Theor Appl Genet (1987) 75:117–126

The nucleotide and deduced amino acid sequences of an HMW glutenin subunit gene from chromosome 1B of bread wheat (*Triticum aestivum* L.) and comparison with those of genes from chromosomes 1A and 1D

N. G. Halford [1], J. Forde [1], O. D. Anderson [2], F. C. Greene [2] and P. R. Shewry [1]

[1] Biochemistry Department, Rothamsted Experimental Station, Harpenden, Herts, UK
[2] USDA Western Regional Research Center, 800 Buchanan St., Albany, CA 94710, USA

Received November 20, 1986; Accepted May 27, 1987
Communicated by R. Riley

Summary. The nucleotide and deduced amino acid sequences of a high molecular weight glutenin subunit gene derived from chromosome 1B of bread wheat (*Triticum aestivum* L.) are reported. The encoded protein corresponds to the y-type subunit 1B9. Comparison of the 5′ upstream untranslated regions of this gene and a previously reported silent y-type gene derived from chromosome 1A showed a deletion of 85 bp in the latter. A sequence present in this region of the 1By 9 gene shows homology with part of the "–300 element" which is conserved in the 5′ upstream regions of other prolamin genes from barley, wheat and maize (Forde BG et al. 1985). It is suggested that the absence of this element is responsible for the lack of expression of the 1Ay gene. Comparison of the derived amino acid sequence with those reported previously for the silent 1Ay gene and the expressed x-type (1Dx2) and y-type (1Dy12) genes derived from chromosome 1D showed that the three y-type proteins are closely related. In contrast the x-type subunit (1Dx2) shows clear differences in the N-terminal region and in the number, type and organisation of repeats in the central repetitive domain.

Key words: Glutenin – Wheat – Amino acid sequences – Chromosomes 1A, 1B and 1D

Introduction

The high molecular weight (HMW) subunits of wheat gluten only account for about 10% of the total gluten proteins, but are currently of considerable interest to wheat breeders, chemists and molecular biologists. This is because of the close association of specific allelic forms with good or poor breadmaking quality (Payne et al. 1984). Genetic studies show that the subunits are specified by single loci on the long arms of the group 1 chromosomes of all three genomes (called the A, B and D genomes) of hexaploid bread wheat (Lawrence and Shepherd 1980, 1981; Payne et al. 1980, 1982). Molecular analyses have shown that two genes are present at each locus (Harberd et al. 1986). However, not all of these genes are expressed. Cultivars of bread wheat have 3 to 5 HMW subunits, which show allelic variation in their electrophoretic mobility (Payne et al. 1981). Two subunits encoded by the D genome are almost always present and these are classified into high M_r 1Dx and low M_r 1Dy subunit types (Payne et al. 1981). A 1Bx subunit is also always present, and this is accompanied in some cultivars by a 1By and/or a 1Ax subunit. 1Ay subunits are never apparently present in bread wheats, but may be present in wild diploid species related to the progenitor of the A genome (Payne et al. 1983).

Although a number of individual HMW subunits have been purified and characterized, the determination of complete amino acid sequences is precluded by their high M_rs (90–140,000 by SDS-PAGE), the presence of repeated sequences and difficulty in preparing sufficiently large amounts of individual proteins. A much easier approach is to determine these sequences indirectly from genomic or cDNA, and the complete nucleotide sequences of three cloned genes have recently been reported (Forde J et al. 1985; Thompson et al. 1985; Sugiyama et al. 1985). Two of these correspond to the 1D subunits, 1Dx2 and 1Dy12 which are present together in a number of cultivars (Payne and Lawrence 1983). The third gene was derived from chromosome 1A and was found to be a silent gene of the 1Ay type. Although the protein sequence of this

gene contained an in-frame ochre stop codon, this is not necessarily related to the lack of expression.

In the present paper we report the first nucleotide and deduced protein sequences of an HMW subunit gene derived from chromosome 1B of bread wheat. Comparison of the 5′ upstream region with that of the silent 1A gene shows an 85 bp deletion in the latter, which may account for its lack of expression. A detailed comparison of the structures of the 1A, 1B, 1Dx and 1Dy subunits also reveals differences in the structural organisation and precise motifs of the repetitive domains and these are discussed in relation to the evolution of the proteins.

Materials and methods

Part of an amplified genomic library of *Triticum aestivum* L. (cv Cheyenne) (Anderson et al. unpublished) was plated on *Escherichia coli* strain 1046 and screened by plaque hybridization using the SmaI/Bam H1 fragment from HMW subunit clone pHSB26 (Forde J et al. 1985) as a probe. Hybridizing plaques were picked, purified and the λ clone DNA's prepared from liquid culture (Maniatis et al. 1982). The clones were analysed by separating EcoRI restriction digests on agarose gels and hybridization to pHSB26. Clone λHMW47 had a 15.8 kb EcoRI fragment, which has been assigned to gene(s) on chromosome 1B (Forde J et al. 1985; Harberd et al. 1986).

A partial restriction map of the insert of λHMW47 was determined and the HMW subunit gene located on three HindIII fragments by hybridization with pHSB26. The three fragments (of about 1,800, 900 and 900 bp) were then subcloned into M13 mp19. Deletions of these fragments were generated from caesium chloride purified DNA in the replicative (i.e. double stranded) form by digestion first with the restriction endonucleases SmaI and KpnI and then with Exonuclease III for periods up to 10 min. The fragments were then digested with S1 nuclease to produce blunt ends and religated. Clones of desired size were selected by running on 1% agarose gels.

To sequence across the HindIII site at position 1,178, a 2.3 kb Bam H1 fragment was cloned into M13mp19. This fragment runs from the Bam H1 site at position 1043 to a site downstream of the gene. Sequence obtained from this clone ran across the HindIII site. To sequence across the HindIII site at postion 305 the 15.8 kb EcoRI fragment from λHMW47 was purified, denatured and annealed to an oligonucleotide primer complementary to the sequence adjoining the HindIII site. This was then used as a sequencing template. All nucleotide sequences were determined using the dideoxy method.

Results and discussion

Identification, isolation and nucleotide sequence of an HMW subunit gene encoded by chromosome 1B

Hybridization of a Southern Blot of *EcoRI*-digested DNA of wheat (cv Cheyenne) with a labelled cDNA insert related to an HMW subunit showed six hybridizing fragments (Forde T et al. 1985). The chromosomal origins of these fragments were determined by comparison of DNA fractions from euploid Cheyenne and aneuploid lines which were nullisomic for chromosome 1A or monosomic for chromosome 1A, 1B or 1D. This showed that two fragments each were derived from chromosomes 1A (6.8 and 8.2 kb), 1B (15.8 and 21.6 kb) and 1D (6.4 and 8.7 kb). These assigments have since been confirmed by Harberd et al. (1986) by analysis of substitution lines of group 1 chromosomes of Cheyenne into Chinese Spring.

