the plant journal

SEB Society for Experimental Biology

*The Plant Journal* (2015) **82,** 794–805 doi: 10.1111/tpj.12851

# Diversification of the celiac disease α-gliadin complex in wheat: a 33-mer peptide with six overlapping epitopes, evolved following polyploidization

**Carmen V. Ozuna[1,†], Julio C. M. Iehisa[1,†,‡], María J. Giménez[1], Juan B. Alvarez[2], Carolina Sousa[3] and Francisco Barro[1,*]**

[1]*Departamento de Mejora Genética, Instituto de Agricultura Sostenible (IAS), Consejo Superior de Investigaciones Científicas (CSIC), Córdoba, E-14080 Spain,*

[2]*Departamento de Genética, Escuela Superior de Ingenieros Agrónomos y Montes, Universidad de Córdoba, Córdoba, E-14071 Spain, and*

[3]*Departamento de Microbiología y Parasitología, Facultad de Farmacia, Universidad de Sevilla, Sevilla, 41012 Spain*

Received 31 January 2015; revised 1 April 2015; accepted 2 April 2015; published online 10 April 2015.

*For correspondence (email fbarro@ias.csic.es).

[†]These authors contributed equally to this work.

[‡]Present address: Departamento de Biotecnología, Facultad de Ciencias Químicas, Universidad Nacional de Asunción, San Lorenzo, Paraguay.

## SUMMARY

**The gluten proteins from wheat, barley and rye are responsible both for celiac disease (CD) and for non-celiac gluten sensitivity, two pathologies affecting up to 6–8% of the human population worldwide. The wheat α-gliadin proteins contain three major CD immunogenic peptides: p31–43, which induces the innate immune response; the 33-mer, formed by six overlapping copies of three highly stimulatory epitopes; and an additional DQ2.5-glia-α3 epitope which partially overlaps with the 33-mer. Next-generation sequencing (NGS) and Sanger sequencing of α-gliadin genes from diploid and polyploid wheat provided six types of α-gliadins (named 1–6) with strong differences in their frequencies in diploid and polyploid wheat, and in the presence and abundance of these CD immunogenic peptides. Immunogenic variants of the p31–43 peptide were found in most of the α-gliadins. Variants of the DQ2.5-glia-α3 epitope were associated with specific types of α-gliadins. Remarkably, only type 1 α-gliadins contained 33-mer epitopes. Moreover, the full immunodominant 33-mer fragment was only present in hexaploid wheat at low abundance, probably as the result of allohexaploidization events from subtype 1.2 α-gliadins found only in *Aegilops tauschii*, the D-genome donor of hexaploid wheat. Type 3 α-gliadins seem to be the ancestral type as they are found in most of the α-gliadin-expressing Triticeae species. These findings are important for reducing the incidence of CD by the breeding/selection of wheat varieties with low stimulatory capacity of T cells. Moreover, advanced genome-editing techniques (TALENs, CRISPR) will be easier to implement on the small group of α-gliadins containing only immunogenic peptides.**

**Keywords: alpha-gliadin, wheat, celiac disease, 33-mer peptide.**

## INTRODUCTION

Wheat is one of the most important crops in the world, with an annual production of about 715 million tons (2013; http://faostat3.fao.org/). Bread wheat (*Triticum aestivum*, $2n = 6x = 42$; genomic code BBAADD) is an allohexaploid species that arose by natural hybridization between emmer wheat (*Triticum turgidum* ssp. *dicoccum*, $2n = 4x = 28$, BBAA), and the diploid *Aegilops tauschii* ($2n = 2x = 14$, DD) (Petersen *et al.*, 2006). In turn, tetraploid emmer wheat is hypothesized to have originated through hybridization between the diploids *T. urartu* (AA) and, possibly, *Ae. speltoides* (SS) (Petersen *et al.*, 2006).

Despite its relatively low protein content (8–15%), wheat is the most important protein source in the human diet. Gluten, the water insoluble fraction of wheat flour protein, is responsible for the bread-making quality of wheat and is mainly composed of two prolamin fractions, called gliadins (α, γ and ω) and glutenins (Shewry, 2009). The ingestion of these proteins is responsible for two important pathologies: (i) celiac disease (CD), a food-sensitive enteropathy with a prevalence of about 0.7–2% in the human population, in genetically predisposed individuals (Rewers, 2005); and (ii) gluten sensitivity, a newly-recognized pathology

with an estimated prevalence of 6% in the USA population (Sapone *et al.*, 2011). In CD, T cells isolated from the intestinal mucosa typically recognize gluten peptides in which specific glutamine residues are converted to glutamate by tissue transglutaminase 2 (tTG2). These modified peptides are able to bind to class II human histocompatibility leukocyte antigen (HLA) molecules DQ2 and DQ8, which stimulate T cells and trigger an inflammatory response in the small intestine leading to flattening of the mucosa (Wieser and Koehler, 2008). Over 90% of CD patients possess HLA-DQ2, encoded by the *DQA1*05* and *DQB1*02* genes (Karell *et al.*, 2003).

The α-gliadin 33-mer is one of the digestion-resistant gluten peptides that is highly reactive to isolated celiac T cells and is the main immunodominant toxic peptide in celiac patients. This peptide is present in the N-terminal repetitive region of α-gliadins and contains six overlapping copies of three different DQ2-restricted T-cell epitopes with highly stimulatory properties (Shan *et al.*, 2002). α-gliadins also contain an additional DQ2-restricted epitope which partially overlaps with 33-mer peptide (Vader *et al.*, 2002). Moreover, the peptide p31–43 of these α-gliadins has been reported to induce the innate immune response necessary to initiate the T-cell adaptive response (Maiuri *et al.*, 1996a, 2003).

The α-gliadins are encoded by the *Gli-2* loci located on the short arms of the homoeologous group 6 chromosomes of wheat (Payne, 1989). The estimated copy numbers of α-gliadins in hexaploid wheat are between 25 and 150 copies (Harberd *et al.*, 1985; Anderson *et al.*, 1997). Analysis of this highly variable multigene family has been performed in tetraploid wheat through RNA-amplicon sequencing applying 454's NGS technology (Salentijn *et al.*, 2013). In this work a comprehensive study combining NGS genomic-amplicon sequencing and Sanger sequencing of the entire fragment of α-gliadins containing immunogenic epitopes has been carried out in diploid, tetraploid and hexaploid wheats. We identify six different types of α-gliadins but only one type contains all the immunogenic peptides and epitopes, and the five other types of α-gliadins do not contain epitopes for the 33-mer peptide.

## RESULTS

### Genes and pseudogenes of α-gliadins

We subjected the genome of domesticated and wild wheat and relatives, including 96 accessions of *Triticum* and *Aegilops* species (Data S1 and Figure S1), to amplicon NGS. The N-terminal repetitive region of α-gliadins, containing three highly immunogenic peptides, was amplified and sequenced (Figure 1). We obtained 200 340 cleaned reads (see Experimental Procedures) with an average of 2087 reads per accession, which were clustered at 99% identity and then a consensus sequence was extracted from each cluster. The high-confidence sequence variants (see Experimental Procedures) were grouped into 999 unique clusters which consisted of 88 736 total reads (Table S1). Multiple alignments of consensus sequences were performed. However, alignments of sequences with frequent insertions/deletions and repeat units, such as α-gliadins, are not accurate using traditional software (Loytynoja and Goldman, 2005; Jordan and Goldman, 2012). In α-gliadins, the repeat unit PFPPQQPYPQPQ or its variants can be found along the entire fragment from the p31–43 peptide to the DQ2.5-glia-α3 epitope. Considering this repeat unit, we manually aligned the consensus sequences obtained by clustering.

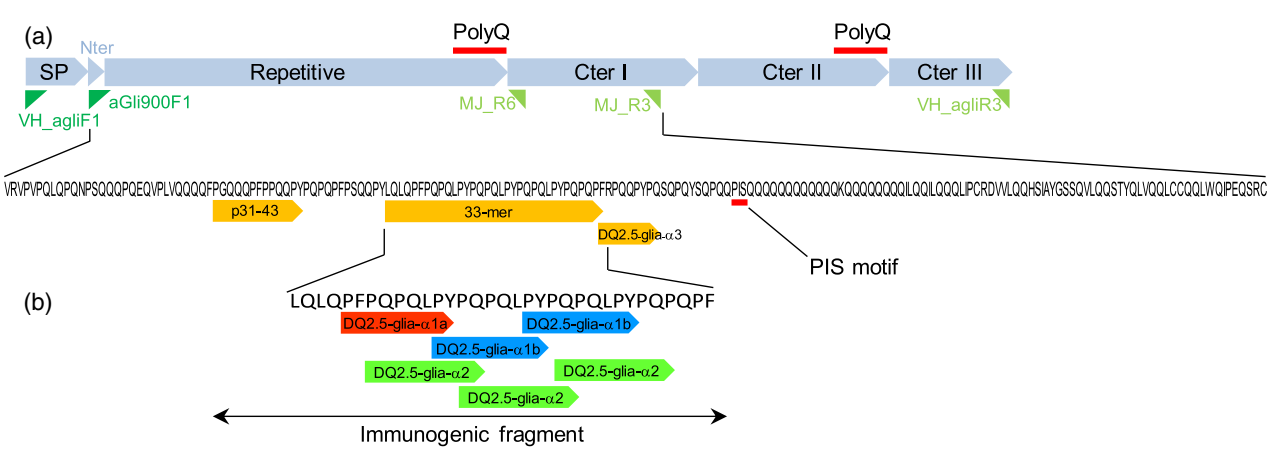


**Figure 1.** Amplicon design.
(a) In the full-length α-gliadin gene (accession number AJ133612), following parts are indicated: signal peptide (SP), transition peptide (Nter), repetitive domain and CterI, CterII and CterIII domains.
(b) Amplicon segment and the main immunotoxic region in which the peptides p31–43, 33-mer and DQ2.5-glia-α3 are indicated. Primers used are indicated by green triangles: aGli900F1, MJ_R6 and MJ_R3 to amplify the amplicon, then VH_agliF1 and VH_agliR3 to amplify the complete α-gliadin gene.

