In vivo footprinting of a low molecular weight glutenin gene (LMWG-1D1) in wheat endosperm

**Michael C.U.Hammond-Kosack,
Michael J.Holdsworth and Michael W.Bevan**[1]

Cambridge Laboratory, John Innes Centre for Plant Science Research, Colney Lane, Norwich, NR4 7UJ, UK

[1]Corresponding author

Communicated by D.H.Northcote

The quality of the wheat grain is determined by the quantity and composition of storage proteins (prolamins) which are synthesized exclusively in endosperm tissue. We are investigating the mechanisms underlying the regulation of expression of a prolamin gene, the low molecular weight glutenin gene LMWG-1D1. The LMWG-1D1 promoter contains the endosperm box, a sequence motif highly conserved in the promoter region of a large number of storage protein genes, which is thought to confer endosperm-specific expression of prolamin genes. Here we show by in vivo DMS footprinting of wheat endosperm tissue that the endosperm box becomes occupied by putative *trans*-acting factors during grain ripening. During early stages of development the endosperm motif within the 5' half of the endosperm box becomes occupied first, followed by binding of a second activity to a GCN4/jun-like motif in the 3' half just prior to the stage of maximum gene expression. Occupancy of the endosperm box is highly tissue-specific: no protection was observed in husk and leaf tissues. Several binding activities were identified in vitro from nuclear protein extracts of wheat endosperm which bind specifically to the endosperm and GCN4/jun motifs identified by in vivo footprinting.

Key words: endosperm box/glutenins/*in vivo* footprinting/tissue-specific expression/wheat storage proteins

Introduction

The endosperm of cereal grain serves as the storage organ of the seed where carbohydrates and proteins are laid down in order to feed the embryo during germination. The storage proteins, collectively called prolamins in wheat, are synthesized at high levels exclusively in the endosperm. Genes encoding storage proteins have been isolated from several cereal species (Colot *et al.*, 1989; Colot, 1990). Synthesis of storage proteins seems to be controlled primarily at the transcriptional level (Bartels and Thompson, 1986; Sørensen *et al.*, 1989; Soave and Salamini, 1984). The prolamins fall into two major groups, the gliadins and glutenins. Whereas the gliadins consist of a complex mixture of monomeric polypeptides, the glutenins form large proteinaceous aggregates, held together by disulfide bonds. These aggregates consist of two monomeric subunits, classified according to size as high molecular weight

glutenins (HMWG, 60–100 kDa) and low molecular weight glutenins (LMWG, 40–50 kDa). Hexaploid wheat contains genes for three to five HMWG subunits, 7–16 LMWG subunits and 15–20 gliadins (Colot, 1990).

The expression of prolamin genes occurs only in the developing endosperm. The *cis*-acting elements responsible for this tissue-specificity have been roughly defined in transgenic tobacco plants for both low and high molecular weight glutenin genes (Colot *et al.*, 1987; Thomas and Flavell, 1990). In particular, for one LMWG gene, LMWG-1D1, it could be shown that two elements, residing between positions (−326 to −160) and (−938 to −371) act in concert to confer expression specifically in the endosperm. Sequence comparison revealed a common motif in both DNA fragments which is not unique to LMWG genes, but can be found in a large number of storage protein genes whose expression is endosperm-specific in a variety of cereal species (Hartings *et al.*, 1990; Kridl *et al.*, 1984; Forde *et al.*, 1985). This consensus sequence has been termed the 'endosperm box' [TGTAAAGTNAATNN-G(A/G)TGAGTCAT]. It contains two conserved motifs (indicated in bold letters) with a high degree of similarity to known *cis*-acting sequences. The 3' conserved motif is very similar to the GCN4-binding motif in yeast (Hill *et al.*, 1986) and the binding site for the mammalian transcription factors jun (Ransone and Verma, 1990) and AP1 (Piette *et al.*, 1988). This motif will subsequently be called the GCN4/jun motif. Interestingly, it also has a strong resemblance to the binding site for the maize regulatory factor opaque-2 (Lohmer *et al.*, 1991). The 5' conserved motif, TGTAAAGT, which resembles the SV40 enhancer core motif TGTGGAAAGT (Mitchell *et al.*, 1987), will be called the endosperm motif. It is plausible to expect at least two protein factors to bind within the wheat endosperm box—one binding to the GCN4/jun-like motif and the other one recognizing the endosperm motif. It is most probably a combination of both factors binding to the endosperm box that contributes to high level tissue-specific gene expression. The *opaque-2* mutation of maize differentially reduces α-zein gene transcription in the endosperm; the transcription of 22 kDa α-zeins is almost totally blocked (90%), whereas the transcription of 19 kDa α-zeins is reduced by 45–70% (Kodrzycki *et al.*, 1989). The *opaque-2* gene encodes a transcriptional activator of the leucine-zipper class (Hartings *et al.*, 1989) which binds to the GCN4/jun motif of the zein gene and wheat LMWG-1D1 endosperm box and activates transcription (Holdsworth *et al.*, unpublished; Schmidt *et al.*, 1992). The different effects of *opaque-2* on α-zein transcription rates are surprising since both genes share the same putative *cis*-acting elements (Thompson and Larkins, 1989). This suggests that while there are common putative *cis*-elements and *trans*-acting factors shared by the storage protein genes, there must also be separate mechanisms governing expression of the different classes of storage

