Molecular phylogenetic analysis of amylase trypsin inhibitors (ATIs) from a selection of ancient and modern wheat

Molecular phylogenetic analysis of amylase trypsin inhibitors (ATIs) from a selection of ancient and modern wheat

Emanuela Simonetti[a,*], Sara Bosi[a], Lorenzo Negri[a], Loredana Baffoni[a], Alberto Masoni[b], Ilaria Marotti[a], Stefano Benedettelli[b], Giovanni Dinelli[a]

a Department of Agricultural and Food Sciences, Alma Mater Studiorum – University of Bologna, 40127 Bologna, Italy

emanuela.simonetti2@unibo.it; sara.bosi@unibo.it; lorenzo.negri4@unibo.it; loredana.baffoni@unibo.it; ilaria.marotti@unibo.it; giovanni.dinelli@unibo.it

b Department of Agriculture, Food, Environment and Forestry (DAGRI), University of Florence, 50144 Firenze, Italy.

alberto.masoni@unifi.it; stefano.benedettelli@unifi.it

* Corresponding author (phone: +39 051 2096672/73/69; fax: +39 051 2096241)

Declarations of interest: none.

Keywords

Wheat Amylase-Trypsin Inhibitors (ATIs); Ancient Wheat; Haplotype; PCA analysis, PCR

Author Contributions: Conceptualization, E.S. and G.D.; data curation and investigation, E.S. and S.B.; methodology, E.S., S.B., L.N., I.M. and L.B.; supervision S.B., A.M., St.Be. and G.D.; validation: S.B. and G.D.; writing original draft E.S.; writing- review & editing S.B., L.N., L.B., I.M., A.M., St.Be. and G.D; funding acquisition G.D. All authors have read and agreed to the published version of the manuscript.

Abstract

Wheat amylase-trypsin inhibitors (ATIs) are a family of wheat proteins, which play an important role in plant defence against pest attacks. Recently, ATIs have been identified as major stimulators of human innate immune cells leading to cause Non-coeliac Wheat Sensitivity. Information about ATI sequence differences among wheat species is scarce, especially considering ancient wheat genotypes. In this study, ten selected wheat accessions with different ploidy level and year of release were used for gene sequencing of four representative alpha-amylase/trypsin inhibitor genes (WMAI, WDAI, WTAI-CM3 and CMx genes). The phylogenetic analysis and the PCA analysis performed on the deduced amino acid sequences of ATI genes evidenced that the ten wheat genotypes can be differentiated on the basis of their ploidy level, but not with respect to ancient or recently developed wheat genotypes. The haplotype analysis based on Nei's genetic distances, beside confirming these results, also allowed the separation of the hulled from the naked wheat genotypes. The genetic sequence differences highlighted in this study among the ten genotypes can be the basis for further studies aimed at identifying proinflammatory sequences in ATI genes. This is the first study analyzing the ATI genetic sequences of a set of ancient and modern wheat genotypes with a different ploidy level.

1. Introduction

Wheat amylase-trypsin inhibitors (ATIs) represent ~2–4% of total wheat protein and are an important family of low molecular weight, water-soluble proteins located in the endosperm of the grain kernel where they play a crucial role in plant defence against pest attacks (Carbonero and García-Olmedo, 1999). ATIs are able to inhibit enzymes of common parasites such as mealworms and mealy bugs from digesting wheat starch and protein. Besides having a protective role, they act as storage proteins (Altenbach et al., 2011).

ATIs are classified into four groups: three groups of alpha-amylase inhibitors differentiated according to their degree of aggregation (monomeric, homodimeric and heterotetrameric forms) and one group of trypsin inhibitors (Carbonero and García-Olmedo, 1999). The three groups of alpha-amylase inhibitors include the wheat monomeric amylase inhibitors (WMAI), often called 0.28 proteins; proteins that form the wheat homodimeric amylase inhibitors group (WDAI), sometimes called 0.19 and 0.53 proteins and the proteins of the wheat heterotetrameric amylase inhibitors (WTAI). The WMAI inhibitors are encoded by genes on the short arm of chromosome 6 of *T. aestivum*, while the genes for WDAI are located in the short arm of chromosome 3 (Carbonero and García-Olmedo, 1999), in particular on 3BS and 3DS chromosomes, though not much is known about homologous common wheat chromosome 3AS (Pandey et al., 2016). The tetrameric inhibitors are often called CM proteins because of their solubility in chloroform/methanol and in modern hexaploid wheat five sub-units have been identified: CM1, CM2, CM3, CM16 and CM17. CM proteins are generally composed by one copy of either CM1 or CM2 sub-unit, encoded by genes on chromosomes 7D or 7B, plus one copy of either CM16 or CM17 sub-unit, encoded by genes on chromosomes 4B and 4D, plus two copies of CM3 sub-unit, also encoded on chromosomes 4B and 4D. WMAI, WDAI and CM proteins have a molecular weight ranging from 13.0 to 13.5 kDa at the subunit level, with the exception of CM3 which shows a mass of about 15.5 kDa (Geisslitz, 2021). All these alpha-amylase inhibitors are active with different specificity against insect, mite and mammalian alpha-amylases, but are not active against cereal enzymes. The WTAI inhibitors are also able to inhibit papain, bovine trypsin and subtisilin. The putative wheat trypsin inhibitors are referred to as CMx proteins, are encoded by genes on the group 4 chromosomes and are monomeric (Altenbach, 2011; Carbonero and García-Olmedo, 1999).

Altenbach et al. (2011) identified 19 different alpha-amylase/protease inhibitors of ~120- to 150-amino acids expressed in the modern wheat genome. The main ATI species in wheat are 0.19 (WDAI) and CM3 (WTAI).

All ATIs are characterized by high homology in their amino acid sequence and by an analogous compact secondary structure showing five or less (often four) intrachain disulphide bridges and a main body of the molecule built around four α-helices arranged in an "up and down" pattern linked together by loop segments (Carbonero and García-Olmedo, 1999).

ATIs are of high interest because they have an important impact on human health. Some ATI proteins are involved in wheat allergies (Tatham and Shewry, 2008) and, recently, ATIs have been identified to play a role in the onset of celiac disease (Huebener et al., 2015) and Non-celiac Wheat Sensitivity (Junker et al., 2012). Several *in vitro* and *in vivo* studies identified ATIs (especially CM3 and 0.19), but not gluten, as a major nutritional trigger of human and murine innate immunity on wheat by the activation of the toll-like receptor 4 (TLR4) complex (Geisslitz, 2021; Junker et al., 2012; Zevallos, 2017).

Thus far, there is little information in literature about the diversity of ATIs between different types of wheat, especially between so-called "ancient" and modern wheat genotypes. Recently, Zevallos et al. (2017) found that modern hexaploid wheat showed higher ATI inflammatory activity than some ancient variants like diploid (einkorn) and tetraploid wheat (emmer, KAMUT® khorasan wheat) or older hexaploid variants like spelt. Another recent paper (Gélinas and Gagnon, 2018) determined the alpha-amylase inhibitory activity against human α-amylase in several different wheat cultivars and showed that the inhibitory potential did not vary with respect to ancient or recently developed wheat cultivars. To the same conclusion came Call et al. (2020) who analysed the ATI concentrations and the trypsin inhibitor activity in a set of different *Triticum* species and EL Hassouni et al. (2021) who determined the content of eight ATI proteins as well as of the total ATI of 149 European old and modern bread wheat cultivars.

However, information about ATI genetic sequences is available only for a few types of wheat, especially for common wheat (*T. aestivum* L. subsp. *aestivum*) and for wild emmer wheat (*T. turgidum* L. subsp. *dicoccoides*) (Pandey et al., 2016; Wang et al., 2010, 2008, 2007), while ATI genetic sequences of some other types of wheat like spelt (*T. aestivum* L. subsp. *spelta*), emmer (*T. turgidum* L. subsp. *dicoccum*) and khorasan wheat (*T. turgidum* L. subsp. *turanicum*) have never been studied. Moreover, gene sequences of some ATIs like WTAI and CMx are very scarce (Liu and Wang, 2012). Lastly, little is known of the genetic diversity of ATIs between genotypes, in particular between ancient and modern wheat genotypes (Geisslitz et al., 2021).

In the present study, ten wheat accessions were selected to cover different *Triticum* species with a different genome composition (diploid, tetraploid, hexaploid) and with a different year of release to determine the genetic sequences of four representative alpha-amylase/trypsin inhibitor genes

(WMAI, WDAI, WTAI-CM3 and CMx genes). The genetic sequence differences among the wheat genotypes were investigated and discussed.

2. Material and methods

2.1 Plant material

In the present study, ten selected accessions belonging to different *Triticum* species i.e. *T. monococcum* L. subsp. *monococcum* (2n = 2x = 14 chromosomes, AA genome), *T. turgidum* L. subsp. *dicoccum*, *T. turgidum* L. subsp. *turanicum*, *T. turgidum* L. subsp. *durum* (all 2n = 4x = 28 chromosomes, AABB genomes), *T. aestivum* L. subsp. *aestivum*, *T. aestivum* L. subsp. *spelta* (all 2n = 6x = 42 chromosomes, AABBDD genomes) were used for the characterization of ATI gene sequences (Table 1). These genotypes were selected to cover different wheat species with a different genome composition (diploid, tetraploid, hexaploid) and with a different year of release. The pedigrees of heritage and modern wheat varieties were retrieved in GRIS – Genetic Resources Information System for Wheat and Triticale (http://wheatpedigree.net) and are shown in Table 1.

2.2 DNA isolation and PCR amplification

Ten seeds for each wheat genotype were grown in sprouters at room temperature in daylight for seven days. The leaf tissues were sampled at the four-leaf stage from ten different plants per genotype, immediately frozen in liquid nitrogen and ground in a mortar with a pestle. Total DNA extraction was performed with the Nucleospin® Plant II kit (Macherey Nagel, Düren, Germany) following manufacturer's instructions. DNA quality and quantity was measured by NanoPhotometer® P-Class (Implen GmbH, München, Germany).

The primers used to amplify WMAI, WDAI, WTAI-CM3 and CMx genes are listed in Table 2. Primers F1, F2, F3, F4, R1 and R2 were previously used by Wang et al. (2008), primers WDAIfor and WDAIrev were previously used by Wang et al. (2007), the other primers were designed *in house* with the software Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/) on the conserved sequences of the coding regions of genes obtained from the GenBank database.

WMAI 0.28 genes were amplified in common wheat samples using F2 and R1 primers and in einkorn, spelt and Peliss genotypes using Mfor3 and Mrev2. Emmer, turanicum and Alzada samples were also amplified using the forward primers F1, F3, F4, Mfor1, Mfor2, Mfor4, Mfor5 and the reverse primers R2, Mrev1, Mrev3, Mrev4, Mrev5, Mrev6. The different combinations of primer sets used for the amplification of WMAI genes are listed in Supplementary Table 1 with the corresponding annealing temperatures used and the expected amplicon sizes.

WDAI 0.53 and 0.19 genes were amplified using WDAIfor and WDAIrev primers with an annealing temperature of 58 °C. The expected amplicon size was 446 bp.

WTAI-CM3 genes were amplified using CM3for and CM3rev primers and an annealing temperature of 55 °C. The expected amplicon size was 647 bp. In order to check for the possible presence of WTAI-CM3 gene sequences also in einkorn, the internal forward primers CM3intfor and CM3intfor2 and the internal reverse primers CM3intrev and CM3intrev2 were designed and used in combinations at the annealing temperature of 50 °C.

CMx1/CMx3 and CMx2 genes were amplified using CMxfor and CMxrev primers and an annealing temperature of 60 °C. The expected amplicon size was 498 bp.

