Multiple Wheat Genomes Reveal Novel Gli-2 Sublocus Location and Variation of Celiac Disease Epitopes in Duplicated -Gliadin Genes

Halstead-Nussloch, Gwyneth; Tanaka, Tsuyoshi; Copetti, Dario; Paape, Timothy; Kobayashi, Fuminori; Hatakeyama, Masaomi; Kanamori, Hiroyuki; Wu, Jianzhong; Mascher, Martin; Kawaura, Kanako; Shimizu, Kentaro K; Handa, Hirokazu

Abstract: The seed protein -gliadin is a major component of wheat flour and causes gluten-related diseases. However, due to the complexity of this multigene family with a genome structure composed of dozens of copies derived from tandem and genome duplications, little was known about the variation between accessions, and thus little effort has been made to explicitly target -gliadin for bread wheat breeding. Here, we analyzed genomic variation in -gliadins across 11 recently published chromosome-scale assemblies of hexaploid wheat, with validation using long-read data. We unexpectedly found that the Gli-B2 locus is not a single contiguous locus but is composed of two subloci, suggesting the possibility of recombination between the two during breeding. We confirmed that the number of immunogenic epitopes among 11 accessions varied. The D subgenome of a European spelt line also contained epitopes, in agreement with its hybridization history. Evolutionary analysis identified amino acid sites under diversifying selection, suggesting their functional importance. The analysis opens the way for improved grain quality and safety through wheat breeding.

DOI: https://doi.org/10.3389/fpls.2021.715985

The following work is licensed under a Creative Commons: Attribution 4.0 International (CC BY 4.0) License.

Originally published at:
Halstead-Nussloch, Gwyneth; Tanaka, Tsuyoshi; Copetti, Dario; Paape, Timothy; Kobayashi, Fuminori; Hatakeyama, Masaomi; Kanamori, Hiroyuki; Wu, Jianzhong; Mascher, Martin; Kawaura, Kanako; Shimizu, Kentaro K; Handa, Hirokazu (2021). Multiple Wheat Genomes Reveal Novel Gli-2 Sublocus Location and Variation of Celiac Disease Epitopes in Duplicated -Gliadin Genes. Frontiers in Plant Science, 12:Epub-Epub.
DOI: https://doi.org/10.3389/fpls.2021.715985
Multiple Wheat Genomes Reveal Novel Gli-2 Sublocus Location and Variation of Celiac Disease Epitopes in Duplicated α-Gliadin Genes

Gwyneth Halstead-Nussloch1†, Tsuyoshi Tanaka2,3†, Dario Copetti1,4, Timothy Paape1,5, Fuminori Kobayashi2, Masaomi Hatakeyama1,6, Hiroyuki Kanamori2, Jianzhong Wu2, Martin Mascher7,8, Kanako Kawaura9, Kentaro K. Shimizu1,9* and Hirokazu Handa2,10*

1 Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland, 2 Institute of Crop Science, National Agriculture and Food Research Organization, Tsukuba, Japan, 3 Research Center for Advanced Analysis, National Agriculture and Food Research Organization, Tsukuba, Japan, 4 Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zürich, Zurich, Switzerland, 5 Brookhaven National Laboratory, Upton, NY, United States, 6 Functional Genomics Center Zurich, Zurich, Switzerland, 7 Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany, 8 German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany, 9 Kihara Institute for Biological Research, Yokohama City University, Yokohama, Japan, 10 Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, Kyoto, Japan

*Correspondence:
Hirokazu Handa
hirokazu@kpu.ac.jp
Kentaro K. Shimizu
kentaro.shimizu@uzh.ch
†These authors have contributed equally to this work

The seed protein α-gliadin is a major component of wheat flour and causes gluten-related diseases. However, due to the complexity of this multigene family with a genome structure composed of dozens of copies derived from tandem and genome duplications, little was known about the variation between accessions, and thus little effort has been made to explicitly target α-gliadin for bread wheat breeding. Here, we analyzed genomic variation in α-gliadins across 11 recently published chromosome-scale assemblies of hexaploid wheat, with validation using long-read data. We unexpectedly found that the Gli-B2 locus is not a single contiguous locus but is composed of two subloci, suggesting the possibility of recombination between the two during breeding. We confirmed that the number of immunogenic epitopes among 11 accessions varied. The D subgenome of a European spelt line also contained epitopes, in agreement with its hybridization history. Evolutionary analysis identified amino acid sites under diversifying selection, suggesting their functional importance. The analysis opens the way for improved grain quality and safety through wheat breeding.

