Nucleotide sequence of a gene from chromosome 1D of wheat encoding a HMW-glutenin subunit

R.D. Thompson, D. Bartels and N.P. Harberd

Plant Breeding Institute, Maris Lane, Trumpington, Cambridge CB2 2LQ, UK

Received 23 August 1985; Accepted 12 September 1985

ABSTRACT

A high molecular weight glutenin gene in hexaploid wheat has been isolated by cloning in bacteriophage lambda and characterized. The gene corresponds to polypeptide 12 encoded by chromosome 1D in the variety "Chinese Spring". The coding sequence predicted contains seven cysteine residues six of which flank a central repetitive region comprising more than 70% of the polypeptide. These findings are related to the role of high molecular weight subunits in the viscoelastic theory of gluten structure.

INTRODUCTION

The developing wheat endosperm is the site of synthesis and deposition of a series of seed storage proteins some of which aggregate into a protein complex known as gluten. In wheat these storage proteins are classified into two groups, the gliadins, which are soluble in aqueous alcohol solutions, and the glutenins, which are alcohol-insoluble. The glutenin fraction is made up of multimeric disulphide-linked aggregates containing two size classes of polypeptides, the high-molecular-weight (HMW) and low-molecular-weight (LMW) subunits. The LMW subunits, \( M_r 34-45 \text{ kD} \), are encoded by genes at the Glu-1 loci on the short arms of the group 1 chromosomes of wheat. The HMW subunits, of fewer types, comprise approximately 10% of the total glutenin aggregate and are encoded by the Glu-1 loci on the long arms of the group 1 chromosomes. They can be distinguished from the gliadins and LMW subunits by their \( M_r \) of 70-90 kD and their relatively higher glycine content (14-21 mole %).

The HMW subunits are believed to be largely responsible for conferring the property of viscoelasticity on dough made with wheat flours. This property distinguishes wheat flour from that made with other cereals. The Ewart model for gluten structure suggests that formation of disulphide bridges between cysteine residues of different subunits results in a network of end-to-end polymers. It is known from amino-acid and DNA sequence data...
that most of the small number of cysteine residues present in each HMW subunit are located close to the subunit termini\textsuperscript{10,11}, although recent evidence suggests that cysteine residues may also be present elsewhere in the molecule\textsuperscript{12}. Estimates of the accessibility of cysteine residues in HMW subunits have indicated that at least some of them are involved in disulphide bond formation\textsuperscript{13}. That disulphide crossbridges are involved in the formation of the elastic glutenin network is implied by the observation that glutenin aggregates are broken down by reduction\textsuperscript{6}.

Comparisons of different wheat varieties have shown that there is substantial intervarietal allelic variation between subunits. This variation is controlled by three complex loci (Glu-Al, Glu-B1, Glu-D1). Each locus encodes up to two different subunits. The pairs of subunits contributed by the Glu-B1 and Glu-D1 loci are subdivided into x and y types on the basis of their electrophoretic mobilities\textsuperscript{14}, NH\textsubscript{2}-terminal amino acid sequences\textsuperscript{15} and chymotrypsin digestion patterns\textsuperscript{16}. Hence Glu-B1 and Glu-D1 are complex loci each controlling an x and y type subunit. Glu-Al controls an x type subunit only in hexaploid wheat\textsuperscript{16}.

Variation between HMW subunits is known to affect the properties of the glutenin aggregates containing them. Presence or absence of particular allelic variants within wheat varieties is correlated with differences in the bread-making quality of flour obtained from them. Comparisons of complete polypeptide sequences of allelic HMW subunits may allow the identification of structural variation associated with differences in bread-making quality. Since these sequences cannot be easily obtained directly from the HMW subunits themselves, cloning and sequencing of wheat DNA fragments containing genes encoding HMW subunits has begun.

Using a characterized cDNA clone complementary to HMW subunit mRNAs\textsuperscript{17}, chromosomal DNA clones of the Glu-1 loci have been isolated from a library of wheat DNA fragments in \*Charon 34. One clone selected from this library has been shown to contain DNA from the Glu-D1 locus and to carry a gene coding for the 1Dy HMW subunit of Chinese Spring.

