HelmholtzZentrum münchen
German Research Center for Environmental Health

PGSB Plant Genome and Systems Biology

www.wheatgenome.org

Reference gene prediction in the hexaploid wheat genome

Sven O. Twardziok, Klaus Mayer, Manuel Spannagl, IWGSC

PAG 2017

HELMHOLTZ ASSOCIATION

Objectives (I)

- **Single, high quality gene annotation for the IWGSC v1.0 Chinese Spring reference assembly**
  - Based on automated, efficient gene modeling tools
  - Compatible with future releases and manual curation
  - Agreed and “useful” gene identifiers
  - Classification of gene models based on their supporting evidence

Objectives (II)

- **Functional annotation of the wheat gene models**
  - Assign potential function to as many gene models as possible
  - Only assign reliable functional descriptors

  Automated functional annotation <-> experimentally verified functions

  - How to transfer existing knowledge on wheat gene functions to now available genomic resources?
  - Controlled vocabulary?

IWGSC gene prediction strategy

Two annotation teams

- PGSB
- INRA - Frederic Choulet, Helene Rimbert, Phillipe Leroy

One quality and consolidation team

- Earlham Institute (formerly TGAC) – David Swarbreck, Luca Venturini

Pipeline comparison and gene model consolidation is currently ongoing!!

Data for structural gene annotation – reference annotations / protein sequences

- *Arabidopsis thaliana* version ARAPORT11
- *Brachypodium distachyon* version 3.1
- Rice version MSU7
- *Sorghum bicolor* version 2.1
- *Seteriaca italica* version 3.1
- All triticeae protein sequences from UniProt (filtered for complete gene models, downloaded 16.10.2016)

Data for structural gene annotation – RNA-seq data sets

ID	#samples	#reads	#mapped	ratio	Availability	Source
E-MTAB-1729	60	2.27E+09	2.03E+09	0.90	public	EBI arrayexpress
E-MTAB-2137	30	2.10E+09	1.92E+09	0.91	public	EBI arrayexpress
E-MTAB-3103	9	5.21E+07	4.22E+07	0.81	public	EBI arrayexpress
NBS-LRRs	17	8.10E+08	7.46E+08	0.92	not public	JIC - Steuernagel
SRP045409	14	2.11E+09	1.76E+09	0.83	public	Sequence Read Archive
INRA 2014	30	3.47E+09	3.14E+09	0.90	Public	https://urgi.versailles.inra.fr
Earlham 2016	6	1.55E+09	1.50E+09	0.96	Public	Earlham Institute

