Plant Communications
Research Article

CellPress
Partner Journal

A recent burst of gene duplications in Triticeae

Xiaoliang Wang[1,2,5], Xueqing Yan[1,2,5], Yiheng Hu[1,2,5], Liuyu Qin[1,2], Daowen Wang[3,*], Jizeng Jia[3,4,*] and Yuannian Jiao[1,2,*]

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

[2]University of Chinese Academy of Sciences, Beijing 100049, China

[3]College of Agronomy, Collaborative Innovation Center of Henan Grain Crops, Henan Agricultural University, Zhengzhou, Henan 450046, China

[4]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[5]These authors contributed equally to this article.

***Correspondence: Daowen Wang (dwwang@henau.edu.cn), Jizeng Jia (jiajizeng@caas.cn), Yuannian Jiao (jiaoyn@ibcas.ac.cn)**

https://doi.org/10.1016/j.xplc.2021.100268

ABSTRACT

Gene duplication provides raw genetic materials for evolution and potentially novel genes for crop improvement. The two seminal genomic studies of *Aegilops tauschii* both mentioned the large number of genes independently duplicated in recent years, but the duplication mechanism and the evolutionary significance of these gene duplicates have not yet been investigated. Here, we found that a recent burst of gene duplications (hereafter abbreviated as the RBGD) has probably occurred in all sequenced Triticeae species. Further investigations of the characteristics of the gene duplicates and their flanking sequences suggested that transposable element (TE) activity may have been involved in generating the RBGD. We also characterized the duplication timing, retention pattern, diversification, and expression of the duplicates following the evolution of Triticeae. Multiple subgenome-specific comparisons of the duplicated gene pairs clearly supported extensive differential regulation and related functional diversity among such pairs in the three subgenomes of bread wheat. Moreover, several duplicated genes from the RBGD have evolved into key factors that influence important agronomic traits of wheat. Our results provide insights into a unique source of gene duplicates in Triticeae species, which has increased the gene dosage together with the two polyploidization events in the evolutionary history of wheat.

Keywords: gene duplication, transposable elements, gene dosage, hexaploid wheat, Triticeae, agronomic traits

Wang X., Yan X., Hu Y., Qin L., Wang D., Jia J., and Jiao Y. (2022). A recent burst of gene duplications in Triticeae. Plant Comm. **3**, 100268.

INTRODUCTION

The Triticeae tribe is one of the largest taxonomic groups in the grasses and comprises many globally important food and forage crops like wheat, barley, and rye. Triticeae crop species, especially polyploid wheat, are more widely used in the agriculture of temperate regions than other cereal crops like maize and rice (He et al., 2019; Pont et al., 2019). It is known that hybridization of the diploid *Triticum urartu* (2n = 2x = 14, AA) and a close lineage of *Aegilops speltoides* (2n = 2x = 14, BB) gave rise to tetraploid wild emmer wheat (*Triticum turgidum* ssp. *dicoccoides,* BBAA), and a further hybridization of a domesticated emmer wheat with the diploid *Aegilops tauschii* (2n = 2x = 14, DD) formed allohexaploid common wheat (*Triticum aestivum,* BBAADD) (Petersen et al., 2006; Marcussen et al., 2014). Tetraploid wheat, especially durum wheat, is becoming a valuable food crop worldwide because of its versatile processing properties and high nutritional value (Maccaferri et al., 2019). Hexaploid bread wheat, which provides about a fifth of the calories consumed by humans and contributes more protein than any other food source, is the most commonly cultivated crop on earth (IWGSC et al., 2018; He et al., 2019; Pont et al., 2019).

In recent years, many large, complex, highly repetitive genomes of Triticeae species have been deciphered (Luo et al., 2017; Mascher et al., 2017; Zhao et al., 2017; IWGSC et al., 2018; Ling et al., 2018; Guo et al., 2020; Jayakodi et al., 2020; Walkowiak et al., 2020; Wang et al., 2020; Li et al., 2021; Rabanus-Wallace et al., 2021; Zhou et al., 2021). Genomes of diploid wheat species (e.g., barley and rye) range from 4.3 Gb to 7.9 Gb in size and contain more than 40,000 annotated genes (Bauer et al., 2017; Mascher et al., 2017; Zhao et al., 2017; Ling et al., 2018; Wang et al., 2020; Li et al., 2021; Rabanus-Wallace et al., 2021; Zhou et al., 2021). The tetraploid

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

emmer genome comprises 10.5 Gb of genomic sequence and 65,012 protein-coding genes (Avni et al., 2017). The genome of the hexaploid bread wheat Chinese Spring (CS) contains 14.5 Gb of sequence and 107,891 high-confidence genes (IWGSC et al., 2018).

In the CS genome, approximately 55% of the homologous genes have been reported to exhibit 1:1:1 correspondence across the three homoeologous subgenomes, and the other 15% have more than one gene copy in at least one of the subgenomes (IWGSC et al., 2018). Furthermore, two genomics studies of *Ae. tauschii*, the donor of the hexaploid wheat D subgenome, revealed an apparently recent burst of gene duplications. The authors speculated that recently duplicated genes were likely to be related to the remarkable genomic enrichment of transposable elements (TEs) (Luo et al., 2017; Zhao et al., 2017). Analysis of intra-genomic synteny of *Ae. tauschii* clearly showed that its most recent whole genome duplication (WGD) was *rho*, which occurred before the divergence of Poaceae species (Tang et al., 2010; Jiao et al., 2014), and these recent duplications were independent and dispersed throughout the genome rather than derived from WGD (Zhao et al., 2017). These recent gene duplications may, at least in part, explain why so many genes in the three subgenomes of CS are not in 1:1:1 correspondence. Therefore, expanded homologous genes in wheat arise not only from polyploidization events but also from recent independent duplications. However, studies regarding the extent, timing, and mechanisms of these recent duplications in different Triticeae species are still lacking. Moreover, it remains unclear whether these duplicates are functionally important for wheat.

The proportions of TEs in these Triticeae genomes are about 80% to 90%, much higher than those of most other grasses (Mascher et al., 2017; Wicker et al., 2018). It has been proposed that TE activities can generate new genes and novel *cis*-regulatory elements and can also modify the epigenetic status of specific genomic regions (Deniz et al., 2019). Occasionally, such activities lead to adaptive effects. For example, Helitrons-like TEs in maize seem to produce new nonautonomous elements for the duplicative insertion of gene segments into new locations that change both the genic and nongenic fractions of the genome, profoundly affecting genetic diversity (Morgante et al., 2005).

Here, we selected a number of representative Triticeae genomes and performed a comprehensive investigation of their recently duplicated genes, classifying them into duplicates from WGD, tandem duplication (TD), proximal duplication (PD), and dispersed duplication (DD) (for definitions, see methods). We discovered a common pattern of a recent burst of gene duplications (RBGD) in these Triticeae genomes and obtained empirical evidence indicating that TEs may have been involved in generating the RBGD. Gene duplications and losses were then examined across the evolutionary history of Triticeae species diversification and allohexaploid wheat formation. Finally, we demonstrated the importance of the RBGD for differentiating the donor genomes of bread wheat and for increasing the genetic dosage, allowing for the evolution of genes that underlie important wheat agronomic traits.

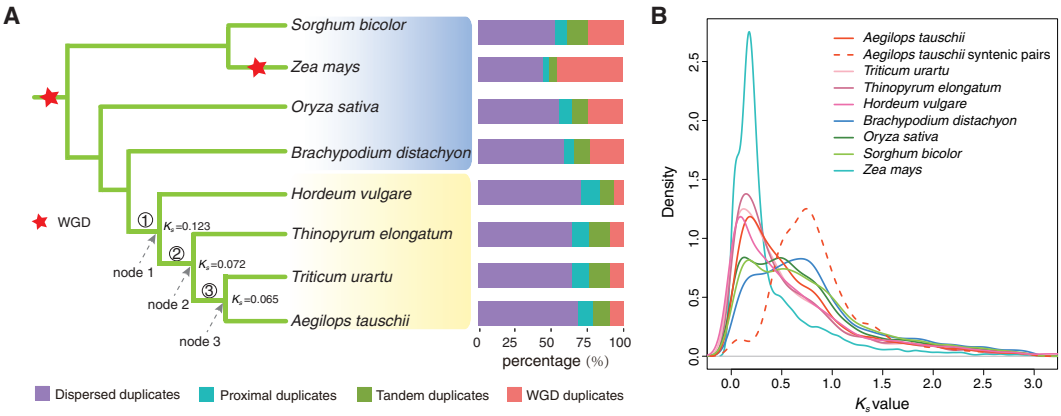
RESULTS

Identification and characterization of recently duplicated genes

We used the best-reciprocal blast approach to retrieve paralogous gene pairs from eight sequenced diploid genomes in the Poaceae: *Sorghum bicolor, Zea mays, Oryza sativa, Brachypodium distachyon, Hordeum vulgare, Thinopyrum elongatum, Ae. tauschii*, and *T. urartu* (Figure 1A; Supplemental Table 1). To distinguish between gene duplications from historical WGD events and those from recent small-scale duplications (SSD), we performed self-genomic comparisons and classified the identified syntenic gene pairs as having arisen from WGD. The remaining duplicates were classified into three categories (TD, PD, and DD) based on the genomic distances between the gene duplicates (see methods) (Supplemental Figure 1; Supplemental Table 2). In general, the proportions of PD and DD gene pairs, which are the result of small-scale duplication events, are about two times higher in Triticeae species than in sorghum, maize, rice, and *Brachypodium* (Wilcoxon test, $p < 0.01$) (Supplemental Table 2). Specifically, we detected 9,044 to 9,787 duplicated gene pairs in the examined Triticeae species: 603 (6.3%) to 924 (10.2%) PD gene pairs and 2254 (23.4%) to 3006 (33.2%) DD gene pairs (Supplemental Table 2).

Synonymous substitution (K_s) analysis clearly showed a peak around 0.2 for all of the Triticeae species we examined (Figure 1B) and indicated that the RBGD is actually a common feature of Triticeae species. The peak K_s values for the syntenic gene pairs in the *Ae. tauschii*, *Oryza*, *Sorghum*, and *Brachypodium* genomes are around 0.75 (Figure 1B), and these duplicates resulted from the *rho* WGD event (Paterson et al., 2004; Tang et al., 2010; Jiao et al., 2014; Wang et al., 2015). A unique K_s peak observed for *Z. mays* reflected a recent WGD in the maize lineage (Schnable et al., 2009). A Gene Ontology (GO)-based analysis revealed functional enrichment of these recently duplicated genes for categories such as protein dimerization activity, xylan metabolic process, catalytic activity, and nucleobase-containing compound metabolic process (Supplemental Figure 2). These categories are distinct from those that are typically retained (and thus enriched) after WGD events in diverse sets of eukaryotes (e.g., kinases, transferases, transporters, transcription regulators, and transcription factors) (Maere et al., 2005; Freeling, 2009; Jiao et al., 2014). In addition to characterizing the distinct functional gene categories of RBGD, these results clearly suggest that RBGDs are apparently common in Triticeae genomes.

We next focused our analyses on the Triticeae by examining the duplicated gene pairs in the four diploid species in detail. We used inter-genomic synteny comparisons to determine whether both of the gene duplicates were located in inter-genomic syntenic blocks. The K_s divergences between *T. urartu* and *H. vulgare, Th. elongatum,* and *Ae. tauschii* were 0.123, 0.072, and 0.065, respectively (Supplemental Figure 3), and we also compared these values with the K_s values of the RBGD gene pairs to date the timing of these gene duplications. Specifically, about half of the genes (1497, 1514, 1431, and 1333 for *H. vulgare, Th. elongatum, T. urartu,* and *Ae. tauschii,* respectively) were duplicated in node 1 (before the differentiation of the Triticeae,



C

	Number of dispersed and proximal duplicates with K_s<0.3	Number of shared duplicates from synteny analysis			Number of shared duplicates from K_s* analysis			Number of Species-specific duplicates
		node 1	node 2	node 3	node 1	node 2	node 3	
H. vulgare	3930	829	-	-	668	-	-	2433
Th. elongatum	3331	664	577	-	850	328	-	910
T. urartu	3904	529	507	136	902	340	43	1864
Ae. tauschii	2857	483	395	115	850	282	62	670

Figure 1. Triticeae species have more recent gene duplicates than other Poaceae species.