In 612 clusters (representing 65 778 out of 88 736 reads), the consensus sequence did not present frame shift or premature stop codon (PSC). These sequences were classified as putative genes and the remaining clusters as pseudogenes. The gene-derived reads ranged in average from 93.4% in *T. monococcum* to 60.0% in *T. polonicum* (Figure S2a). *T. spelta* and S-genome accessions (*Ae. speltoides*, *Ae. searsii*, and *Ae. longissima*) showed high frequency of gene-derived reads.

We also cloned and sequenced the complete sequence of α-gliadins from one accession of each species by the Sanger method (Table 1). The proportion of pseudogenes was higher in tetraploid (average of 76%) and hexaploid (average of 63%) wheats compared with their wild diploid progenitors *T. urartu* (49%), *Ae. speltoides* (36%), and *Ae. tauschii* (21%). The domesticated diploid wheat *T. monococcum* also presented a higher proportion of pseudogenes. In contrast, *Ae. searsii* (S genome) showed the lowest proportion of pseudogenes (12%). In general, the number of pseudogenes per genome is lower in diploids while increased in polyploids. In some pseudogenes, frame shift and/or PSC only appeared downstream of the PIS motif (Figure 1 and Table 1). Thus, the NGS amplicons include sequences without any mutation (real genes) and those containing mutations only downstream of PIS, which we cannot distinguish between them as mutations are out of the amplicon. Considering this observation and based on the proportion of pseudogenes in the Sanger sequencing, we estimated the proportion of genes in amplicons and obtained a result similar to the Sanger sequences (Figure S2b) except in *T. spelta* (higher than Sanger) and *T. durum* (lower than Sanger).

### Types of α-gliadins

From the alignment of consensus sequences obtained by clustering, six types of α-gliadin sequences (named 1–6) were identified which varied mainly in the pattern and number of repeats in the region corresponding to the 33-mer (Figure 2a), although some variants differed in regions other than the 33-mer. In comparison with type 2 α-gliadins, type 1 contained a deletion of PFPPQ, and type 3 a deletion of PYPQPQ. Type 4 sequences had one repeat unit (PFPPQQPYPQPQ or its variant) fewer than type 2. Type 5 also lacked one repeat unit compared with type 3, with an additional deletion of PFPPQQ or its variant. In type 6 α-gliadins, deletion of one repeat unit was observed as compared with type 3. The α-gliadin genes of S-genome diploids (*Ae. speltoides*, *Ae. longissima* and *Ae. searsii*) were mainly composed of type 3 sequences (Figure 2b). In contrast, type 1 α-gliadins predominated in diploids with A (*T. urartu* and *T. monococcum*) and D genomes (*Ae. tauschii*). Type 1 α-gliadins were rare or absent in *Ae. longissima* (Data S2), type 2 were not found in *Ae. speltoides* and *Ae. tauschii* and type 3 was absent in A-genome diploids. Type 4 α-gliadins were found only in one accession of *Ae. tauschii.* Type 1 α-gliadins also predominated in tetraploid and hexaploid wheats, followed by type 6 sequences. These trends were also observed in pseudogenes, except in *Ae. longissima* (higher proportion of type 2 α-gliadins), *Ae. tauschii* (higher proportion of type 3), *T. monococcum* (higher proportion of type 2) and *T. spelta* (dominance of type 6 and very low proportion of type 1). A lower proportion of type 2 and type 3 sequences was also notable in both genes and pseudogenes of tetraploid and hexaploid wheats (Figure 2b).

In general, similar results were obtained from Sanger sequencing (Table S2), with the differences in that type 2 was not found in *T. monococcum*, type 4 was also found in *Ae. speltoides*, *Ae. searsii*, and *T. aestivum*, and type 5 in *T. dicoccum* and *T. durum.* Type 6 was found only in *T. macha* and *T. spelta* in a lower proportion than expected. Other types not found in NGS amplicon sequencing were also found at lower frequencies.

**Table 1** Description of complete α-gliadins genes and pseudogenes sequenced by Sanger

| Genotypes | Genome | Genes | Pseudogenes | Total | Genes/ genome | Pseudogenes/ genome | Pseudogenes with mutation after PIS motif | Maximum length (bp) | Minimum length (bp) |
|---|---|---|---|---|---|---|---|---|---|
| *T. macha* | BBAADD | 38 | 78 | 116 | 12.7 | 26.0 | 40 | 970 | 846 |
| *T. spelta* | BBAADD | 42 | 73 | 115 | 14.0 | 24.3 | 27 | 957 | 657 |
| *T. aestivum* | BBAADD | 43 | 57 | 100 | 14.3 | 19.0 | 34 | 924 | 837 |
| Average | | 41.0 | 69.3 | 110.3 | 13.7 | 23.1 | | 950.3 | 780.0 |
| *T. dicoccum* | BBAA | 14 | 42 | 56 | 7.0 | 21.0 | 13 | 939 | 838 |
| *T. durum* | BBAA | 19 | 60 | 79 | 9.5 | 30.0 | 39 | 939 | 845 |
| Average | | 16.5 | 51.0 | 67.5 | 8.3 | 25.5 | | 939.0 | 841.5 |
| *T. urartu* | AA | 19 | 18 | 37 | 19.0 | 18.0 | 13 | 912 | 831 |
| *T. monococcum* | $A^mA^m$ | 6 | 18 | 24 | 6.0 | 18.0 | 7 | 885 | 852 |
| *Ae. searsii* | $S^sS^s$ | 22 | 3 | 25 | 22.0 | 3.0 | 3 | 933 | 858 |
| *Ae. speltoides* | SS | 14 | 8 | 22 | 14.0 | 8.0 | 7 | 936 | 864 |
| *Ae. tauschii* | DD | 22 | 6 | 28 | 22.0 | 6.0 | 3 | 906 | 840 |
| Average | | 16.6 | 10.6 | 27.2 | 16.6 | 10.6 | | 914.4 | 849.0 |

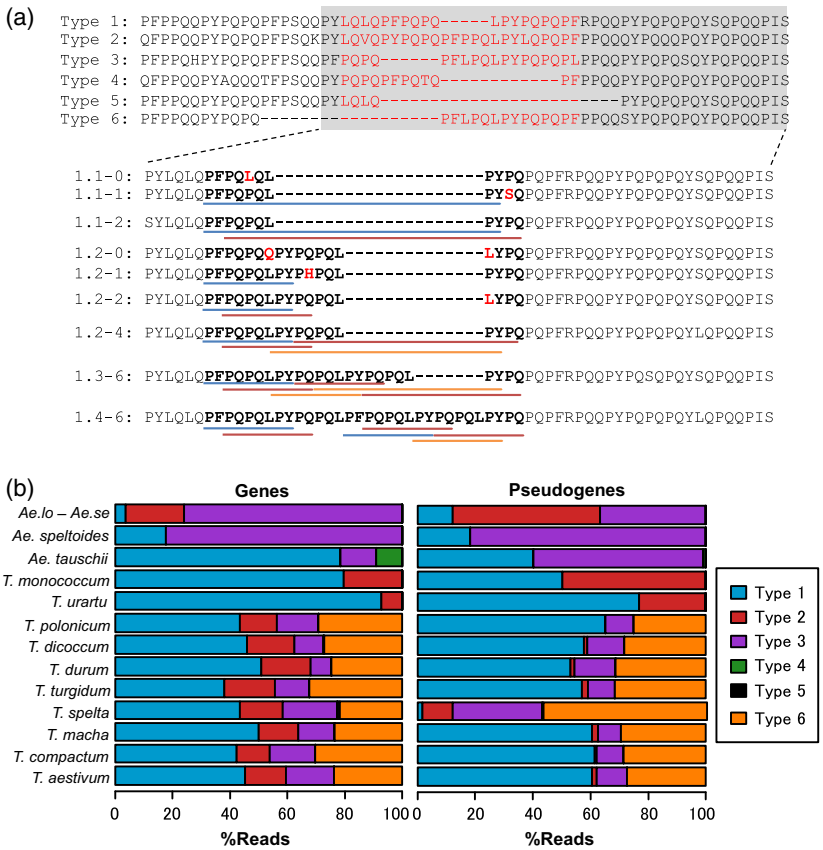


**Figure 2.** Types of α-gliadin.
(a) Alignment of six types of α-gliadin found in NGS amplicon sequencing (top panel), from the first PFPPQQ motif at the p31–43 to PIS motif. Region corresponding to the 33-mer is shaded. Alignment of type 1 subtypes (bottom panel) with different number of CD epitopes. The number after the dot indicates the subtype which represents the number of P(F/Y)PQPQL repeat unit, and that after the hyphen the number of CD epitopes (present in the region indicated in bold). DQ2.5-glia-α1a epitopes are indicated by blue underlines, DQ2.5-glia-α1b by orange and DQ2.5-glia-α2 by red. Amino acid substitutions affecting these three epitopes are indicated in red.
(b) Percentage of different α-gliadin types in reads corresponding to genes and pseudogenes in different species. *Ae.lo–Ae.se: Ae. longissima* and *Ae. searsii.*

