From fragments to the whole: progress in wheat genome annotation

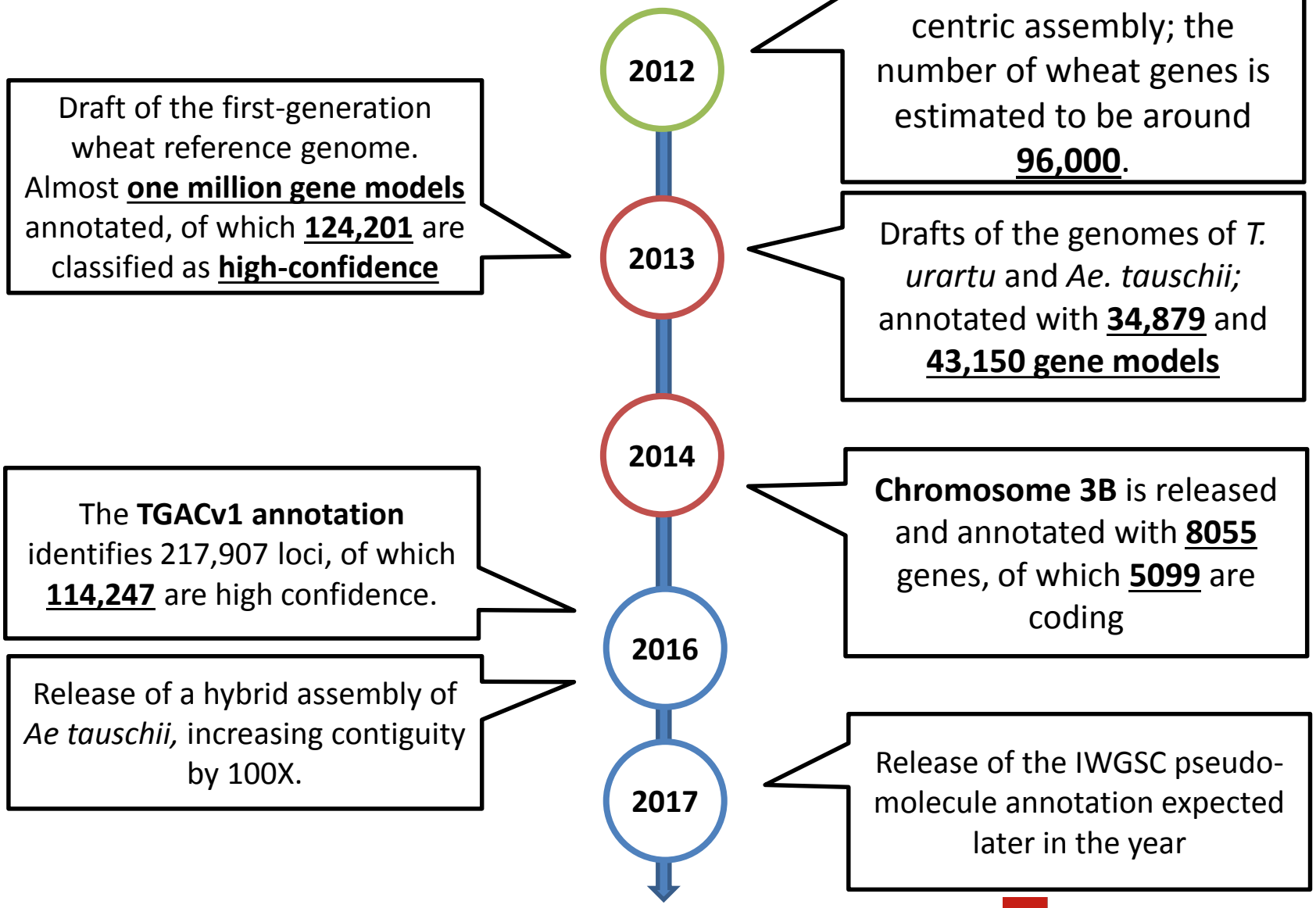
LUCA VENTURINI

Computational Biologist, Swarbreck group

Earlham Institute

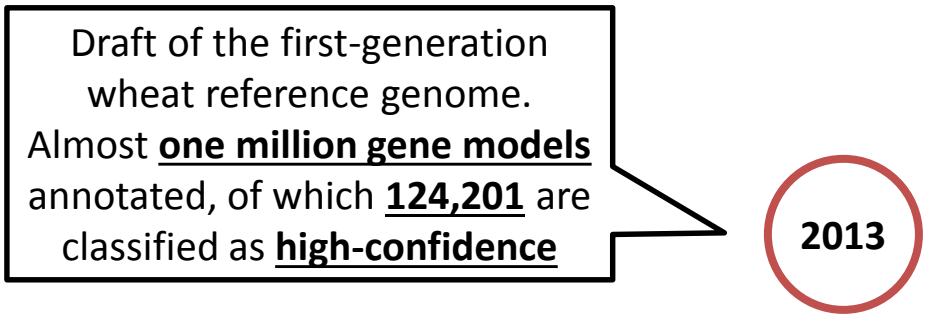
Decoding Living Systems

A timeline of wheat annotations



Earlham Institute

The CSS assembly: a first step towards a reference sequence



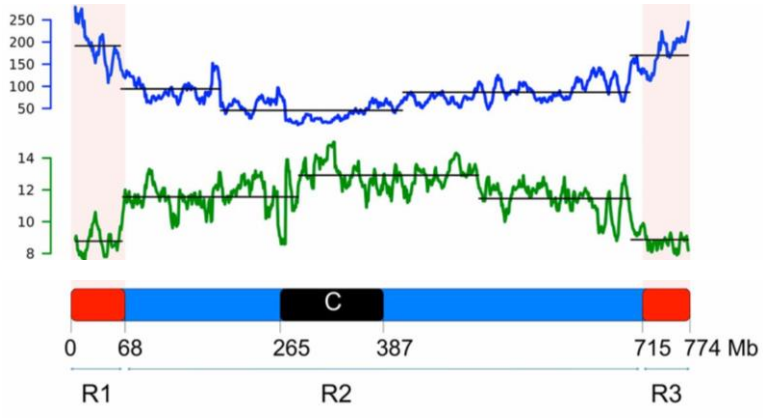
The first draft of the wheat genome, published in 2014, allowed for the first time to anchor a gene catalogue onto its chromosomal locations. It allowed for the first time to analyse genes in terms of their genomic context.

However, the assembly was still not optimal:

- Genes were fragmented (over 1 million low and high confidence genes)
- Out of the 133,090 high confidence genes, 124,201 (93%) could be assigned to a genomic location. However, only 44% were identified as likely full length.

The importance of a contiguous assembly: chr3B

- Having a pseudomolecule, it was possible to define loci completely, and categorise them
- Each gene is linked to its genomic context, allowing to analyse how genes segregate together
- Each gene is linked to long-distance markers, helping in GWAs and breeding



Gene density (in blue) and expression density (in green) along the chr3B pseudomolecule.

	All	Full genes	Pseudogenes
No. of genes	7264	5326	1938
Average size (bps) of coding sequences (± standard deviation)	1095 ± 807	1187 ± 821	840 ± 710
Average number of exons (± standard deviation)	4.2 ± 4.4	4.4 ± 4.6	3.6 ± 3.8
Gene density (kb^{-1})	107	145	400
No. of expressed genes	5185	4125	1060
% genes with alternative splicing	61	63	56

