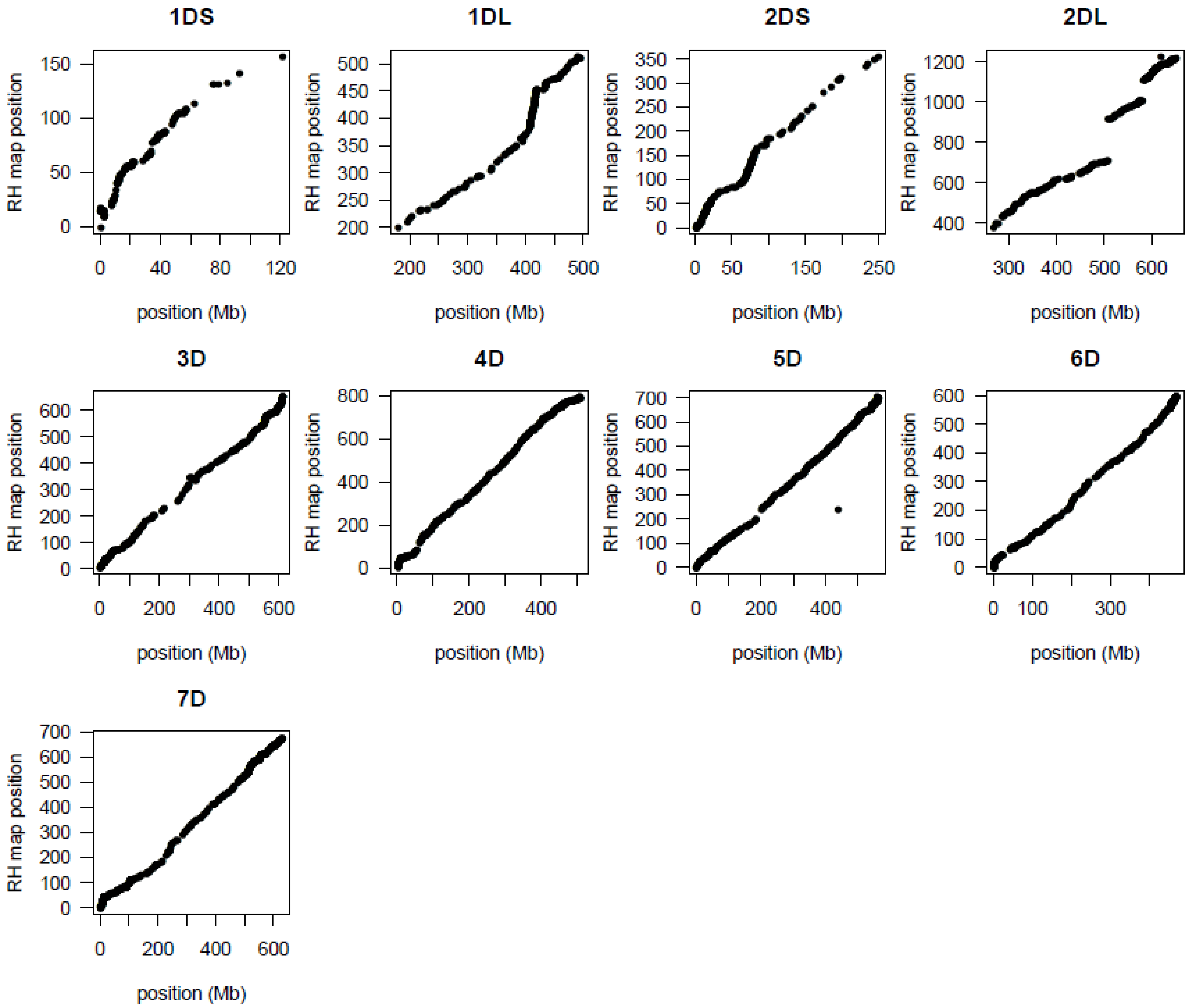


Data provided by Etienne Paux

I W G S C

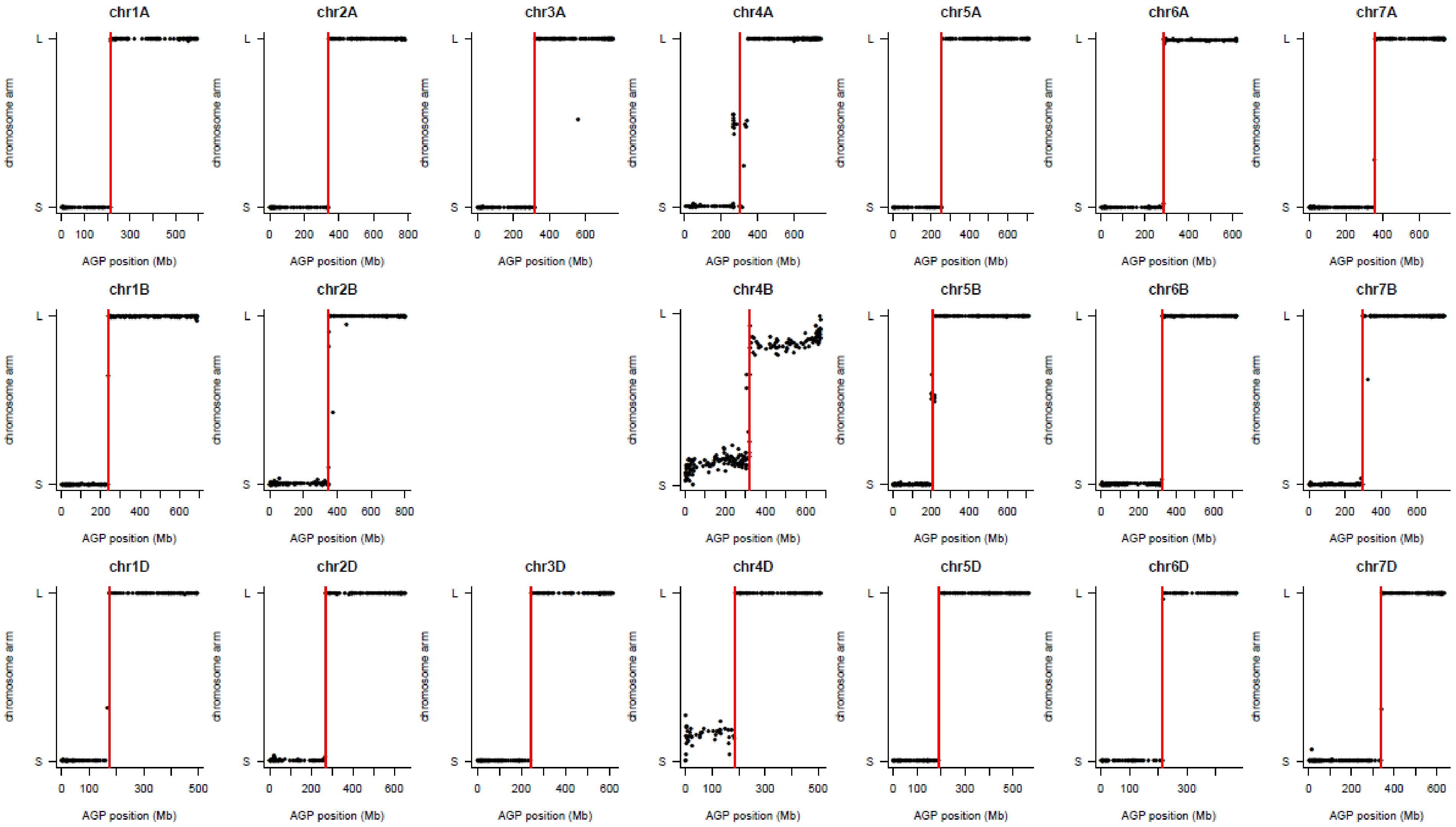
UMR 1095
GDEC
INRA

IWGSC RefSeq / RH Maps

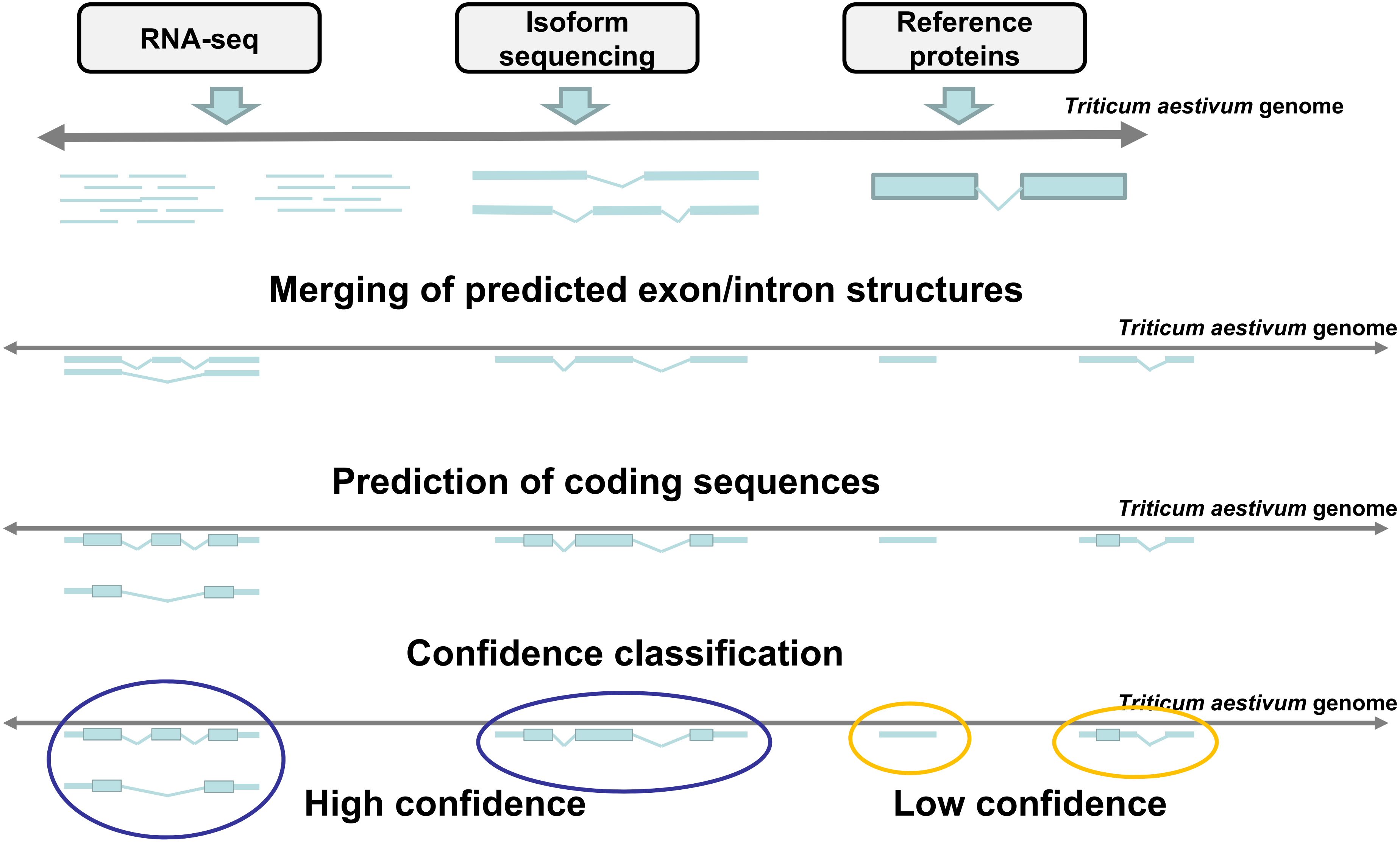


Data provided by Vijay Tiwari

