Theor Appl Genet (1989) 77:689–700

The characterization and comparative analysis of high-molecular-weight glutenin genes from genomes A and B of a hexaploid bread wheat

O. D. Anderson and F. C. Greene

USDA, ARS, Western Regional Research Center, 800 Buchanan Street, Albany, CA 94710, USA

Received September 20, 1988; Accepted December 12, 1988
Communicated by J. Beckmann

Summary. Two high-molecular-weight subunit (HMWS) glutenin genes from the A and B genomes of the hexaploid bread wheat *Triticum aestivum* L. cv Cheyenne have been isolated and sequenced. Both of these genes are of the high M_r class (x-type) of HMW glutenins, and have not been previously reported. The entire set of six HMW genes from cultivar Cheyenne have now been isolated and characterized. An analysis of the Ax and Bx sequences shows that the Ax sequence is similar to the homoeologous gene from the D genome, while the Bx repeat structure is significantly different. The repetitive region of these proteins can be modelled as a series of interspersed copies of repeat modifs of 6, 9, and 15 amino acid residues. The evolution of these genes includes single-base substitutions over the entire coding region, plus insertion/deletions of single or blocks of repeats in the central repetitive domain.

Key words: *Triticum* – Homoeoalleles – Evolution – Dough – Gene library

Introduction

The main wheat storage proteins are divided into two groups, the glutenins, composed of high- and low-molecular-weight families, and the gliadins, composed of α/β, γ, and ω families (Kreis et al. 1985). The high-molecular-weight (HMW) glutenins are of particular interest because of the association of specific alleles of the family with the visco-elastic properties involved in desirable dough formation (Payne et al. 1980; Tatham et al. 1984). One model proposes variations in disulfide linkages among the HMW- and LMW-glutenin subunits which leads to different characteristics of the protein fraction of dough known as gluten (Ewart 1968). However, modelling gluten behavior has been difficult without more information on the primary or higher order structures of the gluten components.

Each of the three genomes (designated A, B, and D) in the hexaploid bread wheats contains two HMW-glutenins, which map together at the *Glu-1* loci on the long arm of the group 1 homoeologous chromosomes (Payne et al. 1982). The two genes at each locus have been classified into x- and y-types according to higher or lower M_r (Payne et al. 1981; Shewry et al. 1984; Payne et al. 1983). A hexaploid bread wheat cultivar thus contains six glutenin genes, although most cultivars only synthesize three to five subunits, presumably because of gene inactivations. The molecular basis of such inactivations are known for two Ay genes which have undergone DNA sequence changes (Forde et al. 1985; Harberd et al. 1987).

The full complement of expressed HMW genes for *Triticum aestivum* cv Cheyenne codes for the following HMW subunits (HMWS); Ax2*, By9, Bx7, Dy10, Dx5. Numbers designate the specific glutenin subunit produced. The Ay gene is silent. The complete sequences of several HMW genes have been reported, including two Dx alleles (Anderson et al. 1989; Sugiyama et al. 1985), two Dy alleles (Thompson et al. 1985; Anderson et al. 1989), one complete (Forde et al. 1985) and one partial (Harberd et al. 1987) Ay gene, and one By gene (Halford et al. 1987). The present paper reports sequences of the Ax and Bx genes from the cultivar Cheyenne. Examples of all six genes have, thus, now been sequenced, including the entire set of HMW-glutenin genes from a single cultivar, i.e., Cheyenne. The Ax2* gene is similar to the previously reported Dx sequences, but the Bx7 gene contains a different pattern of repeats in the repetitive domain. The structures of the genes for the Ax2* and Bx7 HMW

subunits are compared to that of their homoeoallele Dx5, and implications for the evolution of this gene family are discussed.

Materials and methods

Genomic libraries were constructed by ligation of cultivar Cheyenne DNA to EMBL4 EcoRI arms. Wheat DNA was digested to completion with EcoRI and size-fractionated on a 10%–40% sucrose gradient (Maniatis et al. 1982). DNA fractions of the appropriate sizes were ligated to EcoRI + BamHI digested EMBL4 (Frischauf et al. 1983) and plated on P2392 (Stratagene), a P2 lysogen of LE393 allowing spi$^+$ selection. Plaques were transferred to nitrocellulose filters and probed with nick-translated DNA containing the coding region of the Ay gene of cultivar Cheyenne (Forde et al. 1985).

Sequencing was accomplished by subcloning fragments into M13 mp8 and 9 (Messing et al. 1981), then constructing deletion subclones by the method of Dale et al. (1985) and sequencing both strands by the dideoxy procedure (Sanger et al. 1977). Repeated attempts to clone the 5′ EcoRI-Hind III fragment of the Ax2* gene into M13 failed, possibly due to chance incompatibility of some portion(s) of the sequence with survival in M13. The sequence was, however, stable in different plasmid vectors. For this fragment, sequencing in the 3′→5′ direction was accomplished by producing deletions using the exoIII-Mung Bean nuclease procedure with a kit from Stratagene. The 5′→3′ direction was sequenced with oligonucleotide primers made on a Beckman System I Plus DNA Synthesizer and used without purification (Sanchez-Pescador and Urdea 1984). The template for sequencing both orientations was single-stranded plasmid DNA rescued using M13K07 (Mead et al. 1986) infection of Bluescript plasmids (Stratagene).

Nomenclature

The nomenclature system for the HMW-glutenins has changed since this study was begun. Since the current report is the fourth paper in a series describing the structure of the complete set of HMW genes from a single cultivar, we have kept the older nomenclature for simplicity in reading the set of papers. The two nomenclatures are given in Table 1.

Table 1. High-molecular-weight glutenin nomenclature

Subunit		Gene	
Old	New	Old	New
2* or Ax2*	2*	Glu-1Ax2*	Glu-A1-1b
7 or Bx7	7	Glu-1Bx7	Glu-B1-1a
5 or Dx5	5	Glu-1Dx5	Glu-D1-1b
silent	silent	Glu-1Ay	Glu-A1-2a
9 or By9	9	Glu-1By9	Glu-B1-2b
10 or Dy10	10	Glu-1Dy10	Glu-D1-2b

Results and discussion

HMW-glutenin gene isolation and characterization

A genomic library of complete EcoRI digested cultivar Cheyenne DNA was screened with an HMW probe. Clones were isolated containing strongly hybridizing EcoRI inserts of 6.1, 6.7, 8.7, 15.4, and 19–20 kb. Previous analysis of cultivar Cheyenne DNA has associated an approximately 6.8-kb EcoRI fragment with the Ax2* subunit, and an approximately 21.6-kb fragment with the Bx7 subunit (Forde et al. 1985; Harberd et al. 1986). Southern blots (not shown) of both clone λ1B-2 (6.7 kb) and clone λR23 (19–20 kb) found HMW restriction fragment patterns more similar to that already determined for two Dx genes than for the known y-type genes.

