

# SCIENTIFIC REPORTS



OPEN

## Dynamic Evolution of $\alpha$ -Gliadin Prolamin Gene Family in Homeologous Genomes of Hexaploid Wheat

Naxin Huo<sup>1,2</sup>, Tingting Zhu<sup>2</sup>, Susan Altenbach<sup>1</sup>, Lingli Dong<sup>3</sup>, Yi Wang<sup>1</sup>, Toni Mohr<sup>1</sup>, Zhiyong Liu<sup>3</sup>, Jan Dvorak<sup>2</sup>, Ming-Cheng Luo<sup>2</sup> & Yong Q. Gu<sup>1</sup>

Wheat *Gli-2* loci encode complex groups of  $\alpha$ -gliadin prolamins that are important for breadmaking, but also major triggers of celiac disease (CD). Elucidation of  $\alpha$ -gliadin evolution provides knowledge to produce wheat with better end-use properties and reduced immunogenic potential. The *Gli-2* loci contain a large number of tandemly duplicated genes and highly repetitive DNA, making sequence assembly of their genomic regions challenging. Here, we constructed high-quality sequences spanning the three wheat homeologous  $\alpha$ -gliadin loci by aligning PacBio-based sequence contigs with BioNano genome maps. A total of 47  $\alpha$ -gliadin genes were identified with only 26 encoding intact full-length protein products. Analyses of  $\alpha$ -gliadin loci and phylogenetic tree reconstruction indicate significant duplications of  $\alpha$ -gliadin genes in the last ~2.5 million years after the divergence of the A, B and D genomes, supporting its rapid lineage-independent expansion in different Triticeae genomes. We showed that dramatic divergence in expression of  $\alpha$ -gliadin genes could not be attributed to sequence variations in the promoter regions. The study also provided insights into the evolution of CD epitopes and identified a single indel event in the hexaploid wheat D genome that likely resulted in the generation of the highly toxic 33-mer CD epitope.

Polyploidization, an evolutionary process resulting in more than one genome per cell, has played a significant role in the evolutionary history of plants, particularly in agriculturally important crops<sup>1–4</sup>. One of the best examples is bread wheat, which is an annual species in the tribe of the grass family *Poaceae*, and one of the most widely cultivated and most important staple crops worldwide. Bread wheat is an allohexaploid species (*Triticum aestivum* L.  $2n = 6 \times = 42$ ), consisting of three sets of highly related homeologous genomes (A, B, and D), each of which contains seven pairs of chromosomes. The origin and evolution of hexaploid wheat has been well documented through extensive genetic, cytogenetic, taxonomical, and phylogenomic studies<sup>5,6</sup>. The allohexaploid wheat species is believed to have originated from two independent polyploidization events. The first event, occurring 0.36 to 0.5 million years ago, involved the hybridization of two diploid progenitors, an ancestor of *Triticum urartu* ( $2n = 2 \times = 14$ , genome AA) and an unconfirmed species (BB genome) related to *Aegilops speltoides* ( $2n = 2 \times = 14$ , genome SS), which resulted in cultivated allotetraploid emmer wheat (*T. turgidum* ssp. *dicoccum*,  $2n = 4 \times = 28$ , genomes AABB)<sup>7,8</sup>. In the second event, which occurred 8,000–10,000 years ago, an ancestor of the diploid *Aegilops tauschii* (DD genome) hybridized with the allotetraploid to form a hexaploid wheat ( $2n = 6 \times = 42$ )<sup>6</sup>. Recent phylogenomics studies using available genome sequence resources from wheat and its wild diploid ancestor species have provided new insights into the origin and relationship of the wheat A, B and D subgenomes<sup>9</sup>. It has been hypothesized that the D subgenome in hexaploid wheat resulted from a complex history of single or possibly multiple rounds of hybridization between ancestral A and B diploid species<sup>9,10</sup>.

<sup>1</sup>United States Department of Agriculture-Agricultural Research Service, Western Regional Research Center, Albany, California, 94710, USA. <sup>2</sup>Department of Plant Sciences, University of California, Davis, CA, 95616, USA. <sup>3</sup>State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, 100101, China. Naxin Huo and Tingting Zhu contributed equally to this work. Correspondence and requests for materials should be addressed to M.-C.L. (email: [mluo@ucdavis.edu](mailto:mluo@ucdavis.edu)) or Y.Q.G. (email: [yong.gu@ars.usda.gov](mailto:yong.gu@ars.usda.gov))

Because of its relatively recent speciation, wheat represents an excellent system to study evolutionary events that occur in genomes shortly after allopolyploidization. In newly formed allopolyploids following the hybridization of two divergent species, various molecular processes including genetic, epigenetic, and structural subgenome modifications must take place, presumably to couple with the “genome shock” and facilitate the coordination of functions of multiple genomes in a single cell<sup>3,5</sup>. Although the mechanisms underlying these processes are still not clear, a number of investigations have documented active occurrence of the elimination of coding and noncoding DNA sequences, differential miRNA expression, transposon activations, and gene silencing or pseudogenization in nascent allohexaploid wheats<sup>5,11</sup>. The successful speciation of an allopolyploid often provides the new species with better adaptation to a wide range of climates as the heterogeneity from subgenomes could provide genetic advantage (hybrid heterosis) compared to the parental species<sup>4,11–13</sup>.

In addition to enhanced adaptability, the addition of the D genome in allohexaploid wheat is largely responsible for the unique viscoelastic properties required for breadmaking<sup>14,15</sup>. The wheat dough properties are primarily controlled by prolamin storage proteins synthesized and accumulated in endosperm tissue during seed development. Wheat prolamins can be classified into glutenin and gliadin groups. The glutenin group contains HMW-glutenin and LMW-glutenin subunits involved in the formation of gluten polymers through intermolecular disulfide bonds, providing the elasticity to wheat dough. Gliadins, which are monomers, are viscous and give the dough its extensibility. These gliadins are subdivided into  $\alpha$ ,  $\gamma$ ,  $\delta$  and  $\omega$ -gliadins based on electrophoretic mobility<sup>16,17</sup>.

Like prolamins in other cereal crops, wheat prolamin genes belong to large multigene families. Understanding their structure, evolution, and function in relation to baking quality is challenging, but essential for improving the end-use quality of wheat flour<sup>16</sup>. In wheat, the prolamins are primarily mapped to three genomic regions. HMW-glutenin genes are located at the *Glu-1* loci on the long arms of the wheat group 1 homeologous chromosomes<sup>18</sup>, while the short arms of the same chromosomes carry the *Glu-3* loci encoding the LMW-glutenins and the *Gli-1* loci encoding  $\gamma$ -,  $\delta$ -, and  $\omega$ -gliadins<sup>19,20</sup>. Studies have shown that the *Glu-3* and *Gli-1* loci are tightly linked based on genetic and genomic sequence analyses<sup>19,21</sup>. The third genomic region located on the short arms of wheat group 6 chromosomes harbors the *Gli-2* loci encoding  $\alpha$ -gliadins. Gene family members arrayed in clusters are often prone to genetic variations in copy number, sequence polymorphism, and gene expression. Allelic variations in the composition of prolamins have been shown to be strongly correlated with differences in the breadmaking quality in different wheat varieties<sup>22,23</sup>. Both the quality (coding sequence variation) and quantity (differential gene expression) of prolamins contribute to dough properties of flour from different cultivars<sup>16</sup>.

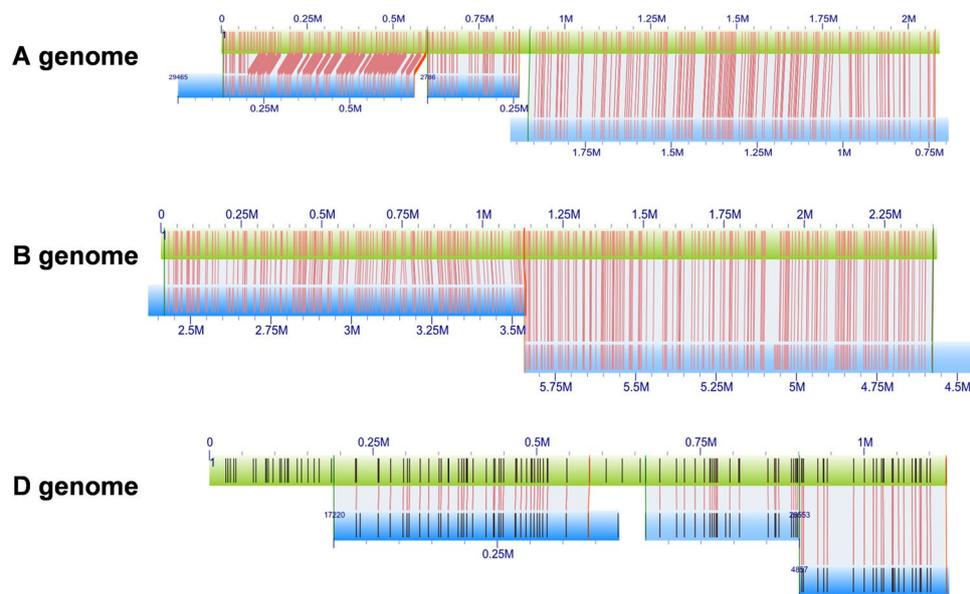
Of the three major genomic regions harboring wheat prolamin genes, the *Gli-2* loci encoding  $\alpha$ -gliadins are the youngest by evolution, as closely related species in Triticeae tribe such as barley and rye do not carry  $\alpha$ -gliadin genes<sup>24</sup>. The  $\alpha$ -gliadin gene family is also the most complex with estimates of copy numbers ranging from 25 to 150 in different wheat cultivars and ancestral species<sup>25,26</sup>. In addition,  $\alpha$ -gliadins are important as they account for 15–30% of the total seed storage proteins in the wheat grain<sup>27</sup>. Unfortunately,  $\alpha$ -gliadins are also major triggers of celiac disease (CD), a food-sensitive autoimmune disorder that impacts about 0.7–2% of the human population worldwide<sup>28</sup>. The most significant T-cell epitopes in celiac patients are PFPQPQLPY (DQ2.5-glia- $\alpha$ 1a), PYPQPQLPY (DQ2.5-glia- $\alpha$ 1b), PQPQLPY (DQ2.5-glia- $\alpha$ 2) and FRPQQPY (DQ2.5-glia- $\alpha$ 3)<sup>28</sup>. In some  $\alpha$ -gliadins, one DQ2.5-glia- $\alpha$ 1a, two DQ2.5-glia- $\alpha$ 1b and three DQ2.5-glia- $\alpha$ 2 epitopes overlap in a 33-mer peptide that is resistant to proteolytic digestion and is highly toxic in celiac patients<sup>28</sup>. A breeding effort to develop wheat with reduced levels of immunologically active epitopes, while retaining baking functionality, has been proposed<sup>29</sup>. A comprehensive understanding of  $\alpha$ -gliadin genomic regions through detailed sequence analyses could facilitate the breeding effort by unraveling their genetic diversity and developing molecular markers for selections in breeding programs.