IWGSC RefSeq / Centromere Positions



IWGSC RefSeq Annotation – General



RNA-seq

Isoform sequencing

Reference proteins

Triticum aestivum genome

Merging of predicted exon/intron structures

Triticum aestivum genome

Prediction of coding sequences

Triticum aestivum genome

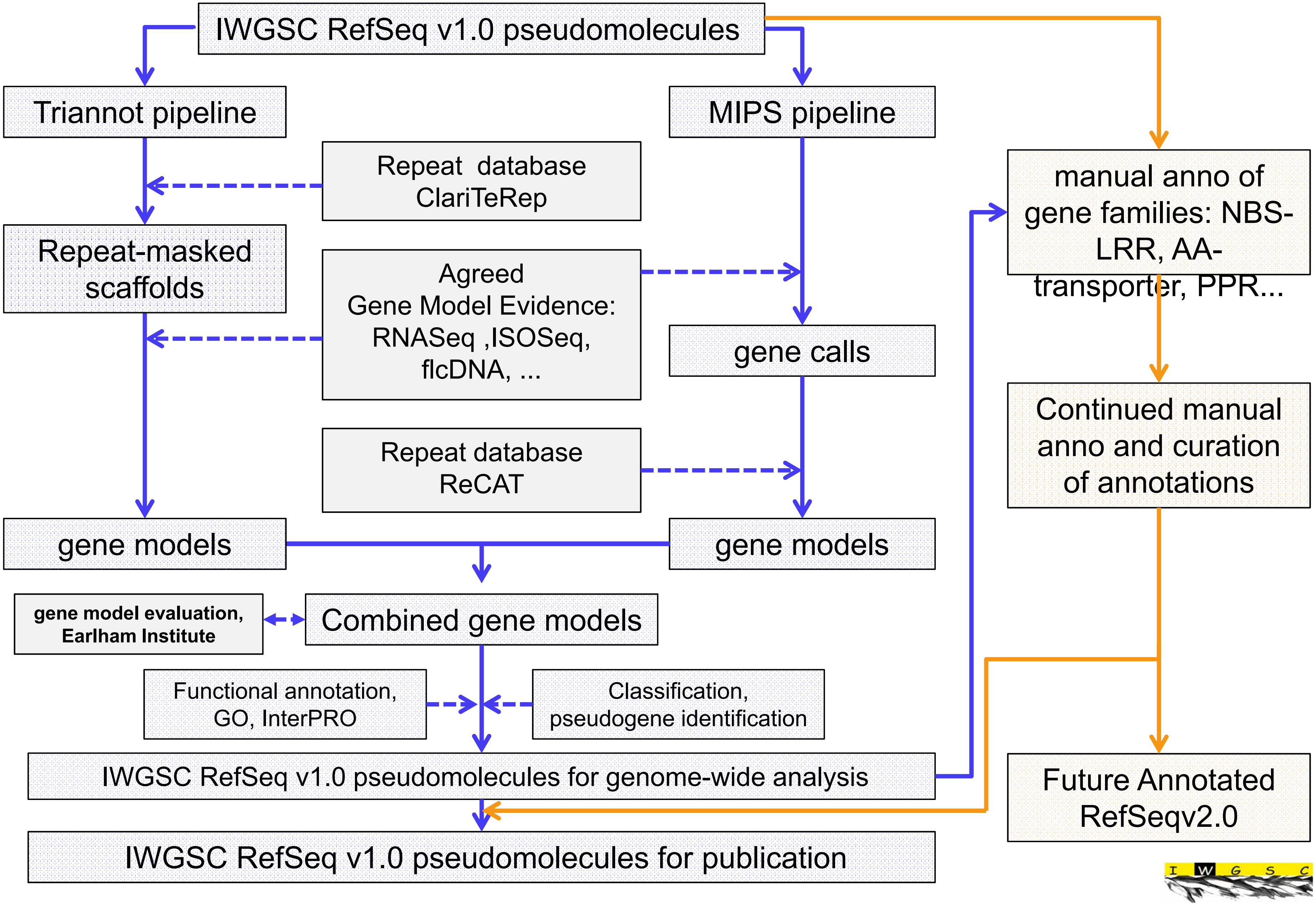
Confidence classification

Triticum aestivum genome

High confidence

Low confidence

IWGSC RefSeq 1.0 Gene / Genome



IWGSC RefSeq v1.0 pseudomolecules

Triannot pipeline

MIPS pipeline

Repeat database
ClariTeRep

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

Repeat-masked scaffolds

Agreed
Gene Model Evidence:
RNASeq ,ISOSeq, flcDNA, ...