All the PCR amplifications were performed with Biometra® ThermoCycler (Biosense srl, Milan, Italy) in 50 µl volume, consisting of 200-400 ng genomic DNA, 1 µM each primer, 1x HotStar HiFidelity PCR Buffer (containing 1.5 mM $MgSO_4$ and 0.3 mM dNTPs), 2.5 U of HotStar HiFidelity DNA Polymerase (QIAGEN, Hilden, Germany). Due to the lack of amplification of WMAI genes in emmer, turanicum and Alzada genotypes and WTAI-CM3 genes in einkorn, $MgSO_4$ was also tested at final concentration of 2 and 3 mM to enhance the stability of primer-template complexes in order to further check the presence of these genes in the mentioned genotypes.

The cycling parameters were 95 °C for 5 min to pre-denature, followed by 45 cycles of 94 °C for 15 sec, annealing temperature (specific for each primer set) for 1 min, and 72 °C for 90 sec, and a final extension at 72 °C for 10 min.

The desired DNA fragment was recovered from 1.5 % agarose gel using QIAquick gel extraction kit (QIAGEN, Hilden, Germany), quantified by NanoPhotometer® P-Class (Implen GmbH, München, Germany), ligated to the pDrive Cloning Vector and used to transform *E. coli* competent cells by using QIAGEN® PCR Cloning Plus Kit (QIAGEN, Hilden, Germany). For each amplified gene (WMAI, WDAI, WTAI-CM3, CMx), five positive clones obtained from each wheat genotype were screened and sequenced with the Sanger method using the Brilliant Dye Terminator 1.1 kit (NimaGen BV, The Netherlands) and with ABI3730 automated sequencer. Sequences were obtained using the service provided by BMR genomics (Padova, Italy).

2.3 Bioinformatics analysis

2.3.1 Haplotypes identification

Open reading frames were obtained by using ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) and EMBL-EBI server (https://www.ebi.ac.uk/Tools/st/emboss_transeq/). A search for similarity was executed with the BLASTN and BLASTP programs available at the National Centre for Biotechnology Information

(NCBI). Haplotypes identification was performed by MEGA software Version 7.0.26 (Kumar, 2016). Multiple sequence alignments were performed using Clustal Omega program (http://www.ebi.ac.uk/Tools/msa/clustalo/), MEGA software Version 7.0.26 and Jalview version 2.11.0 (Waterhouse, 2009). Sequence identity analysis was carried out with SIAS server (http://imed.med.ucm.es/Tools/sias.html). Amino acid sequences of reference proteins were retrieved from the Universal Protein Resource database (Uniprot) (https://www.uniprot.org/).

The signal peptides and the location of their cleavage sites in proteins were determined with SignalP 5.0 (http://www.cbs.dtu.dk/services/SignalP/).

Obtained sequence data have been deposited in EMBL/GenBank Data Libraries under accession nos. MT739524-38 (WMAI genes); MT861996-2029 (WDAI genes); MT887935-58 (WTAI-CM3 genes) and MT897967-88 (CMx genes).

2.3.2 Phylogenetic analysis

The phylogenetic relationships between haplotypes were performed using MEGA software Version 7.0.26. The FASTA multiple sequence alignment was used to infer the Neighbor-Joining (NJ) phylogenetic tree. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted with 1000 bootstrap replicates (Kumar, 2016).

2.3.3 Principal component analysis (PCA)

PCA was used to transform a number of correlated variables into a smaller number of uncorrelated variables called principal components. With this method the original space for variable measurements was projected down onto two low dimensional subspaces. One of these was case-related (ten wheat genotypes), the other was variable-related. The variables were the amino acid substitutions at a specific position. The variable-related subspace was analysed (factor loading) to understand the correlation between the variables and wheat genotypes (principal component).

First, an interim PCA was run with all the variables included and the quality of representation (squared cosine, cos2) of each variable was checked. Second, since a high cos2 indicates a good representation of the variable on the principal component, a definitive PCA was run with those variables with a quality of representation higher than 0.5. The data were not pretreated before submission to PCA since a homogeneous set of variables was used (i.e. amino acid substitutions at

205 a specific position). The PCA analysis was based on correlation matrix and was performed using
206 STATISTICA Software v. 7.1 (StatSoft, Tulsa, Oklahoma, USA).

207 **2.3.4 Haplotype analysis**

208 The program POPGENE 1.31 was used to carry out the haplotype analysis. The genetic distances
209 were calculated to compare the ten wheat genotypes (Nei, 1972) based on the cluster analysis of a
210 0-1 matrix, where the presence or absence of the haplotypes of all the four genes analysed in this
211 study was recorded as 1 and 0, respectively. The dendrogram was constructed based on the genetic
212 distances using UPGMA (unweighted pair group with arithmetic average) method modified from
213 Neighbor procedure of PHYLIP Version 3.5.

3. Results and discussion

3.1 Sequence analysis of monomeric alpha-amylase inhibitors (WMAI)

One desirable PCR product of the expected size was obtained in common wheat samples Judee, Turkey Red, Marquis, Vida (~529 bp using F2 + R1) and in einkorn, spelt and Peliss genotypes (~502 bp using Mfor3 + Mrev2).

Emmer, turanicum and Alzada samples did not produce any PCR product of the expected size with the two sets of primers cited above and other primers were used (see Supplementary Table 1). The primers were designed to cover the most part of the WMAI ORF region. Supplementary Figure 1 shows the gene sequence of 0.28 WMAI and the sequences of all the primers used in this study to amplify the corresponding ORF. All the combinations of primers listed in Supplementary Table 1 were used, but it was not possible to obtain any PCR product corresponding to the ORF of WMAI genes in the tetraploid emmer, turanicum and Alzada wheat samples.

A previous study (Wang et al., 2010) determined the monomeric α-amylase inhibitor gene sequences of 14 populations of wild emmer wheat (*T. dicoccoides*), which is the wild progenitor of modern tetraploid and hexaploid wheat, however, it was not possible to retrieve any monomeric α-amylase inhibitor gene sequence from *T. durum*, *T. turanicum* and *T. dicoccum* in the GenBank non-redundant DNA database (access performed on May 11st, 2021).

Regarding the available studies determining the amount of monomeric α-amylase inhibitors in tetraploid wheat, the results were mixed. Rogniaux et al. (2015) using a targeted MS/MS approach were able to detect and quantify WMAI protein in *T. aestivum*, *T. durum* and *T. monococcum*, but not in KAMUT® khorasan wheat sample (*T. turgidum* ssp. *turanicum*). Geisslitz et al. (2020) used a new targeted LC-MS/MS method and found that emmer had similar contents of ATI 0.28 as common wheat and spelt but obtained mixed results for durum wheat with six out of eight cultivars showing a content near or even below LOD.

The lack of WMAI sequences in *T. durum*, *T. turanicum* and *T. dicoccum* from available databases together with the amplification of WMAI genes in *T. durum* cv. Peliss, but not in other tetraploid wheat in this study, and the mixed results regarding the presence of WMAI protein in tetraploid wheat confirmed that further studies are needed to deepen knowledge about WMAI in tetraploid wheat.

Regarding the other seven wheat genotypes (einkorn, Peliss, spelt, Turkey Red, Judee, Marquis and Vida) the DNA sequences from five clones per each wheat genotype were determined and used to

deduce the full amino acid sequence of the proteins. As described before (Wang et al., 2008), there was no intron in the monomeric α-amylase inhibitor sequences. So, it was possible to isolate the complete coding sequences of monomeric α-amylase inhibitor gene by direct PCR amplification. In this study, no ins/del in the coding region of WMAI were found in any of the wheat samples.

From a total of 35 deduced amino acid sequences, 16 haplotypes were identified (HM1-HM16). Haplotype HM7 was shared by all the hexaploid wheat samples (it occurred in 13 out of 25 hexaploid wheat sequences) and it turned out to be the most abundant WMAI haplotype. No pseudogenes were found (Table 3).

All the putatively functional genes encoded for a 151 amino acid protein (30 amino acid signal peptide and a 121 amino acid mature protein).

In this study, after aligning the 16 monomeric α-amylase inhibitor haplotypes from 7 wheat genotypes and WMAI 0.28 (GenBank: AJ223492), 18 nsSNPs were identified. The frequency of nsSNPs was 1 out of 25.2 bases.

The alignment of the 16 monomeric α-amylase inhibitor haplotypes and the position of the nsSNPs were shown in Supplementary Figure 2.

Sequence alignment showed that all the monomeric α-amylase inhibitors were highly homologous and are part of a monomeric α-amylase inhibitor family including WMAI 0.28 and suggested that these inhibitors might have derived from a very limited number of ancestral genes (Wang et al., 2008). Haplotype HM7, which is the most abundant WMAI haplotype, had the same sequence as WMAI 0.28; the other haplotypes showed a percentage of identity with WMAI 0.28 ranging from 96.7% to 99.3%.

All the deduced proteins of monomeric α-amylase inhibitors had 10 Cys residues which form five disulphide bonds. The disulphide bonds are essential for the inhibitory activity (Carbonero and García-Olmedo, 1999) and in fact the Cys residues were at conserved positions.

Monomeric α-amylase inhibitors are highly active against α-amylase of *Tenebrio molitor* and show a low inhibitory activity against mammalian and some avian α-amylases. The 3-D structure of the complex between α-amylase from *T. molitor* and 0.28 α-amylase inhibitor has been determined (Payan, 2004) and three regions of contact have been identified in the mature protein: the N-terminal segment (residues 1-10); residue 53 (inside the second loop segment) and sequence including residues 103-119, which are part of the fourth loop (shown as residues 31-40, 83 and 133-149 respectively in Supplementary Figure 2). In particular, the last region, which corresponds to the

C-terminal segment of 0.28, plays an important role in filling the central substrate-binding subsites of *T. molitor* α-amylase and in targeting its catalytic residues. Moreover, this region plays an important role in the specificity of 0.28 inhibitor. Most of the haplotypes shared the same sequence as WMAI 0.28 in these regions, with the following exceptions: HM1, HM9, HM10 and HM13.

Other two regions critical for the inhibitory activity have been identified in the mature protein: the N-terminal sequence (positions 1-6) and the sequence after the CRC motif (positions 57-59) (Liu and Wang, 2012). Only haplotype HM1 (found in einkorn) displayed a change at position 6 (shown as residue 36 in Supplementary Figure 2).

The current literature lacks information on the peptide sequences and epitopes responsible for the allergies triggered by wheat albumin and globulin proteins (Tatham and Shewry, 2008). Regarding ATIs, there is only one study which identified a high IgE-binding region at the position 9 to 26 in the mature WMAI protein (Walsh and Howden, 1989). As shown as residues 39-56 in Supplementary Figure 2, only two haplotypes had an amino acid change at this level: HM5 and HM14.

Whether the sequence changes found in this study and cited above are able to reduce or even delete the inhibitory activity and/or the immunogenic properties is not possible to say from these data.

3.2 Sequence analysis of dimeric alpha-amylase inhibitors (WDAI)

The primers WDAIfor and WDAIrev were used to amplify the ORF of the dimeric alpha-amylase genes (WDAI). All the ten wheat genotypes gave the PCR product of the expected size (~509 bp).

From a total of 50 deduced amino acid sequences, 35 haplotypes were identified (HD1-HD35), among which 6 haplotypes represented pseudogenes (indicated with a p after the number), due to the presence of one or more in-frame stop codons because of a single nucleotide polymorphism (substitution, insertion or deletion), although we cannot predict from the genomic data whether these sequences are being expressed (Table 3). All the sequences obtained from the einkorn sample showed an insertion of C at position 160 (from the start codon ATG) which resulted in a premature stop codon and the impossibility to synthesize the correct mature protein. The same insertion of C was observed in the DNA sequences from all the *T. monococcum* accessions and from above all *T. boeoticum* accessions in a previous study (Wang, 2007). The absence of gene sequences coding for functional WDAI genes in *T. monococcum* is in agreement with functional studies using targeted MS/MS approach to detect WDAI proteins (Geisslitz et al., 2020, 2018; Rogniaux et al., 2015).