\section*{MATERIALS AND METHODS}
\underline{Strains and bacteriophages}

\textit{E. coli} strains ED 8800 (rk\textsuperscript{-}mk\textsuperscript{-} SupE SupF lacZ M15 met\textsuperscript{-} RecA56) and DH-1 (gyr A96 RecA1, endA1, thi-1 hsdR17 rk\textsuperscript{-}mk\textsuperscript{+} SupE44) also called WL268, were provided by Dr. N. Murray and Dr. W. Loenen respectively. K803 (SupE met\textsuperscript{-} hsdS\textsuperscript{-} rk\textsuperscript{-}mk\textsuperscript{-}) was provided by Dr. N. Federoff.
The \( \lambda \) derived cloning vector \( \lambda \)-Charon 34\(^{18} \) was provided by Dr. W. Loenen.

**Library construction**

High molecular weight wheat DNA was prepared from the variety Chinese Spring as described previously\(^{17} \). EcoRI partial digestion products of wheat DNA in the 15–20 kb size range (prepared on a sucrose gradient as described by Manlatis et al. (1982)\(^{19} \), a gift of D.C. Baulcombe) were obtained. The \( \lambda \)-Charon 34\(^{18} \) vector was prepared by digestion with BamH1 and EcoRI followed by isopropanol precipitation\(^{20} \). The size-fractionated, wheat DNA EcoRI partial digestion products were ligated into the EcoRI sites of the \( \lambda \)-Charon 34 and \( \lambda \)EMBL4 vectors and the mixture was packaged in vitro\(^{19} \). The packaged mixture was plated on \( E. \) coli strain K803 (RecA\(^{+} \)).

**Clone identification and purification**

After the phage library containing wheat DNA was plated it was screened by plaque hybridization\(^{21} \) using as probe the HMW glutenin cDNA pTag1290\(^{17} \) labelled with \( ^{32} \)P by nick-translation. Hybridizing plaques were picked and purified by several rounds of plaque purification on \( E. \) coli strain WL268 (RecA\(^{-} \)). \( \lambda \) clone DNAs were prepared as described in Manlatis et al.\(^{19} \), following growth of phage on plate lysates. Cloned DNA inserts were subcloned in the pUC9 plasmid vector and grown on \( E. \) coli strain ED 8800 (RecA\(^{-} \)).

**RESULTS**

Molecular cloning of wheat chromosomal DNA fragments containing genes encoding HMW-glutenin subunits

A wheat DNA library was constructed by ligating size-fractionated EcoRI partial digestion products of DNA of the variety Chinese Spring into the EcoRI sites of the cloning vector \( \lambda \)-Charon 34\(^{18} \). The ligations were packaged in vitro and then plated on \( E. \) coli strain K803 (RecA\(^{+} \)). Clones of interest were identified by hybridization to the insert of the HMW glutenin cDNA clone pTag1290\(^{17} \) and then taken through several rounds of plaque purification using \( E. \) coli strain WL268 (RecA\(^{-} \)) as host. Several \( \lambda \)-Charon 34 clones containing sequences hybridizing to the HMW glutenin cDNA were isolated and their DNA was purified for further analysis.

A previous publication\(^{17} \) described the use of the nullisomic-tetrasomic lines of wheat to assign EcoRI fragments hybridizing to HMW glutenin cDNA to the chromosome from which they are derived. Similar analyses have enabled hybridizing fragments obtained by digestion with the restriction enzymes.
Figure 1

Hybridization of the HMW glutenin cDNA (insert of pTag1290) to:
(a) Chinese Spring wheat DNA digested with EcoRI, (b) λC11 DNA digested with EcoRI, (c) Chinese Spring DNA digested with BamHI, (d) λC11 DNA digested with BamHI, (e) Chinese Spring digested with HindIII, (f) λC11 DNA digested with HindIII. The chromosomal locations of the wheat DNA fragments are indicated. Note that HindIII digests of Chinese Spring DNA also contain a small (≈ 0.8 kb) fragment which hybridizes to the HMW-glutenin cDNA (Harberd et al., in prep.). This fragment, derived from chromosome 1A, is not visible in this experiment, because it has migrated too far. The sizes of the hybridizing fragments derived from the λC11 clones are given in kilobase pairs.

endonucleases BamHI and HindIII to be assigned to chromosomes (Harberd et al., in prep.). These data taken together allowed the integrity of putative HMW glutenin genomic DNA clones to be confirmed (Fig. 1). The wheat DNA restriction fragments hybridizing to the HMW glutenin cDNA were compared with the fragments resulting from digestion of the cloned DNA with the same restriction endonuclease. In the example in Figure 1 the cloned DNA digests contain single hybridizing fragments which co-migrate with one of the hybridizing fragments derived from the wheat chromosome 1D. This is observed because there are BamHI and HindIII sites outside the region of hybridization to the cDNA and internal to the EcoRI sites at each end of the cloned fragment (see Fig. 2). Hence λC11 is a clone of a 6.0 kb EcoRI

6836
Figure 2

Map of the 6.0 kb EcoRI fragment in λC11. (a) The BamHI fragment which hybridized to a HMW glutenin cDNA probe (pTag1290)\textsuperscript{17}. (b) The sequenced portion. R = EcoRI, H = HindIII, A = AccI, S = SphI, B = BamHI, X = XbaI.