Choulet et al, Science 2014

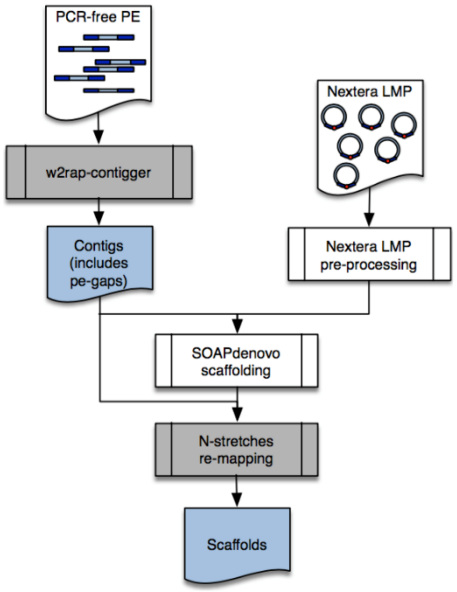
The TGACv1 assembly

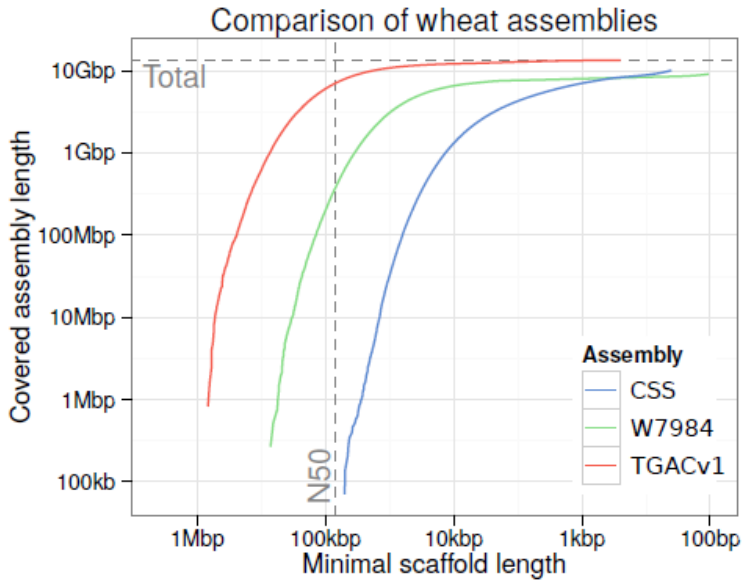
The Earlham Algorithm Development team

Bernardo Clavijo, Gonzalo Garcia Accinelli

Jon Wright

https://github.com/bioinfologics/w2rap





- At EI, we developed novel library preparation methods and novel algorithms to assemble genomes *de novo* quickly and reliably.
- Our assembly captured 60% more of the genomic content, compared with the previous best effort
- The method is completely open, reproducible, and fast.

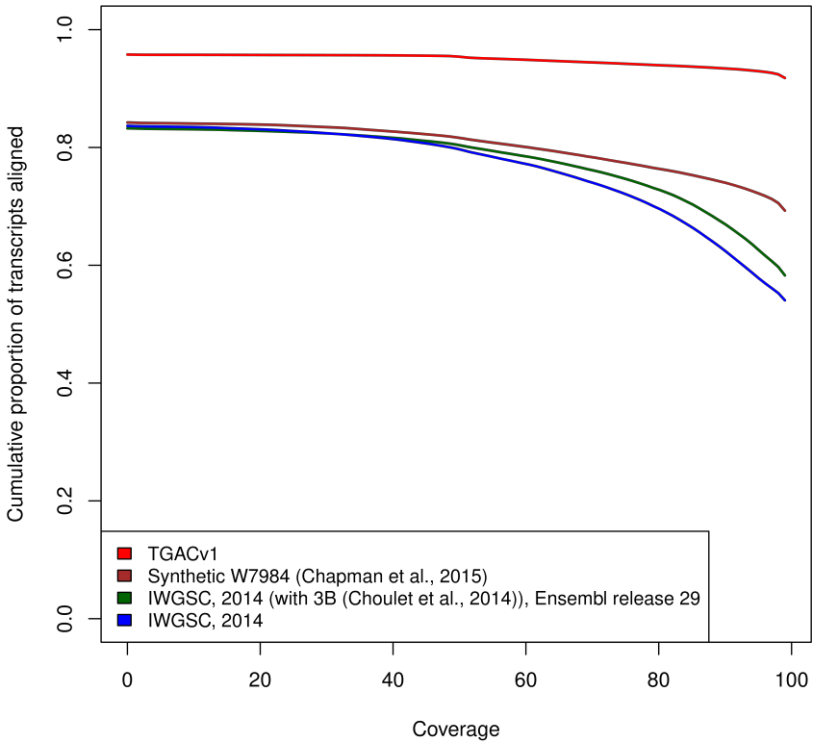
Generation of high-quality input transcript data for genome annotation

The TGACv1 annotation was based on the following sets of data:

- **Two billion Illumina reads from public available datasets**
- **800 million long, strand-specific Illumina dataset**
- **Over one million and a half PacBio full length cDNAs**
- **Protein models from six different species**

	# of sequences	Notes
Public Illumina reads	2,409,760,971	
Internal Illumina dataset	824,241,135	Strand-specific, 250bp PE
IsoSeq	1,509,322	
Protein models	316,385	Six different species

The high proportion of PacBio transcripts aligned to the TGACv1 assembly indicates an excellent representation of the gene space.

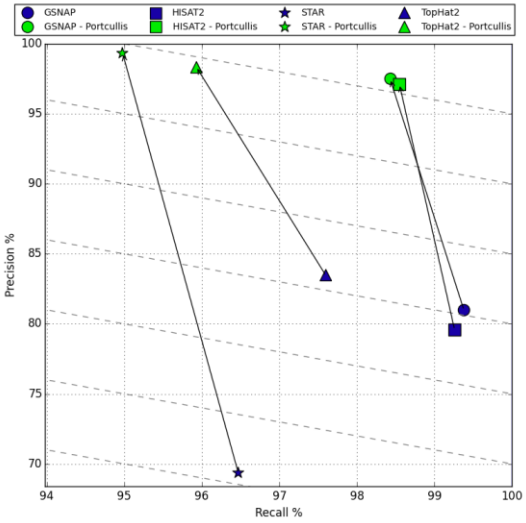


Earlham Institute

PORTCULLIS

https://github.com/maplesond/portcullis

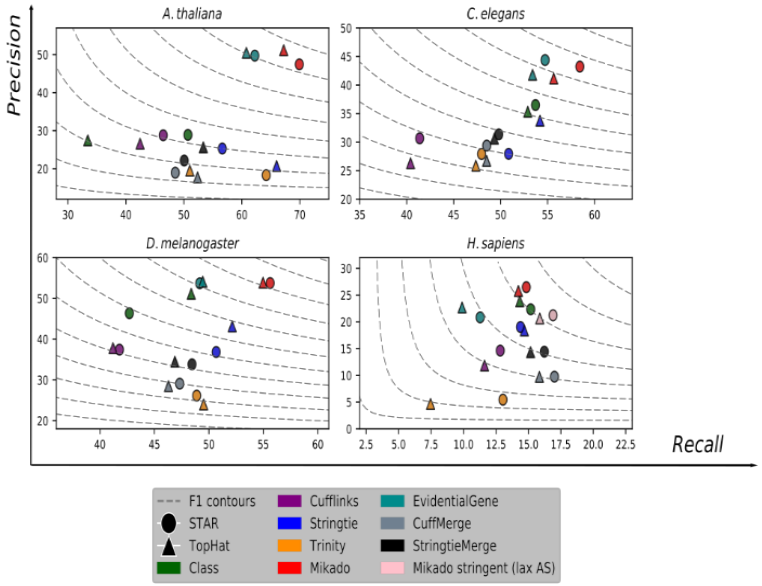
- It allows to distinguish between real and artifactual splicing junctions
- Especially important in datasets with deep sequencing and with polyploidy
- Machine-learning based