All the putatively functional genes were 426 bp long, encoding for a 141 amino acids protein (17 amino acids signal peptide and 124 amino acids mature protein). Haplotype HD18 was found in 6 clones from 5 different wheat genotypes and it was the most abundant.

In this study, after aligning the 35 dimeric α-amylase inhibitor haplotypes from ten wheat genotypes, 69 nsSNPs were identified (Supplementary Figure 3).

The frequency of nsSNPs was 1 out of 6.2 bases. The WDAI gene sequences had more divergence than the WMAI gene sequences of this study, which is in agreement with previous studies (Liu and Wang, 2012).

All the deduced proteins of WDAI but HD1 and ATI 0.53 had 10 Cys residues which formed five disulphide bonds and were at conserved positions.

Studying the modelled complex of human salivary α-amylase (HSA) with 0.19 inhibitor, three spots important for the inhibitory activity of WDAI have been highlighted (Payan, 2004): His47; Ser49 and the sequence Val104 – Val105 – Asp106 – Ala107. Amino acids changes at the level of these three spots can result in a different ability to inhibit human α-amylases. As shown as residues 64, 66 and 121-124 respectively in Supplementary Figure 3, the only haplotypes displaying all the above-mentioned amino acids residues at the three inhibitor spots were HD5, HD9, HD11, HD15 and HD17. Interestingly, all these haplotypes were found only in hexaploid wheat genotypes, which showed in fact the highest α-amylase inhibitory activity *in vitro* (data not published yet).

Considering the percentage of identity with the two amino acid sequences of the well-studied dimeric α-amylase inhibitors 0.19 and 0.53, only the haplotype HD9 had a 100% of identity with 0.19. However, the level of homology was high with both 0.19 and 0.53 (ranged from 87.1% to 100% for 0.19 and from 84.7% to 99.2% for 0.53), with the exception of the pseudogene HD29p.

3.3 Sequence analysis of tetrameric alpha-amylase inhibitors CM3 (WTAI-CM3)

WTAI-CM3 was selected as representative of WTAI genes in this study because of its extended interaction with α-amylases compared to the other CM subunits (Capocchi et al., 2013) and because it was the most abundant expressed WTAI proteins in cv. Butte 86 (Altenbach, 2011).

The sequence of CM3 encoding gene revealed that there is only one ORF encoding the entire protein (Singh et al., 2012).

The primers CM3for and CM3rev were designed *in house* and were used to amplify the ORFs of the tetrameric α-amylase gene CM3 and all the samples but einkorn gave the PCR product of the

expected size (~ 647 bp). Two sets of internal primers were designed *in house* and were used in all the possible combinations with the einkorn sample, but no band of the expected size was amplified. Moreover, using the sequence of inhibitor WTAI-CM3 (NCBI: X17574) a blast search against the GenBank non-redundant DNA database was performed (access performed on May 11th, 2021), but it was not possible to retrieve any WTAI-CM3 sequence from *T. monococcum*.

Some studies were previously performed to quantify allergens abundance in different wheat genotypes and raised doubts on the presence of WTAI-CM3 proteins in diploid wheats. Rogniaux et al. (2015) detected WTAI-CM3 proteins only at very low levels in two *T. monococcum* (cv. Engrain, cv. DV92). Geisslitz et al. (2020, 2018) detected WTAI-CM3 in only four out of eight einkorn cultivars, but in very low concentrations near the LOQ or even below.

From a total of 45 deduced amino acid sequences derived from 9 wheat genotypes, 25 haplotypes were obtained (HC1-HC25), among which one haplotype was identified as a pseudogene (HC20p) due to the presence of one in-frame stop codon (table 3). As described before (Singh et al, 2012) there was no intron in the WTAI-CM3 sequences. In this study, no ins/del in the coding region of WTAI-CM3 was found in any of the wheat samples.

All the putatively functional genes encoded for a 168 amino acid protein (25 amino acid signal peptide and 143 amino acid mature protein).

In this study, after aligning the 25 WTAI-CM3 haplotypes from 9 wheat genotypes, 34 nsSNPs were identified. The frequency of nsSNPs was 1 out of 14.8 bases. Previously, Liu et al. (2012) stated that WTAI genes (CM2, CM3, CM16) are more conserved than WMAI and WDAI genes probably because these subunits are combined to form the tetrameric inhibitor and an amino acid change in one of them can affect the structure and consequently the resulting inhibitory activity. However, in this study results showed that WTAI-CM3 was more conserved than WDAI genes where it was found a snSNPs frequency of 1 out of 6.2 bases, but less conserved than WMAI genes where the snSNP frequency was 1 out of 25.2 bases.

The alignment of the 25 WTAI-CM3 haplotypes and the position of the nsSNPs were shown in Supplementary Figure 4.

Haplotype HC6, which was the most abundant WTAI-CM3 haplotype, had the same sequence as WTAI-CM3 (X17574) studied by García-Maroto et al. (1990), the other haplotypes showed a high percentage of identity with WTAI-CM3 (X17574) ranging from 92.9% to 99.4%.

The most part of the deduced proteins of WTAI-CM3 had 10 Cys residues which form five disulphide bonds and are at conserved positions. The exceptions were HC3 (Tyr instead of the

second Cys of the CRC motif); HC9 (Tyr at position 148); HC11 (Arg at position 29); HC18 (Tyr at position 52).

The sequence DLPGCPRE (amino acid positions 101-108 of the mature protein, shown as residues 126-133 in Supplementary Fig. 4) is the most conserved in CM proteins and may be the characteristic sequence of alpha amylase/ trypsin - bifunctional inhibitors. Most of the haplotypes shared this conserved sequence, with the following exceptions: HC14, HC3, HC4, HC7, HC8, HC21 and HC25 (Supplementary Fig. 4).

Capocchi et al. (2013) proposed a structural model for the emmer tetrameric α-amylase inhibitor and showed that two sequences Phe30-Lys55 and Lys116-Gln122 of the mature protein (shown as residues 55-80 and 141-148 respectively in Supplementary Fig. 4) are important for the inhibitory activity against *T. molitor* α-amylase. In this study four amino acid changes occurred in some haplotypes at level of the first sequence: Thr instead of Met at amino acid position 65 occurred in HC16; Val instead of Ala at amino acid position 68 occurred in HC1; Phe instead of Tyr at amino acid position 71 and Met instead of Gly at position 74 both occurred in HC3, HC7, HC8, HC21, HC25 (Supplementary Fig. 4). Interestingly, the haplotypes which had the last two amino acid changes also had changes at the conserved sequence DLPGCPRE and all these were found only in hexaploid wheat genotypes.

ATI proteins from wheat have been demonstrated to trigger inflammation by eliciting strong innate immune effects *in vitro* and *in vivo* with the activation of the TLR4–MD2–CD14 complex (Junker et al., 2012). Cuccioloni et al. (2016) investigated the interaction between WTAI-CM3 and human TLR4 and predicted the ATI-TLR4 binding interface regions. These sequences are residues 33-44 and 90-100 in the mature protein sequence (shown as residues 58-69 and 115-125 respectively in Supplementary Fig. 4). The haplotypes which showed one or two amino acid changes within these regions were: HC1, HC5, HC8, HC15 and HC16 (Supplementary Figure 4). From the existing evidence it is not possible to say if these changes are able to decrease or delete the ability of CM3-WTAI to bind human TLR4.

Several studies suggested that the apparent absence of alpha-amylase inhibitory activity of *T. monococcum* against mammalian enzymes could be explained by the fact that the corresponding coding genes might be expressed at very low level or even silenced, perhaps because of gene mutations that prevent the translation into the mature protein (García-Maroto et al., 1990). This hypothesis is in agreement with the gene sequencing results of this study. In fact, einkorn did not show any WTAI-CM3 gene sequence and so it should not be able to produce any tetrameric alpha-amylase inhibitor proteins. Moreover, all the WDAI gene sequences showed an insertion of C at

position 160 which resulted in a premature stop codon and the impossibility to synthesize the correct WDAI mature protein. Lastly, einkorn showed WMAI sequences, but it is known that WMAI proteins are highly active against alpha-amylase of *T. molitor* and only weakly inhibits the alpha-amylases from human saliva and pancreas (Payan, 2004). Notwithstanding low levels of ATI proteins and a low alpha-amylase inhibitory activity, Call et al. (2020) found that einkorn samples have high trypsin inhibitory activities and they hypothesized that it could be due to the fact that besides ATIs also lipid transfer proteins (LTPs), grain softness proteins and beta-amylases show high affinity to trypsin.

3.4 Sequence analysis of trypsin inhibitors CMx

The primers CMxfor and CMxrev were designed *in house* and were used to amplify the ORFs of the CMx genes and all the ten wheat genotypes gave the PCR product of the expected size (~ 498 bp). Since CMxfor was designed 9 nucleotides downstream the A of the start codon, these first 9 nucleotides were taken from the reference sequence (GenBank: X75608) and added to each DNA sequence in order to deduce complete protein sequences for subsequent analyses.

From a total of 50 deduced amino acid sequences, 22 haplotypes were obtained (HX1-HX22), among which two haplotypes were identified as pseudogenes due to the presence of one in-frame stop codon (Table 3).

HX20 occurred in 15 out of 50 CMx sequences and it turned out to be the most abundant CMx haplotype.

The deduced proteins from haplotypes HX1-HX10 were 145 amino acids long and were found only in hexaploid wheat genotypes and the deduced proteins from haplotypes HX13-HX20 were 146 amino acids long. Differently, two nucleotide sequences showed a premature stop codon after nucleotide 363 as observed by Sanchez de la Hoz et al. (1994), which resulted in a deduced protein of 121 amino acids (haplotypes HX11 and HX12). Interestingly, emmer wheat showed only the two haplotypes with the deduced protein of 121 amino acids. The homology between the deduced HX11 and HX12 amino acid sequences and those of the other CMx haplotypes was maintained beyond the premature stop codon, up to the second stop codon which appeared in the same position as the other haplotypes. All the deduced proteins showed a 25 amino acid signal peptide.

In this study, after aligning the 22 CMx haplotypes from 10 wheat genotypes, 39 nsSNPs were identified (Supplementary Figure 5). The frequency of nsSNPs was 1 out of 11.2 bases, similar to what was found for CM3 genes in this study.

The most part of the CMx deduced proteins had 10 Cys residues which form up to five disulphide

433 bonds and were at conserved positions. The exceptions were HX3, HX4, HX9, HX11 HX12 and
434 HX17. Interestingly, all the haplotypes found in emmer in this study didn't show all the 10 Cys
435 residues.

436 **3.5 Phylogenetic analysis**

437 The Neighbor-Joining method was used to calculate the phylogenetic distances and to construct the
438 phylogenetic tree among all the haplotypes for each gene (Figure 1).

439 For the 16 WMAI haplotypes the optimal tree with the sum of branch length 0.12061355 was
440 shown (Figure 1A). Cluster III, which displayed all the haplotypes of the tetraploid Peliss and the
441 diploid einkorn, diverged from cluster I, cluster II and HM8, the latter displaying only hexaploid
442 wheat genotypes (in particular cluster II and HM8 displayed only haplotypes from spelt). So, the
443 clusters clearly separated the wheat genotypes according to their ploidy.

444 For the 35 WDAI haplotypes the optimal tree with the sum of branch length 0.96668030 was shown
445 (Figure 1B). HD29p was the most divergent and was highly separated from the other haplotypes.
446 The other pseudogenes from einkorn (HD28p, HD34p, HD35p) diverged together in a separated
447 group within cluster III. Each cluster showed haplotypes from both tetraploid and hexaploid wheat
448 genotypes and from both modern and ancient wheat genotypes, so it was not possible to make any
449 differentiation based on WDAI gene sequences.

450 For the 25 WTAI-CM3 haplotypes the optimal tree with the sum of branch length 0.20342301 was
451 shown (Figure 1C). Two clusters were identified with cluster II displaying only haplotypes from
452 hexaploid wheat genotypes (both modern and ancient genotypes).