protein genes. A similar phenomenon has been observed in the barley *lys3a* mutant, where expression of B- and C-hordein genes is differentially reduced compared with D-hordeins (Kreis *et al.*, 1984; Sørensen *et al.*, 1989), but the resulting low level of transcription is still endosperm-specific (Forde *et al.*, 1981). Although the nature of the *lys3a* gene product has not yet been determined, the similarity of effects between the *lys3a* and *opaque-2* mutations would tend to suggest a similar mechanism, implying that *lys3a* should code for a transcriptional activator with a similar function to opaque-2. The methylation state of B-hordein genes which are down-regulated in *lys3a* mutants has recently been compared for wild type and *lys3a* mutant barley endosperm (Sørensen, 1992). Expression of the B-hordein genes seems to be accompanied by hypomethylation whereas the corresponding CpGs in the same gene in *lys3a* mutant barley are strongly methylated, as they are in leaf tissue where the gene is completely silent. It appears that hypomethylation is correlated with the presence of a wild type allele of the regulatory gene *lys3a*. This suggests that the *lys3a* product may be involved in poising chromatin for gene expression by regulating its methylation state although further characterization of the function of lys3a awaits its molecular cloning.

Despite the common presence of the endosperm box in the 5' region of seed storage protein genes, no direct evidence for the importance of the endosperm box in endosperm-specific gene expression has been reported previously. Here we show, by *in vivo* DMS footprinting of the promoter region of LMWG-1D1 in developing wheat endosperm, that the endosperm box becomes occupied by putative *trans*-acting factors during grain ripening. Ligation-mediated PCR methodology (LMPCR, Mueller and Wold, 1989) was used for the analyses. Our footprinting data suggest that two types of proteins bind sequentially in the endosperm box, whereby the endosperm motif becomes occupied first at 13 days after anthesis (DAA), when the rate of transcription is moderate (Colot *et al.*, 1987), while the GCN4/jun motif becomes occupied at the highest rate of transcription around 18 DAA. We further demonstrate that occupancy of the two sites within the endosperm box is tissue-specific: it was only observed in endosperm tissue but not in leaf or husk tissues. The data presented here are in agreement with a model where a tissue-specific factor binds to the endosperm motif just prior to moderate gene expression followed by binding of a second factor at the GCN4/jun- motif, which confers high levels of expression. This model is further supported by the identification of two DNA binding activities in nuclear protein extracts from developing wheat endosperm that bind specifically to the sites identified by *in vivo* footprinting.