Keywords: α-gliadin, celiac disease epitopes, copy number variation, Gli-2 loci, wheat (Triticum aestivum L.)

INTRODUCTION

Since its origin by allopolyploidization, bread wheat (Triticum aestivum L.) has become a staple crop, providing ∼20% of the calories consumed globally (Shiferaw et al., 2013). Concentrated breeding efforts have increased yield such that the production of bread wheat reached 766 million tons in 2019 (FAOSTAT, 2021). Further selection has made wheat more palatable and increased the quality of desired end-use traits.
Wheat grains are typically processed into flour to make various breads and noodles. Much of the rheological quality of these products relies on gluten formation. Gluten is a complex of two protein families, glutenins and gliadins, which are storage proteins in wheat endosperm. Gliadins are classified into three groups, α-, γ-, and Ω-gliadins, based on their electrophoretic mobility (Shewry and Halford, 2002). α-Gliadins are the most abundant gliadins and represent 15–30% of the wheat seed protein (Gu et al., 2004). Genes encoding α-gliadins are tandemly duplicated and form clusters within each Gli-2 locus. They are located on the short arm of homoeologous chromosome group 6 in *Triticum* species (Qi et al., 2013; Ozuna et al., 2015). Because *T. aestivalis* is allohexaploid (AABBDD), it contains three Gli-2 loci called Gli-A2, Gli-B2, and Gli-D2 (Payne, 1987). High allelic diversity, copy number variation, and expression differences in α-gliadins in bread wheat have been attributed to the combination of tandem and whole genome duplications (Salentinj et al., 2009; Noma et al., 2016; Huo et al., 2018). Although allelic diversity, gene copy number variation and other differences in α-gliadins may be linked to the phenotypic differences for the wheat flour qualities among the cultivars, little is known about precise genomic information for the Gli-2 loci to provide a basis for comparison between cultivars. For example, the general function of α-gliadins in the breadmaking quality is well understood, but the role of individual α-gliadin genes is not entirely clear (Branlard et al., 2001; Brennan, 2009; Metakovsky et al., 2018). α-Gliadins are also the most common proteins that trigger an immune response in patients with celiac disease (CD), one of the widespread wheat-related health disorders (Scherf et al., 2016). The CD reaction is caused by the presence of a variety of peptide sequences called epitopes (Solild et al., 2012; Juhász et al., 2018). The repetitive domain includes the DQ2.5-glia-α1, DQ2.5-glia-α2, and DQ2.5-glia-α3 epitopes (Ozuna et al., 2015), and sometimes, these epitopes overlap to create a 33-mer peptide that is highly immunotoxic to celiac patients (Shan et al., 2002; Huo et al., 2018; Juhász et al., 2018). The immunotoxicity of the 33-mer region was verified by genome-editing (Sánchez-León et al., 2018). Although the three-dimensional structure of a short CD epitope bound to human HLA has been reported (Kim et al., 2004; Petersen et al., 2014, 2016), little is known about the higher-order structure of gliadin proteins because they aggregate in solutions (Urade et al., 2018). Sequence-based characterization of α-gliadin variation within modern hexaploid wheat cultivars will aid in breeding efforts to incorporate both desired end-use quality and lower reactivity for consumers.

Allopolyploidization and tandem duplication have made regions such as Gli-2 difficult to characterize in terms of the genomic organization of and variation within multigene families found in bread wheat. Thus, most variation within α-gliadin gene sequences of different wheat accessions and related species has been detected using bacterial artificial chromosome (BAC) clones, transcriptome analysis, or low-coverage shotgun genome sequencing (Kawaura et al., 2012; Noma et al., 2016; Juhász et al., 2018). High resolution of the structure of homoeologous Gli-2 loci has been described using long-read sequences, but in only one cultivar, Chinese Spring (CS; Huo et al., 2018).

Recently, advances in polyploid genomics enabled the high-quality genome assembly and polymorphism analysis of tandem duplications (Paape et al., 2016, 2018; Avni et al., 2017). Here, using chromosome-level assemblies for 11 accessions including elite bread wheat cultivars and a spelt wheat line in the framework of the “10+ Wheat Genomes Project” (Walkowiak et al., 2020), we began to address the question of global variation in both the structure of and polymorphism within Gli-2 loci among multiple cultivars.

