Theor Appl Genet (1998) 96: 743–750

B. G. Cassidy · J. Dvorak · O. D. Anderson

The wheat low-molecular-weight glutenin genes: characterization of six new genes and progress in understanding gene family structure

Received: 13 May 1997 / Accepted: 1 September 1997

Abstract Although the low-molecular-weight (LMW) glutenin subunits are important for aspects of wheat quality and dough processing, a detailed description of the DNA structure and encoded polypeptides of this multigene family is still lacking. We report progress in obtaining a more thorough description of the LMW-glutenin gene family from a single wheat cultivar ('Cheyenne'). Six new genomic sequences are reported and compared with other LMW-glutenin DNA sequences. Subfamilies of sequences are identified, and an analysis of the repetitive domain of these polypeptides suggests a simple codon pattern with implications for modes of evolution of these repeat motifs. Southern analysis is used to estimate 30–40 members of this gene family in cv 'Cheyenne', and chromosome assignments are made for most restriction fragments, including the six sequenced genes. The known DNA sequences cluster into two groups, and most of the new sequences are tentatively identified as C-type LMW-glutenins. Representatives of the B-genome genes are still lacking.

Key words Wheat · Low-molecular-weight glutenins · Quality · Storage protein · Multigene family

Communicated by G. E. Hart

B. G. Cassidy[1] · J. Dvorak
Department of Agronomy and Range Science, University of California, Davis, CA 95616, USA

O. D. Anderson (✉)
U.S. Department of Agriculture, Agricultural Research Service, Western Regional Research Center, 800 Buchanan Street, Albany, CA 94710, USA

Present address:
[1] The Samuel Roberts Noble Foundation, P.O. Box 2180, Ardmore, OK 73402, USA

Introduction

The wheat endosperm is a major component of the human diet largely due to the unique physical properties of wheat flour water mixtures (Pomeranz 1988). The resulting doughs are made into a wide range of foods: leavened and unleavened breads, pastas, noodles, cakes, etc. The basis of dough functionality is distributed among many different components of the seed: proteins, carbohydrates, lipids. However, the clearest correlations are to the seed proteins, and particularly the prolamines: alcohol-soluble reduced polypeptides rich in proline and glutamine (Macritchie 1992). The high- and low-molecular-weight (HMW and LMW) glutenins are subunits of the insoluble, disulfide-crosslinked matrix critical for dough properties such as viscoelasticity. The important role of the HMW-glutenins has been well established (reviewed in Shewry et al. 1992, 1995), and significant effort has been expended to isolate these genes. Less attention has been paid to the LMW-glutenins genes although their importance to dough properties is also well known (Payne 1987; Gupta et al. 1994; Gupta and Macritchie 1994).

The LMW-glutenins have been mapped to the *Glu-3* loci on the short arms of the group 1 homoeologous chromosomes (Singh and Shepherd 1988). Estimates of the number of LMW-glutenin genes has varied from 10–15 (Harberd et al. 1985) to 35 (Sabelli and Shewry 1991). Only two LMW-glutenin genomic clones have been reported (Pitts et al. 1988; Colot et al. 1989). In addition, two complete and three partial cDNA gene sequences are known (Bartels and Thompson 1983; Okita et al. 1985; Cassidy and Dvorak 1991). More recently, four partial coding sequences from polymerase chain reaction (PCR) products have been reported (D'Ovidio et al. 1992; Van Campenhout et al. 1995). These complete and partial sequences have been from seven different cultivars, both hexaploid and

tetraploid. Three of the sequences are from the hard red winter bread wheat cultivar 'Cheyenne', the same cultivar whose complete set of HMW-glutenin genes has been characterized. The understanding of the function of the LMW-glutenins would be aided by a more complete description of the gene family from a single cultivar, with cv 'Cheyenne' a logical candidate. Therefore, we have begun a project to isolate and characterize a more complete LMW-glutenin gene set from cv 'Cheyenne' and progress in this effort is reported here. Six new LMW-glutenin genomic sequences and all previously reported sequences were used to analyze the LMW-glutenin gene and polypeptide structure. Southern analysis was used to re-examine the number of gene copies with the LMW-glutenin gene family, and most of the new genes are assigned to specific chromosomes.

Materials and methods

Gene isolation and sequencing

LMW-glutenin clones were isolated from the set of complete *Eco*RI digest genomic libraries described by Anderson et al. (1997). The probe used was a 1.1-kb *Xba*I insert of clone pBSM13, which contains the near-complete coding sequence of clone pTdUCD1 (Cassidy and Dvorak 1991) minus the polyA sequence. Plasmid subclones containing LMW-glutenin genes were made using *Eco*RI restriction enzyme. General recombinant techniques were performed according to Maniatis et al. (1982).

Sequence analysis

DNA sequences were determined by two variations of the chain-termination method (Sanger et al. 1977). In the first variation, a set of primers was synthesized complementary to the LMW-glutenin sequence reported in Cassidy and Dvorak (1991) and primers designed from sequences of additional genes as determined in the present study: primer #1F, GCCGTTGTGGCGACA; #2R, GTTTAGCTGCTGCAA; #3R, CCTCATAGCGGGATTG; #4R, AGGATGATGGAGTAG; #5R, CAAAAAGGTACCCTGT; #6R, TCCAACTATATATTACT; #7F, GCCACAACGTCTTGC; #8F, CAAGGTGTCTCCCAA; #9F, AGTGTCAATGTGCCG; #10R, TTGCTCCTGCCATGG; #11F, CATCACAAGCACAAGCATCA; #12R, CTTTATTTGTCACCGCTTC; #AF, ATCCCTGGTTTGGAGAGACCATCGCAGCAA; #BR, TTGCTGCGATGGTCTCTCCAAACCAGGGAT. As the first step in sequencing plasmid DNA, double-stranded plasmid DNA (2 µg) in 18 µl water was denatured with 2 µl 2 *N* NaOH, 1 m*M* EDTA for 5 min at room temperature. The DNA was then neutralized with 3 µl 2 *M* ammonium acetate, pH 4.5, and the DNA precipitated with 100 µl 95% ethanol. After 10 min at −80°C the sample was spun 10 min at 4°C, washed with 70% ethanol at −80°C and spun again for 5 min at 4°C and the DNA pellet air-dried and resuspended in water. Approximately 0.1 µg denatured DNA in 12 µl water was mixed with 1.8 µl 10× annealing buffer (10×: 70 m*M* TRIS pH 7.5, 70 m*M* MgCl, 200 m*M* NaCl, 1 m*M* EDTA) and 50 ng primer in 1–2 µl to a total volume of 15 µl. The mixture was heated at 60°C for 2 min, then cooled to room temperature over 30 min. Unlabeled NTPs and Sequenase enzyme amounts were from Sequenase kits (U.S. Biochemical) according to manufacturer's instructions. In the second sequencing procedure LMW-glutenin containing *Eco*RI fragments were cloned into M13mp8,9. Nested deletions were made according to Dale et al. (1985). Sequence gaps were closed using the primers listed above.