Gene localization and sequencing

The structure of the gene contained in the λC11 clone was determined by restriction endonuclease site mapping and DNA sequencing. The region of the clone containing the HMW glutenin gene was localized on a restriction map of the entire cloned fragment (Fig. 2) by hybridization of the \textsuperscript{32}P-labelled HMW glutenin cDNA sequence\textsuperscript{17} to digests of the cloned DNA. Hybridization was observed to one BamHI fragment of size 2.4 kb. This BamHI fragment, which was also the only BamHI fragment to hybridize to a \textsuperscript{32}P-end-labelled endosperm polyA\textsuperscript{+} RNA probe (prepared as previously described\textsuperscript{22}), was taken and subcloned for sequencing. The DNA in this fragment was sequenced by generation of a set of Bal-31 deletions from both BamHI termini, and subcloning and sequencing of these deletions in M13 vectors\textsuperscript{23,24}. Additional sequence data were obtained by dideoxy sequencing of HindIII and PstI restriction fragments, and from clones generated by sonication\textsuperscript{25}. All clones were propagated on RecA\textsuperscript{-} E. coli hosts and under these conditions, no evidence for sequence instability was apparent. The open reading frame was found to extend beyond one of the BamHI sites and therefore the adjacent BamHI-HindIII fragment was also sequenced (see Fig. 2). The sequence was analysed using the Staden computer programmes\textsuperscript{26,27}.

Gene organisation

The mature amino-terminal protein sequences of several purified HMW-glutenin subunits have recently been determined\textsuperscript{15}. These sequences, together with those from two cDNAs covering COOH termini\textsuperscript{11} enabled us to identify a unique open reading frame encoding a HMW glutenin polypeptide (Fig. 3). The coding sequence is uninterrupted by intron sequences and predicts a mature polypeptide of 68,617 M\textsubscript{r}. The amino acid composition of the predicted polypeptide (eg. 35.5% Glutamine + Glutamic acid acid, 18% Glycine and 11% Proline\textsuperscript{7,10}) matches that previously determined for HMW.
Figure 3

Sequence of a 3095 base pair region of the XC11 insert. The DNA sequence is shown with a one-letter code translation of the reading frame utilized. Possible control elements are overlined. The site of polyadenylation of a related cDNA clone is indicated (+). Putative signal sequences for poly(A) addition are underlined. Arrowed bars indicate the positions of hexamer and nonomer repeat units within the coding sequence.
Figure 4

DIAGON\textsuperscript{27} homology matrix of Glu-D1 sequenced portion against itself. A dot is printed when 11 bases match in a 'window' of 15. Sequence domains A + D within the coding region are indicated. Segment A is the leader sequence, segment B is the non-repetitive amino-terminal portion, segment C is the repetitive region and segment D is the non-repetitive carboxyl terminal region.

Glutenin subunits (32.6\% Glutamine + Glutamic acid, 14.85\% Glycine, 12.82\% Proline). The mature $\text{NH}_2$-terminal sequence is preceded by a 21-residue leader sequence of characteristic amino acid composition\textsuperscript{28}.

The coding sequence of the mature polypeptide is similar to that of other prolamin storage proteins in that it can be divided into a number of distinct segments on the basis of amino acid composition. The HMW subunit has a tripartite structure, consisting of a non-repetitive amino terminal region, an extensive repetitive central region, and a non-repetitive...
Comparison of amino-terminal protein sequences obtained for isolated polypeptides with the mature amino-terminal protein sequence predicted from the gene characterized here.

The non-repetitive regions show homology to the corresponding non-repetitive regions of gliadin genes, and also some homology to sequences in certain globulin storage proteins of dicotyledonous plants. These homologies have been described in detail by Kreis et al. The amino and carboxyl terminal regions contain most of the low abundance residues contained in the HMW subunit, including all but one of the cysteine residues (5 in the amino terminal region, 1 in the carboxyl terminal region).