The 15.8 kb fragment derived from chromosome 1B was isolated from a genomic library of Cheyenne DNA. The HMW subunit gene was located on three HindIII restriction fragments, which were subcloned into M13 mp19. The restriction map of this region and the deletions constructed to sequence the three subclones are shown in Fig. 1. The sequence of 2,996 base pairs contains an open reading frame of 2,115 base pairs encoding a protein of 705 residues (Fig. 2). No introns are present. A TATA box is present beginning at position 660 (85–91 bp upstream of the ATG start codon) and three CCAAT-like sequences beginning at positions 632 (CCAT). 555 (CAAAT) and 516 (CCAAT) (Messing et al. 1983).

Beginning at position 294 (Fig. 2) is a 24 bp sequence (GACATGCTTAGAAGCTTTTAGTGA) which shows considerable homology with part of the –300 element reported by Forde BG et al. (1985) to be conserved in the 5′ flanking regions of B hordeins and other prolamin genes. In addition there are three imperfect direct repeats beginning at positions 122, 147, and 168. These have the consensus sequence GTGAGTCATAGCATA and also have some homology with the –300 element.

In the 3′ flanking region, a single polyadenlylation signal (AATAAA) (Messing et al. 1983) is present at 51 bp downstream from the second (TAG) stop codon, and a second sequence (AATAAT) (Dean et al. 1986) is found a further 50 bp downstream (Fig. 2).

Comparison of the 5′ upstream region with those of other HMW subunit genes

The nucleotide and deduced amino acid sequences of three HMW subunit genes have been reported previously. These are genes for the 1Dx subunit 2. 1Dy subunit 12 and a silent (i.e. unexpressed) 1Ay subunit. Comparison of the 5′ flanking regions of these genes has shown no apparent reason for the silent nature of the 1Ay gene, although the 1Dx2 and 1Dy12 genes have been sequenced for less than 450 bp upstream compared to 800 bp for the silent 1Ay gene. We have sequenced almost 750 bp upstream of the coding sequence of the 1B gene, enabling a more detailed comparison to be made.

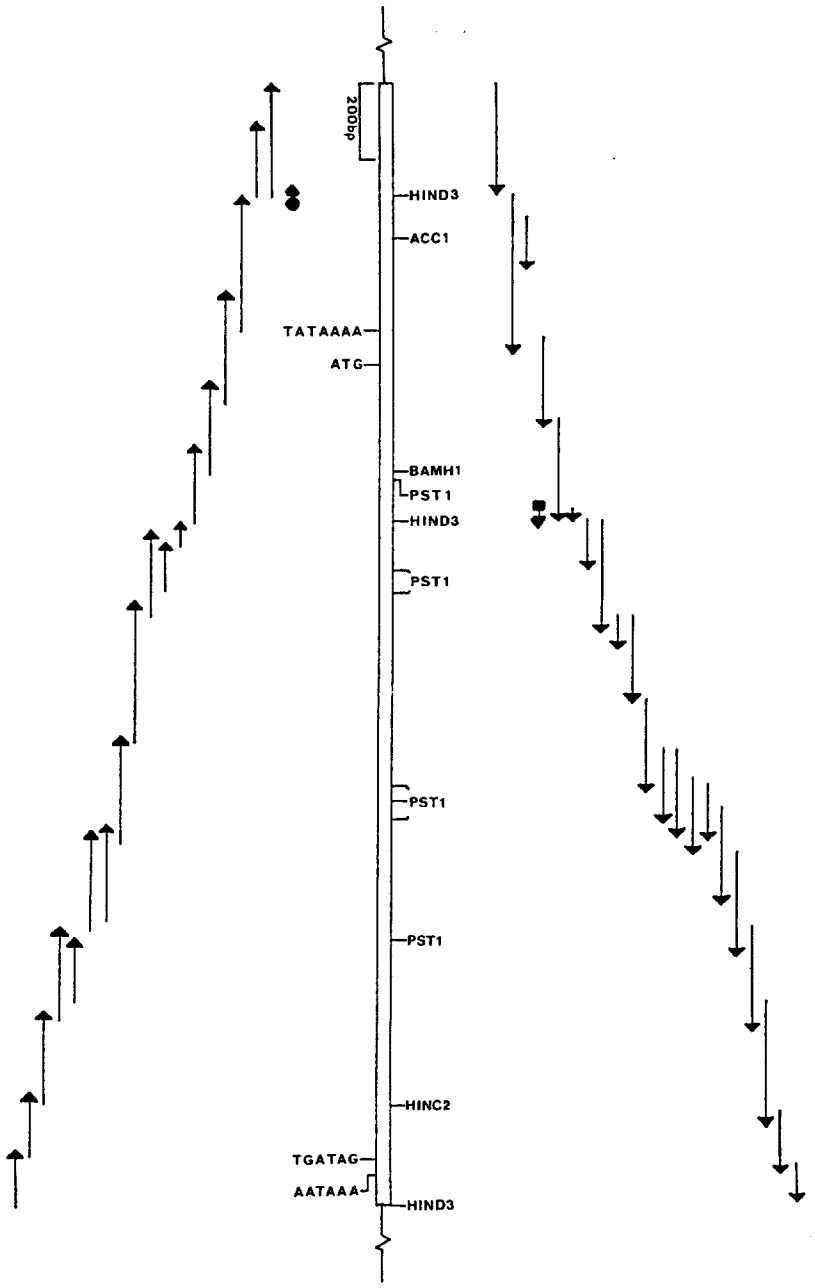


Fig. 1. Restriction map of the sequenced portion of λ HMW 47. The putative TATA box, translation start, stop condons and polyadenylation signals are indicated. *Arrows* show deletions of the HindIII fragments used in sequencing; *arrow with circle* indicates the sequence across the HindIII site at position 305 determined using an oligonucleotide primer; *arrow with square* indicates the sequence across the HindIII site at position 1,178 obtained from the 2.3 kb Bam H1 fragment