On the basis of restriction fragment patterns and the analysis of other HMW genes, clone λ1B-2 was identified as containing the Ax2* gene, and clone λR23 as containing the Bx7 gene. Subsequent subcloning of all the Hind III and Hind III + EcoRI fragments of λR23 suggests a total length of 19.2 kb. The 2-kb discrepancy in the size of the Bx7 insert observed in this paper and of that reported earlier is probably due to difficulty in accurately sizing the original DNA fragments on genomic blots. The EcoRI inserts of clones λ1B-2 and λR23, and their restriction maps are shown in Fig. 1. The Hind III and Hind III + EcoRI fragments containing the coding regions of these two HMW-glutenin genes were subcloned into plasmids and M13, and the sequences were determined for the coding and near flanking portions of the genes (Fig. 2). Both genes contain the putative recognition sequences implicated in transcriptional regulation and found in other HMW-glutenins and storage protein genes (boxed in Fig. 2), i.e., TATA boxes at −91, and '−300' sequences conserved in many cereal storage protein genes and believed to be involved in tissue specific expression (Kreis et al. 1985; Colot et al. 1987). The '−300' sequence is found at approximately −450 in the expressed HMWS genes, but is missing from the silent Ay gene of cultivar Cheyenne. It occurs twice in the Bx7 sequence due to a 53 base tandem exact duplication. Protein gels showing stained cultivar Cheyenne HMWS (not shown) give higher intensities for the Bx7 band relative to the other subunits, while staining intensities of the Dx and Dy band are more nearly equal. This is consistent with higher relative levels of Bx7 expression, but staining artifacts are not ruled out. Similar differences in staining intensities are seen in many cultivars, with the Bx subunit often showing the most intense bands in gels electrophoresis (Galili and Feldman 1983).

The bases just upstream from the initiator codon are highly conserved in eukaryotes [CC(A/G)CC] (Kozak 1987) and are believed to be important in efficient translation initiation. Table 2 compares this region for all six HMW-glutenins of Cheyenne, and with a low-molecular-weight glutenin, and α- and γ-gliadins (three other families of wheat storage proteins genes related to one anoth-

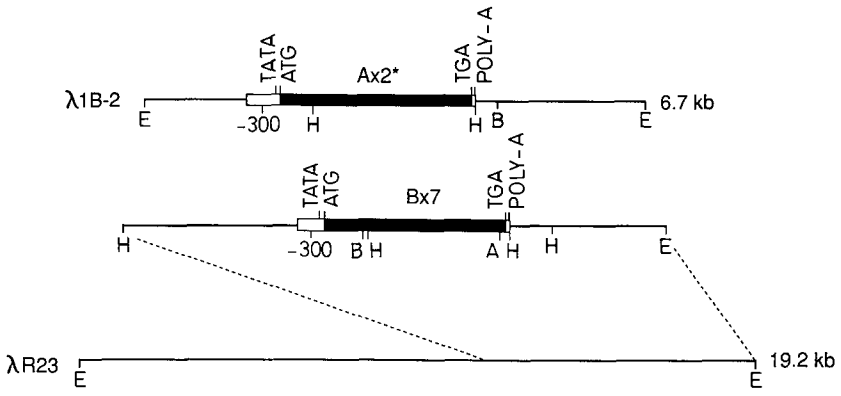


Fig. 1. Clones containing genes for the Ax2* and Bx7 high-molecular-weight glutenin subunits. Clone λ1B-2 contained a 6.7-kb EcoRI fragment encoding the Ax2* gene, and clone λR23 contained a 19.2-kb EcoRI fragment with the Bx7 gene. The region which contains the Bx7 gene is expanded, and transcription and translation control sites of both genes are demarcated and discussed in the text. The sequenced portion of each gene is *boxed*, with the coding region *filled* in. Some restriction sites are shown: A – AccI, B – BamHI, E – EcoRI, H – HindIII

Table 2. Translation start sequences

Gene[a]																		Reference
Ax2*	G	A	C	A	G	T	C	C	A	C	C	G	A	G	A	T	G	A	this paper
Bx7	C	A	C	A	G	T	T	C	A	C	T	G	A	G	A	T	G	G	this paper
Dx5	G	A	T	A	G	T	C	C	A	C	C	G	A	G	A	T	G	G	Anderson et al. 1989
Ay	A	A	C	A	G	T	C	C	A	T	C	G	A	A	A	T	G	G	Forde et al. 1985
By9	A	A	C	A	G	T	C	C	A	T	C	G	A	G	A	T	G	G	Halford et al. 1987
Dy10	G	A	C	A	G	T	C	C	A	C	C	G	A	G	A	T	G	G	Anderson et al. 1989
Consensus:		A	C	A	G	T	C	C	A	C	C	G	A	G	A	T	G	G[b]	
		A	Py	A	G	T	Py	C	A	Py	Py	G	A	G	A	T	G	Pu[c]	
LMW			C	A	C	C	A	A	T	C	C	A	C	C	A	T	G		Colot et al. 1987
α-Gliadin			T	A	C	A	A	A	T	C	C	A	C	C	A	T	G	A	Anderson et al. 1984
γ-Gliadin			A	G	C	A	A	A	T	C	C	A	C	T	A	T	G	A	Bartels et al. 1986
Eurkaryote Consensus[d]										C	C	A/G	C	C	A	T	G	G	Kozak 1987

[a] Gene for designated HMW-subunit
[b] Majority (4 of 5 in active genes)
[c] Consensus for all six HMWS genes
[d] Vertebrate genes

er, but not to the HMWS). While the other wheat storage protein sequences are similar to the eukaryotic consensus, the HMW genes possess three additional bases between the CCACC and the initiator ATG.

Both the Ax2* and Bx7 genes contain a consensus polyadenylation signal (AATAAA) 51 bp downstream of the characteristic double stop of the HMW-glutenins (TGATAG). The existence of multiple polyadenylation sites in the HMW-glutenins is suggested by two HMW cDNA clones with different 3′ ends constructed from polyadenylated RNA (Forde et al. 1983). One of these clones exhibited a poly-A stretch 25 bp 3′ to the AAUAAA, while the other was polyadenylated at least 70 bp downstream of the AAUAAA. The Bx7 gene contains a second potential polyadenylation signal (AATAATAAT) at 118 bp from the translation stop. Potential polyadenylation signals for both the Bx7 and Ax2* genes are underlined in Fig. 2. Multiple polyadenylation signals have been observed in other plants and may be a common feature of such genes (Dean et al. 1986).

The translation of the DNA sequences in Figure 2 into mature protein will produce HMW-glutenins of 86,317 (Ax2*) and 82,873 (Bx7) daltons. The deduced protein sequences for both Ax2* and Bx7 are similar to previously studied HMW-glutenins, i.e., a signal peptide of 21 amino acids followed by an N-terminal domain of 87 amino acid residues for Ax2* and 82 residues for Bx7, a long, repetitive domain of 666 and 647 residues for Ax2* and Bx7, respectively, and finally a C-terminal domain of 42 residues in both genes. The sequences for the N-terminal of the deduced mature proteins coded by

