Identification of Genome Specific Sequence Motifs in α-Gliadins and Wheat Accessions with Less Celiac Disease Epitopes

S. Singh¹*, S. Ram², and S. Narwal²

ABSTRACT

Among gliadins, α-gliadins are important active proteins in triggering celiac disease in human beings owing to the presence of toxic epitopes. A set of 177 α-gliadin gene sequences and the corresponding proteins were analyzed. Twenty accessions of hexaploids including 1, 14, and 5, respectively representing A, B, and D, with no intact CD-epitopes in α-gliadins, were identified. Twenty-two and 13 conserved motifs in non-repetitive domains NR1 and NR2, respectively, of α-gliadins differentiated all the amino acid sequences encoded by A genome of both diploids and hexaploids. Most of the amino acid sequences encoded by D genome (70 of 75 in hexaploids and 13 of 16 in diploids) could be identified by 22 amino acid motif. Large variations and lesser number of intact CD-epitopes was observed for α-gliadins belonging to B genome. As compared to diploids, repeat length of polyglutamine repetitive domain QII of B genome was lower in hexaploids indicating loss of Q residues during evolution of hexaploid wheat. The information can be used in assigning any α-gliadin sequences onto A, B, and D genomes and identifying wheat accessions with lesser CD-epitopes. The result presented here will be useful for the wheat improvement programs aiming for the management of celiac disease in human beings.

Keywords. Conserved motifs, Genome, Triticum aestivum, Wheat improvement.

INTRODUCTION

Wheat is one of the important cereal crops in the world providing energy and nutrition to human beings. Large numbers of end use products such as chapati, bread, biscuit, noodle, and pasta products are made from wheat because of the presence of gluten. Gluten is a visco-elastic complex formed when wheat flour is mixed with water. Gluten is formed by the interactions of gluten proteins in which millions of subunits of glutenins and gliadins are linked through di-sulphide linkages. The glutenins are subdivided into High Molecular Weight (HMW) and Low Molecular Weight (LMW) glutenins and the gliadins into α, β, γ and ω gliadins (Shewry and Halford, 2002). However, some components of gluten induce celiac disease (gluten intolerance) in susceptible human populations across the world. At present, many gluten derived T-cell stimulatory peptides are known which originate from the α and γ gliadins, and the HMW and LMW glutenins (Sjostrom et al., 1998; Vande Wal et al., 1998; Paulsen et al., 1995; Molberg et al., 2003; Anderson et al., 2012). The α-gliadins, with average molecular weight of 31 kDa, are most abundant and represent 15–30% of the total wheat grain proteins. Therefore, α-gliadins are the most consumed storage proteins by human beings (Chen et al. 2008; Gu et al. 2004; Van Herpen et al., 2006).
The α-gliadins are reported to be the most active proteins in triggering celiac disease owing to the presence of several peptides, referred to as toxic epitopes, that constitute the main toxic components in celiac disease (Arentz-Hansen et al., 2000; Vader et al., 2003; Vaccino et al., 2009; Li et al., 2010; Xie et al., 2010; Kawaura et al., 2012; Salentijn et al. 2013; Qi et al., 2013; Li et al., 2014a; Li et al., 2014b). Since the only efficient therapy for celiac disease patients is a life-long gluten-free diet (Mc Manus and Kelleher, 2003), foods made from wheat flour with no or a few toxic epitopes are likely to be better tolerated, thereby improving a celiac disease patient’s quality of life (Molberg et al. 2005; Spaenij-Dekking et al., 2005). However, this would be a formidable challenge to develop “celiac-safe” wheat due to the complex multi-genic control of gluten protein composition (Shewry and Tatham, 2016). To begin with, there is need to identify wheat accessions with less number of celiac disease toxic epitopes (CD-epitopes). Therefore, knowledge of the number of CD-epitopes in α-gliadins and their location with respect to A, B, and D genomes is required. Though earlier reports by Van Herpen et al. (2006) and Vader et al. (2003) showed that size and distribution of toxic epitopes and the number of polyglutamine residues could be used to assign α-gliadin sequences to specific chromosomes, this could not differentiate α-gliadin sequences from large numbers of wheat accessions in this study. In this investigation, efforts were aimed to identify amino acid motifs differentiating α-gliadin sequences representing A, B, and D genomes. This would help to identify accessions with nil or fewer number of CD-epitopes representing diploid progenitors and hexaploid wheats.