Due to the genomic complexity of the *Gli-2* locus regions, the structure and evolution of the  $\alpha$ -gliadin gene family in the hexaploid wheat genome has not been well understood. Although several hundreds of  $\alpha$ -gliadin gene sequences from polyploid wheats and their ancestral diploid species are available in NCBI GenBank, they were primarily obtained by sequencing cDNA or PCR amplified clones, transcriptome RNA-seq analysis, or low-coverage shotgun genome sequencing<sup>30–33</sup>. However, the origin and evolutionary relationships of  $\alpha$ -gliadin family members are often difficult to draw in the absence of the context of their genomic organization. Recently, five  $\alpha$ -gliadin-containing BACs from hexaploid wheat were sequenced to reveal sequence compositions surrounding the  $\alpha$ -gliadin genes<sup>34</sup>. Since the sequences of these BACs were not contiguous and covered portions of the  $\alpha$ -gliadin locus genomic regions, comparisons among the homeologous  $\alpha$ -gliadin regions to reveal sequence changes were limited. A high-quality sequence that covers the entire *Gli-2* locus region has only been reported for *Ae. tauschii*, the progenitor of the wheat D genome<sup>24</sup>. Comparative analysis of orthologous genes from other grass genomes revealed that rapid and dynamic evolution only occurred in the *Ae. tauschii* *Gli-2* region<sup>24</sup>. In this study, we generated contiguous sequences of  $\alpha$ -gliadin genomic regions for the A, B, and D genomes of the hexaploid wheat, cv Chinese Spring using genome sequence contigs constructed from PacBio long reads<sup>35</sup>. The sequence assemblies were improved and validated by aligning them to the Chinese Spring BioNano genome maps. The high-quality sequences generated here allow a detailed comparison of the three homeologous  $\alpha$ -gliadin regions from the wheat A, B, and D subgenomes and provide insights into the structure and evolution of this important gene family.

## Results

### Construction of hexaploid wheat optical genome map and sequence assembly of $\alpha$ -gliadin regions.

Sequences of complex genomic regions like the wheat prolamin gene loci containing tandem gene copies are often difficult to resolve using short sequence reads, particularly in large and highly repetitive genomes. PacBio single molecule real-time (SMRT) technology, which generates long sequence reads (average 10 kb), has proven effective in sequencing such complex genomic regions<sup>36</sup>. The Chinese Spring (CS) hexaploid wheat genome has been



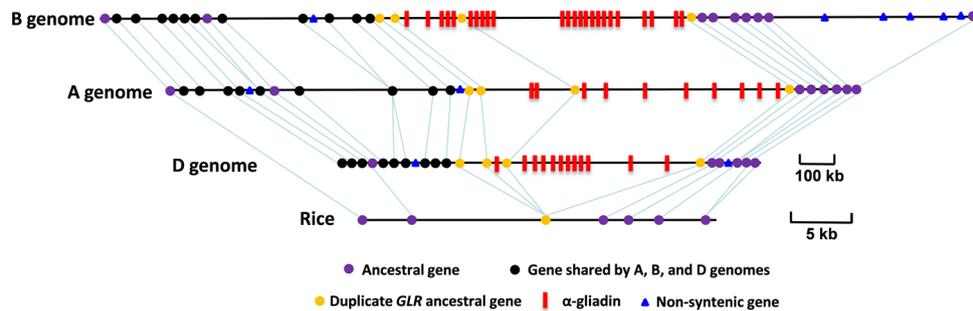
**Figure 1.** Alignment of sequence assemblies of  $\alpha$ -gliadin genomic regions with BioNano genome map contigs. PacBio contigs of hexaploid wheat cv Chinese spring were extracted using genes annotated from the *Ae. tauschii*  $\alpha$ -gliadin regions<sup>24</sup>, digested *in silico* with the restriction endonuclease, Nt.BspQ1, and aligned with the Chinese Spring BioNano genome map (blue bars). After manual editing and validation of the alignments, consensus sequences for  $\alpha$ -gliadin regions were generated for the A,B and D genomes (green bars). Vertical lines represent agreements of sequence motifs of the endonuclease recognition site (GCTTCTTC) between the consensus sequences and BioNano map contigs.

sequenced and assembled using deep sequencing coverage from a combination of short illumina reads and long PacBio reads<sup>35</sup>. The final assembly contained 15 Gb sequence with a weighted average (N50) contig size of 232 kb. However, these large contigs were not anchored to chromosomes to build pseudomolecules. To reconstruct the sequences of the wheat  $\alpha$ -gliadin locus regions, genes annotated from the *Ae. tauschii*  $\alpha$ -gliadin region<sup>24</sup> were used in a BLAST search to identify CS PacBio contigs. Since the CS genome was assembled using either Illumina-PacBio hybrid reads or PacBio only reads<sup>35</sup>, we used BLAST searches against both assemblies to identify overlapping contigs. A total of 38 contigs were identified (Table S1). Based on the initial overlapping sequence assembly and sequence mapping to the wheat chromosome-by-chromosome sequence data, 8, 18, and 12 contigs belong to the A, B, and D genomes, respectively. Contig sizes range from 12 kb to 758 kb (Table S1).

A high-resolution genome map is highly useful in validating and improving genome sequence assembly<sup>36,37</sup>. Here, we constructed a genome-wide enzyme restriction site-based optical BioNano (BNG) map for the hexaploid wheat cv Chinese Spring. A total of 2,947 Gb of data (>20 kb) was obtained from 196 runs (2,456 unique scans) in which 6,796,161 raw molecules (>180 kb) corresponding to 1,927 Gb, representing 113  $\times$  genome equivalents (~17 Gb), were used to *de novo* assemble a BNG map. These raw molecules were assembled into 11,727 BNG contigs, with N50 of 1.69 Mb and maximum BNG contig length of 10.52 Mb. The total length of the BNG map assembly is 14.2 Gb, covering 84% of the genome.

To validate the sequence assembly and reconstruct the genomic regions spanning the entire *Gli-2* loci, *in silico* Nt.BspQ1 restriction maps of the PacBio contigs were generated and used to search against the CS BNG genome map data. Eight BNG contigs were identified for reconstructing the  $\alpha$ -gliadin genomic regions by aligning the maps from BNG data with the sequence contigs, ordering and orientating contigs, resolving inconsistencies, and improving the final assembly (Fig. 1)<sup>37</sup>. We noticed that inconsistencies occasionally occurred at the end of BNG contigs, likely due to the chimeric issues of BNG contigs or inaccuracy in determining the restriction sites at the ends<sup>36</sup>. The final consensus sequences are 2,047,019 (NCBI accession number, MF434820), 2,412,936 (MF434819), and 1,107,773 bps (MF434818) for the A, B and D genome, respectively.