453 For the 22 CMx haplotypes the optimal tree with the sum of branch length 0.27715519 was shown
454 (Figure 1D). A first subdivision was between haplotypes HX1-HX10 (which were all 145 amino
455 acids long and found only in hexaploid wheat genotypes) and HX11-HX22p. Then a further
456 subdivision distinguished HX9 and HX10 (cluster IV) from the other haplotypes of the first group
457 (cluster V), while in the second group three clusters were identified: cluster I with haplotypes found
458 in both tetraploid and hexaploid genotypes, cluster II showing the two 121 amino acids long CMx
459 proteins, cluster III showing only haplotypes found in einkorn.

460 According to the phylogenetic analysis of the haplotypes of all the four genes analysed, it was not
461 possible to differentiate the wheat genotypes with respect to ancient or recently developed wheat
462 genotypes.

3.6 Principal Component Analysis (PCA) of WMAI, WDAI, WTAI-CM3 and CMx sequences

A PCA analysis was performed based on deduced amino acid sequences of all the four genes sequenced in this study (WMAI, WDAI, WTAI-CM3 and CMx).

To visualise relationships between the tested variables, principal component analysis was run for each species in two steps. First, an interim PCA was run including all the variables (160 nsSNPs), and the quality of representation (Cos2) of each variable was checked. Second, a definitive PCA was run with those variables with a quality of representation higher than 0.5 (103 nsSNPs).

The scatter plot reported the projection of cases (10 wheat genotypes) on the first two components PC1 and PC2 and explained the 63.6% of total variance with the first PC accounting for 40.9% and the second PC for 22.7% (Figure 2). Einkorn diverged from all the other genotypes according to the projection of the first component (PC1 axis). The tetraploid wheat genotypes can be divided from the hexaploid wheat genotypes according to the second projection (PC2 axis) with the tetraploid Peliss which was closer to the hexaploid group. This fact can be explained in part by the fact that Peliss was the only tetraploid wheat which was amplified with primers for WMAI genes. It was not possible to differentiate the wheat genotypes with respect to ancient or recently developed wheat genotypes.

3.7 Haplotype analysis

The Nei's genetic distances were calculated for paired comparisons of the ten wheat genotypes (Supplementary Table 2). The values of genetic distances varied from 0.0085 to 0.0321 with an average of 0.022. The genetic distance between turanicum and Alzada was the lowest (0.0085) whereas those between Peliss and both spelt and Vida were the largest.

Figure 3 showed the dendrogram based on the Nei's genetic distances. First, einkorn diverged from the other genotypes and this confirmed the results of the PCA analysis and this is in line with the literature which affirmed that *T. monococcum* did not contribute to the evolution of both tetraploid species of *Triticum* and *T. aestivum* through polyploidization (Zhao et al., 2021).

Then, the group of the naked tetraploid wheat genotypes diverged from the others (in this group turanicum and Alzada are closer to each other than Peliss probably because they lacked WMAI sequences, unlike Peliss). The hulled tetra/hexaploid wheat genotypes diverged from the naked hexaploid wheat genotypes. In this last group it is interesting to note that the heritage genotypes (Turkey Red and Marquis) were closer than the modern genotypes (Judee and Vida). It is also interesting to highlight that emmer was close to spelt and this is in line with recent studies based on

494 HMW glutenin analysis which showed that the European spelt is a result of hybridization between
495 free-threshing hexaploid wheat and domesticated emmer (Blatter et al., 2004).

4. Conclusions

In this study, the sequences of four representative alpha-amylase/trypsin inhibitor genes have been determined in ten wheat genotypes with different ploidy level and year of release, and some of these, for example *T. monococcum* and *T. turgidum* spp. *turanicum*, have never been sequenced for some or all ATI genes before. So, this study has expanded the available information about ATI sequences.

Considering the deduced amino acid sequences of all the four ATI genes studied, both the phylogenetic analysis and the PCA analysis evidenced that the ten wheat genotypes can be differentiated on the basis of their ploidy level. This is in agreement with the fact that ATI genes are present in different chromosomal sets. Overall, einkorn was the most divergent genotype.

Based on both the phylogenetic analysis and the PCA analysis of the haplotypes of the four genes analysed, it was not possible to differentiate the wheat genotypes with respect to ancient or recently developed wheat genotypes. Further studies with a higher number of genotypes per each ploidy level are needed to confirm these results. The haplotype analysis, beside confirming these results, also allowed the separation of the hulled from the naked wheat genotypes.

The literature lacks information on which sequences of ATIs have immunogenic and inflammatory potential (Geisslitz et al., 2021), so it was not possible to make this kind of consideration on the basis of the deduced amino acid sequences. However, the genetic sequence differences highlighted in this study among the ten genotypes can be the basis for further studies aimed at identifying those sequences in ATI genes able to trigger innate and adaptive immune response. Moreover, the sequence variants of ATIs can be useful information for future breeding programs aimed at selecting wheat varieties with potential pest resistance and minimum inflammatory effect on human health.

Finally, it is desirable to measure the inhibitory activities against both alpha-amylase and trypsin enzymes to find possible differences among the ten wheat genotypes and to eventually find a correlation with their genetic sequences.

To the best of our knowledge, this is the first study analyzing the ATI sequences of a set of ancient and modern wheat genotypes with a different ploidy level.

Funding

This work was supported in part by a grant from the Kamut Enterprise of Europe (KEE), Oudenaarde, Belgium. KEE had no role in the study design, data collection, analysis, interpretation of data, or preparation of the manuscript.

References

Altenbach, S., Vensel, W.H., Dupont, F.M., 2011. The spectrum of low molecular weight alpha-amylase/protease inhibitor genes expressed in the US bread wheat cultivar Butte 86. BMC Res. Notes 4.

Blatter, R.H.E., Jacomet, S., Schlumbaum, A., 2004. About the origin of European spelt (*Triticum spelta* L.): allelic differentiation of the HMW Glutenin B1–1 and A1–2 subunit genes. Theor Appl Genet 108, 360–367.

Call, L., Kapeller, M., Grausgruber, H., Reiter, E., Schoenlechner, R., D'Amico, S., 2020. Effects of species and breeding on wheat protein composition. J. Cereal Sci. 93, 102974.

Capocchi, A., Muccilli, V., Cunsolo, V., Saletti, R., Foti, S., Fontanini, D., 2013. A heterotetrameric alpha-amylase inhibitor from emmer (*Triticum dicoccon* Schrank) seeds. Phytochemistry 88, 6–14.

Carbonero, P., García-Olmedo, F., 1999. A Multigene Family of Trypsin/α-Amylase Inhibitors from Cereals. In: P.R. S, Casey R, editors. Seed Proteins. Dordrecht: Springer, 617–633.

Cuccioloni, M., Mozzicafreddo, M., Ali, I., Bonfili, L., Cecarini, V., Eleuteri, A.M., Angeletti, M., 2016. Interaction between wheat alpha-amylase/trypsin bi-functional inhibitor and mammalian digestive enzymes: Kinetic, equilibrium and structural characterization of binding. Food Chem. 213, 571–578.

de la Hoz, P.S., Castagnaro, A., Carbonero, P., 1994. Sharp divergence between wheat and barley at loci encoding novel members of the trypsin/α-amylase inhibitors family. Plant Mol. Biol. 26, 1231–1236.

EL Hassouni, K., Sielaff, M., Curella, V., Neerukonda, M., Leiser, W., Würschum, T., Schuppan, D., Tenzer, S., Longin, C.F.H., 2021. Genetic architecture underlying the expression of eight α-amylase trypsin inhibitors. Theor. Appl. Genet. 134, 3427–3441.

García-Maroto, F., Maraña, C., Mena, M., García-Olmedo, F., Carbonero, P., 1990. Cloning of cDNA and chromosomal location of genes encoding the three types of subunits of the wheat tetrameric inhibitor of insect α-amylase. Plant Mol. Biol. 14, 845–853.

Geisslitz, S., Shewry, P., Brouns, F., America, A.H.P., Caio, G.P.I., Daly, M., D'Amico, S., De Giorgio, R., Gilissen, L., Grausgruber, H., Huang, X., Jonkers, D., Keszthelyi, D., Larré, C., Masci, S., Mills, C., Møller, M.S., Sorrells, M.E., Svensson, B., Zevallos, V.F., Weegels, P.L., 2021. Wheat ATIs: Characteristics and Role in Human Disease. Front. Nutr. 8, 667370.

559 Geisslitz, S., Longin, C.F.H., Koehler, P., Scherf, K.A., 2020. Comparative quantitative LC–
560 MS/MS analysis of 13 amylase/trypsin inhibitors in ancient and modern *Triticum* species. Sci. Rep.
561 10, 14570.
562 Geisslitz, S., Ludwig, C., Scherf, K.A., Koehler, P., 2018. Targeted LC-MS/MS Reveals Similar
563 Contents of α-Amylase/Trypsin-Inhibitors as Putative Triggers of Nonceliac Gluten Sensitivity in
564 All Wheat Species except Einkorn. J. Agric. Food Chem. 66, 12395–12403.
565 Gélinas, P., Gagnon, F., 2018. Inhibitory activity towards human α-amylase in wheat flour and
566 gluten. Int. J. Food Sci. Technol. 53, 467–474.
567 Huebener, S., Tanaka, C.K., Uhde, M., Zone, J.J., Vensel, W.H., Kasarda, D.D., Beams, L., Briani,
568 C., Green, P.H.R., Altenbach, S.B., Alaedini, A., 2015. Specific Nongluten Proteins of Wheat Are
569 Novel Target Antigens in Celiac Disease Humoral Response. J. Proteome Res. 14, 503–511.
570 Junker, Y., Zeissig, S., Kim, S.J., Barisani, D., Wieser, H., Leffler, D.A., Zevallos, V., Libermann,
571 T.A., Dillon, S., Freitag, T.L., Kelly, C.P., Schuppan, D., 2012. Wheat amylase trypsin inhibitors
572 drive intestinal inflammation via activation of toll-like receptor 4. J. Exp. Med. 209, 2395–2408.
573 Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis
574 Version 7.0 for Bigger Datasets. Mol. Biol. Evol. 33, 1870–1874.
575 Liu, Y., Wang, J., 2012. Molecular Evolution of Exogenous Alpha-amylase Inhibitors in *Triticeae* -
576 An Update. Funct. Plant Sci. Biotechnol. 6, 67–74.
577 Nei, M., 1972. Genetic Distance between Populations. American Naturalist 106, 283-292.
578 Pandey, B., Saini, M., Sharma, P., 2016. Molecular phylogenetic and sequence variation analysis of
579 dimeric α-amylase inhibitor genes in wheat and its wild relative species. Plant Gene 6, 48–58.
580 Payan, F., 2004. Structural basis for the inhibition of mammalian and insect α-amylases by plant
581 protein inhibitors. Biochim. Biophys. Acta - Proteins Proteomics 1696, 171–180.
582 Rogniaux, H., Pavlovic, M., Lupi, R., Lollier, V., Joint, M., Mameri, H., Denery, S., Larré, C.,
583 2015. Allergen relative abundance in several wheat varieties as revealed via a targeted quantitative
584 approach using MS. Proteomics 15, 1736–1745.
585 Singh, A., Gaur, A.K., Mishra, D.P., 2012. Molecular cloning and characterization of cm3 gene,
586 from *T. aestivum* and *T. durum* genome. Int. J. Sci. Res. Publ. 2.
587 Tatham, A.S., Shewry, P.R., 2008. Allergens to wheat and related cereals. Clin. Exp. Allergy 38,
588 1712–1726.
589 Walsh, B.J., Howden, M.E.H, 1989. A method for the detection of IgE binding sequences of
590 allergens based on a modification of epitope mapping. J Immunol Methods 121(2), 275-80.

