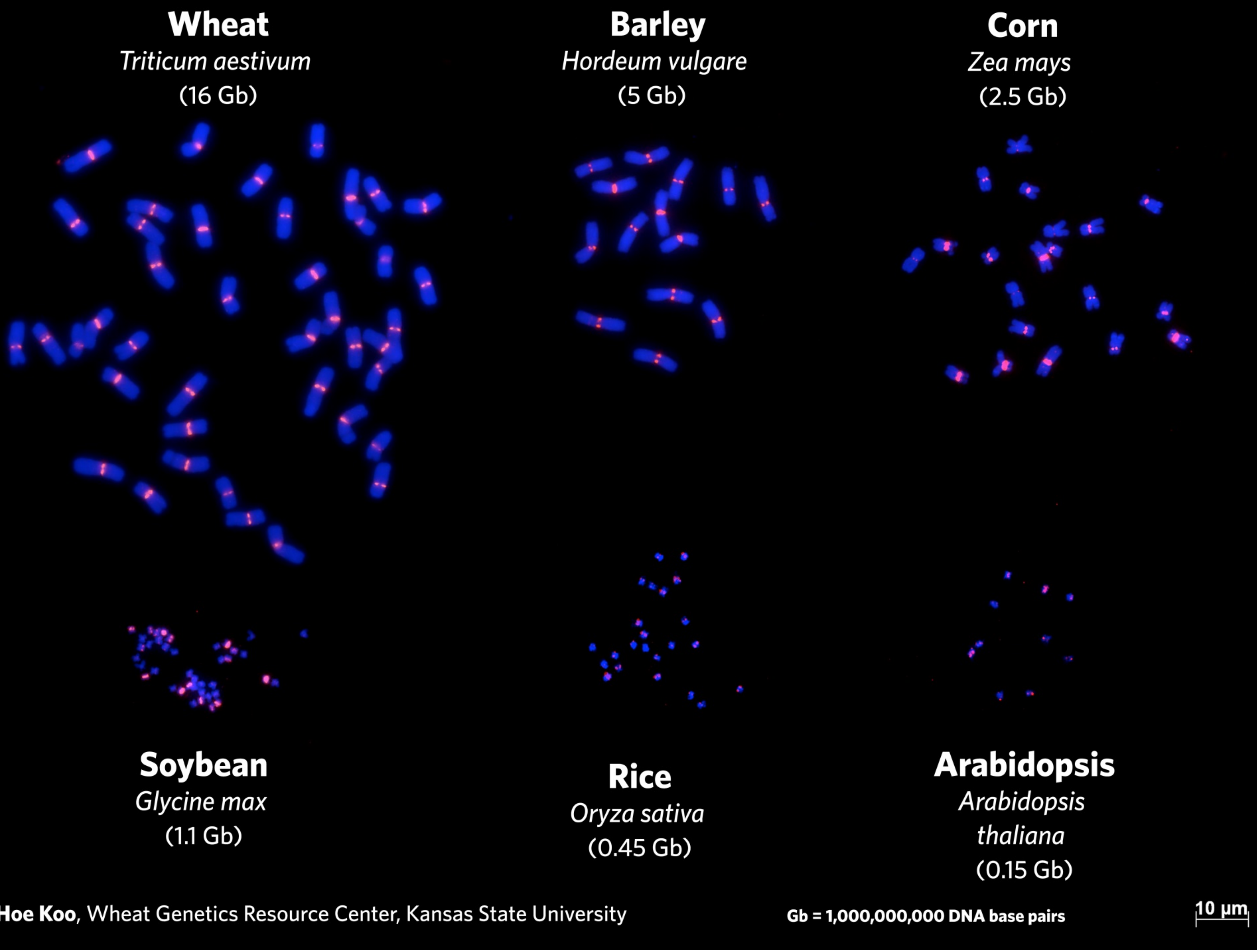
The IWGSC Reference Genome Sequence

Nils Stein, IPK Gatersleben

International Plant & Animal Genome XXV / January 14-18, 2017 - San Diego, CA, USA

Genome sequencing – the ‘wheat’ challenge



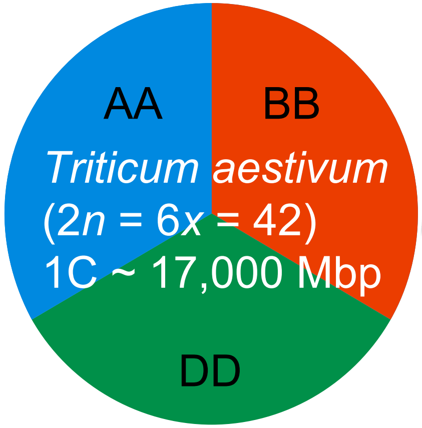
Dal-Hoe Koo, Wheat Genetics Resource Center, Kansas State University