**Syntenic comparison of  $\alpha$ -gliadin regions from the homeologous wheat genomes.** Previously, we annotated 39 genes in the 1.7-Mb *Ae. tauschii*  $\alpha$ -gliadin region<sup>24</sup>. These genes, particularly those located at the ends, were used to define the homeologous regions of the wheat A, B, and D subgenomes. In the wheat homeologous regions, we annotated 33, 54, and 33 genes in the A, B, and D genome, respectively (Fig. 2 and Table S2). When the gene content in the wheat homeologous A, B, and D regions were compared, all the seven ancestral genes identified in the previously investigated grass genomes were also conserved (Fig. 2)<sup>24</sup>. As in the *Ae. tauschii* genome, duplication of ancestral genes were observed in the wheat genomes, resulting in an increase of the total gene number. For example, a single copy ancestral gene encoding a glutamate-like receptor (*GLR*) in rice, *Brachypodium*, and sorghum was duplicated to have four copies in all wheat genomes (Fig. 2), suggesting that these duplications occurred before separation of the A, B, and D genomes. When the three wheat genomes were compared, gene colinearity was generally well maintained. Most non-syntenic or Triticeae-specific genes



**Figure 2.** Synteny comparison of three wheat homeologous  $\alpha$ -gliadin regions and rice orthologous region. Seven ancestral genes that are shared in grass genome<sup>24</sup> are indicated as purple dots and connected by lines except the gene encoding glutamate-like receptor (yellow), which is single copy in rice, but duplicated to have four copies in the A,B and D genomes. Triticeae specific genes that are not shared with rice are shown in black dots. Non-syntenic genes that are not shared among the A,B and D genomes are represented by blue triangles.  $\alpha$ -gliadin genes are indicated with red bars. Note that the size scale bar for wheat and rice is different.

identified when compared with *Brachypodium*, rice, and sorghum were shared among the wheat genomes, indicating that their insertions into this region occurred in the Triticeae lineage. Several genes unique to one wheat genome were also observed (Fig. 2). There was one in the A, two in the D, and six in the B genome. These genes were likely inserted after the divergence of the A, B, and D genome; therefore, they are lineage-specific genes that can be used to develop markers to distinguish the  $\alpha$ -gliadin homeologous loci. In most cases, non-syntenic genes are duplicated and then translocated to different locations at a single gene level<sup>38</sup>. In the B genome, there were five consecutive non-syntenic genes in a 345-kb region (Fig. 2), suggesting a large translocation event.

**Genomic organization of  $\alpha$ -gliadin loci in wheat genomes.** We observed that  $\alpha$ -gliadin genes are highly duplicated in each genome, with 10, 24, and 11 copies in the A, B, and D genome, respectively. All the  $\alpha$ -gliadin genes in the three genomes were located between *GRL2* and *GRL4*, suggesting that  $\alpha$ -gliadin gene duplication occurred after the *GRL* duplication (Table S2). The  $\alpha$ -gliadin genes appeared to have a similar clustering pattern, with no genes intermixing with these prolamin genes except *GRL3*. The number of  $\alpha$ -gliadin gene between *GLR2* and *GLR3* varied; one copy in the D genome, two in the A, and 5 in the B genomes. We also noticed that the distance between two  $\alpha$ -gliadin genes are quite different. Most genes in the B genome cluster close to each other (~10 kb apart), while the genes in the A genome are more dispersed (~71 kb apart) (see discussion below).

To better understand the evolution of these prolamin gene clusters, we performed an in-depth sequence analysis to depict the genomic organizations of the  $\alpha$ -gliadin-containing regions (Figs S1, S2, and S3). In the 696-kb region containing the  $\alpha$ -gliadin genes from the A genome, the sequences between two  $\alpha$ -gliadin genes are mostly composed of repetitive DNA elements. In most cases, a base structure consisting of a DNA transposable element (TE) named *Manor* and two long terminal repeat (LTR) TEs named *Ada* and *Sabrina* are present (Fig. S1), suggesting the duplication of these genes occurred after the formation of this base structure between two  $\alpha$ -gliadin genes. The major differences between two  $\alpha$ -gliadin regions (or intergenic spaces) are caused by the differential insertion of repetitive DNA elements, indicating that these insertion events occurred after the sequence duplications (Fig. S1). The insertion of TEs greatly expanded the intergenic spaces in the A genome, with an average size of ~71 kb in the region from  $\alpha$ -A3 to  $\alpha$ -A10 (Fig. S1). The intergenic region between  $\alpha$ -A1 and  $\alpha$ -A2 is the smallest, about 37 kb, and contains the fewest TEs. The largest gap is between  $\alpha$ -A2 and  $\alpha$ -A3, where the presence of the *GLR3* gene and massive insertions of TEs resulted in a 147-kb separation.

In the 387-kb  $\alpha$ -gliadin region from the D genome, the base sequence structure between genes contained a *Hawi* LTR element that had been expanded by an insertion of 27-kb sequences flanked by an 8-bp direct repeat (Fig. S2). This base structure was the same as the one first identified in the *Ae. tauschii*  $\alpha$ -gliadin region<sup>24</sup>, indicating that the two D genomes from hexaploid wheat and *Ae. tauschii* share the same base structure. When the  $\alpha$ -gliadin regions from *Ae. tauschii* and wheat D genomes were compared, it was found that they shared an average of over 98% sequence identity in the alignment regions. The major differences between these two genomes were in regions where differential TE insertions occurred. For example, the sequence in the regions from  $\alpha$ -D10 and  $\alpha$ -D12 mainly consists of TEs with nested insertion structure. We identified four additional TE insertions that occurred in the region between  $\alpha$ -D10 and  $\alpha$ -D11 only in the Chinese Spring D genome. The *Ae. tauschii* D genome had two additional CACTA elements inserted into a *Sabrina* TE between  $\alpha$ -D<sup>1</sup>11 and  $\alpha$ -D<sup>1</sup>12<sup>24</sup>. Clearly, these differential TE amplifications represent sequence changes that occurred after the divergence of the two D genomes. Another important difference observed is that *Ae. tauschii* has one more  $\alpha$ -gliadin genes than Chinese Spring. Based on sequence organization, we identified that  $\alpha$ -D<sup>3</sup> in *Ae. tauschii* is absent in Chinese Spring. Interestingly, this gene is the most highly expressed  $\alpha$ -gliadin gene in *Ae. tauschii*<sup>24</sup>. At this point, it is unclear if this was a deletion event in Chinese Spring or new gene duplication in *Ae. tauschii*. When we named the  $\alpha$ -gliadin genes for the Chinese Spring D genome, we omitted the use of  $\alpha$ -D3 to describe its absence when compared to the *Ae. tauschii* region.

The 765-kb  $\alpha$ -gliadin region from the B genome contains 24 prolamin genes, more than double the number from the A or D genome (Fig. S3 and Table S2). Most regions between two  $\alpha$ -gliadin genes in the B genome do not have any TE insertion. The base structure with no TE insertion is around 10 kb in size (Fig. S3). In a few



polyQ I domain ranges from 11 to 36 amino acids with averages of 20 for the A genome, 18 for the B genome and 15 for the D genome (Table 1). In comparison, the polyQ II regions of proteins encoded by the B genome are notably larger than those of proteins from the A or D genomes. The average number of glutamines in polyQ II is 23 for B genome  $\alpha$ -gliadins, and 9 and 11 for the A and D genome proteins, respectively.

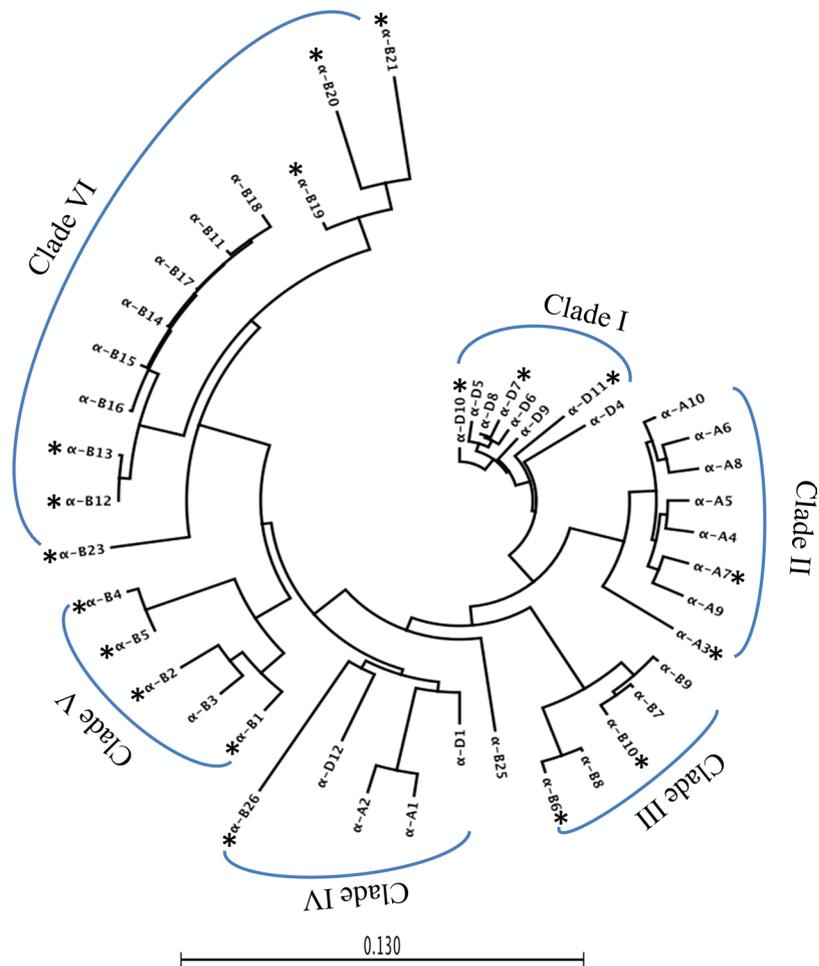
$\alpha$ -gliadin genes have a high rate of pseudogenization. Twenty one (45%) out of 47  $\alpha$ -gliadin genes were identified to be pseudogenes - two genes (20%) from the A genome, 15 (57%) from the B genome, and four (36%) from the D genome. Among the 21 pseudogenes, thirteen (59%) were caused by nucleotide substitutions, generating one or multiple premature stop codons in the coding region (Fig. S4). In most cases, stop codons were the result of C to T changes, altering CAG or CAA codons for glutamine residue into TAG or TAA stop codons. Three pseudogenes from the B genome ( $\alpha$ -B4,  $\alpha$ -B5 and  $\alpha$ -B24) and one from the D genome ( $\alpha$ -D2) contain only partial coding regions likely due to deletion events. We also found that pseudogenes could contain multiple sequence rearrangement events that can all result in pseudogenization. For example,  $\alpha$ -B19,  $\alpha$ -B20 and  $\alpha$ -B22 all contained TE insertions, premature stop codons, and frame shift indels (Figs S2 and S4). In these cases, the initial mechanisms underlying pseudogenization could not be determined.