MIKADO

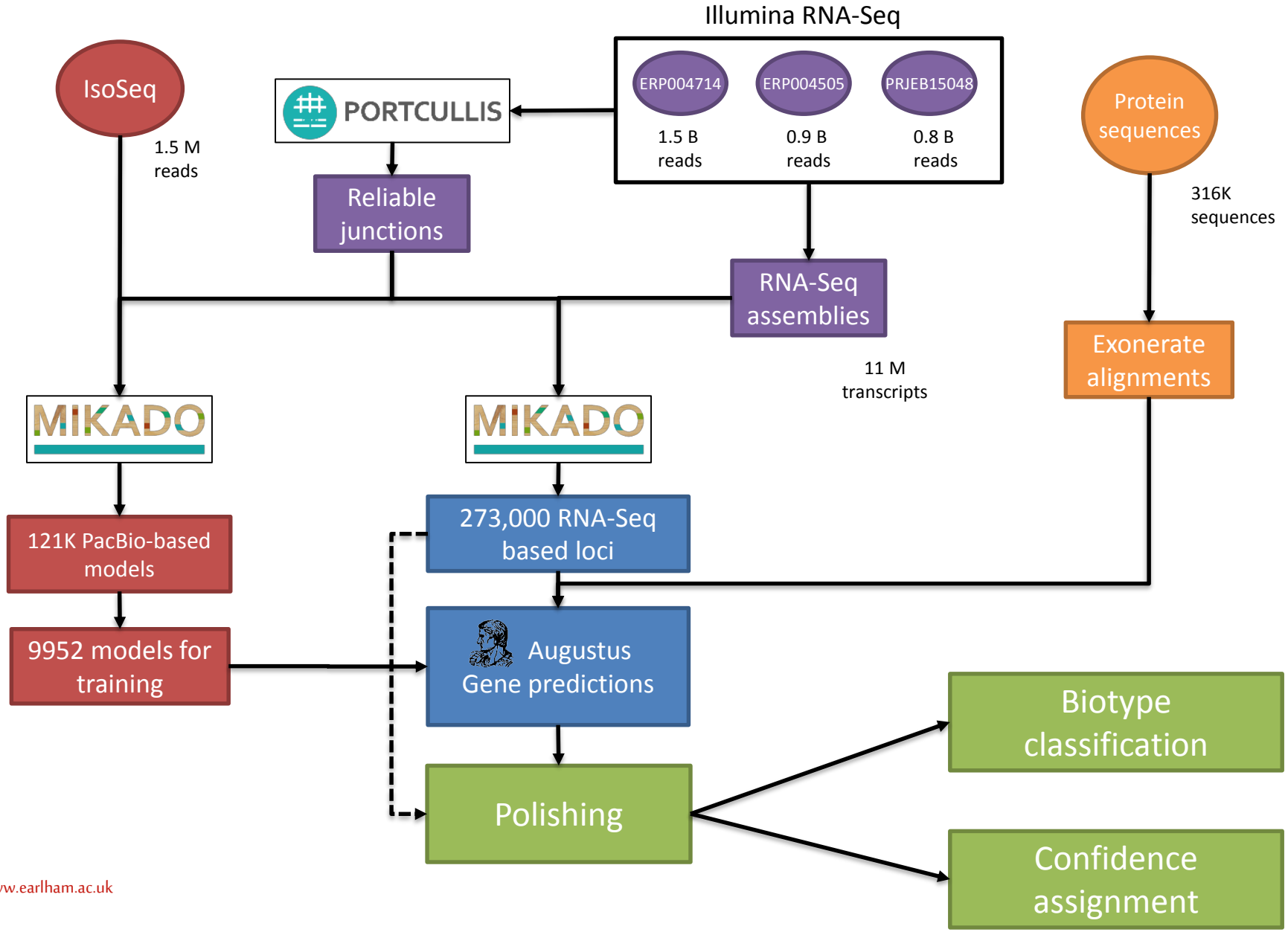
https://github.com/lucventurini/mikado

- Scores transcripts on the basis of intrinsic and extrinsic features (eg. Portcullis junctions)
- Robustly integrate multiple RNA-Seq assemblies
- Detects and resolves chimeric transcripts



Earlham Institute

ANNOTATION PIPELINE



Gene classification and confidence assignment

- We classified each model in two complementary ways:
- Verifying their support using known protein sequences (protein or homology rank)
- Verifying their concordance with RNA-Seq data (transcript or structural rank)

Rank level	Protein	Transcript
1	Over 80% homology	Full support from PacBio models
2	60-80% homology	Full support from Illumina models
3	30-60% homology	Structural congruence greater than 50%
4	Lower than 30% homology	Structural congruence lower than 50%
5	No homology with known proteins	No transcriptomic support

Confidence rankings of coding transcripts

Transcript count	protein rank	transcript rank
66404	P1	T1
43423	P1	T2
20937	P1	T3
10013	P1	T4
21469	P1	T5
3461	P2	T1
3545	P2	T2
3392	P2	T3
2084	P2	T4
6213	P2	T5
1813	P3	T1
4521	P3	T2
3995	P3	T3
3406	P3	T4
12210	P3	T5
781	P4	T1
3116	P4	T2
2846	P4	T3
2494	P4	T4
7484	P4	T5
2079	P5	T1
4638	P5	T2
3944	P5	T3
2915	P5	T4
12364	P5	T5

High confidence
Low confidence

Earlham Institute

Results: transcripts and gene classification

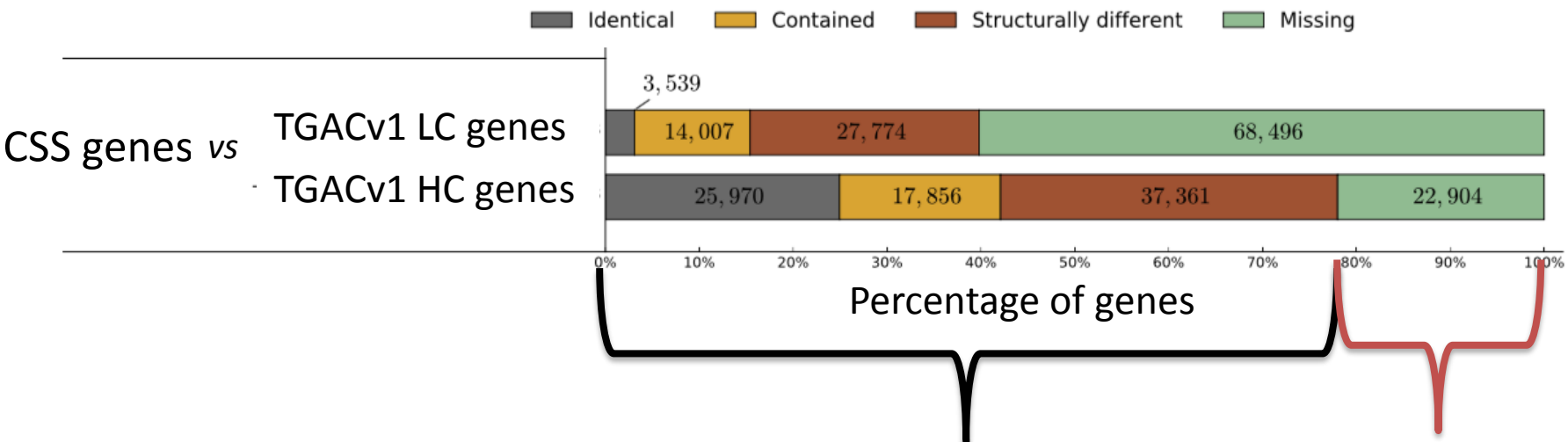
	All TGAC Models	mRNA HC	mRNA LC	ncRNA HC	ncRNA LC	Repeat-associated
Genes	217,907	104,091	83,217	10,156	9,933	10,510
Transcripts	273,739	154,798	85,778	11,591	10,438	11,134
Transcripts per gene	1.26	1.49	1.03	1.14	1.05	1.06
Transcript mean cDNA size (bp)	1,766.12	2,119.52	1,304.53	1,368.24	1,083.98	1,462.71
Exons per transcript	4.48	5.83	2.8	2.58	2.76	2.27
Exon mean size (bp)	394.15	363.73	465.27	530.25	392.24	644.09
Transcript mean CDS size (bp)	1,165.52	1,361.82	839.97	-	-	891.05
Mono-exonic transcripts	60,322 22.04%	19,034 12.30%	30,479 35.53%	3,061 26.41%	3,044 29.16%	4,704 42.25%
Genes with alternative splicing	32,616 14.97%	28,608 27.48%	2,033 2.44%	1,037 10.21%	460 4.63%	478 4.55%

The final set of TGACv1 annotations comprises 217,907 loci, of which 104,091 are classified as high-confidence protein coding genes. Compared with the previous CSS assembly, therefore, our annotation displays:

- A similar number of high confidence genes
- A much decreased number of low-confidence genes, many of which will probably be characterised as pseudogenes in the future
- The explicit characterization of long non-coding RNAs, which were absent from previous catalogues.

A more comprehensive and accurate wheat annotation

- We aligned the CSS/3B (IWGSC) gene models to the TGACv1 assembly and compared against the TGACv1 gene models.

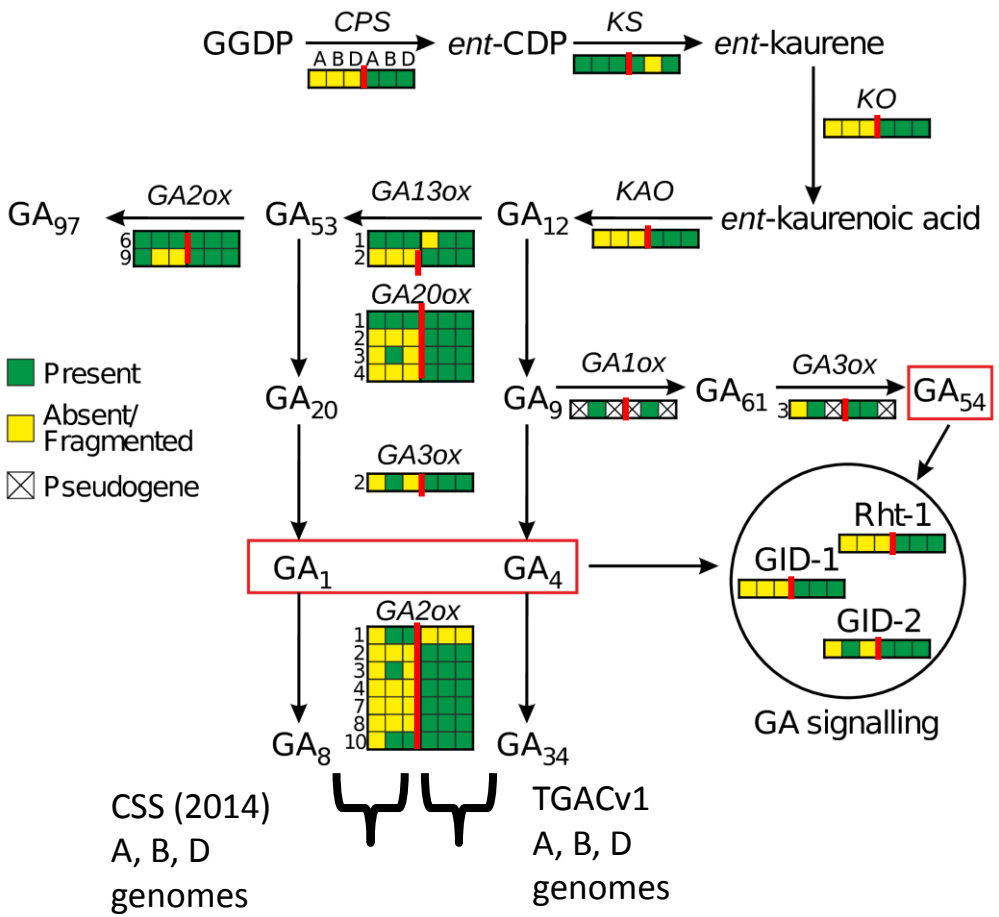


78% of TGACv1 genes overlapped with the full set of CSS genes (LC + HC).