MATERIALS AND METHODS

A set of 177 α-gliadin gene sequences and the corresponding proteins were retrieved from the NCBI database (http://www.ncbi.nlm.nih.gov/protein/). The sequences represented 112 hexaploid bread wheat (T. Aestivum) genotypes and 65 ancestral diploid progenitors of A (Tmonococcum, 30), B (Aegilopspspeltoides & A. Searssii, 19) and D (Aegilopstauschii, 16) genomes. The sequences were assembled and aligned using ClustalW (Thompson et al., 1994) and manually curated using BioEdit version 7.0.9.0 (http://www.mbio.nesu.edu/BioEdit/Bioedit.html). Nucleotide diversity analysis of α-gliadin genes was conducted by Tajima D test using MEGA 6.0 (Tajima, 1989). The Multiple Sequence Alignment (MSA) of all the α-gliadin proteins was performed to construct a phylogenetic tree by MEGA (version 6.0) using default parameters and Maximum Likelihood (ML) method (Tamura et al., 2013). The stability of branch nodes in the ML-tree was measured by performing bootstrap test of 1,000 replications to ensure a high confidence range and accuracy. The phylogenetic tree was further supported by Multiple Expectation-Maximization for Motif Elicitation (MEME) (Timothy et al., 1994). About twenty different conserved motifs ranging from 6 and 50 amino acids in length were detected by MEME software tool.

The identification of the four major immunogenic epitopes (Glia-a, α2, α9 and α20) and polyglutamine repeats (QI & QII) in α-gliadins was done as per the method of Van Herpen et al. (2006). The t-test was used for identifying the significant differences in the repeat length of polyglutamine (QI & QII) repeats in α-gliadins encoded by different genomes.

RESULTS

Identification and Frequency Distribution of Four Major Celiac Disease Toxic Epitopes in α-Gliadins

The entire set of α-gliadin protein sequences belong to hexaploid bread wheat (T. aestivum) and their ancestral A, B and D
Identification of Genome Specific Sequence Motifs

Table 1. Genome-wise distribution of T cell stimulatory (Intact) toxic epitopes presents in hexaploid bread wheat and their diploid genome progenitors. Genome-wise distribution in hexaploid sequences was identified based on the three groups developed in the phylogenetic tree. The details are given in the section on phylogenetic analysis.

<table>
<thead>
<tr>
<th>Species</th>
<th>Genotype</th>
<th>Glia-α</th>
<th>Glia-α2</th>
<th>Glia-α9</th>
<th>Glia-α20</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diploid Progenitors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. monococcum</td>
<td>A</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>24</td>
<td>30</td>
</tr>
<tr>
<td>A. searsii</td>
<td>B</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>A. tauschii</td>
<td>D</td>
<td>14</td>
<td>12</td>
<td>16</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>Hexaploid Wheat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. aestivum</td>
<td>A</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>T. aestivum</td>
<td>B</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>26</td>
</tr>
<tr>
<td>T. aestivum</td>
<td>D</td>
<td>60</td>
<td>60</td>
<td>58</td>
<td>42</td>
<td>75</td>
</tr>
</tbody>
</table>

* N is the total Number of sequences used in the analysis.
Singh et al.

Figure 1. Phylogenetic analysis of amino acid sequences of α-gliadins representing 177 accessions of diploid progenitors and hexaploids. The phylogenetic tree differentiated α-gliadin sequences into three distinct groups representing A (black), B (green) and D (red) genomes.

analyzed to identify genome specific amino acid motifs (Table 3). Two short sequences (motifs) in NR1 and NR2 regions of α-gliadins differentiated most the amino acid sequences representing A, B and D genomes. One of the motifs (22 amino acid sequence) in NR1 domain positioned from 240 to 261 (Figure S1) distinguished all the 30 sequences representing A genome of diploids and 9 of the 11 sequences representing A genome in hexaploids. The motif also clearly differentiated 13 of 16 sequences of diploid (D genome) progenitors and 70 of 75 sequences representing D genome in hexaploids. All the α-gliadin sequences representing A genome of diploid (30) and hexaploid (11) could also be distinguished by 13 amino acid motif (PLGQGSFRPSQQN) in NR2 region. The 13 amino acid motif also distinguished 62 of 75 hexaploid sequences of D genome and 13 of 16 sequences of diploids.
Figure 2. Schematic diagram of motif distribution of α-gliadins. Numbers written below each node are bootstrap values derived from 1,000 replicates. Twenty conserved novel motifs are shown with different colored boxes.
Sequence Polymorphism and Nucleotide Diversity of α-Gliadin Genes

