

Use of PacBio long reads to improve assembly and annotation of prolamins in the wheat reference genome

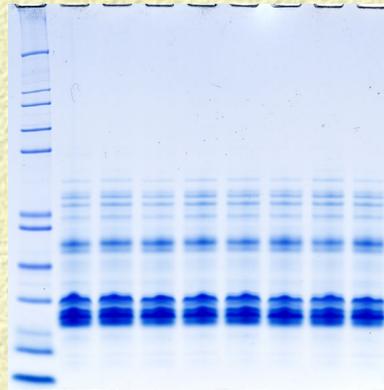
Yong Q. Gu
USDA-Agricultural Research Service
Western Regional Research Center
Albany, California 94710, USA



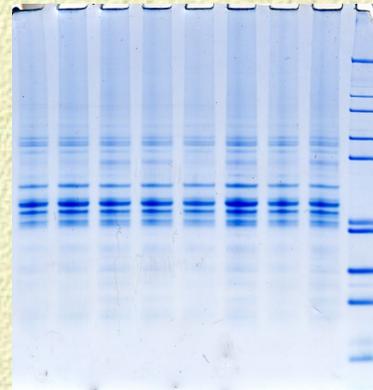
Wheat prolamins



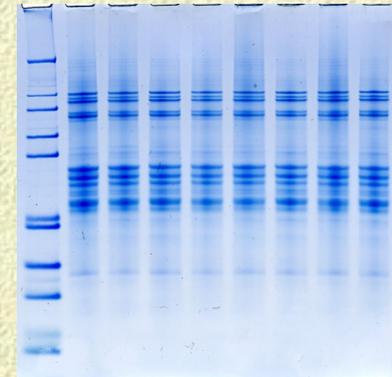
Globulins – 10%



Gliadins – 40%

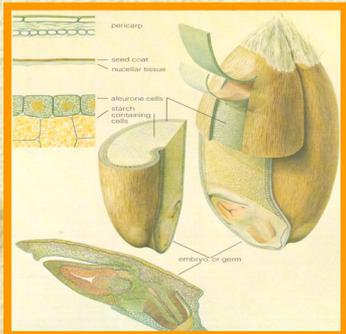


Glutenins- 50%

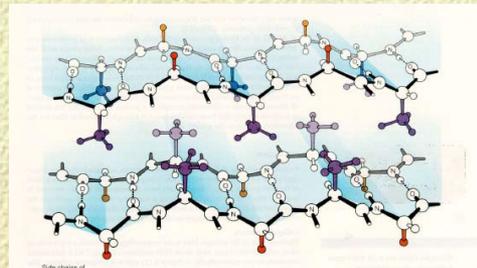


α -gliadin, γ -gliadin, ω -gliadin

HMW-glutenin, LMW-glutenin



Wheat grain



Wheat protein polymer

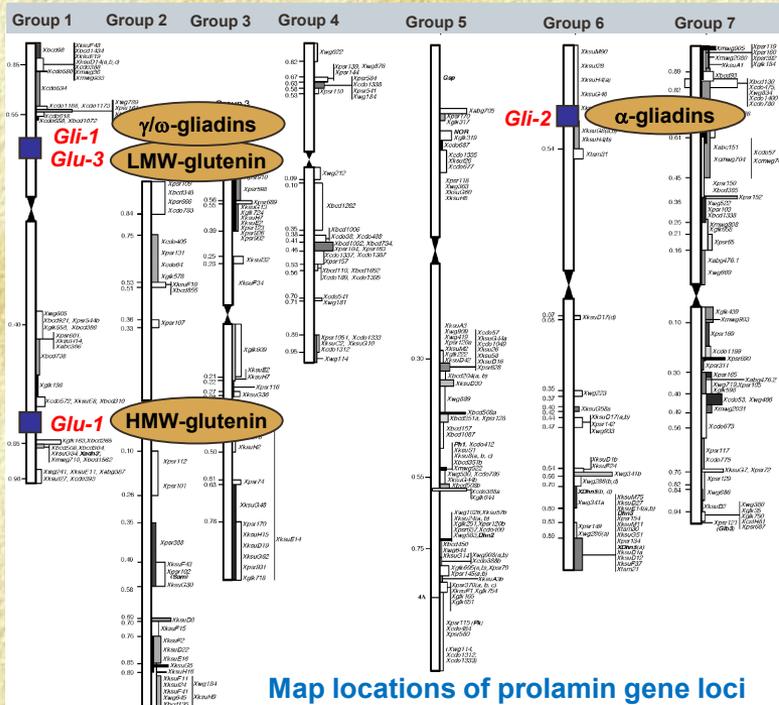


Wheat dough



Wheat bread

Challenges of studying wheat prolamins genes



- Bread wheat is hexaploid containing three highly related subgenomes (AABBDD)
- Prolamin genomic regions are difficult to sequence