The central region is composed of highly repetitive sequence, and is the only region showing extensive repetition in the DIAGON homology plot (segment c in Fig. 4). The sequence consists almost entirely of two repetitive units, one a hexamer related to the amino acid sequence PGQQQQ, the other a nonomer related to the sequence GYYPTSLQQ. These units are interspersed such that single copies of the nonomer repeat separate segments containing several tandemly arranged copies of the hexamer repeat. Occasional length variants (e.g. PGQQ, residues 353-356 of the mature polypeptide) of the basic hexamer and nonomer repetitive units are found in the sequence. Repeat units of both types display considerable variation at both the amino acid and DNA sequence levels. One nonomer repeat variant contains a cysteine residue (residue 525 of the mature polypeptide), probably a substitution for a tyrosine residue via a TAC+TGC codon change.

Comparison of the mature amino terminal sequences of a number of I Dx and I Dy HMW subunits with that of the polypeptide predicted by the gene sequence indicates that this gene encodes the I Dy subunit of Chinese Spring known as subunit 12 (Fig. 5). The sequence is different in two respects to the I D2 sequence; the Arginine for Glutamic Acid at residue 6 and a three codon deletion corresponding to residues 17-19.
Comparison of sequence homology at 5' end of Glu-D1 sequence and α-gliadin gene sequence.

```
GLU C C T T G C T T A T C C A G C T T
α-SLI C C A T G C T T A T C T A G T T T
```

In Figure 3 420 bp of 5' non-coding sequence are also shown. This sequence contains a TATA box at -83 (numbered relative to the A of the starting ATG codon, Fig. 3). When the 5' untranslated region is compared, using the SEQH programme to a 5'-untranslated region of an α-gliadin gene a significant additional region of sequence homology is found (Fig. 6).

The 3'-untranslated sequence shows a high degree of homology to previously reported HMW glutenin cDNA sequences and contains a putative polyadenylation signal at the same position as in those sequences, as well as two additional signals (Fig. 3) which may also be utilised, for example for longer mRNA species similar to that from which pC256 was derived.

**DISCUSSION**

(1) **Clone stability**

This paper describes the isolation, identification and sequencing of a cloned HMW glutenin gene from the Glu-D1 locus of Chinese Spring wheat. This clone was isolated from a wheat DNA library constructed using the λ-Charon 34 vector. Previous attempts to isolate HMW-glutenin gene clones from wheat libraries constructed using the λ EMBL 4 vector and grown on a RecA+ host strain had been unsuccessful due to high levels of clone instability (data not shown). Use of the λ-Charon vector series for cloning wheat DNA has been described by others. The λ-Charon 34-WL268 vector-host system is phenotypically RecA−RecBC− and the resultant reduction in recombination presumably confers stability on otherwise unstable cloned DNA sequences. However, the isolation of stable clones containing 13 different non-storage protein genes of wheat from libraries constructed in λEMBL series vectors and grown on K803 (RecA+) hosts (Baulcombe unpub.) suggests that different DNA sequences from the wheat genome are

---

6842
differentially susceptible to instability during cloning.

The results presented in Fig. 1 demonstrate that no major rearrangements of the 6.0 kb EcoRI fragment occurred during cloning in the \( \lambda \)-charon 34 vector. It is of course not possible to rule out the occurrence of small scale rearrangements causing slight and hence undetectable alterations in DNA fragment mobility.

(2) **Gene structure**

The HMW glutenin gene cloned in \( \lambda \)C11 possesses many of the features commonly found in DNA flanking the coding regions of eukaryotic genes including a TATA box, a translation start sequence and a polyadenylation signal which corresponds well to the canonical plant sequence\(^3\). There is however no sequence corresponding exactly to the CAAT or AGGA\(^3\) boxes found in, for example, the \( \alpha \)-gliadin genes\(^3\),\(^2\). The coding sequence of this gene, like that of other prolamin genes reported, is not interrupted by intervening sequences\(^3\). The reading frame identified in this gene is complete and is not broken by a nonsense codon such as is present in the predicted reading frame of the non-translated HMW-glutenin pseudogene described in the adjacent paper\(^4\). Also, there is good correspondence between the N\(_2\)-terminal amino acid sequence determined from a purified 1Dy subunit\(^1\) and the mature N\(_2\)-terminal sequence of the polypeptide predicted by the gene sequence (Fig. 6). Two HMW glutenin mRNA species, one of 2700 bases and a second of 2200 bases, are specified by chromosome 1D\(^1\). The gene sequenced here is 2153 base pairs from the TATAAA box to the first possible polyadenylation site and therefore must specify the smaller RNA species.