Nucleotide diversity analysis of α-gliadin genes was conducted by Tajima D test (Tajima, 1989) using MEGA 6.0 software (Table 2). Large variations were observed among sequences of α-gliadin genes representing all the three genomes. Among diploid ancestors, B genome progenitors exhibited highest numbers of nucleotide polymorphic sites (ps= 0.31) followed by A genome (ps= 0.16) and D genome (ps= 0.13) progenitors. However, among hexaploids, B and D genome sequences exhibited highest number of nucleotide polymorphic sites (ps= 0.34 and 0.31) followed A genomes (ps= 0.19). Highest Nucleotide diversity (P= 0.08) was observed in B genome donors (Aegilopspeltoides and Aegilopssearsii) and B genome of hexaploids and lowest in D genomes of both diploids (0.02) and hexaploids (0.04). The higher nucleotide diversity in B genome was correlated with the reduced number of intact CD epitopes in α-gliadins and lower nucleotide diversity in D genomes with higher number of intact CD epitopes. D statistics further revealed significant deviations from neutrality distribution of A and D genome sequences of both diploids and hexaploids (P< 0.05). D genome showed highest D test value (-1.51) followed by A (-0.88) and B (-0.58) genomes of diploids. D test value in hexaploid sequences were -1.48 in D, -1.29 in A and 0.07 in B genomes (Table 2).

Table 2. Tajima D test statistics to identify nucleotide diversity of α-gliadin genes in A, B and D genomes of hexaploids wheat and their diploid progenitors.

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome</th>
<th>m</th>
<th>n</th>
<th>S</th>
<th>ps</th>
<th>T</th>
<th>p</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diploid progenitors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. monococcum</td>
<td>A</td>
<td>30</td>
<td>248</td>
<td>40</td>
<td>0.16</td>
<td>0.041</td>
<td>0.03</td>
<td>-0.88</td>
</tr>
<tr>
<td>A. speltoides + A. searsii</td>
<td>B</td>
<td>19</td>
<td>167</td>
<td>52</td>
<td>0.31</td>
<td>0.089</td>
<td>0.08</td>
<td>-0.58</td>
</tr>
<tr>
<td>A. touschii</td>
<td>D</td>
<td>16</td>
<td>265</td>
<td>34</td>
<td>0.13</td>
<td>0.038</td>
<td>0.02</td>
<td>-1.51</td>
</tr>
<tr>
<td>Hexaploid wheat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T. aestivum</td>
<td>A</td>
<td>11</td>
<td>274</td>
<td>52</td>
<td>0.19</td>
<td>0.065</td>
<td>0.05</td>
<td>-1.29</td>
</tr>
<tr>
<td>T. aestivum</td>
<td>B</td>
<td>26</td>
<td>242</td>
<td>75</td>
<td>0.31</td>
<td>0.081</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>T. aestivum</td>
<td>D</td>
<td>75</td>
<td>167</td>
<td>56</td>
<td>0.34</td>
<td>0.069</td>
<td>0.04</td>
<td>-1.48</td>
</tr>
</tbody>
</table>

a m= Number of sequences, n= Total Number of sites, S= Number of Segregating sites, ps= S/n, T= ps/a1, p= Nucleotide diversity and D is the Tajima test statistic.

Table 3. Amino acid motifs in NR1 and NR2 regions of α-gliadins differentiating them into three distinct groups representing A, B and D genomes.

<table>
<thead>
<tr>
<th>Motif</th>
<th>Amino acid motif sequence</th>
<th>Number of α-gliadin sequences differentiated by 22 and 13 amino acid motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>22 Amino acid motif</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A Genome</td>
<td>STYQLLQELCCQHLWQIPEQSQ</td>
<td>30/30 (diploid) and 9/11 (hexaploids)</td>
</tr>
<tr>
<td>B Genome</td>
<td>SYYOLQQLCCQOLIPEQSR</td>
<td>10/19 (diploid) and 12/26 (hexaploids)</td>
</tr>
<tr>
<td>D Genome</td>
<td>STQLYQQLCCQQLWQIPEQSR</td>
<td>13/16 (diploid) and 70/75(hexaploids)</td>
</tr>
<tr>
<td>13 Amino acid motif</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A Genome</td>
<td>PLAGQGSRPSQSN</td>
<td>30/30 (diploid) and 11/11(hexaploids)</td>
</tr>
<tr>
<td>B Genome</td>
<td>PSQGQSFQPSQSN</td>
<td>16/19 (diploid) and 13/26 (hexaploids)</td>
</tr>
<tr>
<td>D Genome</td>
<td>PSSQGQSFQPSQSN</td>
<td>13/16 (diploid) and 62/75 (hexaploids)</td>
</tr>
</tbody>
</table>
DISCUSSION

