

Research article

Open Access

# The $\gamma$ -gliadin multigene family in common wheat (*Triticum aestivum*) and its closely related species

Peng-Fei Qi<sup>1,2</sup>, Yu-Ming Wei\*<sup>1</sup>, Thérèse Ouellet<sup>3</sup>, Qing Chen<sup>1</sup>, Xin Tan<sup>1</sup> and You-Liang Zheng\*<sup>2</sup>

Address: <sup>1</sup>Triticeae Research Institute, Sichuan Agricultural University, Yaan, Sichuan 625014, PR China, <sup>2</sup>Ministry of Education Key Laboratory for Crop Genetic Resources and Improvement in Southwest China, Sichuan Agricultural University, Yaan, Sichuan 625014, PR China and <sup>3</sup>Agriculture & Agri-Food Canada, Eastern Cereal and Oilseed Research Centre, Ottawa, Ontario, K1A 0C6, Canada

Email: Peng-Fei Qi - pengfeiqi@hotmail.com; Yu-Ming Wei\* - ymwei@sicau.edu.cn; Thérèse Ouellet - Therese.Ouellet@agr.gc.ca; Qing Chen - qingchen72@hotmail.com; Xin Tan - kisty560@126.com; You-Liang Zheng\* - ylzheng@sicau.edu.cn

\* Corresponding authors

Published: 21 April 2009

Received: 10 November 2008

BMC Genomics 2009, 10:168 doi:10.1186/1471-2164-10-168

Accepted: 21 April 2009

This article is available from: <http://www.biomedcentral.com/1471-2164/10/168>

© 2009 Qi et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Abstract

**Background:** The unique properties of wheat flour primarily depend on gluten, which is the most important source of protein for human being.  $\gamma$ -Gliadins have been considered to be the most ancient of the wheat gluten family. The complex family structure of  $\gamma$ -gliadins complicates the determination of their function. Moreover,  $\gamma$ -gliadins contain several sets of celiac disease epitopes. However, no systematic research has been conducted yet.

**Results:** A total of 170  $\gamma$ -gliadin genes were isolated from common wheat and its closely related species, among which 138 sequences are putatively functional. The ORF lengths of these sequences range from 678 to 1089 bp, and the repetitive region is mainly responsible for the size heterogeneity of  $\gamma$ -gliadins. The repeat motif **P(Q/L/S/T/I/V/R/A)F(S/Y/V/Q/I/C/L)P(R/L/S/T/H/C/Y)Q<sub>1-2</sub>(P(S/L/T/A/F/H)QQ)<sub>1-2</sub>** is repeated from 7 to 22 times. Sequence polymorphism and linkage disequilibrium analyses show that  $\gamma$ -gliadins are highly diverse. Phylogenic analyses indicate that there is no obvious discrimination between *Sitopsis* and *Ae. tauschii* at the *Gli-1* loci, compared with diploid wheat. According to the number and placement of cysteine residues, we defined nine cysteine patterns and 17 subgroups. Alternatively, we classified  $\gamma$ -gliadins into two types based on the length of repetitive domain. Amino acid composition analyses indicate that there is a wide range of essential amino acids in  $\gamma$ -gliadins, and those  $\gamma$ -gliadins from subgroup SG-10 and SG-12 and  $\gamma$ -gliadins with a short repetitive domain are more nutritional. A screening of toxic epitopes shows that  $\gamma$ -gliadins with a pattern of C9 and  $\gamma$ -gliadins with a short repetitive domain almost lack any epitopes.

**Conclusion:**  $\gamma$ -Gliadin sequences in wheat and closely related *Aegilops* species are diverse. Each group/subgroup contributes differently to nutritional quality and epitope content. It is suggested that the genes with a short repetitive domain are more nutritional and valuable. Therefore, it is possible to breed wheat varieties, the  $\gamma$ -gliadins of which are less, even non-toxic and more nutritional.

## Background

The unique properties of wheat flour primarily depend on seed storage proteins, which mainly consist of gluten [1]. Gluten is glutamine and proline-rich proteins, having a storage function of nitrogen and sulphur. Gluten is traditionally classified into low molecular weight (LMW) glutenins, high molecular weight (HMW) glutenins and the large gliadin group [2]. Gliadins are mainly monomeric proteins of 30–78 kD with poor solubility in dilute salt solutions, and good solubility in 70% ethanol [3]. The gliadins are composed of  $\alpha$ -,  $\gamma$ - and  $\omega$ -types [4]. HMW-glutenin genes locate at the long arms of group 1 chromosomes (*Glu-1* loci) [5]. The  $\alpha$ -gliadins are encoded by the *Gli-2* loci on the short arms of group 6 chromosomes. The  $\gamma$ -gliadins and  $\omega$ -gliadins are encoded by the *Gli-1* loci on the short arms of homeologous chromosome 1, and are tightly linked to the *Glu-3* loci coding for LMW-glutenins [6-8].

Gluten is the most important source of protein for human being, and a wide diversity of food has been developed to take advantage of the properties (i.e. mixing characteristics, dough rheology and baking performance) of wheat flour. It is reported that visco-elastic properties of wheat are affected by the proportions of gluten polymers, and that allelic variation in the composition of the HMW-glutenins is strongly correlated with differences in the bread-making quality [2,9]. Although gliadins comprise of 40–50% of total endosperm storage proteins in wheat, their roles in determining the properties of wheat flour are not well understood yet. The complex family structure of gliadins, compared to HMW-glutenin, complicates the determination of their function. The estimated copy number for  $\gamma$ -gliadin genes is between 15 and 40 in *Triticum aestivum* cv 'Chinese Spring' [10]. Classification based on their primary structure would facilitate further functional studies.  $\gamma$ -Gliadin genes have previously been divided into three hybridization classes [11]. Pistón et al. [12] classified  $\gamma$ -gliadin genes into four groups based on the phylogenetic analysis.

Celiac disease (CD), a widely prevalent autoimmune disease of the small intestine (above 1:200 in most population groups), is induced in susceptible individuals by exposure to dietary gluten [13]. However, given the enormous biological diversity and unique chemistry of gluten, and the absence of satisfactory assays for gluten toxicity, the structural basis for gluten toxicity in CD remains unclear [14]. It has been shown that some native gluten sequences can bind to HLA-DQ2/8 and induce T cell responses. In addition, modification of gluten peptides by the enzyme transglutaminase results in high affinity HLA-DQ2/8 binding peptides that can induce T cell responses [15,16]. The principal toxic components of wheat gluten

are gliadins.  $\gamma$ -Gliadins contain several sets of celiac disease epitopes [17].

The  $\gamma$ -gliadins have been considered to be the most ancient members of the wheat gluten family [18]. Sequence information for  $\gamma$ -gliadin genes in GenBank includes 34 complete/nearly complete open reading frame (ORF) and 66 partial sequences. These sequences come from various wheat and *Aegilops* species. However, no systematic research has been conducted yet.

We have performed an extensive and comparative analysis of  $\gamma$ -gliadin genes from common wheat and its closely related species in order to classify the  $\gamma$ -gliadin genes, and to investigate the diversity of the CD epitopes and nutritional quality of each  $\gamma$ -gliadin group. This study clarifies our understanding of the evolution of the multigene family.

## Results

A total of 170  $\gamma$ -gliadin genes were isolated from common wheat and its closely related species (Table 1). There is no indication of introns interrupting the coding region (Figure 1). Thirty-two of these sequences are pseudogenes, all of which contain one or more internal stop codons or frameshift mutations caused by single nucleotide indels (insertions/deletions). The remaining 138 sequences are putatively functional, with no internal stop codons.

### Sequence polymorphism

ORF lengths of these sequences range from 678 to 1089 bp. It is notable that FJ006717 amplified from *Ae. sharonensis* is the shortest  $\gamma$ -gliadin gene so far reported (678 bp), whose repetitive region has 46 amino acid residues. FJ006692 and FJ006695 isolated from *Ae. speltoides* are the longest (1089 bp), containing 179 amino acid in their repetitive regions.

ORF lengths of the sequences derived from *Ae. sharonensis* are the most variable (678–1020 bp), while the sequences from *T. monococcum* ssp. *aegilopoides* are the most conserved in length. Lengths of the repetitive region (domain II) of these sequences vary greatly as well (138–537 bp) (Table 2). The lengths of ORF and the repetitive region were analyzed for difference on average. It is shown that they have a similar changing trend. Considering the sequence alignment result, it could be concluded that the repetitive region is mainly responsible for the size heterogeneity of the  $\gamma$ -gliadins.

Table 3 shows the three estimates of genetic diversity. Haplotype diversity is very high, with a range from 0.909  $\pm$  0.079 in *Ae. longissima* to 1.000  $\pm$  0.030 in *T. turgidum*. Estimates of  $\pi$  vary from 0.01522  $\pm$  0.00981 in *T. monococcum* ssp. *monococcum* to 0.08129  $\pm$  0.00484 in *T. aestivum*.