- 6665 (29%) are fully supported by PacBio
- 19 810 (86%) have cross species protein similarity support

Earlham Institute

The new assemby and annotation allow to characterize whole gene families in detail



An example of a family which was reconstructed only partially in the CSS assembly is the gibberellic acid (GA) pathway, which plays a central role in plant development.

In the CSS assembly, out of 72 genes, only 23 (32%) could be found as full length sequences.

In the TGACv1 assembly, instead, 67 (93%) of the genes present.

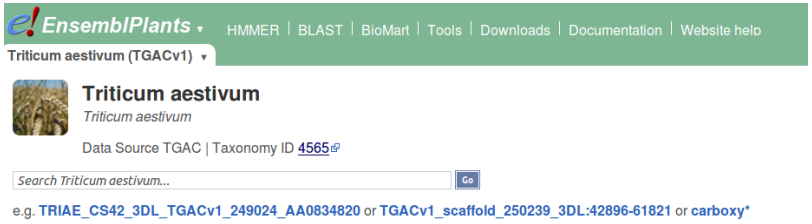
* Analysis courtesy of Andy Phillips

Earlham Institute

Data availability

The reference sequence and annotation can be retrieved at:

Earlham Institute

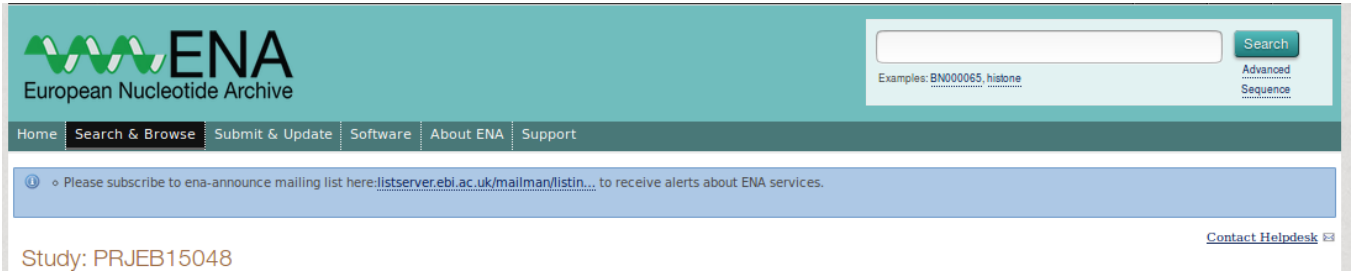


http://opendata.earlham.ac.uk/Triticum_aestivum/TGAC/v1/annotation/

http://plants.ensembl.org/Triticum_aestivum/Info/Index

The RNA sequencing reads can be downloaded from ENA (project PRJEB15048):

http://www.ebi.ac.uk/ena/data/view/PRJEB15048



Manuscript

An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations.

Bernardo J. Clavijo[1*], Luca Venturini[1*], Christian Schudoma[1], Gonzalo Garcia Accinelli[1], Gemy Kaithakottil[1], Jonathan Wright[1], Philippa Borrill[2], George Kettleborough[1], Darren Heavens[1], Helen Chapman[1], James Lipscombe[1], Tom Barker[1], Fu-Hao Lu[2], Neil McKenzie[2], Dina Raats[1], Ricardo H. Ramirez-Gonzalez[1], Aurore Coince[1], Ned Peel[1], Lawrence Percival-Alwyn[1], Owen Duncan[3], Josua Trösch[3], Guotai Yu[2], Dan Bolser[4], Guy Namaati[4], Arnaud Kerhornou[4], Manuel Spannagl[5], Heidrun Gundlach[5], Georg Haberer[5], Robert P. Davey[1,6], Christine Fosker[1], Federica Di Palma[1,6], Andrew Phillips[7], A. Harvey Millar[3], Paul J. Kersey[4], Cristobal Uauy[2], Ksenia V. Krasileva[1,6,8], David Swarbreck[1,6+], Michael W. Bevan[2+] and Matthew D. Clark[1,6+].

**contributed equally to this work*

+corresponding authors

Manuscript available on bioRxiv, and accepted for publication in Genome Research

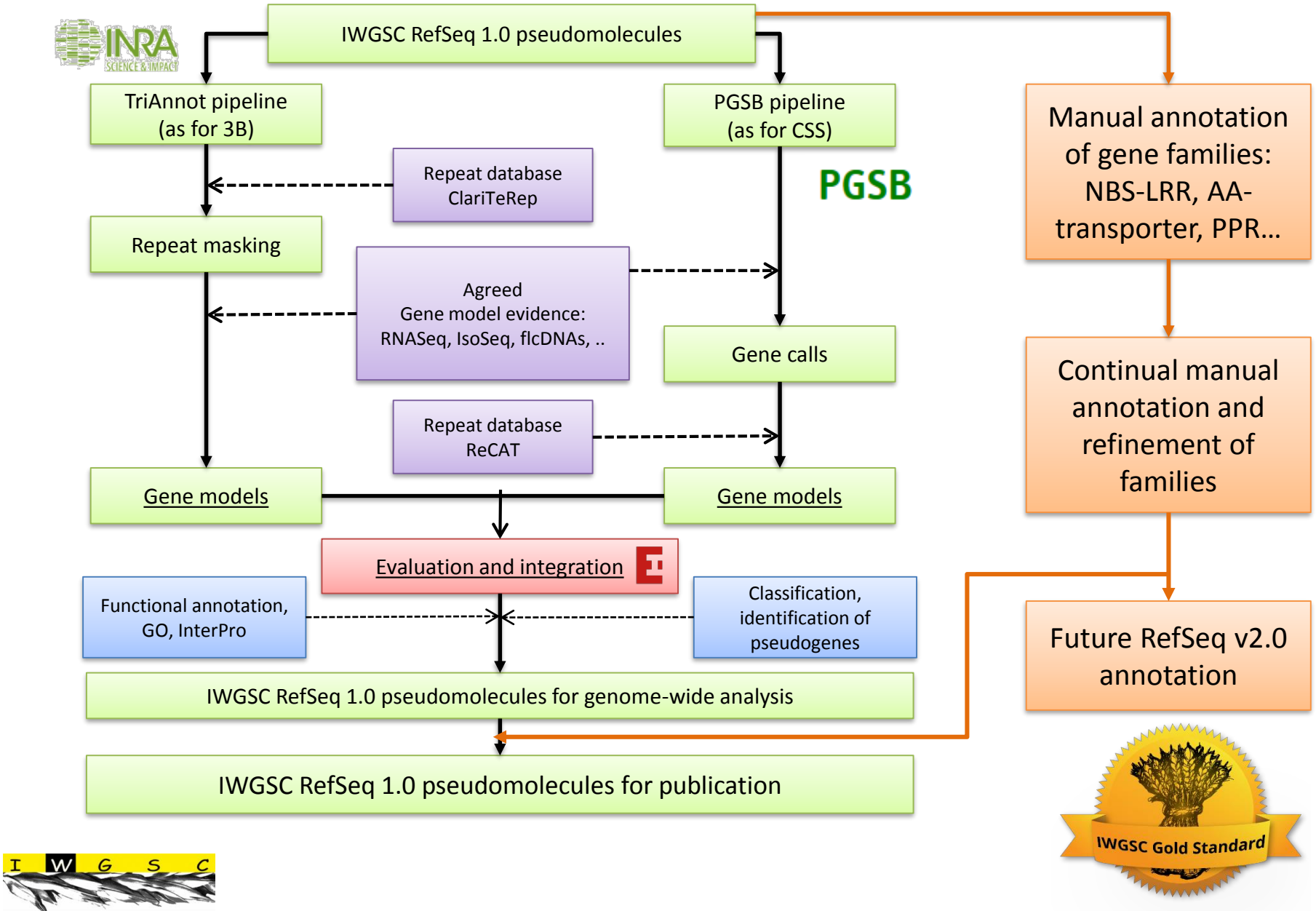
IWGSC RefSeq v1.0 Gene Prediction Strategy