Comparison of the 5′ upstream regions of the four genes shows a high degree of homology, with TATA boxes and CCAAT-like sequences present in similar positions in each (not shown). However, comparison of the additional upstream sequences of the 1Ay and 1B genes provides further information (Fig. 3). Firstly, the three direct repeats which are homologous with part of the "–300 element" are present in similar positions in both (Fig. 3). Secondly, 85 bp of the 1B sequence (positions 289 to 373 in Fig. 2, which correspond to positions –462 to –378 upstream of the ATG codon) are absent from the 1Ay gene. The known upstream sequences of the cv Yamhill 1Dx2 and cv Cheyenne 1Dx5 (Yip and Malpica, unpublished data) genes extend for 68 bp into regions that are homologous to the 85 bp sequence apparently deleted from the silent 1Ay gene and the published sequence of the 1Dy12 gene extends 57 bp into a similar region. We therefore suggest that this deletion in the 5′ flanking region of the 1Ay gene is related to its lack of expression. Detailed analysis of this 85 bp region of the 1B gene shows the presence of a 24 bp sequence that is homologous with

```
1    GAGTTTTCTCTGAAGATCCTATGTTAATTTTAGACATGAATGACCAAAGGTTTCAGTTAGTTGAGTTTTGTCATCGAAAGGTGTTTACATAAGTCCAAAA 100

101  ATTCTACCAGCTTTTGGTACGGTGCGTCATAGCACAGATAAATGTTGTGATTCATTAGATAGATATTGTGAGTCATAGCATGGATTTGTGTTGCCTGGAA 200

201  ATCCAACTACATGACAAGAAACAAACATAAATGGGCTTTTGAAGGATGATTTATCAACTTACCTTATCCATGCAAGCTACCTTCCATTAGTCGGACATGC 300

301  TTAGAAGCTTTTAGTGACCGCAGATCTACAAAAGCAATGGCTAACAGACACATATTCTCCCAAACCCCAAGAAGTATAACCACTTCTCTTAGATAAAAAT 400

401  AGCAGATCGATATACAAACGTCTACACTTCTGCAAACAGTACCCAGAAGTCAGAATTAGGATTGAACCGATTACGTGGCTTTAGCAGACTGTCCAAAAAT 500

501  CTGTTTTGCAAAACTCCAATTGCTCCTTGCTTATCCAGCTTCTTTTGTGCTTGGCAAATTGCTGCTTTTCCAACTGACTTTATTCTTCTCACGTTTCTTC 600

601  TTAGGCTAAACTAACCACACCGTGCACACAACCATTGTCCCGAACCTTCACCACGTCCCTATAAAAGCCCAACCAATCTCCACAATCTCATCATCACCCA 700

                                                                     M  A  K  R  L  V  L  F  A  T  V  V  I  T  L  V  A
701  CAACACCGAGCACCACAAAATAGAGATCAATTCACTAACAGTCCATCGAGATGGCTAAGCGGTTGGTCCTCTTTGCGACAGTAGTCATCACCCTCGTGGC 800
        signal
      L  T  A  A |E  G  E  A  S  R  Q  L  Q  C  E  R  E  L  Q  E  S  S  L  E  A  C  R  Q  V  V  D  Q  Q
801  TCTCACTGCTGCTGAAGGTGAGGCCTCTAGGCAACTACAGTGTGAGCGCGAGCTCCAGGAGAGCTCGCTTGAGGCATGCCGACAGGTCGTGGACCAACAG 900

      L  A  G  R  L  P  W  S  T  G  L  Q  M  R  C  C  Q  Q  L  R  D  V  S  A  K  C  R  P  V  A  V  S  Q  V
901  TTGGCCGGTCGGCTGCCATGGAGCACGGGGCTCCAGATGCGATGCTGCCAGCAGCTCCGAGATGTTAGCGCTAAGTGCCGTCCCGTCGCCGTCAGCCAAG 1000

      V  R  Q  Y  E  Q  T  V  V  P  P  K  G  G  S  F  Y  P  G  E  T  T  P  L  Q  Q  L  Q  Q  V  I  F  W
1001 TCGTAAGACAATATGAGCAAACCGTGGTGCCGCCCAAGGGCGGATCCTTCTACCCTGGCGAGACCACACCACTGCAGCAACTCCAACAAGTAATATTTTG 1100

      G  T  S  S  Q  T  V  Q  G  Y  Y  P  S  V  S  S  P  Q  Q  G  P  Y  Y  P  G  Q  A  S  P  Q  Q  P  G
1101 GGGAACATCTTCACAAACAGTACAAGGGTATTACCCAAGCGTAAGTTCTCCTCAGCAGGGGCCATATTATCCAGGCCAAGCTTCTCCACAACAGCCAGGA 1200

      Q  G  Q  Q  P  G  K  W  Q  E  L  G  Q  G  Q  Q  G  Y  Y  P  T  S  L  H  Q  S  G  Q  G  Q  Q  G  Y  Y
1201 CAAGGGCAACAGCCAGGCAAATGGCAAGAACTGGGACAAGGGCAACAAGGGTACTACCCAACTTCTCTGCATCAGTCAGGACAAGGACAACAAGGGTACT 1300

      P  S  S  L  Q  Q  P  G  Q  G  Q  Q  I  G  Q  G  Q  Q  G  Y  Y  P  T  S  L  Q  Q  P  G  Q  G  Q  Q