**Table 1:  $\gamma$ -Gliadin sequences cloned from wheat and *Aegilops* species**

Species	Accession	Genome formula	Sequenced colonies	Putatively functional (Genbank No.)	Pseudogene (Genbank No.)	Total
<i>T. aestivum</i>	Chinese Spring	AABBDD	41	25 (FJ006589–FJ006613)	4 (FJ006678–FJ006681)	29
SHW	SHW-LI	AABBDD	20	10 (FJ006614–FJ006623)	2 (FJ006682–FJ006683)	12
<i>T. dicoccoides</i>	2--1	AABB	18	8 (FJ006563–FJ006570)	1 (FJ006675)	9
<i>T. dicoccoides</i>	3--1	AABB	18	6 (FJ006571–FJ006576)	1 (FJ006676)	7
<i>T. turgidum</i>	AS2255	AABB	33	12 (FJ006577–FJ006588)	1 (FJ006677)	13
<i>aegilopoides</i>	PI428001	A <sup>m</sup> A <sup>m</sup>	18	9 (FJ006624–FJ006632)	0	9
<i>monococcum</i>	PI191096	A <sup>m</sup> A <sup>m</sup>	18	1 (FJ006633)	10 (FJ006652–FJ006661)	11
<i>T. urartu</i>	PI428222	A <sup>u</sup> A <sup>u</sup>	18	7 (FJ006634–FJ006640)	7 (FJ006662–FJ006668)	14
<i>Ae. speltoides</i>	PI449338	SS	15	11 (FJ006692–FJ006702)	0	11
<i>Ae. bicornis</i>	Clae 47	S <sup>b</sup> S <sup>b</sup>	15	10 (FJ006703–FJ006712)	2 (FJ006669–FJ006670)	12
<i>Ae. longissima</i>	PI 604104	S <sup>s</sup> S <sup>s</sup>	15	11 (FJ006641–FJ006651)	1 (FJ006671)	12
<i>Ae. searsii</i>	PI 599123	S <sup>s</sup> S <sup>s</sup>	15	8 (FJ006684–FJ006691)	2 (FJ006672–FJ006673)	10
<i>Ae. sharonensis</i>	Clae 32	S <sup>sh</sup> S <sup>sh</sup>	15	9 (FJ006713–FJ006721)	1 (FJ006674)	10
<i>Ae. tauschii</i>	AS60	DD	18	11 (FJ006552–FJ006562)	0	11
<b>Total</b>			<b>277</b>	<b>138</b>	<b>32</b>	<b>170</b>

*T. aestivum* originated from the hybridization between *T. turgidum* and *Ae. tauschii*, followed by genome duplication. *T. turgidum* evolved from *T. dicoccoides* that was formed in the intercrossing between *T. urartu* and *Ae. speltoides* (closely relates to the B genome). *T. monococcum* was one representative of the A genome as well. The other four species of *Sitopsis* were used as relatives to *Ae. speltoides*; SHW = synthetic hexaploid wheat; *aegilopoides* = *T. monococcum* ssp. *aegilopoides*; *monococcum* = *T. monococcum* ssp. *monococcum*.

Estimates of  $\theta w$  range from  $0.02584 \pm 0.00336$  in *T. monococcum* ssp. *aegilopoides* to  $0.08053 \pm 0.00584$  in *T. dicoccoides*.  $\theta w$  and  $\pi$  show different trends in the rank of their values.

Table 3 also shows the result of neutrality test. Tajima's D is negative in most species except for *Ae. bicornis* and *Ae. longissima*. Three species/subspecies of diploid wheat show a significant negative Tajima's D due to an excess of variants, which indicates that they significantly departure from an equilibrium neutral model. Fu and Li' D is negative among eight species. The same three species show a significant excess of mutations.

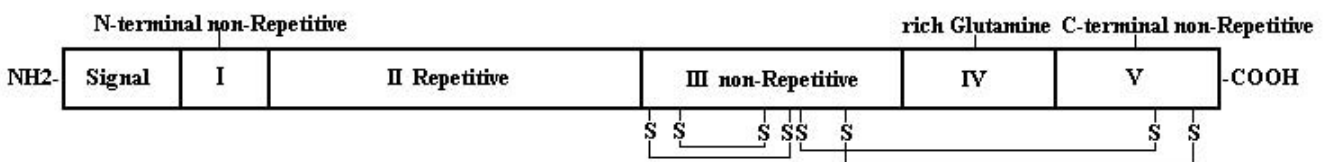
**Linkage disequilibrium (LD) and recombination**

LD between  $\gamma$ -gliadin sequences in different genomes could be estimated by ZnS, which has a range from 0 (equilibrium) to 1 (disequilibrium). The values of ZnS are

in a wide range (Table 3). The highest is 0.7418 in *T. monococcum* ssp. *aegilopoides*, and the lowest is 0.1109 in *T. aestivum*. However, there is a high level of LD in  $\gamma$ -gliadin family. A measure of the minimum number of recombination events for these data yields a range of 0–19.

**Repetitive region**

The repetitive domain is long (encompassing about 45% of total  $\gamma$ -gliadins) and consists of regular short repeats (see Additional file 1). Multiple sequence alignment indicates that the two end of this domain are conserved, and the changes mainly caused by SNP (single nucleotide polymorphism). The internal part is highly variable, mainly resulting from repeat insertions/deletions. These repeat domains have been evolving rapidly, and should not be used as a basis for determining relatedness [19]. To investigate whether the repetitive regions are related to the genomic origin of  $\gamma$ -gliadin genes, the nucleotide



**Figure 1**  
**Model structure of  $\gamma$ -gliadins.** The amino acid sequence starts with a 20-residue signal peptide, followed by a short N-terminal non-repetitive domain (I), a highly variable repetitive domain (II), a non-repetitive domain containing most of the cysteine residues (III), a glutamine-rich region (IV), and the C-terminal non-repetitive domain containing the final two conserved cysteine residues (V). All the eight conserved cysteine residues form intramolecular disulphide bonds.

**Table 2: Length polymorphisms of  $\gamma$ -gliadin sequences**

Species	Range of ORF length	Average <sup>1</sup>	SD <sup>1</sup>	Range of Domain II length	Average <sup>2</sup>	SD <sup>2</sup>
<i>T. aestivum</i>	684–984 bp (300 bp)	857.8 bp	77.4654	183–447 bp (264 bp)	348.3 bp	72.5809
SHW	732–921 bp (189 bp)	859.6 bp	48.4814	198–408 bp (210 bp)	353.6 bp	52.8423
<i>T. dicoccoides</i>	729–1047 bp (318 bp)	967.6 bp	112.382	228–504 bp (276 bp)	436.1 bp	99.6129
<i>T. turgidum</i>	786–909 bp (123 bp)	858.0 bp	24.8478	288–408 bp (120 bp)	349.8 bp	25.5305
<i>aegilopoides</i>	858–864 bp (6 bp)	858.7 bp	2.0000	366–357 bp (9 bp)	358.0 bp	3.0000
<i>monococcum</i>	857–873 bp (16 bp)	859.6 bp	4.6103	356–375 bp (19 bp)	329.0 bp	5.4855
<i>T. urartu</i>	753–903 bp (150 bp)	846.9 bp	31.8889	252–360 bp (108 bp)	343.1 bp	27.9600
<i>Ae. speltoides</i>	909–1089 bp (180 bp)	941.7 bp	72.8136	408–537 bp (129 bp)	431.5 bp	52.1831
<i>Ae. bicornis</i>	908–999 bp (91 bp)	940.4 bp	39.3041	407–456 bp (49 bp)	417.9 bp	14.8352
<i>Ae. longissima</i>	690–957 bp (267 bp)	874.3 bp	77.9943	138–408 bp (270 bp)	343.8 bp	88.5100
<i>Ae. searsii</i>	720–926 bp (206 bp)	890.6 bp	60.3401	219–425 bp (206 bp)	386.3 bp	60.8515
<i>Ae. sharonesis</i>	678–1020 bp (342 bp)	857.4 bp	131.601	138–468 bp (330 bp)	315.2 bp	132.9000
<i>Ae. tauschii</i>	903–1047 bp (144 bp)	1015.1 bp	56.4490	363–504 bp (141 bp)	476.5 bp	49.0171
Total	678–1089 bp (411 bp)			138–537 bp (399 bp)		

ORF refers to the sequences from initial codon to mature stop codon; Average<sup>1</sup> = the average length of ORF; SD<sup>1</sup> = standard deviation of ORF length; Average<sup>2</sup> = the average length of Domain II; SD<sup>2</sup> = standard deviation of the length of Domain II; SHW = synthetic hexaploid wheat; *aegilopoides* = *T. monococcum* ssp. *aegilopoides*; *monococcum* = *T. monococcum* ssp. *monococcum*.

sequences of this domain were used to carry out the phylogenetic analysis. The great majority of the sequences from diploid wheat and *Aegilops* species show a clustering according to their genomic origins (data not shown). Therefore, it would be much better if the repetitive domains were used in the relatedness analysis.

The phylogenetic analysis facilitates us to array peptide repeat pattern in parallel. The reported typical unit of  $\gamma$ -gliadins is PFPQ<sub>1-2</sub>(PQQ)<sub>1-2</sub> [19]. Our consensus repeat motif is similar to it, and modified by the substitution of several residues, i.e. P(Q/L/S/T/I/V/R/A)F(S/Y/V/Q/I/C/L)P(R/L/S/T/H/C/Y)Q<sub>1-2</sub>(P(S/L/T/A/F/H)QQ)<sub>1-2</sub>. The

**Table 3: Nucleotide diversity, neutrality test and minimum number of recombination events in wheat and *Aegilops* species**