(A) Phylogeny of representative Poaceae species (left). Red stars mark two well-acknowledged ancient WGD events. Percentage of recent duplicates classified into four duplication mechanisms in Poaceae species (right).

(B) K_s plot of recent duplicates in major Poaceae crops in (A). The K_s peaks for Triticeae species at ~0.2 suggest a burst of recent gene duplication. The peak K_s value for *Ae. tauschii* syntenic pairs (dashed line) represents the *rho* WGD event, which closely coincides with the K_s peaks for *Oryza*, *Sorghum*, and *Brachypodium*.

(C) Number of recently duplicated gene pairs in *H. vulgare, Th. elongatum, T. urartu,* and *Ae. tauschii* and the phylogenetic timing of their duplications. The duplicates were dated by synteny analyses and K_s analyses. The asterisk indicates that K_s analyses were carried out after synteny analyses.

with a K_s of approximately 0.123), and about a quarter of the genes (905, 847, and 677 for *Th. elongatum, T. urartu,* and *Ae. tauschii,* respectively) were duplicated in node 2 (before the differentiation of *Th. elongatum* and *Triticum*, with a K_s of approximately 0.072) (Figure 1C). A small number of genes (179 and 177 for *T. urartu* and *Ae. tauschii,* respectively) were duplicated in node 3 (before the differentiation of *Triticum*, with a K_s of approximately 0.065) (Figure 1C). These results suggest that a burst of recent gene duplications occurred before the divergence of Triticeae species and that further lineage-specific duplications have also been occurring thereafter.

Possible mechanism of recent gene duplication

Two genomics studies of *Ae. tauschii* proposed that the apparent burst of recently duplicated genes in this species was probably related to the remarkable genomic enrichment of TEs (Luo et al., 2017; Zhao et al., 2017); however, empirical evidence supporting this hypothesis is still lacking. We investigated the particular types of TEs, including both long terminal repeat retrotransposons (LTR-RTs) and DNA transposons, that flanked the recently duplicated genes. Specifically, we identified any TEs that were located within 3,000 base pairs upstream and downstream of all of the recently duplicated genes. We found that the LTR-RTs were the most abundant type (68.3%), followed by the LINE and DNA/CACTA subtypes (Figure 2A). Notably, we found that about 21% to 42% of the TE-flanked, recently duplicated gene pairs possessed intronless duplicates (Figure 2B). Therefore, retrotransposition may be a major mechanism of gene duplication in these Triticeae genomes, as a conspicuous feature of retrotransposition is the formation of an intronless copy of a parental gene (Kim et al., 2017). Here, we show two examples of recently duplicated genes and their flanking sequences in *H. vulgare* and *Ae. tauschii* (Figure 2C). The *H. vulgare* duplicated gene pair *HORVU1Hr1G020310* and *HORVU4Hr1G059030* are located within TEs of the same LTR/Gypsy subtype with 94% sequence identity. Similarly, the duplicated gene pair *evm.model.Contig89.16* and *evm.model.Contig263.36* in *Ae. tauschii* are located within TEs of the same DNA/MULE subtype with 98% sequence identity (Figure 2C).

Given the prevalence of TEs throughout the genomes of Triticeae, we next investigated the chances of two genes duplicated in a WGD being flanked by similar types of TEs. The results revealed a clear trend: a large proportion (38%–42%) of the recently duplicated genes in the four diploid genomes were flanked by TEs of the same subtype (e.g., GYPSY, COPIA, etc.), whereas only ~10% of the syntenic gene pairs (generated by WGD) were flanked by TEs of the same subtype in the four diploid genomes (Figure 2D). When we randomly selected two genes from individual Triticeae genomes, only ~5% of them were flanked by TEs of the same subtype (Figure 2E). Thus, genome-wide empirical evidence supports a major functional contribution of TEs to the generation of RBGDs in Triticeae.

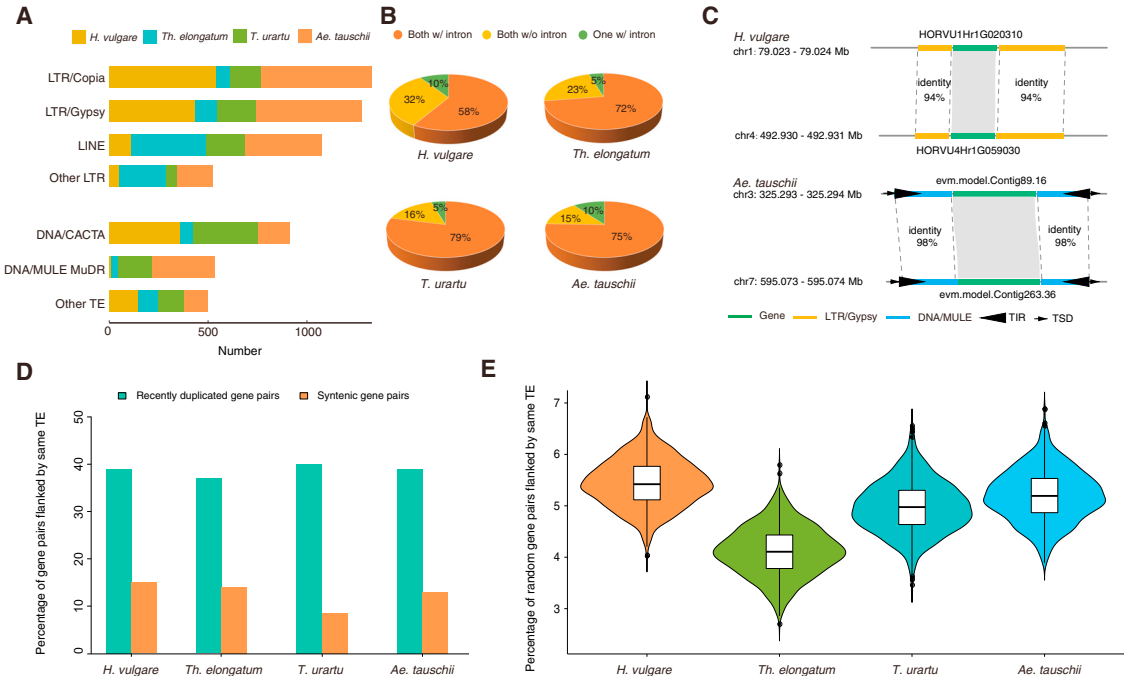


Figure 2. Recently duplicated genes tend to be surrounded by TEs.

(A) Histogram showing the number of TE-flanked, dispersed duplicate gene pairs for each TE type.

(B) Pie chart showing three types of intron distribution within duplicated gene pairs.

(C) Two examples of gene duplicates embedded in TEs of the same subtype with high sequence identity.

(D) Percentage of dispersed and syntenic duplicate gene pairs that are flanked by the same TE type (e.g., Copia, Gypsy, LINE).

(E) Percentage of randomly selected gene pairs that are flanked by the same type of TE. We rarely found two genes flanked by the same type of TE (approximately 5% by chance).

The retention and conservation of the recently duplicated genes

To further understand the genetic contribution of these recently duplicated genes to polyploid wheats, we investigated the retention and diversification of the RBGDs after the formation and diversification of allohexaploid wheat. First, we identified and compared recent duplicates in the genomes of *T. urartu* and *Ae. tauschii* to the corresponding subgenomes of wild and cultivated tetraploid wheat and the subgenomes of hexaploid bread wheat cultivars (Figure 3). We found 1,925 and 2,010 duplicates in subgenome A and B of wild emmer wheat (WEW), and 2,116 and 2,402 duplicates in subgenomes A and B of durum wheat (DEW) (Figure 3B). For hexaploid wheat, there are 2,560, 2,625, and 2,450 duplicates in subgenomes A, B, and D of CS and 2,374, 2,642, and 2,497 duplicates in subgenomes A, B, and D of Jagger (JAG) (Figure 3B and Supplemental Figure 4). JAG and CS are two representative hexaploid wheats that originated in the West and the East, respectively. We found that JAG and CS have about 2,000 co-retained gene pairs in each subgenome (i.e., more than 80% are shared) (Supplemental Figure 4).

We next investigated the retention patterns of these recent gene duplicates after the two successive polyploidization events using CS as a representative hexaploid wheat (Figure 3B and Supplemental Figure 5; Supplemental Tables 3–5). We found that 508, 891, and 1,320 gene pairs were co-retained in the A, B, and D subgenomes after polyploidization events (Figure 3B). We also investigated the duplication times by comparing K_s divergence of these RGBDs with the corresponding species divergence times to separate the species-specific gene pairs into specifically retained or newly duplicated gene pairs in each species. For the A subgenome, 1,108, 450, 369, and 595 gene pairs were specifically retained, and 1,635, 199, 207, and 238 gene pairs were newly duplicated in *T. urartu*, emmer wheat, durum wheat, and CS, respectively. For the B subgenome, 702, 506, and 865 gene pairs were specifically retained, and 263, 270, and 288 gene pairs were newly duplicated in emmer wheat, durum wheat, and CS, respectively. For the D subgenome, 1,203 and 859 gene pairs were specifically retained, and 419 and 217 gene pairs were newly duplicated in *Ae. tauschii* and CS, respectively (Figure 3B). We further compared the particularly well-retained subset with the specifically retained gene pairs and found that the well-retained gene pairs were characterized by their typically higher K_s values (Wilcoxon test, $p < 0.01$) (Supplemental Figure 6). Moreover, genes in the well-retained subset had clearly undergone stronger purifying selection than genes of other duplicated pairs in common wheat that showed no obvious synteny to progenitor genomes (Wilcoxon test, $p < 0.01$) (Supplemental Figure 6).

We next investigated the retention pattern of gene duplicates that were generated before the diversification of the *Triticum* and *Aegilops* species in multiple hexaploid wheat genomes, including JAG, CS, and nine other newly available wheat genomes. In CS, we found that 5,300 of 7,821 gene pairs (1,688, 1,893, and 1,719 in the three subgenomes, respectively) from RBGD were duplicated before the divergence of the *Triticum* and *Aegilops*

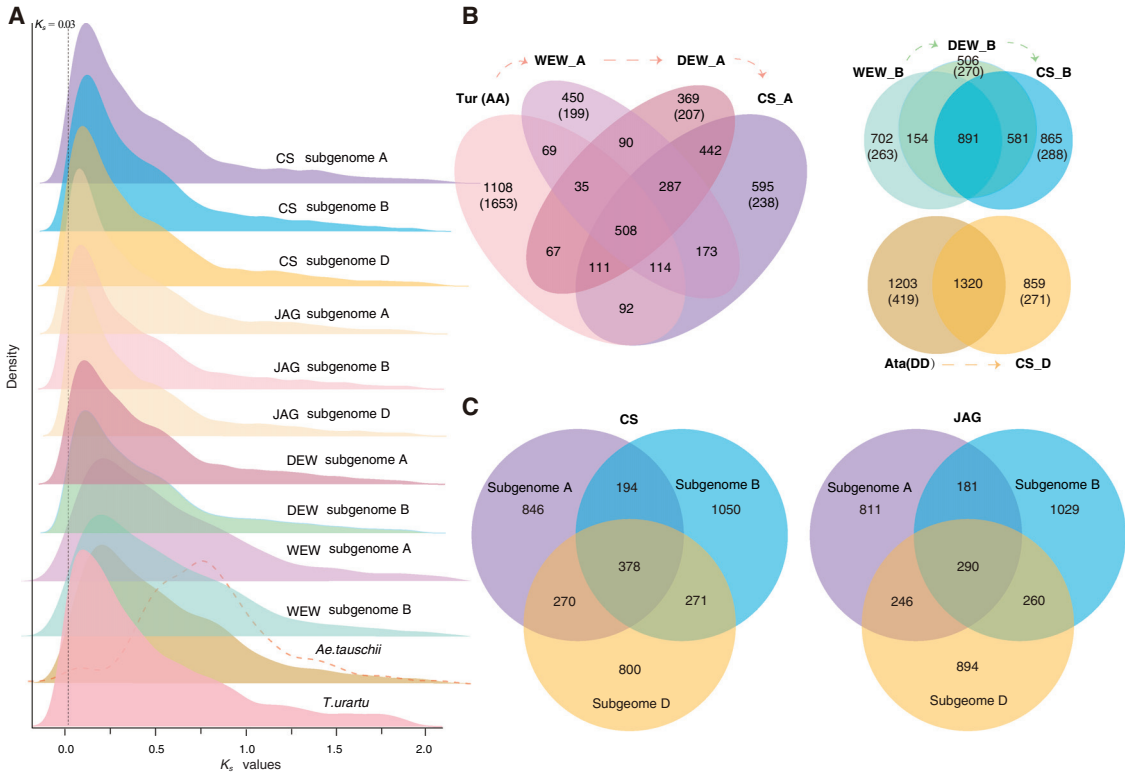


Figure 3. The retention and conservation of the recently duplicated genes in CS.