We also compared the  $\alpha$ -gliadin pseudogenes of the D genomes from Chinese Spring and *Ae. tauschii* to assess if the mutation events in hexaploid wheat were inherited from an ancestral genome. It appears that three pseudogenes ( $\alpha$ -D2 and  $\alpha$ -D'2,  $\alpha$ -D10 and  $\alpha$ -D'10, and  $\alpha$ -D11 and  $\alpha$ -D'11) were shared.  $\alpha$ -D2 and  $\alpha$ -D'2 were both gene fragments.  $\alpha$ -D10 and  $\alpha$ -D'10 from CS all contained multiple premature stop codons and at least one or two premature stop codon mutations were shared by the homologous copy in *Ae. tauschii*. These shared mutations are likely inherited from the ancestral genome. We also identified two independent mutation events in each genome:  $\alpha$ -D'4 and  $\alpha$ -D'12 from *Ae. tauschii* and  $\alpha$ -D7 and  $\alpha$ -D9 from hexaploid D. These pseudogene sequences were all verified by RNA-seq data to ensure that they were not caused by sequencing errors (see discussion below).

**Phylogeny of  $\alpha$ -gliadin genes from the homeologous wheat genomes.** To further infer the evolutionary relationship of  $\alpha$ -gliadin genes, a phylogenetic tree was constructed using the  $\alpha$ -gliadin genes identified from the A, B, and D genomes (Fig. 3). Six major clades were observed (I to VI), with  $\alpha$ -gliadin genes from the same genome generally grouped together. Clade I has only eight  $\alpha$ -gliadin genes from the D genome, clade II contains eight  $\alpha$ -gliadin genes exclusively from the A genome, and B genome formed three separate clades (Fig. 3). The  $\alpha$ -gliadin genes located between two *GLR* genes (*GLR2* and *GLR3*) (Fig. 2) made up Clade V, suggesting duplication of these genes occurred within this defined region. Clade III consisted of  $\alpha$ -gliadin genes from  $\alpha$ -B6 to  $\alpha$ -B10. The largest clade, Clade VI, was composed of 12  $\alpha$ -gliadin genes from the B genome (Fig. 3). Only Clade IV contained  $\alpha$ -gliadin genes from all three genomes, suggesting that these genes represent ancestral forms of the  $\alpha$ -gliadin genes that existed before the separation of A, B, and D genomes. In this clade,  $\alpha$ -D1 and  $\alpha$ -D12 are the first and last  $\alpha$ -gliadin genes in the *Gli-2* locus region from the D genome.  $\alpha$ -B26 and  $\alpha$ -D12 were subgrouped together in Clade IV, supporting the notion that the last  $\alpha$ -gliadin gene in the B genome *Gli-2* locus was moved to its current position through a translocation event. In addition, we found that the first  $\alpha$ -gliadin gene from the B genome and the last  $\alpha$ -gliadin gene from the A genome were not grouped in this clade. It is possible that the corresponding genes have been deleted during evolution.

**Analysis of  $\alpha$ -gliadin gene expressions during grain development.** The expression of  $\alpha$ -gliadin genes was examined using available transcriptome data generated from different seed tissues at different seed developmental stages<sup>9</sup>. It can be challenging to estimate gene expression through sequence assembly of short reads in wheat prolamin genes with highly repetitive domains<sup>30</sup>. Here, we assessed gene expression by mapping Illumina reads to the annotated gene set. Mapping reads to well annotated  $\alpha$ -gliadin genes provides a more accurate view of gene expressions as compared to the *de novo* assembly methods<sup>24</sup>. During this process, we also checked mapped Illumina reads to validate individual  $\alpha$ -gliadin gene sequences derived from the PacBio assembly. In particular, all the mutations in the pseudogenes were confirmed. The RNA-seq transcriptome analysis result indicated that prolamin genes are highly expressed. Based on the reads mapped to the  $\alpha$ -gliadin genes against the total reads mapped to the complete annotated gene sets in wheat, we estimated that  $\alpha$ -gliadin transcripts account for 23% of the total mapped transcript reads in the seed endosperm (SE) at 20 day post-anthesis (Table S3). Even when total RNA-seq reads, instead of total reads mapped to the complete gene set, were used in the calculation, the  $\alpha$ -gliadin transcripts accounted for 13.9%. Among the prolamin genes, the level of expression varies greatly. There are six genes with FPKM values above 15000 at the peak expression, while six other genes had FPKM values under 5000. As expected, the transcript levels of prolamin pseudogenes are very low compared to most intact prolamin genes. The low expression of pseudogenes is likely correlated with the instability of their transcripts during translation<sup>39</sup>. However, we found two pseudogenes,  $\alpha$ -D10 and  $\alpha$ -B13, with FPKM values over 5000 - higher than the FPKM values of some intact prolamin genes (Fig. 4). It is unclear why these two genes have relatively high expression levels. However, it is possible that mRNAs from different pseudogenes might degrade at different rates. We also compared the expression levels of  $\alpha$ -gliadin genes from the A, B, and D genomes to assess differences among the homeologous loci. Although the B genome had the largest number of prolamin genes, it also had a high percentage of pseudogenes (54%). The B genome has 9 expressed genes with FPKM values above 5000 as compared to 6 in the D and 7 in the A genomes. This supports the idea that the B genome contributes a large portion of  $\alpha$ -gliadins to seed protein composition in Chinese Spring.

We also examined the expression of individual  $\alpha$ -gliadin genes in different tissues at different developmental stages (Fig. 4). Some  $\alpha$ -gliadin genes had strong upregulation in expression from 10 to 30 days post-anthesis, while others showed similar expression levels at different developmental stages (Fig. 4), indicating differential responses of gliadin genes during seed maturation. At 20 days post-anthesis, transcriptome data was available from the aleurone layer (AL), starchy endosperm (SE), and transfer cell (TC). Although individual  $\alpha$ -gliadin genes showed different patterns of expression, the highest overall expression of  $\alpha$ -gliadin was in starchy

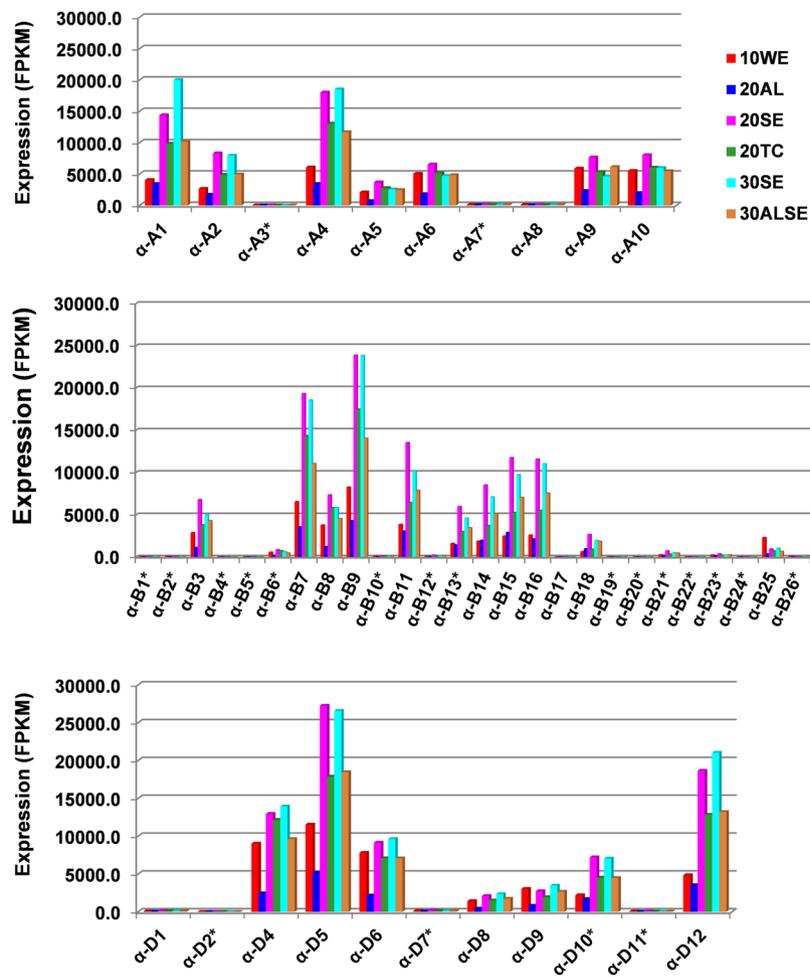


**Figure 3.** Phylogenetic analysis of  $\alpha$ -gliadin genes from hexaploid wheat cv Chinese Spring. Nucleotide sequences of  $\alpha$ -gliadin genes and pseudogenes were aligned and a phylogenetic tree was reconstructed using MEGA6 with the UPGMA methods. The tree was drawn to scale, with branch lengths measured in the number of substitution per site. TE sequences were removed from pseudogenes containing TE insertions before analysis. Pseudogenes are indicated by \*.

endosperm, where starch granules and storage proteins accumulate during seed development. The lowest expression was in the aleurone layer, which primarily accumulates lipid bodies. Transfer cells, which transport sucrose from photosynthetic tissues to the endosperm and embryo, also had significant expression of  $\alpha$ -gliadin genes. It is not clear if the FPKM values reflect true expression levels in TC, or if they may be the result from contamination with starchy endosperm during the process of manual tissue dissection.

