The genome assembly is provided under the generally accepted conditions of the Toronto agreement with particular reference to the condition that the data provided should not be published until the manuscript in preparation is published (Ru et al 2020, submitted) and can be acknowledged appropriately.

Zhengang Ru*[1], Angela Juhasz*[2], Danping Li*[3], Pingchuan Deng*[4], Jing Zhao*[3], Lifeng Gao*[3], Kai Wang*[5], Gabriel Keeble-Gagnere[10], Zujun Yang[6], Guangrong Li[6], , Daowen Wang[7], Utpal Bose[8], Michelle Colgrave[2,8], Chuizheng Kong[3], Guangyao Zhao[3], Xueyong Zhang[3], Xu Liu[3], Guoqing Cui[3], Yuquan Wang[1], Zhipeng Niu[1], Liang Wu[4], Dangqun Cui[4]* Jizeng Jia[3]*, Rudi Appels[9,10]*, Xiuying Kong[3]* (2020). The structure and gene complement of chromosome 1RS.1BL in Chinese wheat variety Aikang58

The assembly was derived from Chinese wheat variety Aikan58 leaf DNA. Genome sequencing used standard illumina sequencing and denovo NR Magic software for a primary assembly (contracted to NovoGene, China), followed by the HiC process for scaffolding the contigs (Jia et al 2020). The alignment of our AK58_chr1RS.1BL_v6 to the reference Chinese Spring 1B is in Ru et al. (2020, submitted) and we note that the orientation of the terminal 13874856 was ambiguous and the orientation provided in the assembly was most compatible with molecular genetics maps as well as genetic marker analyses of recombinant chromosomes (see also, Jia, et al., 2020).

The HMW glutenin locus (at 611.76 Mb) on the long arm was a problem in the standard sequence assembly protocols and the structure was clarified using PCR and a set of primers targeting the X and Y HMW glutenin genes

Annotation of the gene space assigned confidence to a gene in two steps. First, we considered genes with isoforms showing significant homology (BLASTN with e-value < 10^{-10}) to repeat elements library as low confidence genes. Second, we used BLASTP to compare the predicted peptide sequences against known protein datasets (Sit, Pvi, Hvu.HC, Tae, Tur, Bdi, Osa, Sbi, Zma, Ata, CSW.HC, Ttu.HC), as well as the set of predicted proteins from wheat FLcDNAs, and considered hits with an e-value < 10^{-10} as significant. For each gene, we selected the best-matching reference protein as a template sequence and defined the isoform sequence with maximum coverage of the template sequence as a gene representative. Genes were designated high confidence (HC) if they had a significant BLAST hit to reference proteins and if their representative protein had a similarity to the respective template sequence above a threshold which we determined based on the origin of template sequences (>60% for Sit, Pvi, Zma, *S. bicolor* and *O. sativa*; >65% for *B. distachyon*; and >90%

for *H. vulgare*, Ttu_HC, CSW_HC and Ata). Manual annotation of predicted gene models included correction of intron-exon structures based on RNAseq alignments aligned to the genome sequence and viewed in Apollo (Lee et al 2013) and comparison of predicted models to the alignment of IWGSC refseq ver1.0 gene models projected on to the genome sequence.

Retrotransposable elements were annotated using the CLARiTE software (reference)

To predict RGAs a new plant RGAs database was constructed using protein sequences from the RGAdb [1] (*https://bitbucket.org/yaanlpc/rgaugury/src/master/*) and candidate RGAs predicted in *Aegilops tauschii* genome [2]. A total of 61,372 disease resistance related sequences were obtained. Protein sequences of all annotated genes of *AK58* were aligned to the new RGAs database using BLASTP with *e-value* cutoff of 1e-05. Potential RGAs were selected based on rule 80:80:80 (query sequence coverage more than 80%, target sequence coverage more than 80 and identity more than 80%). Seven RGAs-related domains and motifs including NB-ARC, NBS, LRR, TM, STTK, LysM, CC and TIR were searched and identified by RGAugury pipeline [1]. RGA candidates were predicted in *T. aestivum* (Chinese Spring) genome using the same method.

Prolamin super-family genes identified in the wheat reference genome were used to manually annotate the prolamin genes in the AK58 genome sequence. Translated sequences were checked for the presence of signal peptides and the conserved cysteine pattern and Pfam domains as described in Juhasz et al., 2018. Obtained sequences were aligned with gliadins and secalins retrieved from the Uniprot database to confirm the protein sub-types. Expression of genes were analysed using the grain specific transcriptome data set obtained from 4, 10, 15 and 20 DPA grain libraries. Protein level expression of the translated secalins and nsLTPs were analysed using data published by Bose et al., 2019. LC-MS-MS data originally generated from tryptic digests of rye flour protein extracts were re-analysed and protein identification was undertaken using ProteinPilotTM 5.0 software (SCIEX) with the Paragon and ProGroup algorithms (Shilov et al., 2007) with searches conducted against the Poaceae subset of the Uniprot database appended with the identified 1RS gene models and

a contaminant database (Common Repository of Adventitious Proteins).Tryptic peptides were mapped to the secalin and nsLTP sequences in CLC Genomics Workbench v12 (Qiagen, Aarhus, Denmark) using 100% sequence identity to confirm the expression at individual protein levels.