Draft Genome of *Triticum urartu* and its Physical Mapping

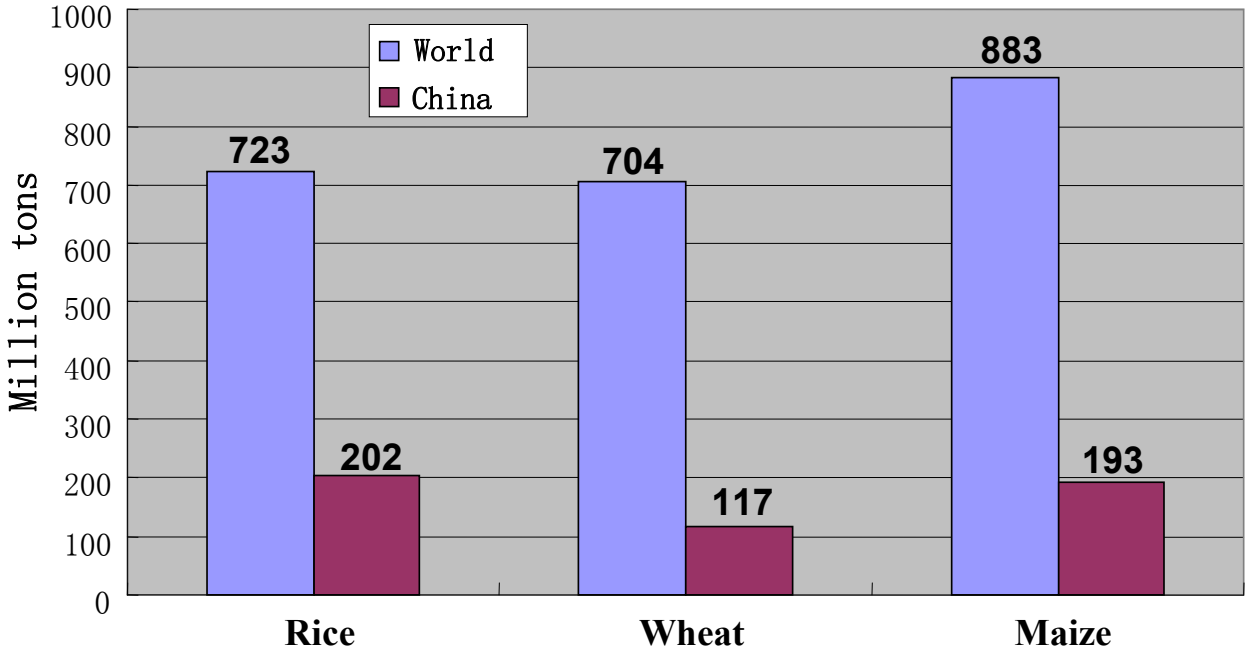
Hong-Qing Ling

hqling@genetics.ac.cn

State Key Laboratory of Plant Cell and Chromosome Engineering
Institute of Genetics and Developmental Biology
Chinese Academy of Sciences

Production of the Main Food Crops in the World and China

(Data from FAOSTAT 2011)



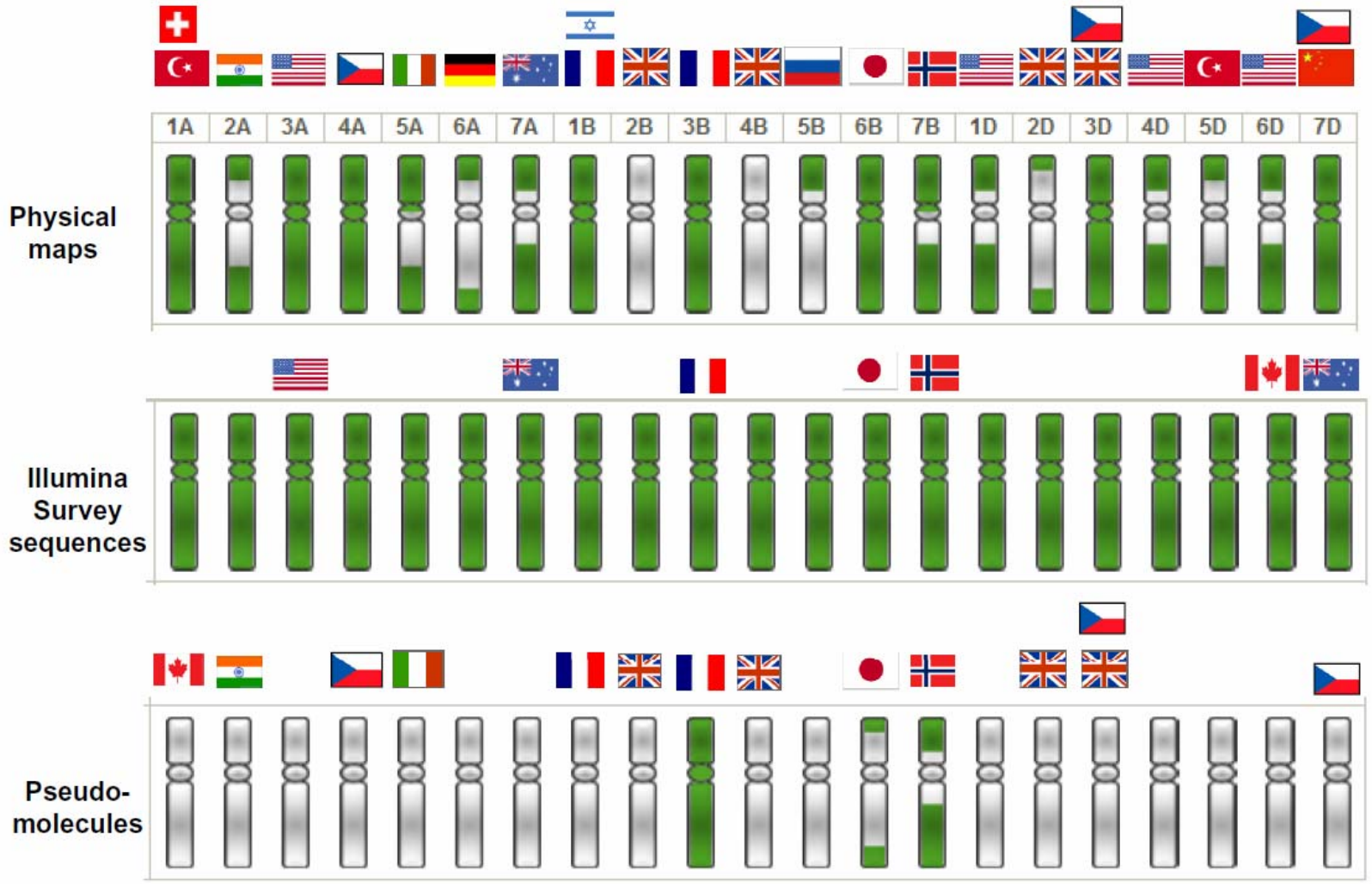
Wheat is one of the most important food crops in the world, feeding about 40% of the world population and providing 20% of total calories and protein in human nutrition.

Characters of Wheat Genome

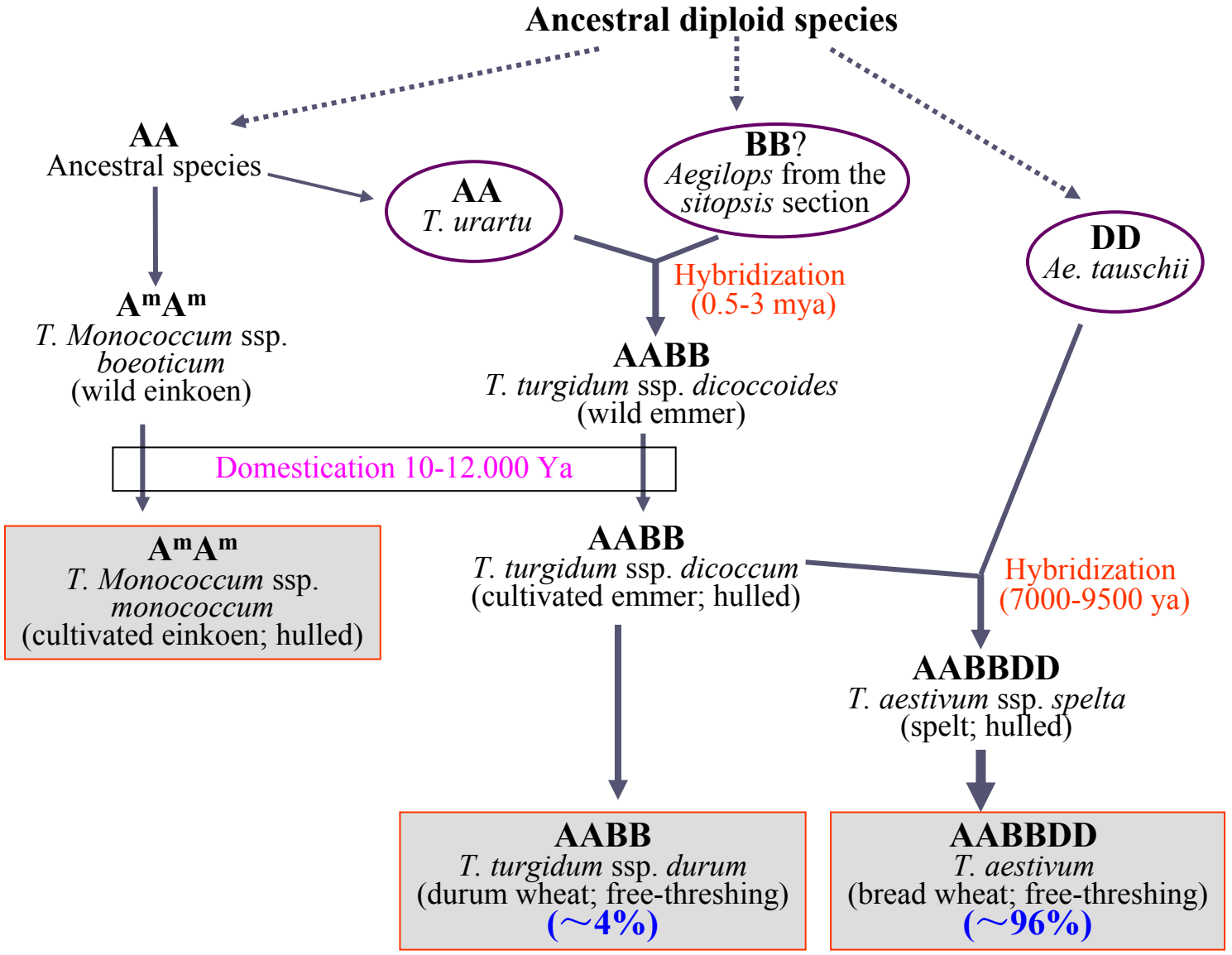
- Bread wheat is an allopolyploid, containing A, B and D subgenomes, which derived from a common ancestor. They show a high similarity each other.
- Bread wheat has a huge genome size (~17 Gb). It is about 8 times larger than maize, 40 times than rice and 100 times than *Arabidopsis thaliana*.
- More than 85% of the genome sequence are repetitive DNA

These characters of wheat make genome study very difficulty. Although some progress has been made, sequencing of wheat genome is still a big challenge.