### Subtypes of type 1 α-gliadins

The type 1 α-gliadins can be divided in subtypes according to the number of P(F/Y)PQPQL repeat units present in the region of 33-mer (Figure 2a), ranging from one (subtype 1.1) to four (subtype 1.4). Subtype 1.1 was found in almost all accessions analyzed and was the predominant class in most. This subtype can contain up to two canonical CD epitopes in the region corresponding to the 33-mer, depending on the presence of amino acid substitutions that affect one or both epitopes (Figure 2a). In hexaploids, tetraploids, and diploids with the A genome, subtype 1.1 with one epitope (subtype1.1-1) was in the majority with a lower proportion of the subtype without epitopes (subtype 1.1-0, Table 2). Hexaploids also contained subtype 1.1-2 with two epitopes. Diploids with the S genome contained only subtype 1.1-0 with no epitopes, except in accession 406 of *Ae. longissima* which had a low proportion of 1.1-1. In contrast, subtype 1.1-2 predominated in *Ae. tauschii* with a low proportion of 1.1-1 and absence of 1.1-0.

Subtypes with more than one P(F/Y)PQPQL repeat unit (1.2 to 1.4) were observed only in species with the D genome such as hexaploid wheat and *Ae. tauschii* (Figure 2a and Table 2). In general, the subtype 1.2 variants with four epitopes (1.2-4) were abundant in these species followed by 1.2-2 in hexaploids and 1.2-1 in *Ae. tauschii.* The variants 1.2-0 and 1.2-1 were absent in hexaploid wheat. The subtype 1.3 with three repeat units is equivalent to the complete 33-mer peptide and contained six epitopes. Although the variant 1.3-6, or 33-mer peptide, was found only in hexaploid wheat at lower frequency, and it was not detected in 10 hexaploid lines. In one accession of *Ae. tauschii*, subtype 1.4-6 was found but in low proportion.

### DQ2.5-glia-α3 variants

We analyzed the variants of DQ2.5-glia-α3 epitope, located downstream of 33-mer in amplicons classified as genes. Three major variants of this epitope were identified and named FR-, FP-, and FS-type according to the first two amino acids of their sequences (Figure 3a and Table S3). Almost all type 1 α-gliadins were associated with FR-type variants, with the canonical DQ2.5-glia-α3 (FRPQQPYPQ) epitope itself the most abundant in all but diploids with the S genome. In these diploids, the most abundant was the variant FRPQQPQPQ which originated from a partial deletion of PYPQ or its variants in some FR-type sequence. Most of the type 2 and type 3 α-gliadin sequences were associated with FP-type variants, the vast majority being the type FPPQQPYPQ. The FS-type variant FSPQQPYPQ was abundant in type 2 α-gliadins of *T. urartu*, and type 3 of *Ae. speltoides* but was also found in other species. A

**Table 2** Percent of Type 1 α-gliadins containing different number of epitopes

| α-Gliadin subtype[a] | *T. aestivum* | *T. compactum* | *T. macha* | *T. spelta* | *T. turgidum* | *T. durum* | *T. dicoccum* | *T. polonicum* | *T. urartu* | *T. monococcum* | *Ae. tauschii* | *Ae. speltoides* | *Ae.lo-Ae. se* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.1-0 | 4.26 | 6.02 | 2.84 | 0.00 | 6.31 | 8.00 | 6.61 | 5.85 | 26.00 | 0.22 | 0.00 | 17.73 | 3.49 |
| 1.1-1 | 24.95 | 23.01 | 25.65 | 25.03 | 29.80 | 43.02 | 39.43 | 37.61 | 66.71 | 79.32 | 0.32 | 0.00 | 0.17 |
| 1.1-2 | 4.71 | 3.86 | 6.40 | 4.39 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 35.55 | 0.00 | 0.00 |
| 1.2-0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.38 | 0.00 | 0.00 |
| 1.2-1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 13.92 | 0.00 | 0.00 |
| 1.2-2 | 2.91 | 1.87 | 4.46 | 2.95 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3.67 | 0.00 | 0.00 |
| 1.2-4 | 7.56 | 6.45 | 8.93 | 9.82 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 23.38 | 0.00 | 0.00 |
| 1.3-6 | 1.50 | 1.08 | 1.76 | 1.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1.4-6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.37 | 0.00 | 0.00 |
| Total | 45.88 | 42.29 | 50.05 | 43.39 | 36.11 | 51.03 | 46.04 | 43.46 | 92.71 | 79.54 | 78.59 | 17.73 | 3.65 |

[a]Number before hyphen is type and subtype of α-gliadin and that after hyphen indicates the number of total epitopes (DQ2.5-glia-α1a, DQ2.5-glia-α1b and DQ2.5-glia-α2) in the region corresponding to the 33-mer.
Percentage was calculated respect to the total number of α-gliadins classified as 'gene' by NGS. *Ae.lo-Ae.se: Ae. longissima* and *Ae. searsi*.

higher abundance of other variants was observed in type 3 sequences, mainly in *T. polonicum*, *T. dicoccum* and *T. turgidum*, all tetraploid wheats. The lack of type 2 α-gliadins in *Ae. speltoides* and *Ae. tauschii* and type 3 in A-genome diploids explain the absence of DQ2.5-glia-α3 variants in their respective sequences. Type 6 α-gliadins, relatively abundant in polyploid wheat, were associated with the FP-type variant FPPQQSYPQ (Table S3).

### p31–43 peptide variants

Two major variants of the peptide p31–43 (LG-, and LP-type), associated with the innate immune response induced by gluten, were found in all diploid and polyploidy species analyzed (Figure 3b and Table S4). Almost all of the type 1 α-gliadins in diploids with the A-genome and tetraploid wheats were found to be associated with the LG-type p31–43 peptide, and in *Aegilops* species with the PG-type. However, in hexaploids around 60% were associated with the LG-type, and 30% with the PG-type. In both cases, the two canonical peptides predominated in all species analyzed. In contrast, type 2 α-gliadin sequences were associated mainly with the PG-type where variants with one mismatch were seen at high frequency. In *Ae. searsii* and *Ae. longissima*, deletion of (L/P)GQQQP produced the variant LVQQQQFPPQQPY and accounted for around 90% of type 2 sequences. The PG-type also predominated in type 3 α-gliadins of many species but with a higher proportion of LG-type and other variants compared with type 2 α-gliadins. However, the LG-type was not found in type 3 α-gliadins from *T. macha* and *Aegilops* species. As in type 1 α-gliadin sequences, the canonical p31–43 peptides predominated in the LG- and PG-types except in diploids with the S genome. As in DQ2.5-glia-α3, the lack of type 2 α-gliadins in *Ae. speltoides* and *Ae. tauschii* and type 3 in A-genome diploids explain the absence of p31–43 variants in their respective sequences. In type 6 α-gliadins, the canonical PG-type peptide accounted for more than 98% of variants (Table S4).