Species	No.	S	Mut	$\pi \pm SD$	$\theta \text{ w} \pm SD$	H	Hd $\pm SD$	Tajima's D	Fu and Li's D	ZnS	Rm
<i>T. aestivum</i>	29	154	180	0.08129 $\pm$ 0.00484	0.07053 $\pm$ 0.00568***	25***	0.983 $\pm$ 0.017***	-0.05382	0.59109	0.1109***	19
SHW	12	130	139	0.06341 $\pm$ 0.01143	0.06321 $\pm$ 0.00554***	11***	0.985 $\pm$ 0.040***	-0.28955	0.4752	0.2358***	6
<i>T. dicoccoides</i>	16	190	212	0.06635 $\pm$ 0.01591	0.08053 $\pm$ 0.00584***	12***	0.917 $\pm$ 0.064***	-1.14004	-1.12294	0.2222***	14
<i>T. turgidum</i>	13	148	160	0.06497 $\pm$ 0.01063	0.06300 $\pm$ 0.00518***	13***	1.000 $\pm$ 0.030***	-0.21148	-0.0347	0.2543***	9
<i>aegilopoides</i>	9	59	60	0.01703 $\pm$ 0.00991	0.02584 $\pm$ 0.00336***	8***	0.972 $\pm$ 0.064***	-1.80203*	-2.05794*	0.7418***	0
<i>monococcum</i>	11	71	71	0.01522 $\pm$ 0.00981	0.02859 $\pm$ 0.00339***	10***	0.982 $\pm$ 0.046***	-2.22668***	-2.74095**	0.6944***	0
<i>T. urartu</i>	14	82	85	0.02016 $\pm$ 0.00659	0.03499 $\pm$ 0.00386***	13***	0.989 $\pm$ 0.031***	-1.97542*	-2.53161*	0.2756***	2
<i>Ae. speltoides</i>	11	116	117	0.03998 $\pm$ 0.01801	0.04371 $\pm$ 0.00406***	9***	0.945 $\pm$ 0.066***	-0.44816	1.13453	0.8323***	2
<i>Ae. bicornis</i>	12	141	146	0.06499 $\pm$ 0.00980	0.05318 $\pm$ 0.00448***	11***	0.985 $\pm$ 0.040***	0.84437	0.02362	0.4440***	1
<i>Ae. longissima</i>	12	99	107	0.06477 $\pm$ 0.00878	0.05331 $\pm$ 0.00536***	9***	0.909 $\pm$ 0.079***	0.57962	0.72021	0.2949***	4
<i>Ae. searsii</i>	10	116	121	0.04333 $\pm$ 0.01678	0.05719 $\pm$ 0.00531***	8***	0.933 $\pm$ 0.077***	-1.35918	-1.34402	0.4010***	2
<i>Ae. sharonesis</i>	10	105	119	0.04776 $\pm$ 0.01564	0.05836 $\pm$ 0.00570***	8***	0.933 $\pm$ 0.077***	-1.37962	-1.35868	0.3258***	5
<i>Ae. tauschii</i>	11	109	114	0.03109 $\pm$ 0.01466	0.04478 $\pm$ 0.00429***	10***	0.982 $\pm$ 0.046***	-1.61405	-1.83433	0.5936***	1

SHW = synthetic hexaploid wheat; *aegilopoides* = *T. monococcum* ssp. *aegilopoides*; *monococcum* = *T. monococcum* ssp. *monococcum*; No. = the number of sequences; S = segregating sites; Mut = total number of mutations; SD = standard deviation; H = haplotypes; Hd = haplotype diversity; \* indicates P < 0.05; \*\* indicates P < 0.01; \*\*\* indicates P < 0.001.

initial and final repeat motifs do not fit the consensus motif but are included since they appear to be related to the consensus [19]. The unit is repeated from 7 to 22 times and interspersed by additional residue(s).

#### Relatedness of $\gamma$ -gliadin genes

Firstly, the  $\gamma$ -gliadin genes cloned from *T. urartu*, *Ae. speltoides* and *Ae. tauschii* were used to carry out the phylogenetic analysis (see Additional File 2). It contains three subtrees and shows a clear clustering of these  $\gamma$ -gliadin genes according to their genomic origins. Exceptions are some genes from *Ae. speltoides* and *Ae. tauschii*. Secondly, to gain a further insight into the relationship between  $\gamma$ -gliadin genes from different genomes, we included the  $\gamma$ -gliadin sequences from the two subspecies of *T. monococcum* and other four *Sitopsis* species i.e. *Ae. sharonensis*, *Ae. bicornis*, *Ae. longissima* and *Ae. searsii* (Table 1) into the phylogenetic analysis. They cover all the diploid species that represent the ancestral genomes of *T. aestivum*. Inclusion of these sequences does not alter the fundamental structure of the phylogenetic tree, and continues to strongly support the division of the three subtrees. For a clear arrangement, we have chosen all the  $\gamma$ -gliadin genes from *Ae. speltoides* and *Ae. tauschii*, and randomly selected three sequences from each of the other diploid species to create a phylogenetic tree (Figure 2). The presence of strongly supported mixed subtrees indicates that there is no obvious discrimination between *Gli-D1* and *Gli-S1* loci, compared with *Gli-A1*. Furthermore, the sequences from the five *Sitopsis* species could not be clearly distinguished in the two subtrees.

To assess the phylogenetic relationship among the  $\gamma$ -gliadin genes from tetraploid wheat, we aligned the sequences of diploid and tetraploid species, and constructed dendrograms. The  $\gamma$ -gliadin sequences that clustered together with the sequences from diploid wheat (96% bootstrap; cluster I) should be closely related to the A genome (see Additional File 3). The sequences that fall in cluster II are supposed to locate at the *Gli-B1* locus of tetraploid wheat. It is interesting that the eleven  $\gamma$ -gliadin sequences of *T. dicoccoides* exactly clustered together with those of *Ae. tauschii* in a small cluster (98% bootstrap), which confirms the close relationship between the *Gli-S1* (*Gli-B1*) and *Gli-D1* loci. Using the characteristics of clusters described above, we could only identify the sequences relating to the A genome (cluster I; see Additional File 4) among those cloned from hexaploid wheat.

#### Cysteine and Classification

The deduced amino acid peptides of the 169 putatively functional  $\gamma$ -gliadin genes (Table 1 and Table 4) were analyzed for differences in the number and placement of cysteine residues [12,19-28]. Most of the peptides (116/169) contain eight cysteine residues that form four

intramolecular disulphide bonds [29]. The eight cysteine residues are located following a conserved pattern (Figure 1) [19], where the fourth and fifth cysteines are consecutive in the polypeptide chain. Fifty-three  $\gamma$ -gliadins show modified patterns of cysteine distribution either by adding/deleting cysteine residues or changing their relative locations (Table 5; Figure 3), among which 1, 45 and 7 peptides contain 10, 9 and 7 cysteine residues, respectively. Differences in the number of cysteine among  $\gamma$ -gliadins are likely to result from point mutations, which involve CGC-TGC, TCC-TGC, TAC-TGC, TTC-TGC, GGT-TGT and TCG-TGC. Besides the number of cysteine, the main difference among the nine patterns resides in the presence of the first cysteine residue either in the non-repetitive region or nearer along in the repetitive domain. It is notable that AJ937838 has an additional cysteine in the glutamine-rich region.

Using the number of cysteine residues as a discrimination factor, we classified the 169 putatively functional  $\gamma$ -gliadin genes into four groups, i.e. C7, C8, C9 and C10 (Table 5). Forty-five sequences of C9 share a high identity (88.93%). A comparison of the relative position of cysteine residues within each group allows to define respectively four, one, three and one cysteine distribution patterns within C7, C8, C9 and C10 (Figure 3). Those patterns with fewer than five members are directly recognized as subgroups (Table 5). Phylogenetic analyses of the deduced mature proteins of the members of pattern C9-P4 and C8-P5 indicate that C9-P4 and C8-P5 should be grouped into four and six subgroups, respectively (see Additional Files 5 and 6). Accordingly, we have totally defined 17  $\gamma$ -gliadin subgroups. The frequencies of the 17 subgroups among the wheat and Aegilops accessions (Table 1) are shown in Additional file 7. The  $\gamma$ -gliadins of each accession prefer to fall into one or two subgroups.

#### Amino acid composition

The proportion of each amino acid differs significantly in  $\gamma$ -gliadins.  $\gamma$ -Gliadins are rich in Gln (33.40%) and Pro (16.64%), followed by Leu (6.59%), and show a 20:10:3 ratio of Q:P:F for total amino acids. Essential amino acids are indispensable for good health, but cannot be synthesized in the body and must be supplied in the diet. There are eight (Trp, Lys, Met, Thr, Phe, Val, Ile and Leu) out of the 20 naturally occurring amino acids considered essential for humans. One of the problems with wheat flour is that they do not provide enough essential amino acids, which is mainly caused by the lack of essential amino acids in seed storage proteins. We analyzed the proportions of the eight essential amino acids in each of the deduced mature  $\gamma$ -gliadin peptides. A wide range of the proportion of each essential amino acid has been observed. The limiting amino acid (the essential amino acid found in the smallest quantity in the foodstuff) in



**Figure 2**  
**Neighbour-joining tree of the  $\gamma$ -gliadin sequences from diploid wheat and *Aegilops* species.** Wider lines indicate the nodes of the three subtrees in Additional File 2; 'p' = pseudogene; AF20184 is the outgroup, which is a 75 K  $\gamma$ -secaline gene.

**Table 4: Putatively functional  $\gamma$ -gliadin genes registered in public database**

Accession	Species	Cultivar	Molecular Type	Reference
<a href="#">AF120267</a>	<i>T. aestivum</i> ssp. <i>spelta</i>	Oberkulmer	Genomic DNA	[20]
<a href="#">AF144104</a>	<i>T. aestivum</i>	Forno	Genomic DNA	[20]
<a href="#">AF234643</a>	<i>T. aestivum</i>	Cheyenne	mRNA	[19]
<a href="#">AF234644</a>	<i>T. aestivum</i>	Cheyenne	mRNA	[19]
<a href="#">AF234646</a>	<i>T. aestivum</i>	Cheyenne	[15]	[19]
<a href="#">AF234647</a>	<i>T. aestivum</i>	Cheyenne	Genomic DNA	[19]
<a href="#">AF234649</a>	<i>T. aestivum</i>	Cheyenne	Genomic DNA	[19]
<a href="#">AF234650</a>	<i>T. aestivum</i>	Cheyenne	mRNA	[19]
<a href="#">AJ133613</a>	<i>T. aestivum</i>	Yamhill	Genomic DNA	[21]
<a href="#">AF175312</a>	<i>T. aestivum</i>	Yamhill	Genomic DNA	[21]
<a href="#">AJ416336</a>	<i>T. aestivum</i>	Mjoelner	Genomic DNA	[22]
<a href="#">AJ416337</a>	<i>T. aestivum</i>	Mjoelner	Genomic DNA	[22]
<a href="#">AJ416338</a>	<i>T. aestivum</i>	Mjoelner	Genomic DNA	[22]
<a href="#">AJ416339</a>	<i>T. aestivum</i>	Mjoelner	Genomic DNA	[22]
<a href="#">AJ937838</a>	<i>T. aestivum</i>	Neepawa	mRNA	
<a href="#">AY338386</a>	<i>T. aestivum</i>	Anza	mRNA	[12]
<a href="#">AY338387</a>	<i>T. aestivum</i>	Anza	mRNA	[12]
<a href="#">AY338388</a>	<i>T. aestivum</i>	Bobwhite	mRNA	[12]
<a href="#">AY338389</a>	<i>T. aestivum</i>	TB175	mRNA	[12]
<a href="#">AY338390</a>	<i>T. aestivum</i>	TB236	mRNA	[12]
<a href="#">AY338391</a>	<i>T. turgidum</i> ssp. <i>durum</i>	TD92	mRNA	[12]
<a href="#">AY338392</a>	<i>T. turgidum</i> ssp. <i>durum</i>	TD92	mRNA	[12]
<a href="#">AY338393</a>	<i>T. turgidum</i> ssp. <i>durum</i>	TD1A23	mRNA	[12]
<a href="#">AY338394</a>	<i>T. turgidum</i> ssp. <i>durum</i>	TD1A23	mRNA	[12]
<a href="#">AY338395</a>	<i>T. turgidum</i> ssp. <i>durum</i> × <i>Hordeum Chilense</i>		mRNA	
<a href="#">D78183</a>	<i>T. aestivum</i>	Hard red spring ICW	Genomic DNA	[23]
<a href="#">EF151018</a>	<i>T. aestivum</i>	Chinese Spring	mRNA	[24]
<a href="#">M13713</a>	<i>T. aestivum</i>	Yamhill	Genomic DNA	[25]
<a href="#">M16064</a>	<i>T. aestivum</i>	Yamhill	Genomic DNA	[26]
<a href="#">M36999</a>	<i>T. aestivum</i>	Yamhill	Genomic DNA	[27]
<a href="#">X77963</a>	<i>T. turgidum</i> ssp. <i>durum</i>	Lira	Genomic DNA	[28]