A

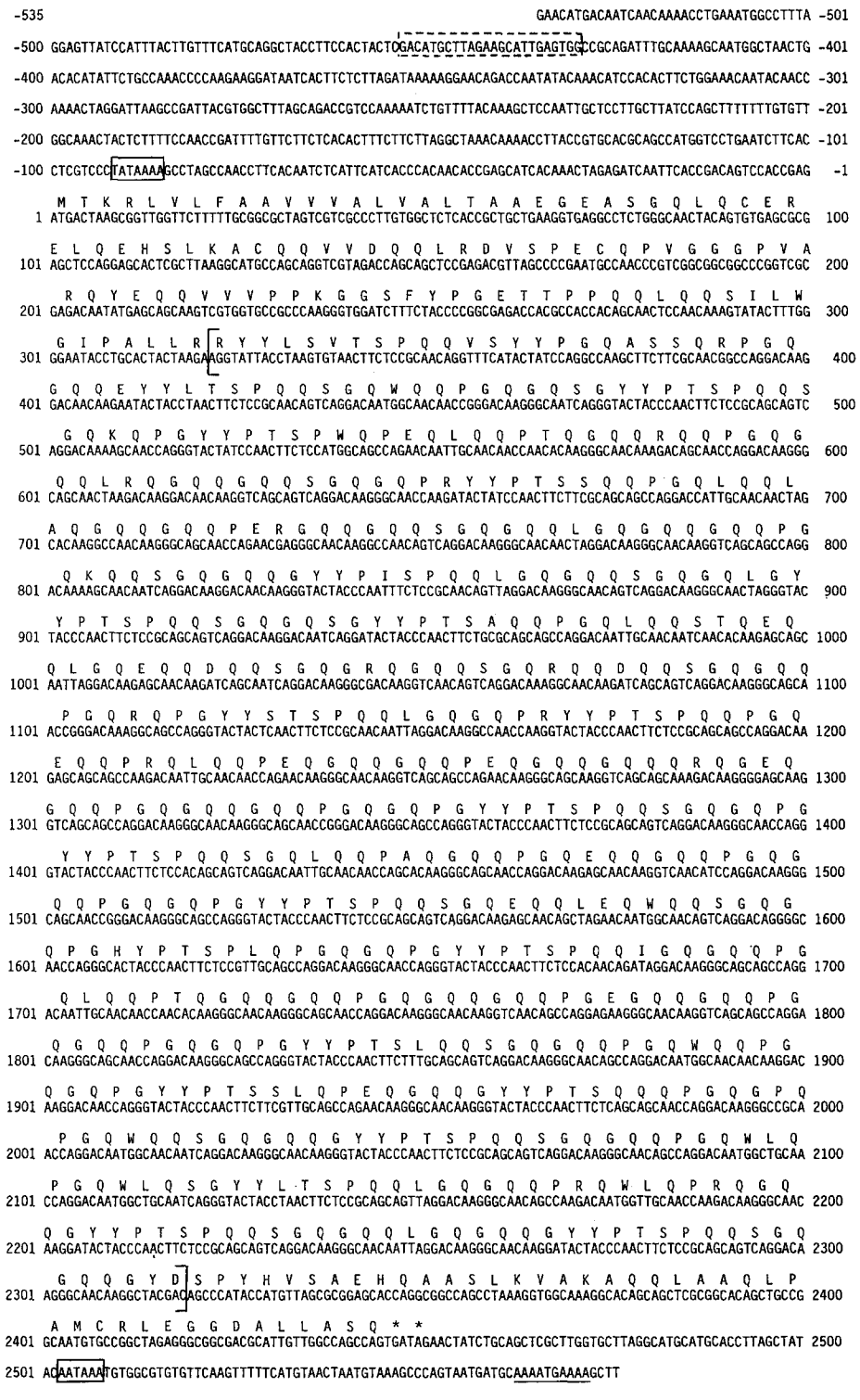
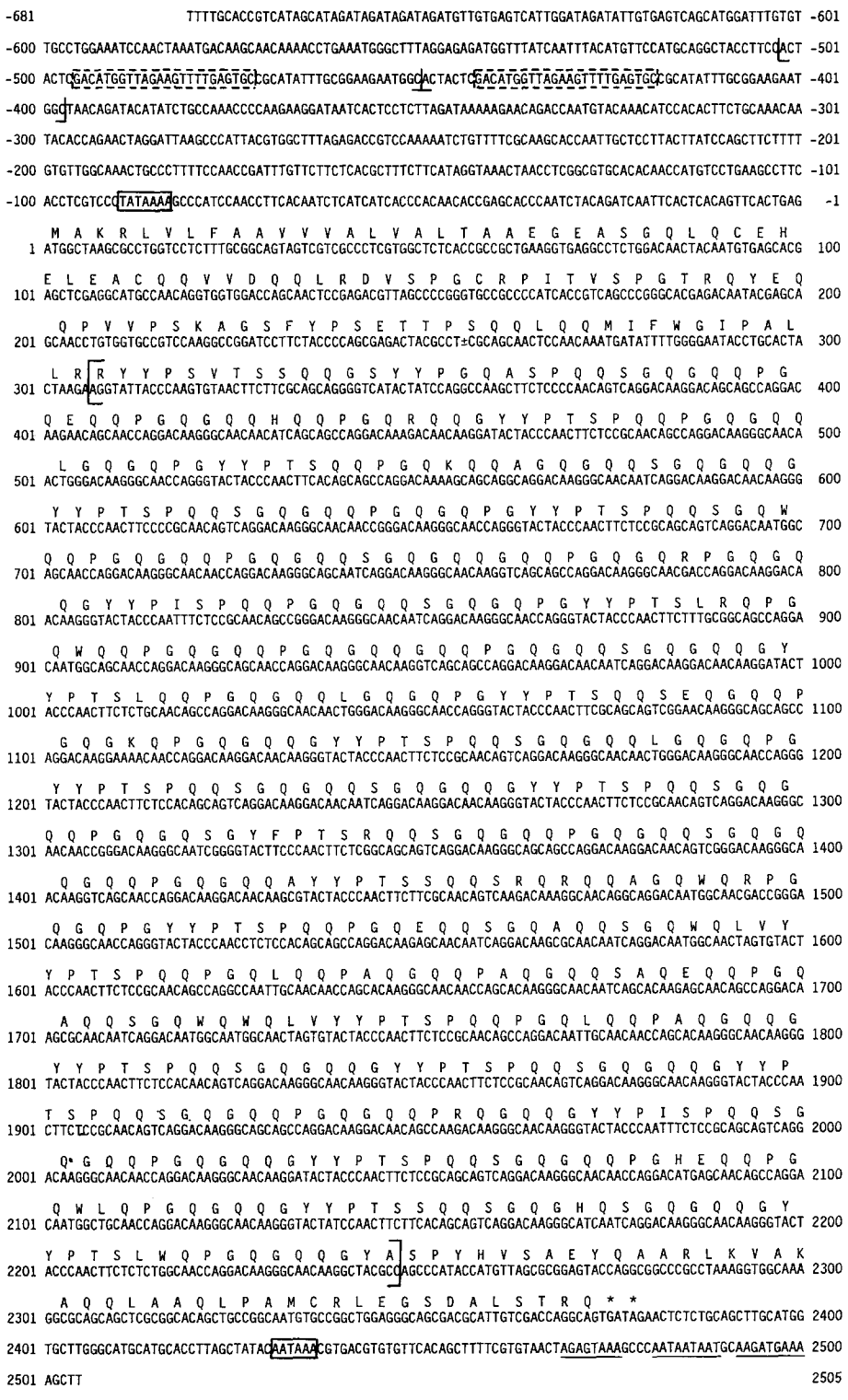