Current Status of IWGSC

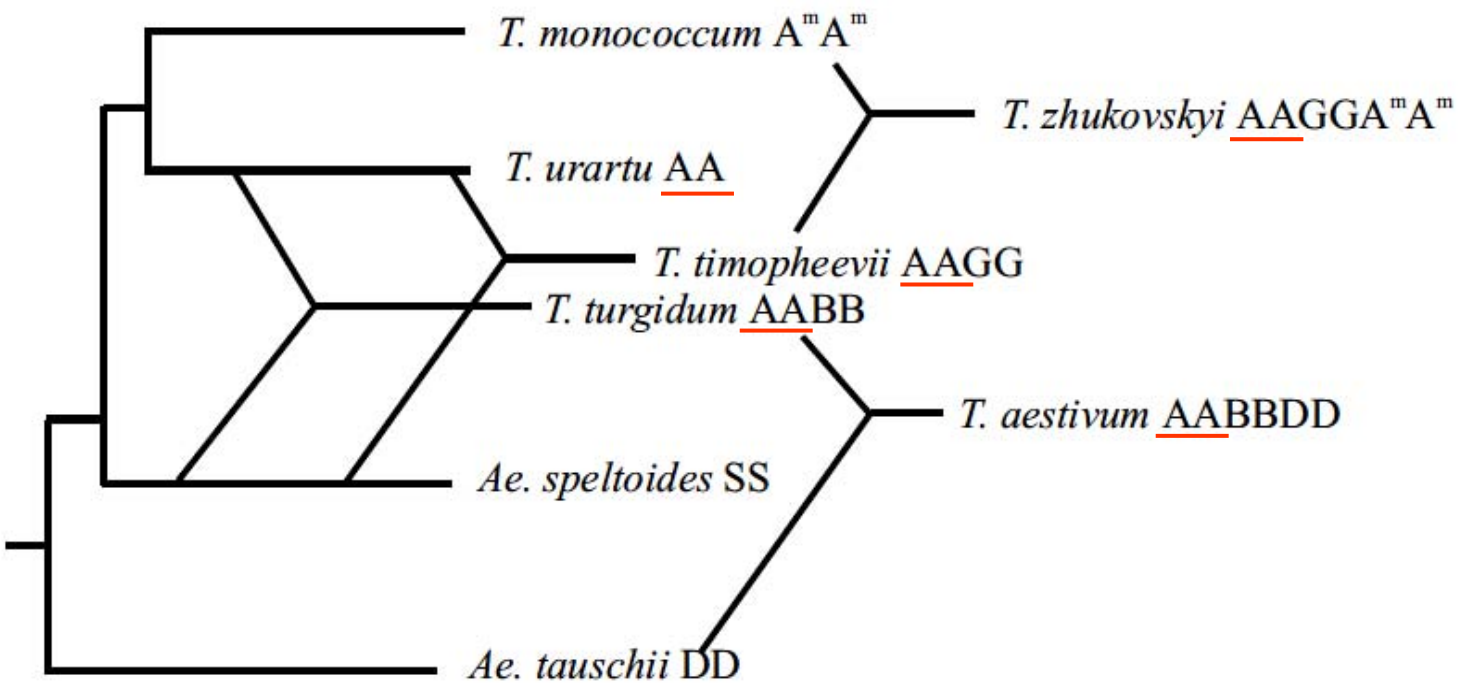


T. urartu Is the Progenitor of Wheat A Genome



Chantret et al. (2005), Plant Cell 17: 1033-1045

The A Genome is a Basic Genome of Wheat



The A genome is a basic genome of bread wheat and other polyploid wheats, plays a central role in wheat evolution, domestication and genetic improvement.

- **Sequencing and assembling of *T. urartu* genome will provide a diploid reference for analysis of polyploid wheat genomes**
- **The genome sequence is also valuable for studying wheat evolution, domestication and even for genetic improvement of wheat**

中国科学院遗传与发育生物学研究所
Institute of Genetics and Developmental Biology, CAS

Genome Sequencing and Assembly

Sequencing Wheat A Genome

The whole genome of *T. urartu* accession G1812 were sequenced, collaborated with BGI-Shenzhen, using shotgun sequencing approach with Illumina's next-generation sequencing platforms.

Sequence Data

Library insert size	Number of libraries	Number of lanes	Average reads length (bp)	Raw data (Gb)	Usable data (Gb)	Effective depth*	Physical data (Gb)	Physical depth*
~200 bp	15	22	111	213.96	164.88	33.38	321.26	65.03
~350 bp	7	16	61	49.00	36.07	7.30	210.98	42.71
~500 bp	10	19	82	121.49	91.87	18.60	576.65	116.73
~700 bp	9	13	79	95.43	64.98	13.15	573.49	116.09
2 kb	6	18	51	56.08	42.67	8.64	1653.53	334.72
5 kb	4	14	53	45.82	32.50	6.58	3046.77	616.76
10 kb	4	5	44	15.51	11.98	2.43	2723.43	551.30
20 kb	2	2	44	12.31	3.53	0.71	1603.93	324.68
Total	57	109	77	609.61	448.49	90.79	10710.05	2168.03

*Calculated with the estimated genome size of 4.94 Gb.

Genome Assembly*

	Contig		Scaffold	
	Size (bp)	Number	Size (bp)	Number
N90	123	5,521,147	126	3,978,271
N80	127	4,715,825	232	410,092
N70	777	715,986	19,387	43,008
N60	1,903	400,253	42,913	27,515
N50	3,422	246,789	63,687	18,663
Longest	82,078		1,066,088	
Total size	3,922,395,337		4,660,785,691	
Total number (≥ 1 kb)		622,928		133,906
Total number (≥ 2 kb)		385,430		81,698
GC ratio (%)	45.388		45.388	

The average length of the contigs containing intact or partial genes was 9.91 kb.

*SOAPdenovo (v.1.05), (Li et al., 2010, Genome Res. 20: 265-272)