1301 ACCCATCTTCTCTGCAGCAACCAGGACAAGGGCAACAGATAGGACAAGGGCAACAAGGATACTACCCAACTTCTCTGCAGCAGCCAGGACAAGGGCAACA 1400

      I  G  Q  G  Q  Q  G  Y  Y  P  T  S  P  Q  H  P  G  Q  R  Q  Q  P  G  Q  G  Q  Q  I  G  Q  G  Q  Q
1401 GATAGGACAAGGACAACAAGGGTACTACCCAACTTCTCCGCAACACCCAGGACAAAGGCAACAACCAGGACAAGGGCAGCAAATAGGACAAGGGCAACAA 1500

      L  G  Q  G  R  Q  I  G  Q  G  Q  Q  S  G  Q  G  Q  Q  G  Y  Y  P  T  S  P  Q  Q  L  G  Q  G  Q  Q  P
1501 CTAGGACAAGGGCGGCAAATAGGACAAGGGCAACAATCAGGACAAGGGCAACAAGGGTACTATCCAACTTCTCCACAGCAGCTAGGACAAGGGCAACAAC 1600

      G  Q  W  Q  Q  S  G  Q  G  Q  Q  G  Y  Y  P  T  S  Q  Q  Q  P  G  Q  G  Q  Q  G  Q  Y  P  A  S  Q
1601 CAGGACAATGGCAACAATCAGGACAAGGGCAACAAGGGTACTACCCAACTTCTCAGCAGCAGCCAGGACAAGGGCAACAAGGGCAGTACCCAGCTTCTCA 1700

      Q  Q  P  G  Q  G  Q  Q  G  Q  Y  P  A  S  Q  Q  Q  P  G  Q  G  Q  Q  G  Q  Y  P  A  S  Q  Q  Q  P
1701 GCAGCAGCCAGGACAAGGGCAACAAGGGCAGTACCCAGCTTCTCAGCAGCAGCCAGGACAAGGGCAACAAGGGCAGTACCCAGCTTCTCAACAACAGCCA 1800

      G  Q  G  Q  Q  G  H  Y  L  A  S  Q  Q  Q  P  G  Q  G  Q  Q  R  H  Y  P  A  S  L  Q  Q  P  G  Q  G  Q
1801 GGACAAGGGCAACAAGGGCACTACCTAGCTTCTCAGCAGCAGCCAGGACAAGGGCAACAACGGCACTACCCAGCTTCTCTGCAGCAACCAGGACAAGGGC 1900

      Q  G  H  Y  T  A  S  L  Q  Q  P  G  Q  G  Q  Q  G  H  Y  P  A  S  L  Q  Q  V  G  Q  G  Q  Q  I  G
1901 AACAAGGGCATTACACAGCTTCTCTGCAGCAACCAGGACAAGGGCAACAAGGGCATTACCCAGCTTCTCTGCAGCAGGTAGGACAAGGACAACAAATAGG 2000

      Q  L  G  Q  R  Q  Q  P  G  Q  G  Q  Q  T  R  Q  G  Q  Q  L  E  Q  G  Q  Q  P  G  Q  G  Q  Q  T  R
2001 ACAGCTAGGACAAAGGCAACAACCAGGACAAGGGCAACAAACAAGACAAGGGCAACAACTAGAACAAGGGCAACAACCAGGACAAGGGCAACAAACAAGA 2100

      Q  G  Q  Q  L  E  Q  G  Q  Q  P  G  Q  G  Q  Q  G  Y  Y  P  T  S  P  Q  Q  S  G  Q  G  Q  Q  P  G  Q
2101 CAAGGGCAACAACTAGAACAAGGGCAACAACCAGGACAAGGGCAACAAGGGTACTATCCAACTTCTCCACAACAGTCGGGACAAGGGCAACAACCAGGAC 2200

      S  Q  Q  P  G  Q  G  Q  Q  G  Y  Y  S  S  S  L  Q  Q  P  G  Q  G  L  Q  G  H  Y  P  A  S  L  Q  Q
2201 AATCACAACAACCAGGACAAGGGCAACAAGGGTACTACTCAAGTTCTCTACAACAGCCAGGACAAGGGCTACAAGGGCACTACCCAGCTTCTCTGCAGCA 2300

      P  G  Q  G  H  P  G  Q  R  Q  Q  P  G  Q  G  Q  Q  P  E  Q  G  Q  Q  P  G  Q  G  Q  Q  G  Y  Y  P
2301 GCCAGGACAAGGACATCCAGGACAAAGGCAACAACCAGGACAAGGGCAACAACCAGAACAAGGGCAACAACCAGGACAGGGGCAACAAGGGTATTATCCA 2400

      T  S  P  Q  Q  P  G  Q  G  K  Q  L  G  Q  G  Q  Q  G  Y  Y  P  T  S  P  Q  Q  P  G  Q  G  Q  Q  P  G
2401 ACTTCTCCGCAGCAGCCAGGACAAGGGAAACAACTAGGACAAGGGCAACAAGGGTACTACCCAACTTCTCCGCAACAGCCAGGACAAGGGCAACAACCAG 2500

      Q  G  Q  Q  G  H  C  P  T  S  P  Q  Q  T  G  Q  A  Q  Q  P  G  Q  G  Q  Q  I  G  Q  V  Q  Q  P  G
2501 GACAAGGGCAACAAGGGCACTGCCCAACTTCTCCGCAGCAGACAGGACAAGCGCAACAACCAGGACAAGGCCAACAAATAGGACAAGTGCAACAACCAGG 2600

      Q  G  Q  Q  G  Y  Y  P  I  S  L  Q  Q  S  G  Q  G  Q  Q  S  G  Q  G  Q  Q  S  G  Q  G  H  Q  L  G
2601 ACAAGGGCAACAAGGGTACTACCCAATTTCTCTGCAACAGTCAGGACAAGGGCAACAGTCAGGACAAGGGCAACAATCAGGACAAGGACACCAACTAGGA 2700

      Q  G  Q  Q  S  G  Q  E  Q  Q  G  Y  D  N  P  Y  H  V  N  T  E  Q  Q  T  A  S  P  K  V  A  K  V  Q  Q
2701 CAAGGGCAGCAATCAGGACAAGAGCAACAAGGCTACGACAACCCATACCATGTTAACACAGAGCAGCAAACGGCCAGCCCAAAGGTGGCAAAGGTGCAGC 2800

      P  A  T  Q  L  P  I  M  C  R  M  E  G  G  D  A  L  S  A  S  Q  *  *
2801 AACCCGCGACACAGCTGCCGATAATGTGTCGGATGGAGGGGGGCGACGCATTGTCGGCTAGCCAGTGATAGAACTCTCTGCAACTTGCATGGTGCTTGGT 2900

2901 ATGCATGCACCTTAGCTATCCAATAAACATGACGTGTGTGCACAGGTTTTCATGTAACTAGAGTAAAACCCAATAATAATGCAAAATGAAAAGCTT  2996
```