wheat is Lys. Its average proportion is 0.69%, with a range of 0.29–1.90%. The values are a little higher than those of Trp (average 0.65%; range 0.00–1.75%). The average proportions of Met, Thr, Phe, Val, Ile and Leu are 1.93%, 2.38%, 4.95%, 5.04%, 5.39% and 6.59%, respectively, which range from 1.10–2.88%, 0.83–4.24%, 2.38–6.01%, 3.89–7.37%, 3.51–7.56% and 5.30–9.13%, respectively.

The repetitive domain is long and variable, consisting mostly of Gln (48.10%), Pro (27.53%) and Phe (9.36%). This domain has a different Q:P:F ratio (5:3:1). To provide more information for wheat breeding, we analyzed the essential amino acid proportions of mature  $\gamma$ -gliadin peptides while ignoring the repetitive domain (The nucleotide sequences left are about 530 bp). The average proportions of six out of the eight essential amino acids significantly increase up (see Additional file 8; t test,  $P < 0.05$ ). Exceptions are Thr and Phe. Eighteen  $\gamma$ -gliadin sequences (belonging to nine subgroups) whose repetitive domains contain fewer than 85 amino acid residues (Table 6) were selected for additional comparison against the whole group. Compared with the other two sets of values (see Additional file 8), theirs are intermediate.

Compared to the essential amino acid proportions of the sequences with a long repetitive domain, the most  $\gamma$ -gliadins with fewer than 85 residues in the repetitive domain contain a higher proportion of total essential amino acids and Lys (see Additional file 9). However, it is noticeable that the corresponding proportions of some  $\gamma$ -gliadins with a long repetitive domain are high as well. We determined the corresponding subgroups of each  $\gamma$ -gliadin (either total proportion of essential amino acid  $\geq 29\%$  or the amount of Lys  $\geq 1\%$ ). Analysis indicated that these  $\gamma$ -gliadins tend to be the members of certain groups, such as SG-10 and SG-12 (Table 5).

#### **Analysis of CD toxic epitopes in $\gamma$ -gliadins**

The perfect matches in the 169 putatively functional genes to the three CD-toxic epitopes of  $\gamma$ -gliadins [17] are shown in Table 7. Each epitope appears at most once in every sequence, and all of them located at the repetitive domain. The result indicates that those subgroups or species contribute differently to the epitope content. Firstly, more  $\gamma$ -gliadins from *T. aestivum*, *Ae. longissima* and *Ae. bicornis* contain CD toxic peptides. Secondly, 23 sequences altogether contain these epitopes. The epitopes are present in seven subgroups (SG-7 (ratio = 2/6); SG-9

**Table 5: Classification of  $\gamma$ -gliadins**

Group	Pattern	Subgroup	GenBank No.						
C10	C10-P1	SG-1	<u>AJ937838</u> <sup>a,T,L</sup>						
C9	C9-P2	SG-2	<u>AF234646</u>						
		SG-3	<u>FJ006638</u>						
	C9-P4	SG-4	<u>FJ006561</u>	<u>FJ006573</u> <sup>T</sup>	<u>FJ006605</u>	<u>FJ006622</u>	<u>FJ006693</u>	<u>FJ006694</u>	
			<u>FJ006696</u>	<u>FJ006697</u>	<u>FJ006698</u>	<u>FJ006699</u>	<u>FJ006700</u>	<u>FJ006701</u>	
			<u>FJ006702</u>						
		SG-5	<u>FJ006641</u>	<u>FJ006643</u>	<u>FJ006650</u>	<u>FJ006704</u>	<u>FJ006706</u>	<u>FJ006710</u>	
			<u>FJ006711</u>	<u>FJ006712</u>	<u>FJ006721</u>				
		SG-6	<u>AF234647</u>	<u>FJ006576</u> <sup>T</sup>	<u>FJ006577</u>	<u>FJ006590</u>	<u>FJ006600</u>	<u>FJ006612</u> <sup>T</sup>	
			<u>FJ006685</u>	<u>FJ006686</u>	<u>FJ006688</u>	<u>FJ006689</u>	<u>FJ006690</u>	<u>FJ006691</u>	
			<u>FJ006592</u>	<u>M36999</u>	<u>X77963</u>				
	C8	C8-P5	SG-7	<u>FJ006596</u>	<u>FJ006602</u>	<u>FJ006604</u> <sup>T,L</sup>	<u>FJ006617</u> <sup>L</sup>	<u>FJ006618</u>	<u>FJ006621</u>
			SG-8	<u>AF175312</u>	<u>AF234650</u>	<u>AY338388</u>	<u>FJ006552</u>	<u>FJ006553</u>	<u>FJ006554</u>
				<u>FJ006555</u>	<u>FJ006556</u>	<u>FJ006557</u>	<u>FJ006558</u>	<u>FJ006559</u>	<u>FJ006560</u>
<u>FJ006562</u>				<u>FJ006563</u>	<u>FJ006564</u> <sup>T</sup>	<u>FJ006565</u>	<u>FJ006566</u>	<u>FJ006568</u>	
<u>FJ006569</u>				<u>FJ006570</u>	<u>FJ006571</u>	<u>FJ006572</u>	<u>FJ006574</u>	<u>FJ006575</u>	
			<u>FJ006595</u>	<u>FJ006603</u>	<u>FJ006692</u>	<u>FJ006695</u>	<u>FJ006707</u>	<u>FJ006713</u>	
			SG-9	<u>AF120267</u> <sup>L</sup>	<u>AF144104</u> <sup>L</sup>	<u>AY338389</u> <sup>T,L</sup>	<u>EF151018</u> <sup>L</sup>	<u>FJ006591</u> <sup>T,L</sup>	
			SG-10	<u>FJ006649</u> <sup>T,L</sup>	<u>FJ006703</u>	<u>FJ006705</u>	<u>FJ006708</u>	<u>FJ006709</u>	<u>FJ006714</u> <sup>T,L</sup>
			<u>FJ006715</u> <sup>L</sup>	<u>FJ006716</u> <sup>T,L</sup>	<u>FJ006717</u>	<u>FJ006718</u> <sup>L</sup>	<u>FJ006719</u> <sup>T,L</sup>	<u>FJ006720</u> <sup>L</sup>	
			SG-11	<u>FJ006642</u> <sup>T,L</sup>					
		SG-12	<u>AF234649</u>	<u>AJ133613</u>	<u>AY338386</u> <sup>T,L</sup>	<u>AY338387</u> <sup>T,L</sup>	<u>AY338390</u>	<u>AY338391</u> <sup>T,L</sup>	
			<u>AY338393</u> <sup>L</sup>	<u>AY338394</u> <sup>T,L</sup>	<u>AY338395</u> <sup>T</sup>	<u>D78183</u>	<u>FJ006567</u> <sup>T,L</sup>	<u>FJ006578</u> <sup>T,L</sup>	
		<u>FJ006585</u> <sup>T,L</sup>	<u>FJ006588</u> <sup>T,L</sup>	<u>FJ006599</u>	<u>FJ006607</u> <sup>T,L</sup>	<u>FJ006619</u> <sup>T,L</sup>	<u>FJ006645</u> <sup>L</sup>		
	<u>FJ006646</u> <sup>L</sup>	<u>FJ006647</u> <sup>L</sup>	<u>FJ006648</u> <sup>L</sup>	<u>FJ006651</u>	<u>FJ006684</u> <sup>T,L</sup>	<u>M13713</u> <sup>L</sup>			
	SG-13	<u>AF234643</u> <sup>T,L</sup>	<u>AF234644</u>	<u>AY338392</u> <sup>T</sup>	<u>FJ006579</u>	<u>FJ006580</u>	<u>FJ006581</u>		
		<u>FJ006582</u>	<u>FJ006583</u>	<u>FJ006584</u>	<u>FJ006586</u>	<u>FJ006587</u>	<u>FJ006589</u>		
		<u>FJ006593</u>	<u>FJ006594</u>	<u>FJ006597</u>	<u>FJ006598</u> <sup>T</sup>	<u>FJ006601</u>	<u>FJ006606</u>		
		<u>FJ006608</u> <sup>T</sup>	<u>FJ006609</u>	<u>FJ006610</u>	<u>FJ006611</u>	<u>FJ006613</u>	<u>FJ006614</u>		
		<u>FJ006615</u>	<u>FJ006616</u>	<u>FJ006620</u>	<u>FJ006623</u>	<u>FJ006624</u>	<u>FJ006625</u>		
		<u>FJ006626</u>	<u>FJ006627</u>	<u>FJ006628</u>	<u>FJ006629</u>	<u>FJ006630</u>	<u>FJ006631</u>		
<u>FJ006632</u>		<u>FJ006633</u>	<u>FJ006634</u>	<u>FJ006635</u> <sup>L</sup>	<u>FJ006636</u>	<u>FJ006637</u>			
<u>FJ006639</u>		<u>FJ006640</u>							
C7	C7-P6	SG-14	<u>AJ416336</u> <sup>L</sup>	<u>AJ416337</u> <sup>T,L</sup>	<u>AJ416338</u> <sup>L</sup>	<u>AJ416339</u> <sup>L</sup>			
	C7-P7	SG-15	<u>M16064</u>						
	C7-P8	SG-16	<u>FJ006687</u> <sup>T</sup>						
	C7-P9	SG-17	<u>FJ006644</u> <sup>L</sup>						

The <sup>a</sup> indicate the representatives of the nine patterns used in Figure 3; <sup>T</sup> and <sup>L</sup> indicate the sequences that contain a higher proportion of total essential amino acids and Lys, respectively. Some sequences are marked more than once. For example, AJ937838<sup>T,L</sup> means AJ937838 has been used in Figure 3, and the proportions of its total essential amino acids and Lys are higher.