(A) K_s plot of recent duplicates in CS and JAG that originated in the West and their progenitor species. Dashed line on the *Ae. tauschii* plot represents the K_s distribution of the syntenic gene pairs that arose from the *rho* WGD event.

(B) Venn diagrams show the retention pattern of the recent duplicates following the evolution of the diploid progenitor, wild emmer, domesticated emmer, and CS wheat. The numbers in parentheses show the number of newly duplicated gene pairs in the progenitors or wheat subgenomes for which no corresponding orthologs were identified in other genomes.

(C) Venn diagrams show the commonly retained recent gene duplicates in the three subgenomes of CS and JAG. These retained gene pairs were duplicated prior to the diversification that led to the diploid AA, BB, and DD species, based on K_s analysis.

species. Among these 5,300 duplicated genes, 378 pairs of genes were well retained after two allopolyploidization events (each set of homologous genes contains six copies in CS), and 846, 1,050, and 800 gene pairs were specifically retained in the A, B, and D subgenomes of CS, respectively (Figure 3C). Similarly, in JAG, 4,978 of 8,573 gene pairs were duplicated before the divergence of the *Triticum* and *Aegilops* species; 290 duplicates were retained in the three subgenomes of JAG, and 811, 1,029, and 894 gene pairs were specifically retained in the A, B, and D subgenomes of JAG, respectively (Figure 3C). Similarly, we identified about 300 co-retained gene pairs and approximately 800, 1,000, and 800 specifically retained gene pairs in the three subgenomes of the other nine sequenced wheat genomes (Supplemental Figure 7). Among these co-retained gene pairs, about 70% to 80% were shared among the hexaploid wheat genomes, whereas CS and JAG shared only 60% of these co-retained gene pairs (Supplemental Tables 6 and 7). A GO-based analysis revealed functional enrichment of these co-retained pairs (~300 pairs) in CS and JAG in categories such as aminoacyl-tRNA ligase activity, tRNA aminoacylation, and tRNA metabolic process. Categories of chromatin modification and histone modification were only enriched in the CS retained duplicates, and the category of transporter activity was specifically enriched in JAG retained duplicates (Supplemental Figure 8).

The diversification patterns of the recently duplicated genes

Given the allohexaploid nature of wheat, we also performed multiple subgenome-specific comparisons of the duplicated gene pairs to investigate any differential regulation and related functional diversity among such pairs in the three subgenomes. First, patterns of functional category enrichment (GO categories) among the retained duplicates differed among the three CS subgenomes; for example, nutrient reservoir activity was enriched in the gene pairs of the A subgenome, macromolecule biosynthetic process was enriched in the gene pairs of the B subgenome, and oxidoreductase activity was enriched in the gene pairs of the D subgenome (Figure 4A). Second, the basic trend from an RNA-sequencing (RNA-seq)-based analysis showed weaker expression for genes of pairs present in a single subgenome compared with genes of pairs whose orthologous gene pairs were retained in two or three subgenomes (Figure 4B). We found that 38% of the subgenome-specific retained duplicates exhibited no expression, a larger percentage than that of the non-subgenome-specific gene pairs (Figure 4B). It was notable that the 378 gene pairs common to all three subgenomes exhibited the highest expression levels (Figure 4B). Third, after reconstructing co-expression modules using the RNA-seq data, we found that about 25% of the subgenome-specific

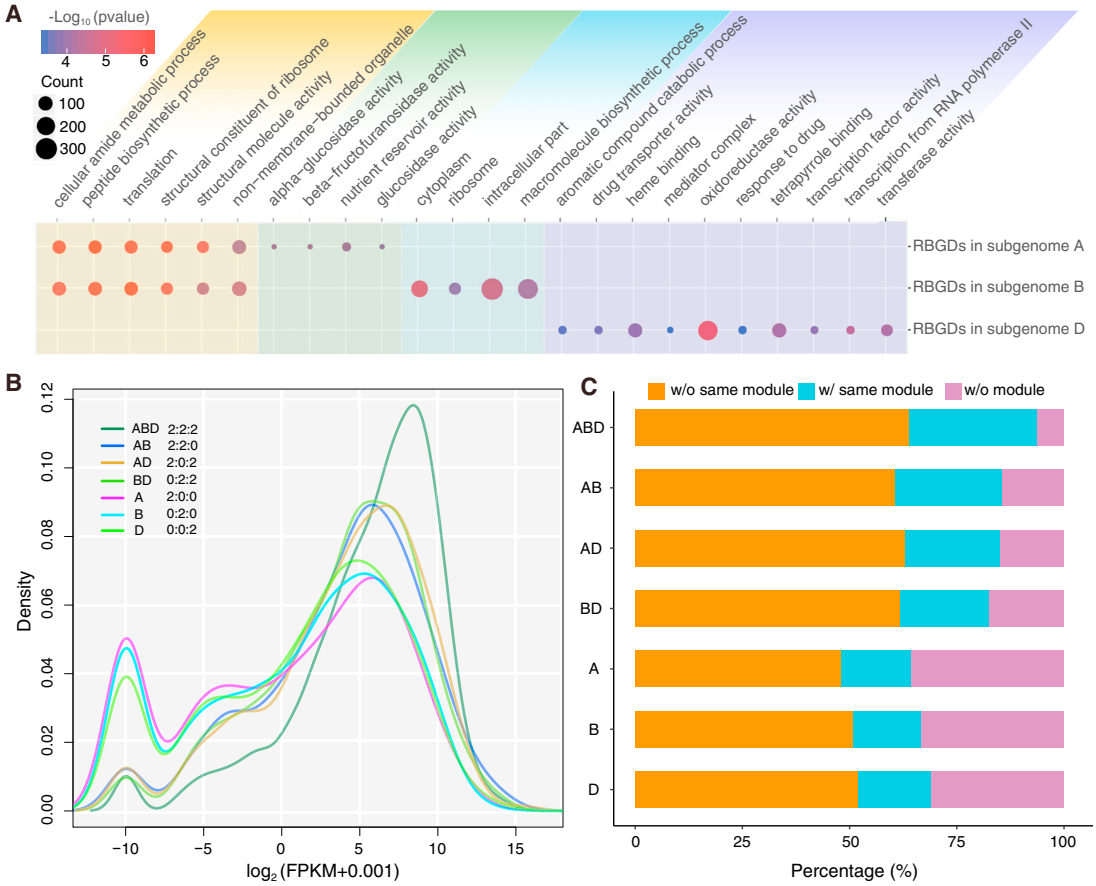


Figure 4. The diversification patterns of the recently duplicated genes in CS.

(A) Different patterns of GO enrichment for recently duplicated genes in the three subgenomes of CS.

(B) Distribution of expression levels for differentially retained genes among the three subgenomes of CS. "ABD" indicates that the three subgenomes all retained duplicates; "AB" indicates that only the A and B subgenomes retained duplicates; and "A" indicates that only the A subgenome retained duplicates.

(C) Differentially retained duplicates assigned to particular modules (same, divergent, or none) in a co-expression network analysis. w/o same module indicates divergent co-expression networks; w/same module indicates the same network; w/o module indicates that neither gene was assigned to a network.

duplicates were not clustered into any modules, compared with only about 10% of the multi-subgenome retained pairs (Figure 4C). Further co-expression network analysis revealed that a larger percentage of the duplicates common to all subgenomes diverged into different modules compared with the subgenome-specific duplicates (73% versus 55%), indicating possible sub- or neo-functionalization of the duplicates over evolutionary time (Figure 4C). Collectively, these analyses emphasize that distinguishing among ancient versus recent duplicates and among subgenome-specific duplicated gene pairs is a viable analytical strategy for isolating specific trends in the regulation and attendant expression divergence of these genes and thus their potential sub- and neo-functionalization.

Evolutionary and expression analyses of *NAC* genes

Several genes derived from the RBGD have been previously identified as agronomically important genes in wheat, e.g., *Sr21*, *Sr33*, and *Sr35*, which specify stem rust resistance (Periyannan et al., 2013; Saintenac et al., 2013; Chen et al., 2018), *Yr10*, which specifies stripe rust resistance (Liu et al., 2014), *Lr1*, which specifies leaf rust resistance (Feuillet et al., 1995), *Pm3B*, which specifies powdery mildew resistance (Brunner et al., 2011), *GPC*, which controls the contents of proteins and health-promoting minerals (iron and zinc) in the grain (Uauy et al., 2006), and phosphomannomutase (*PMM*), which functions in temperature adaptability (Yu et al., 2010) (Figure 5A). In addition, we found that most of these duplicates were derived from the ancestor of the Triticeae (Supplemental Figure 9). We conducted a more systematic study of the evolutionary history of *GPC* genes (encoding *NAC* transcription factors), among which *NAM-B1* is well studied for its function in accelerating leaf senescence and increasing grain protein content in wheat (Uauy et al., 2006). Through phylogenetic and syntenic analyses, we found that a duplication belonging to RBGDs occurred before the divergence of Triticeae species, creating the *NAM-B1* on chromosome 6 from its parental gene on chromosome 2 (Figure 5B). We identified five *NAM* homologs in CS and found that the copy on chromosome 6B was lost. Moreover, we found that similar types of TEs flanked the five *NAM* homologs (two homoeologous pairs in the A and D subgenomes plus one singleton in the B subgenome), indicating potential involvement of TE activity in generating the duplicated functional *NAM-B1* allele before the divergence of Triticeae (Figure 5B and 5C).