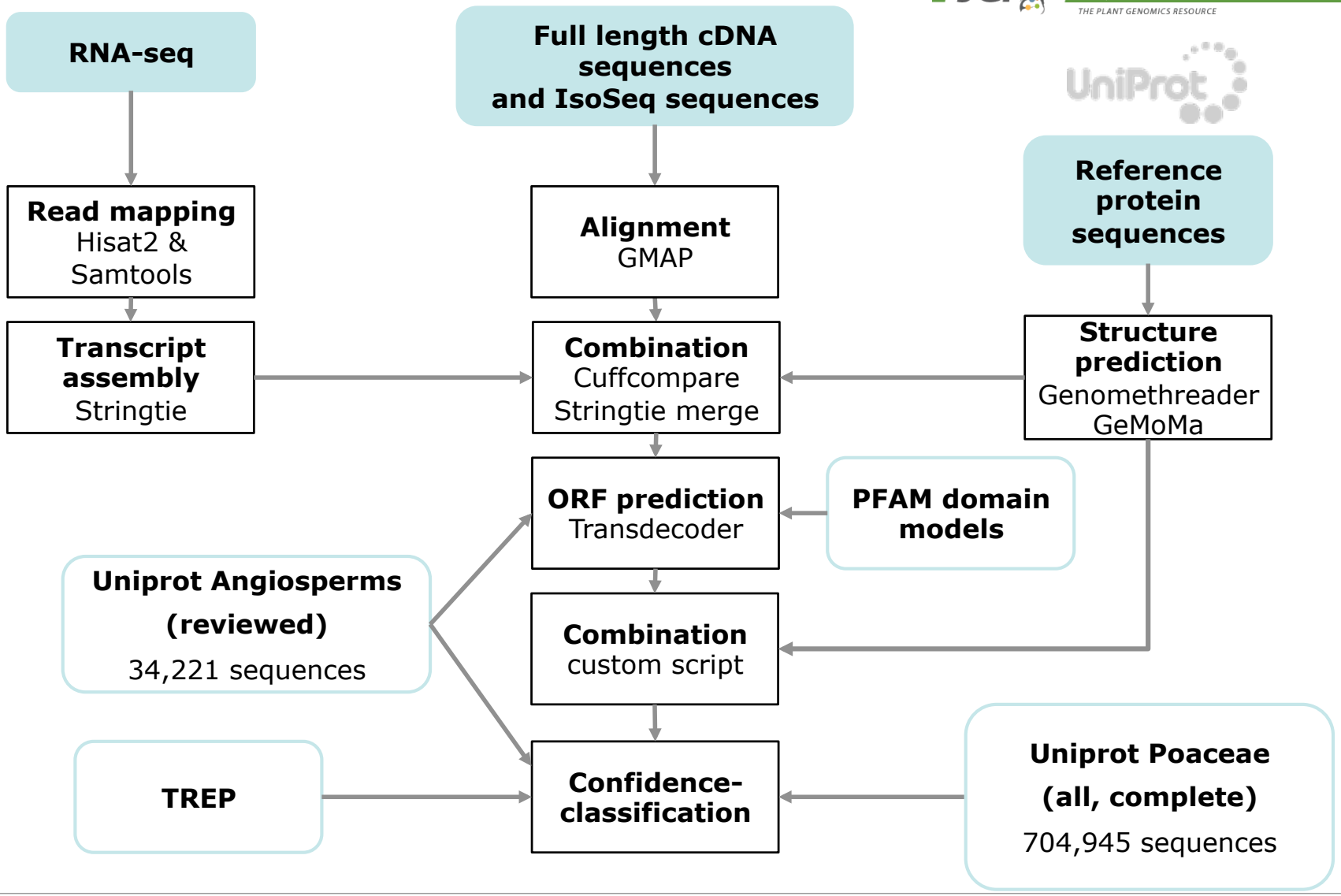
Data for structural gene annotation – PacBio IsoSeq reads

- PacBio isoform sequencing reads from 6 samples (leaf, root, seed, seedling, spike, stem), prov
- overall 817,892 sequences