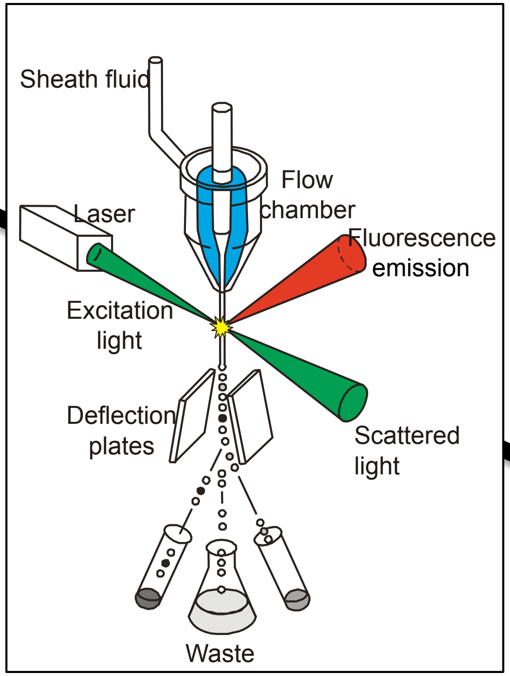
Gb = 1,000,000,000 DNA base pairs

Wheat genome sequencing in 2005

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

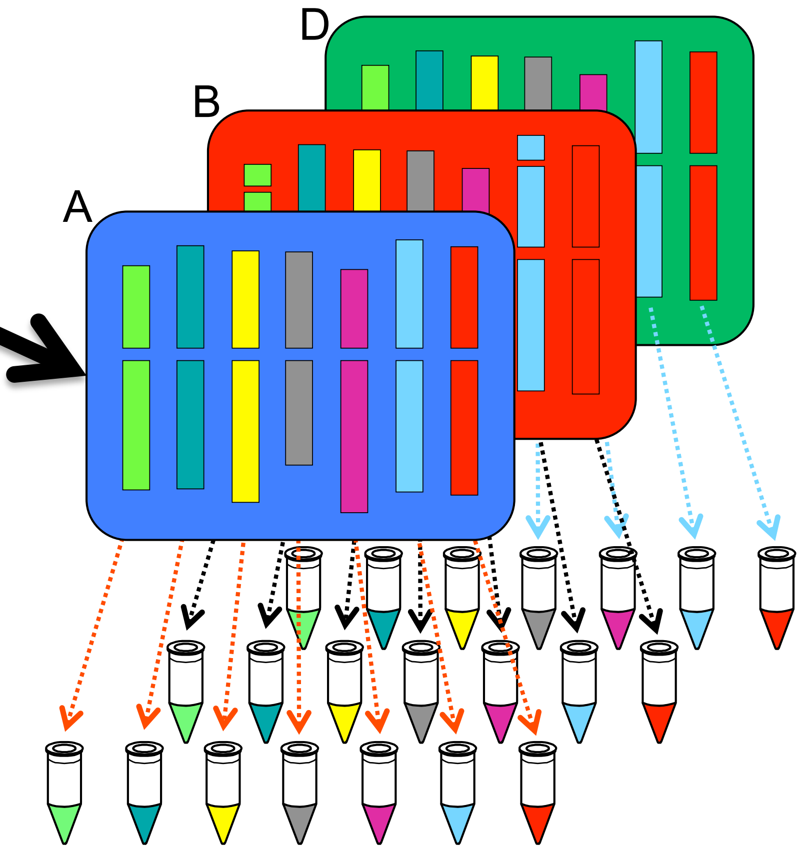
A chromosome-based approach





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

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



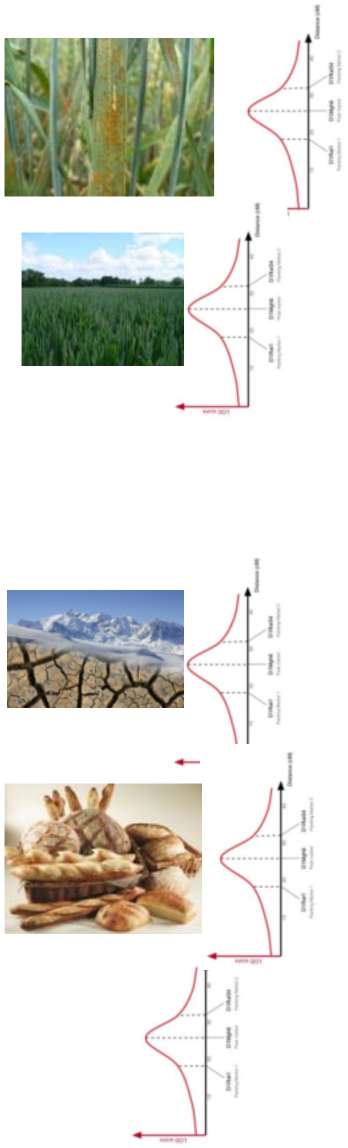
- Chromosomes: 605 - 995 Mbp (3.6 – 5.9% of the genome)
- Chromosome arms: 225 - 585 Mbp (1.3 – 3.4% of the genome)

- **Chromosome specific BAC libraries (2006 - 2012)**
- **Amplified DNA for chromosome survey (2010 - 2011)**

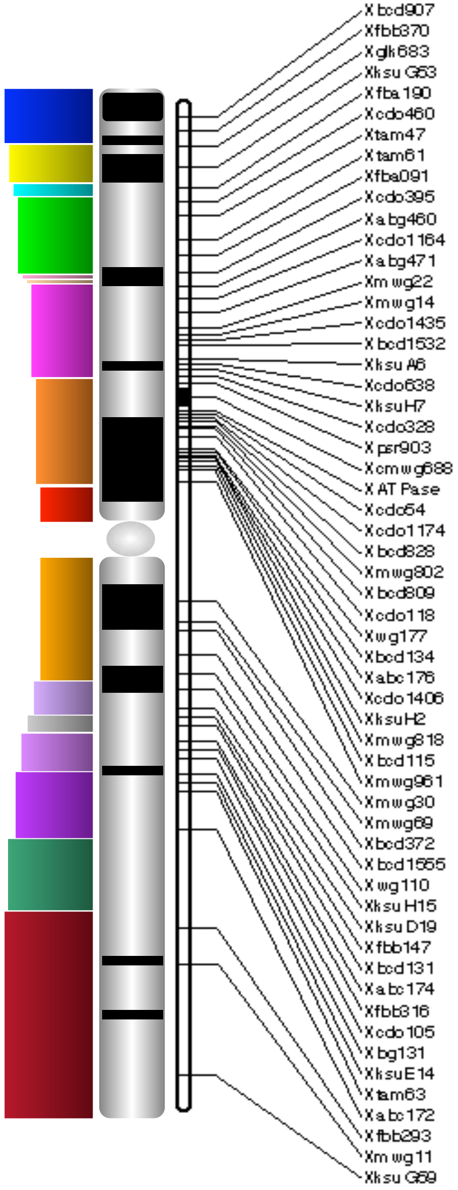
IEB

An integrated and ordered 3B reference sequence

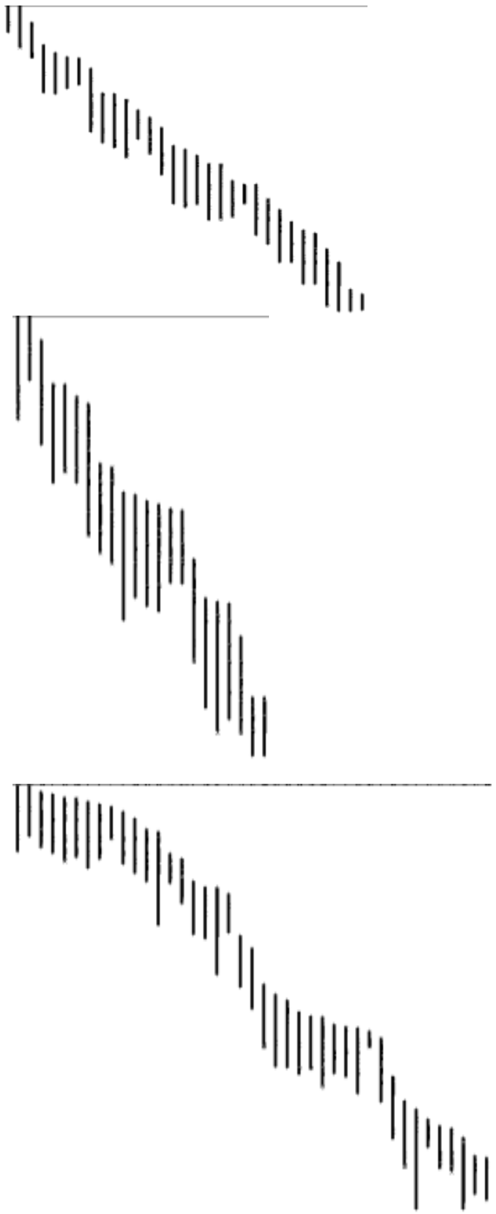
MetaQTL analysis



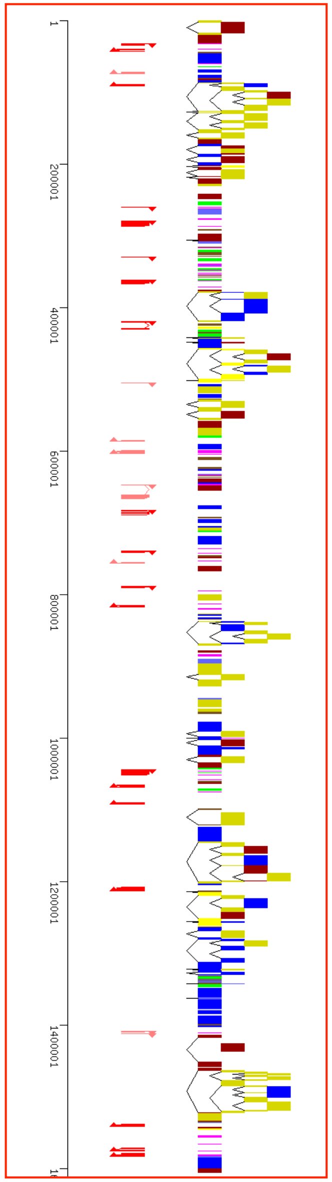
3B consensus map (5000 markers)



3B Physical map



3B pseudomolecule



Courtesy: Catherine Feuillet

The International Wheat Genome Sequencing Consortium

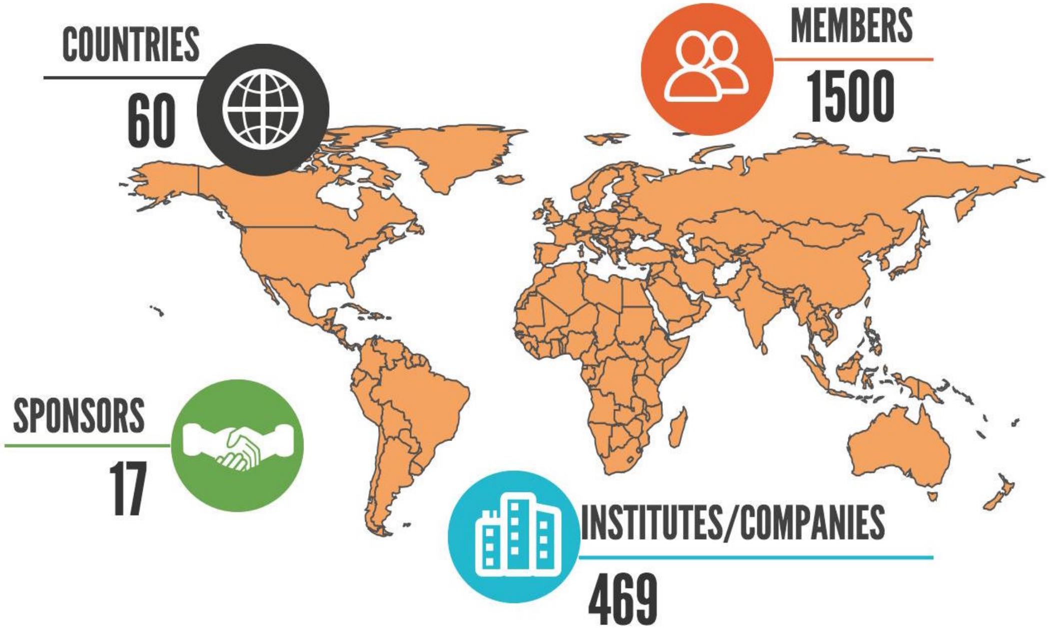
2017

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 60

MEMBERS 1500

SPONSORS 17

INSTITUTES/COMPANIES 469

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.
KeyGene
Limagrain from earth to life
Together, we are Kansas Wheat
University of Saskatchewan Crop Development Centre
RAGT
syngenta
CTAG²