### Abundance of total CD epitopes and their variants

To estimate the potential toxicity for each accession, the gluten T-cell epitopes restricted by HLA-DQ molecules (Sollid *et al.*, 2012) were searched for in predicted amino acid sequences allowing up to two mismatches (Figure 4 and detailed in Data S3). The abundance of each epitope was calculated by multiplying the total number of epitopes found in a given gene by the frequency of that gene in the genome. The diploids with the S genome contained very few or no canonical epitopes. The total abundance of canonical epitopes was lower in tetraploid wheats, followed by hexaploids and *T. urartu*, and *T. monococcum* and *Ae. tauschii* with the highest abundance. The abundance of CD epitope variants with one mismatch was very high with respect to that of the

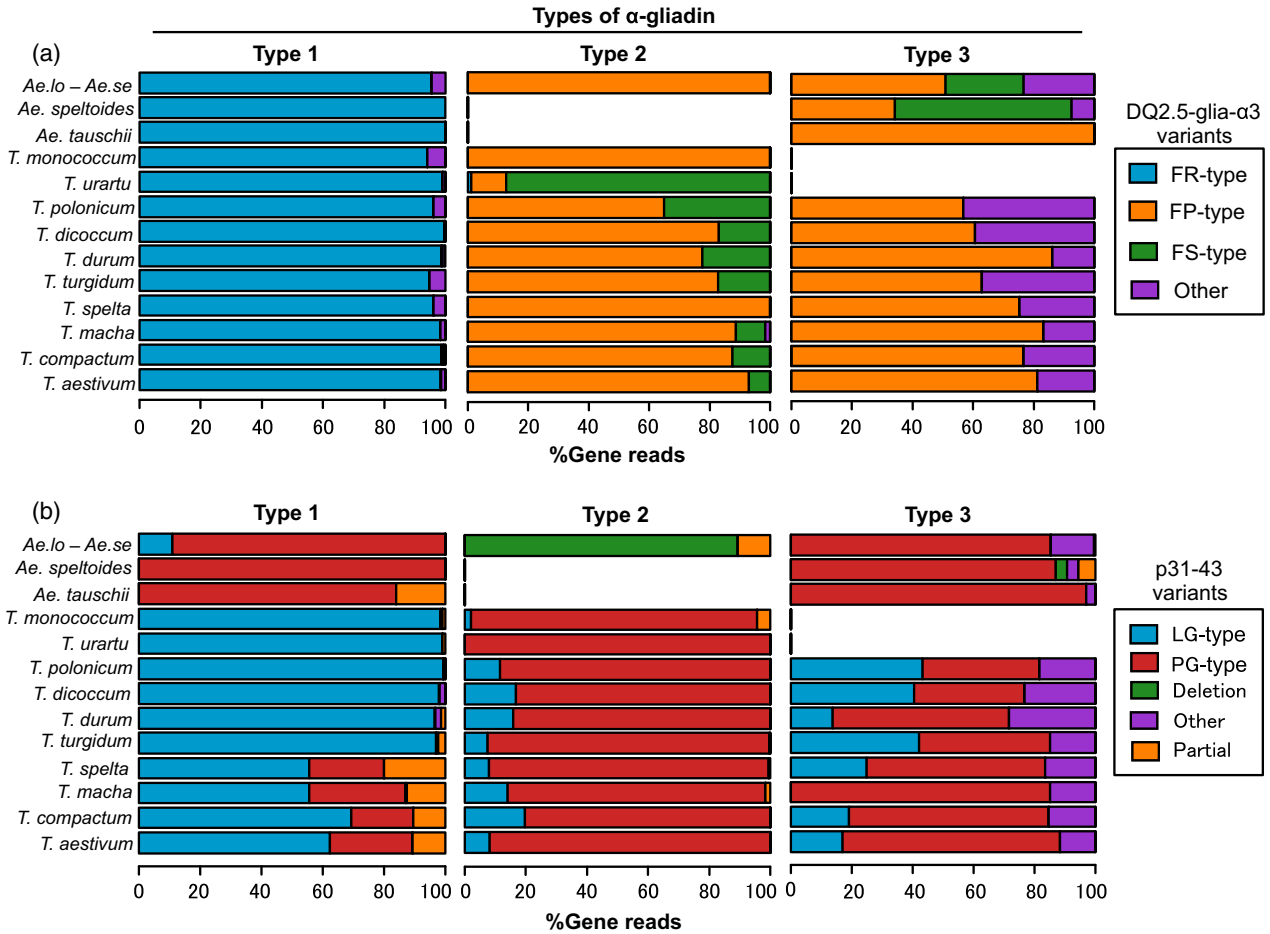


**Figure 3.** Frequency of the DQ2.5-glia-α3 epitope and p31–43 peptide in the three major α-gliadin types.
(a) DQ2.5-glia-α3 variants were grouped into four types according to the first two amino acids: FR-type, FP-type, FS-type and other.
(b) Peptide the p31–43 variants were grouped into four types: in LG-type and PG-type the first two amino acids are respectively LG and PG, variants with partial deletion and other types. Partial sequences in which the p31–43 variant could not determine are indicated as partial. Frequency was determined in reads classified as genes.

canonical sequence, in *T. monococcum*, followed by *Ae. speltoides*, *Ae. searsii* and *T. spelta*, and was low in tetraploid wheats, *T. compactum*, and *T. urartu.* In contrast to the case of canonical epitopes, the abundance of epitopes with two mismatches was higher in diploids with S genome and lower in diploids with A and D genomes. In hexaploid wheats, the abundance was slightly higher than in these latter diploids.

Although the immunogenic capacity of most of the variants with one or two mismatches has not been tested, these amino acid substitutions usually abolish or decrease the T-cell stimulation (Data S4). In addition, five out of 39 variants with one mismatch and 22 out of 39 variants with two mismatches contained proline at positions 2, 4, or 9, and/or positively charged amino acids in positions 4, 6, or 7, which may decrease toxicity (Kim *et al.*, 2004). Our results suggest that diploids with the S genome are the least toxic group, and *Ae. tauschii* and *T. monococcum* two of the most toxic species. Tetraploid wheats might also be considered one of the least toxic.

Based on the abundance of CD epitope variants and the abundance of type 1 α-gliadins with different number of epitopes, we selected putative 'reduced toxicity' accessions. Besides diploids with the S genome, two hexaploid and seven tetraploid wheats were selected (Data S3). Among *T. aestivum* genotypes, the accession THA85 presented the lowest abundance of canonical epitopes and type 1 α-gliadin sequences with more than two epitopes, and a relatively low abundance of type 1 sequences with two epitopes. Another hexaploid, the *T. compactum* accession C2, presented a lower abundance of canonical epitopes. Although tetraploid wheats presented a relatively low level of canonical epitopes compared with hexaploid wheats and lacked type 1 α-gliadins with two or more epitopes, the seven selected accessions contained a low level of canonical epitopes, a relatively

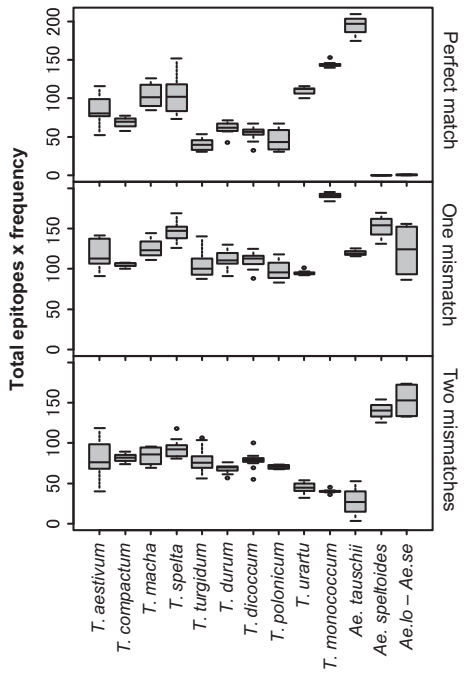


**Figure 4.** Abundance of CD epitopes in different *Triticum/Aegilops* species. Abundance of: canonical epitopes (top panel), variants with one mismatch (middle panel), and with two mismatches (bottom panel) in diploid, tetraploid and hexaploid wheat. Abundance of each epitope or its variants was calculated multiplying the total number of epitopes found in a given gene by its frequency on the genome. The *y*-axis represents the total abundance of all epitopes or their variants.

high level of type 1 α-gliadins without epitopes, and a low level of those with one epitope. These observations may serve to select varieties or accessions with lower toxicity.

### Phylogenetic analysis of α-gliadins

To study the origin of different α-gliadin types in wheat, phylogenetic analysis was performed including the α-gliadin sequences available in the NCBI nucleotide database. In total, 1036 α-gliadin genes and pseudogenes were found in *Triticum–Aegilops* species (Table S5) and 169 genes in other Triticeae species (Figure 5a). As in amplicon sequences, the type 1 α-gliadins predominated in *Triticum–Aegilops*, with subtype 1.1 accounting for 594 hits. In contrast with the NGS amplicon sequencing, type 2 α-gliadins were also found in *Ae. tauschii* and *Ae. speltoides* in a small proportion. No sequences other than type 3 were found in *T. turgidum*, which may be explained by the small number of *T. turgidum* sequences in the database. Type 6 sequences were present in *T. aestivum*, also in *Ae. speltoides*, but at lower frequency than expected by NGS amplicon sequencing. We found that type 6 α-gliadins end with the GIMSTN motif resulting from three nucleotide substitutions, compared with other types (such as types 1, 2 and 3) that end with the motif GIFGTN (Figure S3).