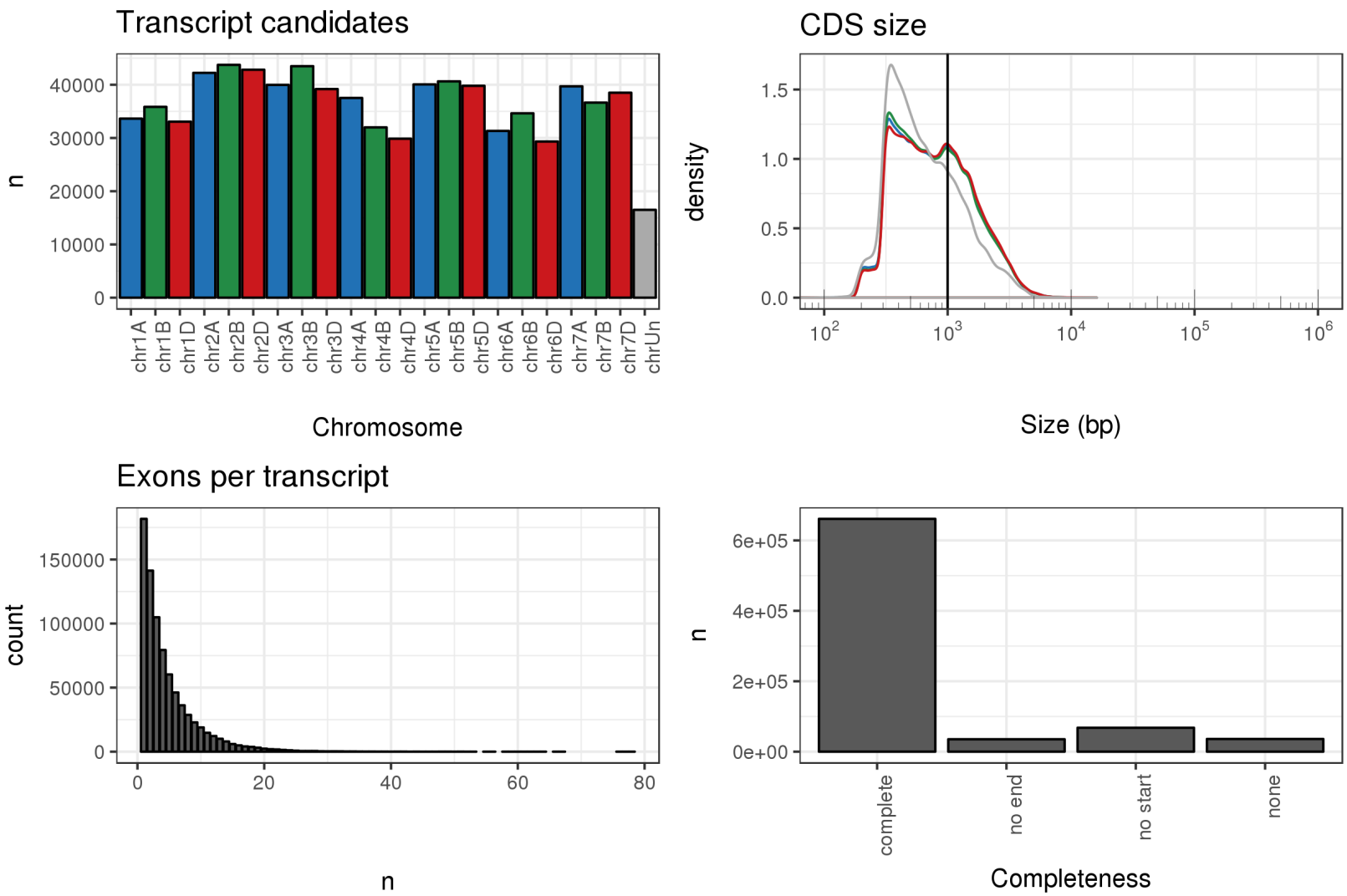
Earlham Institute
Decoding Living Systems

(Tom Barker and Luca Venturini)

PGSB gene annotation pipeline



Gene candidates



800,504 transcripts at 234,588 loci

Confidence classification

TREP: Hypothetical proteins ("PTREP") for the identification of divergent TEs
UniPoa: Uniprot Poaceae protein sequences
UniMag: Uniprot Magnoliophyta protein sequences (validated)

- Protein blast of protein candidates to three databases
- maximal query coverage for TREP
- maximal relative overlap for protein databases