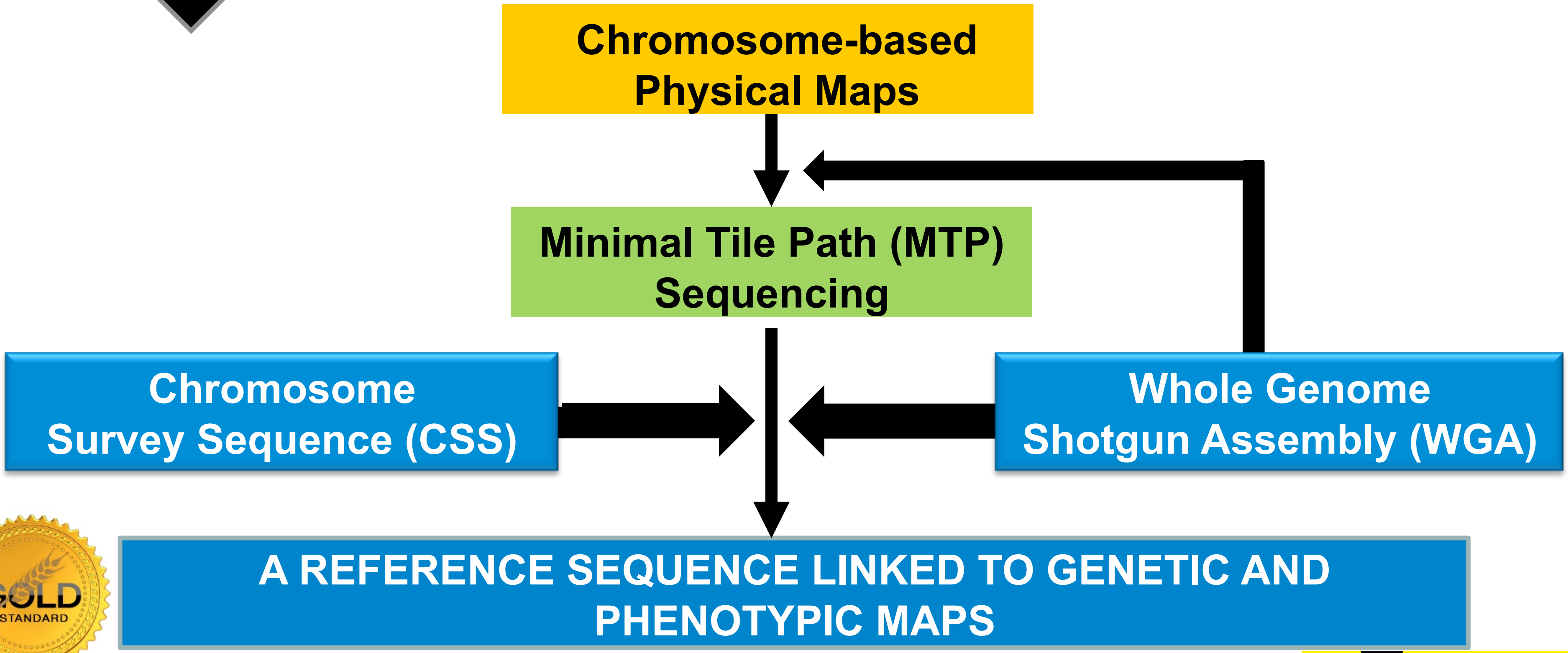
IWGSC

www.wheatgenome.org

IWGSC 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



Chromosome-based Physical Maps

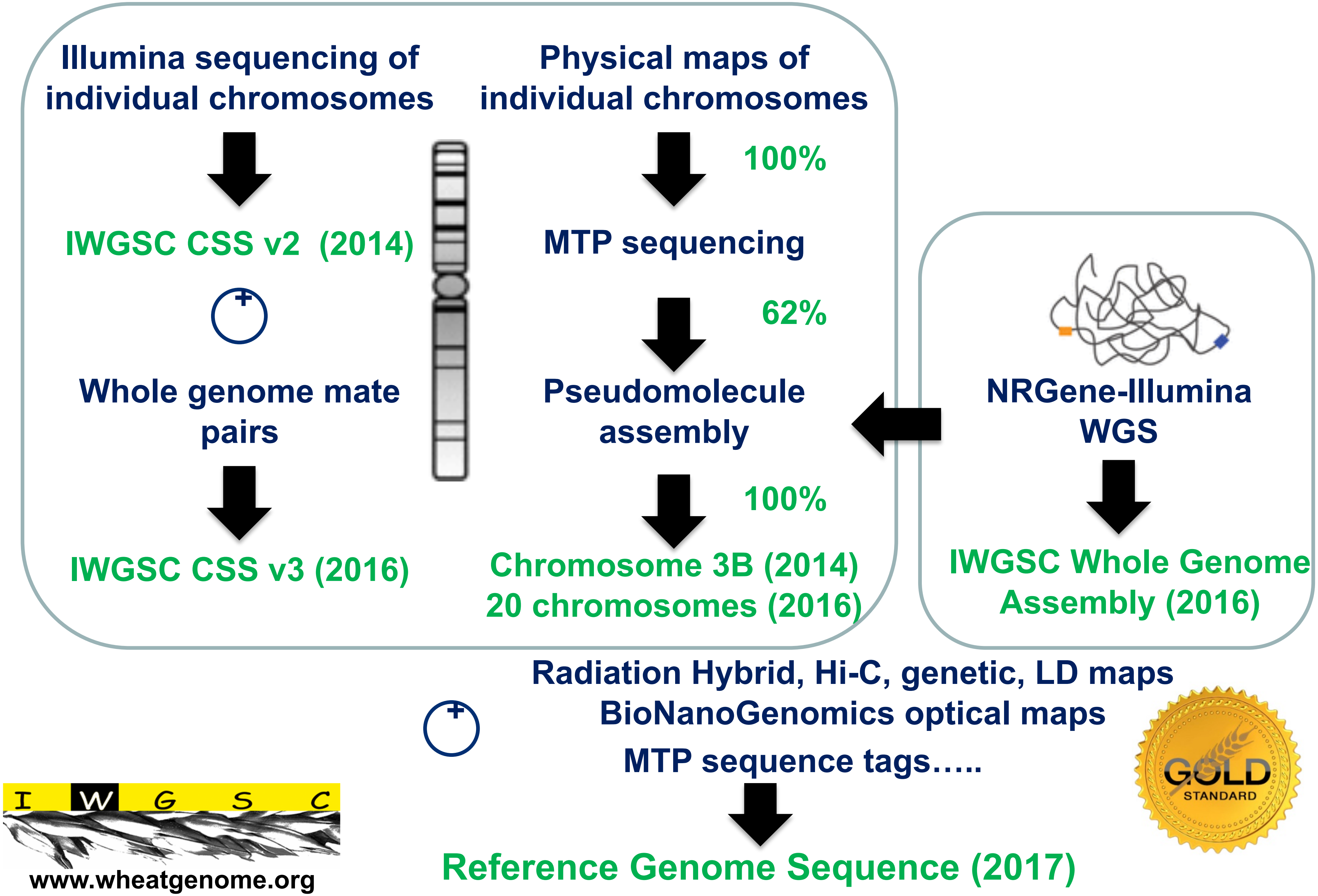
Minimal Tile Path (MTP) Sequencing

Chromosome Survey Sequence (CSS)

Whole Genome Shotgun Assembly (WGA)

A REFERENCE SEQUENCE LINKED TO GENETIC AND PHENOTYPIC MAPS

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

100%

MTP sequencing

62%

Pseudomolecule assembly

100%

Chromosome 3B (2014)
20 chromosomes (2016)

NRGene-Illumina WGS

IWGSC Whole Genome Assembly (2016)