(5/5); SG-10 (4/12); SG-12 (5/34); SG-13 (2/44); SG-14 (4/4) SG-17 (1/1)). However, a majority of them (16/23) are members of C8. Five out of seven members of C7 contain the toxic peptides. Only two out of the 45 sequences of C9 contain the toxic epitopes. One (FJ006591) of the 18  $\gamma$ -gliadins with fewer than 85 residues in the repetitive domain contain a toxic epitope. Finally, occurrences of the three epitopes are at the frequencies of 18/169, 19/169 and 11/169, respectively.

A further look at the sequences reveals that amino acid change(s) in particular epitopes caused by SNP disrupt the continuous peptides. Forty-one analogues of epitopes are shown in Additional file 10. Substitutions of amino acid often concern glutamine and proline residue. Amino acid

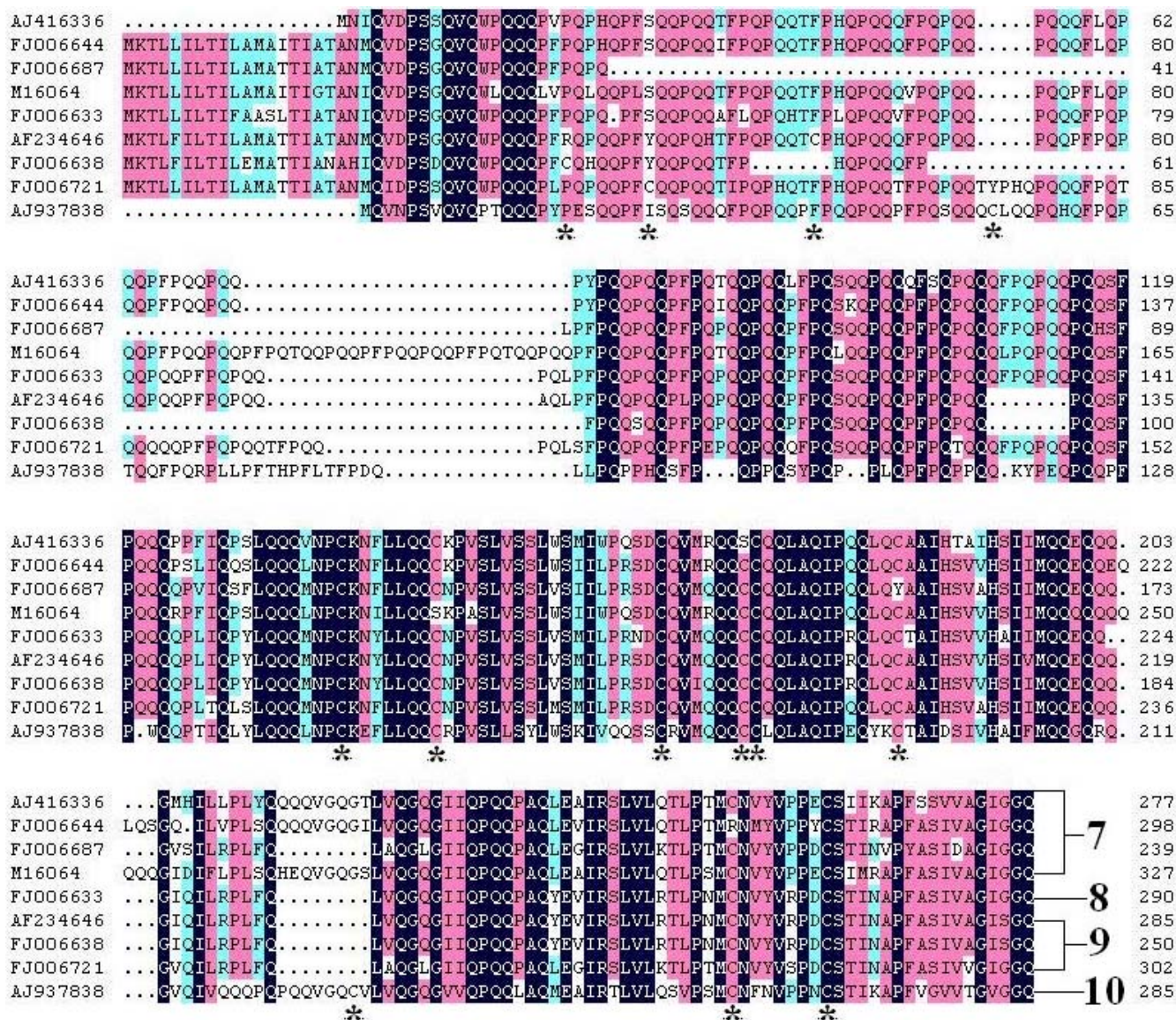
insertions and deletions also occur frequently, which destroy the epitopes (data not shown).

## Discussion

### Diversity of $\gamma$ -gliadin sequences

Sequence diversity between the  $\gamma$ -gliadin genes is due to SNPs and variations in the repetitive region, and the latter is mainly responsible for the size heterogeneity of the  $\gamma$ -gliadins [11,19]. It is comparatively confirmed by our result. Nucleotide diversity ( $\theta_w$  and  $\pi$ ) was estimated to be lower in self-pollinated diploid wheat than open-pollinated *Aegilops* species (Table 3). Besides the large number of segregating sites and mutations, haplotype diversity of  $\gamma$ -gliadin genes is high in every species investigated, which indicates a high nucleotide diversity of  $\gamma$ -gliadin family.





**Figure 3**  
**Alignment of nine  $\gamma$ -gliadin polypeptide sequences, which are representatives of each cysteine distribution pattern.** The gaps in the internal parts of the repetitive domain are not definitely addressed. Asterisks at the bottom of alignment indicate the positions of cysteine. The numbers at the end indicate the number of cysteine residues in each sequence. AJ937838<sup>a</sup>, AF234646<sup>a</sup>, FJ006638<sup>a</sup>, FJ006721<sup>a</sup>, FJ006633<sup>a</sup>, AJ416336<sup>a</sup>, M16064<sup>a</sup>, FJ006687<sup>a</sup> and FJ006644<sup>a</sup> are representatives of the patterns P1 to P9, respectively (Table 5).

Extensive linkage disequilibrium found with different species indicates similar ancestries between freely recombining portions of *Gli-1* loci. Negative values of the neutrality test statistics (Tajima's D, Fu and Li's D) in most species suggest that they are mainly under negative selection. Overall, *Gli-1* loci in different species are diverse, although  $\gamma$ -gliadin is supposed to be the most ancient family among prolamins [18].

**Evolution of  $\gamma$ -gliadin multigene family**

To avoid PCR bias, two forward and two reverse primers are used to amplify the full ORF of  $\gamma$ -gliadin genes. As a result, we isolated 29 unique  $\gamma$ -gliadin genes from *T. aestivum* cv 'Chinese Spring'. Meanwhile, nine to 14 unique genes were cloned from the nine diploid wheat and *Aegilops* species, respectively, with an average of 11 (Table 1). There are 15 to 40 copies of  $\gamma$ -gliadin genes in Chinese

**Table 6:  $\gamma$ -gliadin sequences whose repetitive domains contain fewer than 85 amino acid residues**

Sequence	Number	ORF Length	Origin	Subgroup
<a href="#">AF234643</a>	80	771 bp	<i>T. aestivum</i>	SG-13
<a href="#">AY338392</a>	78	768 bp	<i>T. turgidum</i> ssp. <i>durum</i>	SG-13
<a href="#">FJ006564</a>	80	783 bp	<i>T. dicoccoides</i>	SG-8
<a href="#">FJ006576</a>	76	729 bp	<i>T. dicoccoides</i>	SG-6
<a href="#">FJ006591</a>	73	744 bp	<i>T. aestivum</i>	SG-6
<a href="#">FJ006598</a>	71	714 bp	<i>T. aestivum</i>	SG-13
<a href="#">FJ006607</a>	83	783 bp	<i>T. aestivum</i>	SG-12
<a href="#">FJ006608</a>	61	684 bp	<i>T. aestivum</i>	SG-13
<a href="#">FJ006612</a>	76	729 bp	<i>T. aestivum</i>	SG-6
<a href="#">FJ006619</a>	66	732 bp	SHW	SG-12
<a href="#">FJ006638</a>	84	753 bp	<i>T. urartu</i>	SG-3
<a href="#">FJ006642</a>	61	735 bp	<i>Ae. longissima</i>	SG-11
<a href="#">FJ006649</a>	46	690 bp	<i>Ae. longissima</i>	SG-10
<a href="#">FJ006687</a>	73	720 bp	<i>Ae. searsii</i>	SG-16
<a href="#">FJ006714</a>	46	681 bp	<i>Ae. sharonesis</i>	SG-10
<a href="#">FJ006716</a>	73	762 bp	<i>Ae. sharonesis</i>	SG-10
<a href="#">FJ006717</a>	46	678 bp	<i>Ae. sharonesis</i>	SG-10
<a href="#">FJ006719</a>	56	711 bp	<i>Ae. sharonesis</i>	SG-10
<b>Average</b>	<b>68.3</b>	<b>731.5 bp</b>		

Number = the number of amino acid residues in the repetitive domains.