gene calls

Continued manual anno and curation of annotations

Repeat database
ReCAT

gene models

gene models

gene model evaluation, Earlham Institute

Combined gene models

Functional annotation, GO, InterPRO

Classification, pseudogene identification

IWGSC RefSeq v1.0 pseudomolecules for genome-wide analysis

Future Annotated RefSeqv2.0

IWGSC RefSeq v1.0 pseudomolecules for publication

IWGSC RefSeq Data Access & Availability

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

IWGSC RefSeq Summary

- 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

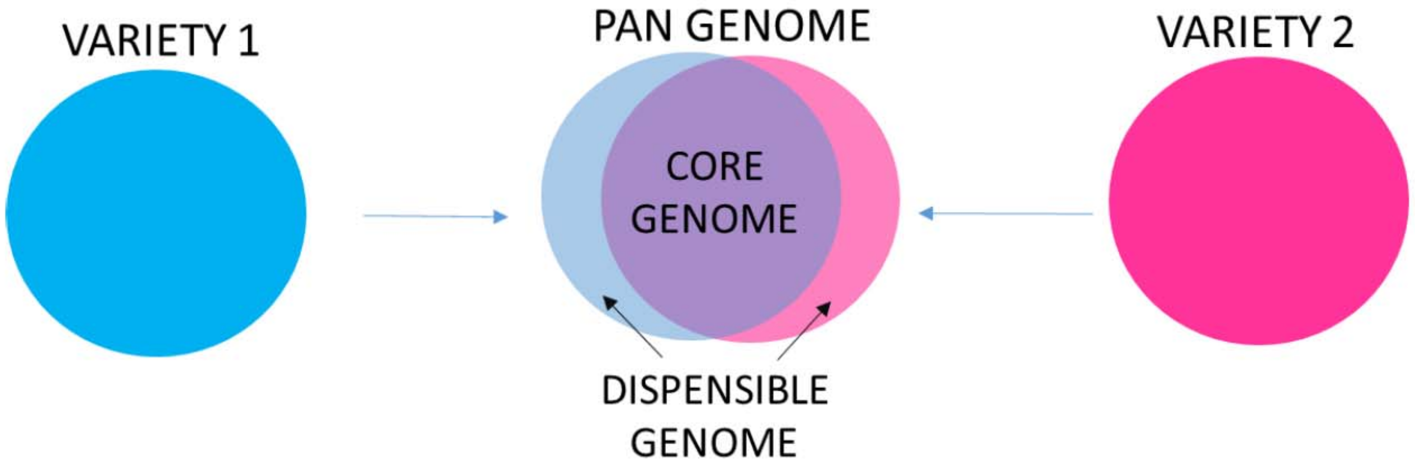
Wheat genome sequencing – work in progress

Reference quality assemblies in progress:

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

	Wild emmer	**Durum**	**CS v0.4**	**CS v1.0**	**CDC Landmark**	**CDC Stanley**
Assembly size:	10.5 Gbp	10.45 Gbp	14.5 Gbp	14.1 Gbp	14.5 Gb	14.5 Gb
Total # scaffolds:	151,912	129,464	138,484	1,601 [4,161]	146,603	166,832
L50:	6.9 Mbp	6.0 Mbp	7.1Mbp	22.8 Mbp	6.1 Mbp	5.9 Mbp
N50 (#sequences):	414	493	566	166	727	684
L90:	1.1 Mbp	1.1 Mbp	1.3 Mbp	4.1 Mbp	1.1 Mbp	1.2 Mb
N90 (#sequences):	1,827	2,019	2,363	718	2,687	2,676