Results

Genomic sequencing of LMWG-1D1

It was necessary to show that the 5' region of LMWG-1D1 could be amplified specifically from total wheat genomic DNA, in particular because there are a large number of homoeologous genes for low molecular weight glutenins in hexaploid wheat (Colot, 1990). This was tested using PCR and various primer sets consisting of Glu3/Glu5/Glu6/Glu9/Glu10 and Glu1 or Glu7/Glu 8 and Glu5. The position of each primer is indicated in Figure 1. Single bands of the
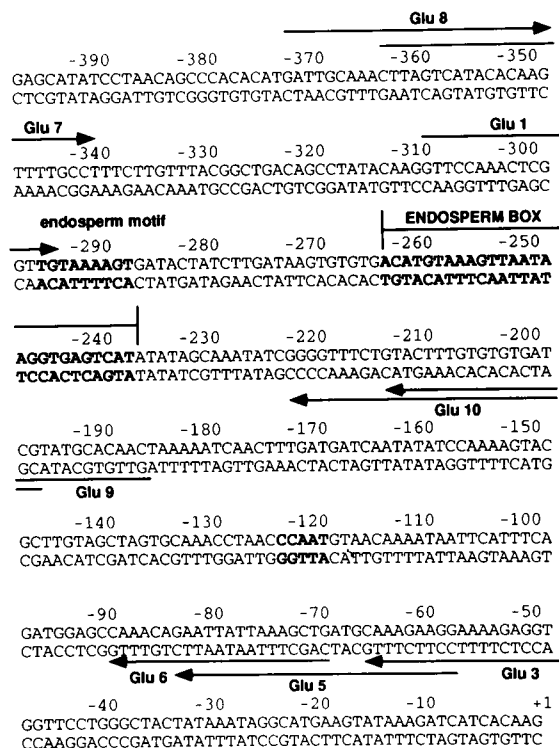


```
                                                          Glu 8
        -390       -380       -370       -360       -350
GAGCATATCCTAACAGCCCACACATGATTGCAAACTTAGTCATACACAAG
CTCGTATAGGATTGTCGGGTGTGTACTAACGTTTGAATCAGTATGTGTTC

  Glu 7                                               Glu 1
        -340       -330       -320       -310       -300
TTTTGCCTTTCTTGTTTACGGCTGACAGCCTATACAAGGTTCCAAACTCG
AAAACGGAAAGAACAAATGCCGACTGTCGGATATGTTCCAAGGTTTGAGC

  endosperm motif                      |  ENDOSPERM BOX
        -290       -280       -270       |-260       -250
GTTGTAAAAGTGATACTATCTTGATAAGTGTGTGACATGTAAAGTTAATA
CAACATTTTCACTATGATAGAACTATTCACACACTGTACATTTCAATTAT

        -240       |-230       -220       -210       -200
AGGTGAGTCATATATAGCAAATATCGGGGTTTCTGTACTTTGTGTGTGAT
TCCACTCAGTATATATCGTTTATAGCCCCAAAGACATGAAACACACACTA

                                            Glu 10
        -190       -180       -170       -160       -150
CGTATGCACAACTAAAAATCAACTTTGATGATCAATATATCCAAAAGTAC
GCATACGTGTTGATTTTTAGTTGAAACTACTAGTTATATAGGTTTTCATG

  Glu 9
        -140       -130       -120       -110       -100
GCTTGTAGCTAGTGCAAACCTAACCCAATGTAACAAAATAATTCATTTCA
CGAACATCGATCACGTTTGGATTGGGTTACATTGTTTTATTAAGTAAAGT

        -90        -80        -70        -60        -50
GATGGAGCCAAACAGAATTATTAAAGCTGATGCAAAGAAGGAAAAGAGGT
CTACCTCGGTTTGTCTTAATAATTTCGACTACGTTTCTTCCTTTTCTCCA

  Glu 6                   Glu 5                       Glu 3
        -40        -30        -20        -10         +1
GGTTCCTGGGCTACTATAAATAGGCATGAAGTATAAAGATCATCACAAG
CCAAGGACCCGATGATATTTATCCGTACTTCATATTTCTAGTAGTGTTC
```

Fig. 1. Sequence of the proximal 5' region of LMWG-1D1. The transcriptional start site is indicated as (+1). The putative *cis*-acting elements (in bold) and the location of all primers used for PCR and LMPCR (see text) are shown.

expected sizes were observed and some of these were sequenced directly using a cycle sequencing methodology as described elsewhere (Hammond-Kosack *et al.*, 1992). The observed sequences showed no redundancy, indicating that the primers used were specific for LMWG-1D1 (data not shown). Initial attempts with shorter primers gave a large number of distinct bands in agreement with the abundance of homoeologous low molecular weight glutenin genes.