中国科学院遗传与发育生物学研究所
Institute of Genetics and Developmental Biology, CAS

Genome Annotation and Comparative Analysis

Analysis of Repetitive Elements

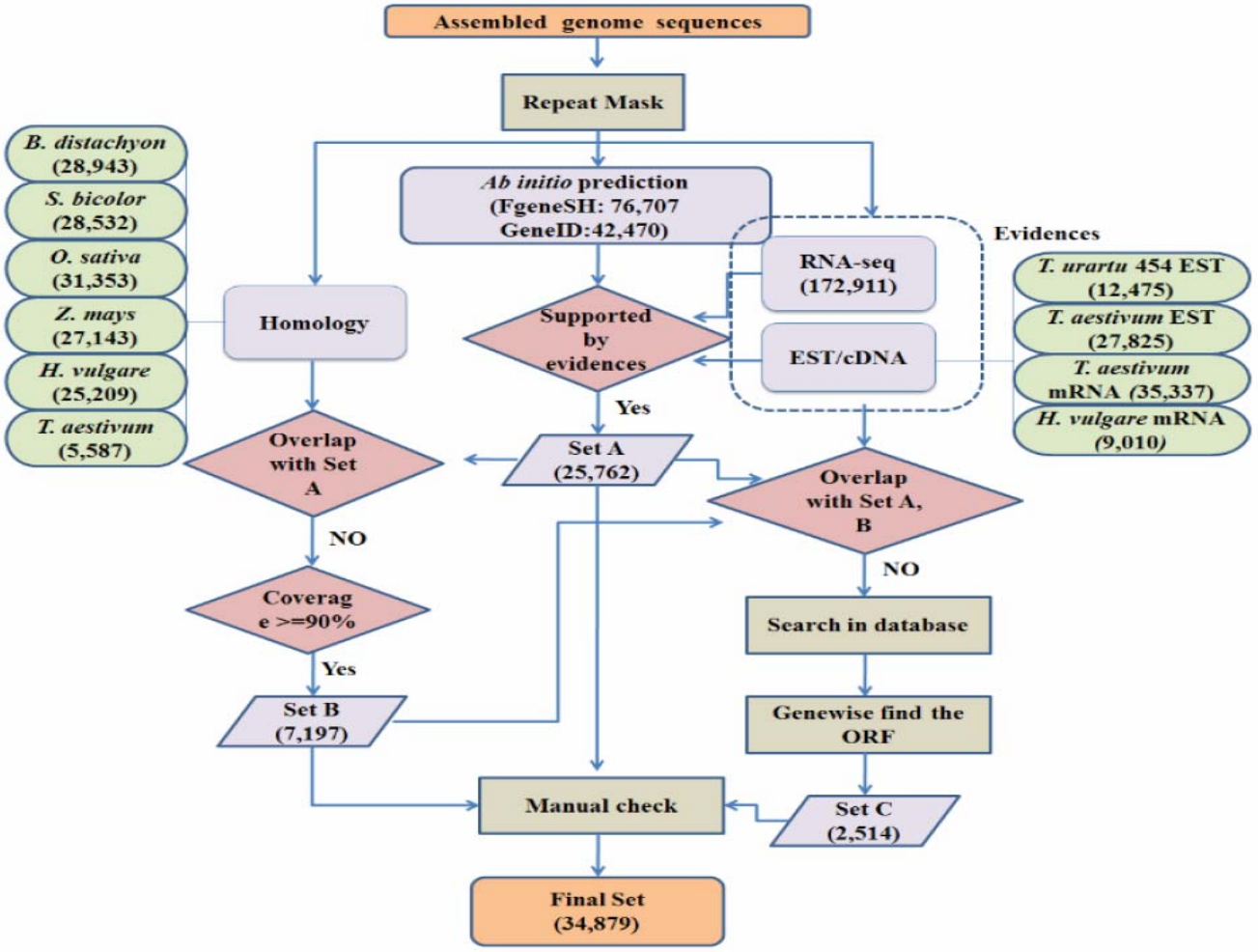
	Percentage of genome (%)					Length (bp)
	Bd	Sb	Os	Zm	Tu	Tu
Class I: Retrotransposon	**21.58**	**50.77**	**21.00**	**76.35**	**49.07**	**1,765,277,214**
LTR-Retrotransposon	18.38	49.70	19.85	75.52	46.66	1,678,595,438
LTR/*Gypsy*	13.77	42.85	16.39	48.43	36.57	1,315,436,369
LTR/*Copia*	4.46	6.81	3.08	26.55	9.89	355,762,130
Other	0.15	0.04	0.38	0.54	0.21	7,396,939
Non-LTR Retrotransposon	3.20	1.07	1.16	0.84	2.41	86,681,776
SINE	0.26	0.08	0.05	0.03	0.07	2,566,147
LINE	2.94	0.98	1.11	0.80	2.34	84,115,629
Class II: DNA Transposon	**5.33**	**7.17**	**5.82**	**5.39**	**9.77**	**351,279,176**
DNA Transposon Superfamily	3.32	4.73	2.75	3.37	7.33	263,725,407
DNA-*CACTA*	1.44	3.67	2.38	2.06	5.44	195,685,200
hAT	0.43	0.26	0.27	0.75	0.37	13,201,783
Harbinger	0.26	0.20	0.08	0.22	0.20	7,238,463
Tc1/Mariner	1.19	0.61	0.03	0.07	0.56	19,993,205
MITE	1.95	2.31	3.07	0.77	1.88	67,769,240
Tourist	0.28	1.47	1.11	0.12	0.32	11,503,170
Stowaway	0.14	0.09	0.60	0.00	0.05	1,839,920
Other	1.53	0.74	1.37	0.66	1.51	54,426,150
Helitron	0.06	0.13	0.00	0.54	0.01	270,967
Tandem repeat	1.89	2.49	2.90	0.86	1.21	43,630,554
Low complexity	0.27	0.19	0.82	0.12	0.10	3,489,764
Unclassified	**8.41**	**5.21**	**0.23**	**0.74**	**8.04**	**289,349,611**
Total content	**37.48**	**65.83**	**30.78**	**82.48**	**66.88**	**2,405,906,001**

Bd, *B. distachyon*; Sb, *S. bicolor*; Os, *O. sativa*; Zm, *Z. mays*; Tu, *T. urartu*.

RNA-Seq and Transcriptome Assembly

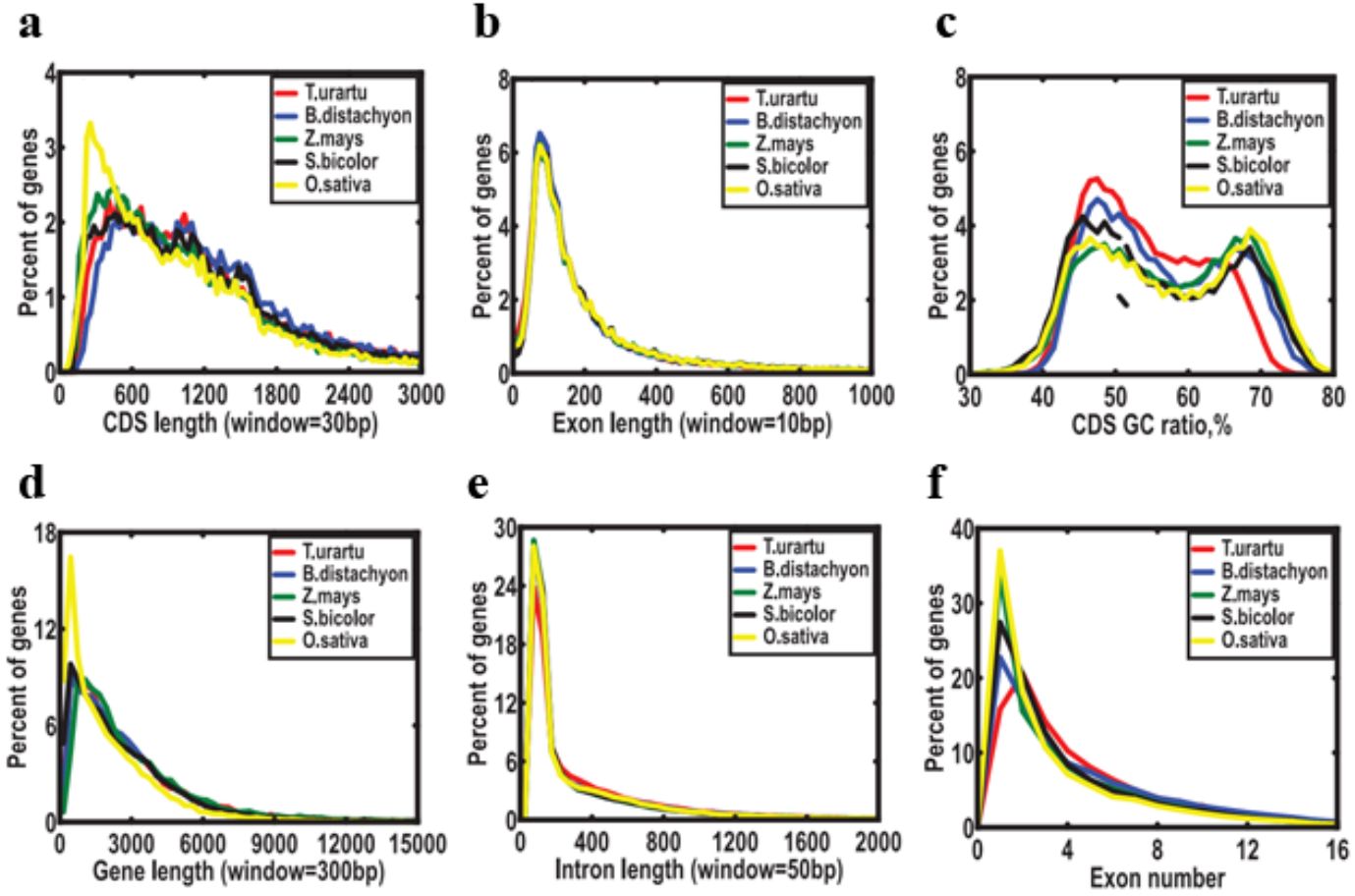
Organ	Usable data (Gb)	Transcripts	Average length (bp)	Maximum length (bp)	Total size of transcripts (bp)
2mR	9.38	54,601	1,096	11,937	59,830,667
YS	18.98	81,950	1,422	15,381	116,559,256
2mL	13.57	55,750	1,311	15,365	73,098,093
5dS	8.58	47,397	1,275	1,437	60,429,264
10dR	4.44	43,350	1,101	15,213	47,732,145
10dL	3.47	36,133	1,045	12,185	37,770,880
7wL	4.02	42,950	1,205	15,205	51,757,961
CL	4.70	50,838	1,208	23,741	61,427,836
Integration	67.14	92,868	1,256	15,378	116,650,180
454-est	**0.34**	**49,935**	**406**	**6,088**	**20,290,712**