**MATERIALS AND METHODS**

**Sequence Resources**

Reference-quality genome assemblies for 9 bread wheat accessions, Arinai*rFor*, CDC Landmark, CDC Stanley, Jagger, Julius, LongReach Lancer, Mace, Norin 61, SY Mattis and one spelt accession, PI190962, released by the “10+ Wheat Genomes Project” (Walkowiak et al., 2020), were accessed through IPK, Germany.\(^1\) We also used the RefSeq v1.0 assembly of CS (International Wheat Genome Sequencing Consortium (IWGSC) et al., 2018), which is available at INRAE, France.\(^2\)

**Identification of Gli-2 Loci and α-Gliadin Sequences**

To identify the location of the Gli-2 loci, BLAST searches were conducted against chromosome assemblies for homoeologous group 6 of the eleven accessions using the α-gliadin gene sequences AS2 and AS7 (for chromosome 6A); AS3, AS4, AS5 and AS6 (for 6B); and AS1, AS8, AS9, AS10, and AS11 (for 6D) as queries (Noma et al., 2016). From the BLAST results, regions with an e-value = 0 and composed of a single exon were selected as candidates for α-gliadin gene copies. The regions were translated into amino acid sequences. Sequences not starting with a methionine residue were discarded as incomplete gene fragments. Sequences that were too diverged based on the sequence alignment or phylogenetic tree were also omitted. Finally, we constructed a codon-based alignment of gliadin gene copies using MUSCLE in MEGA (Kumar et al., 2018). Hi-C data and the alignments of CDC Landmark Oxford Nanopore Technologies (ONT) long-read data were obtained from Walkowiak et al. (2020). The gene coverage values of the long-read alignments were obtained with SAMtools v1.0 (Li et al., 2009) and BEDtools v2.29.0 (Quinlan and Hall, 2010). Read alignments were visualized with IGV v2.8.2 (Robinson et al., 2017).

The evolutionary history of the gene family was inferred from the 429 α-gliadin sequences identified above. Codon positions included were 1st + 2nd + 3rd + Non-coding. All positions containing gaps and missing data were eliminated (complete deletion option). The final dataset contained a total of 524 positions. The tree was estimated using the neighbor-joining method (Saitou and Nei, 1987), and evolutionary distances were computed using the Kimura 2-parameter method (Kimura, 1980).

---

1. https://wheat.ipk-gatersleben.de/
2. https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies
and expressed as the number of base substitutions per site. The rate variation among sites was modeled with a gamma distribution (shape parameter = 2.25). Support for the tree topology was estimated using the bootstrap test with 1,000 replicates and was calculated as the percentage of replicate trees in which the associated taxa clustered together (Felsenstein, 1985). The tree was drawn to scale, with the units for branch lengths being the same as those of the evolutionary distances used to infer the phylogenetic tree.

Celiac Disease Epitope Search and Site Selection Analysis

Using the amino acid sequences of α-gliadin copies without the last stop codon, we searched all sequences for the presence of nine canonical amino acid epitopes previously shown to induce an immunogenic reaction (Sollid et al., 2012; Ozuna et al., 2015).

To test for amino acid sites likely to be under positive selection in the α-gliadin gene family, only full-length sequences were considered for a conservative analysis. Gaps present at the same position in all three Gli-2 loci and sequences containing premature stop codons were discarded. Sequences were also removed if they had no terminal stop codon or were not composed of multiples of three nucleotides, implying frameshifts. Last, sites in regions that were difficult to align (the polyglutamine regions) were not considered, as the uncertain alignments may produce false positive signals. For the selection analysis, a phylogenetic method was applied. First, the most likely phylogenetic tree was estimated using nucleotide alignment and a general time reversible (GTR) + invariant + gamma model in MrBayes (Ronquist et al., 2012). Then, the likelihood of that tree was calculated under different codon substitution models by estimating the non-synonymous and synonymous substitution rate ratios (\(\omega = d_s/d_\lambda\)) for each codon within the alignment. The value of \(\omega\) indicates the type of selection: \(\omega < 1\) indicates negative selection, \(\omega = 0\) indicates neutral evolution, and \(\omega > 1\) indicates positive selection. A likelihood ratio test (LRT) was run between two nested codon substitution models, a null and an alternative model, to determine whether the alternative model of positive selection was supported. The null model (M7) did not allow for sites under positive selection while the alternative model (M8) did allow for positive selection (Yang et al., 2000). Last, the posterior probability of a specific site being under positive selection was estimated using Bayesian empirical Bayes (BEB) (Yang et al., 2005). Sites with a probability > 95% were considered significant. The likelihoods of the codon substitution models and posterior probability calculations were implemented in the CODEML program of the software package PAML4 (Yang, 2007).