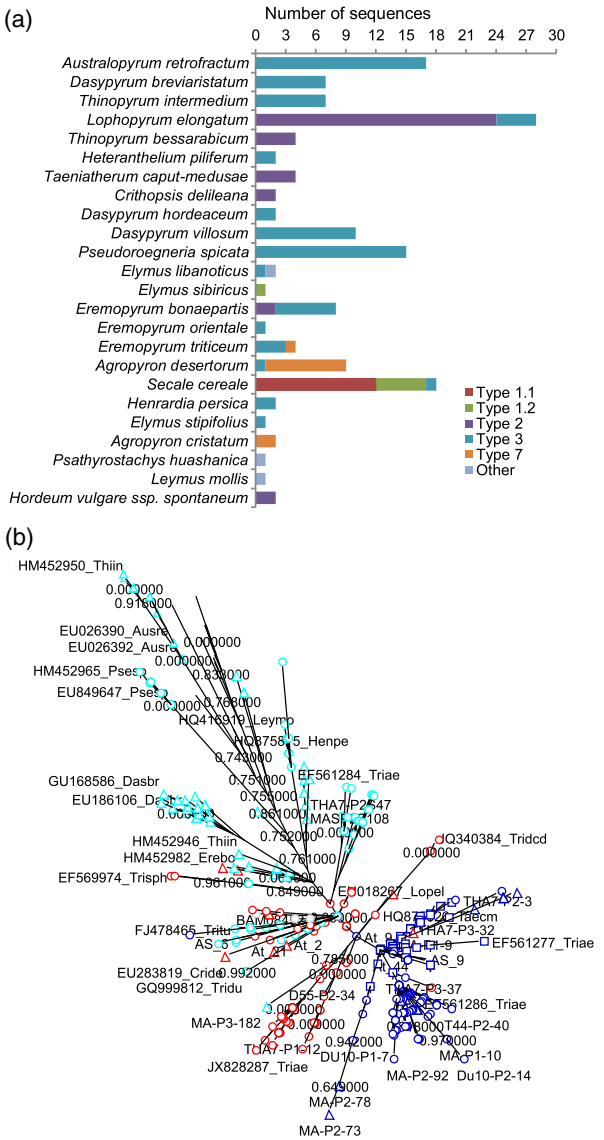


**Figure 5.** Phylogenetic analysis of α-gliadins.
(a) Number of α-gliadin sequences of other Triticeae species by types found in NCBI nucleotide database.
(b) Phylogenetic tree of α-gliadins constructed using N-terminal region. In total, 478 sequences of the tribe Triticeae with the complete N-terminal repetitive region, from Nter to PIS (Figure 1), were used. These included 207 sequences obtained by Sanger method in this study and 271 found in NCBI database (124 of *Triticum/Aegilops* and 147 of other Triticeae species). Type 1 α-gliadins are indicated in blue, type 2 in red, type 3 in cyan, type 6 in gray and other types in black. Circles indicate canonical variants and other variants are indicated by triangles. In the case of type 1, the canonical subtype 1.1 is indicated by circles, variants of subtype 1.1 by triangles and subtypes 1.2 or 1.3 by squares. Number indicates branch support estimated by SH-like approach.

Type 3 α-gliadins predominated in many other Triticeae species, which may indicate that this is the ancestral α-gliadin type (Figure 5a). Type 2 α-gliadins seem to be the main component in *Lophopyrum elongatum* (24 out of 28). Type 1 was found in *Secale cereale* (rye) another highly toxic

cereal forbidden for celiac people. One additional type of α-gliadin (type 7) was found in other Triticeae but not in *Triticum–Aegilops* species and contained a deletion of PFPPQL motif (or its variants) compared with type 2.

Phylogenetic analysis of the N-terminal region (from Nter to the PIS motif; Figure 1) of α-gliadins, using these sequences and those derived from Sanger sequencing in this study, indicated that type 3 α-gliadins are more closely related to type 2 than to type 1 (Figure 5b). We obtained a similar result from constructing a phylogenetic network using complete sequences (Figure S4), validating our classification of the α-gliadins.

## DISCUSSION

### The α-gliadin types and their toxicity

The α-gliadin genes encompass a large multigene family with highly variable and highly immunogenic N-terminal repetitive regions. Two interspersed repeat motifs are readily identified (Shewry and Tatham, 1990). Amplification and NGS sequencing of this region revealed that α-gliadin sequences differed mainly in the number of repeat blocks consisting of two interspersed motifs: PFPPQQ and PYPQPQ. Through alignment in accord with the pattern of these two motifs, we found six types of α-gliadins of which types 1, 2, 3, and 6 were the most abundant. Only the type 1 α-gliadins contain one or more of the canonical 33-mer CD epitopes such as DQ2.5-glia-α1a/b and DQ2.5-glia-α2. In addition, type 1 contains the canonical DQ2.5-glia-α3 epitope. Other main types contain variants of these epitopes, except the canonical DQ2.5-glia-α3, which was rarely found in types 2 and 6. This indicates that type 1 is the most immunogenic of the α-gliadins, with subtypes having a higher number of epitopes, such as 1.4-6 and 1.3-6 (33-mer), being the most immunogenic of the type, followed by 1.2-4. These subtypes were found only in *T. aestivum* and *Ae. tauschii*, which may explain why these species are highly immunogenic (Molberg *et al.*, 2005). In contrast, in diploids with the S genome, although type 1 sequences were present, immunogenic subtypes were absent or very low, explaining the inability to stimulate T cells in CD patients (Molberg *et al.*, 2005). In the A genome, the proline-to-serine (P/S) substitution in DQ2.5-glia-α2 eliminates its toxicity (Molberg *et al.*, 2005; Mitea *et al.*, 2010). Despite the absence of DQ2.5-glia-α1a/b and DQ2.5-glia-α2 epitopes in type 3 α-gliadins, these seem to stimulate α-II-specific T cells (recognizing DQ2.5-glia-α2 epitope) to a lesser extent in some CD patients because type 3 α-gliadins contain peptide W09 described in (Tye-Din *et al.*, 2010).

The subtype 1.3-6 (containing six overlapping epitopes) was found only in the hexaploid wheat and at low frequency, as predicted previously (Molberg *et al.*, 2005) and was absent in some accessions. We found that accessions with relatively low toxicity can be selected based on the abundance of total canonical epitopes and type 1 subtypes with different number of epitopes. Although *T. monococcum* has been found to stimulate only α-I-specific T cells (recognizing DQ2.5-glia-α1a epitope) (Molberg *et al.*, 2005), this epitope was highly abundant and therefore we classified it as one of the most immunogenic species. However, the α-gliadin proteins present in seeds of these selected accessions, in combination with the other gluten proteins, may be still high enough to stimulate T cells in CD patients.

The canonical p31–43 peptide (able to trigger innate immune response), was present in all α-gliadin types except in type 2, where mutated variants are frequent. As the toxicity of p31–43 variants has not been studied, the potential toxicity of type 2 α-gliadins cannot be predicted.

### Origin and evolution of α-gliadin types in wheat

The α-gliadins have been found not only in the genera *Triticum* and *Aegilops* but also in many other species of the tribe Triticeae (Qi *et al.*, 2013). In some Triticeae species such as barley and rye, α-gliadins are not present (Shewry and Tatham, 1990). Based on our results and those from the NCBI database, we hypothesize that after duplication of type 3 α-gliadin (probably the ancestral type), a duplication of the P(F/Y)PQPQ motif in the region corresponding to 33-mer in one of the copies was the origin of the type 2 α-gliadins in *Triticum*, *Aegilops* and other genera (Figure 6). This latter type in turn, after duplication, gave rise to type 1.1 by a deletion of the PF(L/P)PQ motif with one DQ2.5-glia-α1a epitope in *Triticum* and *Aegilops*. Thus, the common ancestor of *Triticum* and *Aegilops* possibly had these three α-gliadins types, which is in agreement with the suggestion that the duplication of certain α-gliadins took place before diploid differentiation (Kawaura *et al.*, 2012). After speciation, diploids with the A genome lost type 3 α-gliadins and *Ae. speltoides* and *Ae. tauschii* could have lost type 2. Recently, Marcussen *et al.* (2014) showed that the homoploid hybridization of diploids with A genome and B genome gave rise to *Ae. tauschii.* According to this, the species specific loss of α-gliadin type might occur after the origin of *Ae. tauschii* or this latter inherited the *Gli-2* locus of B genome. We did not find type 2 from the Sanger sequencing of *T. monococcum.* This might be due to the small number of sequenced clones available since type 2 is less abundant in this species. Because we found type 2 sequences in *Ae. speltoides* and *Ae. tauschii* in the NCBI nucleotide database, some accessions may have this α-gliadin or it may have been introduced by interspecific hybridization. Also at this stage, proline-to-leucine substitution occurred in diploids with the S genome leading to the loss of DQ2.5-glia-α1a and DQ2.5-glia-α2 epitopes, including the P/S substitution at DQ2.5-glia-α2 in A-genome diploids. Duplication of the α-gliadins may have occurred after a deletion event and the establishment of