Alignments were made on the Megalign module of the Lasergene (DNAX) software package using the CLUSTAL V method (Higgins and Sharp 1989). Manual changes in amino acid residues were used to increase alignment clarity after a comparison of encoding DNA sequences.

Southern analysis

Wheat genomic DNA was isolated according to Dvorak et al. (1988). Fifteen micrograms of DNA was digested with *Eco*RI restriction enzyme, separated in 0.7% agarose gels, and transferred for Southern analysis as described by Palmer (1986). Labeled probe for Southern analyses was prepared from the double-stranded 1-kb *Xba*I fragment from LMW-glutenin cDNA clone pTdUCD1 described above. To the annealed mixture was added 1 µl 1 *M* DTT, 5–10 µl dATP (111 TBq/mmol ^{32}P), 1.4 µl dNTP mix minus dATP (to a final concentration of 70 µ*M* from a 1 m*M* stock), and 1 µl Klenow fragment (5 units). The total reaction volume of 20–25 µl was incubated at 30°C. After 1.5 h, 1 µl of 1 m*M* dATP was added and the reaction continued for 30 min, then stopped with 10 µl 50 m*M* EDTA and passed through a Sephadex G50 column to separate the incorporated and unincorporated isotope. Labeled probe for hybridizations was heated at 100°C for 10 min before use. Hybridizations and washes were carried out according to Dvorak et al. (1988).

Results and discussion

Library screening and clone sequencing

Several complete *Eco*RI digest genomic lambda libraries (Anderson et al. 1997) were screened for LMW-glutenin sequences. The main initial screen used libraries chosen as the most likely to contain *Eco*RI LMW-glutenin sequences within 2- to 4-kb *Eco*RI fragments. A smaller screen searched for genes within 5- to 7-kb *Eco*RI fragments. Approximately 10^6 lambda clones were screened on 150-mm plates as described (Anderson et al. 1997). Fifty-three strongly hybridizing plaques were selected for analysis. After eliminating duplicate isolations, 19 unique clones remained. All clones contained LMW-glutenin-hybridizing *Eco*RI bands of 2.6–3.7 kb or 6 kb, in agreement with the subgenomic libraries screened. Many of these fell into groupings that could not be distinguished by restriction patterns. Sequencing was initiated on 12 clones. While 6 unique sequences were determined, the other 6 sequences were identical to other clones over the initially sequenced regions (up to one-third of the coding sequences), and sequencing was terminated as uninformative. Figure 1 shows restriction maps of the 6 known unique clones.

LMW-glutenin sequence analysis

All sequenced clones were confirmed as LMW-glutenin sequences: Table 1 lists all reported LMW-glutenin

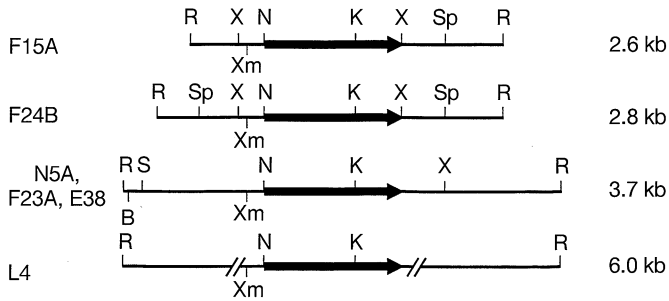


Fig. 1 Restriction maps of 6 new LMW-glutenin gene clones. Restriction maps are shown for the *Eco*RI fragments containing LMW-glutenin genes. Indicated sites on clone L4 were determined by sequencing and not restriction mapping. *B Bam*HI, *K Kpn*I, *N Nco*I, *R Eco*RI, *S Sal*I, *Sp Sph*I, *X Xba*I, *Xm Xma*I. *Arrows* indicate coding sequences. Name of the clone is on the *left*, and size of the *Eco*RI fragment is on the *right*

DNA sequences and GenBank accession numbers. The derived amino acid sequences of the 17 LMW-glutenin DNA sequences in Table 1 were aligned as shown in Fig. 2. The highly variable repetitive domain is not included but is considered separately (Fig. 3). The alignment was used to construct a general model of LMW-glutenin polypeptide structure as shown in Fig. 4. This model has similar features to previous suggested domain arrangements (Colot et al. 1989; Cassidy and Dvorak 1991; Sabelli and Shewry 1991). However, we now have additional sequence information and have concentrated on the domains suggested by the alignments in Fig. 2. An evolutionary comparison among prolamine families will be addressed elsewhere.