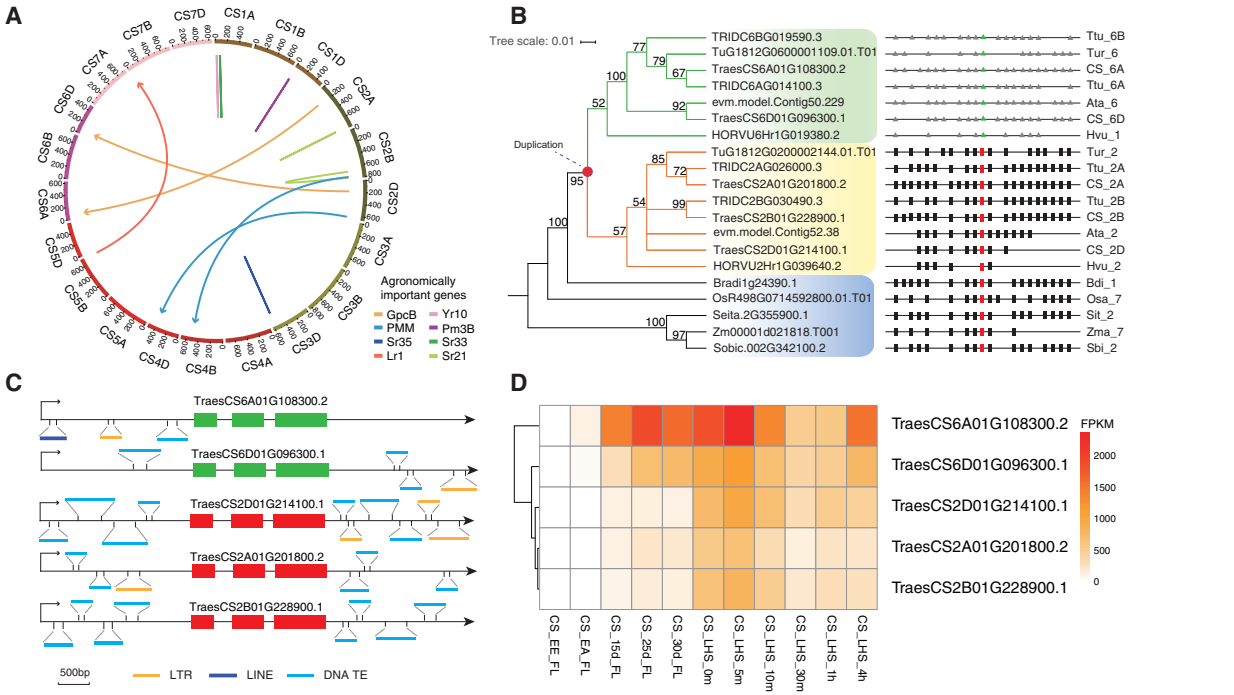


Figure 5. Evolutionary and expression analyses of *NAC* genes.

(A) Circos plot showing eight previously identified important genes that have experienced the RBGD. The known agronomically important genes are associated with stem rust resistance (*Sr21*, *Sr33*, *Sr35*), stripe rust resistance (*Yr10*), leaf rust resistance (*Lr1*), powdery mildew resistance (*Pm3B*), phosphomannomutase (*PMM*), and earlier senescence and higher grain protein, iron, and zinc content (*GPC*). The arrow/line represents the direction of gene duplication from the ancestral gene to the newly duplicated copy.

(B) Maximum likelihood phylogeny of the *NAC* genes and the syntenic regions that contain *NAC* genes in other Poaceae genomes. A red solid circle in the phylogenetic tree represents one of the RBGD duplication events that created a duplicated copy in chromosome 6 of the common ancestor of Triticeae species. The right side of the phylogenetic tree presents the syntenic regions with *NAC* genes. The identified syntenic relationships among genes shown as black and red rectangles suggest that the genes in group I are positionally conserved, and therefore the ancestral copies, whereas the genes in group II that are illustrated as green triangles surrounded by gray triangles are the new, duplicated copies.

(C) Schematic diagrams showing the gene structure and flanking TEs around *NAC* genes in CS. Different types of TEs are indicated by bars with different colors.

(D) Expression levels of *NAC* genes in CS. The duplicated copy of *TraesCS6A01G108300* has the highest expression in the flag leaf among the five *NAC* genes. EE, ear emergence; EA, anthesis; LHS, leaf under heat stress.

We examined the expression pattern of the remaining five *NAM* genes in CS using 100 RNA-seq samples (Ramírez-González et al., 2018). The expression of the *NAM-A1* gene (*TraesCS6A01G108300*), which resulted from duplication, was significantly higher in the flag leaf than that of other *NAM* genes (Figure 5D). This result may reflect the modification of regulatory elements because of the removal of TEs downstream of *TraesCS6A01G108300* or variations in TEs in the upstream region (Figure 5C). Further functional experiments to identify and test the regulatory elements around *TraesCS6A01G108300* may help to unravel the underlying mechanisms that cause increased expression of the novel duplicated gene. However, the case study of *NAM* indicates that the RBGDs may have quickly increased the dosage of agronomically important wheat genes, in addition to the two consecutive allopolyploidization events.

DISCUSSION

Gene duplicates and their duplication mechanisms

Gene duplication provides raw genetic material for evolution and adaptation and is considered to be a driving force in evolution (Ohno, 1970; Adams and Wendel, 2005). Multiple mechanisms have been proposed to generate gene duplicates (Panchy et al., 2016; Qiao et al., 2019; Zhang et al., 2020). Polyploidization is a major source of large-scale gene duplication because it involves the doubling of the entire genome (Soltis et al., 2015; Van de Peer et al., 2017). In this study, we observed a large number of recent gene duplications in all sequenced Triticeae species, a finding that is commonly, if sometimes mistakenly, interpreted as evidence for a WGD event. Genomic synteny comparisons clearly showed that these gene duplicates are the result of independent SSDs rather than a WGD event. However, it is challenging to determine the mechanism if a reference genome is not available, and that is why there are such active controversies (Wang et al., 2019; Zwaenepoel et al., 2019).

In addition to the genomic positions of the duplicated genes, their functional categories can provide another perspective on their possible origins. In an extremely diverse set of eukaryotes, retention of gene duplicates after WGD events was shown to be biased toward certain categories, such as kinases, transferases, transporters, transcription regulators, and transcription factors

(Davis and Petrov, 2005; Maere et al., 2005; Freeling, 2009; Jiao et al., 2011). If no chromosomal genome assembly is available, we can compare the enriched GO categories of the identified gene duplicates with those typically enriched in the duplicates retained after WGD events. In this study, we found apparently distinct functional categories for the RGBD genes in Triticeae species, thus clearly excluding the possibility of their WGD origin. Therefore, the enriched GO pattern of duplicates can serve as complementary evidence to determine whether duplications are the result of an SSD or WGD event.

TE-mediated gene duplication

TEs are widespread components of plant genomes, and expansion in TE numbers can cause dramatic differences in the overall architecture of plant genomes (Arabidopsis Genome Initiative, 2000; Tenaillon et al., 2010; Lisch, 2013; IWGSC et al., 2018). TE activity can cause a broad range of changes in gene expression and function, as well as the evolution of entirely new genes (Kaessmann et al., 2009; Lisch, 2013; Tan et al., 2016; Cerbin and Jiang, 2018). In this study, we found that RBGDs in Triticeae genomes were clearly associated with TEs: a large proportion (38%–42%) of the recently duplicated genes in the four diploid genomes were flanked by TEs of the same subtype and obviously did not result from tandem duplications. We also found that 59% of TEs from the same subtype associated with gene duplications had high identities, greater than 90%. Notably, we found that about 21% to 42% of these same TE-flanked recently duplicated gene pairs had intronless duplicates, which is also powerful evidence, especially for LTR-RT-mediated duplicates. For example, *TraesCS1B01G041800* and *TraesCS6B01G016300* are located beside TEs of the same subtype with 91% sequence identity; the duplicated copy (*TraesCS6B01G016300*) lacks introns (Supplemental Figure 10). These findings suggest that the abundant TEs in Triticeae may have created a large number of new genes via previously reported mechanisms, although other mechanisms such as haplotype recombination may also have contributed to some of these duplications (Jiang et al., 2004; Wang et al., 2006; Kaessmann et al., 2009; Kim et al., 2017).

In this study, we found similar TEs near ~40% of the RBGD genes, and we suspect that the rest of the duplicates may have been generated from other mechanisms or their flanking TEs may have undergone sequence divergence during evolution. In fact, we found that the K_s values of the duplicated genes that were not flanked by TEs of the same subtype were larger than those of duplicates with similar TEs (Wilcoxon test, $p < 0.01$) (Supplemental Figure 11A). Moreover, we also found that the larger the K_s values of the duplicated genes, the lower the identity of their flanking TEs (Supplemental Figure 11B). This trend is consistent with previous reports that only relatively young duplications via TEs can be detected (Jiang et al., 2004; Morgante et al., 2005; Wang et al., 2006; Xiao et al., 2008; Kim et al., 2017; Cerbin and Jiang, 2018). Notably, our reported RBGD includes some duplications that occurred nearly 10 million years ago, and we expect that many other sequence divergences may have occurred and thus erased the signature of the similar TEs (if they existed) over such a long evolutionary period.

Polyploidy advantage of bread wheat and RBGD

Bread wheat has a large, redundant, and allohexaploid genome, making it by far the largest and most complex genome of all sequenced plant species. The genome of the wheat cultivar CS contains 14.5 Gb of sequence and 107,891 high-confidence genes, a larger number of genes than any other sequenced diploid genome. The complexity of the wheat genome is due not only to its allohexaploid nature but also to its enrichment in repetitive sequences and TEs. These features may make a large contribution to its genetic diversity and innovation during evolutionary history, making wheat one of the most complicated genomes.

The advantage of wheat polyploidy may be associated, at least in part, with the increased gene dosage produced by genome merging (Ramírez-González et al., 2018), and the resulting redundant genes may go through mutation robustness, differential gene loss, subgenomic expression dominance, or divergence, which often lead to novel functional molecular networks and ultimately to phenotypic innovations (Wu et al., 2020). As reported previously, 55% of genes exhibit perfect 1:1:1 correspondence across the three subgenomes of CS (Ramírez-González et al., 2018). As we reported here, a recent burst of small-scale gene duplications also occurred during the evolutionary history of speciation and diversification of Triticeae, probably because of TE enrichment in the Triticeae genomes. Thus, in bread wheat, certain functional genes dramatically increased in dosage through both allopolyploidization events and RBGD, and the resulting increased gene dosage may have contributed to the polyploidy advantage of bread wheat. Many previously identified agronomic genes in polyploid wheat species have experienced recent duplications, a finding that highlights the genetic contribution and general importance of RBGD for common wheat.

In conclusion, we revealed a common, recent burst of numerous gene duplications in the Triticeae species, a novel feature of Triticeae that has not been reported for any other clades of green plants. We also provided evidence suggesting that the RBGD resulted from the abundant TEs in Triticeae genomes. By investigating the birth and death patterns of the recently duplicated genes in the Triticeae species, we found that the RBGD began after the origin of Triticeae species, and a large number of young genes may have contributed to their species diversification. Probably because of increased dosage or sub-/neo-functionalization of gene duplicates, several genes have evolved into key factors that function in agronomically important traits of wheat.

METHODS

Genomic data resources

We selected 10 taxa in the Poaceae clade that have whole-genome assemblies: *H. vulgare* (Mascher et al., 2017), *Th. elongatum* (Wang et al., 2020), *Ae. tauschii* (Zhao et al., 2017), *T. urartu* (Ling et al., 2018), *T. turgidum* (Avni et al., 2017; Maccaferri et al., 2019), *T. aestivum* (IWGSC et al., 2018; Walkowiak et al., 2020), *O. sativa* (Goff et al., 2002), *B. distachyon* (Vogel et al., 2010), *Z. mays* (Jiao et al., 2017), and *S. bicolor* (Paterson et al., 2009). Genomic data were downloaded from public repositories or specific project websites (Supplemental Table 1).

Genomic synteny analyses

We performed self-alignment of the protein sequences using BLASTP (Altschul et al., 1997) with parameters "-outfmt 6 -evalue 1e-5", and the

top 15 hits were extracted as an input file for MCScanX (Wang et al., 2012). The intra-genome syntenic blocks were detected using MCScanX with parameters "-e 1e-5 –m 25 –w 5" (Wang et al., 2012). Gene pairs in collinear blocks were identified as whole-genome duplicates.

Paralogous gene detection and classification

We performed genome-wide, all-by-all BLASTP (Altschul et al., 1997) with parameters "-outfmt 6 -evalue 1e-5", and the best reciprocal matches were then extracted as the paralogous genes. For all of the examined Poaceae genomes, we classified the paralogous genes into four categories: tandem duplicated pairs (located within five genetic loci of each other), proximal duplicated pairs (within 5–10 genetic loci), dispersed duplicated pairs (more than 10 genetic loci apart), and duplicated pairs from WGD (gene pairs with evidence of genomic synteny).