Spring [10]. Considering the copy number, it can be concluded that the  $\gamma$ -gliadin sequences we cloned could represent the whole  $\gamma$ -gliadin family. Multiplication of  $\gamma$ -gliadin genes should have occurred in the diploid level, since a large number of  $\gamma$ -gliadin genes have been isolated from a few accessions, which is similar to the result on  $\alpha$ -gliadin genes [30].

The number of  $\gamma$ -gliadin sequences cloned from two accessions of *T. dicoccoides* and AS2255 of *T. turgidum* are 9, 7

and 13, respectively. It seems likely that tetraploid wheat went through a bottleneck in *Gli-1* loci, which is supported by the small copy numbers compared with those of common wheat and diploid species. We could conclude that great changes happened to the *Gli-1* regions in the formation of tetraploid wheat. It is possible that some  $\gamma$ -gliadin sequences disappeared from the genome. It is interesting to note that similar event seems not occur during the formation of *T. aestivum*. Experimental data, based on the simulation of the evolutionary step by synthetic

**Table 7: The distribution of three T cell stimulatory epitopes in  $\gamma$ -gliadins**

FLQPQQFPQPQPYPQQPQQPFPQ			LQPQQFPQPQPYPQQPQ			FSQPQQQFPQPQ		
Gene	Subgroup	Species	Gene	Subgroup	Species	Gene	Subgroup	Species
<a href="#">AFI20267</a>	SG-9	<i>T. spelta</i>	<a href="#">AFI20267</a>	SG-9	<i>T. spelta</i>	<a href="#">AFI20267</a>	SG-9	<i>T. spelta</i>
<a href="#">AFI44104</a>	SG-9	<i>T. aestivum</i>	<a href="#">AFI44104</a>	SG-9	<i>T. aestivum</i>	<a href="#">AFI44104</a>	SG-9	<i>T. aestivum</i>
<a href="#">AJ416336</a>	SG-14	<i>T. aestivum</i>	<a href="#">AJ416336</a>	SG-14	<i>T. aestivum</i>	<a href="#">AJ416336</a>	SG-14	<i>T. aestivum</i>
<a href="#">AJ416339</a>	SG-14	<i>T. aestivum</i>	<a href="#">AJ416337</a>	SG-14	<i>T. aestivum</i>	<a href="#">AJ416337</a>	SG-14	<i>T. aestivum</i>
<a href="#">AY338389</a>	SG-9	<i>T. aestivum</i>	<a href="#">AJ416339</a>	SG-14	<i>T. aestivum</i>	<a href="#">AJ416338</a>	SG-14	<i>T. aestivum</i>
<a href="#">EF151018</a>	SG-9	<i>T. aestivum</i>	<a href="#">AY338389</a>	SG-9	<i>T. aestivum</i>	<a href="#">AJ416339</a>	SG-14	<i>T. aestivum</i>
<a href="#">FJ006586</a>	SG-13	<i>T. turgidum</i>	<a href="#">EF151018</a>	SG-9	<i>T. aestivum</i>	<a href="#">AY338389</a>	SG-9	<i>T. aestivum</i>
<a href="#">FJ006593</a>	SG-13	<i>T. aestivum</i>	<a href="#">FJ006586</a>	SG-13	<i>T. turgidum</i>	<a href="#">EF151018</a>	SG-9	<i>T. aestivum</i>
<a href="#">FJ006644</a>	SG-17	<i>Ae. Longissima</i>	<a href="#">FJ006593</a>	SG-13	<i>T. aestivum</i>	<a href="#">FJ006591</a>	SG-9	<i>T. aestivum</i>
<a href="#">FJ006645</a>	SG-12	<i>Ae. Longissima</i>	<a href="#">FJ006644</a>	SG-17	<i>Ae. longissima</i>	<a href="#">FJ006596</a>	SG-7	<i>T. aestivum</i>
<a href="#">FJ006646</a>	SG-12	<i>Ae. Longissima</i>	<a href="#">FJ006645</a>	SG-12	<i>Ae. longissima</i>	<a href="#">FJ006604</a>	SG-7	<i>T. aestivum</i>
<a href="#">FJ006647</a>	SG-12	<i>Ae. Longissima</i>	<a href="#">FJ006646</a>	SG-12	<i>Ae. longissima</i>			
<a href="#">FJ006648</a>	SG-12	<i>Ae. Longissima</i>	<a href="#">FJ006647</a>	SG-12	<i>Ae. longissima</i>			
<a href="#">FJ006684</a>	SG-12	<i>Ae. Searsii</i>	<a href="#">FJ006648</a>	SG-12	<i>Ae. longissima</i>			
<a href="#">FJ006703</a>	SG-10	<i>Ae. Bicornis</i>	<a href="#">FJ006684</a>	SG-12	<i>Ae. searsii</i>			
<a href="#">FJ006705</a>	SG-10	<i>Ae. Bicornis</i>	<a href="#">FJ006703</a>	SG-10	<i>Ae. bicornis</i>			
<a href="#">FJ006708</a>	SG-10	<i>Ae. Bicornis</i>	<a href="#">FJ006705</a>	SG-10	<i>Ae. bicornis</i>			
<a href="#">FJ006709</a>	SG-10	<i>Ae. Bicornis</i>	<a href="#">FJ006708</a>	SG-10	<i>Ae. bicornis</i>			
			<a href="#">FJ006709</a>	SG-10	<i>Ae. bicornis</i>			

*T. spelta* = *T. aestivum* ssp. *spelta*; *T. turgidum* = *T. turgidum* ssp. *turgidum*

hexaploid wheat (SHW-L1) and its parental lines (*T. turgidum* accession AS2255 and *Ae. tauschii* accession AS60), supports the assumption of  $\gamma$ -gliadin sequences disappearance. Therefore, duplication and subsequent divergence might be important as well at the polyploid level in contrast to the diploid strains.

Pseudogenes are involved in phylogenetic analysis, and they fall into correct clusters, which indicate that the occurrence of pseudogenes should have taken place after the divergence of diploid species.

### Classification

It is the primary structure of peptides that finally determines very specific properties of the ending biomaterials [31]. The structures are important to dough rheology and other aspects of food technology [32]. According to the characteristics of primary structure, i.e. number and placement of cysteine residues and the phylogenetic result, we divided  $\gamma$ -gliadins into 17 subgroups based on the mature peptides (without signal peptide) (Table 5). The different subfamilies are very distinct from each other. The classification of  $\gamma$ -gliadins has essential importance with regard to dough quality, since cysteine residues play a critical role in unique properties of wheat flour. Typical  $\gamma$ -gliadins contain eight cysteine residues. We have also found  $\gamma$ -gliadins containing seven, nine and even ten cysteine residues. Furthermore, we identified nine cysteine distribution patterns. Changes in position and number of cysteine residues might affect the pattern of disulphide bond formation, resulting in failure of forming some intramolecular disulphide bond(s). These cysteine residues would then be available for intermolecular disulphide bond formation and polymer-building [33]. However, we have not known whether these gliadins are chain terminator (only one cysteine residue available for intermolecular disulphide) or chain extenders (subunits with more than one cysteine residues that form intermolecular disulphide bonds), which would presumably have a negative effect on flour quality or allow the formation of stronger dough, respectively [34,35].

Alternatively, we classified  $\gamma$ -gliadins into two types: i.e. repetitive domain < 85 amino acids and repetitive domain  $\geq 85$  amino acids, which are named as sequences with a short repetitive domain (18 sequences) and sequences with a long repetitive domain (151 sequences) respectively. The repetitive domain is rich in glutamine and proline, which is the major sequence variation that discriminates the different  $\gamma$ -gliadins.

### Nutritional quality

It is well known that nutritional quality of food that lack essential amino acids is low, as the body tends to convert the amino acids obtained into fats and carbohydrates.

Therefore, a balance of essential amino acids is necessary for a high degree of net protein utilization (the mass ratio of amino acids converted to proteins: amino acids supplied). The net protein utilization is profoundly affected by the limiting amino acid proportion (the essential amino acid found in the smallest quantity in the foodstuff). The limiting amino acid of wheat is lysine, which is mainly caused by a low level of lysine in gliadin [36], since gliadins account for about half of the total storage proteins [3]. We systematically analyzed the proportions of eight essential amino acids of  $\gamma$ -gliadins, which indicates that subgroup SG-10 and SG-12 and the  $\gamma$ -gliadins with a short repetitive domain contain higher proportions of lysine and total essential amino acids. A wide range of the proportion of each essential amino acid could be seen, which provides the possibility of breeding more nutritional wheat varieties.

### Perspective for wheat breeding programs

The only efficient therapy for CD is a life-long gluten-free diet. Conceivably, a diet based on gluten from a wheat species that expresses no or few T-cell stimulatory gluten peptides should be equally well tolerated by the celiac patients and, importantly, also be beneficial for disease prevention [37,38]. It is also indicated that the genetic differences in gliadins might allow designing strategies for selection and breeding of less toxic wheat varieties [[30,37] and [38]]. Our results indicate that 23 out of the 169 putatively functional sequences contain  $\gamma$ -gliadin epitopes, and that  $\gamma$ -gliadins with a short repetitive domain almost contain no toxic epitopes, with the exception of FJ006591. Meanwhile, 22 sequences out of those with a long repetitive domain contain  $\gamma$ -gliadin epitopes. Obviously, the classification depending on the length of repetitive domain is reflected in the occurrence of toxic epitopes. CD-toxic peptides of  $\gamma$ -gliadins are only found in the repetitive domain, especially the internal part, which is highly variable in length. Those  $\gamma$ -gliadins with a short repetitive domain contain a brief internal part, which means that they are not/nearly not toxic to the population with celiac disease. The two subgroups SG-10 and SG-12, which show a relatively good nutritional quality, present four (4/12) and five (5/34) members containing epitopes. Therefore, it is suggested that the genes with a short repetitive domain are more nutritional and valuable. It is reported that stimulatory epitopes in  $\alpha$ -gliadins from the D genome is the highest, compared to those from the A and B genome [30,38]. However, we have not found any epitope in the  $\gamma$ -gliadins from *Ae. tauschii* (Table 7).