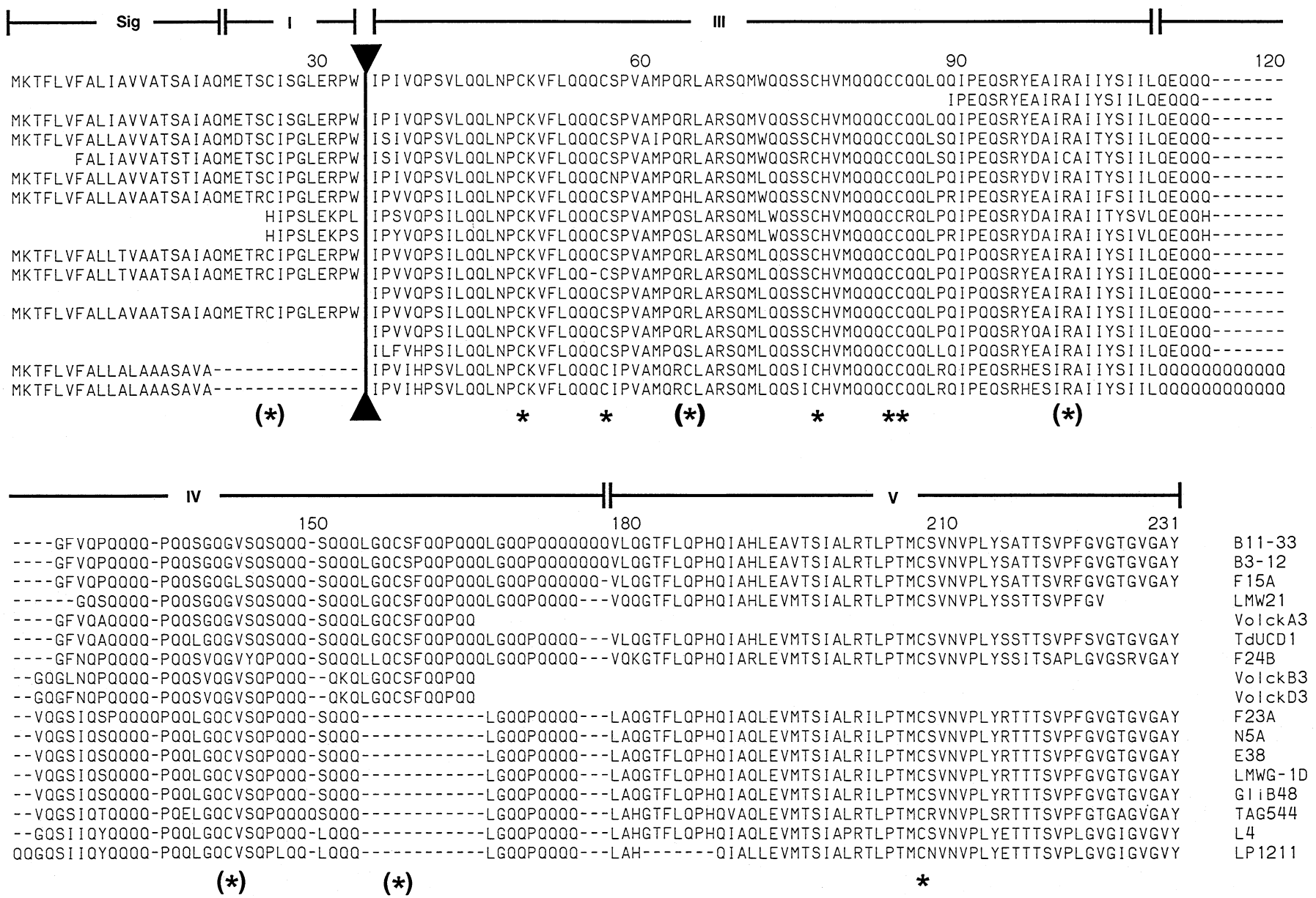
The signal peptide (Sig) is followed in most sequences by a short domain (I) containing the first cysteine residue. The exceptions are LP1211 and L4, which are missing the 13 amino acids making up domain I. Following the repetitive domain (II) is a conserved domain (III) containing five cysteine residues found in all LMW-glutenins. Domain IV is identified by its high glutamine composition; i.e., sequence B11-33 contains 34 glutamines out of 55 residues. Tandem arrays of glutamines are known to be hypervariable (Jennings 1995; Anderson and Greene 1997), and examination of the domain IV alignments in Fig. 2 shows a main mode of variation is within short glutamine runs, likely by DNA slip-mismatching on the glutamine codons during replication. A second mode of variation within domain IV is observed with a 12-residue peptide missing in approximately half the sequences (F23A to LP1211). Domain V is a conserved unique sequence region terminating the LMW-glutenins and containing the sixth and final conserved cysteine.

The LMW-glutenins have been classified as Group B (40–50 k MW) and Group C (30–40 k) (Payne and Corfield 1979). The derived amino acid sequences of the complete coding sequences in Fig. 2 would result in polypeptides ranging from 33.4 kb (TdUCD1) to 39.5 kb (L4), and thus these are mainly Group C LMW-glutenins. The LP1211/L4 pair could be from either class. However, this is assuming standard

Table 1 Wheat LMW-glutenin clones

Clone	Clone type[a]	Cultivar	Accession no.[b]	Reference
Tag544	C[d]	Chinese Spring	–[f]	Bartels and Thompson (1983)
B48	C[d]	Cheyenne	H11335	Okita (1984)
B3-12	C[d]	Cheyenne	H11338	Okita et al. (1985)
B11-33	C	Cheyenne	H11077	Okita et al. (1985)
LP1211	G	Yamhill	X07747	Pitts et al. (1988)
LMWG-1D1	G	Chinese Spring	X13306	Colot et al. (1989)
TdUCD1	C	Mexicali[e]	X51759	Cassidy and Dvorak (1991)
LMW21	P	*T. turgidum*[e]	X62588	D'Ovidio et al. (1992)
VolckA3[c]	P[d]	Chinese Spring	X84959	Van Campenhout et al. (1995)
VolckB3[c]	P[d]	Chinese Spring	X84960	Van Campenhout et al. (1995)
VolckD3[c]	P[d]	Chinese Spring	X84961	Van Campenhout et al. (1995)
F15A	G	Cheyenne	U86028	This paper
F23A	G	Cheyenne	U86027	This paper
E38	G[d]	Cheyenne	U86025	This paper
N5A	G	Cheyenne	U86029	This paper
F24B	G	Cheyenne	U86026	This paper
L4	G	Cheyenne	U86030	This paper

[a] G, Genomic; C, cDNA; P, PCR
[b] GenBank accession numbers
[c] Listed here as sequences and not clones. Sequences were determined by PCR procedures from genomic DNA
[d] Partial coding sequences
[e] Tetraploid wheat; all others are hexaploid wheats
[f] Not submitted to GenBank