RESULTS

Location and Validation of Gli-2 Loci in Accessions

We identified α-gliadin gene copies within 11 wheat assemblies: the 10 reference-quality pseudomolecule assemblies (Walkowiak et al., 2020) plus CS ReSeq v1.0 (International Wheat Genome Sequencing Consortium (IWGSC) et al., 2018). We first examined the chromosomal positions of the α-gliadin copies. Copies that mapped to chromosome 6A in the 11 wheat accessions were assigned as Gli-A2 and were located in single region on the short arm, as expected (Table 1 and Supplementary Table 1). The only exception was Gli-A2 of CDC Landmark, which was split into 2 subloci 7 Mb apart from each other. Similarly, sequences in Gli-D2 mapped to the expected region on chromosome 6D in nine reference-quality assemblies. In LongReach Lancer and CS, we could not identify α-gliadin copies on chromosome 6D; however, those found in scaffolds that were not anchored to a chromosome (chrUn) were assigned to Gli-D2 following the suggestion of Juhász et al. (2018). Surprisingly, the copies found on chromosome 6B showed that the Gli-B2 locus was clearly split into 2 subloci in all accessions. We called them Gli-B2-1 and Gli-B2-2, and they were 12–21 Mb apart from each other on chromosome 6B (Table 1 and Supplementary Table 1). The uniformity of the Hi-C signal along the whole Gli-B2 region and its flanking regions further supported that the bipartite structure of Gli-B2 was not an assembly artifact (Supplementary Figure 1) (Shimizu et al., 2020; Walkowiak et al., 2020). The position of this second locus relative to the well-described locus at ∼43 Mb on chromosome 6B in CS has not been described before, although previous studies mention two sequences that mapped outside that region (Huo et al., 2018; Juhász et al., 2018). The consistency in the Hi-C maps observed among all assemblies supports that Gli-B2 is composed of two parts and opens the possibility of exploiting genetic recombination for breeding purposes.

Although some assemblies showed the subdivision or translocation of several α-gliadin genes compared to those of other accessions, we interpreted them with caution. In Jagger, α-gliadin sequences mapped to a third region (Gli-B2-3) located at the end of the long arm of chromosome 6B. We also found that the sublocus Gli-B2-1 of LongReach Lancer and CDC Landmark was further split into two parts. However, the Hi-C signal of intra- and interchromosomal interactions for these accessions suggested potential misassembly within these regions (Supplementary Figures 1, 2). We note that the 20 Mb regions flanking Gli-B2-1 in CDC Landmark were highly concordant with those in CS, but they were not concordant with those in LongReach Lancer (Supplementary Figure 3). Because the assembly structure and orientation of CS was also supported by additional evidence (International Wheat Genome Sequencing Consortium (IWGSC) et al., 2018), this suggested that the rearrangement in CDC Landmark may represent true biological variation. The location and orientation of these subloci remain interesting cases for further validation to distinguish biological rearrangement from assembly errors.

Next, we checked the accuracy of the assemblies around each single α-gliadin gene copy. We utilized the long-read sequence data from ONT for CDC Landmark that was previously used to validate the assembly (Walkowiak et al., 2020). Though often flanked by assembly gaps, the sequence at and immediately adjacent to each α-gliadin gene copy was supported by continuous alignments of several long reads (Supplementary Figure 4), implying a gene-level correctness.
of each model. The different coverage seen among copies, including those in close proximity, hinted at potential collapses of paralogous copies into a single gene (Supplementary Figure 4B) or the separate assembly of allelic heterozygous copies. To address this possibility, we compared the coverage of the ONT alignments for each α-gliadin gene in the assembly to the median genome-wide gene coverage (32.23 genome equivalents). Of the 34 copies that we manually annotated in CDC Landmark, seven (20.5%) had a mean coverage that clearly deviated from that of other copies. As a comparison, the coverage of the three ADH homoeologs (chosen as a single-copy gene reference) was well within the genome-wide value (Supplementary Table 1). While the three α-gliadin copies at high coverage likely represent collapsed paralogs, the four genes at lower coverage may be haplotype-specific assemblies of heterozygous allelic copies. The long-read data suggested that the assembled α-gliadin sequences were correctly identified, although the exact copy number of ~20% of them may be different.

The number of assembled α-gliadin genes within each Gli-2 locus is reported in Table 1. While most accessions possessed approximately 11 α-gliadin copies in Gli-A2, the accessions Arina/LrFor, Norin 61, SY Mattis, and P1190962 had two to three times as many copies (Table 1). We identified 13–17 copies in Gli-B2 in most accessions, while Arina/LrFor, Julius, SY Mattis and the European spelt P1190962 had only half the number of copies compared to the other accessions (Table 1). For Gli-D2, there were approximately 10 copies in most accessions (Table 1). An extremely high or low copy number for Gli-B2 and Gli-D2 in Jagger, respectively, was possibly an assembly error, as described above. Subsequent analyses in this paper will use the assignment to a particular Gli-2 locus based on previously published assemblies for consistency.