### Conclusion

We systematically characterized the  $\gamma$ -gliadin multigene family in common wheat and its closely related tetraploid and diploid species. It is shown that  $\gamma$ -gliadin family is highly diverse. Phylogenetic analyses indicate a more close

relationship between the *Gli-1* loci of the B(S) and D genomes. According to the differences in primary structure, we have classified  $\gamma$ -gliadins into 17 subgroups, which might reflect their differences in the contributions to the processing qualities of wheat flour. The  $\gamma$ -gliadins with a short repetitive domain are relatively more nutritional, since they contain a higher proportion of essential amino acids. Moreover, these short  $\gamma$ -gliadins almost contain no toxic epitopes. Therefore, it is possible to breed wheat varieties, the  $\gamma$ -gliadins of which are less, even non-toxic and more nutritional.

## Methods

### Plant materials and DNA extraction

*T. aestivum* cv. 'Chinese Spring' (CS) and its closely related wheat and *Aegilops* species (Table 1), one synthetic hexaploid wheat accession (SHW-L1; in the 3<sup>rd</sup>-4<sup>th</sup> generations) and its parental *Ae. tauschii* (AS60) and *T. turgidum* ssp. *turgidum* (AS2255) lines were used.

Genomic DNA was extracted from leaves of single adult plants with a CTAB (Cetyltrimethylammonium bromide) protocol [39].

### Primer design

The PCR primers to isolate  $\gamma$ -gliadin genes from genomic DNA were designed on the conserved parts of the 5' and 3' flanking sequences of  $\gamma$ -gliadin genes retrieved from Genbank <http://www.ncbi.nlm.nih.gov>, which are listed as follows:

Forward1: 5'-TATTAGTTAACGCAAATCCACC/TATG-3'

Forward2: 5'-CTTCACACAAGTAGAGCACAAG-3'

Reverse1: 5'-GATGAATCAGCTAAGCAACGATG-3'

Reverse2: 5'-TCGTTACATCTATTGGTGCATCAG)-3'

### PCR based gene cloning

PCR amplification was conducted in a 25  $\mu$ l volume, consisting of 100 ng genomic DNA, 100  $\mu$ M of each dNTPs, 1.5 mM of Mg<sup>2+</sup>, 2 pmol of each of the four primers, 0.75 U *Taq* polymerase with high fidelity (*TianGen*; P.R. China) and 2.5  $\mu$ l 10 $\times$ buffer (supplied with the *Taq* polymerase). The reactions were run in a PTC-240 (MJ Research, USA) thermal cycler with following program: an initial step of 94°C for 4 min; 35 cycles of 94°C for 45 sec, 57°C for 1 min and 72°C for 80 sec; then a final step of 8 min at 72°C.

The amplified products were separated in 1% agarose gel. The desired fragments were recovered and cloned into pMD-18T vector (*Takara*), then transformed into compe-

tent *E. coli* (JM109) cells. Positive colonies were screened out and sequenced by commercial company (*Invitrogen*).

### DNA sequence analysis

Sequenced clones were confirmed by Blast analysis <http://www.ncbi.nlm.nih.gov>, and aligned using DNAMAN (version 5.2.2; Lynnon Biosoft), Clustal X (version 1.81) [40] and MEGA (version 3.1) [41]. Further modifications to the alignment were done manually. Bootstrap test of phylogenies (1000 replicates; neighbor-joining method) were carried out using MEGA for the nucleotide sequences from initial codon (ATG) to mature stop codon (TGA), on the basis of Kimura 2-parameter distances, complete deletion of gaps. Neighbor-joining trees (1000 replicates) were also constructed for classification of mature proteins on the basis of poisson correction, complete deletion of gaps.

Nucleotide diversity was estimated with three approaches. Indels were excluded from the estimates. The first method used the number of haplotypes to estimate heterozygosity [42]. The second approach used the average number of nucleotide differences per site between two sequences ( $\pi$ ) [42]. The last method used the number of segregating sites to estimate nucleotide diversity per site ( $\theta_w$ ) [42]. The strength of linkage disequilibrium (LD) was estimated using the ZnS statistic [43], which is the average of r<sup>2</sup> (squared correlation coefficient) [44] over all pairwise comparisons. Furthermore, Sequence data analysis (minimum number of recombination events (Rm) [45] and a statistical test of neutrality (Tajima's D and Fu and Li's D) [46,47]) were performed as well. Coalescent simulations were used to test for significant differences [48,49]. DnaSP version 4.50.3 [50] software package was used to complete these analyses.

### Amino acid composition analysis

Amino acid composition data of mature  $\gamma$ -gliadin peptides were determined by MEGA. Statistical analyses were carried out by Statistica version 6.0 <http://www.statsoft.com/>.

### Epitope screening

The program MEGA was used for matching the  $\gamma$ -gliadin epitopes. Only perfect matches were considered.

### Authors' contributions

PFQ, YMW and YLZ designed the experiments. PFQ selected the plant materials and designed the primers. PFQ, QC and XT cloned the  $\gamma$ -gliadin sequences. PFQ, YMW, TO and YLZ analyzed the sequences. PFQ, YMW and TO drafted the manuscript. All authors have read and approved the final manuscript.



## Additional material

### Additional file 1

Alignment of the nucleotide sequences of the 170  $\gamma$ -gliadin genes shown in Table 1.

*Alignment of the nucleotide sequences of the 170  $\gamma$ -gliadin genes shown in Table 1.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-168-S1.meg>]

### Additional File 2

Phylogenic tree of the  $\gamma$ -gliadin gene sequences isolated from *T. urartu*, *Ae. speltoides* and *Ae. tauschii*.

*Phylogenic tree of the  $\gamma$ -gliadin gene sequences isolated from *T. urartu*, *Ae. speltoides* and *Ae. tauschii*. 'p' = pseudogene; *T. urartu*-A, *Ae. speltoides*-S and *Ae. tauschii*-D indicate species-genome; AF20184 is a 75 K  $\gamma$ -secaline gene, which is used as the outgroup.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-168-S2.png>]

### Additional File 3

Phylogenic tree obtained from alignment of the  $\gamma$ -gliadin genes from tetraploid wheat.

*Phylogenic tree obtained from alignment of the  $\gamma$ -gliadin genes from tetraploid wheat.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-168-S3.png>]

### Additional File 4

Phylogenic tree obtained from alignment of the  $\gamma$ -gliadin genes from hexaploid wheat.

*Phylogenic tree obtained from alignment of the  $\gamma$ -gliadin genes from hexaploid wheat. Cluster II contained the  $\gamma$ -gliadin genes from both Gli-B1 and Gli-D1 loci.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-168-S4.png>]

### Additional File 5

Phylogenic analysis of the deduced mature proteins of the  $\gamma$ -gliadins in pattern C9-P4. SG-4 to SG-7 = subgroups described in Table 5.

*Phylogenic analysis of the deduced mature proteins of the  $\gamma$ -gliadins in pattern C9-P4. SG-4 to SG-7 = subgroups described in Table 5.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-168-S5.png>]

### Additional File 6

Neighbour-joining tree of the  $\gamma$ -gliadins in pattern C8-P5, on the basis of alignment of the deduced mature protein sequences.

*Neighbour-joining tree of the  $\gamma$ -gliadins in pattern C8-P5, on the basis of alignment of the deduced mature protein sequences. SG-8 to SG-13 = subgroups described in Table 5.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-168-S6.png>]

### Additional file 7

The frequencies of the 17 subgroups among the wheat and *Aegilops* accessions used in Table 1.

*The frequencies of the 17 subgroups among the wheat and *Aegilops* accessions used in Table 1.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-168-S7.xls>]

### Additional file 8

The average proportions of eight essential amino acids.

*The average proportions of eight essential amino acids. Avg-1, Avg-2 and Avg-3 indicate the average proportions of the eight essential amino acids in the 169 mature  $\gamma$ -gliadins (putatively functional), in the 18  $\gamma$ -gliadins whose repetitive domains contain fewer than 85 amino acid residues (Their repetitive domains are included for data analyses; Table 6) and in the 169 mature  $\gamma$ -gliadins ignoring the repetitive domains, respectively. Data are mean  $\pm$  SD (standard deviation).*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-168-S8.doc>]

### Additional file 9

The  $\gamma$ -gliadins with higher proportions of Lys or total essential amino acids.

*The  $\gamma$ -gliadins with higher proportions of Lys or total essential amino acids. In bold are the  $\gamma$ -gliadins with fewer than 85 residues in the repetitive domain.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-168-S9.xls>]

### Additional file 10

Analogues of CD-toxic epitopes in  $\gamma$ -gliadins.

*Analogues of CD-toxic epitopes in  $\gamma$ -gliadins. The bold and italic letters indicate amino acid substitutions.*

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-168-S10.doc>]

## Acknowledgements

The PI and Clae designated materials were kindly supplied by Germplasm Resources Information Network, USA. AS2255 and SHW-L1 were kindly supplied by Dr. DC Liu, Triticeae Research Institute of Sichuan Agricultural University. This work was supported by the National High Technology Research and Development Program of China (863 program 2006AA10Z179 and 2006AA10Z1F8), and the National Basic Research Program of China (973 Program 2009CB118304).