Flowchart of Gene Prediction

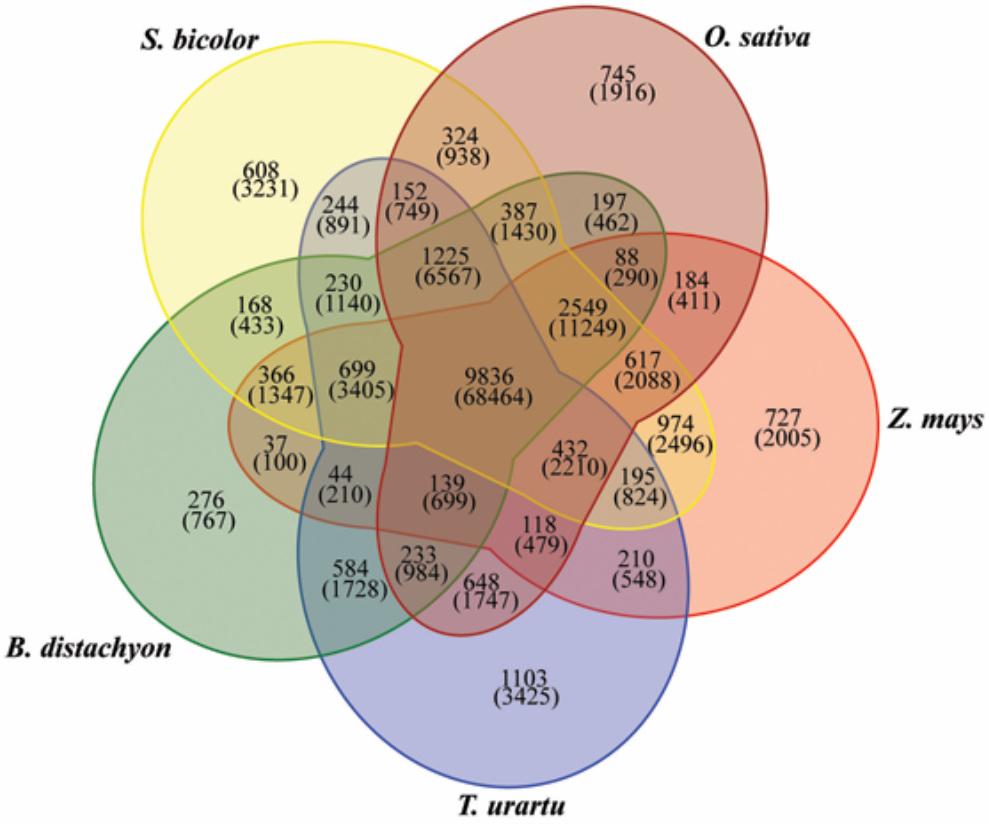


In total, 34,879 protein-coding gene models have been predicted

Comparison of Gene Features

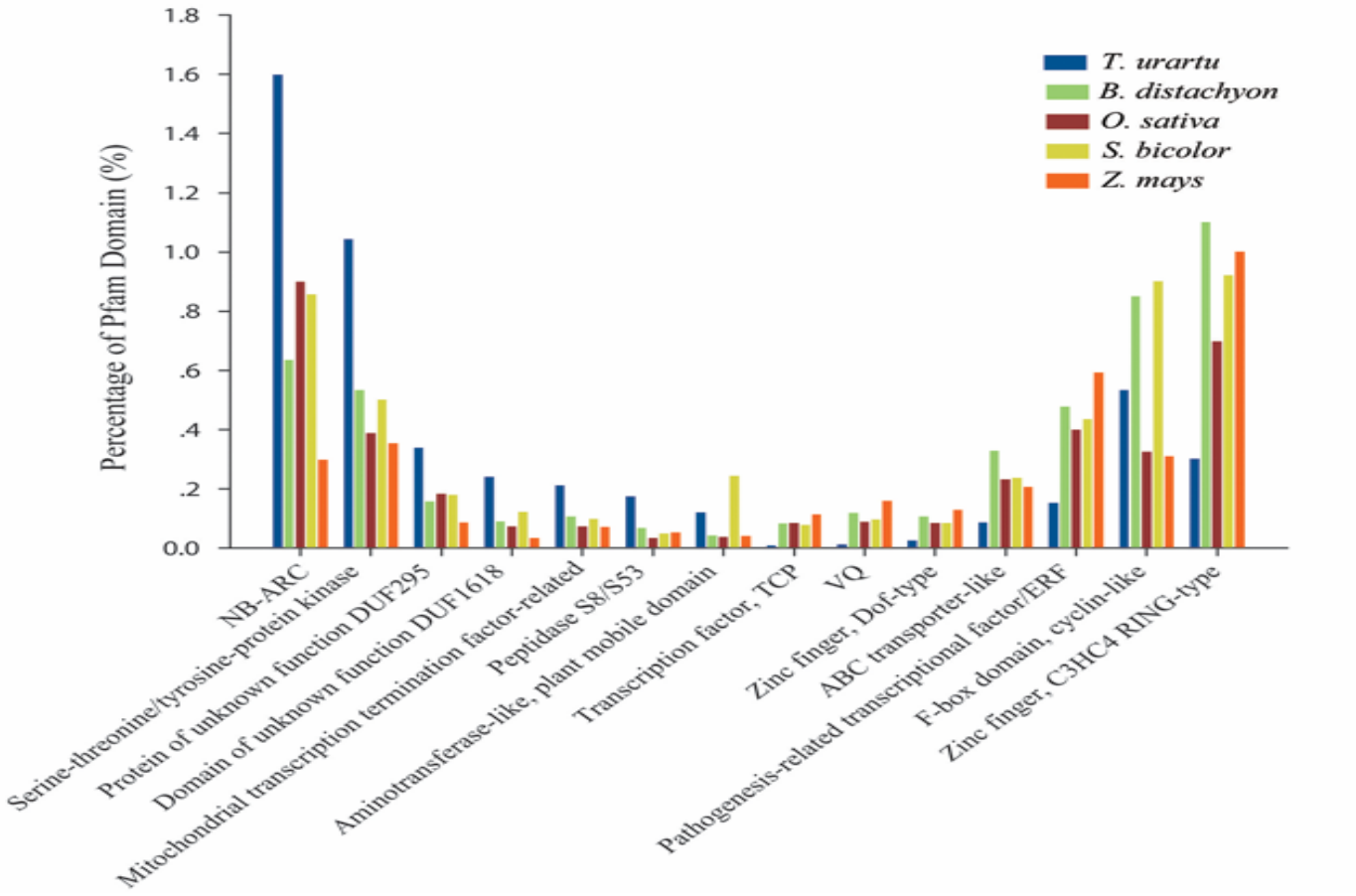


Gene Family Comparison



68464 genes were clustered together in 9836 gene families, which were shared by all five monocots, with 1103 gene families (3425 genes) being specific to the *T. urartu* genome.

Pfam Domain Comparison

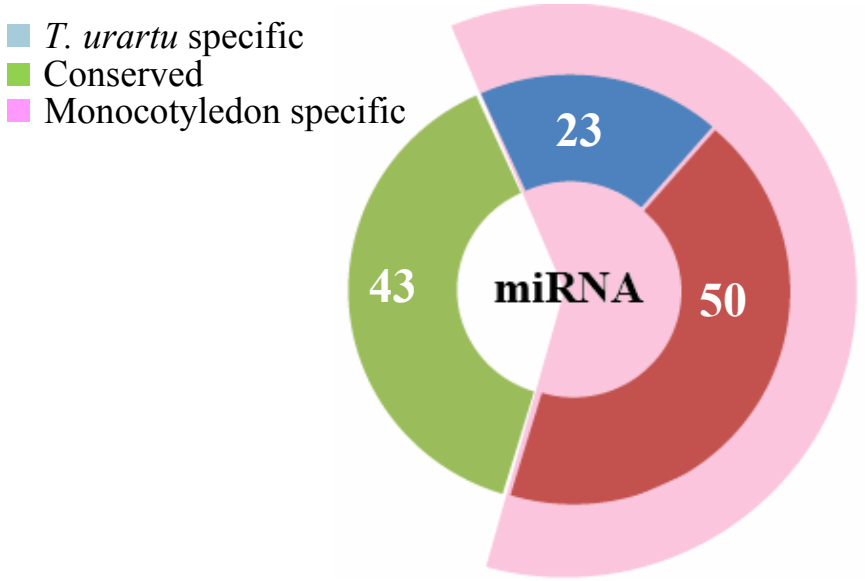


2,067 Pfam domains were shared among the five sequenced monocot species. Of them, 14 Pfam domains had differences in member numbers in *T. urartu* compared to the other four grasses.

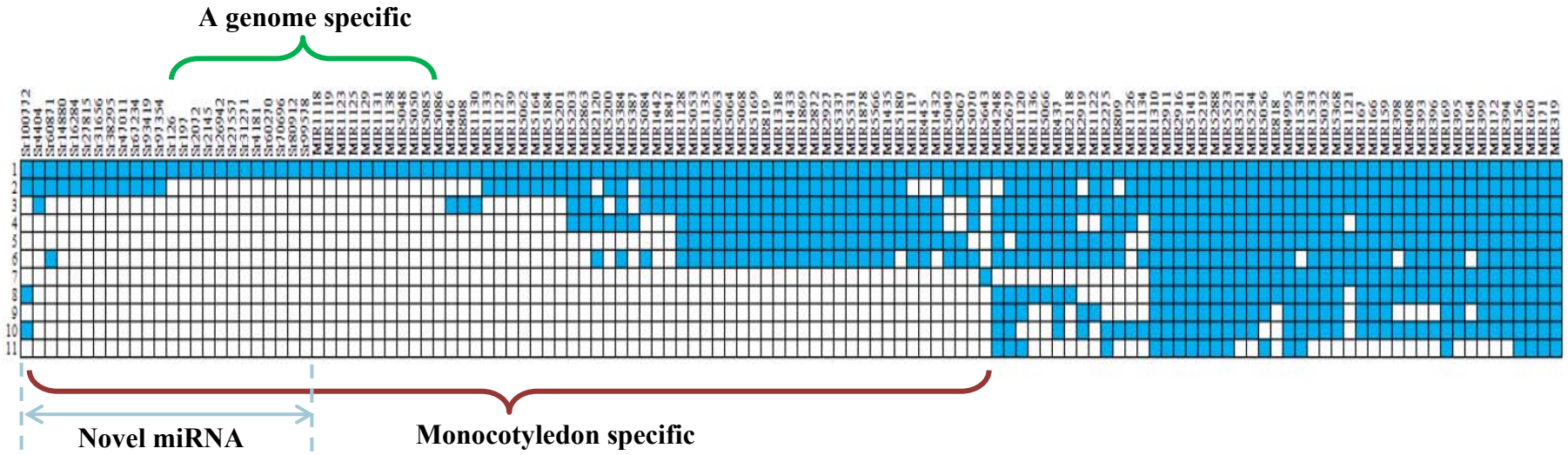
R Proteins in *T. urartu* and Other Grass Genomes

Predicted Protein Domain	*T. urartu*	Brachypodium	Rice	Maize	Sorghum
CC-NBS-LRR	174	121	246	44	96
NBS-LRR	247	40	134	27	92
CC-NBS	69	24	40	19	13
NBS	103	12	40	16	20
TIR-NBS-LRR	0	0	0	0	0
Total	**593**	**197**	**460**	**106**	**211**

Identification of miRNAs



412 conserved and 24 new miRNAs distributed into 116 families were identified. Comparison to the miRNAs of five monocots and five dicots showed that 73 miRNA families were monocot specific, of which 23 were uniquely present in *T. urartu*.