the first part of the "–300 element" sequences present in the 5′-flanking regions of the B hordein of barley (Forde BG et al. 1985) and α- and γ-gliadins of wheat (Anderson et al. 1984; Sumner-Smith et al. 1985; Rafalski 1986) (see Fig. 4). Homology with the α-gliadin sequences extends 14 bp beyond the –300 element. This sequence is longer and shows stronger homology with the –300 element than the three direct repeats present in both the 1B gene and the 1Ay silent gene. It is possible that the "–300 element"-like sequence present in the 85 bp region has a role in the regulation of HMW subunit gene expression, and that the deletion of this element is responsible for the silent nature of the 1Ay subunit gene.

Comparison of the structures of the proteins encoded by the 1Ay, 1Dx, 1Dy and 1B subunit genes

Wheat cv Cheyenne contains two HMW subunits encoded by chromosome 1B, 1Bx7 and 1By9. Subunit 1Bx7 has been purified from three cultivars (including Cheyenne), and the N-terminal amino acid sequence of a fraction from cv Copain determined (Shewry et al. 1984) (Fig. 5). Subunit 1By9 has not been purified and characterized. The deduced N-terminal amino acid sequence of the protein encoded by the 1B subunit gene showed clear differences from that determined directly for subunit 1Bx7, notably the presence of a deletion of three rather than eight residues when compared with the N-terminal sequence of subunit 1Dx2 (Fig. 5). This indicates that the 1B gene encodes the 1By subunit 9, and this conclusion is supported by detailed comparison of the sequence with those of the 1Dy2 and 1Ay subunit genes (see below). The identification of the 1B gene as encoding subunit 1By9 makes it possible for the first time to compare the structures of y-type subunits encoded by all three genomes.

The first 21 residues of the deduced amino acid sequence of the 1By9 subunit constitute a signal peptide (Watson 1984), which is consistent with the synthesis of the HMW subunits on the rough endoplasmic reticulum (Miflin et al. 1983; Forde and Miflin 1983). This is homologous with the signal peptides of the other HMW subunits (three to five amino acid substitutions).

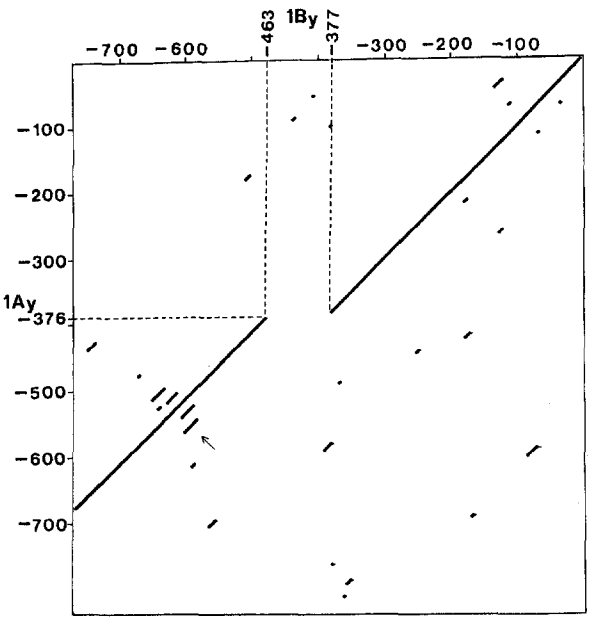


Fig. 3. Homology plot of the 5′ flanking sequences of the 1By9 and silent 1Ay genes using the Wisconsin program DOTPLOT (Window length 21, stringency 14.0). This shows a clear deletion of 85 bp in the 1Ay gene. The three direct repeats indicated by *arrows* correspond to those underlined in **Fig. 2.** The silent 1Ay gene sequence was reported by Forde J et al. (1985)

There is also limited homology with the signal peptides of other cereal prolamins (Kreis et al. 1985).

The deduced amino acid sequence of the mature 1By9 subunit can be divided into three structural domains: a non-repetitive N-terminal domain of 104 residues, a repetitive central domain of 538 residues and a non-repetive domain of 42 residues. The amino acid sequences of these domains are compared with those of the other subunits in Figs. 5 and 6.

The N-terminal amino acid sequence is closely homologous with that of 1Dy12, with only 3 single amino acid substitutions (Fig. 5A). The 1Ay subunit is also closely related, differing in 11 residues. The N-terminal domains of these three subunits all have five cysteine residues which are indicated by arrows in Fig. 5A. The N-terminal domain of the 1Dx subunit differs in the presence of more single amino acid substitutions, the presence of three additional residues between residues 15 and 16 of the y-type subunits and the absence of 18 residues including two cysteines. All four subunits have C-terminal domains of 42 residues, with a single cysteine at position-13. The 1By9 sequence is again more closely related to the 1Ay and 1Dy12 sequences (8 and 7 substitutions respectively) than to the 1Dx2 sequence (11 substitutions).

The repetitive domains of the HMW subunits consist predominantly of two unrelated repeat motifs: a

Fig. 2. Nucleotide and deduced amino sequences of the HMW subunit gene contained in λ HMW 47. The putative TATA box and polyadenylation signals are *boxed.* CCAAT-like sequences are *underlined.* A possible "– 300 element" (Forde BG et al. 1985) is *boxed* with a *broken line* and three direct repeats that also show homology with the "– 300 element" are underlined with *arrows.* The putative signal sequence of the encoded protein is indicated, and nonapeptide (consensus GYYPTSLQQ) and hexapeptide (consensus PGQGQQ) repeat motifs are indicated by *solid* and *broken arrow,* respectively

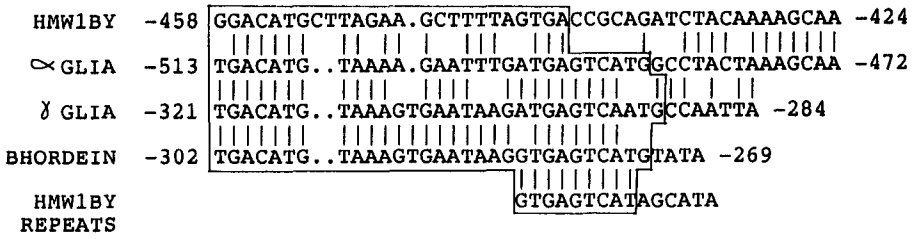


Fig. 4. Comparison of the putative "– 300 element" of the 1By9 gene with those of the α- and γ-gliadin genes of wheat (Sumner-Smith et al. 1985; Rafalski 1986) and the B hordein gene of barley (Forde BG et al. 1985). The consensus sequence of the three direct repeats in the 1By9 subunit gene is also aligned. Initial alignment of the 1By9 subunit and α-gliadin gene sequences was made using the Wisconsin program GAP (gap weight 5.0, length weight 0.30, similarity 65.957 %). The other sequences were aligned manually. The "– 300 element" of Forde BG et al. (1985) is *boxed*