## References

1. Shewry PR, Tatham AS, Lazzeri P: **Biotechnology of wheat quality.** *J Sci Food Agric* 1997, **73**:397-406.
2. Payne PI: **Genetics of wheat storage proteins and the effect of allelic variation on bread-making quality.** *Annu Rev Plant Physiol* 1987, **38**:141-153.
3. Qi PF, Wei YM, Yue YW, Yan ZH, Zheng YL: **Biochemical and molecular characterization of gliadins.** *Mol Biol* 2006, **40**:713-723.

4. Bietz JA, Huebner FR, Sanderson JE, Wall JS: **Wheat gliadin homology revealed through N-terminal amino acid sequence analysis.** *Cereal Chem* 1977, **54**:1070-1083.
5. Payne PI, Law CN, Mudd EE: **Control by homoeologous group I chromosomes of the high-molecular-weight subunits of glutenin, a major protein of wheat endosperm.** *Theor Appl Genet* 1980, **58**:113-120.
6. Payne PI, Holt LM, Jackson EA, Law CN: **Wheat storage proteins: their genetics and potential for manipulation by plant breeding.** *Philos Trans R Soc London Ser B* 1984, **304**:359-371.
7. Payne PI, Jackson EA, Holt LM, Law CN: **Genetic linkage between endosperm storage protein genes on each of the short arms of chromosomes 1A and 1B in wheat.** *Theor Appl Genet* 1984, **67**:235-243.
8. Tatham AS, Shewry PR: **The S-poor prolamins of wheat, barley and rye.** *J Cereal Sci* 1995, **22**:1-16.
9. Shewry PR, Halford NG: **Cereal seed storage proteins: structures, properties and role in grain utilization.** *J Exp Bot* 2002, **53**:947-958.
10. Sabelli P, Shewry PR: **Characterization and organization of gene families at the *Gli-1* loci of bread and durum wheats by restriction fragment analysis.** *Theor Appl Genet* 1991, **83**:209-216.
11. Okita TW, Cheesbrough V, Reeves CD: **Evolution and heterogeneity of the  $\alpha$ - $\beta$ -type and  $\gamma$ -type gliadin DNA sequences.** *J Biol Chem* 1985, **260**:8203-8213.
12. Pistón F, Dorado G, Martín A, Barro F: **Cloning of nine  $\gamma$ -gliadin mRNAs (cDNAs) from wheat and the molecular characterization of comparative transcript levels of  $\gamma$ -gliadin subclasses.** *J Cereal Sci* 2006, **43**:120-128.
13. Shan L, Molberg Ø, Parrot I, Hausch F, Filiz F, Gray GM, Sollid LM, Khosla C: **Structural basis for gluten intolerance in celiac disease.** *Science* 2002, **297**:2275-2279.
14. McAdam SN, Sollid LM: **Getting to grips with gluten.** *Gut* 2000, **47**:743.
15. Koning F: **The molecular basis of celiac disease.** *J Mol Recognit* 2003, **16**:333-336.
16. Dieterich W, Esslinger B, Schuppan D: **Pathomechanisms in celiac disease.** *Int Arch Allergy Immunol* 2003, **132**:98-108.
17. Gänzle MG, Laponen J, Gobetti M: **Proteolysis in sourdough fermentations: mechanisms and potential for improved bread quality.** *Trends Food Sci Tech* 2008, **19**:513-521.
18. Shewry PR, Tatham AS: **The prolamins storage proteins of cereal seeds: structure and evolution.** *Biochem J* 1990, **267**:1-12.
19. Anderson OD, Hisa CC, Torres V: **The wheat  $\gamma$ -gliadin genes: characterization of ten new sequences and further understanding of  $\gamma$ -gliadin gene family structure.** *Theor Appl Genet* 2001, **103**:323-330.
20. von Büren M, Lüthy J, Hübner P: **A spelt-specific  $\gamma$ -gliadin gene: discovery and detection.** *Theor Appl Genet* 2000, **100**:271-279.
21. Arentz-Hansen EH, McAdam SN, Molberg Ø, Kristiansen C, Sollid LM: **Production of a panel of recombinant gliadins for the characterisation of T cell reactivity in coeliac disease.** *Gut* 2000, **46**:46-51.
22. Arentz-Hansen H, McAdam SN, Molberg Ø, Fleckenstein B, Lundin KE, Jorgensen TJ, Jung G, Roepstorff P, Sollid LM: **Celiac lesion T cells recognize epitopes that cluster in regions of gliadins rich in proline residues.** *Gastroenterology* 2002, **123**:803-809.
23. Maruyama N, Ichise K, Katsube T, Kishimoto T, Kawase S, Matsumura Y, Takeuchi Y, Sawada T, Utsumi S: **Identification of major wheat allergens by means of the *Escherichia coli* expression system.** *Eur J Biochem* 1998, **255**:739-745.
24. Bartels D, Altsaar I, Harberd NP, Barker RF, Thompson RD: **Molecular analysis of gamma-gliadin gene families at the complex *Gli-1* locus of bread wheat (*T. aestivum* L.).** *Theor Appl Genet* 1986, **72**:845-853.
25. Rafalski JA: **Structure of wheat gamma-gliadin genes.** *Gene* 1986, **43**:221-229.
26. Sugiyama T, Rafalski A, Soell D: **The Nucleotide sequence of a wheat gamma-gliadin genomic clone.** *Plant Sci* 1986, **44**:205-209.
27. Scheets K, Hedgcoth C: **Nucleotide sequence of a gamma gliadin gene: Comparisons with other gamma gliadin sequences show the structure of gamma gliadin genes and the general primary structure of gamma gliadins.** *Plant Sci* 1988, **57**:141-150.
28. D'Ovidio R, Simeone M, Masci S, Porceddu E, Kasarda DD: **Nucleotide sequence of a gamma-type glutenin gene from a durum wheat: Correlation with a gamma-type glutenin subunit from the same biotype.** *Cereal Chem* 1995, **72**:443-449.
29. Müller SW, Wieser H: **The location of disulphide bonds in  $\alpha$ -type gliadins.** *J Cereal Sci* 1995, **22**:21-27.
30. van Herpen TWJM, Goryunova SV, Schoot J van der, Mitreva M, Salentijn E, Vorst O, Schenk MF, van Veelen PA, Koning F, van Soest LJM, Vosman B, Bosch D, Hamer RJ, Gilissen LMWJ, Smulders MJM: **Alpha-gliadin genes from the A, B and D genomes of wheat contain different sets of celiac disease epitopes.** *BMC Genomics* 2006 [<http://www.biomedcentral.com/1471-2164/7/11>].
31. Elmorjani K, Thiévin M, Michon T, Popineau Y, Hallet JN, Guéguen J: **Synthetic genes specifying periodic polymers modeled on the repetitive domain of wheat gliadins: conception and expression.** *Biochem Biophys Res Commun* 1997, **239**:240-246.
32. DuPont FM, Vensel W, Encarnacao T, Chan R, Kasarda DD: **Similarities of omega gliadins from *Triticum urartu* to those encoded on chromosome 1A of hexaploid wheat and evidence for their post-translational processing.** *Theor Appl Genet* 2004, **108**:1299-1308.
33. Masci S, Rovelli L, Kasarda DD, Vensel WH, Lafiandra D: **Characterization and chromosomal localization of C-type low-molecular-weight glutenin subunits in the bread wheat cultivar Chinese Spring.** *Theor Appl Genet* 2002, **104**:422-428.
34. Tao HP, Cornell DG, Kasarda DD: **Surface and optical properties of wheat glutenin monolayers.** *J Cereal Sci* 1989, **10**:5-18.
35. Ikeda TM, Nagamine T, Fukuoka H, Yano H: **Identification of new Low molecular weight glutenin subunit genes in wheat.** *Theor Appl Genet* 2002, **104**:680-687.
36. Kasarda DD, Bernardin JE, Nimmo CC: **Wheat proteins in Advances in Cereal Science and Technology. Volume 1.** Edited by: Promeranz Y. St. Paul, MN: American Association of Cereal Chemists; 1976:158-236.
37. Molberg Ø, Uhlen AK, Jensen T, Flåte NS, Fleckenstein B, Arentz-Hansen H, Raki M, Lundin KEA, Sollid LM: **Mapping of gluten T cell epitopes in the bread wheat ancestors: implications for celiac disease.** *Gastroenterology* 2005, **128**:393-401.
38. Spaenij-Dekking L, Kooy-Winkelaar Y, van Veelen P, Wouter Drijfhout J, Jonker H, van Soest L, Smulders M, Bosch D, Gilissen L, Koning F: **Natural variation in toxicity of wheat: potential for selection of nontoxic varieties for celiac disease patients.** *Gastroenterology* 2005, **129**:797-806.
39. Doyle JJ, Dolyle JL: **A rapid DNA isolation procedure from small quantities of fresh leaf tissues.** *Phytochem Bull* 1987, **19**:11-15.
40. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25**:4876-4882.
41. Kumar S, Tamura K, Nei M: **MEGA 3: Integrated software for molecular evolutionary genetics analysis and sequence alignment.** *Brief Bioinform* 2004, **5**:50-163.
42. Nei M: *Molecular Evolutionary Genetics* New York, Columbia University Press; 1987.
43. Kelly JK: **A test of neutrality based on interlocus associations.** *Genetics* 1997, **146**:1197-1206.
44. Hill WG, Robertson A: **Linkage disequilibrium in finite populations.** *Theor Appl Genet* 1968, **38**:226-231.
45. Hudson RR, Kaplan NL: **Statistical properties of the number of recombination events in the history of a sample of DNA sequences.** *Genetics* 1985, **111**:147-164.
46. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123**:585-595.
47. Fu YX, Li WH: **Statistical tests of neutrality of mutations.** *Genetics* 1993, **133**:693-709.
48. Hudson RR: **Gene genealogies and the coalescent process.** *Oxf Surv Evol Biol* 1990, **7**:1-44.
49. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical recipes in C. The art of Scientific Computing* Cambridge, Cambridge University Press; 1992.
50. Rozas J, Sanchez-Del Barrio JC, Messeguer X, Rozas R: **DnaSP, DNA polymorphism analyses by the coalescent and other methods.** *Bioinform* 2003, **19**:2496-2497.