Tandem gene duplication



Repeat sequences in the prolamins genes



- Characterizing the expression of individual genes is difficult
- Marker development for genotyping is difficult

Gene	Copy Number
HMW-glutenin	4 to 6
LMW-glutenin	10 to 30
α-gliadin	20 to 150
γ-gliadin & ω-gliadin	15 to 40

Complexity of prolamin sequences



Identification of a complete set of prolamin genes in hexaploid wheat

Illumina short reads have issues in assembling prolamin gene regions

- High copy gene family members
- Repetitive domains
- High rate of pseudogenes
- Polyploid genome with high repetitive DNA content



(GIGA)ⁿ
SCIENCE

GigaScience, 6, 2017, 1–7

doi: [10.1093/gigascience/gix097](https://doi.org/10.1093/gigascience/gix097)

Advance Access Publication Date: 23 October 2017

Data Note

DATA NOTE

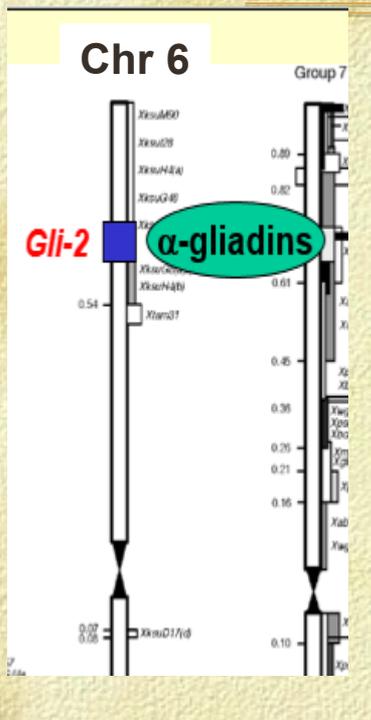
The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*

Aleksey V. Zimin^{1,2}, Daniela Puiu¹, Richard Hall³, Sarah Kingan³, Bernardo J. Clavijo⁴ and Steven L. Salzberg^{1,5,*}