### Genomic Variation of α-Gliadin Genes

We then assessed the relationship between all α-gliadin copies identified in the 11 accessions using phylogenetic analysis. According to the clustering pattern, α-gliadin copies were classified into three main clades named 1, 2, and 3 (Figure 1). Clades 1 and 3 showed a compact structure and included copies mostly from Gli-A2 to Gli-D2, respectively. In Gli-D2, unlike other subgenome loci, there was little difference in copy number between accessions and the genetic distances between branches were shorter. Limited allelic diversity at the Gli-D2 locus is consistent with the lower diversity of the coding sequences in the D subgenome (International Wheat Genome Sequencing Consortium (IWGSC), 2014; Jordan et al., 2015; Walkowiak et al., 2020). Clause 2 mostly contained copies from Gli-B2 but also included sub-clades of Gli-A2 and Gli-D2, although with weak branch support (Figure 1). As mentioned above, we found 2 subloci in Gli-B2, i.e., Gli-B2-1 and Gli-B2-2 (Table 1). In the phylogenetic tree (Figure 1), the two gliadin sequences encoded in Gli-B2-2 formed subclades distinct from other sequences in Gli-B2-1, indicating that the split of Gli-B2 was shared among all wheat accessions and that the genes in the two subloci experienced different histories. These data further support the bipartite structure of Gli-B2.

### Phylogenetic Analysis of α-Gliadin Copies

The α-gliadin copies that mapped to Jagger Gli-B2-3 clustered in clade 3, which is composed of Gli-D2 copies (Figure 1, blue arrows). This is consistent with the possible misassembly from Gli-D2 to the end of chromosome 6BL in this cultivar (see also the previous section; Supplementary Figure 2). If we reassigned these sequences as a part of Gli-D2, the copy numbers of Gli-B2 and Gli-D2 in Jagger are closer to the average copy number found within the other accessions. In the
case of LongReach Lancer and CS, we assigned all α-gliadin copies in chrUn as copies of Gli-D2 following the suggestion of Juhász et al. (2018). However, we found that several copies clustered with those assigned to Gli-A2 or Gli-B2 in the other accessions (Figure 1, orange arrows). Despite these potential misclassifications, we were able to show that there were clear variations among accessions.

Among the accessions with the largest differences in copy numbers, we observed distinct clustering patterns. Branches from accessions with the highest copy number for Gli-A2, such as ArinaLrFor, Norin 61, SY Mattis and PI190962, were clearly separated from the branches of the other seven accessions. Similarly, we found distinct clusters containing copies of Gli-B2-1 from ArinaLrFor, Julius, and SY Mattis. These three accessions, in addition to PI190962, contained the lowest copy number within this locus. These examples highlight potential differences in evolutionary and/or breeding history between accessions and that the gene duplications or losses in some cultivars did not originate independently but were likely from a common ancestor.

Focusing on the spelt wheat, PI190962, we observed no clear association with the other accessions for Gli-A2 and Gli-B2-1. Rather, most of the copies in PI190962 formed their own branches or small clusters. This was not the case for Gli-D2, where the PI190962 copies were positioned on the same branches as those for other bread wheats. Interestingly, the α-gliadin copies in Gli-B2-2 in PI190962 also clustered with those of the other accessions (Figure 1, purple highlight). The spelt accession, PI190962, used in this study is a Central European spelt, which has been suggested to have originated from the introgression of a hulled tetraploid emmer wheat into bread wheat during the migration of bread wheat from the Fertile Crescent to Europe. Therefore, the A and B subgenomes between bread wheat and European spelt had higher sequence divergence,
while the D subgenome showed greater sequence similarity (Blatter et al., 2004; Dvorak et al., 2012). Our observation of the separation of α-gliadin sequences in PI190962 from those of other bread wheats in Gli-A2 and Gli-B2-1 supports that this Central European spelt accession had an introgressed origin with a tetraploid emmer wheat, which was recently shown to be distinct from the origins of Iberian spelt (Abrouk et al., 2021). This result also indicates that the introgressed loci from chromosome 6B of emmer wheat may be confined to the region encoding Gli-B2-1, further supporting a different evolutionary history for the two Gli-B2 loci identified in this study.