A N-TERMINAL AMINO ACID SEQUENCES

B C-TERMINAL AMINO ACID SEQUENCES

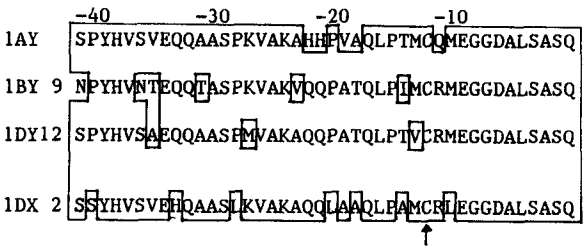
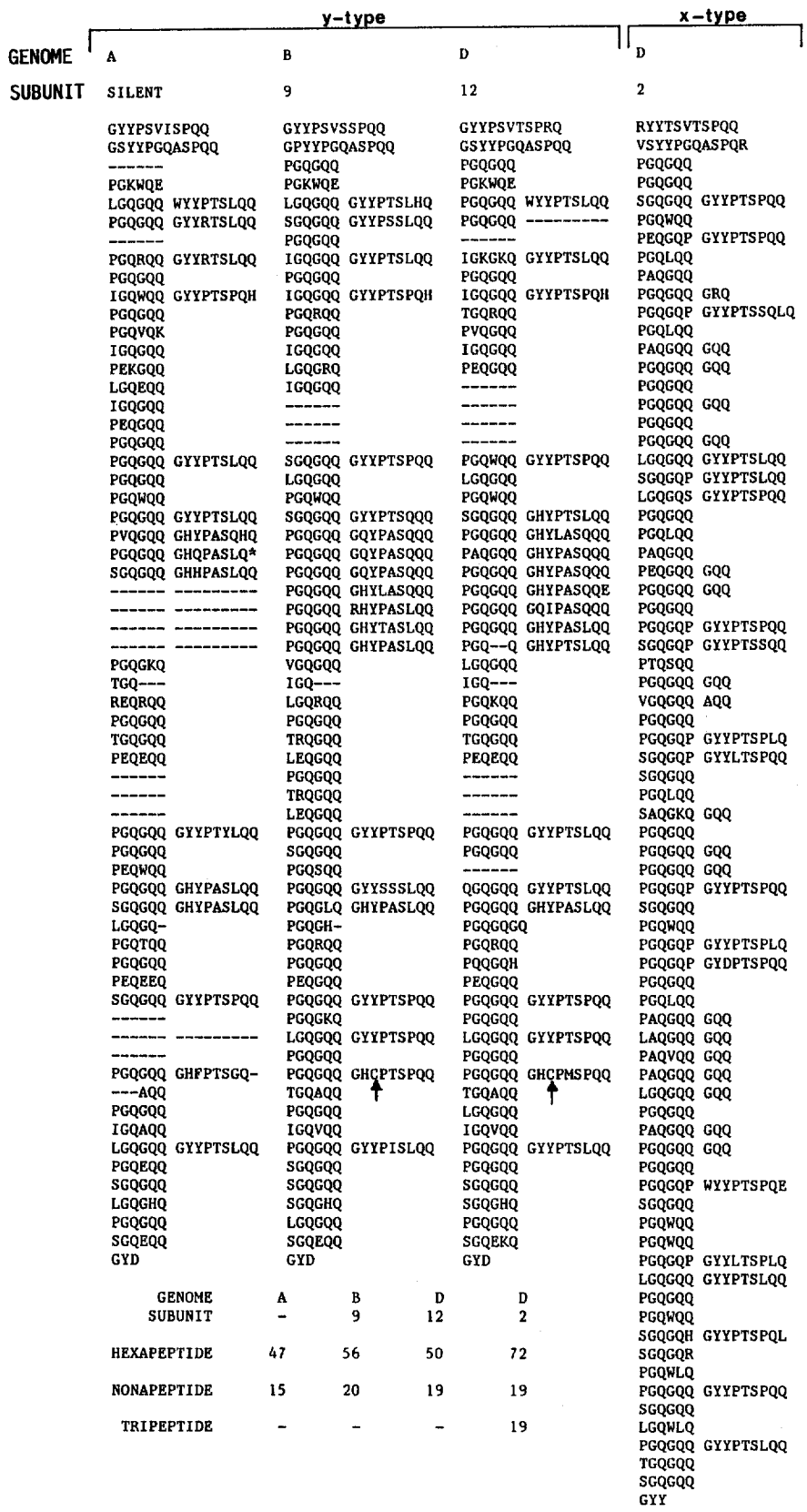


Fig. 5. Comparison of the amino and carboxy-terminal domains of HMW subunits. The sequences of the 1Ay, 1Dy12 and 1Dx2 proteins used and in Figs. 6 and 7 were deduced from genomic DNA by Forde J et al. (1985), Thompson et al. (1985) and Sugiyama et al. (1985) respectively. The partial N-terminal sequence of the 1Bx7 subunit was determined directly by Shewry et al. (1984). Identical residues are *boxed* and deletions represented by *dashes*. Cysteine residues are indicated by *arrows*



GENOME	A	B	D	D
SUBUNIT	-	9	12	2
HEXAPEPTIDE	47	56	50	72
NONAPEPTIDE	15	20	19	19
TRIPEPTIDE	-	-	-	19

Fig. 6. The amino acid sequences of the central domains of the 1Ay, 1By9, 1Dy12 and 1Dx2 subunits arranged to demonstrate their repetitive block structure. The sequences of the y-type subunits are aligned as described in the text, and the cysteine residues present in the 1Dy12 and 1By9 subunits indicated by *arrows*. The sequence of the 1Dx2 subunit can not be aligned with those of the y-type subunits. The total numbers of tripeptides, hexapeptides and nonapeptides in the subunits are tabulated

hexapeptide (consensus pro gly gln gly gln gln) and a nonapeptide. The latter has a consensus sequence of gly tyr tyr pro thr ser leu gln gln in the y-type subunits and gly tyr tyr pro thr ser pro gln gln in the x-type. In addition, the 1Dx2 subunit contains a number of tripeptides (consensus gly gln gln) which may correspond to the second half of the hexapeptide motif. The repetitive domains of all four subunits start with two longer repeats (10 and 11 residues) that may correspond to degenerated nonapeptides. With this exception, the nonapeptide motif is never present in tandem arrays, but is always interspersed with hexapeptides. In contrast, the hexapeptides may be present in tandem arrays of up to eight units. The numbers of tri, hexa and nonapeptides present in the four subunits are summarised in Fig. 6 B. The major difference between the 1Dx subunit 2 and the three y-type subunits is the presence of more hexapeptides and 20 tripeptides within the former, which render it considerably larger than the y-type subunits. The three y-type subunits differ from each other in the numbers of hexa- and nonapeptides.