Sig
I
III
30
60
90
120
MKTFLVFALIAVVATSAIAQMETSCISGLERPW IPIVQPSVLQQLNPCKVFLQQQCSPVAMPQRLARSQMWQQSSCHVMQQQCCQQLQQIPEQSRYEAIRAIIYSIILQEQQQ-------
IPEQSRYEAIRAIIYSIILQEQQQ-------
MKTFLVFALIAVVATSAIAQMETSCISGLERPW IPIVQPSVLQQLNPCKVFLQQQCSPVAMPQRLARSQMVQQSSCHVMQQQCCQQLQQIPEQSRYEAIRAIIYSIILQEQQQ-------
MKTFLVFALLAVVATSAIAQMDTSCIPGLERPW ISIVQPSVLQQLNPCKVFLQQQCSPVAIPQRLARSQMWQQSSCHVMQQQCCQQLSQIPEQSRYDAIRAITYSIILQEQQQ-------
FALIAVVATSTIAQMETSCIPGLERPW ISIVQPSVLQQLNPCKVFLQQQCSPVAMPQRLARSQMWQQSRCHVMQQQCCQQLSQIPEQSRYDAICAITYSIILQEQQQ-------
MKTFLVFALLAVVATSTIAQMETSCIPGLERPW IPIVQPSVLQQLNPCKVFLQQQCNPVAMPQRLARSQMLQQSSCHVMQQQCCQQLPQIPEQSRYDVIRAITYSIILQEQQQ-------
MKTFLVFALLAVAATSAIAQMETRCIPGLERPW IPVVQPSILQQLNPCKVFLQQQCSPVAMPQHLARSQMWQQSSCNVMQQQCCQQLPRIPEQSRYEAIRAIIFSIILQEQQQ-------
HIPSLEKPL IPSVQPSILQQLNPCKVFLQQQCSPVAMPQSLARSQMLWQSSCHVMQQQCCRQLPQIPEQSRYDAIRAIITYSVLQEQQH-------
HIPSLEKPS IPYVQPSILQQLNPCKVFLQQQCSPVAMPQSLARSQMLWQSSCHVMQQQCCQQLPRIPEQSRYDAIRAIIYSIVLQEQQH-------
MKTFLVFALLTVAATSAIAQMETRCIPGLERPW IPVVQPSILQQLNPCKVFLQQQCSPVAMPQRLARSQMLQQSSCHVMQQQCCQQLPQIPQQSRYEAIRAIIYSIILQEQQQ-------
MKTFLVFALLTVAATSAIAQMETRCIPGLERPW IPVVQPSILQQLNPCKVFLQQ-CSPVAMPQRLARSQMLQQSSCHVMQQQCCQQLPQIPQQSRYEAIRAIIYSIILQEQQQ-------
IPVVQPSILQQLNPCKVFLQQQCSPVAMPQRLARSQMLQQSSCHVMQQQCCQQLPQIPQQSRYEAIRAIIYSIILQEQQQ-------
MKTFLVFALLAVAATSAIAQMETRCIPGLERPW IPVVQPSILQQLNPCKVFLQQQCSPVAMPQRLARSQMLQQSSCHVMQQQCCQQLPQIPQQSRYEAIRAIIYSIILQEQQQ-------
IPVVQPSILQQLNPCKVFLQQQCSPVAMPQRLARSQMLQQSSCHVMQQQCCQQLPQIPQQSRYQAIRAIIYSIILQEQQQ-------
ILFVHPSILQQLNPCKVFLQQQCSPVAMPQSLARSQMLQQSSCHVMQQQCCQQLLQIPQQSRYEAIRAIIYSIILQEQQQ-------
MKTFLVFALLALAAASAVA---------- IPVIHPSVLQQLNPCKVFLQQQCIPVAMQRCLARSQMLQQSICHVMQQQCCQQLRQIPEQSRHESIRAIIYSIILQQQQQQQQQQQ
MKTFLVFALLALAAASAVA---------- IPVIHPSVLQQLNPCKVFLQQQCIPVAMQRCLARSQMLQQSICHVMQQQCCQQLRQIPEQSRHESIRAIIYSIILQQQQQQQQQQQ
(*)
*
*
(*)
*
**
(*)
IV
V
150
180
210
231
----GFVQPQQQQ-PQQSGQGVSQSQQQ-SQQQLGQCSFQQPQQQLGQQPQQQQQQQVLQGTFLQPHQIAHLEAVTSIALRTLPTMCSVNVPLYSATTSVPFGVGTGVGAY B11-33
----GFVQPQQQQ-PQQSGQGVSQSQQQ-SQQQLGQCSPQQPQQQLGQQPQQQQQQQVLQGTFLQPHQIAHLEAVTSIALRTLPTMCSVNVPLYSATTSVPFGVGTGVGAY B3-12
----GFVQPQQQQ-PQQSGQGLSQSQQQ-SQQQLGQCSFQQPQQQLGQQPQQQQQQ-VLQGTFLQPHQIAHLEAVTSIALRTLPTMCSVNVPLYSATTSVRFGVGTGVGAY F15A
------GQSQQQQ-PQQSGQGVSQSQQQ-SQQQLGQCSFQQPQQQLGQQPQQQQ---VQQGTFLQPHQIAHLEVMTSIALRTLPTMCSVNVPLYSSTTSVPFGV LMW21
----GFVQAQQQQ-PQQSGQGVSQSQQQ-SQQQLGQCSFQQPQQ VolckA3
----GFVQAQQQQ-PQQLGQGVSQSQQQ-SQQQLGQCSFQQPQQQLGQQPQQQQ---VLQGTFLQPHQIAHLEVMTSIALRTLPTMCSVNVPLYSSTTSVPFSVGTGVGAY TdUCD1
----GFNQPQQQQ-PQQSVQGVYQPQQQ-SQQQLLQCSFQQPQQQLGQQPQQQQ---VQKGTFLQPHQIARLEVMTSIALRTLPTMCSVNVPLYSSITSAPLGVGSRVGAY F24B
--GQGLNQPQQQQ-PQQSVQGVSQPQQQ--QKQLGQCSFQQPQQ VolckB3
--GQGFNQPQQQQ-PQQSVQGVSQPQQQ--QKQLGQCSFQQPQQ VolckD3
--VQGSIQSPQQQQPQQLGQCVSQPQQQ-SQQQ----------LGQQPQQQQ---LAQGTFLQPHQIAQLEVMTSIALRILPTMCSVNVPLYRTTTSVPFGVGTGVGAY F23A
--VQGSIQSQQQQ-PQQLGQCVSQPQQQ-SQQQ----------LGQQPQQQQ---LAQGTFLQPHQIAQLEVMTSIALRILPTMCSVNVPLYRTTTSVPFGVGTGVGAY N5A
--VQGSIQSQQQQ-PQQLGQCVSQPQQQ-SQQQ----------LGQQPQQQQ---LAQGTFLQPHQIAQLEVMTSIALRILPTMCSVNVPLYRTTTSVPFGVGTGVGAY E38
--VQGSIQSQQQQ-PQQLGQCVSQPQQQ-SQQQ----------LGQQPQQQQ---LAQGTFLQPHQIAQLEVMTSIALRILPTMCSVNVPLYRTTTSVPFGVGTGVGAY LMWG-1D
--VQGSIQSQQQQ-PQQLGQCVSQPQQQ-SQQQ----------LGQQPQQQQ---LAQGTFLQPHQIAQLEVMTSIALRILPTMCSVNVPLYRTTTSVPFGVGTGVGAY GliB48
--VQGSIQTQQQQ-PQELGQCVSQPQQQQSQQQ----------LGQQPQQQQ---LAHGTFLQPHQVAQLEVMTSIALRTLPTMCRVNVPLSRTTTSVPFGTGAGVGAY TAG544
--GQSIIQYQQQQ-PQQLGQCVSQPQQQ-LQQQ----------LGQQPQQQQ---LAHGTFLQPHQIAQLEVMTSIAPRTLPTMCSVNVPLYETTTSVPLGVGIGVGVY L4
QQGQSIIQYQQQQ-PQQLGQCVSQPLQQ-LQQQ----------LGQQPQQQQ---LAH-------QIALLEVMTSIALRTLPTMCNVNVPLYETTTSVPLGVGIGVGVY LP1211
(*)
(*)
*