+

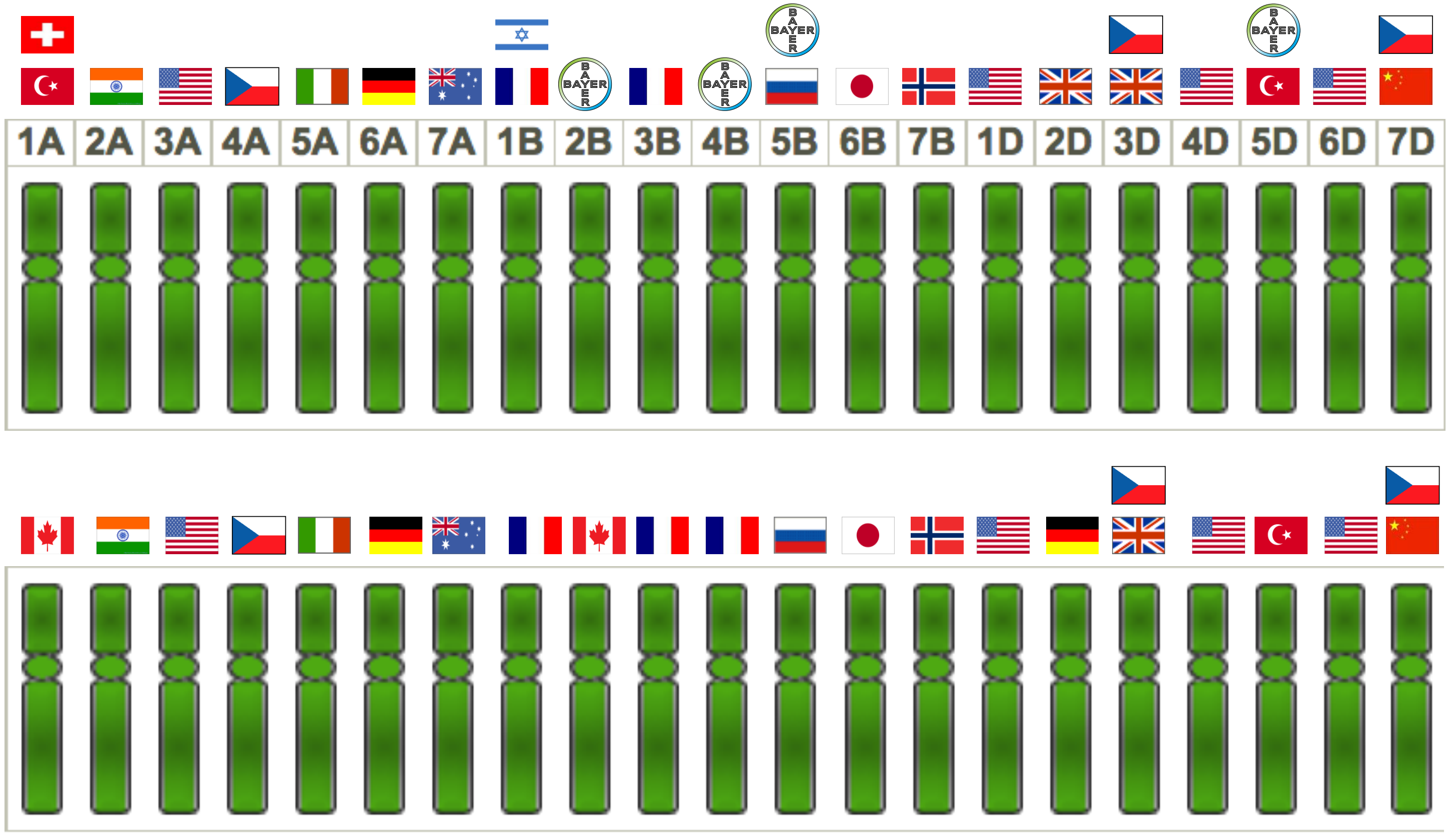
Radiation Hybrid, Hi-C, genetic, LD maps
BioNanoGenomics optical maps
MTP sequence tags…..

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 and pseudo-molecule sequences available at IWGSC repository: https://wheat-urgi.versailles.inra.fr

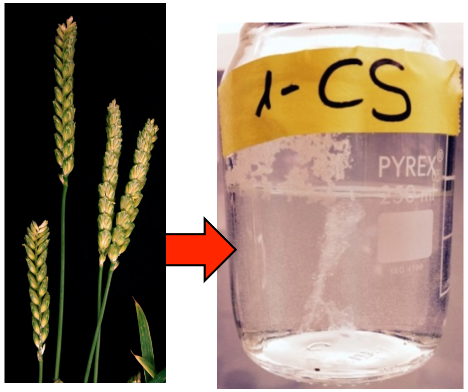
IWGSC

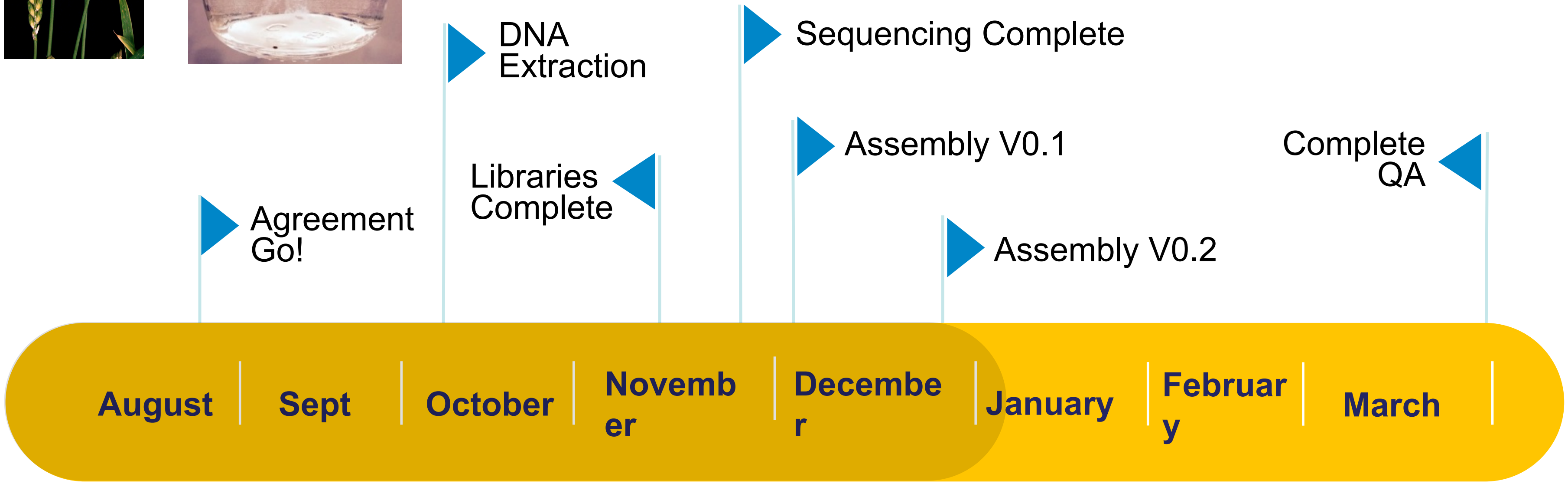
www.wheatgenome.org

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

illumina

DeNovoMAGIC™

~2 months from data accumulation to completion of first assembly

IPK GATERSLEBEN · UNIVERSITY OF SASKATCHEWAN · GIFS GLOBAL INSTITUTE FOR FOOD SECURITY · K-STATE Kansas State University · TEL AVIV UNIVERSITY · INRA SCIENCE & IMPACT · NRGene · IWGSC