The repetitive domains of the 1Ay, 1By9 and 1Dy12 subunits are aligned in Fig. 6 A to illustrate their structural relationship. The alignment was made by manual comparison, with regions of clear homology (e.g. truncated repeat motifs or unusual amino acid substitutions) being aligned first. This necessitated the introduction of some gaps. We appreciate that any manual comparison is to some extent subjective and that at least some aspects of the alignment presented in Fig. 6 A could be subject to different intepretations. However, we know of no statistical procedure which would provide a more valid alignment of these sequences and believe our interpretation to be substantially correct. The differences between the y-type sequences are consistent with the three proteins having evolved from a single ancestral protein, by a process involving the intragenic duplication of either two primordial repeat motifs or blocks of such repeat motifs. In addition, deletions of single or small numbers of residues have also occurred.

The repetitive block structure of the 1Dx subunit 2 is also presented in Fig. 6 A and is significantly different from that of the y-type subunits. The 1Dx5 HMW glutenin gene, and its encoded protein subunit, is similar to 1Dx2 in both structure and size (Yip and Malpica, unpublished data). The differences between the 1Dx and y-type subunits are consistent with divergence of the ancestral gene to x- and y-types prior to the establishment of polyploidy.

There are also differences in the extent and type of amino acid substitutions at different positions of the nonapeptide and hexapeptide repeat motifs and in the different subunits (Fig. 7). Position 3 of the hexapeptide (94%–100% Gln) is highly conserved in all the subunits, while positions 1 and 4 of the y-type subunits and 1 and 6 of 1Dx2 are most variant. Similarly postions 6 and 8 of the nonapeptide motif are conserved in all subunits and positions 2 and 5 in the 1Dx2 subunit. The former is always tyrosine in the 1Dx2, but up to 47% histidine in the y-type subunits. The silent 1Ay subunit differs from the 1By9 and 1Dy12 subunits in the presence of a higher proportion of leucine at position 7 of the nonapeptide. This position is also variant in 1Dx2 subunit, with proline being the major residue instead of leucine. A point of interest is the presence of a single cysteine residue at position 3 of a nonapeptide in the 1By 9 and 1Dy 12 subunits. This nonapeptide is towards the C-terminal end of the repetitive domain and is arrowed in Fig. 6 A. The tripeptide repeat motif present in the 1Dx2 subunit is very highly conserved (Fig. 6 A).

It is not known whether the differences in amino acid substitutions result from evolutionary constraints imposed by the role of specific residues in protein conformation. This is certainly implied for residues which are invarient or highly conserved in all the subunits. If so, their precise role is not readily determined. Tatham et al. (1984, 1985) proposed, on the basis of circular dichroism spectroscopy and computer predictions, that the HMW subunits have an unusual secondary structure with no α-helix or β-sheet but rich in β-turns. A β-turn is where the polypeptide chain turns 180°. It involves four residues, with the actual turn occurring between residues 2 and 3. Different amino acids vary in their probabilities of being at positions 1, 2, 3 and 4 of a turn, and the formation of a turn depends on all four amino acids being in favoured positions. It is possible that the conservation of specific amino acids at some positions of the hexapeptide and nonapeptide repeat motifs of the HMW subunits is related to their positions in the predicted β-turns. However, structural prediction indicates the presence of overlapping β-turns, with many residues involved in two or even three predicted turns. It is unlikely that all these turns will actually form, and direct experimental evidence is required to clarify the stuation.

Relationship of HMW subunit structure to gluten structure and grain technology

It has been suggested that the HMW subunits of glutenin have a role in determining the elasticity of wheat gluten (Tatham et al. 1985) and that allelic variation in HMW subunit structure is responsible for differences in breadmaking quality between cultivars (Payne et al. 1984). The structural features discussed above, notably the predominantly terminal locations of cysteine reidues and the β-turn rich conformation of

A-Hexapeptide

	1	2	3	4	5	6
1Ay	Pro 62 Leu 11 Ser 11 Ile 9 Thr 4 Arg 2 – 2	Gly 83 Glu 13 Val 2 – 2	Gln 94 Lys 4 – 2	Gly 66 Glu 11 Trp 9 Ala 4 Arg 4 Thr 2 Val 2 – 2	Gln 92 Glu 2 His 2 Lys 2 – 2	Gln 92 – 4 Glu 2 Lys 2
1By9	Pro 54 Leu 14 Ser 14 Ile 11 Thr 5 Val 2	Gly 91 Glu 5 Arg 4	Gln 98 Lys 2	Gly 82 Arg 5 Trp 4 Ala 2 Glu 2 Ser 2 Val 2 – 2	Gln 89 His 4 Arg 2 Lys 2 Leu 2 – 2	Gln 94 – 4 Glu 2
1Dy12	Pro 66 Ile 10 Leu 8 Ser 8 Thr 6 Gln 2	Gly 88 Glu 6 Ala 2 Arg 2 Val 2	Gln 96 Lys 4	Gly 76 Trp 6 Arg 4 Glu 4 – 4 Ala 2 Lys 2 Val 2	Gln 90 Lys 4 – 4 His 2	Gln 92 Glu 2 Gly 2 His 2 – 2
1Dx2	Pro 72 Ser 17 Leu 8 Val 3 Thr 1	Gly 83 Ala 13 Glu 3 Thr 1	Gln 100	Gly 81 Trp 10 Leu 7 Ser 1 Val 1	Gln 96 Leu 3 Lys 1	Gln 79 Pro 17 Arg 1 His 1 Ser 1

B-Tripeptide

	1	2	3
1Ay			
1By9			
1Dy12			
1Dx2	Gly 95 Ala 5	Gln 100	Gln 100

C-Nonapeptide

	1	2	3	4	5	6	7	8	9
1Ay	Gly 93 Trp 7	Tyr 60 His 40	Tyr 80 Gln 7 His 7 Phe 7	Pro 87 Arg 13	Thr 67 Ala 33	Ser 93 Tyr 7	Leu 73 Pro 13 Gln 7 Gly 7	Gln 93 His 7	Gln 80 His 7 * 7 – 7
1By9	Gly 95 Arg 5	Tyr 55 His 30 Gln 15	Tyr 95 Cys 5	Pro 85 Leu 5 Thr 5 Ser 5	Thr 45 Ala 40 Ser 10 Ile 5	Ser 100	Leu 45 Pro 30 Gln 25	Gln 95 His 5	Gln 95 His 5
1Dy12	Gly 95 Trp 5	Tyr 47 His 47 Gln 5	Tyr 90 Cys 5 Ile 5	Pro 95 Leu 5	Thr 58 Ala 37 Met 5	Ser 100	Leu 47 Pro 26 Gln 26	Gln 100	Gln 90 Glu 5 His 5
1Dx2	Gly 95 Trp 5	Tyr 100	Tyr 95 Asp 5	Pro 90 Leu 10	Thr 100	Ser 100	Pro 68 Leu 21 Ser 10	Gln 84 Leu 16	Gln 84 Leu 10 Glu 5