More wheat genome sequences – towards wheat pan-genome analysis



10 wheat genomes (and more) project:

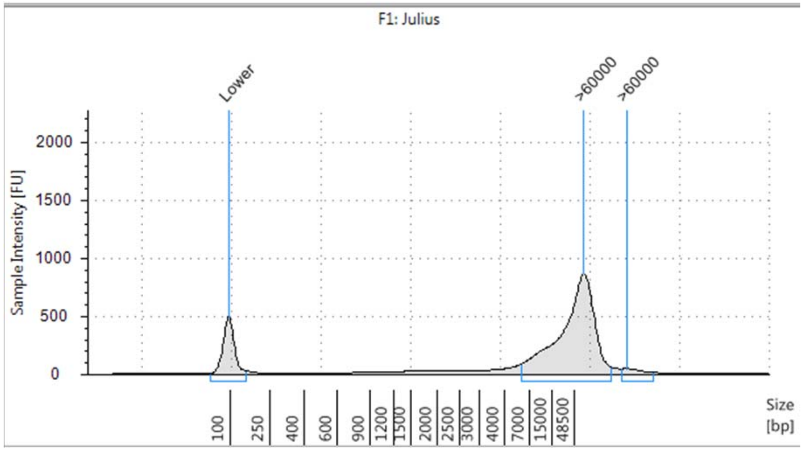
NRGene refseq quality assemblies

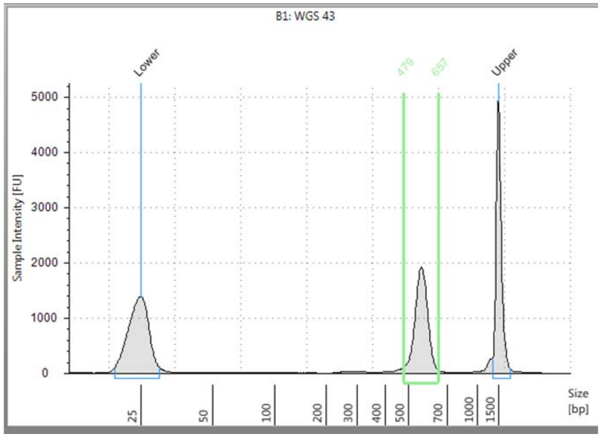
- 2 Canadian varieties (CDC Landmark, CDC Stanley)
- 1(+2) US varieties (Jagger)
- 1 German winter wheat variety (Julius)
- 1 Swiss winter wheat variety (Arina)
- (2) Australian varieties (tbd)
- 1(+1) Chinese variety (Zhongmai175)

Earlham Institute assemblies

- CS, Cadenza, Paragon, Kronos, Robigus, Claire, Alchemy, Brompton, Hereward, Rialto, Soissons, Xi19