Fig. 2 A and B. Primary structure of the Ax2* **A** and Bx7 **B** genes. TATA boxes and polyadenylation signals are *boxed.* Potential additional polyadenylation signals are *underlined.* The '−300' sequences (see text) are indicated by *dashed boxes,* and are found at about −450 to the start codon. The tandem duplication indcluding the '−300' elements in Bx7 are indicated by *half-brackets.* The repetitive portion of the coding sequences are enclosed by *brackets.* The double translation stop is indicated by *asterisks*

B



clones λ1B-2 and λR23 agree best with the N-terminal protein sequences reported for the Ax1 and Bx7 alleles from cultivar Copain (Table 3), and λ1B-2 is similar to the Ax sequence of *T. monococcum,* an ancestor of the A genome (Shewry et al. 1984). These comparisons confirm the identity of the clones in the current study. The amino acid composition of all three Cheyenne x-type subunits is given in Table 4, and shows the characteristic high Gln, Gly, Pro content of the HMW glutenins. Together these make up about 65% of the amino acid residues and are

Table 3. Known N-terminal amino acid sequences of HMWS. (The unusual abbreviations for amino acids, N and Z, mean: N = not determined and Z = Glu or Gln)

HMWS	Source	Method	Sequence	Reference
Dx2	Highbury, Copain, Brigand	Protein	EGEASEQLQCERELQEL-QERELKACQQVPDEQLZD	Shewry et al. 1984
Dx2	Yamhill	DNA	EGEASEQLQCERELQEL-QERELKACQQVMDQQLRD	Sugiyama et al. 1985
Dx5	Cheyenne	DNA	EGEASEQLQCERELQEL-QERELKACQQVMDQQLRD	Anderson et al. 1989
Dx5	Cheyenne	Protein	EGEASEQLQCENELQEL-QENELKACQQVNDEQLED	Shewry et al. 1984
Dx	*Aegilops squarrosa*	Protein	EGEASEQLQCERELQEL-QENELKAC	Shewry et al. 1984
Dy10	Cheyenne	DNA	EGEASRQLQCERELQE----SSLEACRQVVDQQLAG	Anderson et al. 1989
Dy10	Cheyenne	Protein	EGEASRQLQCERELQE----NNLEAC	Shewry et al. 1984
Dy12	Copain	Protein	EGEASRQLQCERELQE----NQLKACQQV	Shewry et al. 1984
Dy12	Chinese Spring	DNA	EGEASRQLQCERELQE----SSLEACRQVVDQQLAG	Thompson et al. 1985
Ax1	Copain	Protein	EGEASGQLQCENELQENN----LKACQQV	Shewrey et al. 1984
Ax2*	Cheyenne	DNA	EGEASGQLQCERELQEHS ----LKACQQVVDQQLRD	this paper
Ax	*Triticum monococcum*	Protein	EGEASGQLQCERELEENN---- LKAC	Shewry et al. 1984
Ay	Cheyenne	DNA	EGEASKQLQCERELQE----SSLEACRLVVDQQLAS	Forde et al. 1985
Bx7	Copain	Protein	EGEASGQLQCENELE---------ACQQVNDQQL	Shewry et al. 1984
Bx7	Cheyenne	DNA	EGEASGQLQCEHELE---------ACQQVVDQQLRD	this paper
By9	Cheyenne	DNA	EGEASRQLQCERELQE----SSLEACRQVVDQQLAG	Halford et al. 1987

Table 4. Amino acid composition of x-type HMWS [a]

Amino acid	Ax2*	Bx7	Dx5 [a]
Ala	3.0	3.1	3.0
Arg	2.5	2.1	1.1
Ans	0.0	0.0	0.0
Asp	0.7	0.4	0.5
Cys	0.5	0.5	0.6
Gln	33.8	34.0	35.7
Glu	3.9	1.7	2.0
Gly	17.1	18.6	19.8
His	0.5	0.6	0.5
Ile	0.5	0.6	0.5
Leu	5.3	2.9	4.4
Lys	0.9	0.6	0.7
Met	0.2	0.3	0.2
Phe	0.2	0.4	0.2
Pro	11.6	12.1	13.2
Ser	7.5	8.8	5.9
Thr	3.4	3.5	3.0
Trp	1.1	1.2	1.1
Tyr	6.0	7.0	5.7
Val	2.1	1.4	1.7
Acidic	4.6	2.1	2.5
Basic	3.4	2.7	1.8
Aromatic	7.3	8.6	7.0
Hydrophobic	15.4	13.8	13.8
Mol. Wt [b]	86.3	82.9	88.1

[a] Anderson et al. 1989
[b] Kilodaltons

found mainly in the simple repeat motif, PGQGQQ. While the three subunits have a similar overall amino acid content, the Ax2* subunit contains considerably more total charged residues, i.e., 8.0% compared to 4.8% and 4.3% for Bx7 and Dx5, respectively.

As in all other HMW-glutenin genes and all other known cereal storage protein genes, neither the Ax2* nor Bx7 genes contains introns, making the HMW-glutenin genes some of the longest known examples of non-interrupted coding sequences.

The 5′ nonrepetitive DNA sequences of the three x-type genes are similar except for a 9-amino acid segment (Fig. 3 A). The DNA sequence of this region is shown in Fig. 3 B for the three x-type gens, along with the amino acid residues of Dx5. At the DNA level this region of Dx5 is seen as an overlapping series of short repeats, which by deletion/duplication, may form the basis for the relative variability of this portion of the HMWS homoeoalleles. The 3′ nonrepetitive sequence is of identical length in all three genes, with single base changes causing a few amino acid differences (Fig. 3 C).

The cysteine residues within the HMWS figure prominently in models of dough formation. The Ax2* and Bx7 sequences are similar to published Dx sequences in three respects: no cysteines occur within the repetitive portion of the genes; three are present in the 5′ nonrepetitive region; a fourth occurs in the 3′ nonrepetitive sequence.