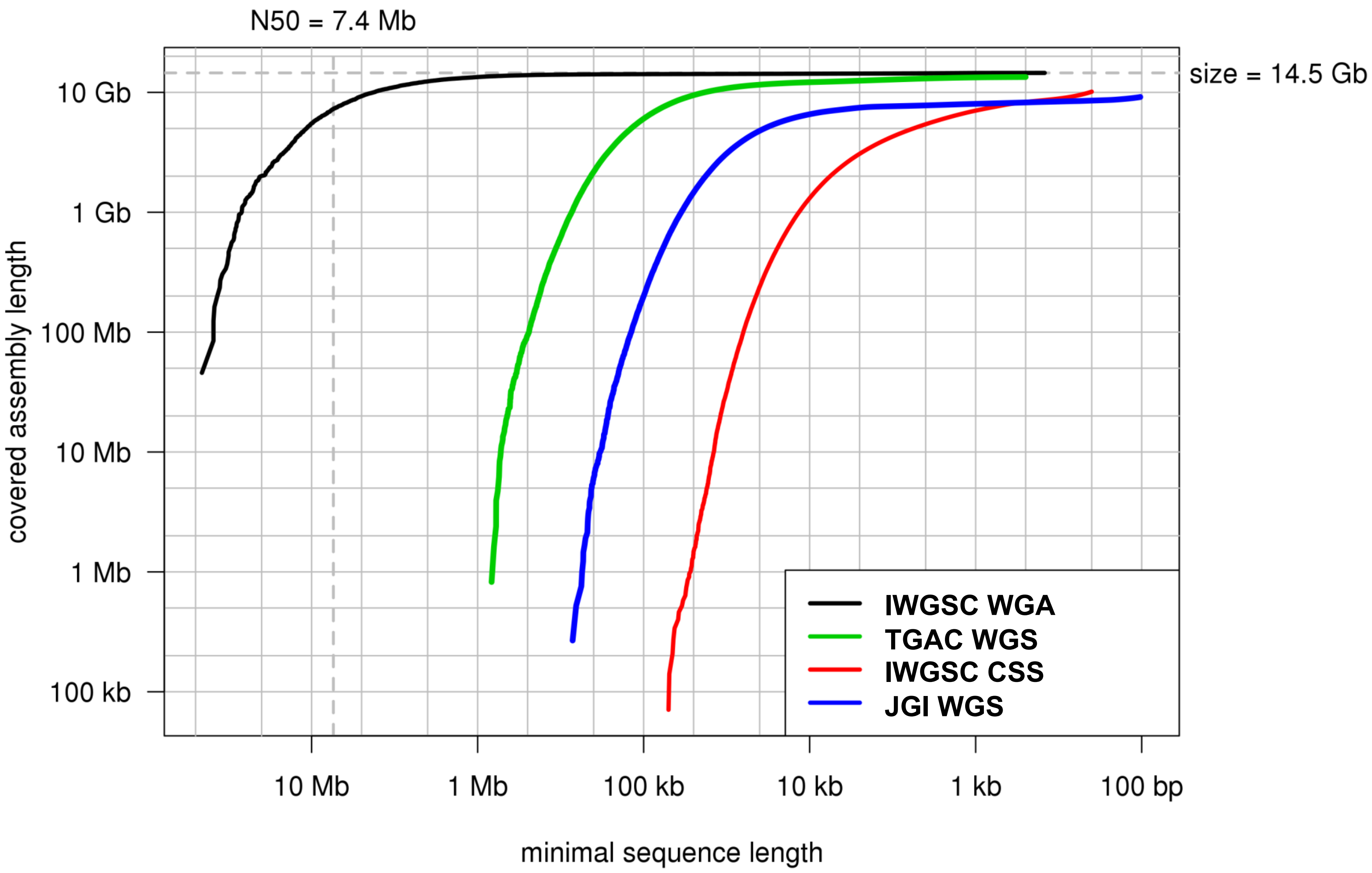
IWGSC Whole Genome Assembly Project

De novo assembly:

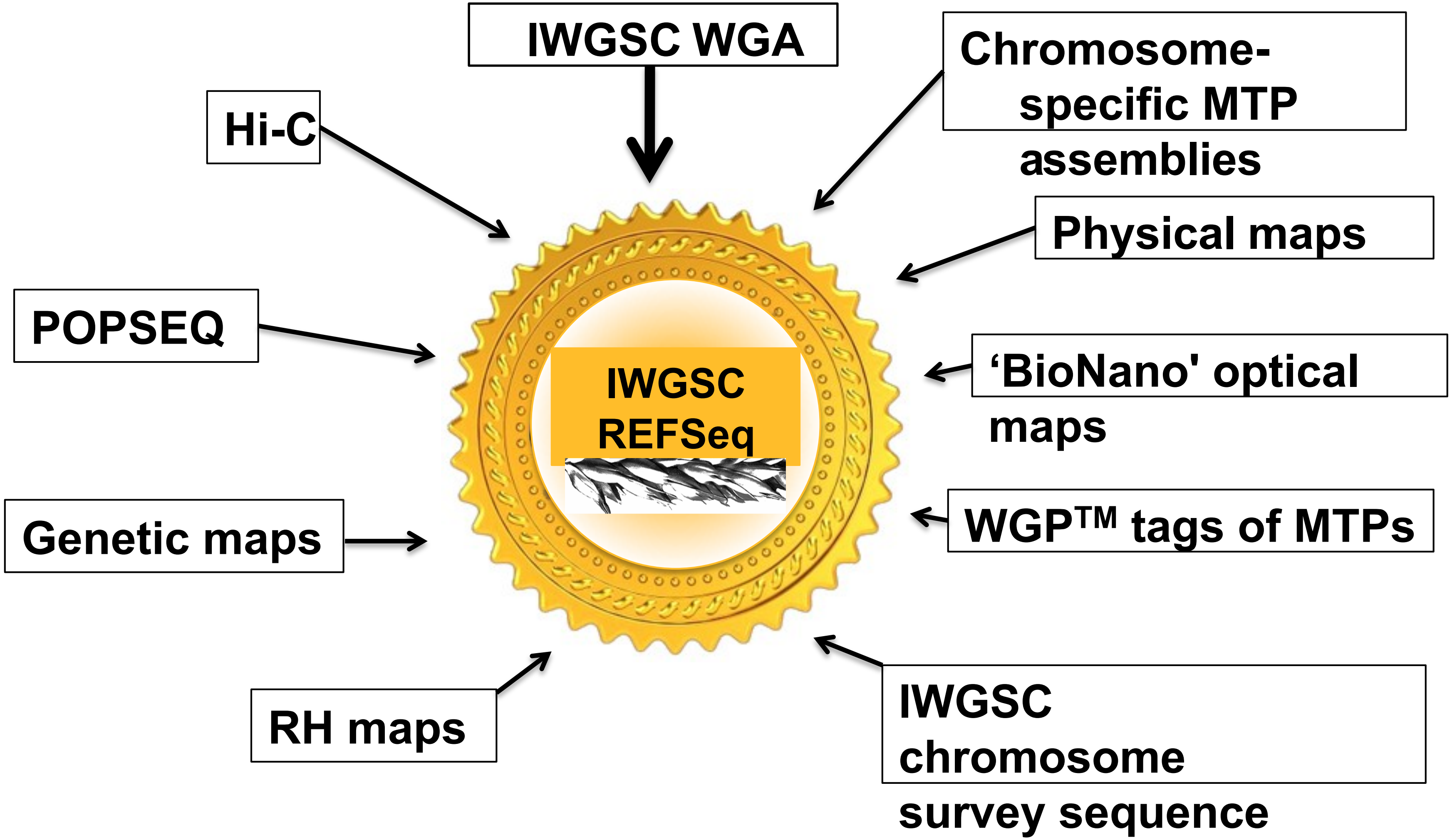
- NRGene's DeNovoMagic-2 platform, total run time < 3 weeks, 1Tb RAM computer
- Illumina short-read sequencing data only (200 x coverage)