Identification and Frequency Distribution of CD Toxic Epitopes in α-Gliadins

There were large variations in the number of intact (T cell stimulatory) epitopes of α-gliadins representing A, B, and D genomes of bread wheat and its progenitors (Table 1). Intact Glia-α epitope was absent in α-gliadins encoded by A genome of both the diploid progenitors and hexaploid sequences. Intact Glia-a2 was absent in all the α-gliadins encoded by A and B genomes of diploids. However, both the Glia-andGlia-a2 sequences were present in higher frequencies in D genome progenitors and encoded by D genomes of hexaploids. Intact Glia-a9 and Glia-a20 were present in high proportions in A and D while absent or rarely present in B genome encoded α-gliadins. Although intact Glia-a2 and Glia-a9 epitopes were absent in B genome diploid progenitors, 7 accessions in hexaploids exhibited both the epitopes in intact form in α-gliadins encoded by B genomes. Earlier reports indicated the absence of both Glia-a2 and Glia-a9 epitopes in all the diploid and hexaploid sequences (Van Herpen et al., 2006). Overall, α-gliadinin D genome progenitors (Aegilops tauschii) and encoded by D genome of bread wheat exhibited all four intact CD epitopes. Several other reports also showed the presence of all epitopes in intact form in α-gliadins encoded by D genome (Van Herpen et al., 2006; Li et al., 2010; Xie et al., 2010; Salentijn et al., 2013; Li et al., 2014a; Salentijn et al., 2009; Van den Broeck et al., 2010). Surprisingly, 5 accessions of hexaploids exhibited variant form of all the epitopes (none having intact epitopes) in α-gliadin encoded by D genomes. The accessions are TA[513129818][gb|AGO17661.1]; TA[47883564][gb|ABQ52125.1]; TA[421932488][gb|AFX69622.1]; TA[376341626][gb|AFB35199.1]; and TA[513129824][gb|AGO17663.1]. Since D genome in hexaploids is the major source of intact CD toxic epitopes, these accessions can be very useful source of improving wheat for reducing toxicity load among CD patients.

Phylogenetic Analysis and Assignment of α-Gliadin Sequences onto A, B and D Genomes

Both the diploid and hexaploid sequences of α-gliadins were phylogenetically clustered into 3 distinct groups representing A, B, and D genomes. Similarly, motif analysis showed three distinct groups of sequences represented by A, B, and D genomes. The characteristic feature of α-gliadins is the presence of two polyglutamine repetitive domains (QI & QII) and the size of which influences the visco-elastic properties of doughs. This is because of intermolecular interactions of the large numbers of glutamine side chains, which act as good hydrogen bond donors and acceptors (Masci et al., 2000). Earlier reports indicated significant differences in the length of polyglutamine (QI and QII) repeats among all the three genomes, which could be used to distinguish genome specific sequences of α-gliadins (Van Herpen et al., 2006; Xie et al., 2010). However, in the present investigation, QI and QII repeat lengths could not distinguish α-gliadins representing different genomes. QII repeat length of α-gliadins showed significant differences between α-gliadins encoded by A and B between B and D genomes while no difference was observed between A and D genomes in diploids (Figure 3). However, in hexaploids, QII repeat length in α-gliadins was significantly different between A and B and between A and D genomes, however, no differences were observed between B and D genomes. As compared to diploids, QII repeat length of B genome was lower in hexaploids indicating loss of Q residues during early phase of evolution of hexaploid wheat (Figure 3).
Since QI and QII length could not differentiate α-gliadin sequences encoded by different genomes, all the sequences were further analyzed to identify amino acid motifs distinguishing α-gliadin sequences representing A, B, and D genomes. Specific 22 and 13 amino acid motifs were found present in NR1 and NR2 domains, respectively. These motifs could differentiate α-gliadin sequences representing A and D genomes. The 22 amino acid motif distinguished all the 30 sequences representing A genome and 13 out of 16 sequences of D genomes of diploid progenitors. The motif also distinguished 70 of the 75 sequences representing D genomes and 9 out of 11 sequences of D genomes of diploid progenitors. The motif also distinguished 70 of the 75 sequences representing D genomes and 9 out of 11 sequences representing A genome of hexaploids. Surprisingly, all the 70 sequences of D genome had valine at 245 positions and all the 9 sequences of A genome had glutamic acid (E), Histidine (H) and glutamine (Q) at 247, 252 and 261 positions, respectively. Similarly, 13 amino acid motif in NR2 region at 332-344 positions also differentiated all the α-gliadin sequences representing A genome of both diploids and hexaploids. Overall, both 22 and 13 amino acid motifs could distinguish α-gliadin sequences representing A and D genomes and the remaining sequences could be grouped into B genome. In hexaploid wheat, the donor of the B genome has been the most controversial and is still relatively unknown, in spite of a large number of attempts to identify the parental species (Huang et al., 2002). This may be associated with the higher diversification rate of the B genome compared to the A(u) genome in the polyploid wheat (Petersen et al., 2006), the incomplete chromosome pairing between B genome chromosomes and any diploid species (Blake et al., 1999) and the fact that the B genome is relatively diverged from its putative diploid progenitors (Talbert et al., 1995).