LP1211	L4	VolckD3	VolckB3
ISQQQ	ISQQQQQ	PLPLQQ	PLPLQQ
APPFSQQQQ	PPFSQQQQ	LIWYHQQQ	ILWYQQQQ
PPFSQQQQ	PQFSQQ	PIQQQPQ	PIQQQPQ
PPFSQQQQ	PPFSQQQQ	PFPQQ	PFPQQ
SPFSQQQQQ	PPFSQQQQQ	PPCSQQQQ	PPCSQQQQ
PPFAQQQQ	PPFAQQQQ	PPLSQQQQ	PPLSQQQQ
PPFSQQ	PPFSQQ	PPFSQQQ	PPFSQQQ
PPISQQQQ	PPFSLQQQ	PPFSQQ	PPFSQQQQ
PPFSQQQQ	PPFSQQQQ	ILPILPQQ	PILPQQ
PQFSQQQQ	PQFSQQQQ	PPFSQQQQPQ	PPFSQQQQQ
PPYSQQQQ	PPYSQQQQ	FSQQQQ	FPQQQQ
PPYSQQQQ	PPYSQQQQ	PFPQQQQ	PLLPQQ
PPFSQQQQ	PPFSQQQQ	PLLLQQ	PPFSQQQ
PPFSQQQQQ	PPFSQQQQQ	PPFSQQ	PPFSQQQQQ
PPFTQQQQQQQQQQ	PPFTQQQQQQQQQQ	RPPFSQQQQQ	PPFSQQQQQ
PFTQQQQ	PFTQQQQ	PVLPQQ	PILLQQ
PPFSQQ	PPFSQQ	PPFSQQQQQQ	PPFSQHQQ
PPISQQQQ	PPISQQQQ	PILPQQ	PVLPQQQ
PPFLQQQR	PPFSQQQQ	PPFSLHQQ	
PPFSRQQQ	PQFSQQQQ	PVLPQQQ	

LMWG-1D1 N5A, F23A , E38	VolckA3	F24B	LMW21
PLPPQQ	PLPPQQ	PLQQKE	PLPPQQ
TFPQQ	TLFPQQQ	TFPQQ	TFPQQ
PLFSQQQQQQ	PFPQQQ	PPSSQQQQ	PPFSQQQQQ
LFPQQ	PPFSQQQ	PFPQQ	PFPQQ
PSFSQQQ	PSFSQQQ	PPFLQQQ	PSFSQQQ
PPFWQQQ	PPFSQQQ	PSFSQQ	PILPQG
PPFSQQQ	PILPE	PLFSQKQQ	PPFPQQTQ
PILPQQ	PPFSLQQQ	PVLPQQ	PVLPQQ
PPFSQQQQ	PVLPQQ	PAFSQQQQ	SPFSQQQQ
LVLPQQ	SPFSQQQ	TVLPQQ	LILPPQQQQQ
PPFSQQQQ	LVLPPQQQQQ	PAFSQQQHQQ	LPQQQ
PVLPPQQ	LPQQQ	LLQQQ	
SPFPQQQQQHQQ	LPQQQ		
LVQQQ			

F15A, B11-33	TdUCD1	TAG544
PLPPQQ	FPQQQ	.
SFSQQ	PFPQQQQ	.
PPFSQQQQQ	PPFSQQQQ	.
PLPQQ	PSFLQQQ	.
PSFSQQQ	PILPQQ	PPFSQQQ
PPFSQQQ	LPFSQQQQ	PVHPQQ
PILSQQ	PVLPQQ	PPFSQQQQ
PPFSQQQQ	SPFSQQQ	PILPQQ
PVLPQQ	LVLPPQQQQYQQ	PPFSQQQQQ
SPFSQQQQ	VLQQQ	PVLPQQQ
LVLPPQQQQQQ		
LVQQQ		

Fig. 3 Repeat motifs comprising the repetitive domains of LMW-glutenins. The amino acid repetitive domains of reported LMW-glutenin sequences are arranged vertically to emphasize the repeat motifs. Only part of the repeat domain of clone TAG544 was reported

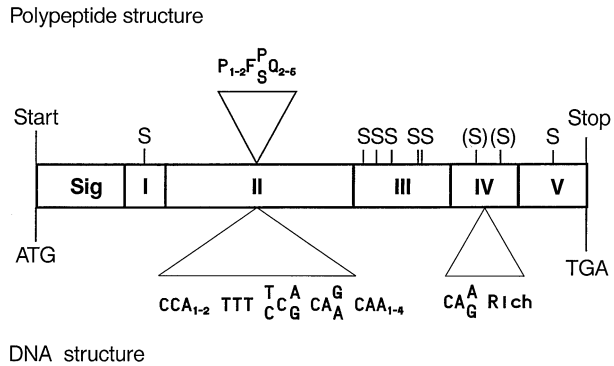


Fig. 4 Model of LMW-glutenin gene and polypeptide structure. The main structural domains of the LMW-sequence are shown in *boxes*. Specific features are indicated *above* (protein) and *below* (DNA). *Sig* Signal peptide, *S* sulfhydryl groups of cysteine residues

polypeptide behavior in SDS-PAGE. It is known that the repetitive region of the HMW-glutenin behaves anomalously in such conditions (D'Ovidio et al. 1997). A similar direct comparison of genes and protein migrations is needed for the other prolamines such as the LMW-glutenins.