Statistical test

The Wilcoxon test was used to evaluate differences between groups (Supplemental Figures 6 and 11A). Taking Supplemental Figure 6 as an example, we divided the duplicates into two groups based on whether they were conserved. We then tested the significance of differences in K_a, K_s, and K_a/K_s between these two groups of data. A p value of <0.05 was considered to be statistically significant: NS (not significant) $p > 0.05$, *$p < 0.05$, **$p < 0.01$.

Synonymous substitution (K_s) analysis

For each pair of homologous genes, protein sequences were aligned using MUSCLE (Edgar, 2004) with default parameters, and nucleotide sequences were then forced to fit the amino acid alignments using PAL2NAL (Suyama et al., 2006). Finally, K_s values were calculated using the Nei-Gojobori algorithm (Nei and Gojobori, 1986) implemented in the codeml package of PAML (Yang, 1997).

TE annotation

The repetitive sequences were identified using a combination of repeat homology searching and *ab initio* prediction approaches. For homology searching, Repbase (2018) (Bao et al., 2015) was used to search against the genome using RepeatMasker (Tarailo-Graovac and Chen, 2009) with default parameters. For *ab initio* predictions, a consensus sequence library was built using RepeatModeler (http://repeatmasker.org/RepeatModeler/) with the parameters "-engine ncbi." Then LTR_harvest (Ellinghaus et al., 2008), LTR_finder (Xu and Wang, 2007), and LTR_retriever (Ou and Jiang, 2018) were used to build an LTR library with default parameters. Both libraries were then used to annotate the genome using RepeatMasker, and the detected TEs were combined to obtain the final TE annotation. A wheat TE reference library named ClariTeRep, described previously (Daron et al., 2014), was also used to annotate the TEs of *Triticum* genomes.

Phylogenetic analysis

A phylogenetic tree was constructed for the Poaceae homologs of the *T. turgidum NAC* gene (GenBank accession No. ABI94352.1). To identify the homologs in other species, the amino acid sequences of the *T. turgidum NAC* genes were used as a query to search against the other eight species with a previously reported method (Jiao et al., 2014). Protein sequences were aligned using MUSCLE (Edgar, 2004) with default parameters. The maximum likelihood trees were then constructed using the JTT+G4 model implemented in IQ-TREE, and bootstrap supports were evaluated by ultrafast bootstrapping testing (1,000 replicates) (Nguyen et al., 2015).

Conservation of the recently duplicated gene pairs

We used both inter-genomic synteny comparisons and K_s analysis to date all of the recently duplicated gene pairs detected in the three subgenomes of CS. The inter-genome syntenic blocks were detected using MCScanX with the default parameters. Then, if a pair of duplicated genes in CS had collinear genes in the genomes of progenitors of CS or other early diverging species (e.g., *H. vulgare*), we considered that this pair of genes was duplicated before the speciation and were therefore retained and conserved duplicates. If no syntenic relationship was detected, we further dated the duplication by calculating the K_s value and comparing it with the K_s values of speciation among the Triticeae species.

GO enrichment analysis

To find the enriched GO terms in dispersed duplicates and syntenic genes, we used the R package topGO and calculated the p values of GO terms with the default method "weight01." Fisher's exact test in combination with the "classic" algorithm of this R package was used to test for overrepresented GO terms. Statistical enrichment of GO terms was evaluated by comparing the sample (duplicated genes) with the background (all annotated genes) based on Fisher's exact test, and adjusted p values ($p < 0.01$) were calculated by the Benjamini and Hochberg (false-discovery rate) method (Ashburner et al., 2000).

Gene expression analysis and co-expression module construction

RNA-seq data for 100 diverse CS samples from different tissues, growth conditions, and developmental stages were mapped to the CS genome using STAR with default parameters (Dobin et al., 2013), and RSEM was used to estimate gene expression levels (Li and Dewey, 2011). Read counts for each gene were normalized to the sequencing depth of the samples using DESeq2 with default parameters (Love et al., 2014).

All expressed genes were used to build a co-expression network with the WGCNA R package (Langfelder and Horvath, 2008). A soft power threshold of five was used because it was the lowest power for which the scale-free topology fit index reached 0.9. The blockwise module function in WGCNA was used to construct blockwise in two blocks, with a maximum block size of 46,000 genes. Other parameters for the blockwise module function were set as follows: maxPOutliers = 0.05, TOMType = "unsigned," mergeCutHeight = 0.15, and minimum module size $\geq$30. The most highly correlated genes identified by the signedKME() function were considered central to the module.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at *Plant Communications Online*.

FUNDING

We thank the National Natural Science Foundation of China (Grant number 31870209) and the Key Science and Technology Program of Henan Province (201300110800) for research funding.

AUTHOR CONTRIBUTIONS

Y.J. and J.J. initiated and conceived the study; Y.J. and X.W. performed the principal gene duplication data analyses; Y.H., X.Y., and L.Q. performed some preliminary analyses and helped with the discussion of the research and final figures; Y.J., X.W., X.Y., and Y.H. drafted the manuscript; D.W. contributed to the discussion and editing of the manuscript. All authors contributed to and approved the final manuscript.

ACKNOWLEDGMENTS

The authors declare no competing interests.

Received: May 6, 2021
Revised: November 9, 2021
Accepted: December 9, 2021
Published: December 11, 2021

REFERENCES

Adams, K.L., and Wendel, J.F. (2005). Polyploidy and genome evolution in plants. Curr. Opin. Plant Biol. **8**:135–141.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

International Wheat Genome Sequencing Consortium (WGSC); IWGSC RefSeq principal investigators, Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., IWGSC Whole-Genome Assembly Principal Investigators, and Pozniak, C.J., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science **361**:eaar7191.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**:796–815.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25**:25–29.

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S.O., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., Wiebe, K., et al. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science **357**:93–97.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. Mobile DNA **6**:11.

Bauer, E., Schmutzer, T., Barilar, I., Mascher, M., Gundlach, H., Martis, M.M., Twardziok, S.O., Hackauf, B., Gordillo, A., Wilde, P., et al. (2017). Towards a whole-genome sequence for rye (*Secale cereale* L.). Plant J. **89**:853–869.

Brunner, S., Hurni, S., Herren, G., Kalinina, O., von Burg, S., Zeller, S.L., Schmid, B., Winzeler, M., and Keller, B. (2011). Transgenic *Pm3b* wheat lines show resistance to powdery mildew in the field. Plant Biotechnol. J. **9**:897–910.

Cerbin, S., and Jiang, N. (2018). Duplication of host genes by transposable elements. Curr. Opin. Genet. Dev. **49**:63–69.

Chen, S., Zhang, W., Bolus, S., Rouse, M.N., and Dubcovsky, J. (2018). Identification and characterization of wheat stem rust resistance gene *Sr21* effective against the Ug99 race group at high temperature. PLoS Genet. **14**:4.

Daron, J., Glover, N., Pingault, L., Theil, S., Jamilloux, V., Paux, E., Barbe, V., Mangenot, S., Alberti, A., Wincker, P., et al. (2014). Organization and evolution of transposable elements along the bread wheat chromosome 3B. Genome Biol. **15**:546.

Davis, J.C., and Petrov, D.A. (2005). Do disparate mechanisms of duplication add similar genes to the genome? Trends Genet. **21**:548–551.

Deniz, Ö., Frost, J.M., and Branco, M.R. (2019). Regulation of transposable elements by DNA modifications. Nat. Rev. Genet. **20**:417–431.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29**:15–21.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**:1792–1797.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics **9**:18.

Feuillet, C., Messmer, M., Schachermayr, G., and Keller, B. (1995). Genetic and physical characterization of the *LR1* leaf rust resistance locus in wheat (*Triticum aestivum L.*). Mol. Gen. Genet. **248**:553–562.

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu. Rev. Plant Biol. **60**:433–453.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science **296**:92–100.

Guo, W., Xin, M., Wang, Z., Yao, Y., Hu, Z., Song, W., Yu, K., Chen, Y., Wang, X., Guan, P., et al. (2020). Origin and adaptation to high altitude of Tibetan semi-wild wheat. Nat. Commun. **11**:5085.

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., Forrest, K., Fritz, A., Hucl, P., Wiebe, K., et al. (2019). Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat. Genet. **51**:896–904.

Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V.S., Gundlach, H., Monat, C., Lux, T., Kamal, N., Lang, D., Himmelbach, A., et al. (2020). The barley pan-genome reveals the hidden legacy of mutation breeding. Nature **588**:284–289.

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., and Wessler, S.R. (2004). Pack-MULE transposable elements mediate gene evolution in plants. Nature **431**:569–573.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature **473**:97–100.

Jiao, Y., Li, J., Tang, H., and Paterson, A.H. (2014). Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. Plant Cell **26**:2792–2802.

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.S., et al. (2017). Improved maize reference genome with single-molecule technologies. Nature **546**:524–527.

Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. Nat. Rev. Genet. **10**:19–31.

Kim, S., Park, J., Yeom, S.I., Kim, Y.M., Seo, E., Kim, K.T., Kim, M.S., Lee, J.M., Cheong, K., Shin, H.S., et al. (2017). New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. Genome Biol. **18**:210.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics **9**:559.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics **12**:323.

Li, G., Wang, L., Yang, J., He, H., Jin, H., Li, X., Ren, T., Ren, Z., Li, F., Han, X., et al. (2021). A high-quality genome assembly highlights rye genomic characteristics and agronomically important genes. Nat. Genet. **53**:574–584.

Ling, H.Q., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., Cao, Y., Gao, Q., Zheng, S., Li, Y., et al. (2018). Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. Nature **557**:424–428.

Lisch, D. (2013). How important are transposons for plant evolution? Nat. Rev. Genet. **14**:49–61.

Liu, W., Frick, M., Huel, R., Nykiforuk, C.L., Wang, X., Gaudet, D.A., Eudes, F., Conner, R.L., Kuzyk, A., Chen, Q., et al. (2014). The stripe rust resistance gene *Yr10* encodes an evolutionary-conserved and unique CC-NBS-LRR sequence in wheat. Mol. Plant **7**:1740–1755.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**:550.

Luo, M., Gu, Y., Puiu, D., Wang, H., Twardziok, S., Deal, K., Huo, N., Zhu, T., Wang, L., Wang, Y., et al. (2017). Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature **551**:498–502.

Maccaferri, M., Harris, N.S., Twardziok, S.O., Pasam, R.K., Gundlach, H., Spannagl, M., Ormanbekova, D., Lux, T., Prade, V.M., Milner, S.G., et al. (2019). Durum wheat genome highlights past domestication signatures and future improvement targets. Nat. Genet. **51**:885–895.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. U S A **102**:5454–5459.

Marcussen, T., Sandve, S.R., Heier, L., Spannagl, M., Pfeifer, M., International Wheat Genome Sequencing Consortium., Jakobsen, K.S., Wulff, B.B., Steuernagel, B., Mayer, K.F., et al. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. Science **102**:5454–5459.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S.O., Wicker, T., Radchuk, V., Dockter, C., Hedley, P.E., Russell, J., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. Nature **544**:427–433.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., and Rafalski, A. (2005). Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat. Genet. **37**:997–1002.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. **32**:268–274.

Ohno, S. (1970). Evolution by Gene Duplication (Berlin: Springer).

Ou, S., and Jiang, N. (2018). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. **176**:1410–1422.

Panchy, N., Lehti-Shiu, M., and Shiu, S.H. (2016). Evolution of gene duplication in plants. Plant Physiol. **171**:2294–2316.

Paterson, A.H., Bowers, J.E., and Chapman, B.A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc. Natl. Acad. Sci. U S A **101**:9903–9908.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. Nature **457**:551–556.

Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. Nat. Rev. Genet. **18**:411–424.