Sequencing of a german winter wheat: „Julius“



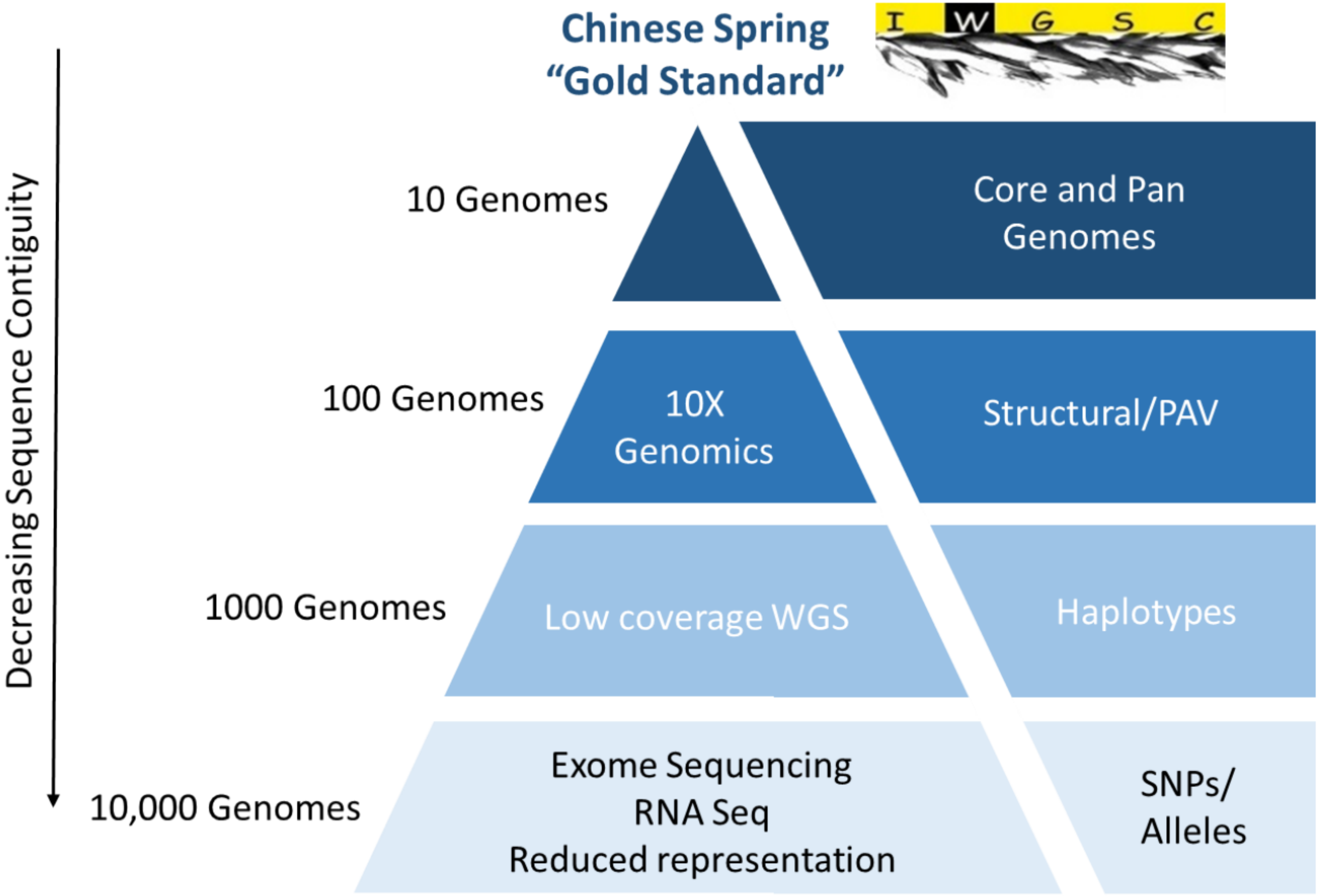


- Sequencing of 800 nt PE and MP libraries done
- Sequencing of 450 nt PE libraries ongoing

- Assembly, End of January 2017

- Hi-C/10x genomics based scaffolding, March 2017

Wheat genome sequencing – the future



Chinese Spring
“Gold Standard”

IWGSC

Decreasing Sequence Contiguity

10 Genomes

Core and Pan Genomes

100 Genomes

10X Genomics

Structural/PAV

1000 Genomes

Low coverage WGS

Haplotypes

10,000 Genomes

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

CTAG²
Canadian Triticum Applied Genomics

GenomeCanada

GenomePrairie

SaskWheat
DEVELOPMENT COMMISSION

Alberta Wheat
COMMISSION
GROWTH.
INNOVATION.
RESULTS.