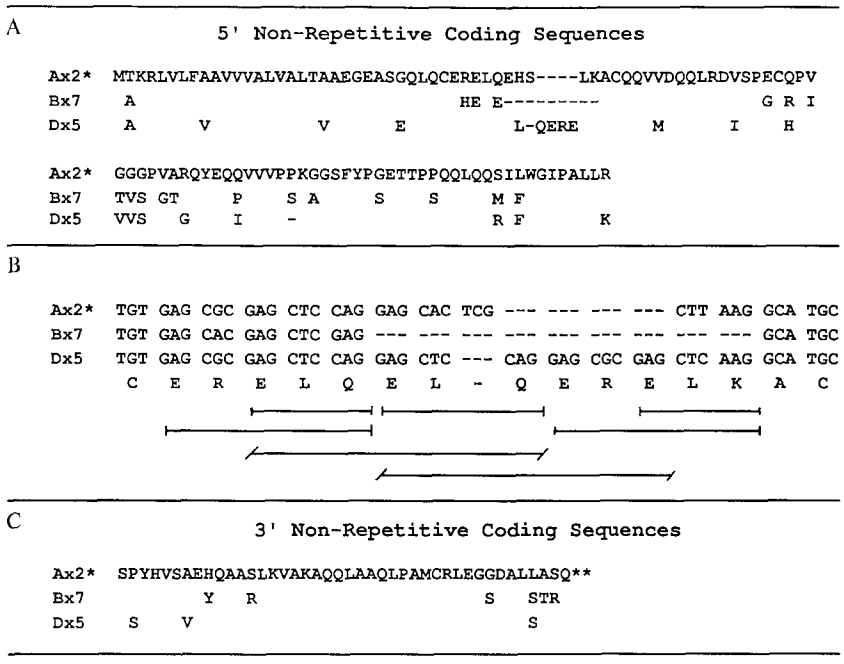


Fig. 3A–C. Comparison of nonrepetitive coding sequences of the x-type HMWS. The sequences flanking the central repetitive region of all three x-type HMWS are aligned for comparison. **A** The 5′ nonrepetitive regions of the three x-type subunits of cultivar Cheyenne. The amino acid sequence of Ax2* is shown along with the sites of difference of Bx7 and Dx5. **B** The DNA sequence of the portion of the 5′ region between the first and second cysteine residues. *Brackets* under the sequences delineate three sets of similar sequences in the DNA gene. The DNA sequence is taken from Greene et al. (unpublished results). **C** The 3′ nonrepetitive region of the three x-type subunits of cultivar Cheyenne

The Dx5 sequence is thus far unique in that in contains a fifth cysteine at the beginning of the repetitive domain (Anderson et al. 1989).

Repeat structure

The dominant feature of the HMW subunits is the central repetitive region made up of repeats of a few fundamental 6–15 amino acid residue motifs. The repeat structures of the three x-type HMW-glutenins of the wheat cultivar Cheyenne are diagrammed in Fig. 4. The repeats of the Ax2* and Dx5 sequences were manually aligned as reported for the y-type sequences (Halford et al. 1987). As the authors pointed out, the alignments are somewhat subjective, and in some places may not reflect the true history of the region. However, by aligning by similar repeat structure and unusual amino acid combinations, we believe this to be a reasonable approximation of the relationship between Ax2* and Dx5, at least for the terminal portions. The Bx7 subunit is significantly different, and cannot be aligned with the other two sequences.

Previously (Halford et al. 1987) we described the basic repeat motifs of the HMWS as: a hexapeptide of PGQGQQ, a nonapeptide of GYYPTSPQQ, and a number of highly conserved GQQ tripeptides associated with the hexapeptide in the 1Dx2 gene, but absent in the sequenced y-type genes. We have now reported sequences from the final two of the six HMWS gene types, plus alleles from the *Glu-D1* locus (Anderson et al. 1989). Further analysis suggests a modification of the model of repeat motif structure. The following features of the known sequences indicate that changes in the repetitive region involve either the hexapeptide motif or a hexapeptide accompanied by either the tripeptide or nonapeptide motifs: (1) the tripeptide and nonapeptide motifs always occur between hexapeptides; (2) runs of up to 10 hexapeptides, or hexapeptides plus one of the other two motifs are a common feature in both the x-type and y-type subunits; (3) when alleles are compared, the differences in the repetitive region suggest that the tripeptide and nonapeptide motifs delete or duplicate mainly in concert with adjoining hexapeptide motifs; (4) hexapeptide motifs occurring both before and after nonapeptides have similar residue preferences, which are slightly different from hexapeptides occurring either with other hexapeptides or with tripeptides. The repetitive region of the HMW-glutenins is, thus, seen as composed of a series of interspersed copies of two (y-type genes) or three (x-type genes) simple units (Fig. 4), all of which begin with the hexapeptide PGQGQQ, followed by: class A, no further residues; class B, the tripeptide GQQ; class C, the nonapeptide GYYPTSPQQ.

When Ax2* and Dx5 are compared (Fig. 4), the repeat pattern between lines 1–19 and 48–80 indicates the clearest relationship between the two genes. Lines 20–47 also show some possible conservation; e.g., the center of the repeat region can be reasonably aligned by repeat classes, but the amino acid differences indicate the probability of different origins. In the Dx5 and Ax2* sequences, lines 53–62 show a run of class B repeats, particularly in the Dx5 sequence. Similar such runs have been seen in the y-type subunits for class A and C repeats