HC1 : UniMag and complete

HC2 : not UniMag and (UniPoa and not TREP) and complete

LC1 : (UniMag or (UniPoa and not TREP)) and not complete

LC2 : not UniMag and not UniPoa and not TREP and complete

REP : not (UniMag and complete) and TREP

none: not UniMag and not UniPoa and not TREP and not complete

Gene prediction results

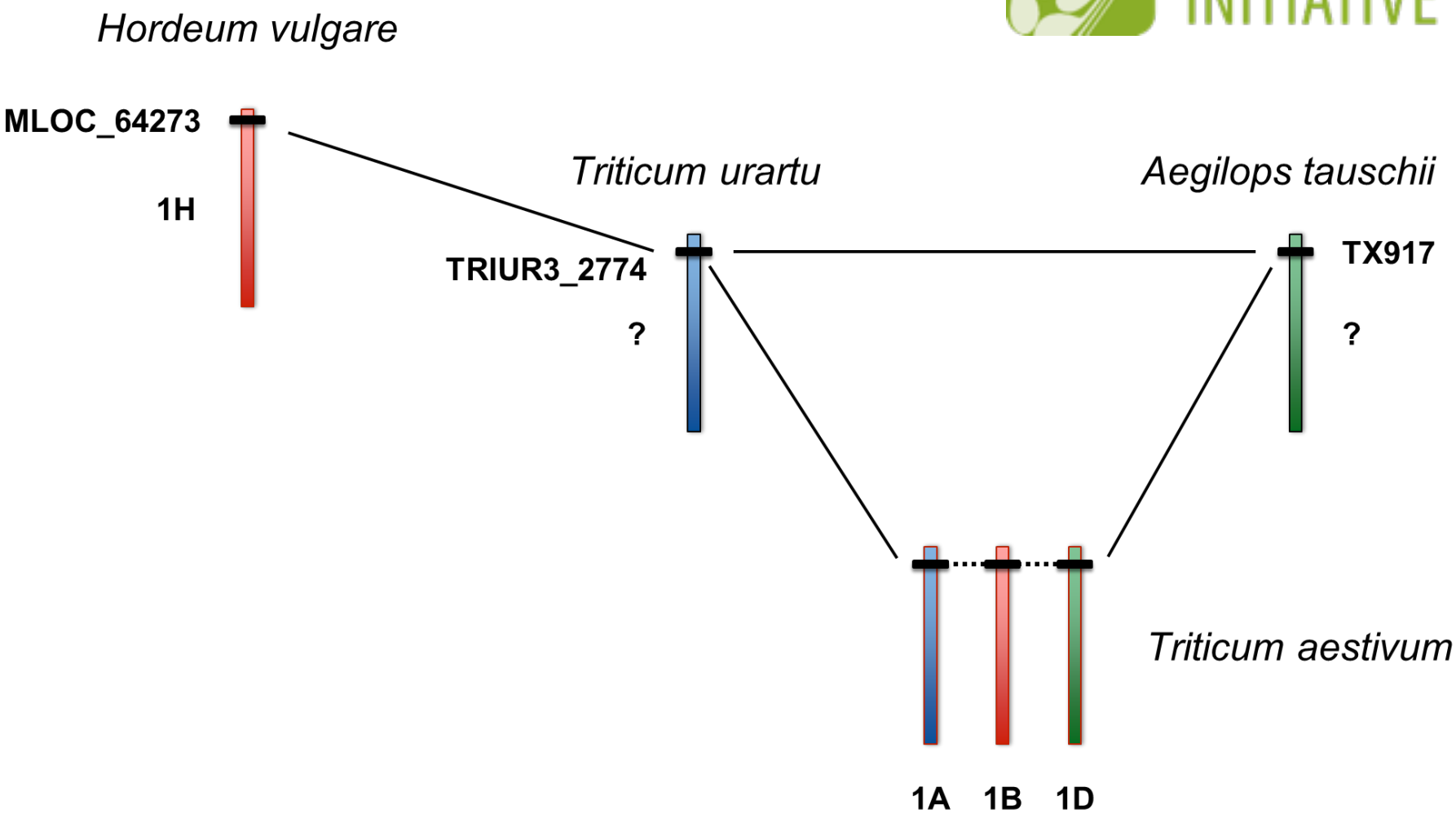
	HC	LC
Number of genes	104,696	100,947
Mean locus size (bp)	4,546	1,825
Median locus size (bp)	2,021	566
Number of single transcript genes	48,526	82,791
Number of multi transcript genes	56,170	18,156
Number of transcripts	297,971	134,126
Mean transcripts per gene	2.8	1.3
Mean transcript size (bp)*	1,258	611
Median transcript size (bp)*	1,071	453
Mean exons per transcript*	4.5	1.7
Median exons per transcript*	3	1
Number of single exon transcripts*	29,939	66,828
Number of multi exon transcripts*	74,757	34,119