Predicted Target Genes of New miRNAs

miRNA name	Target gene number	InterPro	Description of target gene
Sr126	15	IPR000863	Sulfotransferase domain
		IPR003690	Mitochodrial transcription termination factor-related
		IPR002866	Maturase, MatK
Sr197	9	IPR002885	Pentatricopeptide repeat
Sr2072	14	IPR000767	Disease resistance protein
		IPR002182	NB-ARC
		IPR012871	Protein of unknown function DUF1677
		IPR000058	Zinc finger, AN1-type
Sr2145	1	IPR000432	DNA mismatch repair protein MutS, C-terminal domain
Sr26942	4	IPR009057	Homeodomain-like
		IPR003100	Argonaute/Dicer protein, PAZ
Sr4181	1	IPR007749	Protein of unknown function DUF677

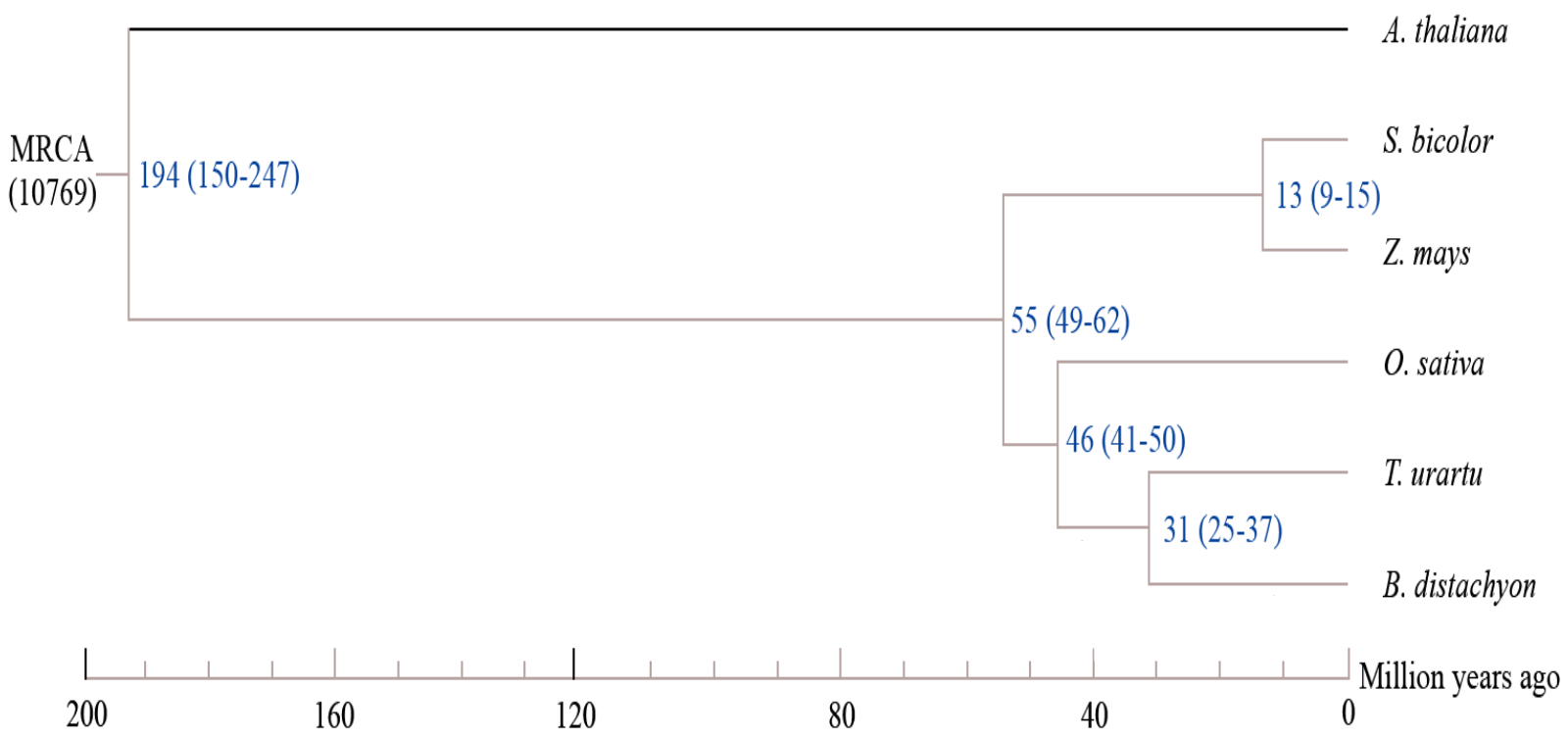
Scaffold Anchoring, Synteny Analysis

Assigning Scaffolds on Chromosome Bins

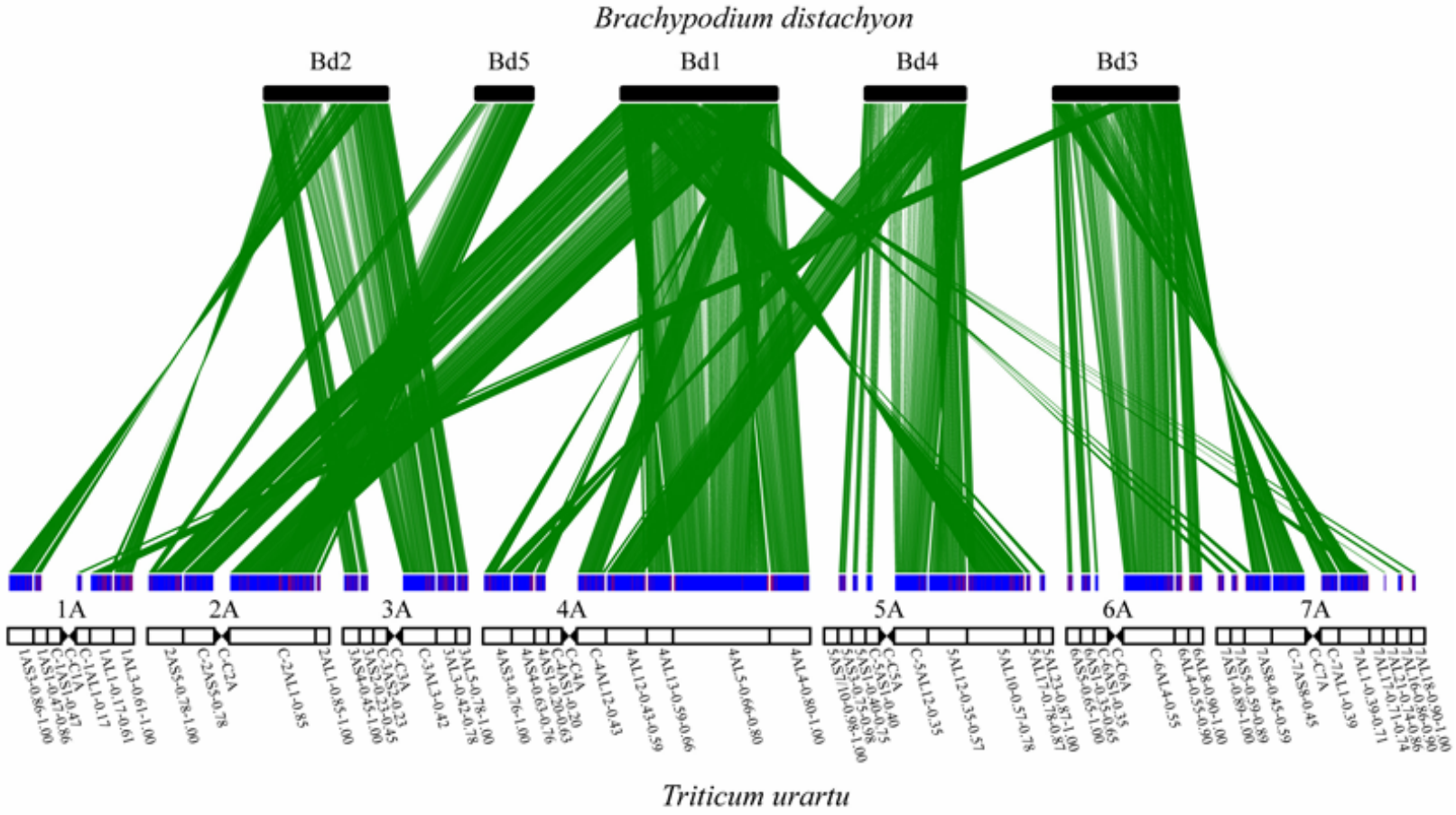
The scaffolds and gene models of *T. urartu* were assigned to chromosome bins using genetically mapped bread wheat ESTs20 as queries to search for homologous sequences in the *T. urartu* assembly. A total of 8,715 scaffolds, harboring 14,578 genes (41.8% of the total predicted genes) were mapped to 45 chromosomal bins of the wheat A genome.