Evidence from comparison of restriction fragments associated with the Glu-Dlx and Glu-Dly alleles from several wheat varieties (Harberd, in prep.) suggests that these genes are present in single copies in the wheat genome. These considerations indicate that the gene contained in the \( \lambda \)C11 clone is active at the level of transcription and translation and is the gene from the Glu-D1 locus encoding the 1Dy HMW-glutenin subunit of Chinese Spring known as subunit 12\(^8\),\(^1\).

It is known that gliadins and glutenins are synthesised coordinately during endosperm development (Bartels, in prep.) and expression of the genes encoding them is therefore likely to be subject to a common regulatory mechanism. Summer-Smith et al.\(^3\) have noted the presence of a 17 bp direct repeat sequence present in two copies at -579 and -387 in the 5' untranslated region of three \( \alpha \)-gliadin genes, and have suggested that this
sequence may be recognised by a developmentally regulated effector of gene expression. This sequence is part of a 56 bp sequence showing about 80% homology within the 5' untranslated region of the α-gliadin gene. A sequence displaying strong homology to this repeat exists at -210 in the HMW-glutenin gene sequence described above (Fig. 3, 6) and in the sequence presented by Forde et al. These homologous regions may indeed therefore be involved in the regulation of storage protein gene expression. Equivalent sequences, sometimes termed enhancers, have been identified in the 5' non-coding regions of members of other multigene families. In certain cases they have been shown to confer specificity of expression on coding sequences.

(3) HMW-glutenin polypeptides: structure and function

Wheat storage proteins are synthesised on membrane bound polysomes and the α-gliadins have been shown to possess leader sequences. The HMW-glutenin signal sequence lacks obvious homology with those reported for the gliadins but is of characteristic amino acid composition, possessing a lysine residue at position 2 in the sequence, followed by a stretch of hydrophobic residues and with alanine preceding the mature NH$_2$-terminal sequence.

The mature HMW-subunit contains three distinct regions, a non-repetitive NH$_2$-terminal region, a central region consisting of repetitive sequence and a non-repetitive -COOH terminal region. The sequences of the NH$_2$- and -COOH terminal non-repetitive regions display a high propensity towards formation of α-helix according to the rule for prediction of secondary structure from primary amino acid sequence. The cysteine residues contained in these regions are likely to be available for disulphide bridge formation between or within the HMW subunits and other aggregating polypeptides. The central region, when tested for hydrophobicity using the parameters described by Chou and Fasman is more hydrophobic than the non-repetitive NH$_2$- and -COOH terminal regions. Hydrophobic stretches are punctuated by several small hydrophilic pockets. The secondary structure of this region is a reflection of its unusual amino acid content. Computer predictions made from previous sequences of the region have suggested that tetraplets, which occur in both six-mer and nine-mer repeats, are involved in the formation of β-turn structures and that these multiply stacked β-turns give the molecule elastic properties.

The occurrence of a cysteine residue within the repetitive central region of this subunit is of interest. There are no cysteines in the
repetitive regions of the partial sequences reported previously\(^1\), or in the central region of the gene sequence reported by Forde et al.\(^1\). The molecular environment of this cysteine is hydrophobic and this may reduce the probability of formation of disulphide bridges or affect the propensity for disulphide exchange. Whether this extra cysteine residue is involved in the viscoelasticity differences between proteins from different alleles can only be established by more comparisons.

The HMW subunits clearly possess many of the features predicted for them in the models which account for gluten viscoelasticity by inter-molecular end-to-end disulphide cross-linking connecting polypeptides with extensive elastic central regions in an elastic network\(^9\). The glutenin aggregate is a complex mixture of polypeptides and each of these presumably contribute in one way or another to gluten properties. The isolation of the genes for all the glutenin components including the HMW-subunit gene reported here, and LMW glutenin will enable a detailed assessment of the nature of the molecular interactions in gluten to be made. In particular, it will soon be possible to compare the sequences of allelic polypeptides known to differ in their effects on gluten properties and to characterise the structures associated with these differences.

ACKNOWLEDGEMENTS

We thank Dr. A. Tatham, Dr. N. Murray and Dr. E. Coen for advice and assistance during the course of this work and are grateful to Dr. R.B. Flavell for guidance. NPH is supported by a training fellowship from the UK Medical Research Council and DB is supported by EEC Contract GBI-4-027-UK. We are grateful to the authors of the accompanying paper for discussions and the opportunity to see a draft of their paper before publication.

REFERENCES