There is still only limited flanking DNA sequence available for the LMW-glutenins. However, the proximal non-coding DNA sequences of the LMW-glutenins are different from the other evolutionarily related gliadin families by two characteristics (not shown). The LMW-glutenin coding regions are terminated by double stop codons. A double stop is also characteristic of the HMW-glutenins, the second major polypeptide class making up the disulfide cross-linked gluten matrix, but not of other gliadins. Second, LMW-glutenin sequences have what appears to be a double TATA box. The significance of these doublings is unknown.

All LMW-glutenin sequences thus far reported have no obvious defects to prevent polypeptide synthesis. This lack of obvious defects is interesting since at least some pseudogenes exist in the γ-gliadin (Rafalski 1986) and HMW-glutenin gene families (Forde et al. 1985), and approximately half of the α-gliadin gene family are pseudogenes as judged by internal stop codons (Anderson and Greene 1997). Further study of the other gliadin super-family members will clarify this observation

Cysteine residues

Kasarda (1989) has theorized that the LMW-glutenins have six cysteines involved in intramolecular disulfide bonds and two cysteines free to form intermolecular bonds. This would explain their ability to form both mixed polymers with the HMW-glutenins and polymers strictly composed of LMW-glutenins. Another consequence would be that the addition or subtraction of one cysteine via mutation may create a LMW-glutenin with only a single free cysteine. This mutated LMW-glutenin could then function

Fig. 2 Alignment of LMW-glutenin amino acid sequences. All reported LMW-glutenin sequences were converted to amino acid sequences and aligned. *Brackets* above the alignment indicate sequence domains. *Sig* Signal peptide. The repetitive domain is not included (*vertical bracket* indicates site of repetitive domain). *Asterisks* indicate cysteine residues; *asterisks with parentheses* are cysteine residue positions not found in all sequences

as a chain terminator in polymer formation. We have already noted that LP1211 and L4 (Fig. 2) are missing one conserved cysteine, but they maintain a total of eight by an additional residue not found in other LMW-glutenins. Similarly, VolckB3 and VolckD3 have an extra cysteine in the repeat domain, but are missing the 5′ cysteine (amino acid position 25 in Fig. 2). VolckA3 contains a cysteine not found in other sequences, but since the sequence is incomplete the accurate total cysteine complement is not known.

The single cysteine of domain IV is actually contributed by two different codon positions. These different cysteines likely arose as single base mutations that added or subtracted a cysteine, followed by a second, complementary mutation. It will be interesting to determine if this cysteine is necessary for correct intramolecular crosslinking and/or secondary structure.

Repeat domain structure

The cereal prolamines are known for their regions of tandem variations of short peptide motifs rich in prolamine and glutamine. The wheat LMW-glutenins contain a single such domain (II; not shown in Fig. 2) with 11–20 repeats, as shown in Fig. 3. The repetitive domain begins and ends with several glutamines. Thus, the display of a repeat as PFSQQQ could also be considered as QQQPFS. We prefer the former pattern (Fig. 3), but there is likely no correct choice. Kreis et al. (1985) and Shewry and Tatham (1990) suggested that the repeats within the N-terminus of the LMW-glutenins had a consensus of two heptapeptide motifs: PQQPPPFS and QQQQPVL. Cassidy and Dvorak (1991) suggested a peptide repeat pattern of $PPFSQQQ_n$. Colot et al. (1989) concentrated their analysis on the DNA structure and suggested a repeat motif based on CAA CAA CAA $C^{C}_{A}CA$ TTT $^{C}_{T}CA$ ($QQQ^{P}_{Q}PF^{P}_{S}$). Similarly, our assignment of repeat composition is based on the DNA sequence, which we believe gives a more accurate history of repeat evolution (Anderson and Greene 1997), although selection on the polypeptides may keep the repeat variation within specific bounds.

The analysis of all reported DNA sequences suggests the repeat structure of all LMW-glutenin genes (Table 1) can be reduced to the following pattern of codons: CCA_{1-2} TTT $^{T}_{C}C^{A}_{G}$ $CA^{A}_{G}CAA_{1-4}$. Many of the genes have a specific variation on this basic pattern, presumably due to the tendency of the repeat to homogenize within an individual gene. Most variations from this pattern can be explained by single base changes in glutamine codons (e.g. CAA → CAC, CAA → CCA, CAA → CGA); i.e., at the end of the LP1211 repeats there are two R (arginine) residues that likely originated from CAA → CGA transitions.

Chromosomal origin or RFLPs

Figure 5 shows the restriction pattern of LMW-glutenin sequences within genomic DNA probed with a LMW cDNA. The genomic DNAs are from the cultivars 'Cheyenne' and 'Chinese Spring', and several genetic stocks of 'Chinese Spring' (including the three 'Cheyenne' chromosome 1 substitutions, a nullisomic for chromosome 1A, and three nullisomic-tetrasomic lines of the group 1 chromosomes). The chromosome substitution lines, in which chromosomes 1A, 1B, and 1D of 'Cheyenne' are independently substituted for the corresponding 'Chinese Spring' chromosomes, and nullisomic-tetrasomic lines of 'Chinese Spring', in which one member of group 1 is deleted and replaced by another member of the same group, were used to assign restriction fragment length polymorphisms (RFLPs) to specific chromosomes. The 'Chinese Spring' *Eco*RI pattern is similar to that of Bartels et al. (1986) and Sabelli et al. (1992) except that we resolve two larger bands (nos. 15 and 16). Although most restriction size fragments occur in both cultivars, there are five band differences, at least some of which likely represent RFLP shifts of the same gene sequences. Most bands can be