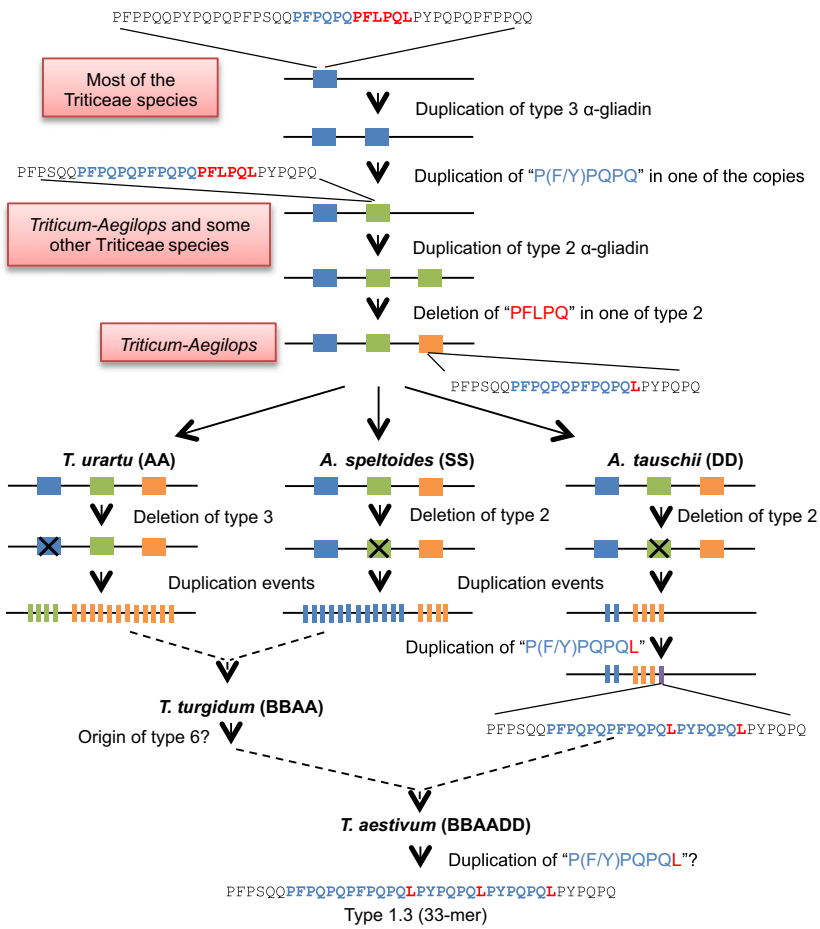


**Figure 6.** Proposed model of α-gliadin evolution in wheat.
Type 3 α-gliadins are indicated by blue boxes, type 2 by green boxes, subtype 1.1 in orange and 1.2 in purple. The repeat motifs P(F/Y)PQPQ and PF(L/P)PQ are indicated respectively by blue and red letters. Duplication of the α-gliadins might occurred after deletion events indicated by boxes with cross. Type 6 α-gliadin might be originated in tetraploid wheat and the subtype 1.3 after the allohexaploidization event. Broken lines indicate hybridization and polyploidization events.

these genome-specific variants, since some types and/or subtypes are absent in some species. In tetraploid wheat, type 1 (especially subtype 1.1-1) and type 2 α-gliadins were mainly contributed by *T. urartu* and type 3 by *Ae. speltoides.* In the D genome of *Ae. tauschii*, at least one duplication of P(F/Y)PQPQL has occurred resulting in a more toxic α-gliadin subtype. Although subtype 1.4 was found in *Ae. tauschii*, in this study and the NCBI database, no subtype 1.3 was found, suggesting that this subtype was originated after an allohexaploidization event. As type 6 α-gliadin was found only in polyploid wheat, it seems to have been originated after the hybridization between *T. urartu* and *Ae. speltoides.* However, one α-gliadin from *Ae. speltoides* corresponded to type 6, indicating that it could be originated before allotetraploidization. Due to its similarity to type 2 α-gliadin, type 6 might be originated from deletion of 'PFPSQQPYLQLQPYPQPQ' or its variant. The type 6 α-gliadin differs from other types in the presence of GIMSTN motif at the end of coding region. For this reason, the commonly used primers targeting the GIFGTN motif (van Herpen *et al.*, 2006; Mitea *et al.*, 2010; Xie *et al.*, 2010; Qi *et al.*, 2013; Li *et al.*, 2014) can rarely amplify it, explaining the low presence of this sequence found by our Sanger sequencing and the NCBI database. Similarly, the high proportion of type 1 in pseudogenes found by Sanger sequencing in *T. spelta* indicates that for some reason (such as the low efficiency of amplification using our primer sets) we could not detect it in the amplicon.

Pseudogenization of members of multigene families occurs relatively frequently (Kambere and Lane, 2007). In previous reports, approximately 50–87% of α-gliadins have been found to be pseudogenes (Anderson and Greene, 1997; Xie *et al.*, 2010), even in the diploid ancestors of polyploid wheat (van Herpen *et al.*, 2006). However, our findings suggest that after polyploidization the proportion of pseudogenes increased compared with their diploid ancestors, especially *Ae. speltoides* and *Ae. tauschii.* In hexaploid wheat, genetic redundancy created by polyploidization may allow an accelerated accumulation of mutations, leading to pseudogenization of duplicated genes (Akhunov *et al.*, 2013). This could explain the low number of pseudogenes in wild diploids as compared with polyploid wheat.

In this work six types of α-gliadins were identified in diploid and polyploid wheats but only one contains all the immunogenic peptides and epitopes, and five types of α-gli-

adins that do not contain epitopes for the 33-mer, which is the most immunogenic peptide described so far. These findings are important for reducing the incidence of CD by the breeding/selection of wheat varieties containing α-gliadins with low stimulatory capacity of T cells. Moreover, advanced genome-editing techniques (TALENs, CRISPR) will be easier to implement on the small group of α-gliadins containing only the most immunogenic subtypes of α-gliadins.

## EXPERIMENTAL PROCEDURES

### Plant materials

Ancient and modern wheat varieties were used. Thirty-four accessions of hexaploid wheat included 18 non-commercial lines of *T. aestivum* ssp. *spelta*, *T. aestivum* ssp. *macha* and *T. aestivum* ssp. *compactum*, and 16 commercial lines of *T. aestivum* ssp. *aestivum* or bread wheat. Thirty eight accession of tetraploid wheat of which 28 were non-commercial lines included *T. turgidum* ssp. *turgidum*, *T. turgidum* ssp. *dicoccum and T. turgidum* ssp. *polonicum*; 10 were commercial lines of *T. turgidum* ssp. *durum* or durum wheat. Finally 24 accessions of diploid wheat and *Aegilops* included *T. monococcum* ssp. *monococcum*, *T. urartu*, *Ae. speltoides*, *Ae. longissima*, *Ae. searsii* and *Ae. tauschii*. All accessions are detailed in the Data S1. Plants were grown during 2011–2012 in field conditions. The fertilization was done according to agricultural practices in the region.

The non-commercial and commercial lines were selected based on the polymorphism (different allelic variants) observed in the A-PAGE gel, all non-commercial accessions selected were different in the α-gliadin fraction. The 96 genotypes are therefore a representative of the diploid, tetraploid and hexaploid genomes. The commercial lines included in the study were cultivated in Spain and some of them are inbred materials to develop new cultivars in the region.

### Genomic DNA extraction and 454 amplicon sequencing

Leaf tissue of individual plants was harvested, frozen in liquid nitrogen, and stored at -80°C until DNA extraction. Genomic DNA was extracted using the CTAB method (Murray and Thompson, 1980) with minor modifications.

For amplification of α-gliadins, gene specific primers were designed based on the sequences present in the NCBI database using Primer3Plus software (Untergasser *et al.*, 2007), and then *in silico* approach was applied to check their specificity. For PCR conditions, see Methods S1. Preparation of the 454 amplicon library and sequencing was carried out according to the manufacturer's instructions (GS FLX Titanium/emPCR kit XLR70 for the LibA and emPCR kit XL+ for the LibL) at the Unidad de Genómica Cantoblanco of Fundación Parque Científico de Madrid (FPCM, Spain).

### Amplicon sequence clustering

The α-gliadin amplicon sequences (217 118 total reads comprising 107 Mbp and average of 491 bp) were preprocessed to remove adaptors and barcode sequences, trim nucleotides with quality values <20, and reads with length <100 bp were eliminated using seq_crumbs and ngs_backbone cleaning software (http://bioinf.comav.upv.es/). Then, all reads (200 340 cleaned reads comprising 49 Mbp) were mapped to the reference α-gliadin sequence (accession number AJ133612) using Geneious version 7.0.6 (Biomatters Ltd., Auckland, New Zealand; available at http://www.geneious.com/) for correct orientation (5′–3′) and to remove polyQ and downstream sequences (Figure 1). Sequence AJ133612 was used as reference because the complete 33-mer peptide was described in this α-gliadin gene (Arentz-Hansen *et al.*, 2000). In this step, reads shorter than 60 bp were discarded for further analyses. The maximum read length was 384 bp with an average of 224.9 bp and median of 243 bp. The number of reads per accession ranged from 739 to 5311 with an average of 2086.9.