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

WGA Assembly Statistics



Concerted integration of resources: RefSeq v1.0



IWGSC WGA

Hi-C

Chromosome-specific MTP assemblies

Physical maps

POPSEQ

'BioNano' optical maps

IWGSC REFSeq

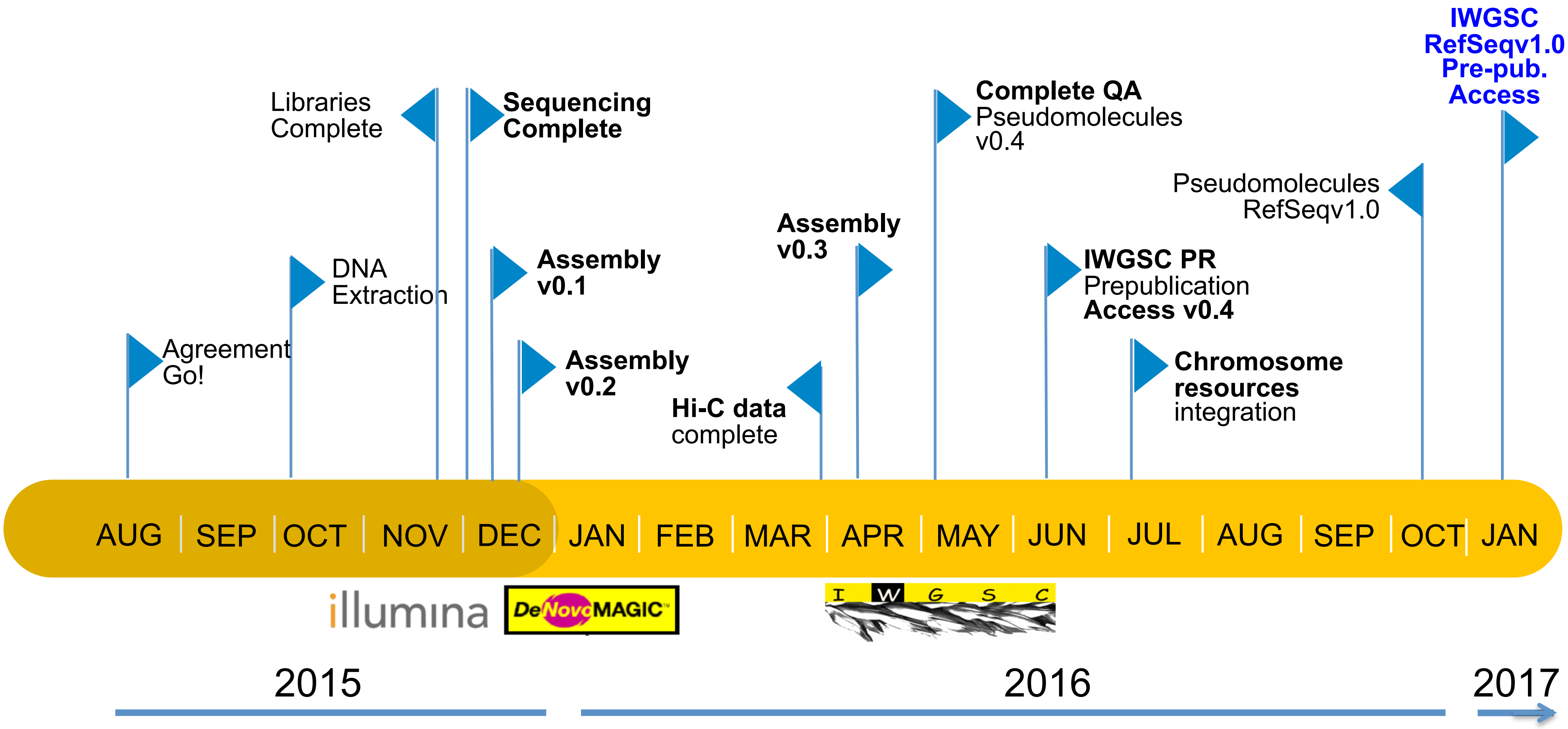
Genetic maps

WGP™ tags of MTPs

RH maps

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
RefSeqv1.0

**IWGSC
RefSeqv1.0
Pre-pub.
Access**

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

2015

2016

2017

IWGSC RefSeq Project

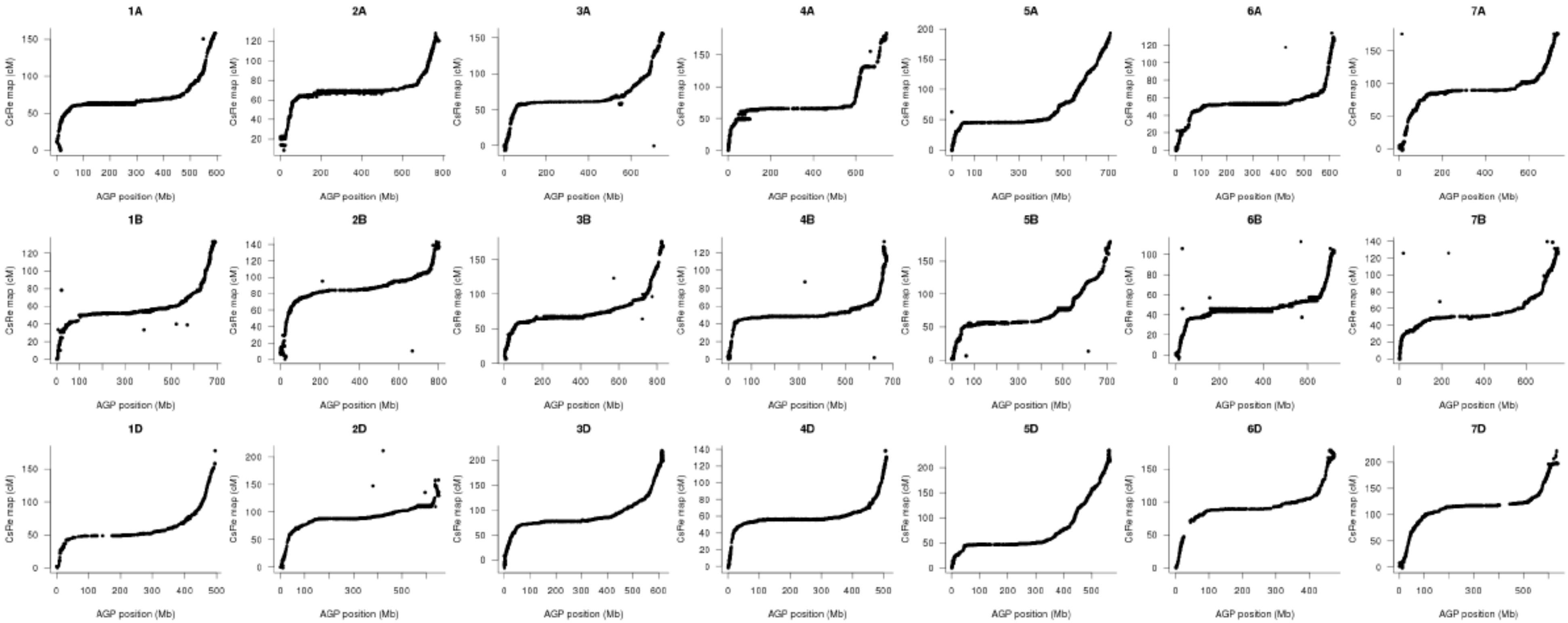
- Physical maps for all chromosomes
  - 1,839,128 BACs, 47,810 contigs, 380,675 singletons
- WGP tags (mostly from MTP BACs) for all chromosomes except 3B
  - 4,305,249 unique tags, 693,697 BACs
- BAC sequence assemblies for 8 chromosomes (1A, 1B, 3B, 3D, 6B, 7A, 7B, 7D) and partial MTP data for two 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

	WGA v0.4	**RefSeq 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 v1.0 / CS-Renan Genetic Map