Scaffolds on Chromosome 1

Phylogenetic Relationship of *T. urartu* and Four Sequenced Grasses



Synteny Analysis between *B. distachyon* and *T. urartu*



Wheat A genome

Genome Expansion

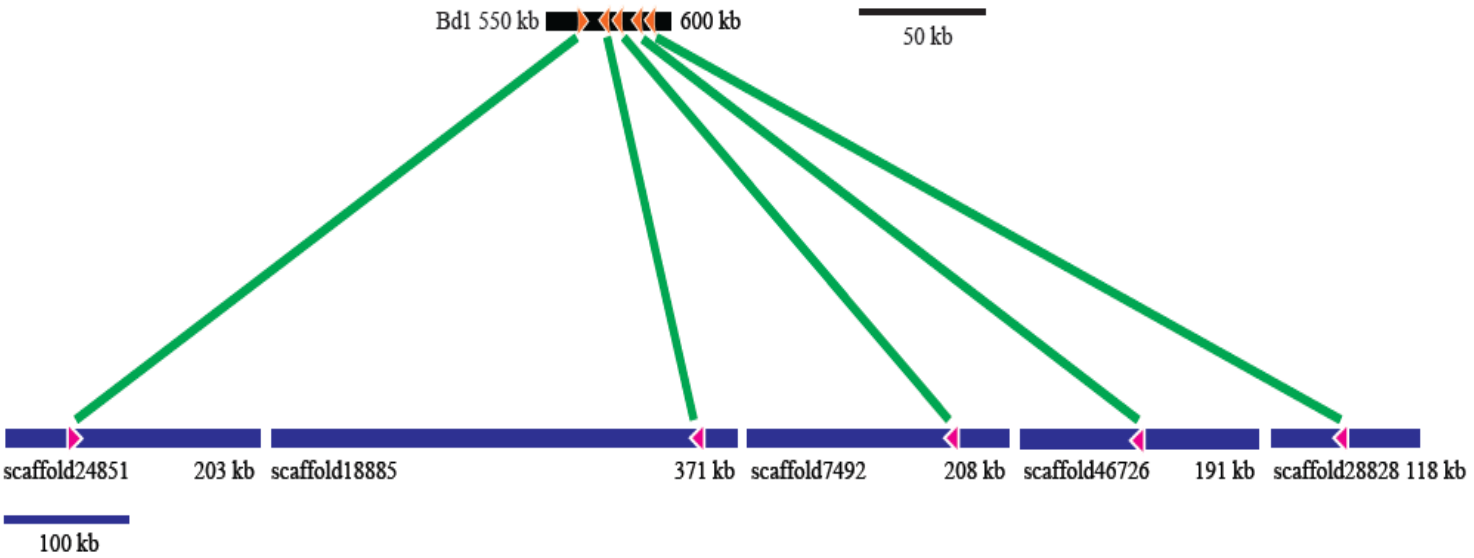
Species	Genome size	Fold
T. urartu	5,000 Mb	
B. distachyon	272 Mb	18.4
O. sativa	466 Mb	10.7
S. bicolor	730 Mb	6.8
Z. mays	2,300 Mb	2.2

Analysis of Synteny Blocks in Gene-rich Regions between *T. urartu* and *B. distachyon*

Region	*T. urartu* (bp)	*B. distachyon* (bp)	Ratio
Intergenic	24,128,306	10,827,354	2.23
Gene	20,118,863	18,821,875	1.07
CDS	7,848,156	8,380,692	0.94
Intron	12,270,707	10,441,183	1.18
Total	108,613,201	78,120,333	1.39

The 2,498 synteny blocks contained 7,559 genes, indicating that 21% of *T. urartu* genes had similarly sized intergenic spaces to those of in *B. distachyon*.

Intergenic Space Expansion in *T. urartu* Genome

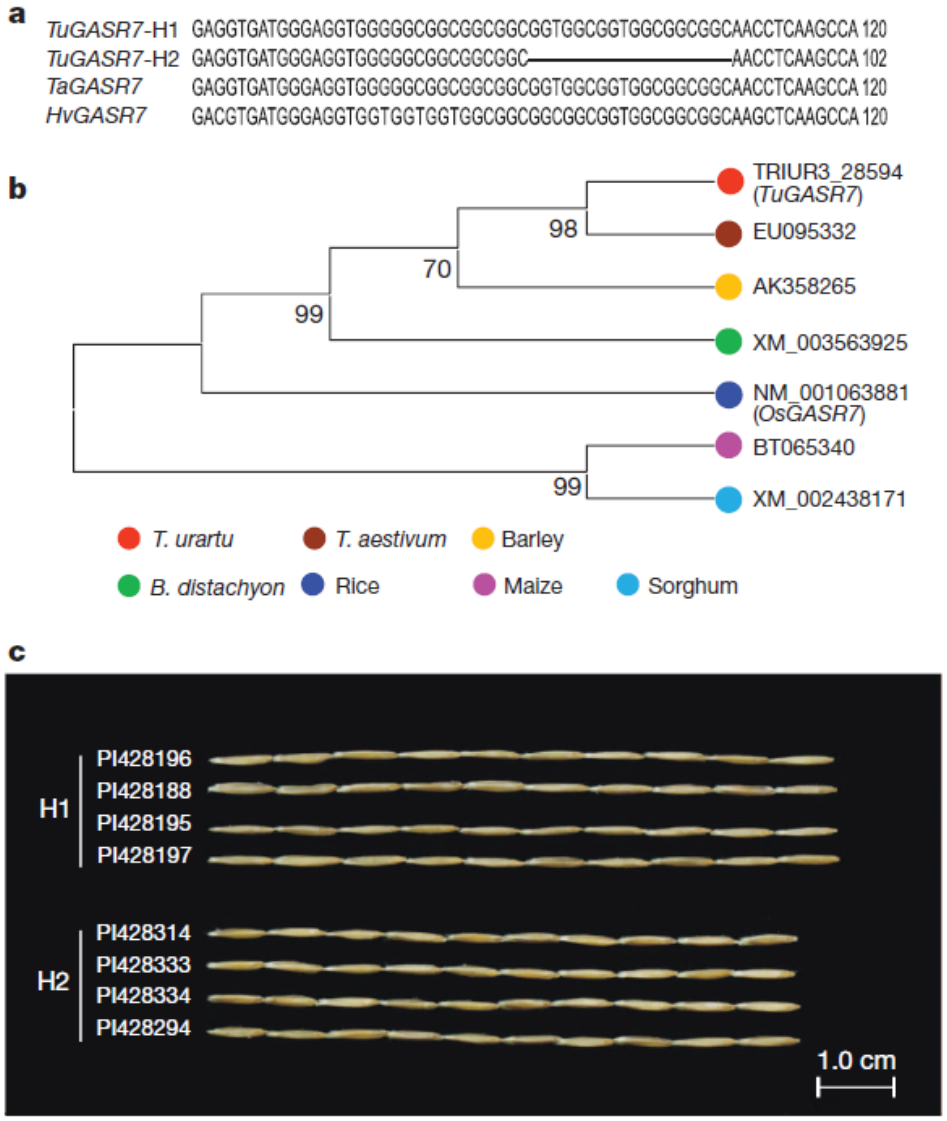


A synteny block containing 5 genes in a 50-kb region on chromosome 1 of *B. distachyon*. Their orthologous genes of *T. urartu* were separated in five scaffolds with a total length of >1091 kb. The intergenic spaces were expanded >21 fold in *T. urartu* by repetitive DNA (*Gypsy*, *Copia*), compared to *B. distachyon*.

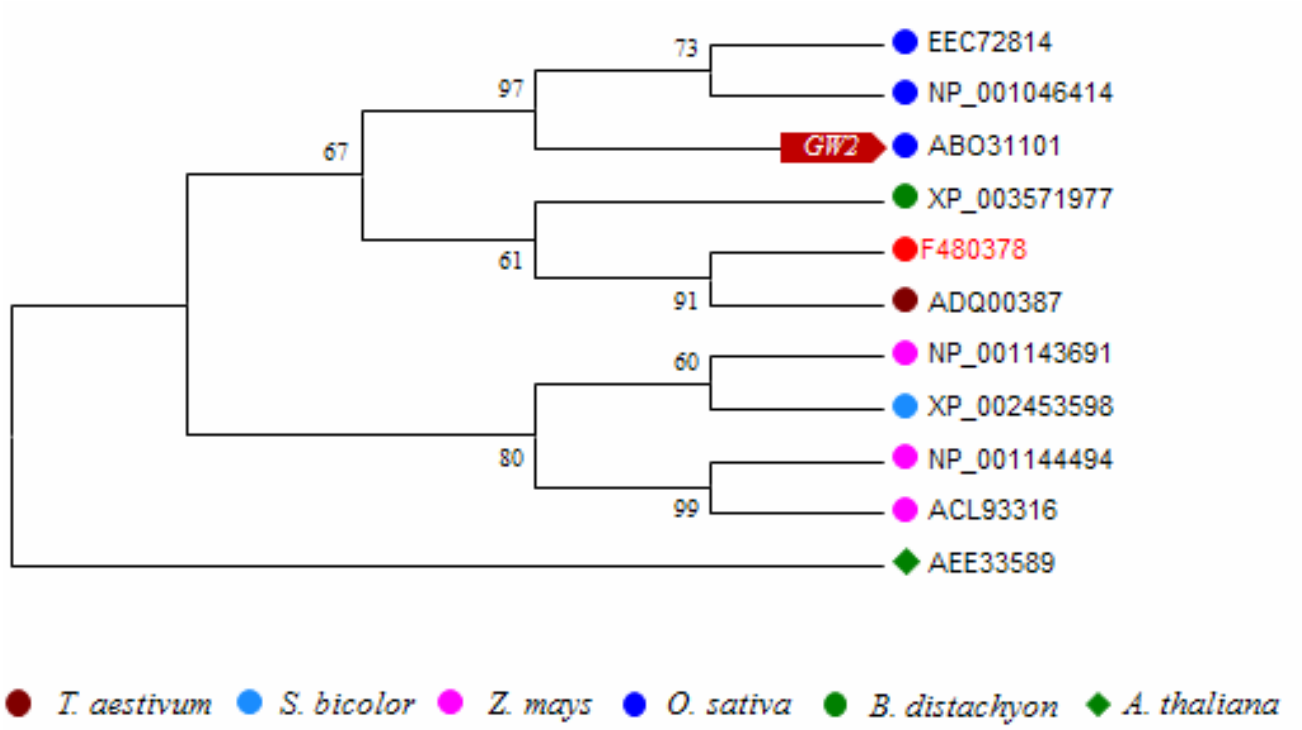
中国科学院遗传与发育生物学研究所
Institute of Genetics and Developmental Biology, CAS

Assessing the Utility of the Genome Sequence

TuGASR7, a Gene Controlling Seed Length



Homologous gene of rice *GW2*



Powdery Mildew Resistance Genes

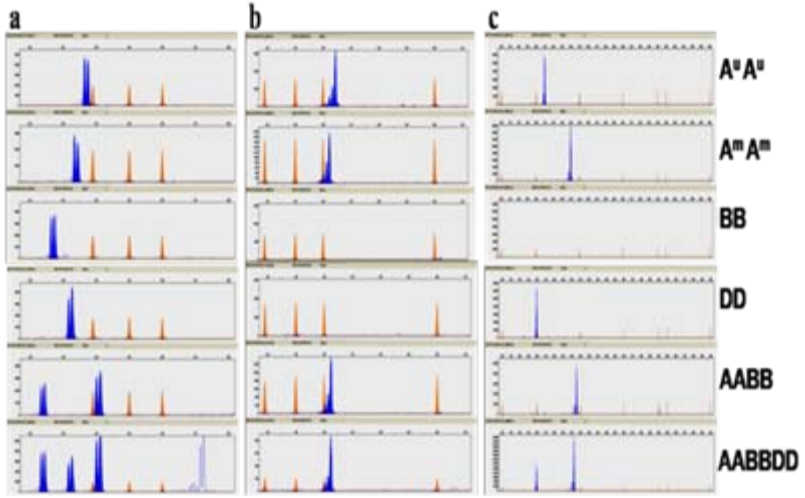
- 36 homologs of *Pm3b* in *T. urartu* genome have been identified. Fifteen of them revealed a high sequence identity (Evalue = 0).
- 21 of the 36 *Pm3b*-homologs have been assigned to chromosome. Of them, 9 homologs were assigned to the chromosome 1AS of wheat.