**Celiac Disease Epitope Copy Number and Positive Selection in α-Gliadins**

Specific epitopes found in α-gliadins can induce reactions in patients with CD and gluten intolerance. Therefore, the search for new alleles and/or copy number variations that may cause weaker or no reaction is beneficial in breeding programs. Among the amino acid sequences produced by α-gliadin genes from the 11 wheat accessions, we found polymorphic sites within three major immunogenic regions, p31-43, the 33-mer, and the DQ2.5-glia-α3 peptide, using the established nomenclature (Sollid et al., 2012; Ozuna et al., 2015). The presence of epitope sequences showed a subgenome-specific pattern within the 11 accessions (Figure 2), and the count of CD epitopes in each accession mirrored the total α-gliadin copy number present in each locus (Table 1). Gli-A2 contained mostly DQ2.5-glia-α1b, DQ2.5-glia-α3 and p31-43-LG epitopes. Variants of the latter two epitopes were also present, but at low frequency, in four accessions. The B subgenome encoded the fewest epitopes, the highest proportion of which were p31-43-PG. Among all accessions, the largest variety of CD epitopes was present in Gli-D2 and included several that overlapped in a single gene copy. The toxic 33-mer sequence that contains six epitopes (33-mer 1.3-6) was found in the Gli-D2 sequence of 5 accessions, including once in PI190962 and twice each in CS and LongReach Lancer (Figure 1, red curve and Figure 2).

In general, a single known epitope sequence was not found in the genes of all three subgenomes and sequences with multiple overlapping epitopes were restricted to Gli-D2. For example, DQ2.5gla-α1a was present in both Gli-A2 and Gli-D2. The sequence encoding both DQ2.5gla-α1a and DQ2.5gla-α2 (PFPQPQLPYPQ) was found only in Gli-D2 due to a P to S substitution (PFPQPQLPYSQ) in Gli-A2. No DQ2.5-glia-α-type epitopes were present in Gli-B2, except for the potential misassembly or translocation of regions from chromosomes 6D to 6B in Jagger. The patterns we observed reflected those of
previous studies reporting the presence of specific epitopes in the subgenomes of hexaploid wheat (van Herpen et al., 2006; Salentijn et al., 2009; Sollid et al., 2012; Ozuna et al., 2015; Noma et al., 2016; Juhász et al., 2018).

Global prevalence of CD has increased (Singh et al., 2018) and this has been attributed, by some, to modern breeding practices. Due to its hybridization history, spelt wheats contain different gliadin and glutenin contents and has been subject to less intensive selection than modern bread wheats (Dubois et al., 2016; Escarnot et al., 2018), prompting the idea it could be less reactive for consumers. We observed that the numbers and distribution patterns of the immunogenic epitopes in the particular spelt accession, PI190962, were similar to those of other bread wheat accessions (Figure 2), including one copy of the 33-mer peptide that was identified in Gli-D2. Although the study of Asian and other spelts (Blatter et al., 2004; Dvorak et al., 2012) would be necessary to draw conclusions about spelt diversity, the data from this single accession of spelt did not support the claim that spelt (as a species) could produce weaker reactions in people with CD, in agreement with previous genetic

---

**FIGURE 3** Sites under positive selection in α-gliadin. Amino acids under selection were detected by estimating the ratio of non-synonymous (dN) to synonymous substitutions (dS) in each codon in the α-gliadin alignment for each subgenome. The panels show the Bayesian posterior probability of an amino acid site being under positive selection for the gliadin genes of each subgenome separately and the alignment of all three together (ABD). Highlighted in blue are the sites with a posterior probability > 95%. The different conserved domains (from Noma et al., 2016) are marked at the top of the panels. Sites that were difficult to align in the polyglutamine domains are excluded.
studies (Ozuna et al., 2015; Dubois et al., 2016; Ruiz-Carnicer et al., 2019). Recent studies investigating overall protein and gluten content of both modern and old hexaploid wheat as well as “ancient” varieties including spelt, emmer, and einkorn showed no conclusive role of modern breeding techniques in the increased prevalence of CD. Rather, they exemplified the high variability of gluten content between all varieties, new and old, and reiterate the importance of environmental factors in overall protein content of wheat and its relatives (Escarnot et al., 2018; Geißlitz et al., 2019; Call et al., 2020; Pronin et al., 2021). Our results on the genetic variability are in line with these protein-based studies and, taken together, show the tools to identify low immunoreactive varieties are well developed. These studies not only proposed suitable varieties for further breeding already but also motivate a more comprehensive characterization of wheat and its relatives to tap into existing variability for breeding (Shewry, 2018).