To optimize reaction conditions, LMPCR was carried out initially with plasmid DNA (p326VC) containing 326 bp of 5' flanking DNA of LMWG-1D1 which had been linearized with *Hind*III (called p326H). The amount of this segment of DNA within the total wheat DNA corresponds to 0.24 pg for 10 μg chromosomal DNA, corresponding to 2.5 pg linearized plasmid DNA. Three different amounts of plasmid DNA, 5 ng, 5 pg and 50 fg, were used and the result of LMPCR using primer set Glu3/Glu5/Glu6 and the linker (Ib + II) is shown in Figure 2. Linker primer Ib was designed to give a similar T_m to the other primers used in these studies and thus is different from the one used originally (Mueller and Wold, 1989). To mimic the conditions for LMPCR of the endogenous wheat gene, 15 μg of salmon sperm DNA were added to each sample to increase the complexity of the sample before carrying out the DMS modification and LMPCR. Using only 50 fg of gene-specific DNA, the G-specific pattern can be clearly observed in an 8 h autoradiographic exposure. Figure 2 also shows the typical gap between the end-labelled primer GLU6 and the start of the sequence pattern. This gap is 25 nucleotides long and this length corresponds exactly to the length of the longer linker primer Ib. This gap would be expected only if the universal linker had been successfully ligated to the blunt
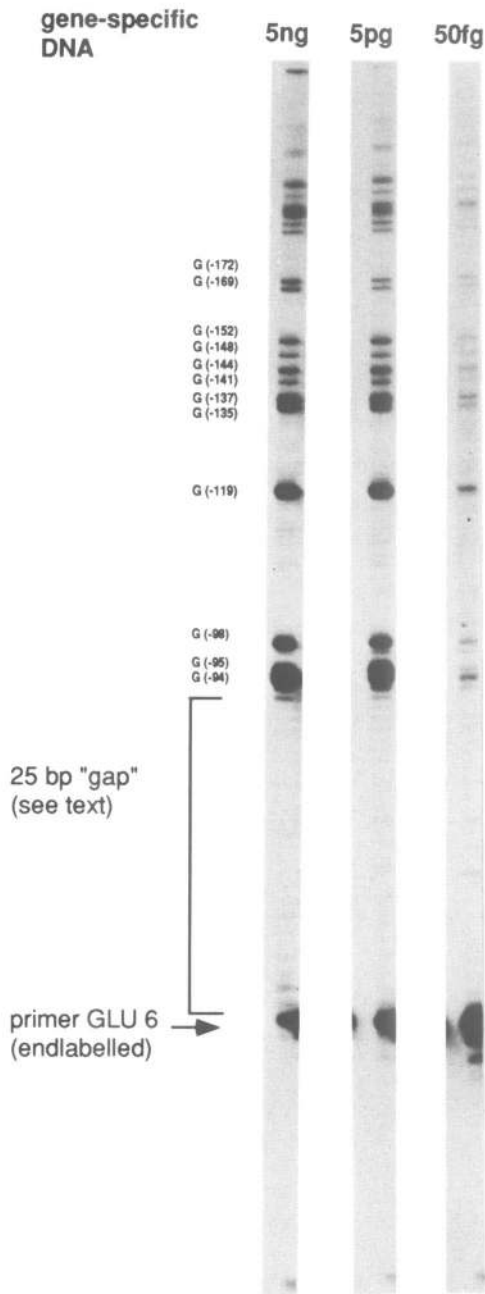
gene-specific DNA 5ng 5pg 50fg

G (-172)
G (-169)

G (-152)
G (-148)
G (-144)
G (-141)
G (-137)
G (-135)

G (-119)

G (-98)
G (-95)
G (-94)

25 bp "gap"
(see text)

primer GLU 6
(endlabelled)

Fig. 2. LMPCR sequencing of DMS-treated p326H plasmid DNA. The unlabelled linearized plasmid DNA was reacted with DMS and subjected to LMPCR using primer set Glu3/Glu5/Glu6 thereby visualizing the top strand. The end-labelled third gene specific primer, Glu6, can be seen at the bottom of the sequencing gel. The 25 bp gap shows that the 25 nucleotide linker (Ib + II) had ligated onto genomic DNA. Note that although faint, all G-specific bands can still be observed when sequencing only 50 fg of p326H.

ends of gene-specific cleavage products. The gap therefore shows that the ligation step works efficiently with very small amounts of gene-specific DNA. The sensitivity was extremely high, as the gene-specific sequence could still be identified in the sample with a 100-fold less LMWG-1D1 DNA than would be present in 10 μg chromosomal DNA (see above). Therefore, it was envisaged that under ideal conditions, even 100 ng total chromosomal DNA should be sufficient for genomic sequencing and *in vivo* footprinting.