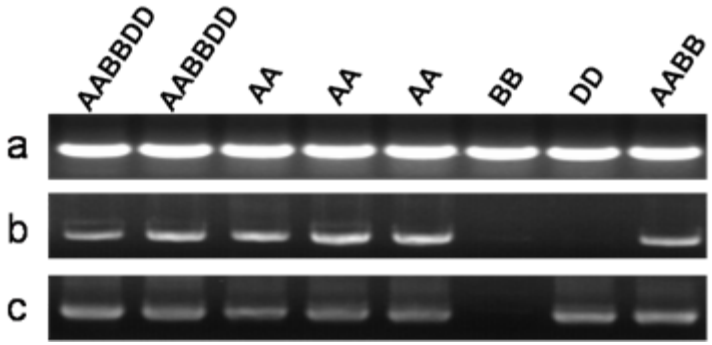
Molecular Marker Development

Marker type	Number	A-specific
SSR	166,309	33.6%
ISBP	739,534	10.2%
SNP (G1812/DV2138)	3,422,189	-

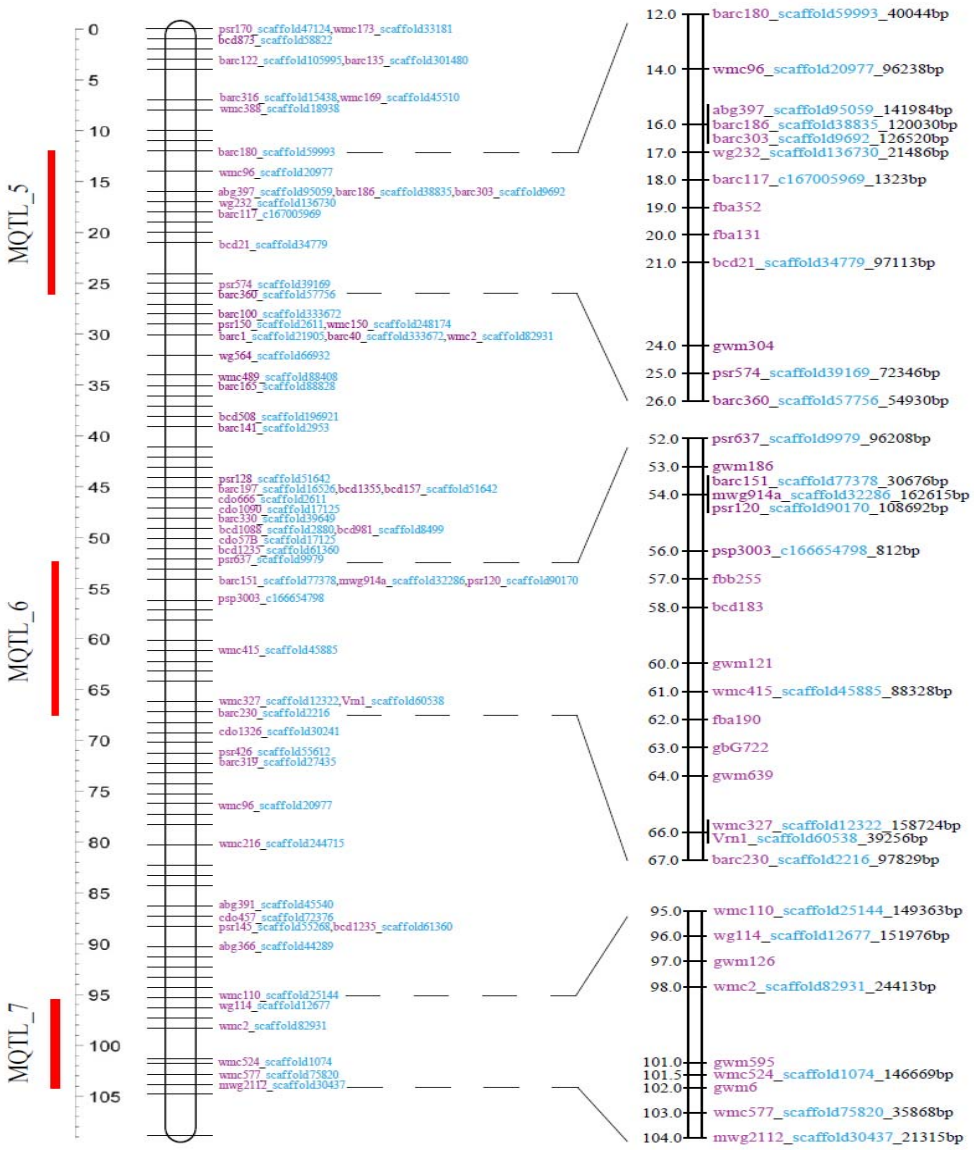
SSR marker



ISBP marker



Dissection of Meta-QTLs Controlling Grain Size



LETTER

OPEN

doi:10.1038/nature11997

Draft genome of the wheat A-genome progenitor *Triticum urartu*

Hong-Qing Ling[1]*, Shancen Zhao[2,3]*, Dongcheng Liu[1]*, Junyi Wang[1,2]*, Hua Sun[1]*, Chi Zhang[2]*, Huajie Fan[1], Dong Li[2], Lingli Dong[1], Yong Tao[2], Chuan Gao[2], Huilan Wu[1], Yiwen Li[1], Yan Cui[1], Xiaosen Guo[2], Shusong Zheng[1], Biao Wang[1], Kang Yu[1], Qinsi Liang[2], Wenlong Yang[1], Xueyuan Lou[1], Jie Chen[2], Mingji Feng[2], Jianbo Jian[2], Xiaofei Zhang[1], Guangbin Luo[1], Ying Jiang[2], Junjie Liu[2], Zhaobao Wang[2], Yuhui Sha[2], Bairu Zhang[1], Huajun Wu[4], Dingzhong Tang[1], Qianhua Shen[1], Pengya Xue[1], Shenhao Zou[1], Xiujie Wang[4], Xin Liu[1], Famin Wang[1], Yanping Yang[1], Xueli An[1], Zhenying Dong[1], Kunpu Zhang[1], Xiangqi Zhang[1], Ming-Cheng Luo[5], Jan Dvorak[5], Yiping Tong[1], Jian Wang[2], Huanming Yang[2], Zhensheng Li[1], Daowen Wang[1], Aimin Zhang[1] & Jun Wang[2,6,7]

Bread wheat (*Triticum aestivum*, AABBDD) is one of the most widely cultivated and consumed food crops in the world. However, the complex polyploid nature of its genome makes genetic and functional analyses extremely challenging. The A genome, as a basic genome of bread wheat and other polyploid wheats, for example, *T. turgidum* (AABB), *T. timopheevii* (AAGG) and *T. zhukovskyi* (AAGGA^{m}A^m), is central to wheat evolution, domestication and genetic improvement[1]. The progenitor species of the A genome is the diploid wild einkorn wheat *T. urartu*[2], which resembles cultivated wheat more extensively than do *Aegilops speltoides* (the ancestor of the B genome[3]) and *Ae. tauschii* (the donor of the D genome[4]), especially in the morphology and development of spike and seed. Here we present the generation, assembly and analysis of a whole-genome shotgun draft sequence of the *T. urartu* genome. We identified protein-coding gene models, performed genome structure analyses and assessed its utility for analysing agronomically important genes and for developing molecular markers. Our *T. urartu* genome assembly provides a diploid reference for analysis of polyploid wheat genomes and is a valuable resource for the genetic

200 base pairs (bp) to 65.8 kb, with an average length of 9.91 kb. The assembly was evaluated by comparisons with published bacterial artificial chromosome and expressed sequence tag (EST) sequences and by validation with PCR (Supplementary Information), and both indicated that the draft sequence had extensive genome coverage with high accuracy. The distribution of GC content in the *T. urartu* genome was comparable with those in the genomes of rice[12], maize[13], sorghum[14] and *Brachypodium distachyon*[15] (Supplementary Information).

Genome annotation of the assembly was performed as described in Supplementary Information. About 66.88% of the *T. urartu* assembly was identified as repetitive elements, including long terminal repeat retrotransposons (49.07%), DNA transposons (9.77%) and unclassified elements (8.04%) (Supplementary Information). The proportion of repetitive DNA was lower than the roughly 80% previously reported[16], which is probably due to a decreased incorporation of repeat sequence reads into the assemblies.

To facilitate gene prediction, we generated a 116.65-megabase (Mb) transcriptome of *T. urartu* with 67.14 Gb of RNA-Seq data from eight different tissues and treatments using the HiSequation (2000)

Nature 469: 87-90 (2013)

Physical Mapping and Superscaffold Construction

BAC Libraries and Physical Mapping

- **Library Construction**

Three genomic BAC libraries of *T. urartu* G1812 were constructed using *Hind*III, *Eco*RI and *Mbo*I. They contain 470,000 BAC clones with an average insert size of 120 kb.