The cleaned reads were clustered using USEARCH version 7.0.1090 (Edgar, 2010) with cluster_fast mode, 99% homology and extracting a consensus sequence for each cluster because the consensus tends to correct the sequencing errors. In total, 56 093 unique clusters were obtained; the maximum cluster size (number of reads per cluster) was 10 501, the number of singletons (clusters with only one read) was 38 416 and in average contained 3.6 reads per cluster. To extract the high-confidence sequence variants for each accession, accessions with less than five reads in a given cluster were removed from that cluster, thus each cluster contained five or more reads per accession. After this filtering, we obtained 999 unique clusters with an average of 88.82 reads per cluster and 49 clusters per accession (Table S1). The number of total reads in the 999 clusters was 88 736 (average of 924 reads per accession).

### Multiple alignment and classification of sequences

Manual alignment of the consensus sequence of all clusters was performed using the software Geneious according to the pattern of two interspersed motifs 'PFPPQQ' and 'PYPQPQ'. First, the Nter to p31–43 segments and regions close to PIS motif (Figure 1) were located and aligned. Next, the remaining repetitive region was aligned considering the frequent Q to P, and P to S or L substitutions. Two of the characteristic motifs are 'LQL' at the beginning and 'LPY' at the middle of 33-mer peptide.

We classified the consensus sequences in genes or pseudogenes based on the absence or presence of PSC and frame shifts, respectively. Then, according to the number of the two repeat motifs at the region corresponding to 33-mer we classified the genes and pseudogenes into six types of α-gliadins. If the sequence variation was found outside the 33-mer region, we considered as variant of a given α-gliadin type. Type 1 sequences included those differing in the number of repeat motif 'P(F/Y)PQPQL', which can be one up to four repeats. These variants were named subtypes e.g. subtype 1.1, where the first number indicates the α-gliadin type and the number after dot the subtype which coincides with the number of repeats.

### PCR amplification of complete α-gliadin gene and sequencing by Sanger

One accession of each subspecies was choose to sequence by Sanger method, in all cases we chose each one that present higher content of gliadin according to GlutenTox ELISA-Sandwich (Biomedal S.L., Sevilla, Spain). For PCR conditions, see Methods S1. The full-length DNA sequences were ligated into pGEM-T Easy vector (Promega, Madison, WI, USA) and cloned into *Escherichia coli* DH5α cells. We sequenced 48 clones of full-length α-gliadin genes per haploid genome, i.e. 48, 96 and 144 clones of diploid, tetraploid and hexaploid species, respectively. In total, 912 clones were sequenced.

Seqman from DNASTAR (Madison, WI, USA) was used to assemble the sequences from the Sanger sequencing. The N-terminal repetitive region was manually aligned as described for amplicon sequences, and the remaining C-terminal region was aligned using the CLUSTALW algorithm (Thompson *et al.*, 1994) using Geneious.

 

### Estimation of genes in NGS amplicon sequences

From pseudogenes found in Sanger sequencing, the number of sequences with mutation (PSC and/or frame shift) at the N-terminal repetitive region (from signal peptide to PIS motif), downstream of PIS region, and at both regions were determined. Because gene of amplicons includes sequences without any mutation (the real genes) and those containing mutation only downstream of PIS, the percent of reads corresponding to genes were multiplied by correction factor to estimate the real percentage of genes. The correction factor was determined dividing the number of clones (sequenced by Sanger method) classified as gene by the sum of clones classified as genes and pseudogenes with mutation only downstream of PIS.

### Search of CD epitopes and peptide p31–43 variants

From alignment of consensus sequences, regions corresponding to p31–43, 33-mer and DQ2.5-glia-α3 were extracted. At the region corresponding to 33-mer, we searched the presence of three canonical CD epitopes: DQ2.5-glia-α1a (PFPQPQLPY), DQ2.5-glia-α1b (PYPQPQLPY), DQ2.5-glia-α2 (PQPQLPYPQ) (Sollid *et al.*, 2012). Only the type 1 α-gliadins contained these epitopes. Thus, we represented e.g. subtype 1.1-1 where the number after hyphen is the number of canonical CD epitope contained in this subtype. From DQ2.5-glia-α3 region, we extracted peptides matching to the canonical epitope 'FRPQQPYPQ' and its variants. Taking the first two amino acid sequence of this epitope, we divided into FR-type, FP-type (such as 'FPPQQPYPQ'), FS-type (such as FSPQQPYPQ) and other types. From the p31–43 region, we also extracted sequences matching to the canonical peptides 'LGQQQPFPP QQPY' and 'PGQQQPFPPQQPY' which are immunogenic in CD patients (Maiuri *et al.*, 1996b, 2003). As for DQ2.5-glia-α3 epitope, p31–43 variants were classified according to the first two amino acid sequences such as LG-type (e.g. LGQQQPFPPQQPY), PG-type (e.g. PGQQQPFPPQQPY) and deletion types derived from partial deletion of this segment. There were partial sequences in which p31–43 variants were not able to determine, and we classified as 'partial' in Figure 3(b).

### Abundance of total CD epitopes and its variants

From amplicon consensus sequences classified as genes, amino acid sequence were predicted and BLASTP searched against CD epitopes database (Sollid *et al.*, 2012) using blastall parameters –F F, –W 2 and *E*-value <1. Only the blast hits with up to two mismatches and having nine amino acid length were used for analysis. Because CD epitopes in the database (Sollid *et al.*, 2012) are described in deamidated form, the conversion of glutamine (Q) residues to glutamate by tTG were considered on query sequences of the BLASP results. tTG recognizes Q residues in Q-X-P or Q-X-X-(phenylalanine, tyrosine, tryptophan, methionine, leucine, isoleucine or valine) sequence but not Q-P and Q-X-X-P, where X is any amino acid other than P (Fleckenstein *et al.*, 2002; *Vader et al.,* 2002). Abundance of each individual epitope was calculated multiplying the total number of epitopes found in a given gene by the frequency of that gene in the genome. Total abundance of canonical epitopes, variants with one or two mismatches was calculated as a sum of abundance of the canonical epitopes or its variants.

### Phylogenetic analysis

The α-gliadin genes and pseudogenes of *Triticum/Aegilops* species and genes of other Triticeae species were downloaded from NCBI nucleotide database. Sequences lacking N-terminal repetitive region were discarded because α-gliadin type cannot be determined. The N-terminal repetitive region of these sequences, in addition to those obtained by Sanger method in this study, was also manually aligned as in amplicons. The remaining C-terminal region, excluding the two polyQ regions, was aligned using the ClustalW algorithm (Thompson *et al.*, 1994) implemented in Geneious. Aligned N-terminal nucleotide sequences were translated and the protein sequences were subjected to phylogenetic analysis using PHYML version 20130513 (Guindon *et al.*, 2010) with following parameters: Jones–Taylor–Thornton (JTT) amino acid substitution model, subtree pruning and regrafting (SPR) approach for tree topology search, maximum-likelihood estimates for proportion of invariable sites and gamma shape parameter, optimization of tree topology, branch length and substitution rate parameters, and branch support was estimated with non-parametric Shimodaira-Hasegawa (SH)-like procedure. Phylogenetic network was constructed using the translated full-length coding region of α-gliadins (without polyQ regions) with SPLITSTREE 4.13.1 (Huson and Bryant, 2006) and branch support was estimated by 1000 bootstrap replicates.

### Data availability

All 454 amplicon sequences were submitted to NCBI sequence read archive under accession number SRP051484. Sanger sequences were deposited in GenBank under the accession numbers KP405234 to KP405835.

## ACKNOWLEDGEMENTS

The Spanish Ministry of Economy and Competitiveness (Project AGL2013-48946-C3-1-R), the European Regional Development Fund (FEDER), and Junta de Andalucía (Project P11-AGR-7920) supported this work. The technical assistance of Ana García is acknowledged.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

**Figure S1.** Polyploid evolution of wheat.

**Figure S2.** Percent of reads classified as gene in different subspecies.

**Figure S3.** Alignment of the C-terminal end motif of α-gliadin.

**Figure S4.** Phylogenetic network of α-gliadins constructed using complete coding sequence.

**Table S1.** Statistics after sequence clustering at 99% identity.

**Table S2.** Number of clones classified in each α-gliadin types in genes and pseudogenes found by Sanger sequencing.

**Table S3.** Abundance (%) of the main DQ2.5-glia-α3 variants in each type of alpha-gliadins.

**Table S4.** Abundance (%) of the main p31-43 variants in each type of α-gliadins.

**Table S5.** Number of α-gliadin genes and pseudogenes of *Triticum-Aegilops* species by types found in NCBI nucleotide database.

**Data S1.** Accessions and origin of *Triticum* and *Aegilops* used in this study.

**Data S2.** Abundance of different alpha-gliadin types in genes and pseudogenes of 96 accessions.

**Data S3.** Total abundance of CD epitope variants and type 1 alpha-gliadins with different number of epitopes in each individual.