- Coordination and oversight – Jane Rogers, IWGSC

- Two annotation teams:
  - **INRA-GDEC** – Frédéric Choulet, Hélène Rimbert, Philippe Leroy
  - **PGSB** – Sven Twarzdiok, Klaus Mayer, Manuel Spannangl

- Evaluation and integration team
  - **Earlham Institute** – David Swarbreck, Luca Venturini, Gemy Kaithakottil

IWGSC RefSeq v1.0 Annotation Approach



Characteristics of the two annotations

	PGSB		INRA			
	PGSB HC	PGSB All	INRA HC	INRA LC	INRA Pseudo	INRA all
Number of genes	104,696	205,643	65,884	41,342	73,044	180,270
Number of transcripts	297,971	432,097	65,884	41,342	73,044	180,270
Number of monoexonic genes	24,231	88,313	23,677	18,683	34,501	76,861
Average transcripts per gene	2.85	2.10	1.00	1.00	1.00	1.00
Average CDS length	1,384	1,145	1,110	1,231	766	998
CDS exons per transcript	6.32	4.96	4.08	4.04	2.84	3.57

The pipelines underlying the two annotations utilise different approaches and input data:

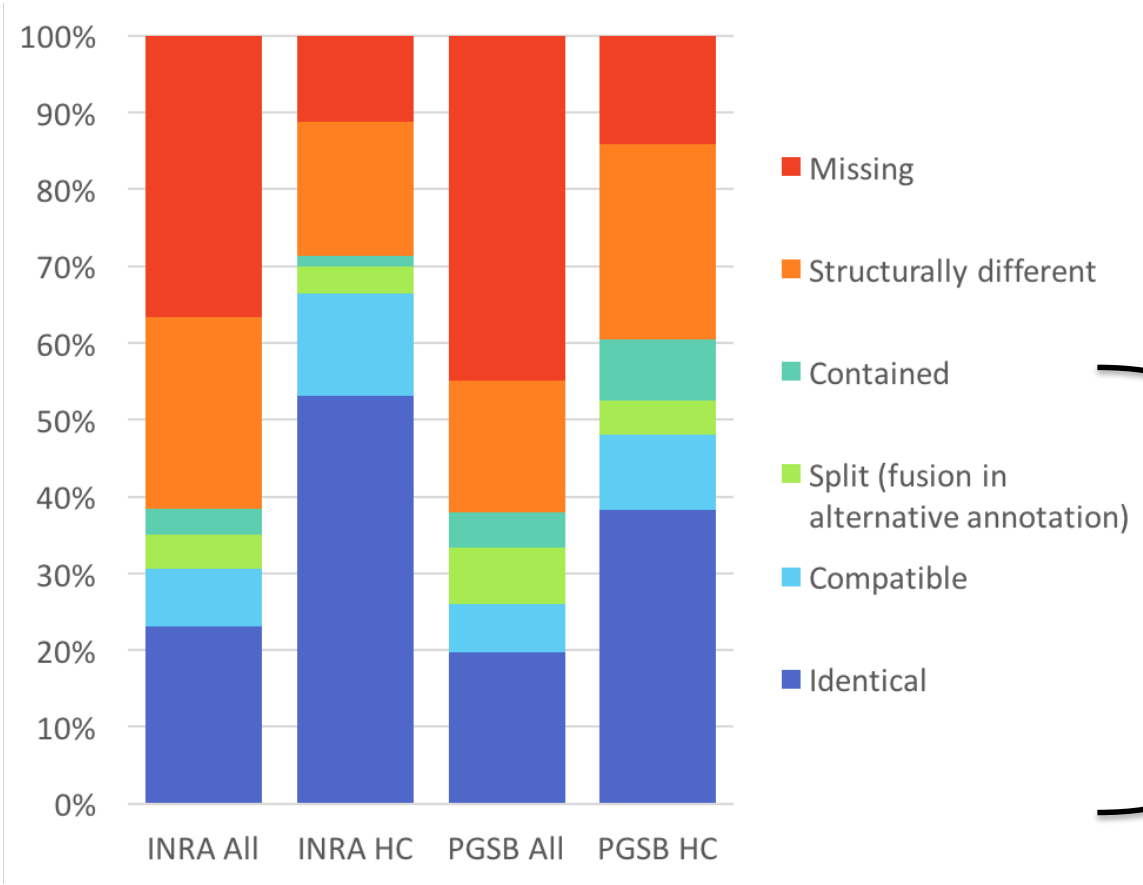
- The PGSB annotation:
  - is an alignment-driven approach (based on protein and wheat transcriptome data)
  - Captures splice variants
  - Classifies genes as high and low based on homology
- The INRA annotation:
  - is an evidence-guided approach (utilising gene predictions and aligned evidence)
  - provides an annotation of the UTR of transcripts
  - Identifies pseudogenes

Each approach will have specific strengths and weaknesses.

*preliminary results

Majority of high confidence genes are identified in both annotations

*preliminary results



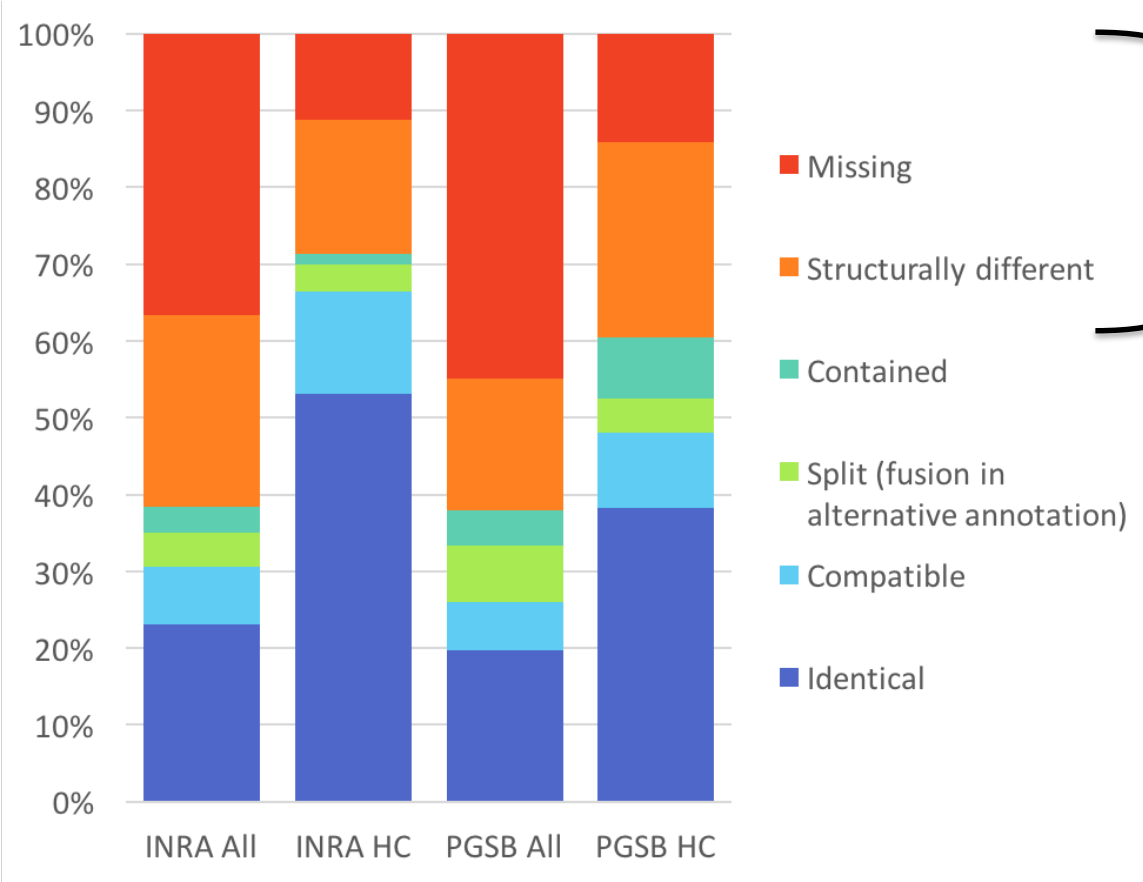
Over 80% of the genes defined as high confidence by INRA/PGSB overlapped a gene in the alternative annotation

A substantial percentage of genes show highly similar structures for the high-confidence gene sets (in blue here) ..