We also compared the expression of the homologous  $\alpha$ -gliadin pairs from the D genomes from Chinese Spring and *Ae. tauschii*<sup>24</sup>. It was found that their expression can be quite different. As previously described, the corresponding gene of  $\alpha$ -D'3, representing the highest expression in *Ae. tauschii*<sup>24</sup>, was not present in the hexaploid D genome. The second highest expressed gene  $\alpha$ -D12 in Chinese Spring was a pseudogene ( $\alpha$ -D'12) in *Ae. tauschii*. The homologs  $\alpha$ -D'1 and  $\alpha$ -D1 are both intact genes. In Chinese Spring, it is highly expressed, while in *Ae. tauschii*, it has very low expression levels for an intact gene. This data suggests that the expression patterns of individual  $\alpha$ -gliadin genes have changed dramatically since the divergence of the two D genomes.

**Analysis of  $\alpha$ -gliadin CD epitopes.**  $\alpha$ -Gliadins are the major triggers of CD disease. Analysis of translated protein sequences indicated that fifteen of the  $\alpha$ -gliadins (60%) contain epitopes capable of stimulating T-cells from CD patients (Table 1)<sup>28</sup>. Proteins encoded by the D genome contain the greatest number of CD epitopes. All of the major CD epitopes are represented in this group of proteins (Table 1, Fig. 5A,B). Four proteins contain three to eight CD epitopes ( $\alpha$ -D4,  $\alpha$ -D5,  $\alpha$ -D6,  $\alpha$ -D8) including the DQ2.5-glia- $\alpha$ 1a and DQ2.5-glia- $\alpha$ 2 epitopes that have been found to be immunodominant<sup>40</sup>.  $\alpha$ -D5 also contains the 33-mer toxic peptide that is particularly immunogenic and has been shown to be resistant to proteolytic digestion (Shan *et al.*, 2002). In comparison, the protein encoded by  $\alpha$ -D1 contains only the DQ8/DQ8.5-glia- $\alpha$ 1 epitope while the protein encoded by  $\alpha$ -D12 does not contain any CD epitopes. All proteins encoded by the A genome contain either one ( $\alpha$ -A1,  $\alpha$ -A2,  $\alpha$ -A9) or two ( $\alpha$ -A4,  $\alpha$ -A5,  $\alpha$ -A6,  $\alpha$ -A8,  $\alpha$ -A10) CD epitopes. Six contain immunodominant DQ2.5-glia- $\alpha$ 1a epitopes (shown in red in Fig. 5A), five have DQ2.5-glia- $\alpha$ 3 epitopes (shown in blue in Fig. 5A) and two have DQ8/DQ8.5-glia- $\alpha$ 1 epitopes (Fig. 5B). Proteins encoded by the B genome are the least immunogenic. Of the ten



**Figure 4.** Expression of  $\alpha$ -gliadin genes at different grain development stages and tissues. Transcriptome data generated from different grain development stages and tissues were downloaded from published results<sup>55</sup>. 10WE represents whole embryo at 10 days post-anthesis, 20AL, aleurone layer at 20 days, 20SE, seed endosperm at 20 days, 20TC, transfer cell at 20 days, 30SE, seed endosperm at 30 days, and 30ALSE, aleurone plus endosperm at 30 days. Using CLC genomic workbench RNA-seq analysis toolbox, the RNA-seq reads were mapped to the annotated wheat gene set<sup>50</sup> except that the  $\alpha$ -gliadin genes were replaced with the  $\alpha$ -gliadin genes identified in this study. FPKM values were calculated using the functions in the toolbox. Pseudogenes are labeled with an \* symbol.

proteins encoded by the B genome, only two contain CD epitopes.  $\alpha$ -B3 and  $\alpha$ -B25 contain single copies of the DQ8/DQ8.5-glia- $\alpha$ 1 epitope (Table A1, Fig. 5B). Of the 25  $\alpha$ -gliadins, two from the A genome, 11 from the B genome and two from the D genome contain CSTT near the C-terminus of the protein, a motif first noted by Wang *et al.*<sup>30</sup> (Table 1). Consistent with their findings, the CSST motif is found solely in proteins without CD epitopes or those that contain only epitopes that bind the DQ8 and DQ8.5 human leukocyte antigen proteins ( $\alpha$ -A1,  $\alpha$ -A2,  $\alpha$ -D1).

One interesting finding is that the  $\alpha$ -gliadin 33-mer toxic peptide is only present in one  $\alpha$ -gliadin gene ( $\alpha$ -D5) from the wheat D genome, while the homologous copy ( $\alpha$ -D<sup>5</sup>) from the progenitor D genome of *Ae. tauschii* does not contain this peptide sequence (Fig. S5). Sequence alignment at both nucleotide and amino acid levels shows that there is a 21-nucleotide indel event that coincides with the position of the 33-mer, suggesting that a single indel event resulted in the formation of the specific 33-mer peptide (Fig. S5). Whether it is a deletion in *Ae. tauschii* or insertion in hexaploid wheat remains unclear. However, previous studies based on sequencing of multiple diploid and polyploid wheat lines indicated that the 33-mer epitope is only present in hexaploid wheat<sup>31</sup>. Therefore, we propose that the occurrence of the 33-mer epitope was caused by an insertion event that occurred in the D genome after the allohexaploidization. It is notable that  $\alpha$ -D5 is the most highly expressed  $\alpha$ -gliadin in Chinese Spring.

## Discussion

Wheat prolamins are an important nutrition source for humans worldwide and contribute unique functional properties that make it possible to produce a wide range of food products. Generating high-quality sequences in genomic regions harboring wheat prolamins loci for comparative analysis is challenging due to the polyploidy

A Genome		B	
			DQ8/DQ8.5-glia- $\alpha$ 1
$\alpha$ -A1	PQPQPPF-PQLPYPQ-----PQFFPPQQYPQPQTHLQPQQPIS	$\alpha$ -A1	<u>QGSFQPSQQ</u>
$\alpha$ -A2	PQPQPPF-PQLPYPQ-----PQFFPPQQYPQPQTHHLQPQQPIS	$\alpha$ -A2	<u>QGSFQPSQQ</u>
$\alpha$ -A4	LQLQPPFPQQLPYSQ-----PQFFRPQQYPQPQPQYSQPQHPIS	$\alpha$ -A4	<u>QGSFRPSQQ</u>
$\alpha$ -A5	LQLQPPFPQQLPYSQ-----PQFFRPQQYPQPQPQYSQLQQPIS	$\alpha$ -A5	<u>QGSFRPSQQ</u>
$\alpha$ -A6	LQLQPPFPQQLPYSQ-----PQFFRPQQYPQPQPQYSQPQQPIS	$\alpha$ -A6	<u>QGSFRPSQQ</u>
$\alpha$ -A8	LQLQPPFPQQLPYSQ-----PQFFRPQQYPQPQPQYSQPQQPIS	$\alpha$ -A8	<u>QGSFRPSQQ</u>
$\alpha$ -A9	LQLQPPFPQQLPYSQ-----PQFFPQQYPQPQPQYSQPQQPIS	$\alpha$ -A9	<u>QGSFRPSQQ</u>
$\alpha$ -A10	LQLQPPFPQQLPYSQ-----PQFFRPQQYPQPQPQYSQPQEPIS	$\alpha$ -A10	<u>QGSFRPSQQ</u>
D Genome		$\alpha$ -B3	<u>QGSFQPSQQ</u>
$\alpha$ -D1	PQPQPPF-SQLPYPQ-----PQFFPPQQYPQPQPQYPQPQQPIS	$\alpha$ -B7	<u>QVSFQPSQL</u>
$\alpha$ -D4	LQLQPPFPQQLPYPQ-----PQFFRPQQYPQPQPQYSQPQQPIS	$\alpha$ -B8	<u>QVSFQPSQL</u>
$\alpha$ -D5	LQLQPPFPQQLPYPQPQLPYPQPQPFRPQQYPQSQPQYSQPQQPIS	$\alpha$ -B9	<u>QVSFQPSQL</u>
$\alpha$ -D6	MQLQPPFPQQLPYPQPQLPYPQ-----PQFFRPQQYSQPQPQYSQPQQPIS	$\alpha$ -B11	<u>QGFQPSQQ</u>
$\alpha$ -D8	LQLQPPFPQQLPYPQPQLPYPQ-----PQFFRPQQYPQPQPQYSQPQQPIS	$\alpha$ -B14	<u>QGFQPSQQ</u>
$\alpha$ -D12	PQPQPPF-PQLPYPQ-----TQFFPPQQYPQPQPQYPQPQQPIS	$\alpha$ -B15	<u>QGFQPSQQ</u>
		$\alpha$ -B16	<u>QGFQPSQQ</u>
		$\alpha$ -B17	<u>QGFQPSQQ</u>
		$\alpha$ -B18	<u>QGFQPSQQ</u>
		$\alpha$ -B25	<u>QGSFQPSQQ</u>
		$\alpha$ -D1	<u>QGSFQPSQQ</u>
		$\alpha$ -D4	<u>QGFQPSQQ</u>
		$\alpha$ -D5	<u>QGSFQPSQQ</u>
		$\alpha$ -D6	<u>QGSFQPSQQ</u>
		$\alpha$ -D8	<u>QGSFQPSQQ</u>
		$\alpha$ -D12	<u>QGSFQPSQQ</u>