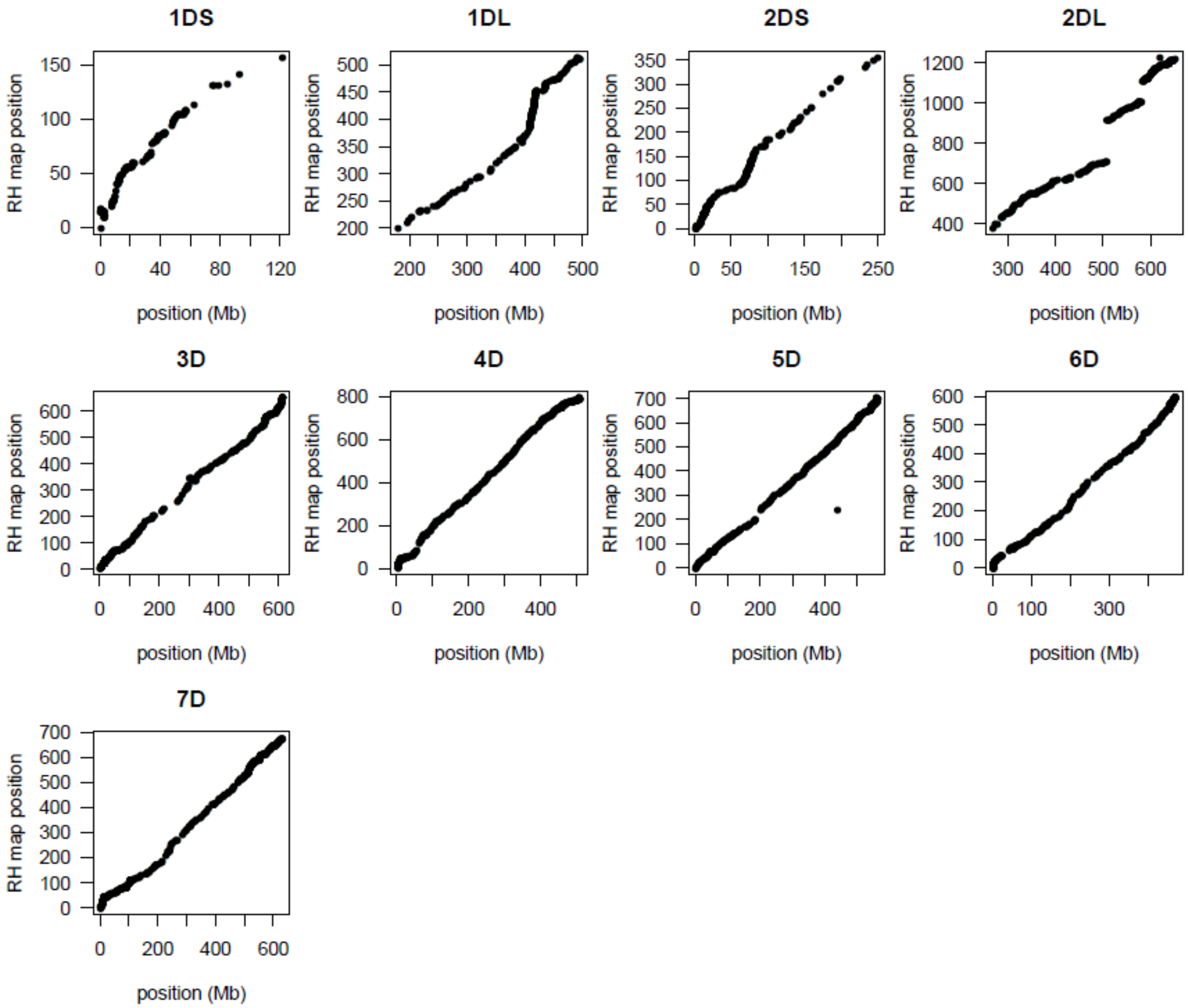


Data provided by Etienne Paux

I W G S C

UMR 1095
GDEC
INRA

IWGSC RefSeq v1.0 / RH Maps



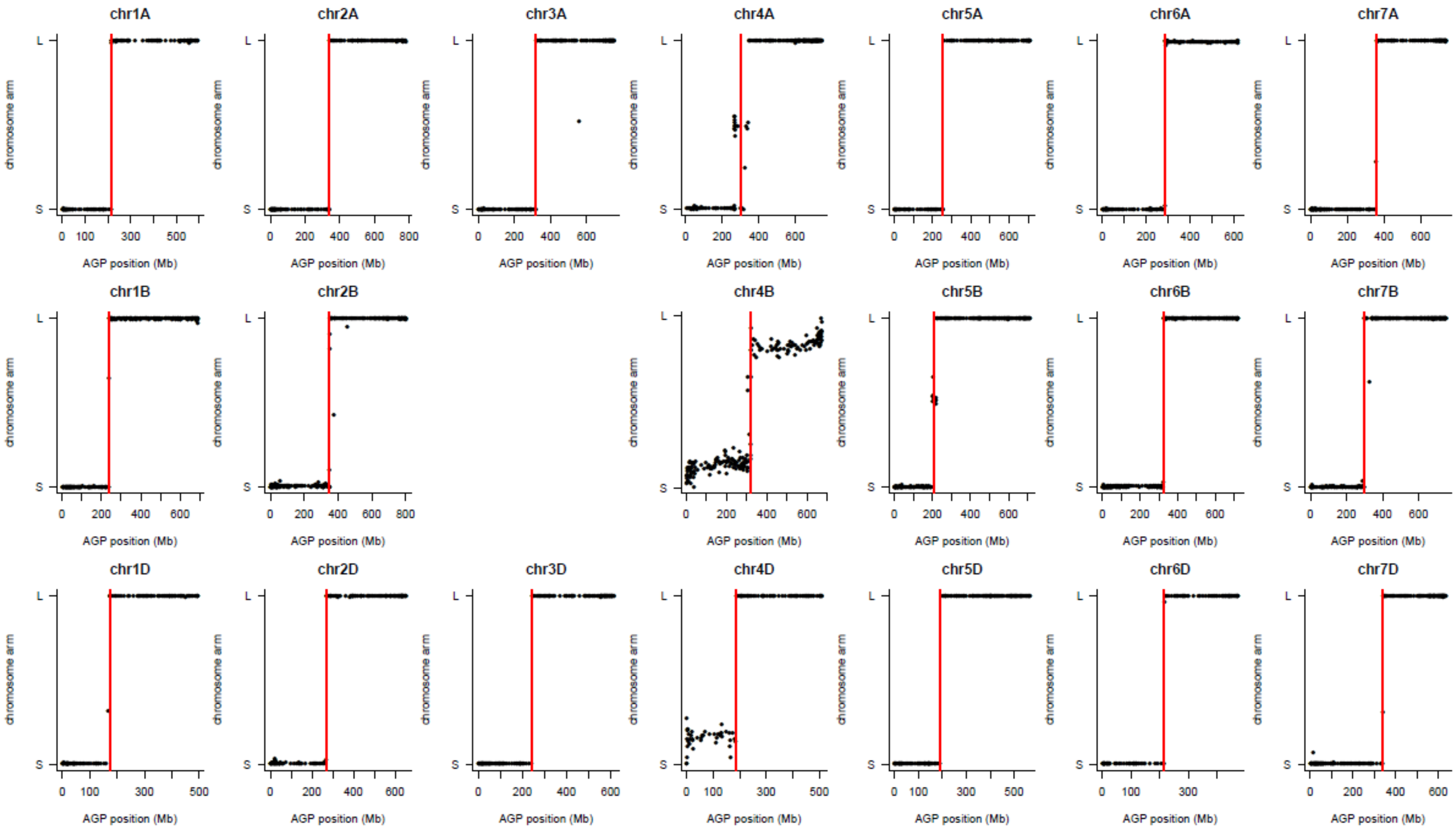
Data provided by Vijay Tiwari

I W G S C

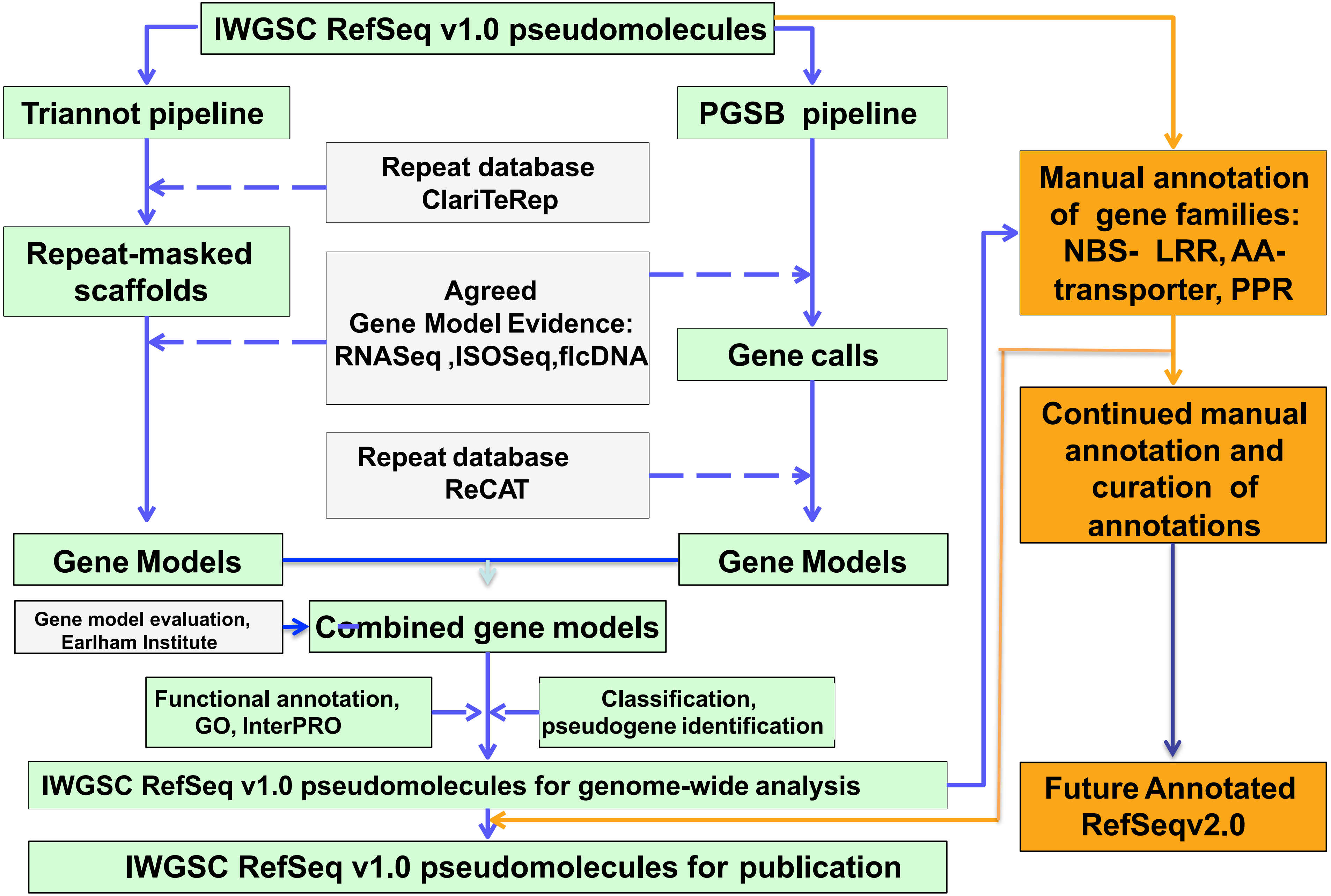
K-STATE Kansas State University

UNIVERSITY OF MARYLAND

IWGSC RefSeq v1.0 / Centromere Positions



IWGSC RefSeq v1.0 Gene / Genome annotation



IWGSC RefSeq v1.0 pseudomolecules

Triannot pipeline

PGSB pipeline

Repeat database
ClariTeRep

Repeat-masked scaffolds

Agreed
Gene Model Evidence:
RNASeq ,ISOSeq,flcDNA

Gene calls

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

IWGSC RefSeq v1.0 pseudomolecules for publication

Manual annotation of gene families:
NBS- LRR, AA-transporter, PPR

Continued manual annotation and curation of annotations

Future Annotated RefSeqv2.0

IWGSC RefSeq Data Access & Availability

FEEDBACK | CO

Pre-publication data access

IWGSC WGA v0.4: June 13, 2016

IWGSC RefSeq v1.0: January 14, 2017

Planned manuscript submission:

late spring / early summer 2017

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

IWGSC RefSeq Summary

- IWGSC has accomplished its goal of generating a reference sequence for hexaploid bread wheat in a little over 10 years after its initiation.

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

- Draft pseudomolecule sequences were made accessible pre-publication in summer 2016.

- The reference sequence will be submitted for publication in 2017.

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

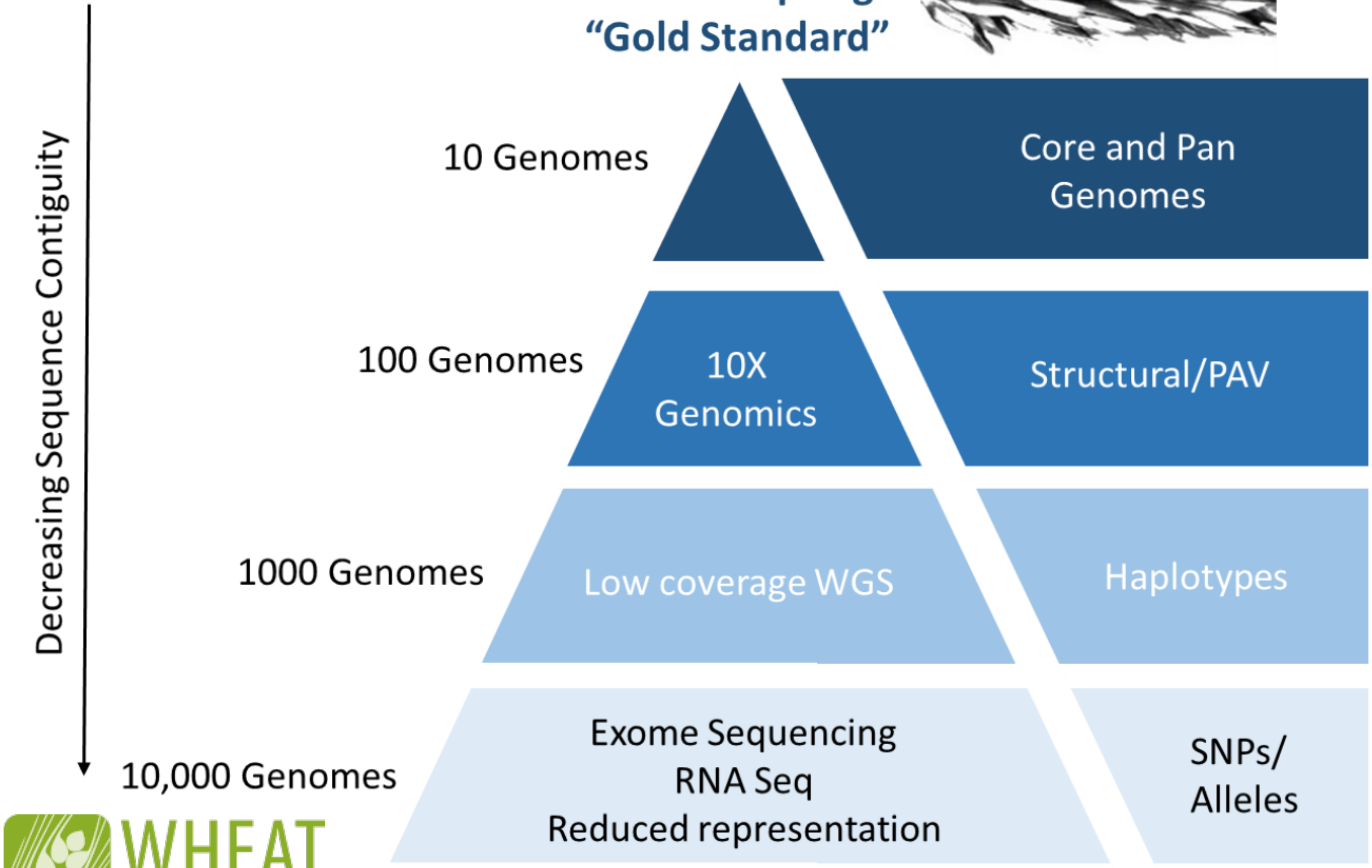
The reference sequence is expected to:

- reduce time and improve success of cloning genes and QTL
- provide unlimited access to high quality DNA markers for MAS and GS
- facilitate exploration of diversity in genetic resources for pre-breeding

I W G S C

Wheat genome sequencing – the future

Chinese Spring
“Gold Standard”



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

WHEAT INITIATIVE

Summary and Conclusions

- IWGSC RefSeqv1.0 is publicly available for BLAST and download today