* for one representative transcript per gene

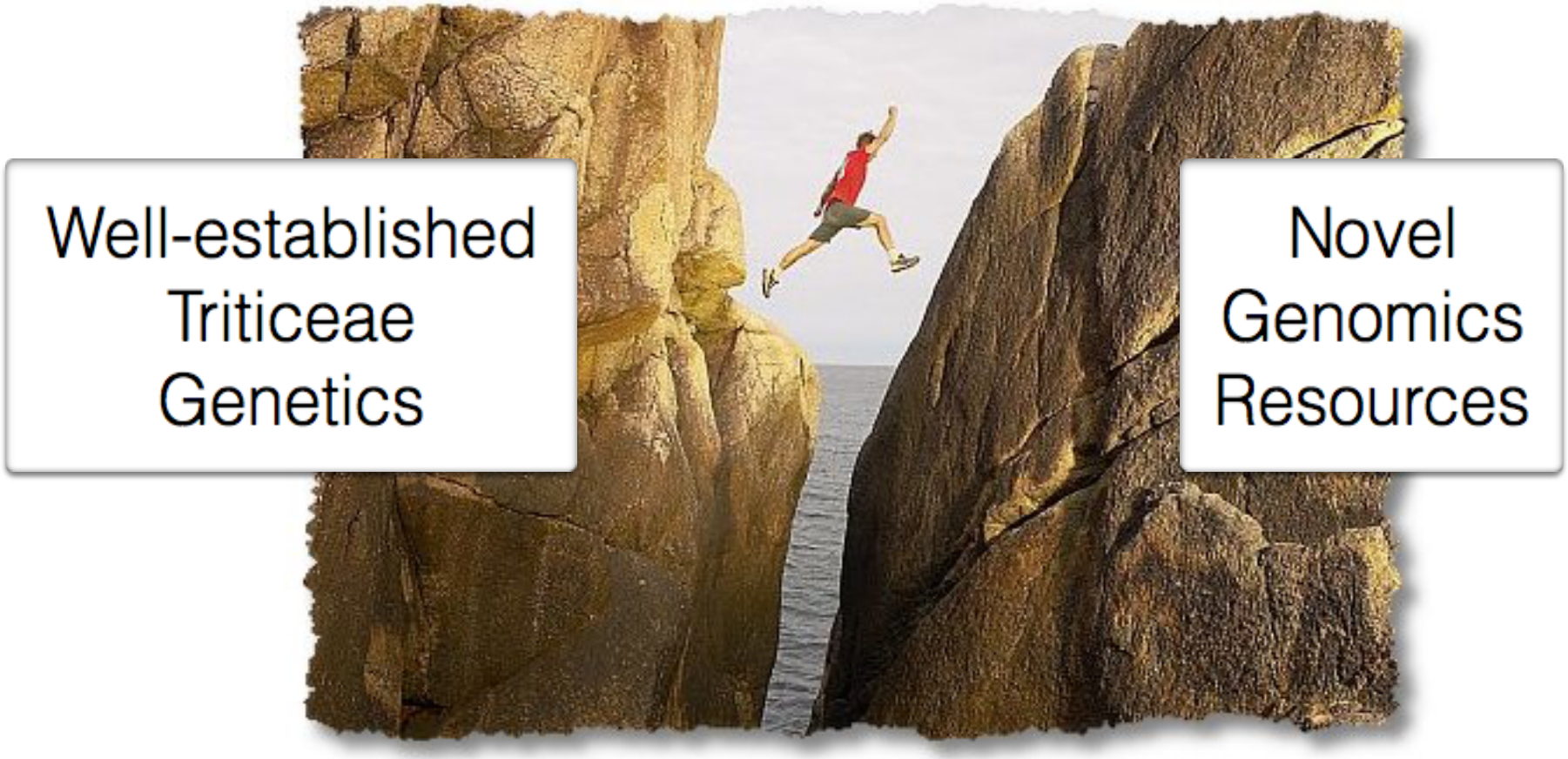
Gene names - *T. aestivum*

Slides from Mario Caccamo (NIAB)