Wang, J.-R., Wei, Y.-M., Deng, M., Nevo, E., Yan, Z.-H., Zheng, Y.-L., 2010. The impact of single nucleotide polymorphism in monomeric alpha-amylase inhibitor genes from wild emmer wheat, primarily from Israel and Golan. BMC Evol. Biol. 10, 170.

Wang, J.R., Wei, Y.M., Yan, Z.H., Zheng, Y.L., 2008. SNP and haplotype identification of the wheat monomeric α-amylase inhibitor genes. Genetica 134, 277–285.

Wang, J.-R., Wei, Y.-M., Yan, Z.-H., Zheng, Y.-L., 2007. Sequence variations and haplotype identification of wheat dimeric alpha-amylase inhibitor genes in einkorn wheats. Biochem. Genet. 45, 803–14.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J., 2009. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189–1191.

Zevallos, V.F., Raker, V., Tenzer, S., Jimenez-Calvente, C., Ashfaq-Khan, M., Rüssel, N., Pickert, G., Schild, H., Steinbrink, K., Schuppan, D., 2017. Nutritional Wheat Amylase-Trypsin Inhibitors Promote Intestinal Inflammation via Activation of Myeloid Cells. Gastroenterology 152, 1100–1113.

Zhao, X., Fu, X., Yin, C., Lu, F., Engineering, C., 2021. Wheat speciation and adaptation: perspectives from reticulate evolution. aBIOTECH 2, 386–402.

Tables

Table 1

List of wheat genotypes used in this study.

***Triticum* species**	**Genotypes**	**Genome formula**	**Year of release**	**Typology**	**Pedigree**	**Provenance**	**Name used in this study**
T. monococcum L. subsp. *monococcum*	Local landrace	AA	/	Hulled, ancient wheat	N.D.	Montana State University (Montana, USA)	Einkorn
T. turgidum L. subsp. *dicoccum*	Local landrace	AABB	/	Hulled, ancient wheat	N.D.	Montana State University (Montana, USA)	Emmer
T. turgidum L. subsp. *turanicum**	QK-77	AABB	/	Naked, ancient wheat	N.D.	Kamut International (Montana, USA)	Turanicum
T. turgidum L. subsp. *durum*	Peliss	AABB	1900	Naked, heritage durum wheat	LV-Oran	University of Saskatchewan (Canada)	Peliss
T. turgidum L. subsp. *durum*	Alzada	AABB	2004	Naked, modern durum wheat	Mohawk/Kofa	Montana State University (Montana, USA)	Alzada
T. aestivum L. subsp. *spelta*	Local landrace	AABBDD	/	Hulled, ancient wheat	N.D.	Montana State University (Montana, USA)	Spelt
T. aestivum L. subsp. *aestivum*	Turkey Red	AABBDD	1873	Naked, heritage hard red common wheat	(S)Crimean	Hartland Mills (Kansas, USA)	Turkey Red
T. aestivum L. subsp. *aestivum*	Judee	AABBDD	2011	Naked, modern hard red common wheat	93-X-312-E-14/NuHorizon	Montana State University (Montana, USA)	Judee
T. aestivum L. subsp. *aestivum*	Marquis	AABBDD	1913	Naked, heritage hard red common wheat	Hard-Red-Calcutta/Red-Fife	Montana State University (Montana, USA)	Marquis
T. aestivum L. subsp. *aestivum*	Vida	AABBDD	2006	Naked, modern hard red common wheat	Scholar/Reeder	Montana State University (Montana, USA)	Vida

* sold under the KAMUT® brand. KAMUT® is a registered trademark of Kamut International, Ltd, and Kamut Enterprises of Europe, bv.

Table 2

List of primers used in this study.

Gene	Name	Sequence
WMAI (NCBI GenBank: AJ223492)	Mfor1	5’ – CACCACTTATATCCAAGGACCA – 3’
	Mfor2	5’ – AGCTTGCTTTGATTCTGCTGAT – 3’
	Mfor3	5’ – ACTAAATTGAAACAATGTGGAT – 3’
	F2	5’ – ATGTGGATGAAGACCGKGTT – 3’
	F4	5’ – ATGCTCGTGGCGACAACAAT – 3’
	F3	5’ – ACAACAATGGCGGTCGAGTA – 3’
	F1	5’ – CATAACAGTGGTCCTTGGAGT – 3’
	Mfor4	5’ – GCAATGGTGAAGCTCCAGT – 3’
	Mfor5	5’ – AGCTGGCCGACATCAACA – 3’
	Mrev1	5’ – TCTCTGAGAGGACACATACACCA – 3’
	R1	5’ – CACGCACCGCACCATTACTT – 3’
	Mrev2	5’ – CACCGCACCAATTAAGAT – 3’
	Mrev3	5’ – CGCACCAATTAAGATGCAGA - 3’
	Mrev4	5’ – TTAAGATGCAGATTCGCTTGAC – 3’
	R2	5’ – GACTAGRYGTCCGKATACGC – 3’
	Mrev5	5’ – GATGGGCACCTTGCAGAC – 3’
	Mrev6	5’ – GCACCTCCTTCCCCTCAC – 3’
WDAI (NCBI GenBank: AK330823)	WDAIfor	5’ – CTATGTATGCTCGTGGCGAC - 3’
	WDAIrev	5’ – ACTCATTYGCTTGACTAGGC – 3’
WTAI-CM3 (NCBI GenBank: X17574)	CM3for	5’ – CGAACCAGACTTGGCTAGAATA – 3’
	CM3intfor2	5’ – ACAACAAACTTGTGGCACCT – 3’
	CM3intfor	5’ – GCGCTGCGCTACTTCATA – 3’
	CM3rev	5’ – ATTCATAGCAGATAGCCCACAC – 3’
	CM3intrev2	5’ – TGTGAATGGTCGCCAAGT – 3’
	CM3intrev	5’ – GCCGCTCTCACCAACAT – 3’
CMx (NCBI GenBank: X75608.1)	CMxfor	5’ – AAGCACCAGCTCATCCTCTC – 3’
	CMxrev	5’ – ATACACATATGCGATTCGTCCA – 3’

Table 3

Distribution of WMAI, WDAI, WTAI-CM3 and CMx haplotypes in the ten wheat genotypes. The numbers under parenthesis indicate how many times the haplotype occurred in the 5 sequences analysed per each wheat genotype.

GENOTYPES	WMAI HAPLOTYPES	WDAI HAPLOTYPES	WTAI-CM3 HAPLOTYPES	CMx HAPLOTYPES
Einkorn	HM1 (1), HM2 (1), HM3 (3)	HD28p (2), HD29p (1), HD34p (1), HD35p (1)	/	HX13 (3), HX14 (1), HX21p (1)
Emmer	/	HD8 (1), HD18 (1), HD26 (2), HD27 (1)	HC6 (3), HC16 (1), HC17 (1)	HX11 (4), HX12 (1)
Turanicum	/	HD18 (2), HD20 (1), HD21 (1), HD22 (1)	HC6 (1), HC9 (1), HC10 (1), HC11 (1), HC12 (1)	HX17 (1), HX18 (1), HX20 (3)
Peliss	HM3 (3), HM4 (1), HM5 (1)	HD1 (1), HD2 (1), HD4 (1), HD18 (1), HD32 (1)	HC6 (4), HC18 (1)	HX20 (5)
Alzada	/	HD8 (1), HD18 (1), HD23 (1), HD24 (1), HD25 (1)	HC6 (2), HC13 (1), HC14 (1), HC15 (1)	HX19 (1), HX20 (4)
Spelt	HM6 (1), HM7 (1), HM8 (1), HM9 (1), HM10 (1)	HD15 (2), HD31 (1), HD30p (2)	HC6 (2), HC7 (1), HC22 (1), HC23 (1)	HX3 (1), HX8 (1), HX11 (3)
Turkey Red	HM7 (3), HM11 (1), HM12 (1)	HD5 (1), HD6 (1), HD7 (1), HD8 (1), HD9 (1)	HC6 (2), HC19 (1), HC25 (1), HC20p (1)	HX1 (1), HX5 (1), HX8 (2), HX9 (1)
Judee	HM7 (3), HM13 (1), HM14 (1)	HD9 (1), HD10 (1), HD11 (1), HD12 (1), HD33 (1)	HC1 (1), HC2 (1), HC3 (1), HC4 (1), HC5 (1)	HX6 (1), HX15 (1), HX16 (1), HX20 (2)
Marquis	HM7 (4), HM15 (1)	HD13 (1), HD14 (1), HD15 (2), HD16 (1)	HC6 (2), HC7 (1), HC21 (1), HC24 (1)	HX7 (1), HX8 (1), HX10 (1), HX20 (1), HX22p (1)
Vida	HM7 (2), HM15 (2), HM16 (1)	HD3 (1), HD9 (1), HD17 (1), HD18 (1), HD19p (1)	HC6 (2), HC7 (1), HC8 (2)	HX2 (1), HX4 (1), HX8 (3)

Supplementary Table 1

Combinations of primer sets used to amplify WMAI genes. For each primer set the expected amplicon size (bp) and annealing temperature used (°C) are listed.

Reverse / Forward	R1	R2	Mrev2	Mrev3	Mrev5	Mrev6	Mrev4	Mrev1
F1	445 bp 50°C	374 bp 62°C	404 bp 50°C	401 bp 50°C				
F2	529 bp 58°C	458 bp 50°C	488 bp 50°C	485 bp 56°C				
F3	478 bp 50°C	407 bp 56°C	437 bp 52°C	434 bp 55°C				
F4	490 bp 50°C	419 bp 50°C	449 bp 50°C	446 bp 55°C				
Mfor3	543 bp 54°C	472 bp 50°C	502 bp 50°C	502 bp 50°C				
Mfor4				329 bp 54°C	240 bp 54°C	169 bp 54°C		
Mfor5				262 bp 58°C	173 bp 54°C	102 bp 54°C		
Mfor1							550 bp 50°C	
Mfor2								623 bp 50°C

Supplementary Table 2

Nei's genetic distance matrix of WMAI, WDAI, WTAI-CM3 and CMx haplotypes in the ten wheat genotypes.

Wheat genotype	Einkorn	Emmer	Turanicum	Peliss	Alzada	Spelt	Turkey Red	Judee	Marquis
Emmer	0.0274								
Turanicum	0.0224	0.0206							
Peliss	0.0296	0.0287	0.0167						
Alzada	0.0249	0.0205	**0.0085**	0.0124					
Spelt	0.0264	0.0133	0.0221	**0.0321**	0.0229				
Turkey Red	0.0256	0.0220	0.0212	0.0312	0.0212	0.0192			
Judee	0.0247	0.0272	0.0169	0.0286	0.0177	0.0237	0.0166		
Marquis	0.0281	0.0254	0.0212	0.0294	0.0211	0.0174	0.0130	0.0157	
Vida	0.0273	0.0237	0.0212	**0.0321**	0.0229	0.0201	0.0130	0.0210	0.0148

Figure captions

Figure 1.

Phylogenetic analysis of WMAI (A), WDAI (B), CMx (C) and WTAI-CM3 (D) haplotypes. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) was shown next to the branches. In red are indicated in which genotypes the haplotypes occurred. *T. turgidum* L. subsp. *durum* and *T. aestivum* L. subsp. *aestivum* genotypes are indicated with the abbreviations DU and AE respectively.

Figure 2.

Principal component analysis based on deduced WMAI, WDAI, WTAI-CM3 and CMx amino acid sequences. The scatter plot reports the projection of cases (10 wheat genotypes) on the first two components PC1 and PC2 (accounting for 63.6% of total variability): the diameter of each balloon is proportional to the variability observed for the wheat genotype. Each colour of the balloons corresponds to a different wheat ploidy: green for diploid wheat; red for tetraploid wheat; blue for hexaploid wheat.

Figure 3.

The dendrogram of the ten wheat genotypes based on Nei's genetic distance matrix.