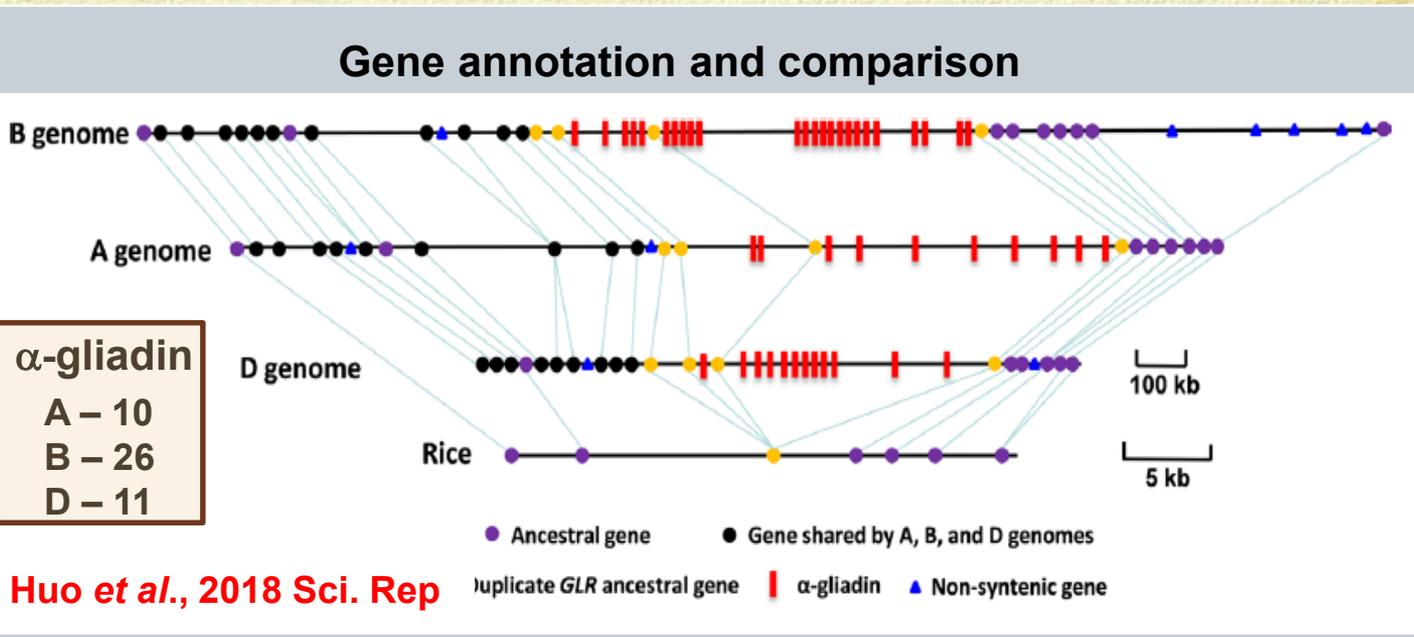
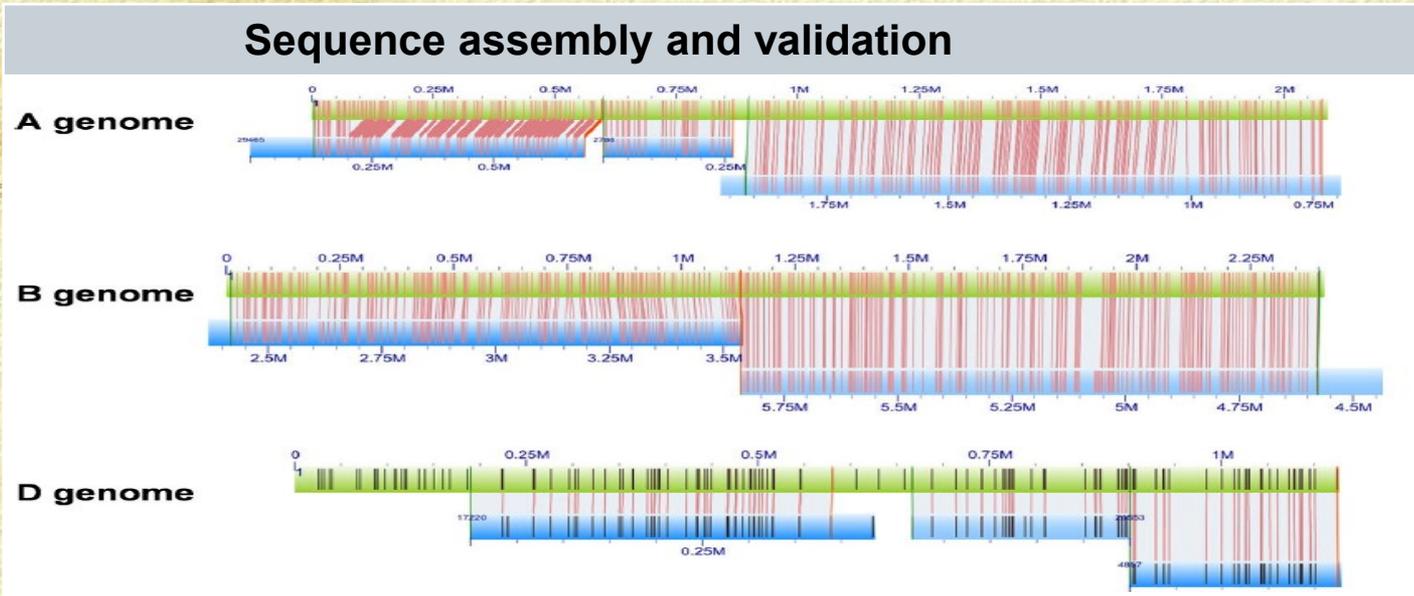
**Chinese Spring genome assembly with 40x PacBio long reads
N50 contig size over 230 kb.**