**Figure 5.** CD epitope analysis of  $\alpha$ -gliadins in hexaploid wheat cv Chinese Spring. (A) Major CD epitopes in the repetitive portion of  $\alpha$ -gliadins encoded by the A and D genomes. DQ2.5-glia- $\alpha$ 1a and DQ2.5-glia- $\alpha$ 3 epitopes are shown in red and blue, respectively. DQ2.5-glia- $\alpha$ 1b epitopes are indicated by the blue lines above sequences, DQ2.5-glia- $\alpha$ 2 epitopes are underlined in red and the 33-mer toxic peptide is underlined in black. Epitopes shown or underlined in red were found to be immunodominant by Tye-Din *et al.*<sup>40</sup>  $\alpha$ -gliadins encoded by the B genome contain none of these epitopes. (B) Sequence variation in regions of the DQ8/DQ8.5-glia- $\alpha$ 1 epitope in  $\alpha$ -gliadins encoded by the A,B and D genomes. Epitope sequence is underlined. Amino acids that differ from the canonical epitope are shown in red.

nature of the genome, high repetitive DNA content, and large and complex gene families. In this work, we used a BioNano genome map and the PacBio contigs to reconstruct three homeologous  $\alpha$ -gliadin genomic regions of the A, B, and D genomes from the hexaploid wheat cv Chinese Spring. This work represents the first in-depth view of structural organization of a full complement of  $\alpha$ -gliadin genes from the genetic background of a single hexaploid wheat species. Our detailed comparative analysis revealed dynamic evolution of the  $\alpha$ -gliadin gene family in hexaploid wheat.

**Gene collinearity in homeologous *Gli-2* regions in hexaploid wheat.** Comparative genomics studies have shown that Triticeae genome has a greater number of dispersed genes than other genomes such as rice, *Brachypodium*, and sorghum and is evolving an order of magnitude faster with regard to the structural rearrangement of its chromosomes<sup>38,41,42</sup>. It has been hypothesized that the exceptionally high amount of very similar TEs can result in frequent recombination errors and cause gene duplication and structural chromosome changes<sup>43</sup>. So far, most comparative analyses of Triticeae genomes were performed against small and compact grass genomes such as rice, *Brachypodium*, and sorghum that have diverged at least 25 million year ago. Large scale comparative analyses among Triticeae genomes could reveal recent evolution rates in Triticeae species, but only a few studies have been reported. Analyses of homeologous genomic regions of BAC insert sizes from the wheat A, B, and D genomes revealed that although gene collinearity is well maintained, the intergenic regions are completely divergent due to differential insertions and deletions of TEs<sup>18,44</sup>. Recently, the short arm sequences of *Ae. tauschii* 3D chromosome (At3DS) were compared with the homeologous regions from the Chinese Spring 3B chromosome (Ta3BS). In the syntenic regions, 2222 and 3101 genes were annotated for At3DS and Ta3BS, respectively. Collinearity analysis identified 1296 collinear gene pairs between the homeologous regions, representing 58.32% of genes in At3DS and 47.79% in Ta3BS. The low percentage of collinear genes between the two genomes is surprising, but supports the notion of the fast evolution of Triticeae genomes<sup>38,42,43</sup>. One potential problem with genome-wide syntenic analyses is that only annotated genes were used in the comparison. Gene annotation is complicated and can result in variations in gene numbers when different annotation pipelines are used because of difficulty in recognizing short protein-coding genes, distinguishing functional genes from non-functional pseudogenes, and assembling regions containing repeats and duplications<sup>45</sup>. For synteny comparisons, it is important to include both functional genes and pseudogenes to reveal genomic structure changes. In the homeologous  $\alpha$ -gliadin regions, genes were annotated using a combination of the automated TriAnnot pipeline and manual annotation, since precise prediction of the full complement of wheat prolamins is difficult because of premature stop codons, frame shift mutations, microsatellites from repeat codons for glutamine, and gene disruptions

by sequence deletions and TE insertions. Our syntenic analysis indicated that for the non-prolamins genes, in addition to the syntenic relation of ancestral genes shared among grass species, ten Triticeae-specific genes were conserved among the three homeologous genomes (Fig. 2). Non-syntenic genes for each genome were detected, but not as frequently as described in the comparison of At3DS and Ta3BS<sup>38,42</sup>. Therefore, in-depth comparative analyses will be needed to have a better understanding of genomic structural variations in the Triticeae genomes.

**Lineage-specific evolution of  $\alpha$ -gliadin genes in A, B, and D genomes.** Tandem gene duplication is one of the major gene duplication mechanisms in eukaryotes, as illustrated by the prevalence of gene family clusters. Tandem gene duplication arises through unequal crossing over, which often results from homologous recombination between paralogous sequences<sup>46</sup>. The expansion of  $\alpha$ -gliadin genes clearly represents such an example of gene duplication. Unlike copy number increases *via* whole genome duplication, tandem gene duplication occurs more frequently and can result in rapid expansion of gene family numbers.  $\alpha$ -gliadin genes are the youngest group of prolamins in wheat and its ancestral species<sup>24</sup>. They are not present in some closely related species, such as rye and barley, which separated from wheat about 7 to 11 MYA<sup>7</sup>. Comparative analysis on the syntenic regions from several grass species indicated that the  $\alpha$ -gliadin loci originated from translocation of a prolamin gene sequence located on the short arm of chromosome 1. This translocation event likely occurred in a lineage after the separation of wheat-related ancestral species from rye and barley<sup>24</sup>. The wheat A, B, and D homeologous chromosomes have an estimated divergence time of about 2.3–2.4 MYA<sup>7</sup>. Our analysis showed that despite the high collinearity of non-prolamins genes, most  $\alpha$ -gliadin genes in one genome had no obvious orthologous counterparts in the other homeologous chromosomes. This observation was primarily based on genomic structural organization of  $\alpha$ -gliadin genes in the A, B, and D genomes (Figs S1, S2, and S3). It was found that the base structure for  $\alpha$ -gliadin gene duplication was different in size and TE insertions (Figs S1, S2, and S3), suggesting duplications in expanding the gene family members were not shared events among A, B, and D genomes. The phylogenetic tree shows that most  $\alpha$ -gliadin genes from the same genome are grouped into the same clades, indicating that genes from the same chromosome are more closely related. Taken together, our results indicate that the duplication and expansion of  $\alpha$ -gliadin in the three genomes occurred independently in the last ~2.4 million years.

The phylogenetic tree also identified one clade that contained  $\alpha$ -gliadin genes from all three homeologous chromosomes (Fig. 3). These genes could represent ancestral gene forms. It appears that two copies of  $\alpha$ -gliadin genes were present before the divergence of the A, B and D genomes. Rapid amplification of lineage-specific  $\alpha$ -gliadin genes took place between the two ancestral copies. In the course of  $\alpha$ -gliadin gene evolution in different genomes, sequence rearrangements have occurred to these ancestral genes. For instance, the orthologous counterpart of  $\alpha$ -D1 in the B genome and  $\alpha$ -D12 in the A genome have been deleted. The B genome counterpart ( $\alpha$ -B26) of  $\alpha$ -D12 has been translocated to a different location. This implies dynamic changes of the ancestral copies after the formation of new prolamin copies. This observation is consistent with the studies on maize zein storage protein genes that suggested that after duplication, older copies of prolamin genes are often deleted or disrupted<sup>47</sup>. In maize, it was also found that once genes are duplicated, expression of the donor genes is reduced relative to new copies. Epigenetic regulation might contribute to the silencing of the older copies<sup>48</sup>.