	Ax2*	Dx5	Bx7
1	RYYLSVTSPQQV	RYYPSVTCPQQV	RYYPSVTSSQQG
2	SYYPGQASSQR	SYYPGQASPQR	SYYPGQASPQQ
3	------	PGQGQQ	SGQGQQ
4	PGQGQQ EYYLTSPQQ	PGQGQQ GYYPTSPQQ	PGQEQQ
5	SGQWQQ	PGQWQQ	PGQGQQ HQQ
6	PGQGQS GYYPTSPQQ	PEQGQP RYYPTSPQQ	PGQRQQ GYYPTSPQQ
7	SGQKQP GYYPTSPWQ	------ ---------	PGQGQQ
8	PEQLQQ	SGQLQQ	LGQGQP GYYPTS-QQ
9	PTQGQQ RQQ	------ ---	PGQKQQ
10	PGQGQQ	PAQGQQ	AGQGQQ
11	LRQGQQ GQQ	PGQGQQ GQQ	SGQGQQ GYYPTSPQQ
12	SGQGQP RYYPTSSQQ	PGQGQP GYYPTSSQLQ	SGQGQQ
13	PGQLQQ	PGQLQQ	PGQGQP GYYPTSPQQ
14	LAQGQQ GQQ	PAQGQQ GQQ	SGQWQQ
15	PERGQQ GQQ	PGQGQQ GQQ	PGQGQQ
16	SGQGQQ	PGQGQQ	PGQGQQ
17	LGQGQQ GQQ	PGQGQQ	SGQGQQ GQQ
18	PGQKQQ	PGQGQQ	PGQGQR
19	SGQGQQ GYYPISPQQ	LGQGQQ GYYPTSLQQ	PGQGQQ GYYPISPQQ
20	LGQGQQ	------	PGQGQQ
21	SGQGQL GYYPTSPQQ	SGQGQP GYYPTSLQQ	SGQGQP GYYPTSLRQ
22	SGQGQS GYYPTSAQQ	LGQGQS GYYPTSPQQ	PGQWQQ
23	------	PGQGQQ	PGQGQQ
24	PGQLQQ	PGQLQQ	PGQGQQ GQQ
25	STQEQQ	PAQGQQ	PGQGQQ
26	LGQEQQ DQQ	PEQGQQ GQQ	SGQGQQ GYYPTSLQQ
27	SGQGRQ GQQ	PGQGQQ GQQ	PGQGQQ
28	SGQRQQ DQQ	------ ---	LGQGQP GYYPTS-QQ
29	SGQGQQ	PGQGQQ	SEQGQQ
30	PGQRQP GYYSTSPQQ	PGQGQP GYYPTSPQQ	PGQGKQ
31	LGQGQP RYYPTSPQQ	SGQGQP GYYPTSSQQ	PGQGQQ GYYPTSPQQ
32	PGQEQQ	PTQSQQ	SGQGQQ
33	PRQLQQ	------	LGQGQP GYYPTSPQQ
34	PEQGQQ GQQ	------ ---	SGQGQQ
35	PEQGQQ GQQ	PGQGQQ GQQ	SGQGQQ GYYPTSPQQ
36	QRQGEQ GQQ	VGQGQQ AQQ	SGQGQQ
37	PGQGQQ GQQ	PGQGQQ ---	PGQGQS GYFPTSRQQ
38	PGQGQP GYYPTSPQQ	PGQGQP GYYPTSPQQ	SGQGQQ
39	SGQGQP GYYPTSPQQ	SGQGQP GYYLTSPQQ	PGQGQQ
40	SGQLQQ	SGQGQQ	SGQGQQ GQQ
41	PAQGQQ	PGPLQQ	PGQGQQ AYYPTSSQQ
42	PGQEQQ GQQ	SAQGQK GQQ	SRQRQQ
43	PGQGQQ	PGQGQQ	AGQWQR
44	------ ---	PGQGQQ GQQ	PGQGQP GYYPTSPQQ
45	------ ---	PGQGQQ GQQ	PGQEQQ
46	PGQGQP GYYPTSPQQ	PGQGQP GYYPTSPQQ	SGQAQQ
47	SGQEQQ	SGQGQQ	SGQWQL VYYPTSPQQ
48	LEQWQQ	PGQWQQ	PGQLQQ
49	SGQGQP GHYPTSPLQ	PGQGQP GYYPTSPLQ	PAQGQQ
50	PGQGQP GYYPTSPQQ	PGQGQP GYDPTSPQQ	PAQGQQ
51	IGQGQQ	PGQGQQ	SAQEQQ
52	PGQLQQ	PGQLQQ	PGQAQQ
53	PTQGQQ GQQ	PAQGQQ GQQ	SGQWQWQL VYYPTSPQQ
54	PGQGQQ GQQ	LAQGQQ GQQ	PGQLQQ
55	PGEGQQ GQQ	PAQVQQ GQR	PAQGQQ GYYPTSPQQ
56	------ ---	PAQGQQ GQQ	SGQGQQ GYYPTSPQQ
57	------ ---	PGQGQQ GQQ	SGQGQQ GYYPTSPQQ
58	------ ---	LGQGQQ GQQ	SGQGQQ
59	------ ---	PGQGQQ GQQ	PGQGQQ
60	------ ---	PGQGQQ GQQ	PRQGQQ GYYPISPQQ
61	------ ---	PGQGQQ GQQ	SGQGQQ
62	------ ---	PGQGQQ GQQ	PGQGQQ GYYPTSPQQ
63	PGQGQQ	PGQGQQ	SGQGQQ
64	PGQGQP GYYPTSLQQ	PGQGQP WYYPTSPQE	PGHEQQ
65	SGQGQQ	SGQGQQ	PGQWLQ
66	PGQWQQ	PGQWQQ	PGQGQQ GYYPTSSQQ
67	PGQGQP GYYPTSSLQ	PGQGQP GYYLTSPLQ	SGQGHQ
68	PEQGQQ GYYPTSQQQ	LGQGQQ GYYPTSLQQ	SGQGQQ GYYPTSLWQ
69	PGQGPQ	PGQGQQ	PGQGQQ GYA
70	PGQWQQ	PGQWQQ	
71	SGQGQQ GYYPTSPQQ	SGQGQH GYYPTSPQL	
72	SGQGQQ	SGQGQR	
73	PGQWLQ	PGQWLQ	
74	PGQWLQS GYYLTSPQQ	PGQGQQ GYYPTSPQQ	
75	LGQGQQ	SGQGQQ	
76	PRQWLQ	LGQWLQ	
77	PRQGQQ GYYPTSPQQ	PGQGQQ GYYPTSLQQ	
78	SGQGQQ	TGQGQQ	
79	LGQGQQ GYYPTSPQQ	------ ---------	
80	SGQGQQ GYD	SGQGQQ GYY	

Fig. 4. Repetitive region of the x-type HMWS. The amino acid residues of all three x-type HMWS of cultivar Cheyenne are ordered by repeat classes. The Ax2* and Dx5 (Anderson et al. 1989) repeats are aligned manually and some gaps inserted. The Bx7 gene is not aligned. the *solid line block* in the Bx7 sequence encloses the repetitive portion of the cultivar Chinese Spring Bx7 HMWS (Thompson et al. 1983). The *internal dotted block* is missing in the Chinese Spring sequence. The two *bracketed blocks* toward the C-terminus of the Bx7 sequence delineate an apparent duplication of a single block. Lines are numbered to facilitate discussion in the text