IWGSC RefSeq

*preliminary results

Some low confidence genes are present in only one annotation

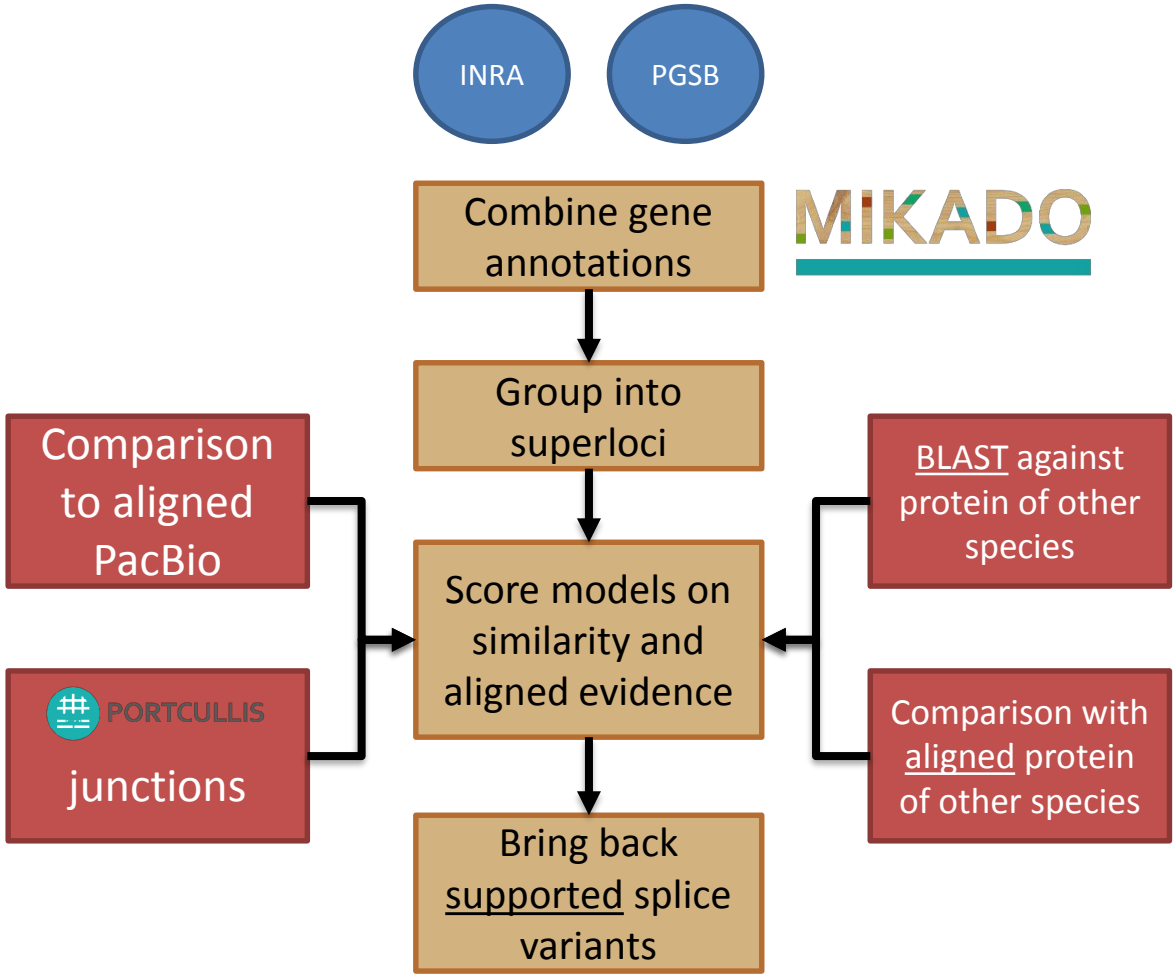


.. however, many low-confidence genes are present in only one of the two annotations (red fractions).

The differences between the two annotations reflect the challenge of annotating a transcriptionally complex polyploid species as well as differences in the datasets utilised and the annotation approaches (eg. sensitivity for detecting pseudogenes)

*preliminary results

An integrative approach to improve wheat annotation



By utilising an approach that selects gene models from across the two annotations we can exploit the strength of each annotation pipeline.

Gene models were selected based on the extent of evidence support and intrinsic gene characteristics

Including support from:

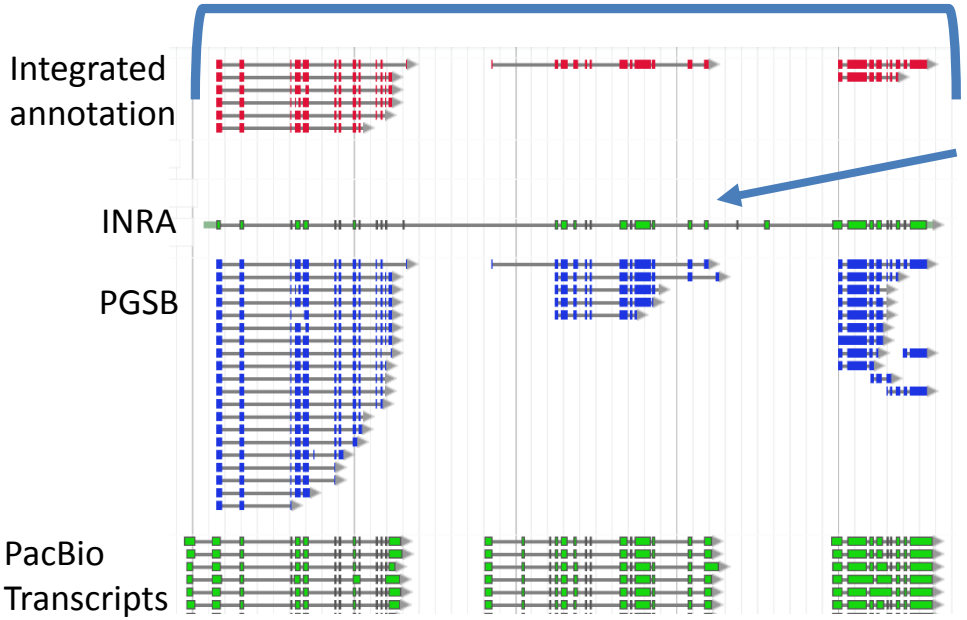
- PacBio transcripts
- High quality RNA-Seq assemblies
- Aligned proteins
- Validated junctions
- Sequence homology with known proteins

IWGSC

IWGSC RefSeq

Integrating both annotations allows us to identify and resolve errors

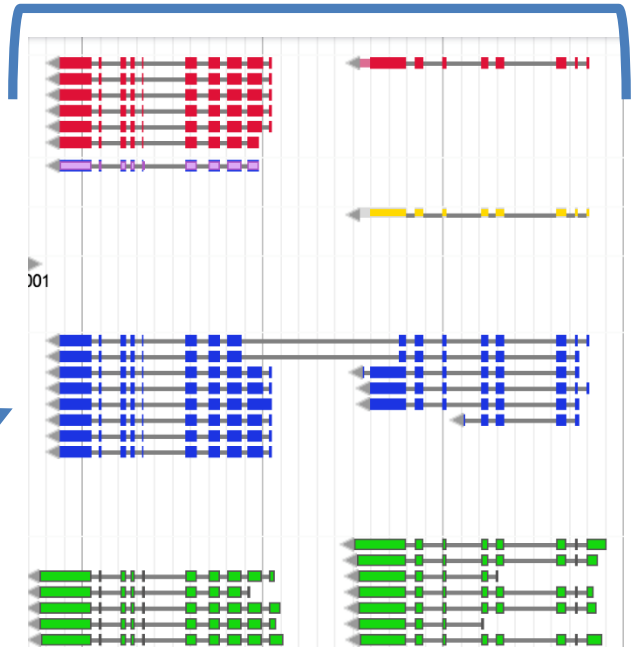
By “cherry picking” from the two annotations we can resolve issues where genes were incorrectly fused in one of the original annotations.



Correctly defined as 3 genes in the integrated gene set

Incorrect INRA fusion

Correctly defined as 2

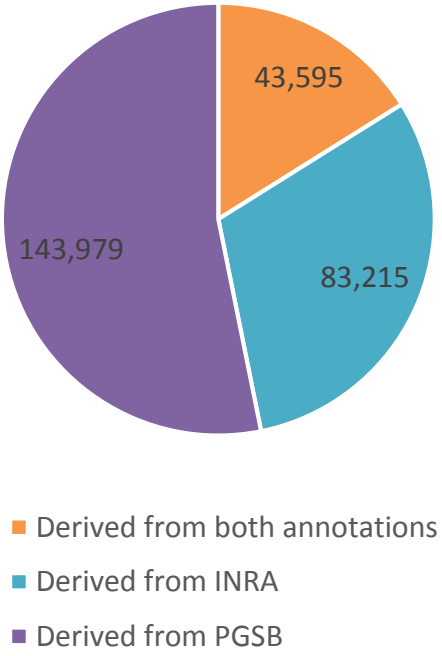


Incorrect PGSB fusion

IWGSC RefSeq

Preliminary IWGSC v1.0 gene annotation results

	High Confidence	Low Confidence	All
Number of genes	110,790	158,793	270,789
Number of transcripts	137,056	162,011	299,076
Number of monoexonic genes	31,660	104,364	136,263
Average transcripts per gene	1.24	1.02	1.10
Average CDS length	1,323.77	604.3	934.01
CDS exons per transcript	5.27	1.86	3.42
Average CDS exon length	251.29	325.15	273.02
Average CDS intron length	476.92	799.06	538.81



The integrated annotation combines the two annotation, removing redundancy and retaining the best models from both datasets.