**Expression regulation of tandem duplicated  $\alpha$ -gliadin genes.** One salient feature of the  $\alpha$ -gliadin family is that the expression level of each member can be quite different (Fig. 4). One reason is pseudogenization, resulting in dramatic reduction of transcripts of  $\alpha$ -gliadin pseudogenes that account for 46% of the total 47 genes identified in Chinese Spring. Many  $\alpha$ -gliadin pseudogenes have promoter sequences with the same *cis*-regulatory elements as intact genes<sup>24,34</sup>. Therefore, the initiation of transcription might start normally. The low expression level likely results from regulation at the post-translational level by a mechanism called nonsense-mediated mRNA decay (NMD). The NMD mechanism involves degradation of premature termination codon-containing transcripts to prevent production of truncated proteins that could function in dominant-negative or other deleterious mechanisms<sup>39</sup>.

Even among the full-length intact genes, differences in expression were sometimes more than 50-fold (Fig. 4). Examination of  $\alpha$ -gliadin promoter sequence regions does not provide an good explanation, as all the *cis*-regulatory elements controlling prolamin specific expression are highly conserved among the  $\alpha$ -gliadin genes<sup>24,34</sup>. Likewise, our genomic sequence analysis of  $\alpha$ -gliadin gene duplicates did not detect major sequence rearrangements in the promoter regions. Differential expression also could be the effect of subgenome dominance that causes higher expression of homeologous gene sets from the dominant subgenome in polyploid species<sup>49</sup>. However, in our transcriptome data analysis, differential expression was observed among the family members from each subgenome, and the levels of highly expressed genes in the three subgenomes were very similar (Fig. 5), suggesting that subgenome dominance doesn't play significant role in regulating  $\alpha$ -gliadin expression. Expression divergence is also a common phenomenon among the maize prolamin genes. Recently, it has been shown that epigenetic changes through DNA methylation in both promoter and coding regions contributed to expression divergence of prolamin genes<sup>48</sup>. The impact of epigenetics on wheat prolamin gene expression remains to be investigated. One interesting observation is that homologous  $\alpha$ -gliadin gene pairs from the diploid *Ae. tauschii* and Chinese Spring D genome displayed significant difference in expression. Our analysis showed that except for the deletion of  $\alpha$ -D3 in CS and a few differential nested TE insertions, the genomic structural organization of the two homologous  $\alpha$ -gliadin regions is highly conserved. We also noticed that while  $\alpha$ -D3 in CS is deleted, the homologous counterpart (*a*-D'3) has the highest expression in the *Ae. tauschii* genome<sup>24</sup>. Whether deletion of the highest expression gene could have resulted in change of DNA methylation status in the  $\alpha$ -gliadin region is not clear. Finally, the difference in expression could be caused by reprogramming of gene regulation after the integration of the A, B, and D genomes into the same nuclei during allopolyploidization<sup>3</sup>.

In conclusion, our study represents a comprehensive analysis of three homeologous  $\alpha$ -gliadin loci in a hexaploid wheat species. The rapid and dynamic duplication of  $\alpha$ -gliadin genes, as well as the expression divergence among each copy and frequent mutations resulting in pseudogenes, could provide an explanation to the great genetic diversity in this important genomic region<sup>32</sup>. The identification of a full complement of  $\alpha$ -gliadin genes from a single genetic background not only allows us to understand their evolutionary relationship but also provides a better view of CD epitope distribution and expression in wheat. For instance, we identified that a single indel event in the hexaploid wheat D genome resulted in generation of the 33-mer highly toxic CD epitope. The knowledge gained from this study will facilitate proteomic studies that provide insights into the roles of the  $\alpha$ -gliadins in flour functionality and human health as well as the development of novel strategies for breeding elite wheat varieties with improved end-use traits and reduced immunogenic potential for human consumption.

## Materials and Methods

**De novo BioNano genome map assembly and analysis.** High molecular weight (HMW) DNA was isolated from young leaves (grown in darkness) of hexaploid wheat (*Triticum aestivum* L.) genotype ‘Chinese Spring’ by Amplicon Express (Pullman, WA). The nicking endonuclease *Nt.BspQI* (New England BioLabs, Ipswich, MA) was chosen to label high-quality HMW DNA molecules at specific sequence motifs (GCTCTTC) based on sequences of the publicly available hexaploid wheat genome<sup>50</sup>. The nicked DNA molecules were then stained according to the instructions of IrysPrep Reagent Kit (BioNano Genomics, San Diego, CA). The stained DNA sample was loaded onto the nanochannel array of IrysChip (BioNano Genomics) and was automatically imaged by Irys system (BioNano Genomics). Raw DNA molecules >20 kb were collected and converted into BNX files by AutoDetect software to obtain basic labeling and DNA length information. The filtered raw DNA molecules in BNX format were aligned, clustered, and assembled into the BioNano genome (BNG) map by using the BioNano Genomics assembly pipeline as described in previous publications<sup>51,52</sup>. The *P* value thresholds used for pairwise assembly, extension/refinement, and merge stages were  $1 \times 10^{-10}$ ,  $1 \times 10^{-11}$ , and  $1 \times 10^{-15}$ , respectively. The initial BNG map was then checked for potential chimeric BNG contigs and was further refined.

**Sequence analysis and gene annotation.** To identify the homeologous  $\alpha$ -gliadin gene regions from hexaploid wheat cv Chinese Spring, genes previously annotated in the *Ae. tauschii*  $\alpha$ -gliadin region<sup>24</sup> were used for BLASTN search against the genomic sequence contigs of Chinese Spring generated using PacBio read-only assembly and hybrid assembly of BacBio and Illumina reads<sup>35</sup>. Sequence contigs with high stringent matches (*E* value less than  $1e^{-150}$ ) were downloaded. To compare the sequences with the BNG map, the extracted sequences were digested *in silico* according to the restriction site of *Nt.BspQI* by using Knickers. The alignment of sequence assemblies with the BNG map was computed with RefAligner, and the visualization of the alignment was performed with snapshot in IrysView. Software and packages used can be obtained from BioNano Genomics (<http://www.bionanogenomics.com/support/software-updates/>). Manual check and editing are involved to improve the final assembly by aligning, merging, and reorienting contigs<sup>37</sup>.

For sequence annotation, the final assembled  $\alpha$ -gliadin genomic sequences for the A, B, and D genomes were first submitted to TriAnnot pipeline for automated gene annotation<sup>53</sup>. The annotated genes were then compared with the gene contents from the *Ae. tauschii*  $\alpha$ -gliadin regions<sup>24</sup>. In addition, a homology search was performed against the NCBI nonredundant databases using BLASTN, BLASTX, and TBLASTX algorithm to verify annotated genes and identify missed genes and pseudogenes. Because gene annotation often includes transposable elements, only genes that have homology in other monocots were included. DNA repetitive elements were annotated with DNASTar MegAlign dotplot analysis and by comparison with the TREP database (<http://botserv2.uzh.ch/kelldata/trep-db/index.html>).

**Transcriptome data analysis.** A total of 176.5 Gbp Chinese Spring RNA-seq data derived from 3 time points (10, 20 and 30 days post-anthesis) and 3 main cell types/combinations (Seed endosperm, transfer cell and aleurone layer) were downloaded from NCBI (ERP004505). The Chinese Spring coding sequences (CDS) (TGAC v1.0) were downloaded from EnsemblPlants ([ftp://ftp.ensemblgenomes.org/pub/plants/release-37/fasta/triticum\\_aestivum/cds/](ftp://ftp.ensemblgenomes.org/pub/plants/release-37/fasta/triticum_aestivum/cds/)). The annotated  $\alpha$ -gliadin gene sequences along with the TGAC CDS (minus the  $\alpha$ -gliadin genes) were used as reference for RNA-Seq analysis using the CLC Genomic Workbench (v8.5) RNA-Seq Analysis Toolbox. Because of the high nucleotide similarities of  $\alpha$ -gliadin genes, stringent mapping parameters with mismatch cost 2, insertion and deletion cost 3, length fraction 0.9, similarity 0.99 were employed in read mapping. The FPKM values were also calculated using the function in the CLC Toolbox. RNA-seq alignment was manually reviewed to confirm the assembly of  $\alpha$ -gliadin gene sequences, including mutation sites causing pseudogenization.

**Phylogenetic tree analysis.** Nucleotide sequences of  $\alpha$ -gliadin gene coding regions were extracted and aligned using MUSCLE using default settings, with manual modification. Pseudogenes that were disrupted by TE insertions but contained the full-length gene sequences were included by removing the TE sequences. Phylogenetic analysis were constructed using the UPGMA method in the MEGA6 program<sup>54</sup>. The tree was drawn to scale, with branch lengths measured by the bootstrap method with 1000 replications.

**Data availability.** All the sequence data set and analysis results obtained in this work are available from the corresponding authors on reasonable request.