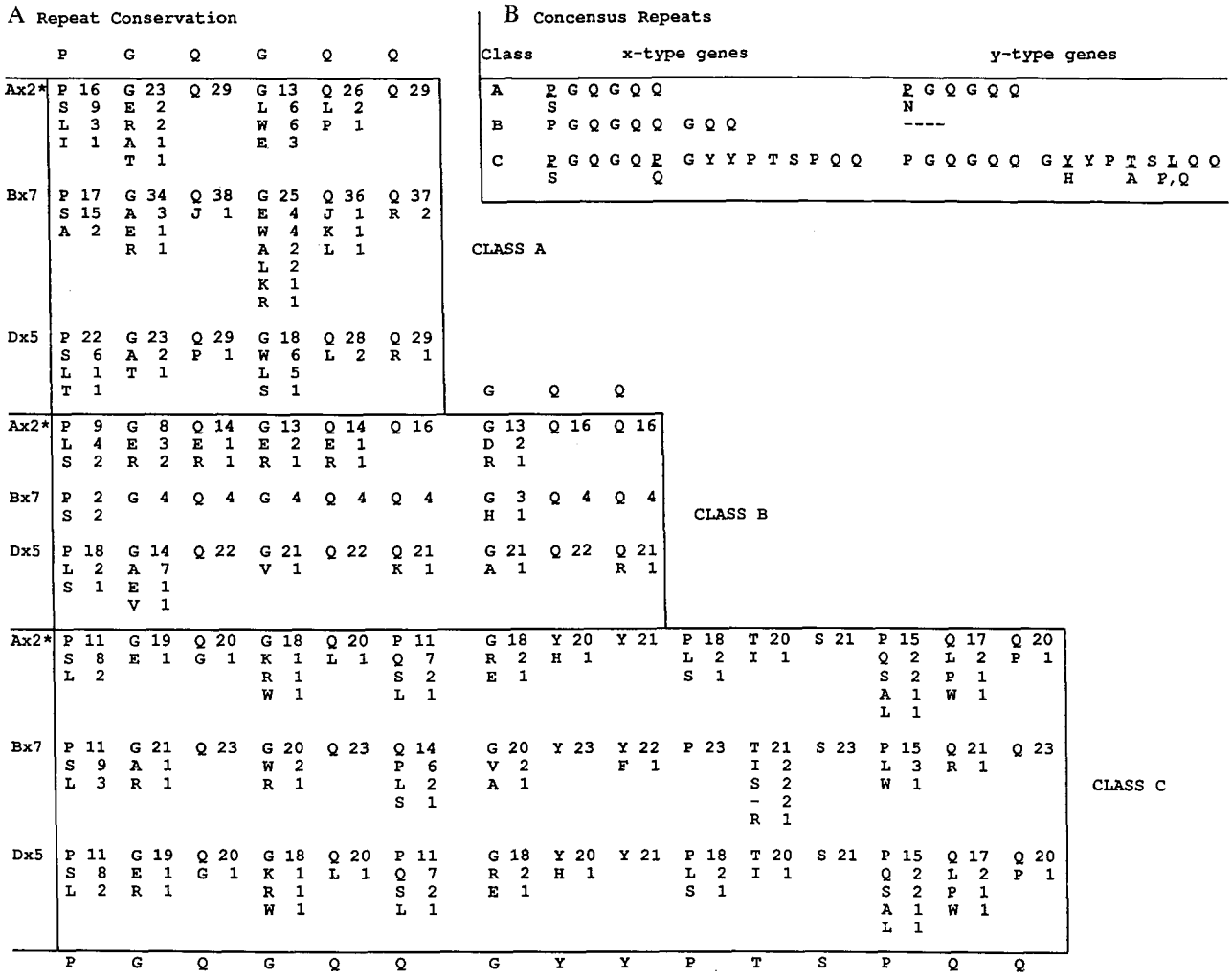


Fig. 5 A and B. Amino acid residue distribution in repeat classes. **A** The total amino acid distribution in the three repeat classes (A, B, and C) is listed vertically for each position for the three x-type HMWS of cultivar Cheyenne (Fig. 4). The hexapeptides are aligned for all three classes. The consensus residues are given at the top for the class A hexapeptide and at the bottom for the class C hexapeptide + nonapeptide. The class B pattern is the hexapeptide + tripeptide. **B** The consensus residues the class A, B, and C repeats for both x-type and y-type (Halford et al. 1987) subunits

(Halford et al. 1987; Anderson et al. 1989). The Bx7 repeat sequence is too divergent from those of Ax2* and Dx5 for alignment, with the possible exception of the most C-terminal portion of the region, which shows some similarity with the other x-type subunits.

A tally of the three motifs in all three x-type subunits gives the following distribution of A, B and C classes: Ax2*, 28-16-21; Bx7, 39-4-23; Dx5, 35-19-19. The main differences are that Bx7 contains relatively more of the class A repeats, but only 4 of the class B. Figure 5 shows the distribution of amino acids at each position in the three repeat classes of the x-type subunit of cultivar Cheyenne, and gives the consensus repeats for both the x- and y-type genes. It may be suspected that the more conserved a specific position in a repeat, the more significant a role is played by the amino acid in that position. However, not enough is known about how the structure of the wheat storage proteins is related to constraints of seed development or utilization during germination, to distinguish conservation by selection from conservation by DNA sequence structure constraints. All of the glutamine positions are highly conserved (95%), but the single most stable position is the serine in the nonapeptide in the class C repeats. Only one variant occurs at this position in all six HMWS genes of cultivar Cheyenne, and that is in the silent Ay gene. Also relatively invariant are the two tyrosines in the class C nonapeptide of the x-type subunits; tyrosine occurs in 64 and 63 out of the 65 positions, respectively (the y-type subunits often have histidine in the first position). The main difference at specific residue sites among the consensus x-type repeats is that the Ax and Dx genes favor proline at position 6 of

the class C repeat, while the Bx gene favors glutamine, as do the y-type genes (Halford et al. 1987).

Evolution of the repeat domain

By comparing sequences of homoeologous genes from all three genomes with each other and with alleles of the same loci (Halford et al. 1987; Anderson et al. 1989; Anderson and Greene, current paper), the following modes of sequence evolution are proposed for the repeat region of the HMW-glutenin genes, starting with the simplest mechanism: (1) Single base changes: leading to either no effect or amino acid substitutions. (2) Deletions or additions within a repeat: most seem to be loss/gain of three codons forming the GQQ tripeptide, but there are a few examples of 1–2 codon duplications/deletions. (3) Single repeat changes: suggesting a single repeat has been either duplicated or deleted. (4) Deletions or duplications of blocks of repeats: while it is difficult to distinguish this mechanism from a series of single repeat changes, especially when examining runs of the same repeat class, some differences are most easily explained by a single event removing or duplicating a block of repeats. The three clearest examples occur in the two B genome HMWS of cultivar Cheyenne. In Fig. 4, the solid blocked portion of Bx7 matches a region reported for an unknown HMWS from cultivar Chinese Spring (Thompson et al. 1983). The amino acid residue and DNA sequences of the Chinese Spring Bx7 subunit both match the Cheyenne Bx7 subunit, except for the apparent deletion of a block of repeats (dotted block in Fig. 4). A second example is an apparent duplication within the Cheyenne Bx7 sequence. Lines 45–49 possess unusual amino acid residue positions and combinations very similar to lines 51–55 (bracketed), and this is probably due to a five repeat block tandem duplication. Similarly, an examination of the By9 sequence (Halford et al. 1987) indicates a three hexapeptide block exact tandem duplication in the center of the repeat region.

In addition to the above discussed methods of repeat change, there may be mechanisms favoring certain series of events, such as expansion or contraction of one type of repeat. It may be that runs are particularly susceptible to sequence alteration, e.g., unequal crossover, slip-mismatching, etc.

While it is possible to account for most differences between genes at different loci and between alleles, a few repeats do not fit easily into the above categories. Lines 8 and 28 (Fig. 4) of Bx7 illustrate one such problem. The repeat sequences are identical, with enough characteristic amino acid combinations to make it likely to have a common origin, yet they are widely separated. The current distance between them could be explained either by divergence of the intervening sequence after an original duplication or by the occurrence of a specific duplication across a greater distance. There is no simple model for the latter.

The repeat arrangement in Fig. 4 emphasizes changes occurring strictly in repeat units. Note, however, that the repeat borders may not delineate the boundaries of repeat changes. If a crossover, etc., occurs between bases at the same position within two repeats, the exact position of the crossover may not be evident unless unusual amino acid or base combinations are present to serve as markers. That at least some of the repeat changes do occur precisely with whole consensus repeats is suggested in lines 1–4. There, the two non-consensus repeats related to the nonapeptide motif beginning the repetitive portion of the sequence are too different to be easily involved in sequence interactions with the following repeats, yet the following repeats do show changes between homoeoalleles. It is also interesting that the repetitive portion of the HMW proteins not only begins with two peptides related to the nonapeptide part of the type C repeat, but also ends with the first nine amino acids of the type C repeat.

As noted above, the Bx repeat region is significantly different from the Ax and Dx pattern, which may indicate a more distant ancestry. Payne et al. (1981) and Galili and Feldman (1983) observed that the *Glu-B1* alleles show significantly higher variation in mobility on SDS-PAGE than proteins from the *Glu-A1* and *Glu-D1* loci, and it was suggested that the origin of the B genome is polyphyletic. Both of these observations could be explained if the B genes change at a faster rate than the A and D genes. We know of no mechanism for such a difference, but it may be supported by the observation that the only three clear examples of repeat block duplication occur in genes from the B genome, implying that the duplications have occurred recently enough that subsequent changes have not obscured the events.