A BioNano genome map was generated for the hexaploid wheat cv Chinese Spring

Editing and validation of sequence assemblies in prolamin regions with BioNano maps



Copy number: 20 ~ 150



Expression of prolamin genes

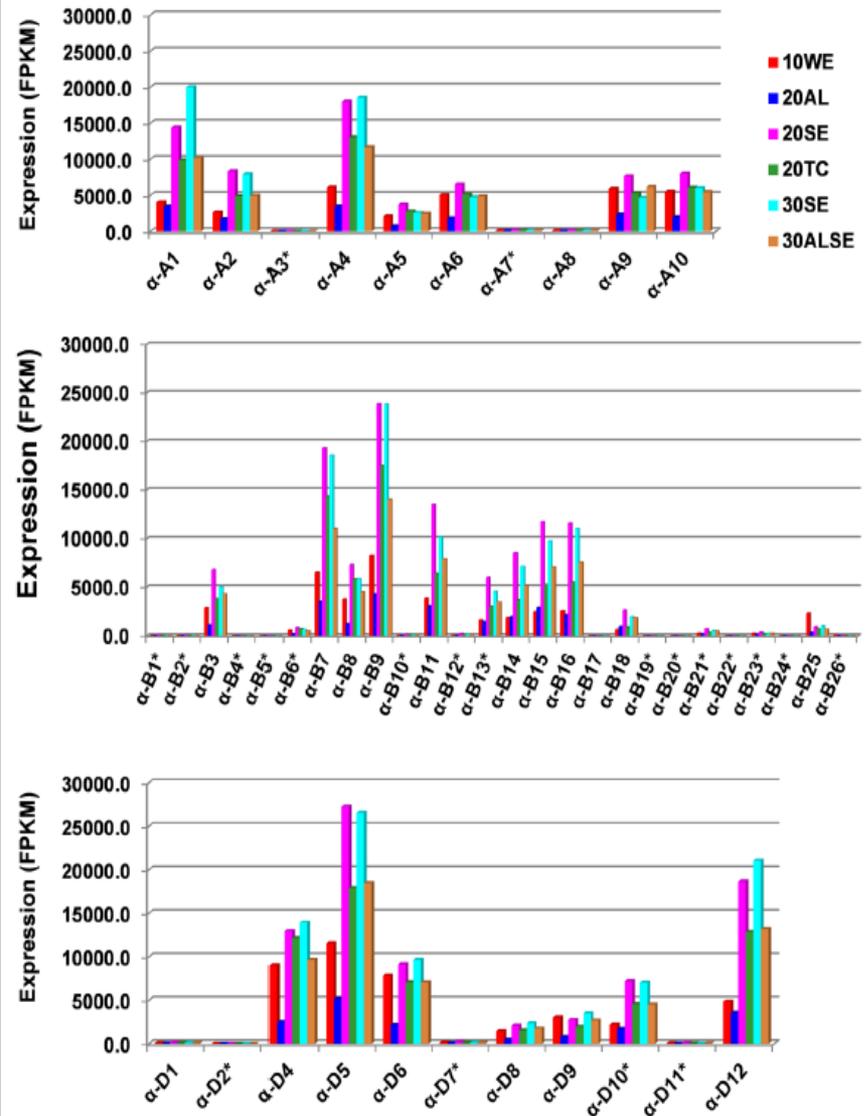
Approach

1. Map transcriptome reads to the annotated genes
2. Examine the alignments manually
3. Validate pseudogene sequences
4. Count reads to determine relative expression levels