Government
of
Saskatchewan

Western Grains
Research Foundation

I W G S C

Murdoch
UNIVERSITY

NSF

USDA

illumina

With support from

Federal Ministry
of Food
and Agriculture

Federal Office for
Agriculture and Food

by decision of the
German Bundestag

INRA
SCIENCE & IMPACT

VICTORIA
State
Government

Economic Development,
Jobs, Transport
and Resources

IWGSC Chromosome Leaders

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

UMR 1095
GDEC
INRA

Etienne Paux, Frédéric Choulet

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

Jaroslav Dolezel, Hana Simkova, Miroslav Valarik, Jan Bartos

Murdoch
UNIVERSITY

Rudi Appels

NIAS

Hirokazu Handa

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

Luigi Cattivelli

Sabancı Universitesi

Hikmet Budak

K-STATE
Kansas State University

Bikram Gill

Bayer CropScience

Catherine Feuillet
John Jacobs

Universität Zürich UZH

Beat Keller

University of Haifa

Abraham Korol

NMBU
Norwegian University of Life Sciences

Odd-Arne Olsen

UNIVERSITY OF SASKATCHEWAN

Curtis Pozniak

IPK
GATERSLEBEN

Nils Stein
Thorsten Schnurbusch

THE FEDERAL RESEARCH CENTER
Institute of Cytology and Genetics
OF SIBERIAN BRANCH OF THE RUSSIAN ACADEMY OF SCIENCES

Elena Salina

PUNJAB AGRICULTURAL UNIVERSITY

Kuldeep Singh

NORTHWEST A&F UNIVERSITY

Song Weining

Earlham Institute

Matt Clark

IWGSC Team Leaders

IWGSC Sequence Repository

URGI **Michael Alaux**

BAC Libraries

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

Jaroslav Dolezel, Hana Simkova

BAC Library Pools

CNRGV **Hélène Bergès**

WGP Tags

Bayer CropScience

John Jacobs

RNASeq

John Innes Centre
Unlocking Nature's Diversity

Cristobal Uauy

Genetic Maps

K-STATE Kansas State University

Jesse Poland

WGA PIs

IPK GATERSLEBEN **Nils Stein**

University of Saskatchewan

Curtis Pozniak
Andrew Sharpe

K-STATE Kansas State University **Jesse Poland**

UMR 1095 GDEC INRA **Frédéric Choulet**

NRGene **Gil Ronen**

illumina

Mike Thompson

IWGSC

Kellye Eversole
Jane Rogers

Tel Aviv University **Assaf Distelfeld**

Pseudomolecule Team

UMR 1095 GDEC INRA **Frédéric Choulet**

Victoria State Government | Economic Development, Jobs, Transport and Resources

Gabriel Keeble-Gagnere

IPK GATERSLEBEN **Martin Mascher**

Annotation Team

UMR 1095 GDEC INRA

Philippe Leroy
Frédéric Choulet

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Klaus Mayer
Manuel Spannagl

Earlham Institute **David Swarbreck**

RH Mapping

University of Maryland

Vijay Tiwari

Thanks to IWGSC Sponsors!

Arcadia BIOSCIENCES

ACPFG

biogemma

BAYER Bayer CropScience

CSIRO

CIMMYT MR International Maize and Wheat Improvement Center

Céréales Vallée

DUPONT® The miracles of science™

FLORIMOND DESPREZ

INRA SCIENCE & IMPACT

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

KWS ®

KeyGene

Limagrain from earth to life

Together, we are KANSAS WHEAT

MONSANTO ®

UNIVERSITY OF SASKATCHEWAN

Crop Development Centre

RAGT

syngenta

CTAG²
Canadian Triticum Applied Genomics

I W G S C

An international initiative to sequence the genome of the durum wheat cultivar Svevo

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

Luigi Cattivelli
Nicola Pecchioni

Consiglio Nazionale delle Ricerche

Aldo Ceriotti
Luciano Milanesi
Gabriella Sonnante

ALMA MATER STUDIORUM A.D. 1088

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

IPK GATERSLEBEN

Nils Stein
Martin Mascher

MONTANA STATE UNIVERSITY

Hikmet Budak

אוניברסיטת תל-אביב

University of Tel Avi
Assaf Distelfeld

UNIVERSITY OF SASKATCHEWAN

Curtis Pozniak

HelmholtzZentrum münchen
German Research Center for Environmental Health

Klaus Mayer

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

With support from

Federal Ministry
of Food
and Agriculture

by decision of the
German Bundestag

Federal Office for
Agriculture and Food