- Similar resources are in progress for wheat species at all ploidy levels
- Additional high 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 IWGSC will stay committed to communicating progress in wheat research to the community and providing information about work in progress and data accessibility
- Academia and Industry must be aware of the developments and get their bioinformatics data storage and analysis infrastructure in place to efficiently accommodate multi-genome information for breeding and research

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

NIAS

Hirokazu Handa

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

Luigi Cattivelli

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

Elena Salina

Nikolai Ravin

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

Murdoch University

Rudi Appels

Sabancı Universitesi

Hikmet Budak

IPK Gatersleben

Nils Stein
Thorsten Schnurbusch

University of Saskatchewan

Curtis Pozniak
Andrew Sharpe

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**

BAC WGP Tags

Bayer CropScience **John Jacobs**

Genetic Maps

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

RH Mapping

University of Maryland **Vijay Tiwari**

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**

Tel Aviv University **Assaf Distelfeld**

illumina

Mike Thompson

IWGSC

Kellye Eversole
Jane Rogers

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

Manuel Spannagl, Klaus Mayer

Earlham Institute **David Swarbreck**

RNASeq

John Innes Centre
Unlocking Nature's Diversity

Cristobal Uauy

IWGSC Sponsors

Arcadia BIOSCIENCES

ACPFG

biogemma

Bayer CropScience

CSIRO

CIMMYT
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.

KeyGene

Limagrain
from earth to life

Together, we are
KANSAS WHEAT

UNIVERSITY OF SASKATCHEWAN
Crop Development Centre

RAGT

syngenta

CTAG²
Canadian Triticum Applied Genomics

IWGSC

Acknowledgements - WHEATSEQ

IPK
GATERSLEBEN

HelmholtzZentrum münchen
German Research Center for Environmental Health

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

Federal Office for
Agriculture and Food

by decision of the
German Bundestag