Supplementary Figure 1.

The gene sequence of monomeric α-amylase inhibitor 0.28 (AJ223492) and the primers used to amplify WMAI genes. The sequences of signal peptide and mature protein are highlighted in red and blue respectively. The stop codon is marked with an asterisk.

Supplementary Figure 2.

Multiple sequence alignment of 16 WMAI haplotypes obtained from seven wheat genotypes and WMAI 0.28 (GenBank: AJ223492). The sequences related to the signal peptide have been included (residues 1 – 30 highlighted with a light blue box). The dots indicate conserved residues and the letters correspond to the substituted amino acid residues for each alignment gap. Blue arrows highlight the positions of the 10 Cys residues. The amino acids at the three inhibitor spots (Payan et al., 2004) are highlighted with yellow boxes. The two regions critical for the inhibitory activity (Garcia-Maroto et al., 1991) are marked with red rectangles.

Supplementary Figure 3.

Multiple sequence alignment of 35 WDAI haplotypes obtained from ten wheat genotypes with WDAI 0.19 (Uniprot: P01085) and 0.53 (Uniprot: P01084). The sequences related to the signal peptide have been included (residues 1 – 17 highlighted with a light blue box). The dots indicate conserved residues and the letters correspond to the substituted amino acid residues for each alignment gap. The pseudogenes have been included: in-frame stop codons are indicated with asterisks *; insertions and deletions are indicated with blue and red triangles respectively and the sequences after these changes until the end of the ORF of the corresponding functional gene is shown. Blue arrows highlight the positions of the 10 Cys residues. The amino acids at the three inhibitor spots are highlighted with yellow boxes.

Supplementary Figure 4.

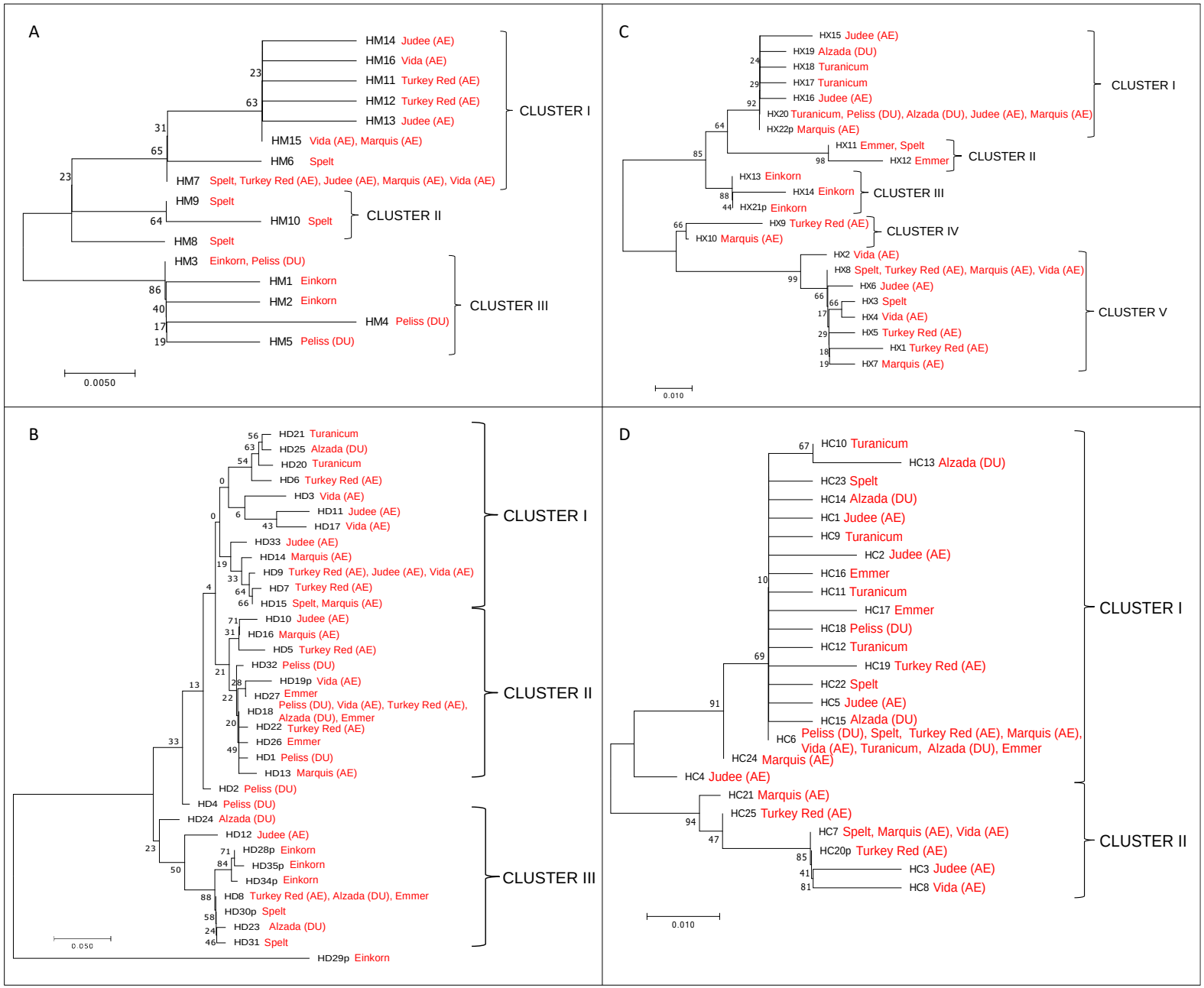
Multiple sequence alignment of 25 WTAI-CM3 haplotypes obtained from nine wheat genotypes and WTAI-CM3 (GenBank: X17574). The sequences related to the signal peptide have been included (residues 1 – 25 highlighted with a light blue box). The dots indicate conserved residues and the letters correspond to the substituted amino acid residues for each alignment gap. The asterisk * shows in-frame stop codons of pseudogene. Blue arrows highlight the positions of the 10 Cys residues.

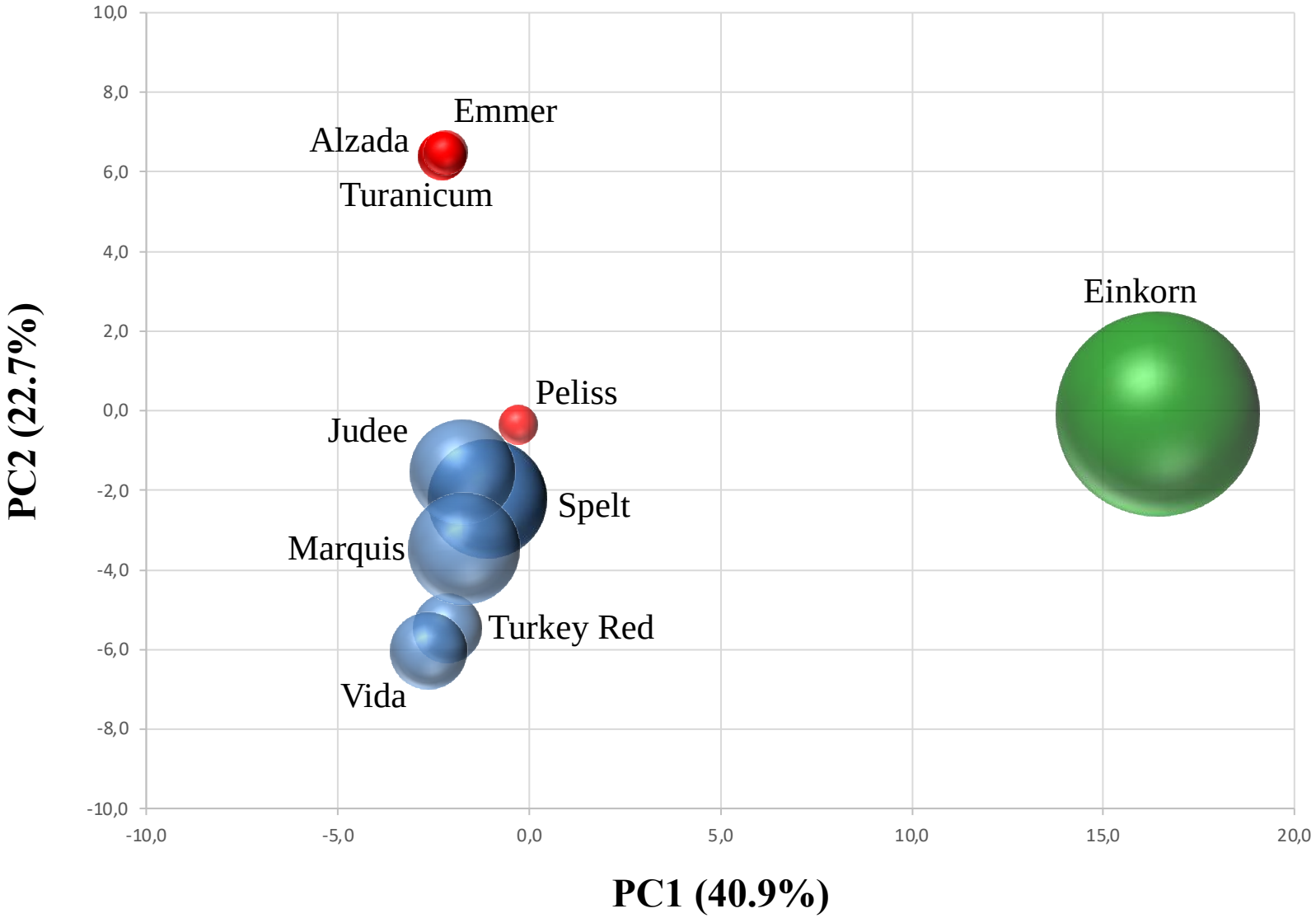
The conserved sequence DLPGCPRE is marked with a red rectangle, the amino acid residues of contact with TMA with yellow boxes and the ATI-TLR4 binding interface regions with blue rectangles.

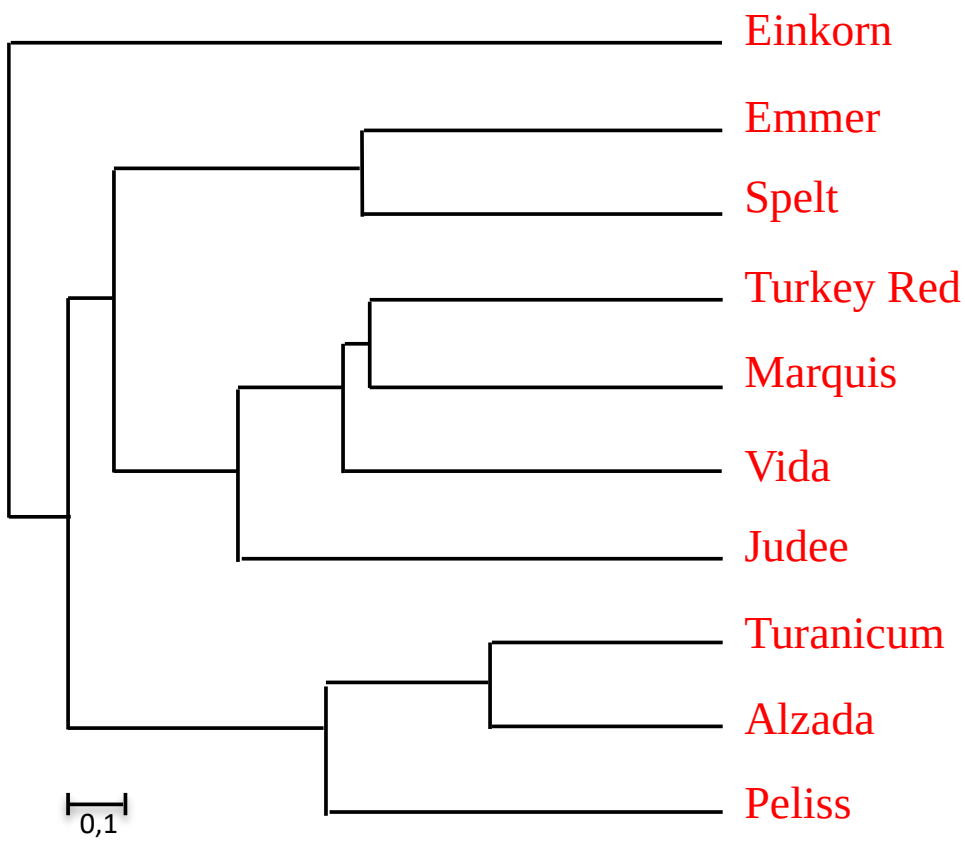
Supplementary Figure 5.