**Nucleotide Diversity of α-Gliadin Genes**

Tajima D test statistics was conducted to identify nucleotide diversity of α-gliadin genes. B genome donors (*Aegilops speltoides* and *Aegilops searsii*) and B genome of hexaploids exhibited higher nucleotide diversity and reduced number of intact CD epitopes in α-gliadins as compared to A and D genome sequences. The statistics revealed significant deviations from neutrality distribution of A and D genome sequences of both diploids and hexaploids (P < 0.05). Highest D test value (negative) was exhibited by D genome sequences. The expectation for a neutral locus in a population is zero value of Tajima’s D.
ACKNOWLEDGEMENTS

The financial assistance received from DBT, Government of India with the grant no. BT/PR5675/AGIII/103/863/2012 is gratefully acknowledged. The authors are thankful to the Director, Indian Institute of Wheat & Barley Research Institute (IIWBR), Karnal, for providing basic facilities. We also thank NCBI and various authors for making available sequence data of α-gliadin gene in various genomic backgrounds of bread wheat in the public domain.

REFERENCES


Identification of Genome Specific Sequence Motifs


شناسایی توالی موتیف های ویژه زنوم در α-Gliadins و نمونه های بیماری سلیبک

گنگم با ایپی توب های بیماری سلیبک

س. سینگ، س. رام، و س. ناروال

چکیده

در هیپبی Gliadinّب، α- Gliadins بِ خبطز حضَر اپی تَپ ّبی سوی، پزٍتئیيّبی فعبل ٍهْوی در شزٍع بیوبری سلیبک در افزاد بشز ّستٌذ. در ایي پضٍّش، هجوَعِ ای شبهل 755 تَالی صًی α- Gliadin پزٍتئیيّبی هتٌبظز آىّب تجشیِ شذ. در ًتیجِ، 02ًوًَِ ثبت شذُ ّگشاپلَیذی شبهل 7، 72، ٍ 3، بِ تزتیب ًوبیٌذُ صًَم A، B، ٍ D بِزٍى CD-epitopes بی سبلن ٍ دست ًخَردُ در α- Gliadin شٌبسبیی شذ. سپس، 00ٍ 71(هَتیف حفبظت شذُ) بِ تزتیب در شناسایی بود. مشاهدات حاکی از تغییرات زیاد و تعداد کمتر repeat length نخورده در α-Gliadin مربوط به زنوم B بود. در مقایسه با دیپلود ها، طُل تکرار (repeat length) شناسایی بود. مشاهدات حاکی از تغییرات زیاد و تعداد کمتر α-Gliadin دامنه های غیر تکراری از β1-4NR2 و α-Gliadin در D از دیپلود ها و هگزاپلیود ها را تمايز بِای کردند. بیشتر توالی آمینو اسیدهای رمز گذاری شده با زنوم D (70 تا از 75 هگزا پلیپ و 33 تا از 16 دیپلود) با 22 موتیف آمیتواسید قابل conserved motifs حفاظت شده ( به ترتیب در به هیپبی CD-epitopes بب صًَم) بِ حفاظت بیوبری سلیبک در افزاد بشز ّذف هذیزیت کزدى بیوبری سلیبک در افزاد بشز هفیذ ببشذ.