We used a method to identify selection on amino acid-changing substitutions (PAML; Yang, 2007). This method estimates the ratio of amino-acid replacement mutations (non-synonymous substitutions, dN) compared with synonymous substitutions (dS). When the dN/dS ratio is greater than 1, it indicates positive or diversifying selection. Many positions showed a posterior probability higher than 0.75. Among them, in Gli-B2, we found two codon positions that were above the 95% significance level: one in unique domain I and another in unique domain II (Figure 3). When all Gli-2 loci were analyzed together, the position in unique domain I remained significant (Figure 3 and Supplementary Table 2). The other amino acid position that was significant in Gli-B2 domain II was just below the threshold when A, B and D were analyzed together (Figure 3).

DISCUSSION

The importance of bread wheat in human nutrition and its role in disease warrant the characterization of genetic and structural variation within the gene family encoding gliadin, which forms the gluten protein structure together with glutenin. However, this research has been challenging due to the complexity of the loci caused by tandem and homoeologous duplications. Here, we characterized the diversity of α-gliadin gene copies and their organization within Gli-2 loci in chromosome-scale assemblies of 11 globally distributed bread and spelt wheat accessions. Long-read data supported that the assembled gliadin coding regions were correct, and 80% of them were assembled as a single copy with high confidence. The remaining 20% may possibly be collapsed, highly similar paralogs or independently assembled alleles of a gene copy. Unexpectedly, we found a bipartite structure of the Gli-B2 loci in all assemblies, which was supported by Hi-C data and evolutionarily supported by phylogenetic analysis. This suggests that further expansion of the variation at the gliadin locus through chromosomal recombination using the segregation of these subloci may be applicable for future wheat breeding. Using the PAML method, we detected amino acid positions that were under diversifying selection, suggesting that polymorphisms at these positions may be relevant for functional differences, such as those involved in interactions with glutenins (Li et al., 2014). This warrants further functional validation via amino acid substitution experiments.

Previous reports describe the subgenome specificity of sequences with CD epitopes, and those that cause the strongest immune cell reactions occur mostly in the A and D subgenomes and their respective progenitors. On the other hand, the wheat B subgenome, barley and several other Triticeae species contain epitopes that produce relatively weak responses from their α-gliadin and related proteins (van Herpen et al., 2006; Juhász et al., 2018). Our results not only reflect this subgenome specificity but also show that epitopes causing gluten-related reactions are unevenly distributed among accessions covering a wide range of wheat diversity (Walkowiak et al., 2020). The D subgenome is the only identified source of the toxic 33-mer epitope within bread wheat, and its presence has been detected at low frequency in the germplasm of the D progenitor Aegilops tauschii (Schaart et al., 2021). Current efforts to incorporate this knowledge into breeding safer varieties include the generation of synthetics and Gli-D2 deletion lines (Camerlenko et al., 2017; Li et al., 2018), the development of probes to quickly confirm the presence of reactive epitopes (Dubois et al., 2017), and the genome-editing to reduce the immunotoxic 33-mer (Sánchez-León et al., 2018). Our study can inform these efforts. Our results show the reduced frequency of reactive epitopes in some accessions but also show that reactive epitopes are present in spelt, which is consistent with a previous study (Escarnot et al., 2018), indicating that detailed cultivar-specific analysis is needed. While the immunogenic effects of many of the polymorphic epitopes have not been directly tested, our main findings indicate that resources for breeding less reactive wheat are already present in elite germplasm.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: IPK, Germany (https://wheat.ipk-gatersleben.de/) INRAE, France (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies).

AUTHOR CONTRIBUTIONS

KKS and HH conceived the study. TT, FK, HK, JW, and KK performed the identifications and phylogenetic analyses of α-gliadin genes. GH-N, DC, TP, and MH performed the epitope identification and evolutionary analyses within α-gliadins. DC and MM performed and curated the bioinformatic data. KK, KKS, and HH acquired the funding. GH-N, DC, KKS, and HH wrote the original draft. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by JST CREST (JPMJCR1603), the Swiss National Science Foundation 31003A_182318, MEXT
ACKNOWLEDGMENTS

We thank Sean Walkowiak and Curtis Pozniak (Crop Development Centre, University of Saskatchewan, Saskatoon, Saskatchewan, Canada) for sharing the long-read alignments, and the IT Support Group D-HEST, ETH Zürich, for computational support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.715985/full#supplementary-material

REFERENCES

Abrouk, M., Athiyanann, N., Müller, T., Pailles, Y., Stritt, C., Roulin, A. C., et al. (2021). Population genomics and haplotype analysis in spelt and bread wheat identifies a gene regulating glume color. *Commun. Biol.* 4, 1–11. doi: 10.1038/s42003-021-01908-6