Both PGSB and INRA contributed significantly to the final integrated annotation

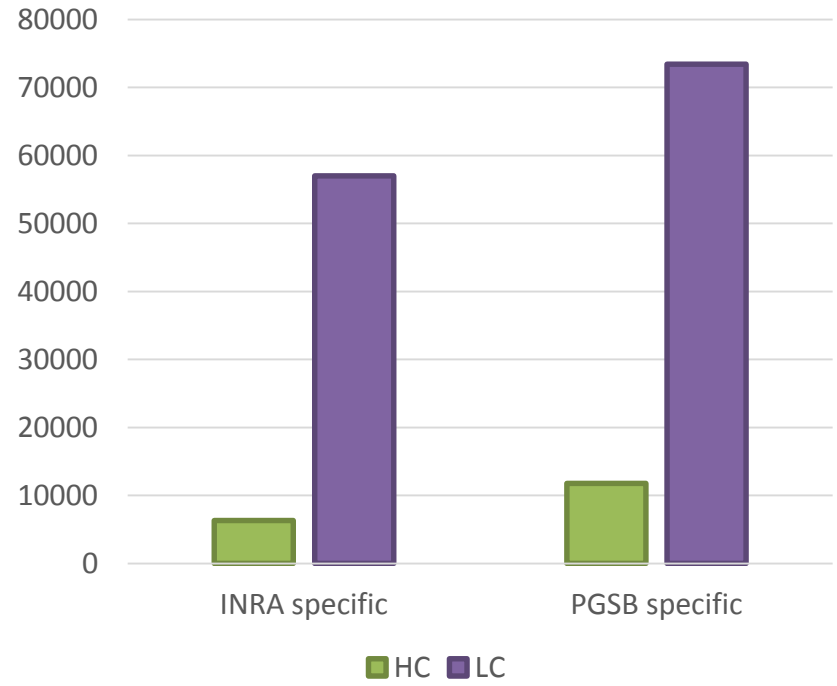
<u>The number of high confidence genes is similar to previous estimates, and to those found in the TGACv1 annotation.</u>

IWGSC RefSeq

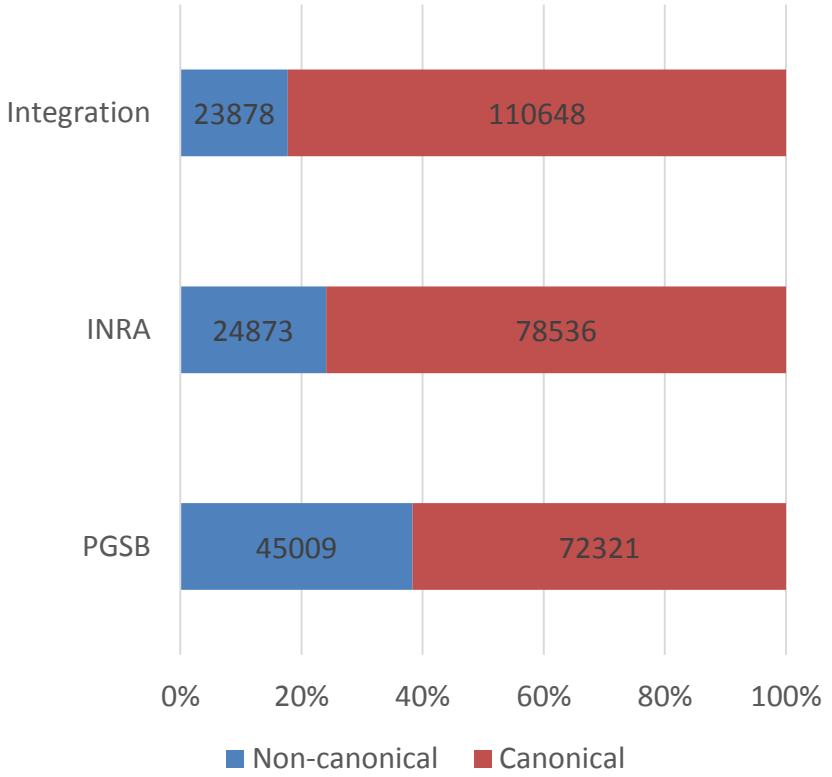
*preliminary results

The integration provides a more comprehensive representation of genes and potential pseudogenes

- The combined annotation contains 6320 high-confidence genes that were absent in PGSB, and 11,780 high-confidence genes that were not represented in the INRA annotation.

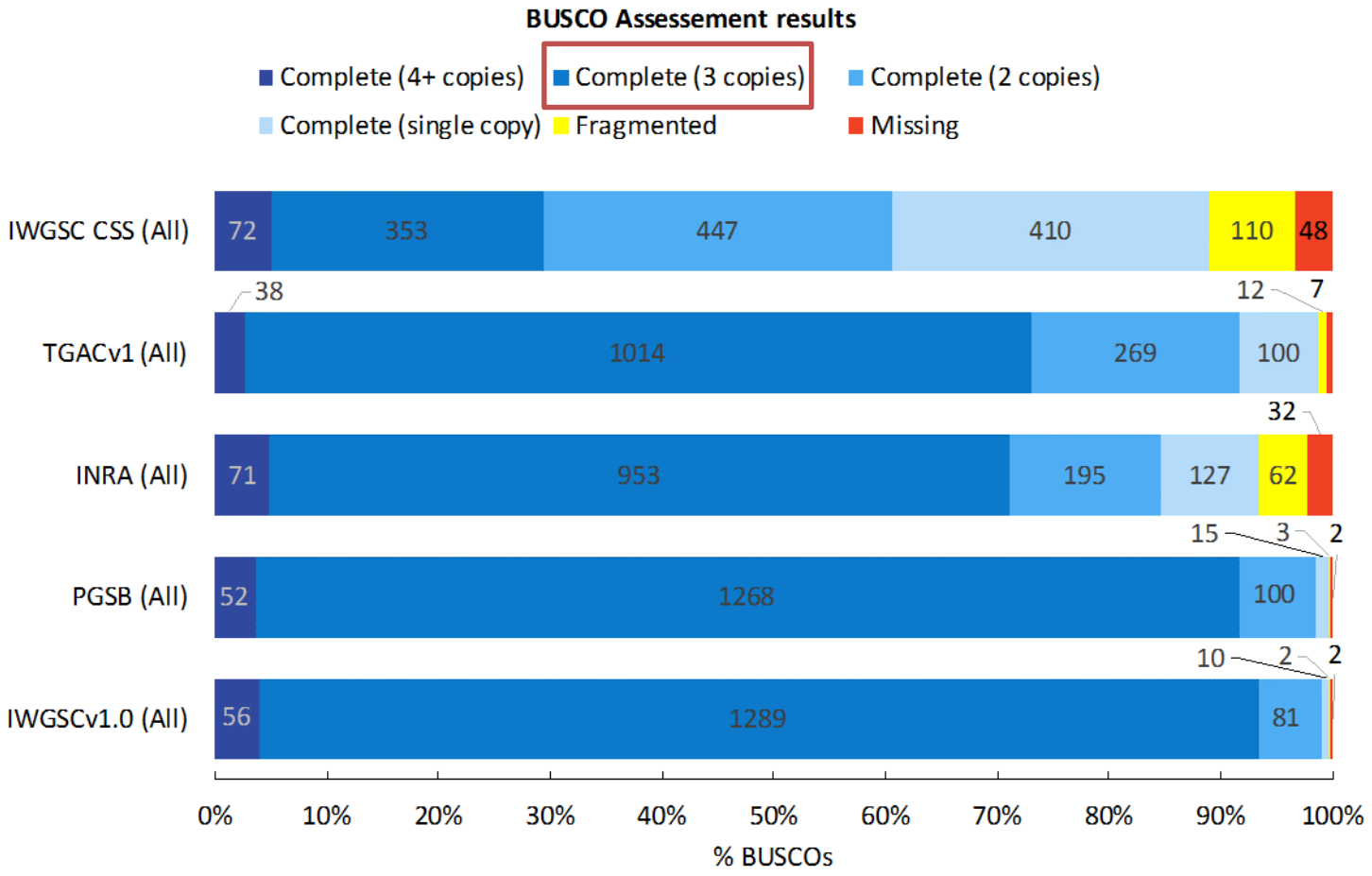


The integration removes many potential incorrect splice variants



*preliminary results

As annotations improve, more of the triplet highly-conserved homeologs are captured