Names in Context



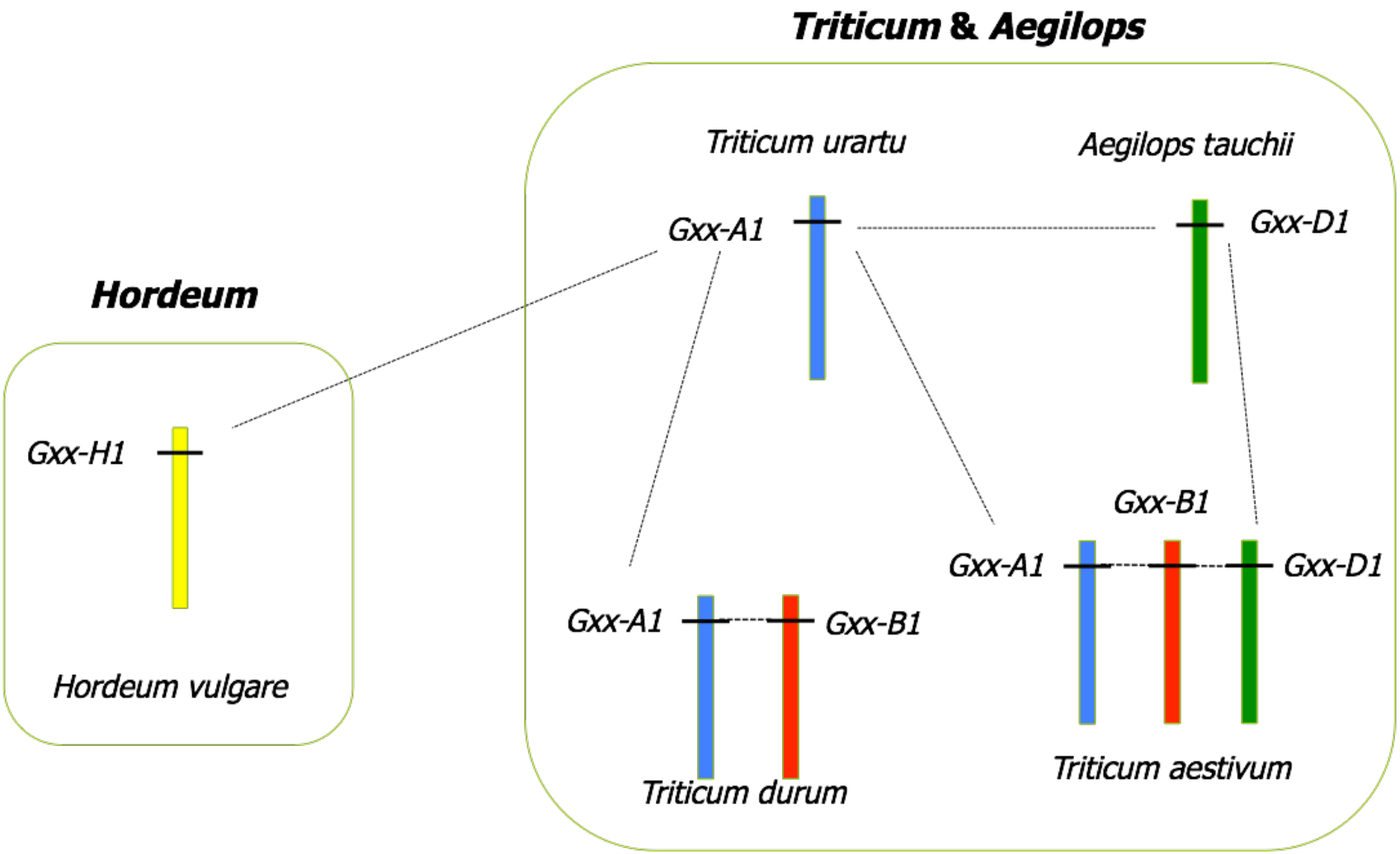
Bridging the Gap



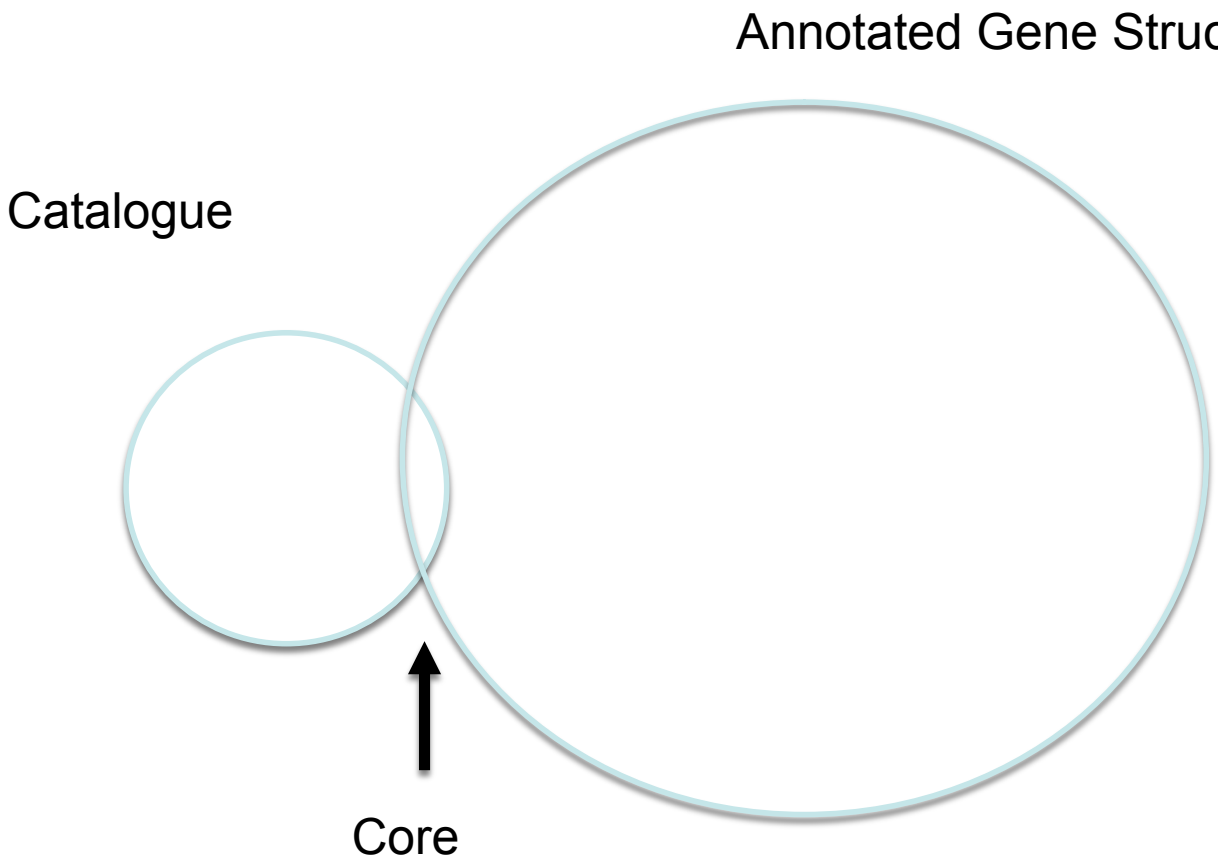
Workshop Objectives
Munich, October 2016

- How can the wheat gene catalog support the provision of gene symbols across the Triticeae tribe.
- Propose guidelines, criteria and schema for the annotation of gene symbols across the Triticeae species.
- Identify gaps in the current datasets to inform and guide projects for Triticeae species in the generation of genomics resources.

Scheme - Gene Symbol Gxx*



Genes vs Genes



Category I. The Catalogue provides sequences (either DNA or peptide) that could be used to **unambiguously** identify a gene structure in a sequence reference. This the priority category for the work on the nomenclature.

Moving forward

Submitted abstracts to several meetings

PAG poster - **P0350**

Group meets regularly over conference call

Funding for meeting in 2017

Look for opportunities to fund more work!