Periyannan, S., Moore, J., Ayliffe, M., Bansal, U., Wang, X., Huang, L., Deal, K., Luo, M., Kong, X., Bariana, H., et al. (2013). The gene *Sr33*, an ortholog of barley *Mla* genes, encodes resistance to wheat stem rust race Ug99. Science **341**:786–788.

Petersen, G., Seberg, O., Yde, M., and Berthelsen, K. (2006). Phylogenetic relationships of *Triticum* and *Aegilops* and evidence for the origin of the A, B, and D genomes of common wheat (*Triticum aestivum*). Mol. Phylogenet. Evol. **39**:70–82.

Pont, C., Leroy, T., Seidel, M., Tondelli, A., Duchemin, W., Armisen, D., Lang, D., Bustos-Korts, D., Goué, N., Balfourier, F., et al. (2019). Tracing the ancestry of modern bread wheats. Nat. Genet. **51**:905–911.

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., and Paterson, A.H. (2019). Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. Genome Biol. **20**:38.

Rabanus-Wallace, M.T., Hackauf, B., Mascher, M., Lux, T., Wicker, T., Gundlach, H., Baez, M., Houben, A., Mayer, K., Guo, L., et al. (2021). Chromosome-scale genome assembly provides insights into rye biology, evolution and agronomic potential. Nat. Genet. **53**:564–573.

Ramírez-González, R.H., Borrill, P., Lang, D., Harrington, S.A., Brinton, J., Venturini, L., Davey, M., Jacobs, J., van Ex, F., Pasha, A., et al. (2018). The transcriptional landscape of polyploid wheat. Science **361**:eaar6089.

Saintenac, C., Zhang, W., Salcedo, A., Rouse, M.N., Trick, H.N., Akhunov, E., and Dubcovsky, J. (2013). Identification of wheat gene *Sr35* that confers resistance to Ug99 stem rust race group. Science **341**:783–786.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science **326**:1112–1115.

Soltis, P.S., Marchant, D.B., Van de Peer, Y., and Soltis, D.E. (2015). Polyploidy and genome evolution in plants. Curr. Opin. Genet. Dev. **35**:119–125.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. **34**:W609–W612.

Tan, S., Cardoso-Moreira, M., Shi, W., Zhang, D., Huang, J., Mao, Y., Jia, H., Zhang, Y., Chen, C., Shao, Y., et al. (2016). LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. Genome Res. **26**:1663–1675.

Tang, H., Bowers, J.E., Wang, X., and Paterson, A.H. (2010). Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc. Natl. Acad. Sci. U S A **107**:472–477.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics, Chapter 4. https://doi.org/10.1002/0471250953.bi0410s25.

Tenaillon, M.I., Hollister, J.D., and Gaut, B.S. (2010). A triptych of the evolution of plant transposable elements. Trends Plant Sci. **15**:471–478.

Uauy, C., Distelfeld, A., Fahima, T., Blechl, A., and Dubcovsky, J. (2006). A *NAC* gene regulating senescence improves grain protein, zinc, and iron content in wheat. Science **314**:1298–1301.

Vogel, J.P., Garvin, D.F., Mockler, T.C., Schmutz, J., Rokhsar, D., Bevan, M.W., Barry, K., Lucas, S., Harmon-Smith, M., Lail, K., et al. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature **463**:763–768.

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M.T., Brinton, J., Ramirez-Gonzalez, R.H., Kolodziej, M.C., Delorean, E., Thambugala, D., et al. (2020). Multiple wheat genomes reveal global variation in modern breeding. Nature **588**:277–283.

Wang, W., Zheng, H., Fan, C., Li, J., Shi, J., Cai, Z., Zhang, G., Liu, D., Zhang, J., Vang, S., et al. (2006). High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell **18**:1791–1802.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. **40**:e49.

Wang, X., Wang, J., Jin, D., Guo, H., Lee, T.H., Liu, T., and Paterson, A.H. (2015). Genome Alignment spanning major poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. Mol. Plant **8**:885–898.

Wang, H., Guo, C., Ma, H., and Qi, J. (2019). Reply to Zwaenepoel et al.: meeting the challenges of detecting polyploidy events from transcriptomic data. Mol. Plant **12**:137–140.

Wang, H., Sun, S., Ge, W., Zhao, L., Hou, B., Wang, K., Lyu, Z., Chen, L., Xu, S., Guo, J., et al. (2020). Horizontal gene transfer of *Fhb7* from fungus underlies *Fusarium* head blight resistance in wheat. Science **368**:eaba5435.

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramírez-González, R.H., De Oliveira, R., International Wheat Genome Sequencing Consortium, Mayer, K., Paux, E., et al. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol. **19**:103.

Wu, S., Han, B., and Jiao, Y. (2020). Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. Mol. Plant **13**:59–71.

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E.J., and van der Knaap, E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science **319**:1527–1530.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35**:W265–W268.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555–556.

Yu, C., Li, Y., Li, B., Liu, X., Hao, L., Chen, J., Qian, W., Li, S., Wang, G., Bai, S., et al. (2010). Molecular analysis of phosphomannomutase (*PMM*) genes reveals a unique PMM duplication event in diverse Triticeae species and the main PMM isozymes in bread wheat tissues. BMC Plant Biol. **10**:214.

Zhang, X., Li, X., Zhao, R., Zhou, Y., and Jiao, Y. (2020). Evolutionary strategies drive a balance of the interacting gene products for the *CBL* and *CIPK* gene families. New Phytol. **226**:1506–1516.

Zhao, G., Zou, C., Li, K., Wang, K., Li, T., Gao, L., Zhang, X., Wang, H., Yang, Z., Liu, X., et al. (2017). The *Aegilops tauschii* genome reveals multiple impacts of transposons. Nat. Plants **3**:946–955.

Zhou, Y., Bai, S., Li, H., Sun, G., Zhang, D., Ma, F., Zhao, X., Nie, F., Li, J., Chen, L., et al. (2021). Introgressing the *Aegilops tauschii* genome into wheat as a basis for cereal improvement. Nat. Plants **7**:774–786.

Zwaenepoel, A., Li, Z., Lohaus, R., and Van de Peer, Y. (2019). Finding evidence for whole genome duplications: a reappraisal. Mol. Plant **12**:133–136.

Plant Communications, Volume *3*

Supplemental information

A recent burst of gene duplications in Triticeae

Xiaoliang Wang, Xueqing Yan, Yiheng Hu, Liuyu Qin, Daowen Wang, Jizeng Jia, and Yuannian Jiao

Supplemental Information

A recent burst of gene duplications in Triticeae

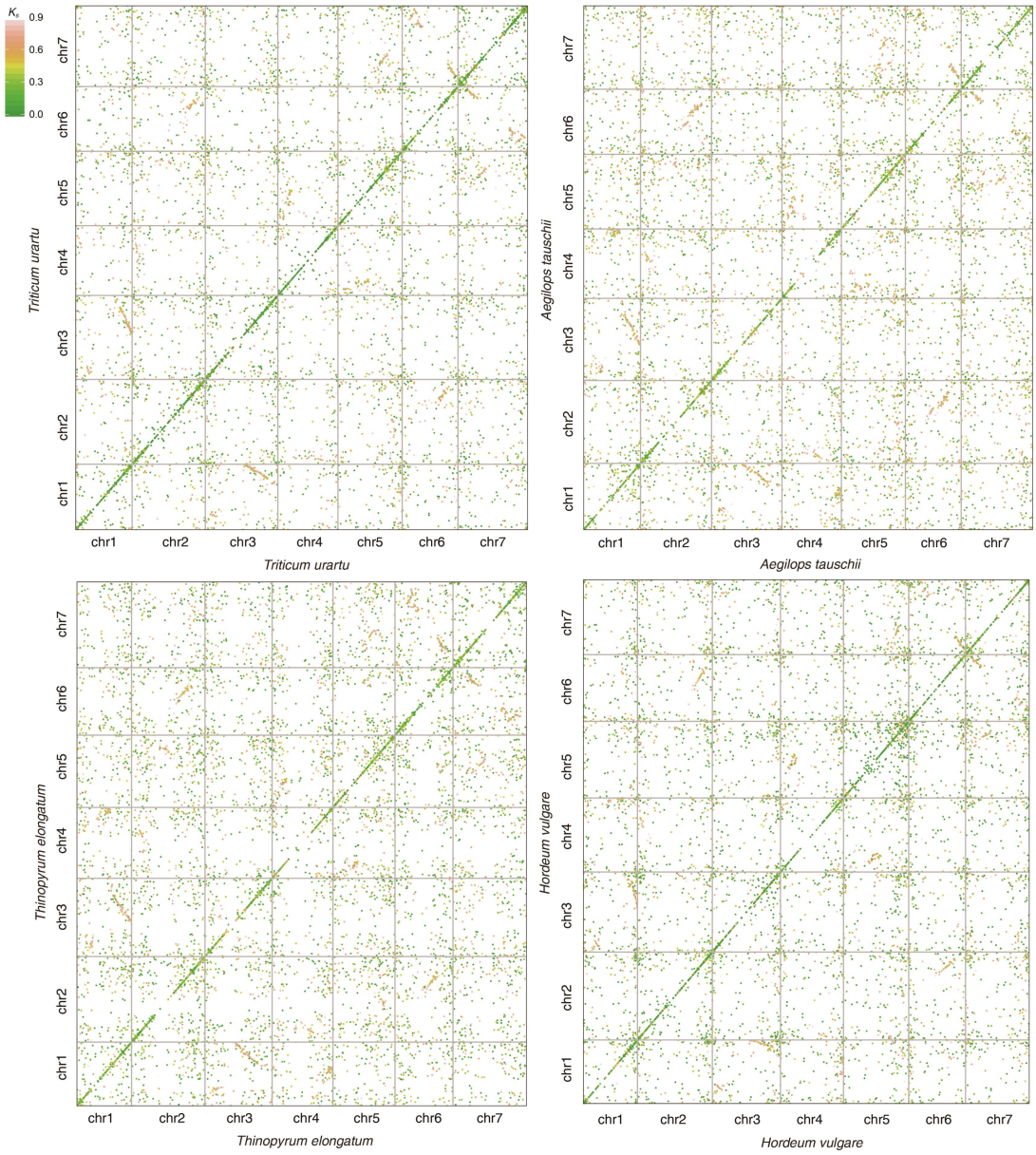
Xiaoliang Wang[1,2#], Xueqing Yan[1,2#], Yiheng Hu[1,2#], Liuyu Qin[1,2], Daowen Wang[3*], Jizeng Jia[3,4*], Yuannian Jiao[1,2*]

[1]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China.

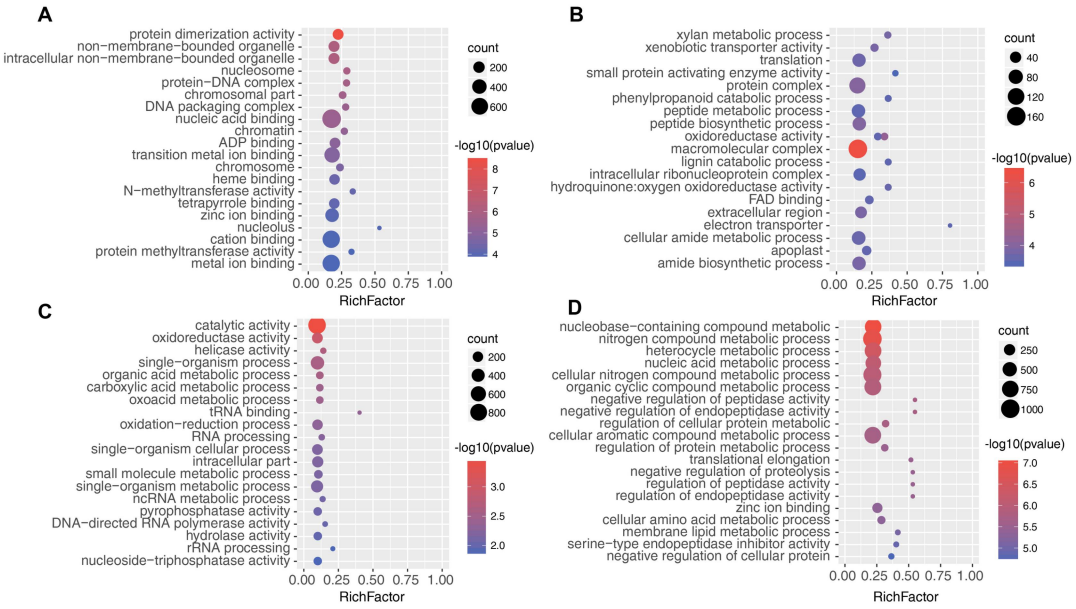
[2]University of Chinese Academy of Sciences, Beijing 100049, China.