Identification of unknown sequences and HMW nomenclature

Now that examples of all six HMW genes have been sequenced, it will be possible to assign unknown protein or cDNA sequences to specific loci, even with only partial sequence information. For example, N-terminal amino acid sequence data can identify specific subunits (Table 3). Similarly, partial DNA sequences (which usually yield C-terminal information) should also be sufficient for identification. Two partial cDNA sequences derived from endosperm mRNA of the cultivar Sentry have been reported to code for HMWS of unknown identity (Forde et al. 1983). One clone, pc256, has a repeat pattern similar to Ax2*. In addition, the 3′ nonrepetitive sequence is closer to Ax2* than the other five genes. A second clone, pc237, which only contains the 3′ nonrepetitive sequence, is similar to Dx5. Thus, these two par-

tial cDNAs code for the 3′ portions of the Ax1 and Dx3 subunits, the Ax and Dx alleles of cultivar Sentry. Similarly, a partial sequence from the cultivar Chinese Spring (Thompson et al. 1983) is similar to the Bx7 gene. It may now even be possible to determine allelic differences from mRNA preparations without resorting to cloning using specific sequencing primers on total DNA or mRNA preparations.

Even though the cultivars Cheyenne and Chinese Spring contain Bx subunits designated 7 by similar mobilities on SDS-PAGE, we have noted above a difference in their DNA sequences. It seems reasonable to assume that even subunits with the same characteristics will often contain both single amino acid residue and repeat unit changes. Thus, similarity in the physical properties of the subunits, such as apparent molecular weights from gel electrophoresis, are not themselves sufficient to conclude identity of two HMWS; DNA or protein sequence is required.

Relationship to dough quality

Studies have shown that the 1D genes correlate both positively and negatively with bread-making quality (Payne et al. 1984; Branlard and Dardevet 1985), depending on the specific alleles present in the wheat variety. The Ax genes are also associated with good quality, while the B genes (both x- and y-types) have not shown clear correlations. The finding of an additional cysteine residue in the good quality Dx5 gene (Anderson et al. 1988) when compared to a poorer quality related allele (Dx2) supports a model of a more complex disulfide cross-linked glutenin matrix leading to improved dough visco-elasticity. Both the Ax2* and Bx7 subunits lack the extra cysteine residue found in Dx5 (but not in the poorer quality subunit Dx2). However, there is still no conclusive connection between the extra cysteine and quality, nor an indication whether some other, less obvious, structural feature is responsible. Both of the sequences reported here, the Bx with its different repeat pattern, and the Ax with its increased charge density, have properties that could affect their participation in the glutenin complex or in the non-covalent interactions with members of the other families of storage proteins.

The complete characterization of all the HMW-glutenin genes from a single cultivar should aid in future research to understand the relationship between individual gene products and quality contribution to dough formation. Further work needs to be done on how specific structural features of these proteins relate to desired quality characteristics. One potential approach is the use of genes, both native and modified, to produce individual proteins (or combinations of proteins) through biotechnological methods. In addition, the detailed comparative study of the DNA sequences of the different HMW genes from the three homoeologous loci and between y-type and x-type genes will contribute to our understanding of the evolution of this gene family.

Acknowledgements. We wish to thank D. Kasarda, A. Blechl, and C. Fornari for discussions and suggestions, and T. Stephens for excellent technical assistance.

References

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

Anderson OD, Greene FC, Yip RE, Halford NG, Shewry PR, Malpica-Romero J-M (1989) Nucleotide sequences of the two high-molecular-weight glutenin genes from the D-genome of a hexaploid bread wheat, *Triticum aestivum* L. cv Cheyenne. Nucleic Acids Res 17:461–462

Bartels D, Altosaar I, Harberd NP, Barker RF, Thompson RD (1986) Molecular analysis of γ-gliadin gene families at the complex locus of bread wheat (*T. aestivum* L.). Theor Appl Genet 72:845–853

Branlard G, Dardevet M (1985) Diversity of grain protein and bread wheat quality. II. Correlation between high molecular weight subunits of glutenin and flour quality characteristics. J Cereal Sci 3:345–354

Colot V, Robert LS, Kavanagh TA, Bevan MW, Thompson RD (1987) Localization of sequences in wheat endosperm protein genes which confer tissue-specific expression in tobacco. EMBO J 6:3559–3564

Dale RMK, McClure BA, Houchins JP (1985) A rapid single-stranded cloning strategy for producing a sequential series of overlapping clones for use in DNA sequencing: application to sequencing the corn mitochondrial 18 S rDNA. Plasmid 13:31–40

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

Ewart JAD (1968) A hypothesis for the structure and rheology of glutenin. J Sci Food Agric 19:617–623

Forde J, Forde BG, Fry RP, Kreis M, Shewry PR, Miflin BJ (1983) Identification of barley and wheat cDNA clones related to the high-M_r polypeptides of wheat gluten. FEBS Lett 162:360–366

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

Frischauf A-M, Lehrach H, Poustka A, Murray N (1983) Lambda replacement vectors carrying polylinker sequences. J Mol Biol 170:827–842

Galili G, Feldman M (1983) Genetic control of endosperm proteins in wheat: 2. Variation in high molecular weight glutenin and gliadin subunits of *Triticum aestivum*. Theor Appl Genet 66:77–86

Halford NG, Forde J, Anderson OD, Greene FC, Shewry PR (1987) 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. Theor Appl Genet 75:117–126

Harberd NP, Bartels D, Thompson RD (1986) DNA restriction-fragment variation in the gene family encoding high molecu-

lar weight (HMW) glutenin subunits in wheat. Biochem Genet 24:579–596

Harberd NP, Flavell RB, Thompson RD (1987) Identification of a transposon-like insertion in a *Glu-1* allele of wheat. Mol Gen Genet 209:326–332

Kozak M (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acid Res 15: 8125–8148

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 for wheat, barely and rye. Oxford Surv Plant Mol Cell Biol 2:253–317

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor/NY

Mead DA, Szczesna-Skorupa E, Kemper B (1986) Single-stranded DNA 'blue' T7 promoter plasmids: a versatile tandem promoter system for cloning and protein engineering. Protein Eng 1:67–74

Messing J, Crea R, Seeburg PH (1981) A system for shotgun DNA sequencing. Nucleic Acids Res 9:309–321

Payne PI, Harris PA, Law CN, Holt LM, Blackman JA (1980) The high-molecular-weight subunits of glutenin: structure, genetics and relationship to bread-making quality. Ann Technol Agric 29:309–320

Payne PI, Holt LM, Law CN (1981) Structural and genetical studies on the high-molecular-weight subunits of wheat glutenin. Part 1: allelic variation in subunits amongst varieties of wheat (*Triticum aestivum*). Theor Appl Genet 60:229–236

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

Payne PI, Holt LM, Thompson RD, Bartels D, Harberd NP, Harris PA, Law CN (1983) The high-molecular-weight glutenins: classical genetics, molecular genetics and the relationship to bread-making quality. Proc 6th Int Wheat Genet Symp 827–834

Payne PI, Holt LM, Jackson EA, Law CN (1984) Wheat storage proteins: their genetics and their potential for manipulation by plant breeding. Philos Trans R Soc London Ser B 304:359–371

Sanchez-Pescador R, Urdea MS (1984) Use of unpurified primers in sequencing. DNA 3:339–343

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. PNAS 74:5463–5467

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

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

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

Thompson RD, Bartels D, Harberd NP, Flavell RB (1983) Characterization of the multigene family coding for HMW glutenin subunits in wheat using cDNA clones. Theor Appl Genet 67:87–96

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