Acknowledgements

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.
IPK Gatersleben
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
IWGSC

Acknowledgements

IPK
Nils Stein
Martin Mascher

IPK GGR / GED
Axel Himmelbach

IPK CSF
Andreas Houben

IPK BIT
Sebastian Beier
Uwe Scholz

Earlham Institute
David Swarbreck
Luca Venturini
Matt Clark

PGSB
Sven Twardziok
Heidrun Gundlach
Georg Haberer
Thomas Lux
Marius Felder
Klaus Mayer

UMN
Gary Muehlbauer

IEB Olomouc
Hana Simkova
Jaroslav Dolezel

INRA
Fred Choulet
Phillipe Leroy
Etienne Paux
Michael Alaux
Hadi Quesneville
Helene Rimbert

IWGSC: Kellye Eversole, Jane Rogers

NIAB: Mario Caccamo

Bayer: Catherine Feuillet
Univ. of Zurich: Thomas Wicker & Beat Keller
Univ. of Udine: Michele Morgante et al.
KWS: V Korzun
CNRGV: H Berges, A Bellec
Haifa: A Korol, Z Frenkel
KSU: Jesse Poland
UIA,USDA: Roger Wise
USASK: Curtis Pozniak
U.Tel Aviv: Assaf Distelfeld

Thank you for your attention!