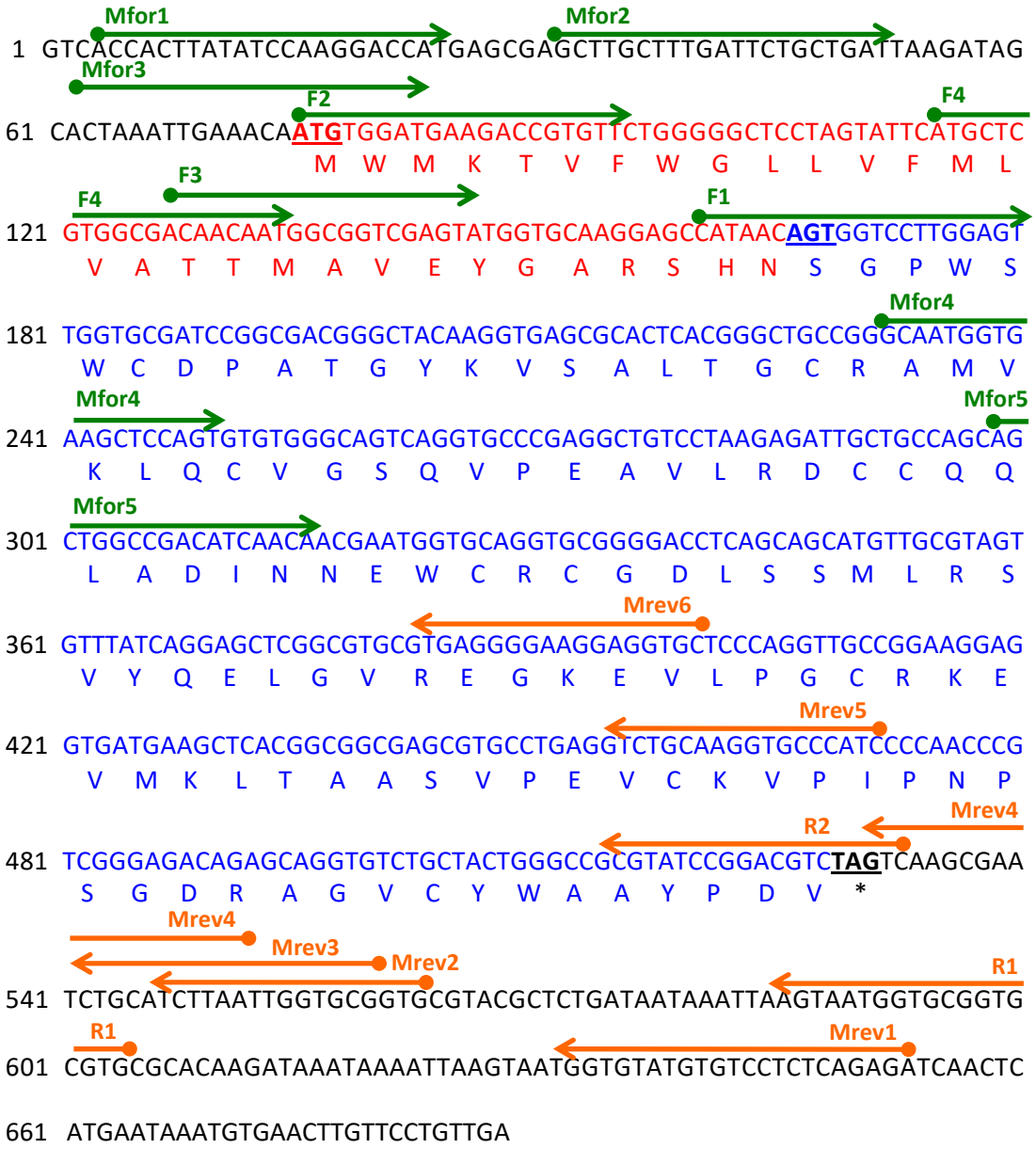
Multiple sequence alignment of 22 CMx haplotypes obtained from different wheat genotypes with CMx1/CMx3 (deduced from the sequence GenBank: X75608) and with CMx2 (deduced from the sequence GenBank: X75609). The sequences related to the signal peptide have been included (residues 1 – 24 highlighted with a light blue box).

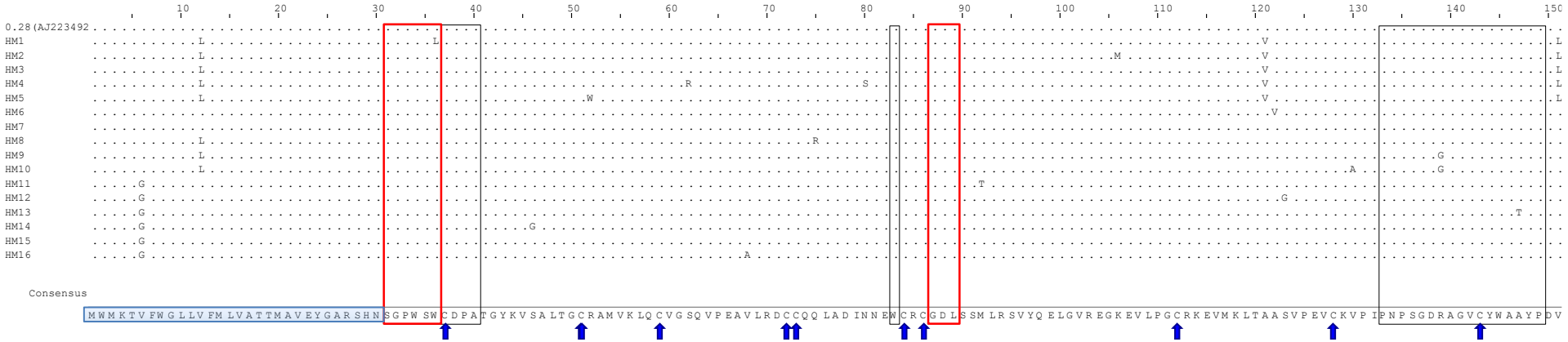
The asterisks * show in-frame stop codons. The red triangle shows the positions of the deletion found in the nucleotide sequence used to deduce the amino acid sequence. The dots indicate conserved residues and the letters correspond to the substituted amino acid residues for each alignment gap. Blue arrows highlight the positions of the 10 Cys residues.