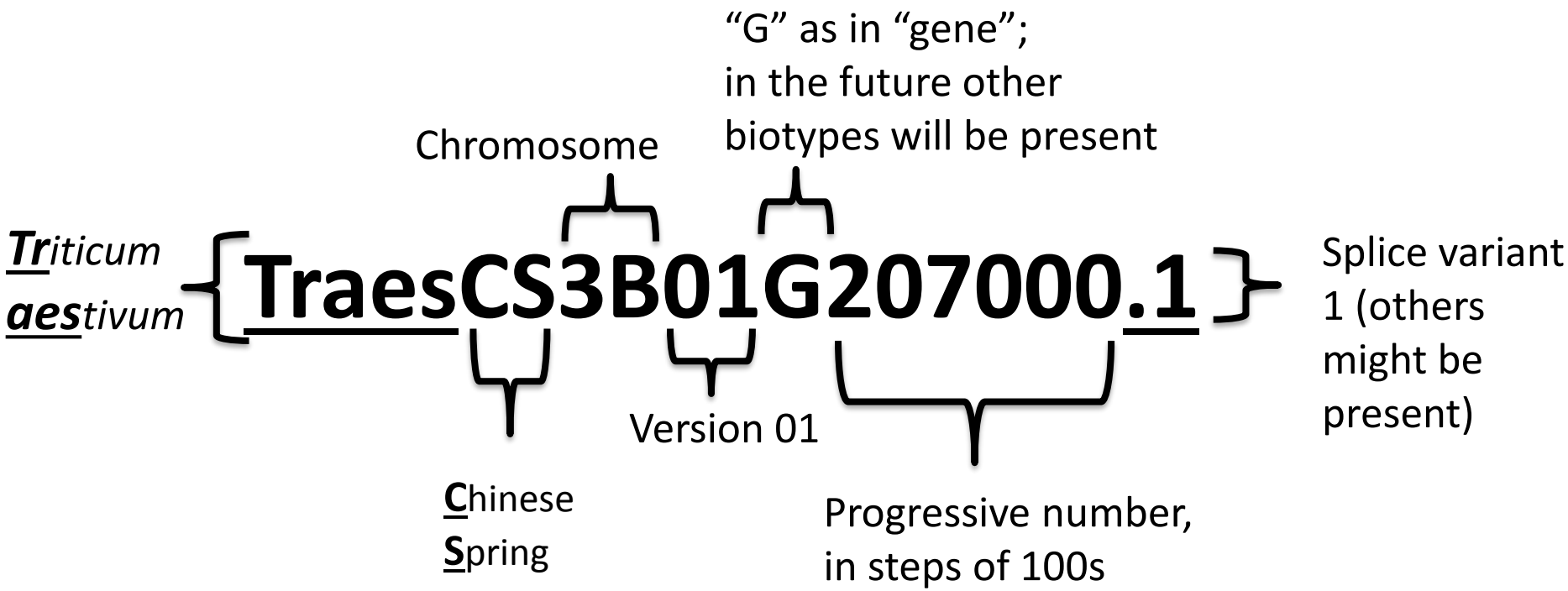
An important quality check for gene annotations is to verify that they contain expected genes that are conserved across lineages – and when the organism is polyploidy, that the correct number of copies are present.

For wheat, the expectation is to find **three copies** for each of these highly conserved genes.

*preliminary results

IWGSC RefSeq v1.0 gene identifiers

For each gene locus, one transcript/isoform/splice variant has been selected as the representative transcript, which is the longest transcript/isoform/splice variant in the respective confidence class.



Conclusions

- More contiguous and complete wheat genome assemblies have enabled the full gene space to be captured
- The forthcoming IWGSC gene annotation will be the most comprehensive wheat annotation to date
- This annotation will be instrumental to perform global analyses such as:
  - Marker positioning for GWAS and breeding
  - Evolution analyses on many gene families
  - Gene regulation across multiple tissues
  - Defining precisely translocation events among different chromosome arms
- IWGSC RefSeq v1.0 focuses mainly on coding genes, however, in future releases we can expect a better definition of other important genomic elements such as:
  - Long and short non-coding RNAs
  - Pseudogenes
  - Expansions of the splicing isoform catalogue
  - Manual revision of specific gene families
- Finally, this assembly and annotation will be the cornerstones upon which future IWGSC projects for high quality functional annotation and for resequencing the breadth of global germplasm diversity

I W G S C

IWGSC RefSeq

Acknowledgments

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

IWGSC Chromosome Leaders:

UMR 1095 GDEC INRA — Etienne Paux, Frédéric Choulet

K-STATE Kansas State University — Bikram Gill

Murdoch University — Rudi Appels

Institute of Experimental Botany of the AS CR, v. v. i. — Jaroslav Dolezel, Hana Simkova, Miroslav Valarik, Jan Bartos

Bayer CropScience — Catherine Feuillet, John Jacobs

Sabancı Üniversitesi — Hikmet Budak

IPK Gatersleben — Nils Stein, Thorsten Schnurbusch

Universität Zürich[UZH] — Beat Keller

University of Saskatchewan — Curtis Pozniak, Andrew Sharpe

NIAS — Hirokazu Handa

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

University of Haifa — Abraham Korol

Punjab Agricultural University — Kuldeep Singh

NORTHWEST A&F UNIVERSITY — Song Weining

The Federal Research Center Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences — Elena Salina

Norwegian University of Life Sciences — Odd-Arne Olsen

Earlham Institute — Matt Clark

Nikolai Ravin

I W G S C

IWGSC RefSeq

IWGSC RefSeq v1.0 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 — Jesse Poland

RH Mapping

University of Maryland — Vijay Tiwari

WGA PIs

IPK Gatersleben — Nils Stein

University of Saskatchewan
Curtis Pozniak
Andrew Sharpe

K-State — 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

PGSB
Manuel Spannagl, Klaus Mayer

Earlham Institute — David Swarbreck

RNASeq

John Innes Centre, Unlocking Nature's Diversity — Cristobal Uauy

IWGSC

IWGSC RefSeq

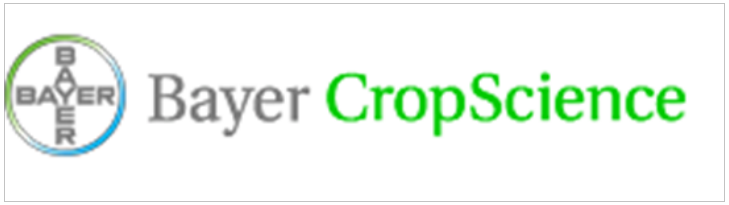
IWGSC RefSeq v1.0 Annotation Team

- Coordination and oversight – Jane Rogers, IWGSC

- Two annotation teams:
  - **INRA** – Frédéric Choulet, Hélène Rimbert, Philippe Leroy
  - **PGSB** – Sven Twarzdiok, Klaus Mayer, Manuel Spannangl

- Evaluation and integration team
  - **Earlham Institute** – David Swarbreck, Luca Venturini, Gemy Kaithakottil

Thanks to IWGSC Sponsors!



Acknowledgments

Earlham Institute

- Bernardo Clavijo and his team
  - Gonzalo Garcia Accinelli
  - Jon Wright
- David Swarbreck and his team
  - Gemy George Kaithakottil
  - Daniel Mapleson
- Ksenia Krasileva and her team
  - Christian Schudoma
  - Dina Raats
- Matt Clark and his team
  - George Kettleborough
  - Aurore Coince
  - Ned Peel
  - Lawrence Percival-Alwyn
- Darren Heavens and his team
  - James Lipscombe
- Helen Chapman
- Tom Barker
- Robert Davey
- Christine Fosker
- Federica di Palma

John Innes Centre
Unlocking Nature's Diversity

- Cristobal Uauy and his team
  - Philippa Borrill
  - Ricardo H. Ramirez-Gonzalez
- Mike Bevan and his team
  - Fu Hao-Lu
  - Neil Mckenzie
- Guotay Yu

plant energy biology
ARC Centre of Excellence

- Harvey Millar and his team
  - Owen Duncan
  - Joshua Troesch

EMBL

- Paul Kersey and his team
  - Dan Bolser
  - Guy Namaati
  - Arnaud Kerhornou

PGSB

- Manuel Spannangl
- Heidrun Gundlach
- Georg Haberer

ROTHAMSTED RESEARCH

- Andy Phillips