**Data S4.** List of CD epitopes variants with one mismatch found.

**Methods S1.** PCR conditions for 454 amplicon and Sanger sequencing.

## REFERENCES


**Akhunov, E.D., Sehgal, S., Liang, H. *et al.*** (2013) Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. *Plant Physiol.* **161**, 252–265.

**Anderson, O.D. and Greene, F.C.** (1997) The a-gliadin gene family. II. DNA and protein sequence variation, subfamily structure, and origins of pseudogenes. *Theor. Appl. Genet.* **95**, 59–65.

**Anderson, O.D., Litts, J.C. and Greene, F.C.** (1997) The a-gliadin gene family. I. Characterization of ten new wheat a-gliadin genomic clones, evidence for limited sequence conservation of flanking DNA, and Southern analysis of the gene family. *Theor. Appl. Genet.* **95**, 50–58.

**Arentz-Hansen, E.H., McAdam, S.N., Molberg, Ø., Kristiansen, C. and Sollid, L. M.** (2000) Production of a panel of recombinant gliadins for the characterisation of T cell reactivity in coeliac disease. *Gut*, **46**, 46–51.

**Edgar, R.C.** (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

**Fleckenstein, B., Molberg, O., Qiao, S.W., Schmid, D.G., von der Mulbe, F., Elgstoen, K., Jung, G. and Sollid, L.M.** (2002) Gliadin T cell epitope selection by tissue transglutaminase in celiac disease. Role of enzyme specificity and pH influence on the transamidation versus deamidation process. *J. Biol. Chem.* **277**, 34109–34116.

**Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O.** (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321.

**Harberd, N.P., Bartels, D. and Thompson, R.D.** (1985) Analysis of the gliadin multigene loci in bread wheat using nullisomic-tetrasomic lines. *Mol. Gen. Genet.* **98**, 234–242.

**Huson, D.H. and Bryant, D.** (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267.

**Jordan, G. and Goldman, N.** (2012) The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* **29**, 1125–1139.

**Kambere, M.B. and Lane, R.P.** (2007) Co-regulation of a large and rapidly evolving repertoire of odorant receptor genes. *BMC Neurosci.* **8**(Suppl 3), S2.

**Karell, K., Louka, A.S., Moodie, S.J., Ascher, H., Clot, F., Greco, L., Ciclitira, P.J., Sollid, L.M. and Partanen, J.** (2003) HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the european genetics cluster on celiac disease. *Hum. Immunol.* **64**, 469–477.

**Kawaura, K., Wu, J., Matsumoto, T., Kanamori, H., Katagiri, S. and Ogihara, Y.** (2012) Genome change in wheat observed through the structure and expression of alpha/beta-gliadin genes. *Funct. Integr. Genomics*, **12**, 341–355.

**Kim, C.Y., Quarsten, H., Bergseng, E., Khosla, C. and Sollid, L.M.** (2004) Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. *Proc. Natl Acad. Sci. USA*, **101**, 4175–4179.

**Li, Y., Xin, R., Zhang, D. and Li, S.** (2014) Molecular characterization of α-gliadin genes from common wheat cultivar Zhengmai 004 and their role in quality and celiac disease. *Crop J.* **2**, 10–21.

**Loytynoja, A. and Goldman, N.** (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl Acad. Sci. USA*, **102**, 10557–10562.

**Maiuri, L., Ciacci, C., Ricciardelli, I., Vacca, L., Raia, V., Auricchio, S., Picard, J., Osman, M., Quaratino, S. and Londei, M.** (2003) Association between innate response to gliadin and activation of pathogenic T cells in coeliac disease. *Lancet*, **362**, 30–37.

**Maiuri, L., Picarelli, A., Boirivant, M., Coletta, S., Mazzilli, M.C., De Vincenzi, M., Londei, M. and Auricchio, S.** (1996a) Definition of the initial immunologic modifications upon in vitro gliadin challenge in the small intestine of celiac patients. *Gastroenterology*, **110**, 1368–1378.

**Maiuri, L., Troncone, R., Mayer, M., Coletta, S., Picarelli, A., De Vincenzi, M., Pavone, V. and Auricchio, S.** (1996b) In vitro activities of A-gliadin-related synthetic peptides: damaging effect on the atrophic coeliac mucosa and activation of mucosal immune response in the treated coeliac mucosa. *Scand. J. Gastroenterol.* **31**, 247–253.

**Marcussen, T., Sandve, S.R., Heier, L., Spannagl, M., Pfeifer, M., The International Wheat Genome Sequencing Consortium, Jakobsen, K.S., Wulff, B.B.H., Steuernagel, B., Mayer, K.F.X. and Olsen, O-A.** (2014) Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, **345**, 1250092.

**Mitea, C., Salentijn, E.M.J., van Veelen, P. *et al.*** (2010) A universal approach to eliminate antigenic properties of alpha-gliadin peptides in celiac disease. *PLoS One*, **5**, 1–9.

**Molberg, Ø., Uhlen, A.K., Jensen, T., Flæte, N.S., Fleckenstein, B., Arentz-Hansen, H., Raki, M., Lundin, K.E.A. and Sollid, L.M.** (2005) Mapping of gluten T-cell epitopes in the bread wheat ancestors: Implications for celiac disease. *Gastroenterology*, **128**, 393–401.

**Murray, M.G. and Thompson, W.F.** (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**, 4321–4325.

**Payne, P.I.** (1989) Genetics of wheat sorage proteins and the effect of allelic variation on bread-making quality. *Annu. Rev. Plant Physiol.* **38**, 141–153.

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

**Qi, P.F., Chen, Q., Ouellet, T., Wang, Z., Le, C.X., Wei, Y.M., Lan, X.J. and Zheng, Y.L.** (2013) The molecular diversity of alpha-gliadin genes in the tribe Triticeae. *Genetica*, **141**, 303–310.

**Rewers, M.** (2005) Epidemiology of celiac disease: What are the prevalence, incidence, and progression of celiac disease? *Gastroenterology*, **128**, S47–S51.

**Salentijn, E.M.J., Esselink, D.G., Goryunova, S.V., van der Meer, I.M., Gilissen, L.J.W.J. and Smulders, M.J.M.** (2013) Quantitative and qualitative difference in celiac disease epitopes among durum wheat varieties identified through deep RNA-amplicon sequencing. *BMC Genom.* **14**, 905.

**Sapone, A., Lammers, K.M., Casolaro, V. *et al.*** (2011) Divergence of gut permeability and mucosal immune gene expression in two gluten-associated conditions: celiac disease and gluten sensitivity. *BMC Med.* **9**, 23.

**Shan, L., Molberg, O., Parrot, I., Hausch, F., Filiz, F., Gray, G.M., Sollid, L.M. and Khosla, C.** (2002) Structural basis for gluten intolerance in celiac sprue. *Science*, **297**, 2275–2279.

**Shewry, P.R.** (2009) Wheat. *J. Exp. Bot.* **60**, 1537–1553.

**Shewry, P.R. and Tatham, A.S.** (1990) The prolamin storage proteins of cereal seeds: structure and evolution. *Biochem. J.* **267**, 1–12.

**Sollid, L.M., Qiao, S.W., Anderson, R.P., Gianfrani, C. and Koning, F.** (2012) Nomenclature and listing of celiac disease relevant gluten T-cell epitopes restricted by HLA-DQ molecules. *Immunogenetics*, **64**, 455–460.

**Thompson, J.D., Higgins, D.G. and Gibson, T.J.** (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.

**Tye-Din, J.A., Stewart, J.A., Dromey, J.A. *et al.*** (2010) Comprehensive, quantitative mapping of T cell epitopes in gluten in celiac disease. *Sci. Transl. Med.* **2**, 41ra51.

**Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R. and Leunissen, J.A.** (2007) Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* **35**, W71–74.

**Vader, W., Kooy, Y., van Veelen, P., de Ru, A., Harris, D., Benckhuijsen, W., Peña, S., Mearin, L., Drijfhout, J.W. and Koning, F.** (2002) The gluten response in children with celiac disease is directed toward multiple gliadin and glutenin peptides. *Gastroenterology*, **122**, 1729–1737.

**van Herpen, T.W., Goryunova, S.V., van der Schoot, J. *et al.*** (2006) Alpha-gliadin genes from the A, B, and D genomes of wheat contain different sets of celiac disease epitopes. *BMC Genom.* **7**, 1.

**Wieser, H. and Koehler, P.** (2008) The biochemical basis of celiac disease. *Cereal Chem.* **85**, 1–13.

**Xie, Z., Wang, C., Wang, K., Wang, S., Li, X., Zhang, Z., Ma, W. and Yan, Y.** (2010) Molecular characterization of the celiac disease epitope domains in alpha-gliadin genes in *Aegilops tauschii* and hexaploid wheats (*Triticum aestivum* L.). *Theor. Appl. Genet.* **121**, 1239–1251.