- **Physical Mapping**

In collaboration with Keygene, 451,584 BAC clones were analyzed using the whole genome profiling approach. Of them, 345,233 BAC clones were available and used for constructing the physical map. The total size was 8x of *T. urartu* genome.

Physical Mapping Results

	LS-WGP
Total No. of BACs in FPC	345,233
Contig No.	12,137
No. and % BACs in contigs	323,058 (94%)
No. and % Singleton BACs	22,175 (6%)
Coverage (Mbp)	**4,688**
Average contig size (BACs)	26.6
Average contig size (kb)	386
N50 contig size (BACs)	55
N50 contig size (kb)	656

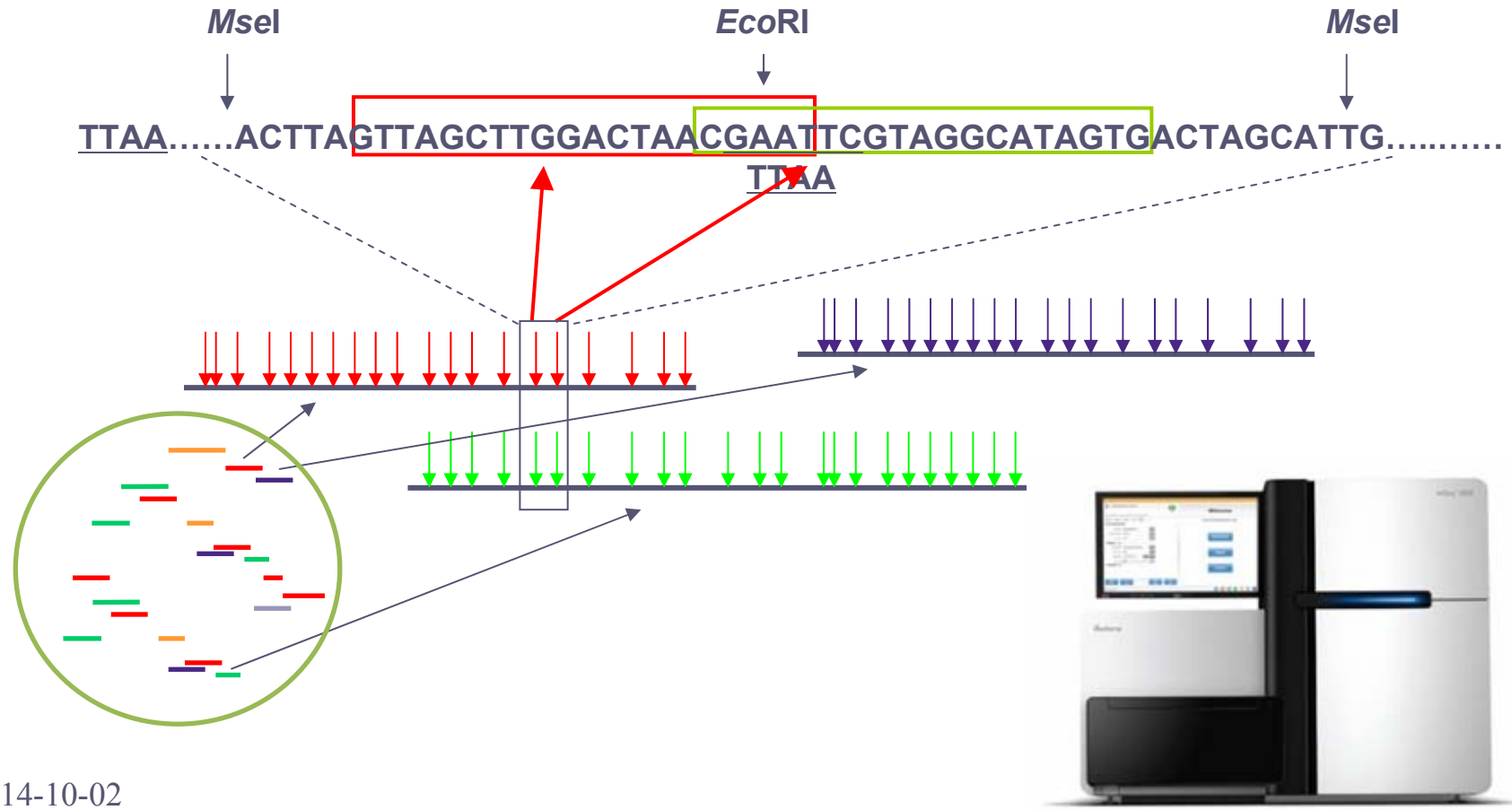
Evaluation of BAC contigs

(10 largest BAC contigs)

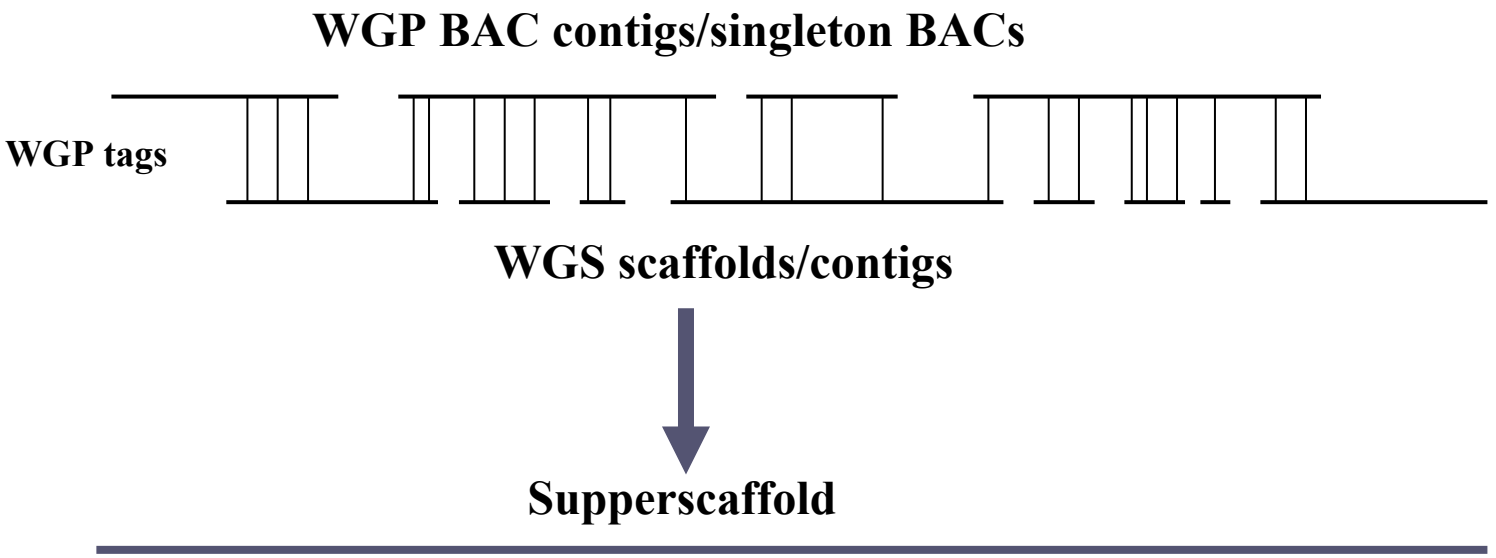
	Minimum tiling path	Sequencing & assembling	%
BAC Nos	154	152	99%
Total length of BACs (TLBS)	26,357,760	20,505,754	78%
Average length of BACs	171,154	134,906	79%
Nos of overlapped BACs	144	115	80%
Total length of overlapped sequence (TLOBS)	4,419,584	3,163,170	72%
Average length of overlapped sequences	30,692	27,506	90%
TLOBS/TLBS	17%	15%	92%

Whole Genome Profiling
(Keygene N.V.)

Sequence-based physical mapping BAC clones using Illumina Genome Analyzer II / HiSeq 2000



Integration of WGP and WGS Data



Superscaffold Building Results

	Scaffolds	Superscaffolds
Number	133,687	13,899
Total Bases with Ns	3,597,407,679	4,975,825,774
Total Bases without Ns	2,864,495,294	2,531,764,367
Largest (nt)	1,066,088	19,663,984
Smallest (nt)	1,001	8,416
Average (nt)	26,909	357,999
N50 size (nt)	**89.810**	**509,168**
N50 index	**11,635**	**2,850**
N80 size (nt)	39.151	252,480
N80 index	29,490	6,994
N95 size (nt)	8,856	126,240
N95 index	54,258	11,036

Ongoing Works

1) Constructing a high resolution genetic map with SNPs

- Generated 500 F_2 plants by crossing of *T. urartu* G1812 with G3146
- Sequencing the 500 F2 plants to identify SNPs
- Constructing a high resolution genetic map

2) Anchoring the BAC contigs/scaffolds on the genetic map to finish the physical mapping of wheat A genome.

3) Sequencing BACs (70,000) to complete the wheat A genome.

Summary

- The A genome is a basic genome of bread wheat and other polyploid wheats, played a central role in wheat evolution, domestication and genetic improvement.
- We sequenced, assembled and annotated the wheat A genome using a whole-genome shotgun sequencing strategy. 34,789 protein-coding gene models has been predicted.
- We analyzed the genome structure, and assessed its utility for analyzing agronomically important genes and for developing molecular markers.
- BAC contigs have been constructed using WGP technology. We integrated the WGS and WGP data and constructed superscaffolds. The average size of supperscaffolds was 358 kb, and its N50 size reached to 509 kb.
- *T. urartu* genome assembly provides a diploid reference for analysis of polyploid wheat genome and is a valuable resource for the genetic improvement of wheat.

Acknowledgments:

Academician Zhensheng Li

Dr. Daowen Wang

Dr. Aimin Zhang

All peoples worked on this project

Collaboration: BGI-Shenzhen

UC Davis (Prof. Jan Dvorak, Dr. Mingcheng Luo)

Keygene N.V. (Dr. Michiel van Eijk)

Sequencing Center of IGDB (Dr. Chengzhi Liang)

Funding: Chinese Ministry of Science and Technology

Thanks !