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S. O., Gundlach, H., et al. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* 357, 93–97. doi: 10.1126/science.aaf0032

Blatter, R. H. E., Jacomet, S., and 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. doi: 10.1007/s00122-003-1441-7

Branlard, G., Dardevet, M., Saccomano, R., Lagoutte, F., and Gourdon, J. (2001). Genetic diversity of wheat storage proteins and bread wheat quality. *Euphytica* 119, 59–67. doi: 10.1023/A:1017586220359

Brennan, C. (2009). Gliadin and Glutenin: the unique balance of wheat quality. *Int. J. Food Sci. Technol.* 44, 1865–1866. doi: 10.1111/j.1365-2621.2007.01703.x

Call, L., Kapeller, M., Grausgruber, H., Reiter, E., Schoenlechner, R., and D’Amico, S. (2020). Effects of species and breeding on wheat protein composition. *J. Cereal Sci.* 103:2974. doi: 10.1016/j.jcs.2020.102974

Camerlengo, F., Sestili, F., Silvestri, M., Colaprico, G., Margiotta, B., Ruggeri, R., et al. (2017). Production and molecular characterization of bread wheat lines with reduced amount of α-tgliadins. *BMC Plant Biol.* 17:248. doi: 10.1186/s12870-017-1211-3

Dubois, B., Bertin, P., and Mingote, D. (2016). Molecular diversity of α-gliadin expressed genes in genetically contrasted spelt (*Triticum aestivum ssp. spelta*) accessions and comparison with bread wheat (*T. aestivum ssp. aestivum*) and related diploid Triticum and Aegilops species. *Mol. Breeding* 36:152. doi: 10.1007/s11032-016-0569-5

Dubois, B., Bertin, P., Muhovsky, Y., Escarnot, E., and Mingote, D. (2017). Development of TaqMan probes targeting the four major celiac disease epitopes found in α-gliadin sequences of spelt (*Triticum aestivum ssp. spelta*) and bread wheat (*Triticum aestivum ssp. aestivum*). *Plant Methods* 13:72. doi: 10.1186/s13007-017-0222-2

Dvorak, J., Deal, K. R., Luo, M.-C., You, F. M., von Borstel, K., and Dehghani, H. (2012). The origin of spelt and free-threshing hexaploid wheat. *J. Hered.* 103, 426–441. doi: 10.1093/jhered/esr152

Escarnot, E., Gofflot, S., Sinnaeve, G., Dubois, B., Bertin, P., and Mingote, D. (2018). Reactivity of gluten proteins from spelt and bread wheat accessions towards A1 and G12 antibodies in the framework of celiac disease. *Food Chem.* 268, 522–532. doi: 10.1016/j.foodchem.2018.06.094

FAOSTAT (2021). Available online at: http://www.fao.org/faostat/en/#data/QC (accessed February 8, 2021).

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 793–791. doi: 10.2307/2408678

Geisslitz, S., Longin, C. F. H., Scherf, K. A., and Koehler, P. (2019). Comparative study on gluten protein composition of ancient (einkorn, emmer and spelt) and modern wheat species (*Durum* and *Common Wheat*). *Food* 84:09. doi: 10.3390/foods8090409

Gu, Y. Q., Crossman, C., Kong, X., Luo, M., You, F. M., Coleman-Derr, D., et al. (2004). Genomic organization of the complex α-gliadin gene loci in wheat. *Theor. Appl. Genet.* 109, 648–657. doi: 10.1007/s00122-004-1672-2

Huo, N., Zhu, T., Altenbach, S., Dong, L., Wang, Y., Mohr, T., et al. (2018). Dynamic evolution of α-Gliadin prolamin gene family in homoeologous genomes of hexaploid wheat. *Sci. Rep.* 8:5181. doi: 10.1038/s41598-018-13570-5

International Wheat Genome Sequencing Consortium (IWGSC) (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum* genome). *Science* 345:1231788. doi: 10.1126/science.1251788

International Wheat Genome Sequencing Consortium (IWGSC), Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361:eaaar7191. doi: 10.1126/science.aar7191

Jordan, K. W., Wang, S., Lun, Y., Gardiner, L.-J., MacLachlan, R., Hucl, P., et al. (2015). A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol.* 16:48. doi: 10.1186/s13059-015-0606-4
Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Halstead-Nussloch, Tanaka, Copetti, Paape, Kobayashi, Hatakeyama, Kanamori, Wu, Mascher, Kawaura, Shimizu and Handa. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.