[3]College of Agronomy, Collaborative Innovation Center of Henan Grain Crops, Henan Agricultural University, Zhengzhou, Henan 450046, China.

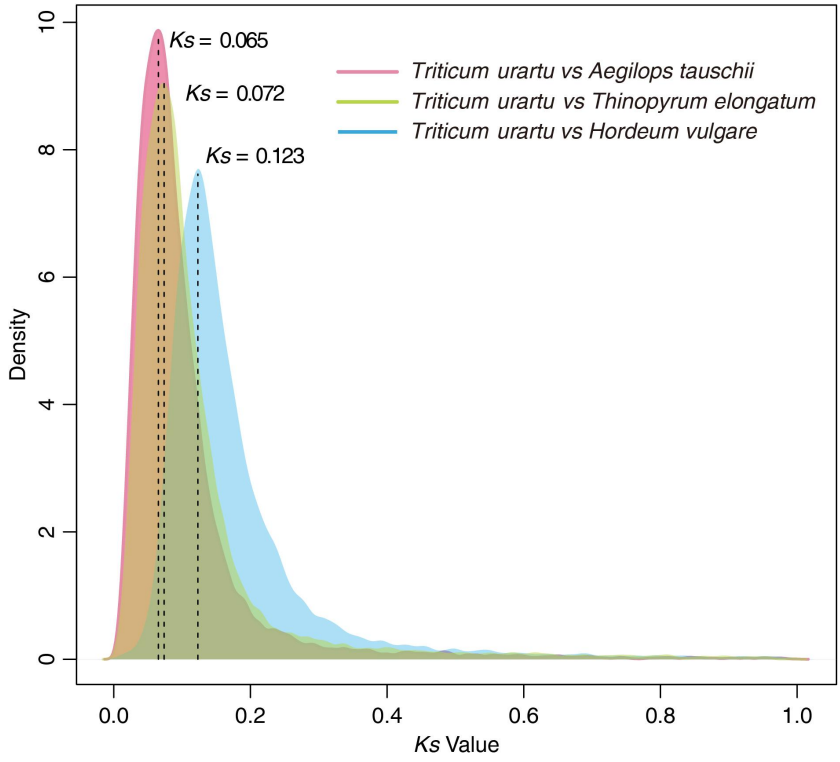
[4]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China.



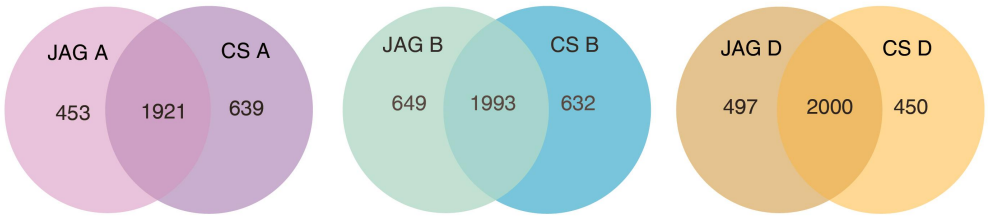
Supplementary Figure 1. Dot plots of genomic or subgenomic self-comparisons. Intra-genomic dot plots of gene pairs retrieved from all against all best reciprocal hits in *T. urartu, Ae. tauschii*, *Th. elongatum* and *H. vulgare* genome respectively. K_s value of each gene pair was shown by plotting different colors.



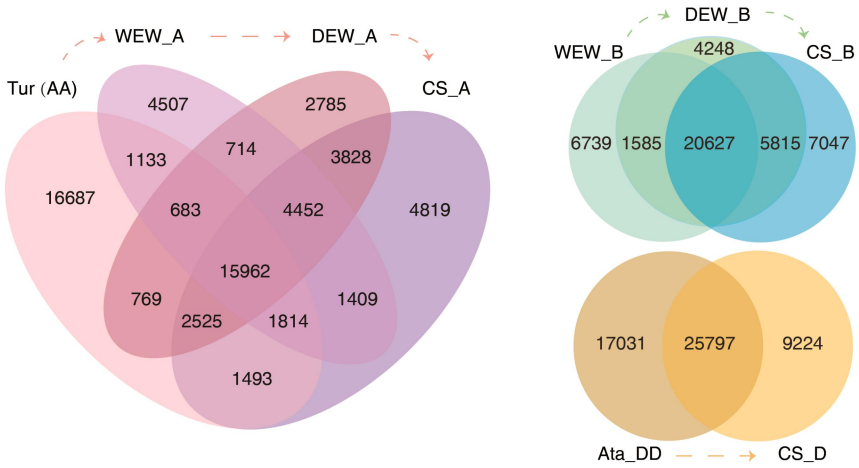
Supplementary Figure 2. Significantly enriched GO terms for the recently duplicated genes in (A) *T. urartu,* **(B)** *Ae. tauschii*, **(C)** *Th. elongatum* and **(D)** *H. vulgare* genome. The results are sorted according to significance, of which GO term of protein dimerization activity, xylan metabolic process, catalytic activity, and nucleobase-containing compound metabolic are the most significant respectively.



Supplementary Figure 3. Density plot of K_s values of the best reciprocal hits comparing *T. urartu* to *Ae. tauschii*, *Th. elongatum* and *H. vulgare* respectively. The three K_s values correspond to the differentiation of the Triticeae with the K_s around 0.123, the differentiation of *Th. elongatum* and *Triticum* with the K_s around 0.072, and the differentiation of *Triticum* with the K_s around 0.065 respectively.



Supplementary Figure 4. Venn diagrams show commonly-retained and specific-retained recent duplicates in wheat cultivars of JAG and CS.



63

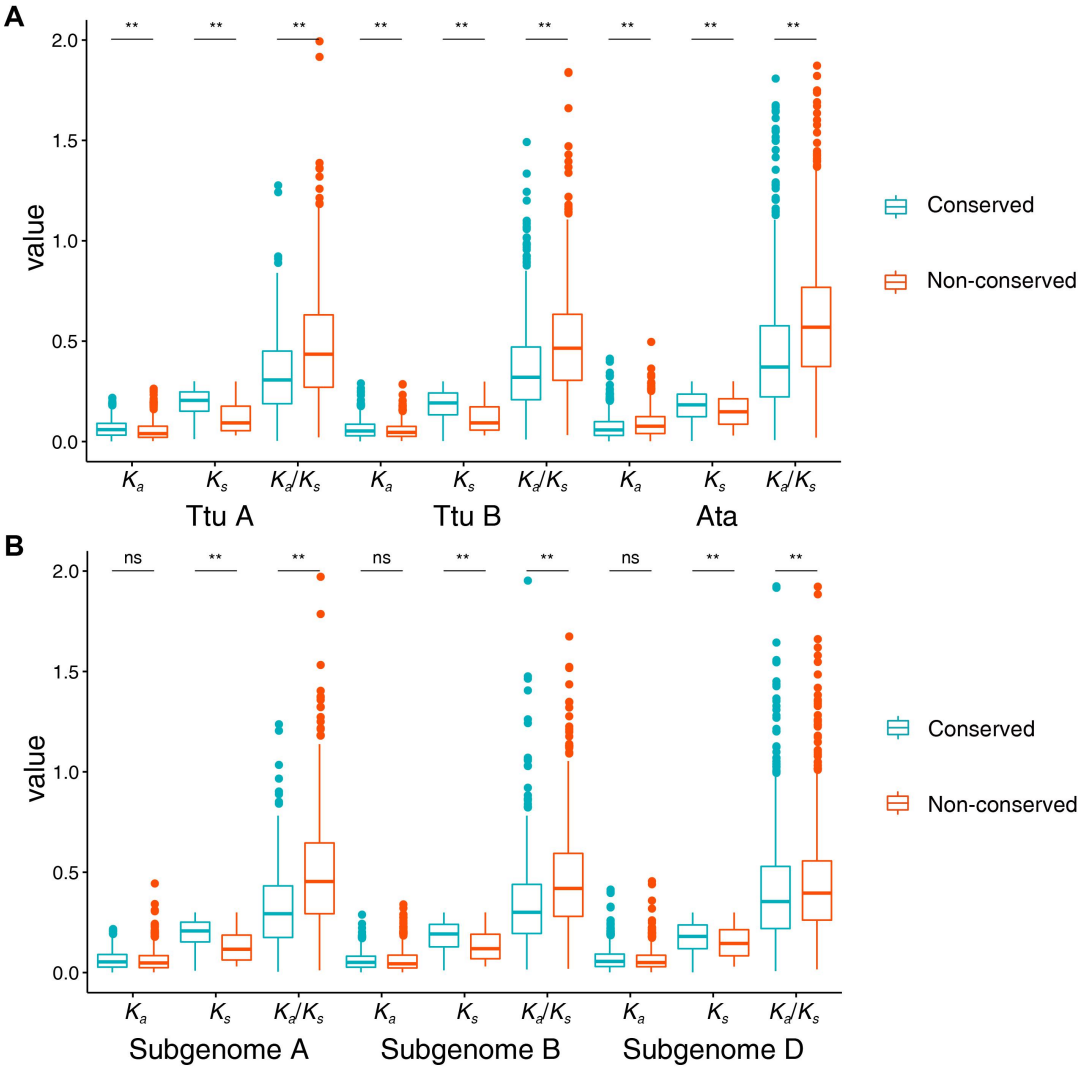
64 **Supplementary Figure 5. Venn diagram shows the numbers of orthologous in CS**