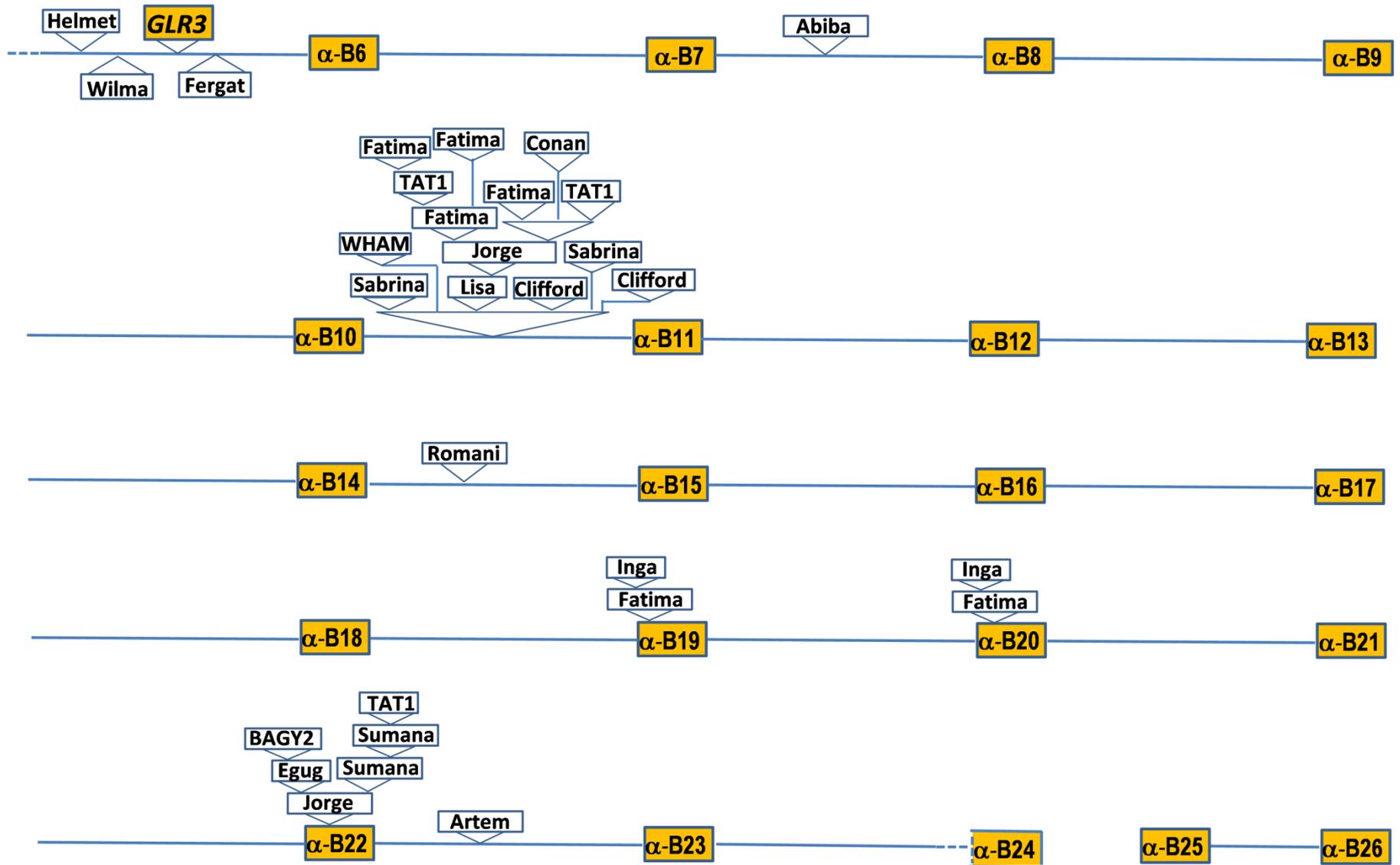
Prolamin genes are intronless. Mapping reads to the complete set of prolamin genes provides more robust and accurate view of the expression of individual genes

Huo *et al.*, 2018 *Sci. Rep*

Expression analysis by transcriptomics



Annotation of the α -gliadin genomic region of the B genome in Chinese Spring



Annotation of wheat prolamins genes

	HMW	Delta	Gamma	LMW	Omega	Alpha
Intact	4	2	11	10	5	26
Pseudogene	2	3	3	7	14	21
Total	6	5	14	17	19	47

A complete set of prolamins genes in hexaploid wheat cv Chinese Spring

- **Use the manually annotated prolamins gene sequences to align with the ReSeqv1 and RefSeqv2**
- **Identify and verify discrepancies of prolamins genes between our published data and RefSeq.**
- **Make corrections and submission of re-annotated prolamins genes in RefSeqV2**

Annotation of wheat prolamins genes

	Size (Mb)	Version 1.0	Version 2.0	Paper
Glu-3 & Gli-1A	5.33	295	142	9
Glu-3 & Gli-1B	6.53	368	74	4
Glu-3 & Gli-1D	5.64	360	196	16
Gli-2A	2.05	132	56	1
Gli-2B	2.41	259	83	3
Gli-2D	1.10	147	28	1