Fig. 7. Comparison of the amino acid sequences of the repeat motifs present in the repetitive domains of the 1Ay, 1By9, 1Dy12 and 1Dx2 subunits. The amino acids present at each position of the motifs are expressed as a percentage of those present at that position in each individual subunits

the repetitive domain, may be relevant to the role of the subunits in elastic gluten polymers, as discussed by Tatham et al. (1985). However, it is not yet possible to interpret differences in subunit structure in terms of quality associations. This is because subunits 2, 12 and 9 are all associated with poor breadmaking quality. Evaluation of the role of specific subunits in gluten structure and functionality must await the determination of amino acid sequences of allelic subunits with contrasting quality associations and chemical studies of the structures of these subunits and their interactions with other gluten proteins.

Acknowledgements. We are grateful to Drs. B. J. Miflin, A. S. Tatham, B. G. Forde and M. Kreis for advice and discussions. The study was supported by NATO Travel grant No. 324/82 to P.R.S.

References

Anderson OD, Litts JC, Gautier MF, Greene FC (1984) Nucleic acid sequence and chromosome assignment of a wheat storage protein gene. Nucleic Acids Res. 12:8129–8144

Dean C, Tamaki S, Dunsmuir P, Favreau M, Katayama C, Dooner H, Bedbrook J (1986) mRNA transcripts of several plant genes are polyadenylated at multiple sites *in vivo*. Nucleic Acids Res 14:2229–2240

Forde BG, Heyworth A, Pywell J, Kreis M (1985) Nucleotide sequence of a B1 hordein gene and the identification of possible upstream regulatory elements in endosperm storage protein genes from barley, wheat and maize. Nucleic Acids Res 13:7327–7339

Forde J, Miflin BJ (1983) Isolation and identification of mRNA for the high-molecular-weight storage proteins of wheat endosperm. Planta 157:567–576

Forde J, Malpica J-M, Halford NG, Shewry PR, Anderson OD, Greene FC, Miflin BJ (1985) The nucleotide sequence of a HMW glutenin subunit gene located on chromosome 1A of wheat (*Triticum aestivum* L.) Nucleic Acids Res 13:6817–6832

Harberd NP, Bartels D, Thompson RD (1986) DNA restriction-fragment variation in the gene family encoding high molecular (HMW) glutenin subunits of wheat. Biochem Genet 24:579–595

Harris PA, Law CN (1983) The high-molecular-weight subunits of glutenin: classical genetics, molecular genetics and the relationship to bread-making quality. Proc 6th Int Wheat Genet Symp, Kyoto, pp 827–834

Kreis M, Shewry PR, Forde BG, Forde J, Miflin BJ (1985) Structure and evolution of seed storage proteins and their genes with particular reference to those of wheat, barley and rye. Oxford Surv Plant Mol Cell Biol 2:253–317

Lawrence GJ, Shepherd KW (1980) Variation in glutenin protein subunits in wheat. Aust J Biol Sci 33:221–223

Lawrence GJ, Shepherd KW (1981) Inheritance of glutenin protein subunits of wheat. Theor Appl Genet 60:333–337

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning, a laboratory manual. Cold Spring Harbour Laboratory, New York

Messing J, Geraghty D, Heidecker G, Hu N-T, Kridl J, Rubenstein I (1983) Plant gene structure. In: Kosuge T, Meredith CP, Hollaender A (eds) Genetic engineering of plants: an agricultural perspective. Plenum

Miflin BJ, Field JM, Shewry PR (1983) Cereal storage proteins and their effect on technological properties. In: Daussant J, Mosse J, Vaughan J (eds) Seed proteins. Academic Press, London, pp 255–319

Payne PI, Lawrence GJ (1983) Catalogue of alleles for the complex gene loci, Glu-A1, Glu-B1 and Glu-D1 which code for the high-molecular weight subunits of glutenin in hexaploid wheat. Cereal Res Commun 11:29–35

Payne PI, Law CN, Mudd EE (1980) Control by homoeologous group 1 chromosomes of the high-molecular-weight subunits of glutenin, a major protein of wheat endosperm. Theor Appl Genet 58:113–120

Payne PI, Holt LM, Law CN (1981) Structural and genetical studies of the high molecular weight subunits of wheat glutenin: I. Allelic variation in subunits amongst varieties of wheat (*T. aestivum* L.). Theor Appl Genet 60:229–241

Payne PI, Holt LM, Worland AJ, Law CN (1982) Structural and genetical studies on the high-molecular-weight subunits of wheat glutenin. 3. Telocentric mapping of the subunit genes on the long arms of the homeologous group 1 chromosomes. Theor Appl Genet 63:129–138

Payne PI, Holt LM, Thompson RD, Bartels D, Harberd NP, Rafalski JA (1986) Structure of wheat-gamma-gliadin genes. Gene 43:221–229

Shewry PR, Field JM, Faulks AJ, Parmar S, Miflin BJ, Dietler MD, Lew EJ-L, Kasarda DD (1984) The purification and N-terminal amino acid sequence analysis of the high molecular weight gluten polypeptides of wheat. Biochim Biophys Acta 788:23–34

Sugiyama T, Rafalski A, Peterson D, Söll D (1985) A wheat HMW glutenin subunit gene reveals a highly repeated structure. Nucleic Acids Res 13:8729–8737

Sumner-Smith M, Rafalski JA, Sugiyama T, Stoll M, Söll D (1985) Conservation and variability of wheat alpha/beta gliadin genes. Nucleic Acids Res 11:3905–3916

Tatham AS, Shewry PR, Miflin BJ (1984) Wheat gluten elasticity: a similar molecular basis to elastin? FEBS Lett 177:205–208

Tatham AS, Miflin BJ, Shewry PR (1985) The beta-turn conformation in wheat gluten proteins: relationship to gluten elasticity. Cereal Chem 62:405–412

Thompson RD, Bartels D, Harberd NP (1985) Nucleotide sequence of a gene from chromosome 1D of wheat encoding a HMW-glutenin subunit. Nucleic Acids Res 13:6833–6846

Watson MEE (1984) Compilation of published signal sequences. Nucleic Acids Res 12:5145–5164