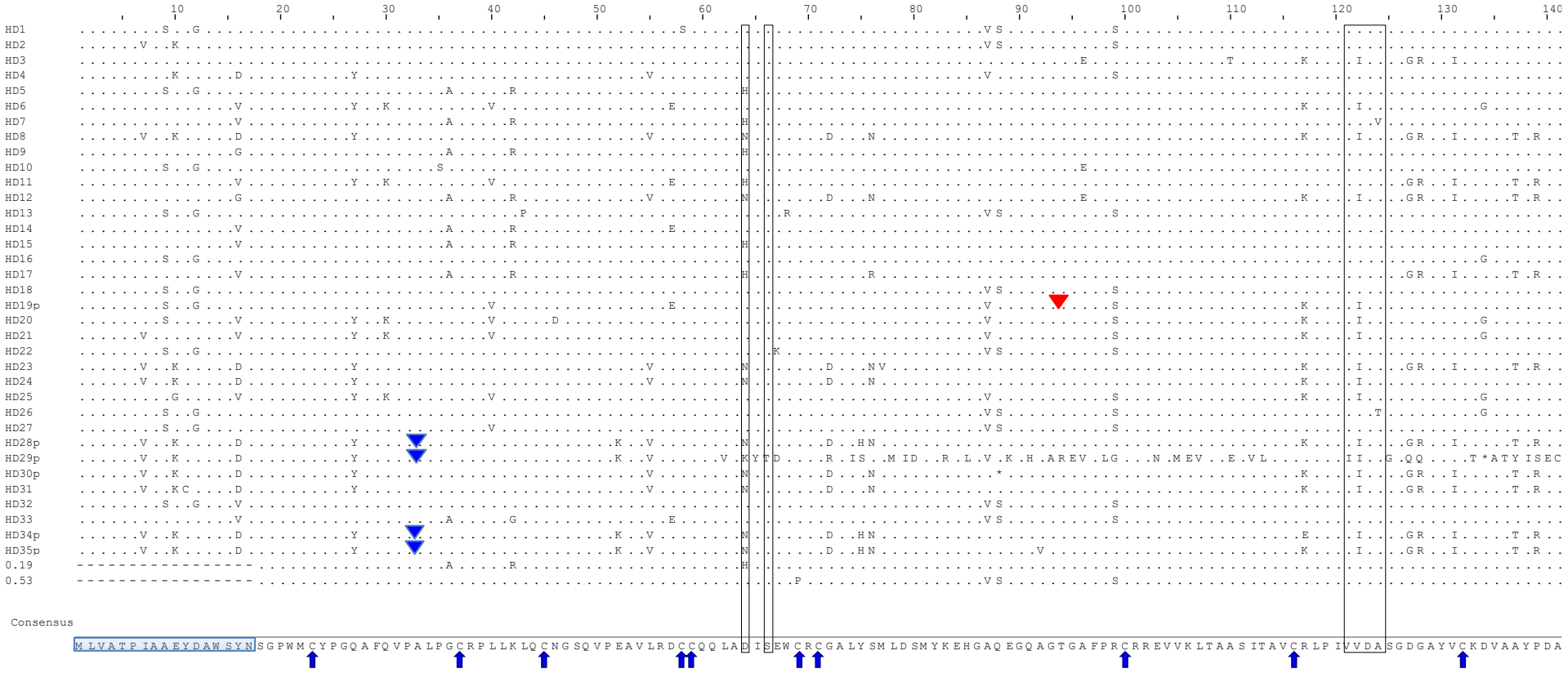

A
HM14 Judee (AE)
HM16 Vida (AE)
HM11 Turkey Red (AE)
HM12 Turkey Red (AE)
HM13 Judee (AE)
HM15 Vida (AE), Marquis (AE)
HM6 Spelt
HM7 Spelt, Turkey Red (AE), Judee (AE), Marquis (AE), Vida (AE)
HM9 Spelt
HM10 Spelt
HM8 Spelt
HM3 Einkorn, Peliss (DU)
HM1 Einkorn
HM2 Einkorn
HM4 Peliss (DU)
HM5 Peliss (DU)
CLUSTER I
CLUSTER II
CLUSTER III
0.0050
C
HX15 Judee (AE)
HX19 Alzada (DU)
HX18 Turanicum
HX17 Turanicum
HX16 Judee (AE)
HX20 Turanicum, Peliss (DU), Alzada (DU), Judee (AE), Marquis (AE)
HX22p Marquis (AE)
HX11 Emmer, Spelt
HX12 Emmer
HX13 Einkorn
HX14 Einkorn
HX21p Einkorn
HX9 Turkey Red (AE)
HX10 Marquis (AE)
HX2 Vida (AE)
HX8 Spelt, Turkey Red (AE), Marquis (AE), Vida (AE)
HX6 Judee (AE)
HX3 Spelt
HX4 Vida (AE)
HX5 Turkey Red (AE)
HX1 Turkey Red (AE)
HX7 Marquis (AE)
CLUSTER I
CLUSTER II
CLUSTER III
CLUSTER IV
CLUSTER V
0.010
B
HD21 Turanicum
HD25 Alzada (DU)
HD20 Turanicum
HD6 Turkey Red (AE)
HD3 Vida (AE)
HD11 Judee (AE)
HD17 Vida (AE)
HD33 Judee (AE)
HD14 Marquis (AE)
HD9 Turkey Red (AE), Judee (AE), Vida (AE)
HD7 Turkey Red (AE)
HD15 Spelt, Marquis (AE)
HD10 Judee (AE)
HD16 Marquis (AE)
HD5 Turkey Red (AE)
HD32 Peliss (DU)
HD19p Vida (AE)
HD27 Emmer
HD18 Peliss (DU), Vida (AE), Turkey Red (AE), Alzada (DU), Emmer
HD22 Turkey Red (AE)
HD26 Emmer
HD1 Peliss (DU)
HD13 Marquis (AE)
HD2 Peliss (DU)
HD4 Peliss (DU)
HD24 Alzada (DU)
HD12 Judee (AE)
HD28p Einkorn
HD35p Einkorn
HD34p Einkorn
HD8 Turkey Red (AE), Alzada (DU), Emmer
HD30p Spelt
HD23 Alzada (DU)
HD31 Spelt
HD29p Einkorn
CLUSTER I
CLUSTER II
CLUSTER III
0.050
D
HC10 Turanicum
HC13 Alzada (DU)
HC23 Spelt
HC14 Alzada (DU)
HC1 Judee (AE)
HC9 Turanicum
HC2 Judee (AE)
HC16 Emmer
HC11 Turanicum
HC17 Emmer
HC18 Peliss (DU)
HC12 Turanicum
HC19 Turkey Red (AE)
HC22 Spelt
HC5 Judee (AE)
HC15 Alzada (DU)
HC6 Peliss (DU), Spelt, Turkey Red (AE), Marquis (AE), Vida (AE), Turanicum, Alzada (DU), Emmer
HC24 Marquis (AE)
HC4 Judee (AE)
HC21 Marquis (AE)
HC25 Turkey Red (AE)
HC7 Spelt, Marquis (AE), Vida (AE)
HC20p Turkey Red (AE)
HC3 Judee (AE)
HC8 Vida (AE)
CLUSTER I
CLUSTER II
0.010

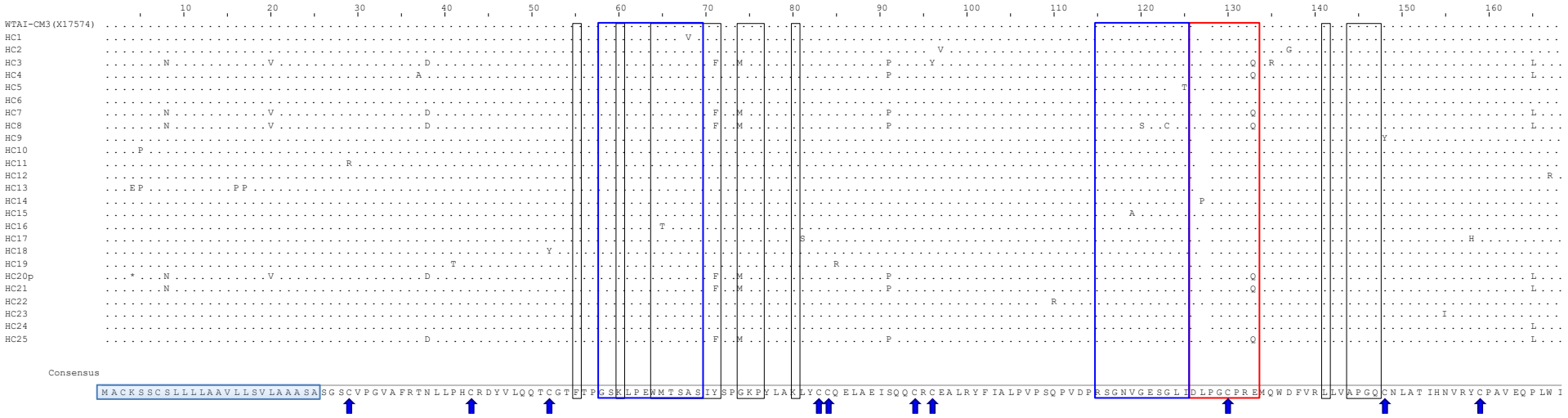

10,0
8,0
6,0
4,0
2,0
0,0
-2,0
-4,0
-6,0
-8,0
-10,0
-10,0
-5,0
0,0
5,0
10,0
15,0
20,0
PC2 (22.7%)
PC1 (40.9%)
Emmer
Alzada
Turanicum
Einkorn
Peliss
Judee
Spelt
Marquis
Turkey Red
Vida

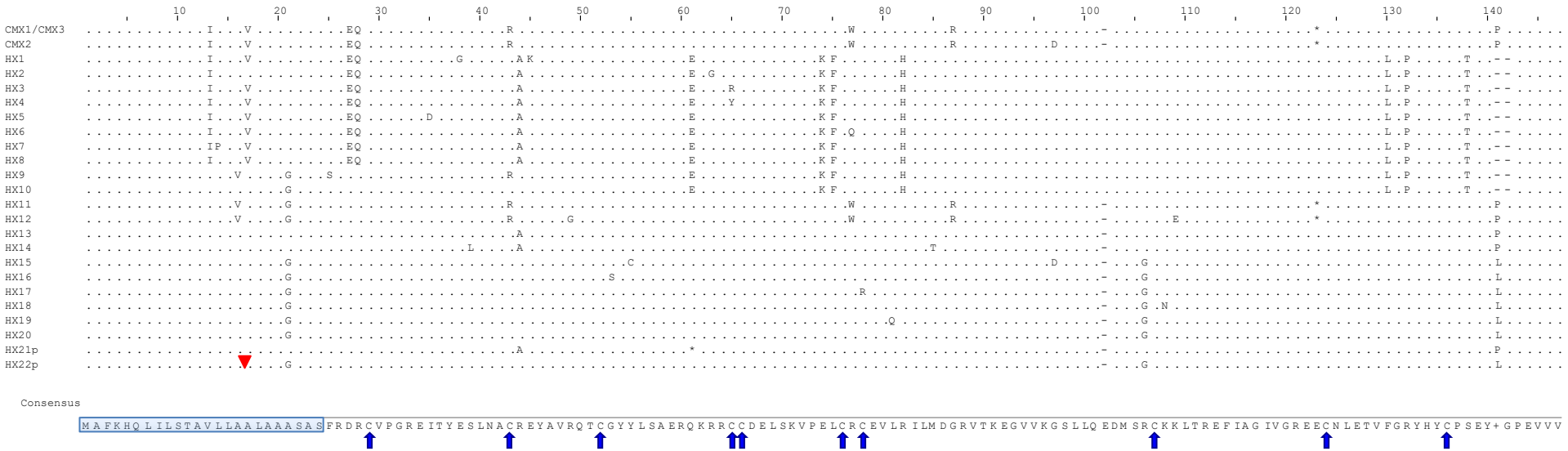

Einkorn
Emmer
Spelt
Turkey Red
Marquis
Vida
Judee
Turanicum
Alzada
Peliss
0,1


Mfor1
Mfor2
1 GTCACCACTTATATCCAAGGACCATGAGCGAGCTTGCTTTGATTCTGCTGATTAAGATAG
Mfor3
F2
F4
61 CACTAAATTGAAACAATGTGGATGAAGACCGTGTTCTGGGGGCTCCTAGTATTCATGCTC
M W M K T V F W G L L V F M L
F4
F3
F1
121 GTGGCGACAACAATGGCGGTCGAGTATGGTGCAAGGAGCCATAACAGTGGTCCTTGGAGT
V A T T M A V E Y G A R S H N S G P W S
Mfor4
181 TGGTGCGATCCGGCGACGGGCTACAAGGTGAGCGCACTCACGGGCTGCCGGGCAATGGTG
W C D P A T G Y K V S A L T G C R A M V
Mfor4
Mfor5
241 AAGCTCCAGTGTGTGGGCAGTCAGGTGCCCGAGGCTGTCCTAAGAGATTGCTGCCAGCAG
K L Q C V G S Q V P E A V L R D C C Q Q
Mfor5
301 CTGGCCGACATCAACAACGAATGGTGCAGGTGCGGGGACCTCAGCAGCATGTTGCGTAGT
L A D I N N E W C R C G D L S S M L R S
Mrev6
361 GTTTATCAGGAGCTCGGCGTGCGTGAGGGGAAGGAGGTGCTCCCAGGTTGCCGGAAGGAG
V Y Q E L G V R E G K E V L P G C R K E
Mrev5
421 GTGATGAAGCTCACGGCGGCGAGCGTGCCTGAGGTCTGCAAGGTGCCCATCCCCAACCCG
V M K L T A A S V P E V C K V P I P N P
Mrev4
R2
481 TCGGGAGACAGAGCAGGTGTCTGCTACTGGGCCGCGTATCCGGACGTCTAGTCAAGCGAA
S G D R A G V C Y W A A Y P D V *
Mrev4
Mrev3
Mrev2
R1
541 TCTGCATCTTAATTGGTGCGGTGCGTACGCTCTGATAATAAATTAAGTAATGGTGCGGTG
R1
Mrev1
601 CGTGCGCACAAGATAAATAAAATTAAGTAATGGTGTATGTGTCCTCTCAGAGATCAACTC
661 ATGAATAAATGTGAACTTGTTCCTGTTGA


10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
0.28(AJ223492
HM1
HM2
HM3
HM4
HM5
HM6
HM7
HM8
HM9
HM10
HM11
HM12
HM13
HM14
HM15
HM16
Consensus
MWMKTVFWGLLVFMLVATTMAVEYGARSHNSGPWSWCDPATGYKVSALTGCRAMVKLQCVGSQVPEAVLRDCCQQLADINNEWCRCGDLSSMLRSVYQELGVREGKEVLPGCRKEVMKLTAASVPEVCKVPIPNPSGDRAGVCYWAAYPDV


10 20 30 40 50 60 70 80 90 100 110 120 130 140
HD1
HD2
HD3
HD4
HD5
HD6
HD7
HD8
HD9
HD10
HD11
HD12
HD13
HD14
HD15
HD16
HD17
HD18
HD19p
HD20
HD21
HD22
HD23
HD24
HD25
HD26
HD27
HD28p
HD29p
HD30p
HD31
HD32
HD33
HD34p
HD35p
0.19
0.53
Consensus
MLVATPIAAEYDAWSYNSGPWMCYPGQAFQVPALPGCRPLLKLQCNGSQVPEAVLRDCCQQLADISEWCRCGALYSMLDSMYKEHGAQEGQAGTGAFPRCRREVVKLTAASITAVCRLPIVVDASGDGAYVCKDVAAYPDA


WTAI-CM3(X17574)
HC1
HC2
HC3
HC4
HC5
HC6
HC7
HC8
HC9
HC10
HC11
HC12
HC13
HC14
HC15
HC16
HC17
HC18
HC19
HC20p
HC21
HC22
HC23
HC24
HC25
Consensus


10 20 30 40 50 60 70 80 90 100 110 120 130 140
CMX1/CMX3
CMX2
HX1
HX2
HX3
HX4
HX5
HX6
HX7
HX8
HX9
HX10
HX11
HX12
HX13
HX14
HX15
HX16
HX17
HX18
HX19
HX20
HX21p
HX22p
Consensus
MAFKHQLILSTAVLLAALAAASASFRDRCVPGREITYESLNACREYAVRQTCGYYLSAERQKRRCCDELSKVPELCRCEVLRILMDGRVTKEGVVKGSLLQEDMSRCKKLTREFIAGIVGREECNLETVFGRYHYCPSEY+GPEVVV