65 **subgenomes and their progenitor genomes.**

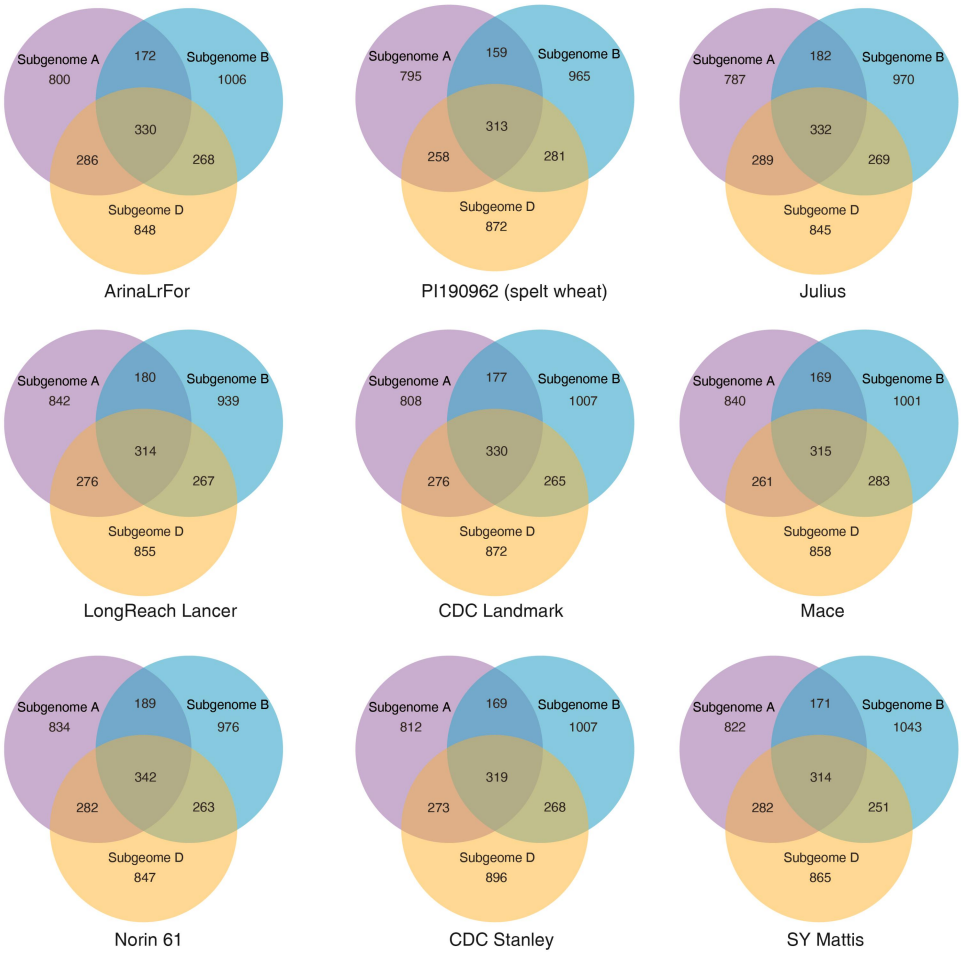
66

67

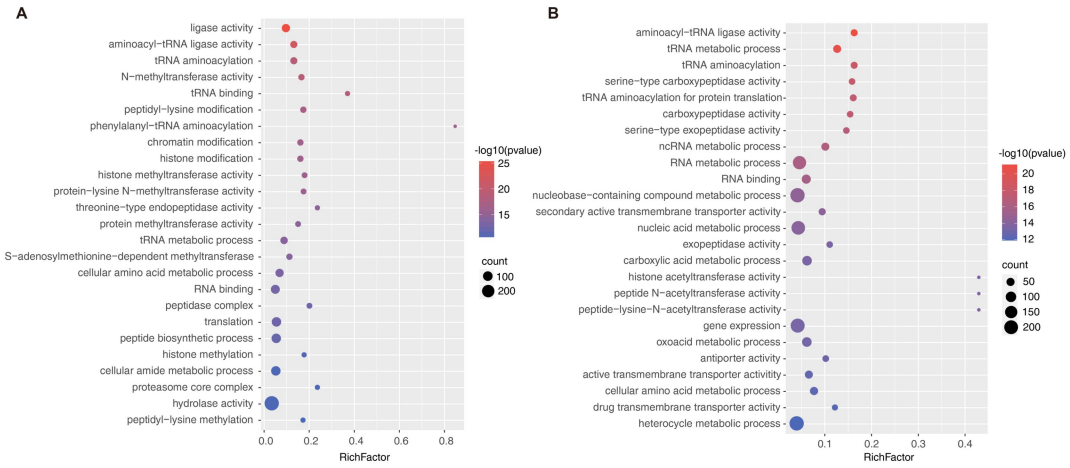
68



Supplementary Figure 6. Sequence divergence and selection analyses of the orthologous gene pairs between CS and its progenitor species. (A) The K_a, K_s, and K_a/K_s analyses of recent duplicates in progenitor genomes. Conserved gene pairs refer to gene pairs in the progenitor genomes requiring both genes have corresponding genes in CS, while non-conserved ones mean the gene pairs that have no corresponding genes in CS. **(B)** The K_a, K_s, and K_a/K_s analyses of recent duplicates in three subgenomes of CS. Conserved gene pairs refer to gene pairs that have corresponding genes with the progenitor genomes of CS, while non-conserved ones mean those that are not corresponding with the progenitor genome of CS. ns (not significant) $P > 0.05$, *$P < 0.05$, ** $P < 0.01$ in Wilcoxon test.



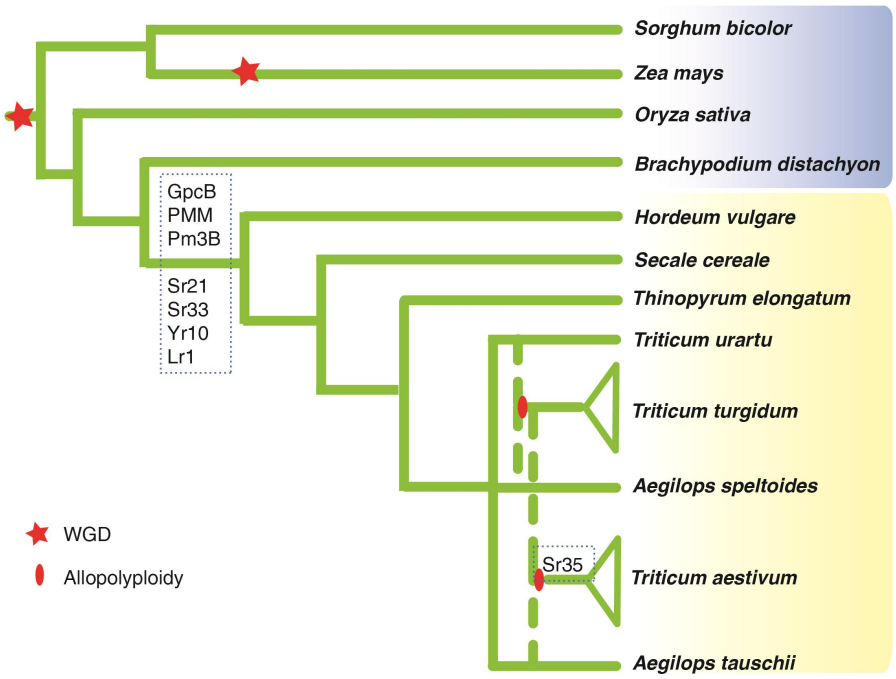
Supplementary Figure 7. Venn diagram shows the commonly retained recent gene duplicates for the three subgenomes of the nine wheat genomes.



89

90 **Supplementary Figure 8. Significantly enriched GO terms for well retained**
91 **genes in the three subgenomes of (A) CS and (B) JAG.** The results are sorted
92 according to the significance, of which GO term of ligase activity is the most
93 significant.

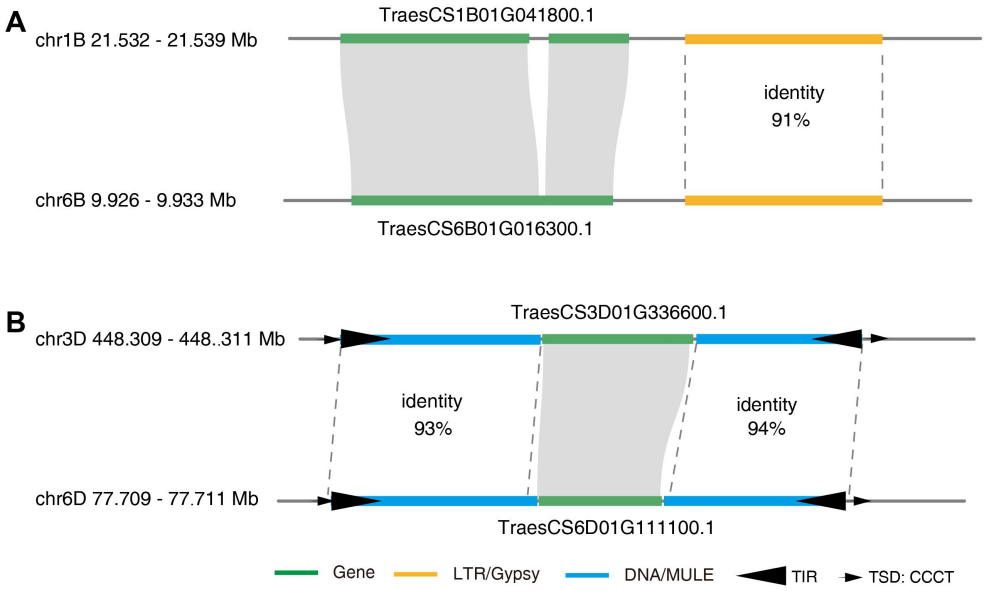
94



95

96 **Supplementary Figure 9. Duplication timing of the eight previously identified**
97 **agronomically important genes in the RBGD.**

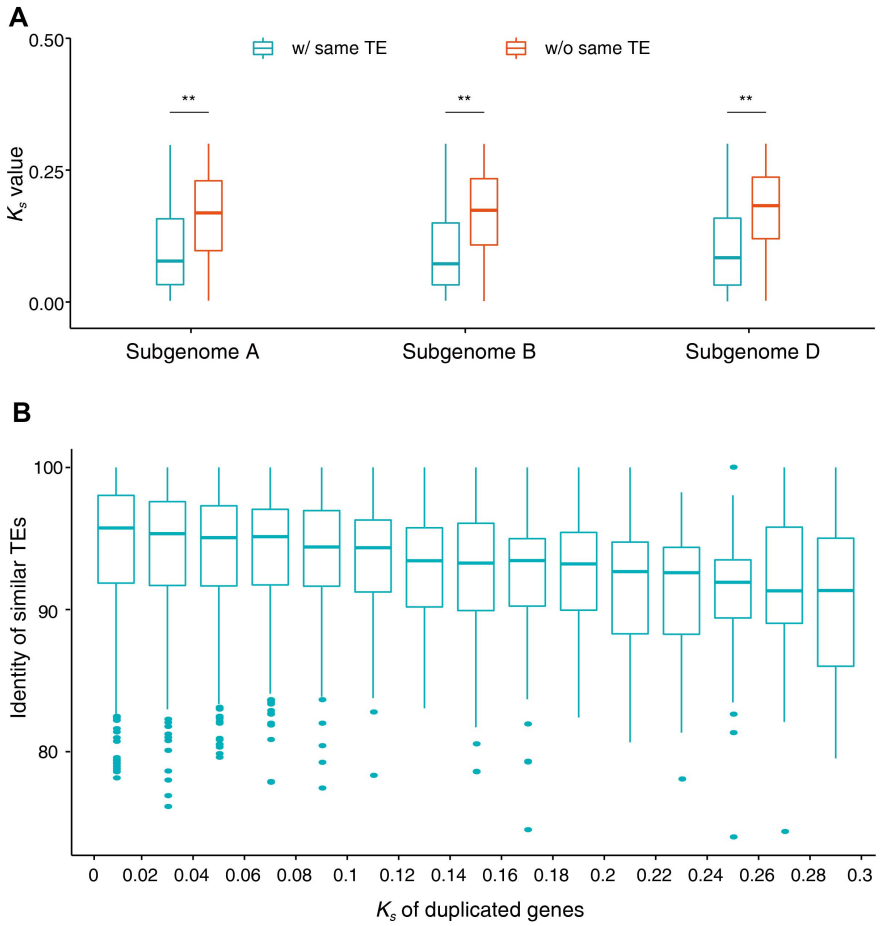
98



99

100 **Supplementary Figure 10. Two examples of gene duplication potentially derived**
101 **from TE activity in CS genome.** Identity of gene pairs is greater than 90%. **(A)** The
102 genes of *TraesCS1B01G041800.1* and *TraesCS6B01G016300.1* locate beside TEs of
103 the same subtype and with 91% sequence identity; the duplicated copy
104 (*TraesCS6B01G016300.1*) exhibits intron-less. **(B)** *TraesCS3D01G336600.1* and
105 *TraesCS6D01G111100.1* locate within TEs of the same subtype (DNA/MULE) and
106 with 94% sequence identity. TIR: terminal inverted repeat, TSD: target site
107 duplication.

108



Supplementary Figure 11. Evolutionary rate of recent duplicate gene pairs in CS. **(A)** K_s of duplicates that are either flanked or not flanked by a given TE type. ** $P <$ 0.01 in Wilcoxon test. **(B)** Sequence identity of similar TEs during the evolution process. The abscissa are K_s values in each 0.02 window of recently duplicated genes with the same subtype of TE.