## References

- Soltis, D. E. & Soltis, P. S. Polyploidy: recurrent formation and genome evolution. *Trends Ecol. Evol.* **14**, 348–352 (1999).
- Wendel, J. F. Genome evolution in polyploids. *Plant Mol. Biol.* **42**, 225–249 (2000).
- Chen, Z. J. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu. Rev. Plant Biol.* **58**, 377–406 (2007).
- Alix, K., Gérard, P. R., Schwarzacher, T. & Heslop-Harrison, J. S. Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Ann. Bot.* **120**, 183–194 (2017).
- Feldman, M. & Levy, A. A. Genome evolution due to allopolyploidization in wheat. *Genetics* **192**, 763–774 (2012).
- Salamini, F., Ozkan, H., Brandolini, A., Schafer-Pregl, R. & Martin, W. Genetics and geography of wild cereal domestication in the near east. *Nat Rev Genet* **3**, 429–441 (2002).
- Huang, S. *et al.* Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proc. Natl. Acad. Sci. USA* **99**, 8133–8138 (2002).
- Dvorak, J. & Akhunov, E. D. Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the Aegilops-Triticum alliance. *Genetics* **171**, 323–332 (2005).
- Marcussen, T. *et al.* Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345** (6194) (2014).
- El Baidouri, M. *et al.* Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*). *New Phytol.* **213**, 1477–1486 (2017).
- Li, A. L., Geng, S. F., Zhang, L. Q., Liu, D. C. & Mao, L. Making the Bread: Insights from Newly Synthesized Allohexaploid Wheat. *Mol Plant* **8**, 847–859 (2015).
- Comai, L. The advantages and disadvantages of being polyploid. *Nat Rev Genet* **6**, 836–846 (2005).
- Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat Rev Genet* **18**, 411–424 (2017).
- Shewry, P. R. *et al.* The high molecular weight subunits of wheat glutenin and their role in determining wheat processing properties. *Adv. Food Nutr. Res.* **45**, 219–302 (2003).
- Hsam, S. L. K., Kieffer, R. & Zeller, F. J. Significance of *Aegilops tauschii* Glutenin Genes on Breadmaking Properties of Wheat. *Cereal Chemistry Journal* **78**, 521–525 (2001).
- Shewry, P. R., Halford, N. G., Belton, P. S. & Tatham, A. S. The structure and properties of gluten: an elastic protein from wheat grain. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **357** (2002).
- Anderson, O. D., Dong, L., Huo, N. & Gu, Y. Q. A new class of wheat gliadin genes and proteins. *PLoS one* **7**, e21139 (2012).
- Gu, Y. Q., Coleman-Derr, D., Kong, X. & Anderson, O. D. Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four triticeae genomes. *Plant Physiol.* **135**, 459–470 (2004).
- Dong, L. *et al.* Rapid evolutionary dynamics in a 2.8-Mb chromosomal region containing multiple prolamin and resistance gene families in *Aegilops tauschii*. *Plant J.* **87**, 495–506 (2016).
- D'Ovidio, R. & Masci, S. The low-molecular-weight glutenin subunits of wheat gluten. *Journal of Cereal Science* **39**, 321–339 (2004).
- Singh, N. K. & Shepherd, K. W. Linkage mapping of genes controlling endosperm storage proteins in wheat. *Theor. Appl. Genet.* **75**, 628–641 (1988).
- Bonafede, M. D., Tranquilli, G., Pflüger, L. A., Peña, R. J. & Dubcovsky, J. Effect of allelic variation at the *Glu-3/Gli-1* loci on breadmaking quality parameters in hexaploid wheat (*Triticum aestivum* L.). *Journal of cereal science* **62**, 143–150 (2015).
- Zhang, X. *et al.* Novel insights into the composition, variation, organization, and expression of the low-molecular-weight glutenin subunit gene family in common wheat. *J. Exp. Bot.* **64**, 2027–2040 (2013).
- Huo, N. *et al.* New insights into structural organization and gene duplication in a 1.75-Mb genomic region harboring the alpha-gliadin gene family in *Aegilops tauschii*, the source of wheat D genome. *Plant J.* **92**, 571–583 (2017).
- Anderson, O. D. & Greece, F. C. The a-gliadin gene family. II DNA and protein sequence variation, subfamily structure, and origin of pseudogenes. *Theor Appl Genet* **95**, 59–65 (1997).
- Harberd, N. P., Bartels, D. & Thompson, R. D. Analysis of the gliadin multiple locus on bread wheat using nullisomic-tetrasomic lines. *Mol. Gen. Genet.* **198**, 234–242 (1985).
- Altenbach, S. B. *et al.* Differential effects of a post-anthesis fertilizer regimen on the wheat flour proteome determined by quantitative 2-DE. *Proteome Science* **9**, 46 (2011).
- Sollid, L. M., Qiao, S.-M., Anderson, R. P., Gianfrani, C. & Konig, F. Nomenclature and listing of celiac disease relevant gluten T-cell epitopes restricted by HLA-DQ molecules. *Immunogenetics* **64**, 455–460 (2012).
- Shewry, P. R. & Tatham, A. S. Improving wheat to remove coeliac epitopes but retain functionality. *J Cereal Sci* **67**, 12–21 (2016).
- Wang, D. W. *et al.* Genome-wide analysis of complex wheat gliadins, the dominant carriers of celiac disease epitopes. *Sci Rep* **7**, 44609 (2017).
- Ozuna, C. V. *et al.* Diversification of the celiac disease  $\alpha$ -gliadin complex in wheat: a 33-mer peptide with six overlapping epitopes, evolved following polyploidization. *The Plant J* **82**, 794–805 (2015).
- Dubois, B., Bertin, P. & Mingeot, D. Molecular diversity of  $\alpha$ -gliadin expressed genes in genetically contrasted spelt (*Triticum aestivum* ssp. *spelta*) accessions and comparison with bread wheat (*T. aestivum* ssp. *aestivum*) and related diploid Triticum and Aegilops species. *Mol. Breed.* **36**, 152 (2016).
- Kawaura, K., Mochida, K. & Ogiwara, Y. Expression profile of two storage-protein gene families in hexaploid wheat revealed by large-scale analysis of expressed sequence tags. *Plant Physiol.* **139**, 1870–1880 (2005).
- Kawaura, K. *et al.* Genome change in wheat observed through the structure and expression of alpha/beta-gliadin genes. *Functional & integrative genomics* **12**, 341–355 (2012).
- Zimin, A. V. *et al.* The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *Gigascience* **6**, 1–7 (2017).
- Dong, J. *et al.* Analysis of tandem gene copies in maize chromosomal regions reconstructed from long sequence reads. *Proc. Natl. Acad. Sci. USA* **113**, 7949–7956 (2016).
- Hastie, A. R. *et al.* Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex *Aegilops tauschii* genome. *PLoS one* **8**, e55864 (2013).
- Glover, N. M. *et al.* Small-scale gene duplications played a major role in the recent evolution of wheat chromosome 3B. *Genome Biol* **16**, 188 (2015).
- Hug, N., Longman, D. & Cáceres, J. F. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res.* **44**, 1483–1495 (2016).
- Tye-Din, J. A. *et al.* Comprehensive, quantitative mapping of T cell epitopes in gluten in celiac disease. *Science Translational Medicine* **2**, 41ra51–41ra51 (2010).
- Luo, M. C. *et al.* A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. *Proc. Natl. Acad. Sci. USA* **110**, 7940–7945 (2013).
- Xie, J. *et al.* Sequencing and comparative analyses of *Aegilops tauschii* chromosome arm 3DS reveal rapid evolution of Triticeae genomes. *Journal of genetics and genomics* **44**, 51–61 (2017).
- Luo, M.-C., Gu, Y. Q., Puiu, D. & Hao, W. Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature* **551**, 498–502 (2017).
- Chantret, N. *et al.* Molecular basis of evolutionary events that shaped the *hardness* locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *The Plant Cell* **17**, 1033–1045 (2005).
- Mudge, J. M. & Harrow, J. The state of play in higher eukaryote gene annotation. *Nat Rev Genet* **17**, 758–772 (2016).

46. Panchy, N., Lehti-Shiu, M. & Shiu, S. H. Evolution of gene duplication in plants. *Plant Physiol.* **171**, 2294–2316 (2016).
47. Xu, J. H. & Messing, J. Organization of the prolamin gene family provides insight into the evolution of the maize genome and gene duplications in grass species. *Proc. Natl. Acad. Sci. USA* **105**, 14330–14335 (2008).
48. Miclaus, M., Xu, J. H. & Messing, J. Differential gene expression and epiregulation of alpha zein gene copies in maize haplotypes. *PLoS Genet* **7**, e1002131 (2011).
49. Woodhouse, M. R. *et al.* Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc. Natl. Acad. Sci. USA* **111**, 5283–5288 (2014).
50. Clavijo, B. J. *et al.* An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res.* **27**, 885–896 (2017).
51. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat Biotech* **30**, 771–776 (2012).
52. Cao, H. *et al.* Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *GigaScience* **3**, 34 (2014).
53. Leroy, P. *et al.* TriAnnot: A versatile and high performance pipeline for the automated annotation of plant genomes. *Front Plant Sci* **3**, 5 (2012).
54. Tamura, K., Stecher, G., Peterson, D., Filipiński, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).
55. Pfeifer, M. *et al.* Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* **345** (2014).

## Acknowledgements

This work was supported from funding from National Science Foundation grants (IOS 0822100 and IOS 1238231) and USDA-Agricultural Research Service CRIS project (5325-21000-019). Mention of trade names or commercial project is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. USDA is an equal opportunity provider and employer.

## Author Contributions

N.H. and T.Z. performed the experiment and generated data; N.H., T.Z., S.A., L.D., Y.W., T.M., Z.L., J.D., M.-C.L., and Y.Q.G. participated in the data analysis as well as in preparation of the manuscript; Y.Q.G., M.-C.L. and J.D. managed the research project; N.H., T.Z., S.A. and Y.Q.G. wrote the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-23570-5>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018