Gap numbers in the prolamins genomic regions in different CS sequence assemblies

Annotation of wheat prolamins

Version 1.0	Delta	Gamma	LMW	HMW	Omega	Alpha
Total	5	14	17	6	19	47
Match	4	12	14	0	0	33
Gap	1	2	3	5	15	3
Fragment	0	0	0	1	2	2
Unanchored	0	0	0	0	2	8

29 prolamin genes have gaps in version 1

Version 2.0	Delta	Gamma	LMW	HMW	Omega	Alpha
Total	5	14	17	6	19	47
Match	5	13	17	5	4	36
Gap	0	1	0	1	15	1
Fragment	0	0	0	0	0	3
Unanchored	0	0	0	0	(+2Gap)	6(+1Frag)

18 prolamin genes have gaps in version 2

The wheat prolamin superfamily

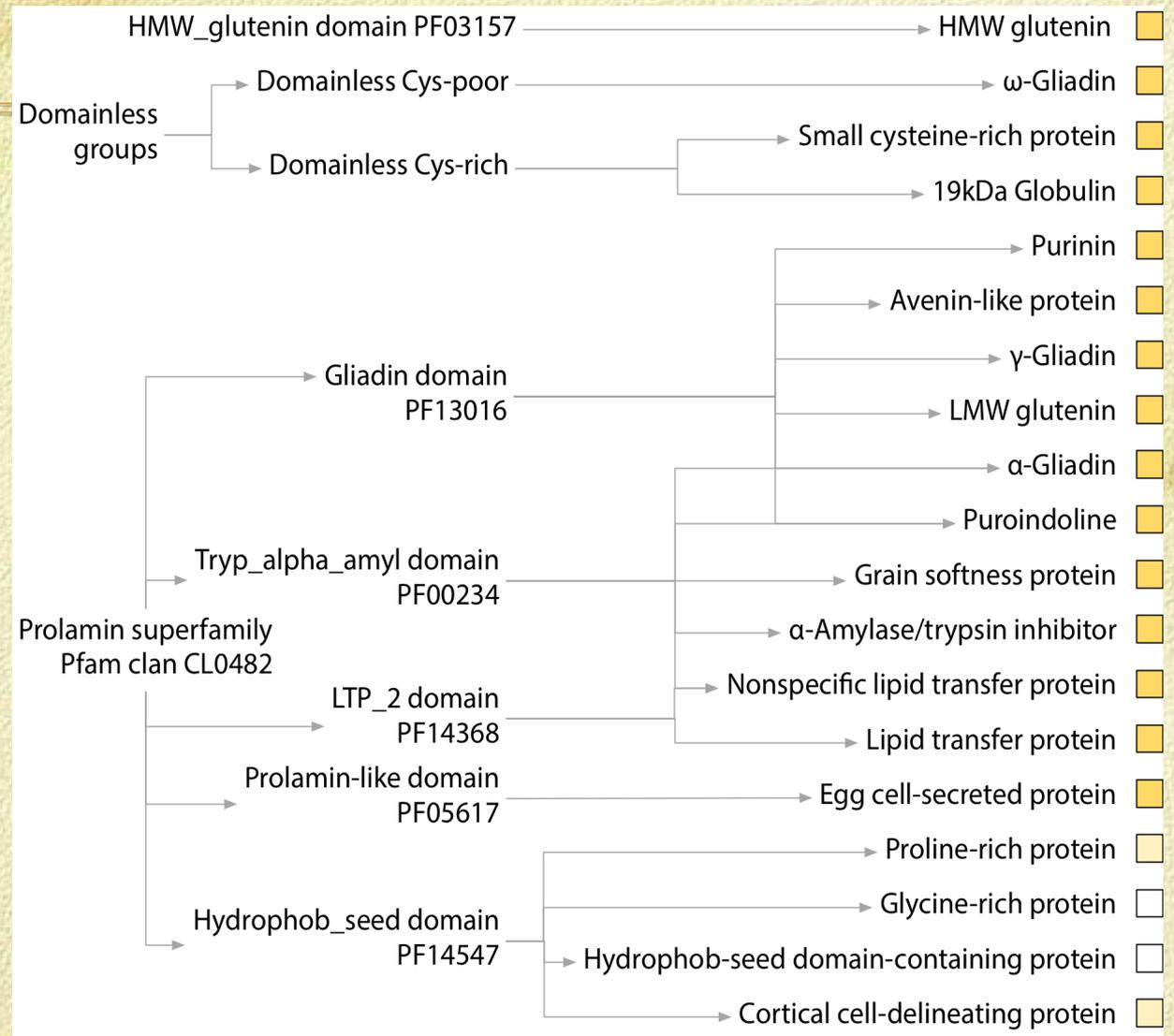
Pfam domains Prolamin clan CL0482

Gliadin
Tryp-alpha-amyl
LTP2
Hydrophob_seed
Prolamin-like

HMW glutenins

Domain-less protein groups

omega gliadins
small cysteine-rich
proteins

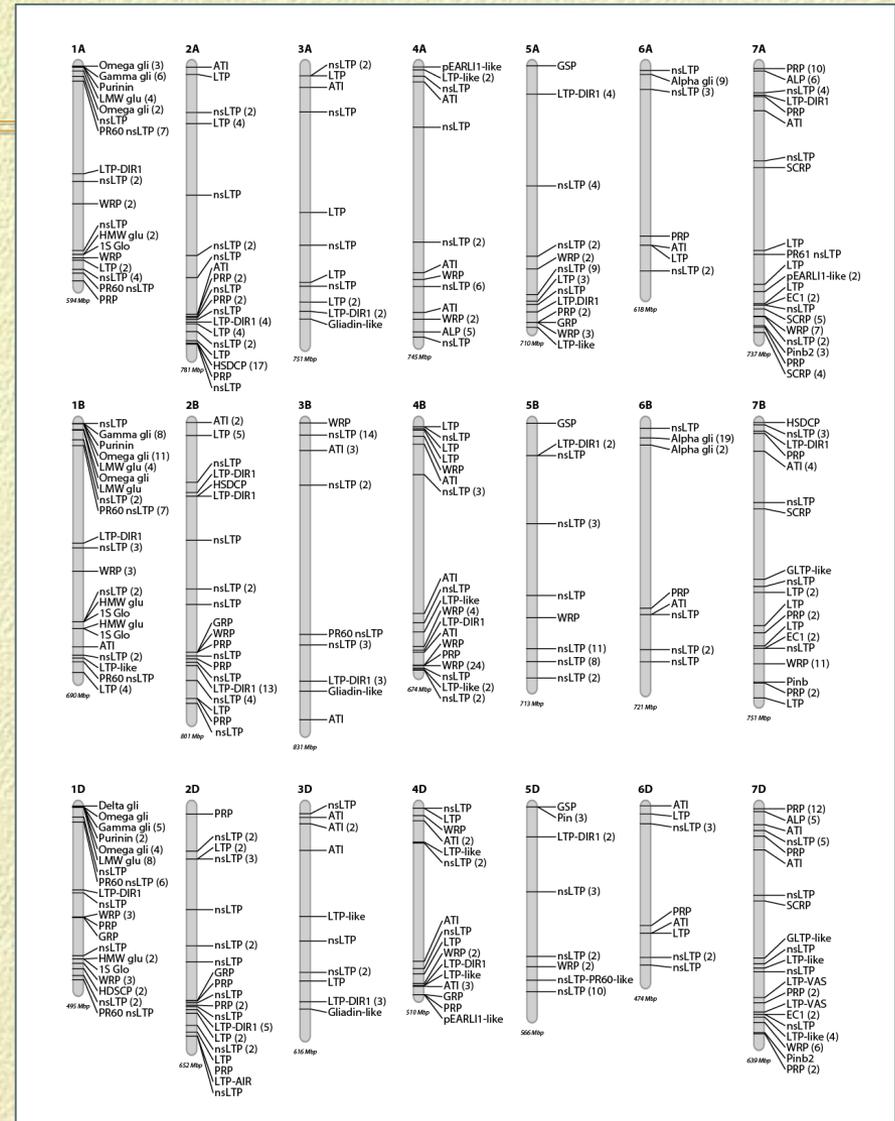


First comprehensive map of wheat prolamins



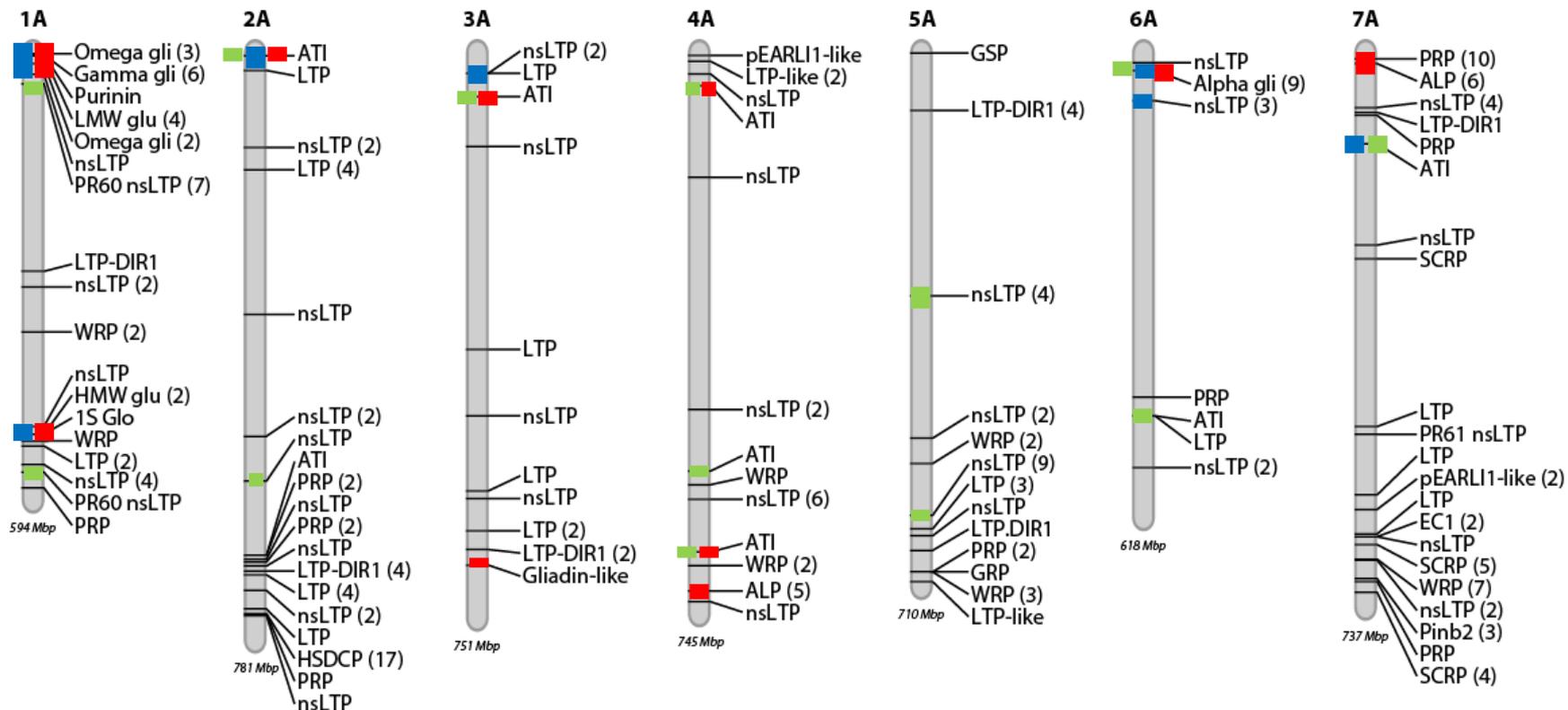
IWGSC, Science 2018

- Within the IWGSC RefSeq v1 annotation, 731 proteins were manually corrected, including 135 proteins that were added as a completely new sequence
- Expressed everywhere e.g. in roots, leaves, spike, pollen or grain



Immune responsive proteins are highly enriched in the glutenin and gliadin loci

A genome loci



- Celiac disease
- Wheat allergy
- Baker's asthma

Acknowledgments

USDA-ARS, WRRRC

Naxin Huo
Susan Altenbach
Shengli Zhang

Edith Cowan University

Angela Juhasz

University of Melbourne

Rudi Appels

University of California, Davis

Ming-Cheng Luo
Tingting Zhu

Chinese Academy of Sciences

Daowen Wang,
Liling Dong