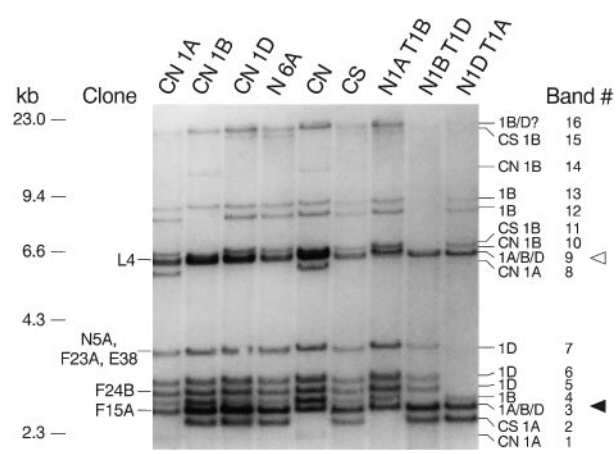


Fig. 5 Southern analysis of LMW-glutenin sequences in the wheat genome. *Eco*RI restriction digests of total wheat DNA were separated on agarose gels, transferred to a membrane, and the membrane probed with a LMW-glutenin coding sequence probe. DNA is from 'Cheyenne' group 1 chromosome substitutions into 'Chinese Spring' (e.g. *CN1A*), a 'Chinese Spring' nullisomic 6A (*N6A*) to rule out cross-hybridization with the 6A α-gliadin genes (Anderson et al. 1997), cv 'Cheyenne' (*CN*) and 'Chinese Spring' (*CS*), and 'Chinese Spring' nullisomic-tetrasomics for the group 1 chromosomes (e.g. *N1AT1B* is missing the 1A chromosome pair and has four copies of the 1B chromosomes). Bands are numbered on the *right* for reference in the text. Individual bands associating only with cv 'Cheyenne' or cv 'Chinese Spring' are indicated *before* the chromosome name. Two possible chromosome assignments are indicated by *slashes*. Bands including fragments from at least two of the three chromosomes are indicated by all *three names separated by slashes* (*open arrowhead* for the 6.0-kb band, *closed arrowhead* for the 2.6-kb band). Clones assigned in the present study and size markers (*Hin*dIII restriction of λ phage DNA) are indicated on the *left*

assigned to a specific chromosome based on either their loss in 'Chinese Spring' aneuploid lines or the 'Cheyenne' chromosome substitution into 'Chinese Spring' lines. Two bands, nos. 3 and 9, seem to contain sequences from at least two of the three group 1 chromosomes. Where the band is assigned in 'Chinese Spring', and without contrary evidence, it is assumed the same assignment holds for 'Cheyenne'. The 6.0-kb band may also contain some signal to the large 6-kb *Eco*RI α-gliadin gene family that cross-hybridizes to LMW-glutenin sequences enough to appear as one to three copies using a LMW-glutenin probe (Anderson et al. 1997). The one anomaly in the analysis is the 8.2-kb B-genome band which is missing in the 'Cheyenne' 1B substitution into 'Chinese Spring' even though the band is present in both Cheyenne and 'Chinese Spring'. The other difference is the relative intensity of some bands; i.e., band no. 4 is relatively more intense in 'Cheyenne' and may indicate increased copies of these sequences.

Gene copy numbers

Using densitometry of blots such as Fig. 5 we estimate 25–30 LMW-glutenin genes, assuming that the larger B-genome bands represent single copies. However, as discussed in Anderson et al. (1997), such numbers will always be low estimates since the more divergent members of a gene family will hybridize less well to any single probe. Therefore, we believe an estimate of 30–40 LMW-glutenin genes for cv 'Cheyenne' and 'Chinese Spring' is a more reasonable estimate, and is similar to that of Sabelli and Shewry (1991). We know that 'Cheyenne' has at least 9 genes since that is the number of different gene sequences thus far reported (Table 1). In addition, it is known from the present report that some of the bands contain several genes. The F23A, N5A, and E38 clones are all contained within 3.7-kb *Eco*RI restriction fragments and have nearly identical sequences (Figs. 1–3). Finally, at least seven *Eco*RI bands from cv 'Cheyenne' in Fig. 5 are not yet represented in isolated clones, including all the B-genome genes.

Relatedness of sequenced genes

An alignment of DNA sequences (coding and proximal non-coding) was used to estimate relatedness of the analyzed sequences, as shown in Fig. 6. The sequences cluster into two branches whose member placement generally agrees with the amino acid sequence variations shown in Fig. 2. The only significant difference is that using the amino acid sequences moves LP1211 and L4 into a separate major branch (not shown). Finally, when tentative chromosome assignments are made, the major branches tend to place genes within a branch on the same chromosome.

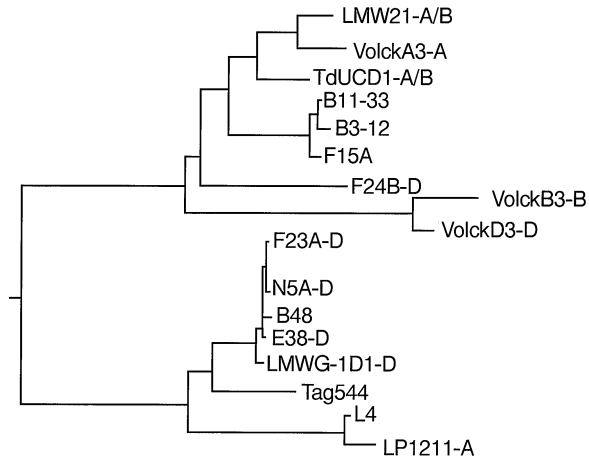


Fig. 6 Phylogenetic tree of LMW-glutenin sequences. All reported LMW-glutenin DNA sequences were compared by Clustal analysis. Sequences used were all available DNA sequences from the beginning of the TATA box to 160 bp downstream of the stop codons, but excluding the repetitive domain similar to as was done in Fig. 2. Chromosome assignments are indicated *following* the clone names. LMW-1D1 was assigned by Colot et al. (1989). LP1211 was assigned by Van Campenhout et al. (1995). TdUCD1 and LMW21 are both from durum wheats and therefore must originate from either the A or B genomes

Further work is needed to search for clones of bands not yet represented by known sequences, particularly the B-genome genes found in the larger *Eco*RI fragments, and a more thorough search for clones in the 6-kb range is necessary.

Acknowledgments Thanks to Timothy Stephens and Elizabeth Ferrero for technical assistance and Cheryl Hsia for help in figure preparation.

References

Anderson OD, Greene FC (1997) The α-gliadin gene family. II. DNA and protein sequence variation, subfamily structure, and origins of pseudogenes. Theor Appl Genet 95:59–65

Anderson OD, Litts JC, Greene FC (1997) The α-gliadin gene family. I. Characterization of ten new wheat α-gliadin genomic clones, evidence for limited sequence conservation of flanking DNA, and Southern analysis of the gene family. Theor Appl Genet 95:50–58

Bartels D, Thompson RD (1983) The chacterization of cDNA clones coding for wheat storage proteins. Nucleic Acids Res 11: 2961–2977

Bartels D, Altosaar I, Harberd NP, Barker RF, Thompson RD (1986) Molecular analysis of γ-gliadin gene families at the complex *Gli-1* locus of bread wheat (*T. aestivum* L.). Theor Appl Genet 72:845–853

Cassidy BG, Dvorak J (1991) Molecular characterization of a low-molecular-weight glutenin cDNA clone from *Triticum durum*. Theor Appl Genet 81:653–660

Colot V, Bartels D, Thompson R, Flavell R (1989) Molecular characterization of an active wheat LMW glutenin gene and its

relation to other wheat and barley prolamin genes. Mol Gen Genet 216 : 81–90
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 18S rDNA. Plasmid 13 : 31–40
Dvorak J, McGuire PE, Cassidy B (1988) Apparent sources of the A genomes of wheats inferred from polymorphism in abundance and restriction fragment length of repeated nucleotide sequences. Genome 30 : 680–689
D'Ovidio R, Tanzarella OA, Porceddu E (1992) Nucleotide sequence of a low-molecular-weight glutenin from *Triticum durum*. Plant Mol Biol 18 : 781–784
D'Ovidio R, Anderson OD, Masci S, Skerritt, J, Porceddu E (1997) Construction of novel wheat high-M_r glutenin subunit gene variability: modification of the repetitive domain and expression in *E. coli*. J Cereal Sci 25 : 1–8
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
Gupta RB, Macritchie F (1994) Allelic variation at glutenin subunit and gliadin loci, *Glu-1*, *Glu-3*, and *Gli-1* of common wheats. II. Biochemical basis of the allelic effects on dough properties. J Cereal Sci 19 : 19–29
Gupta RB, Paul JG, Cornish GB, Palmer GA, Bekes F, Rathjen AJ (1994) Allelic variation at glutenin subunit and gliadin loci, *Glu-1*, *Glu-3*, and *Gli-1*, of common wheats. I. Its additive and interaction effects on dough properties. J Cereal Sci 19 : 9–17
Harberd NP, Bartels D, Thompson RD (1985) Analysis of the gliadin multigene loci in bread wheat using nullisomic-tetrasomic lines. Mol Gen Genet 198 : 234–242
Higgins DG, Sharp PM (1989) Fast and sensitive multiple sequence alignments on a microcomputer. CABIOS 5 : 151–153
Jennings C (1995) How trinucleotide repeats may function. Nature 378 : 127
Kasarda DD (1989) Glutenin structure in relation to wheat quality. In: Pomeranz Y (ed) Wheat is unique. Am Assoc Cereal Chem, St. Paul, Minn., pp 277–302
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
Macritchie F (1992) Physicochemical properties of wheat proteins in relation to functionality. Adv Food Nutr Res 36 : 1–87
Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning. A laboratory manual. Cold Spring Harbor Press, New York
Okita TW (1984) Identification and DNA sequence analysis of a γ-gliadin cDNA plasmid from winter wheat. Plant Mol Biol 3 : 325–332
Okita TW, Cheesbrough V, Reeves CD (1985) Evolution and heterogeneity of the α/β-type and γ-type gliadin DNA sequences. J Biol Chem 260 : 8203–8213
Palmer JD (1986) Isolation and structural analysis of chloroplast DNA. Methods Enzymol 118 : 167–186
Payne PI (1987) Genetics of wheat storage proteins and the effect of allelic variation on breadmaking quality. Annu Rev Plant Physiol 38 : 141–153
Payne PI, Corfield KG (1979) Subunit composition of wheat glutenin proteins isolated by gel filtration in a dissociating medium. Planta 145 : 83–88
Pitts EG, Rafalski JA, Hedgcoth C (1988) Nucleotide sequence and encoded amino acid sequence of a genomic gene region for a low molecular weight glutenin. Nucleic Acids Res 16 : 11376
Pomeranz Y (1988) Composition and functionality of wheat flour components. In: Pomeranz Y (ed) Wheat chemistry and technology, vol 2. Am Assoc Cereal Chem, St Paul, Minn., pp 219–370
Rafalski JA (1986) Structure of wheat gamma-gliadin genes. Gene 43 : 221–229
Sabelli PA, Shewry PR (1991) Characterization and organization of gene families at the *Gli-1* loci of bread and durum wheats by restriction fragment analysis. Theor Appl Genet 83 : 209–216
Sabelli PA, Lafiandra D, Shewry PR (1992) Restriction fragment analysis of 'null' forms at the *Gli-1* loci of bread and durum wheats. Theor Appl Genet 83 : 428–434
Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74 : 5463–5467
Shewry PR, Tatham AS (1990) The prolamin storage proteins of cereal seeds: structure and evolution. Biochem J 267 : 1–12
Shewry PR, Halford NG, Tatham AS (1992) High molecular weight subunits of wheat glutenin. J Cereal Sci 15 : 105–120
Shewry PR, Tatham AS, Barro F, Barcelo P, Lazzeri P (1995) Biotechnology of breadmaking: unraveling and manipulating the multi-protein gluten complex. Biotechnology 13 : 1185–1190
Singh NK, Shepherd KW (1988) Linkage mapping of genes controlling endosperm storage proteins in wheat. 1. Genes on the short arms of group 1 chromosomes. Theor Appl Genet 75 : 628–641
Van Campenhout S, Stappen JV, Sagi L, Volckaert G (1995) Locus-specific primers for LMW glutenin genes on each of the group 1 chromosomes of hexaploid wheat. Theor Appl Genet 91 : 313–319