DNA: plasmid chromosomal
1 2 3 4   1 2 3 4

G (-172)
G (-169)

G (-152)
G (-148)
G (-144)
G (-141)
G (-137)
G (-135)

G (-119)

G (-98)
G (-95)
G (-94)

25 bp "gap"
(see text)

primer GLU 6
(endlabelled)
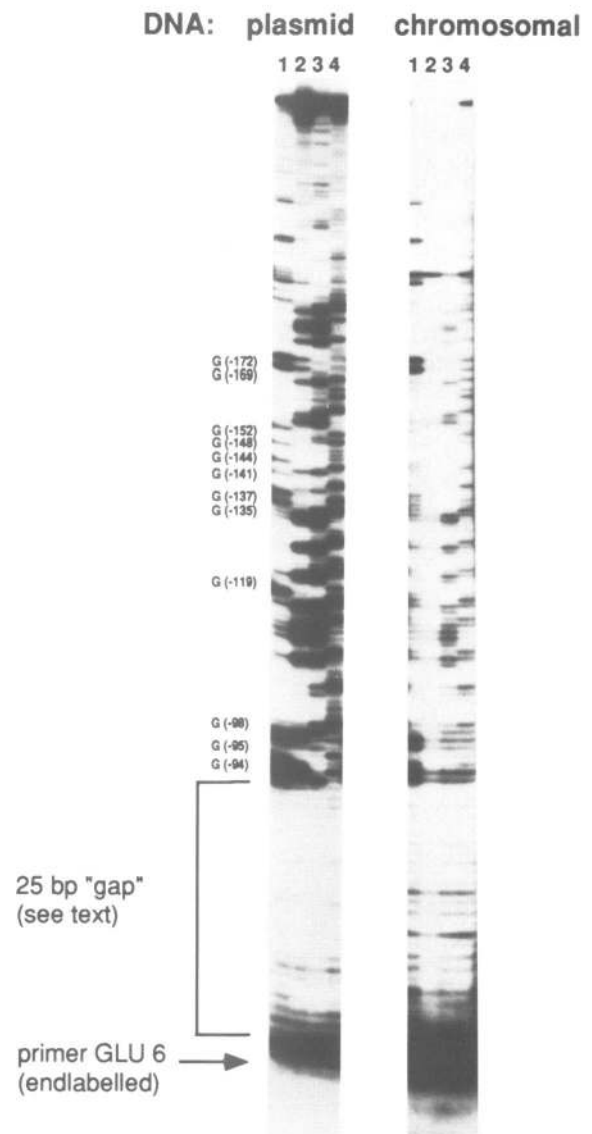
Fig. 3. Genomic sequencing of the LMWG-1D1 promoter from Chinese Spring chromosomal DNA. Lanes 1–4 show the DNA reacted with the G-, GA-, AC- and TC-specific chemicals, respectively. The sequenced strand is the top strand using the primer set Glu3/Glu5/Glu6.

Using the same set of primers the promoter region of LMWG-1D1 could be sequenced directly from *Triticum aestivum* (cv Chinese Spring) total genomic DNA (10 μg) using the Maxam−Gilbert sequencing methodology and LMPCR. Furthermore, the Maxam−Gilbert sequencing pattern is the same for chromosomal and plasmid DNA (p326H) although the similarity is slightly obscured by differing band intensities (Figure 3; lane 1, G-specific Maxam−Gilbert reaction; lane 2, GA; lane 3, AC; lane 4, TC).

In vivo footprinting of LMWG-1D1 with DMS

Grains were harvested 15 ± 4 days after anthesis (DAA) and the endosperm was isolated by carefully peeling off all protective layers around the endosperm and removing the embryo. Between 5 and 20 grains were used per treatment. The endosperm was immersed in 1% DMS for 5 min at

room temperature and washed several times in MS salts before DNA preparation. Different concentrations and times of treatment of endosperm tissue with DMS were attempted. Optimal results giving adequate reaction to reveal fragments in the range of 20−200 bases upon LMPCR were obtained with 1% DMS for the time indicated. However, in some circumstances, different degrees of *in vivo* reaction occurred despite control of the duration of treatment. This led to differences in the overall abundance and distribution of reacted G residues between different treatments. This is most probably an intrinsic problem associated with reacting large masses of tissue with DMS, where penetration of the reagent is not even. Parallel samples were treated under identical conditions without DMS (0%, Figure 4) in order to be able to differentiate between bands generated by *in vivo* DMS

treatment and bands resulting from other forms of degradation. As can be seen in Figure 4 the background is still rather high, particularly in the bottom half of the gel. However, the G-specific bands can be identified even in this region of the gel, indicating that the DMS has reacted with the chromosomal DNA in a sequence-specific manner. One G residue, denoted by an arrow, G (−119), cannot be detected above the background in the DMS treated samples (compare lanes 1 and 3). This suggests that a protein may be bound at or near G (−119), protecting it from reaction with DMS. The endosperm box (hatched) is located in the top half of the gel, where the background is reduced. In order to visualize the reacted G residues in this part of the gel, it had to be autoradiographed for longer. The G-specific bands are more distinct and most G residues are equally
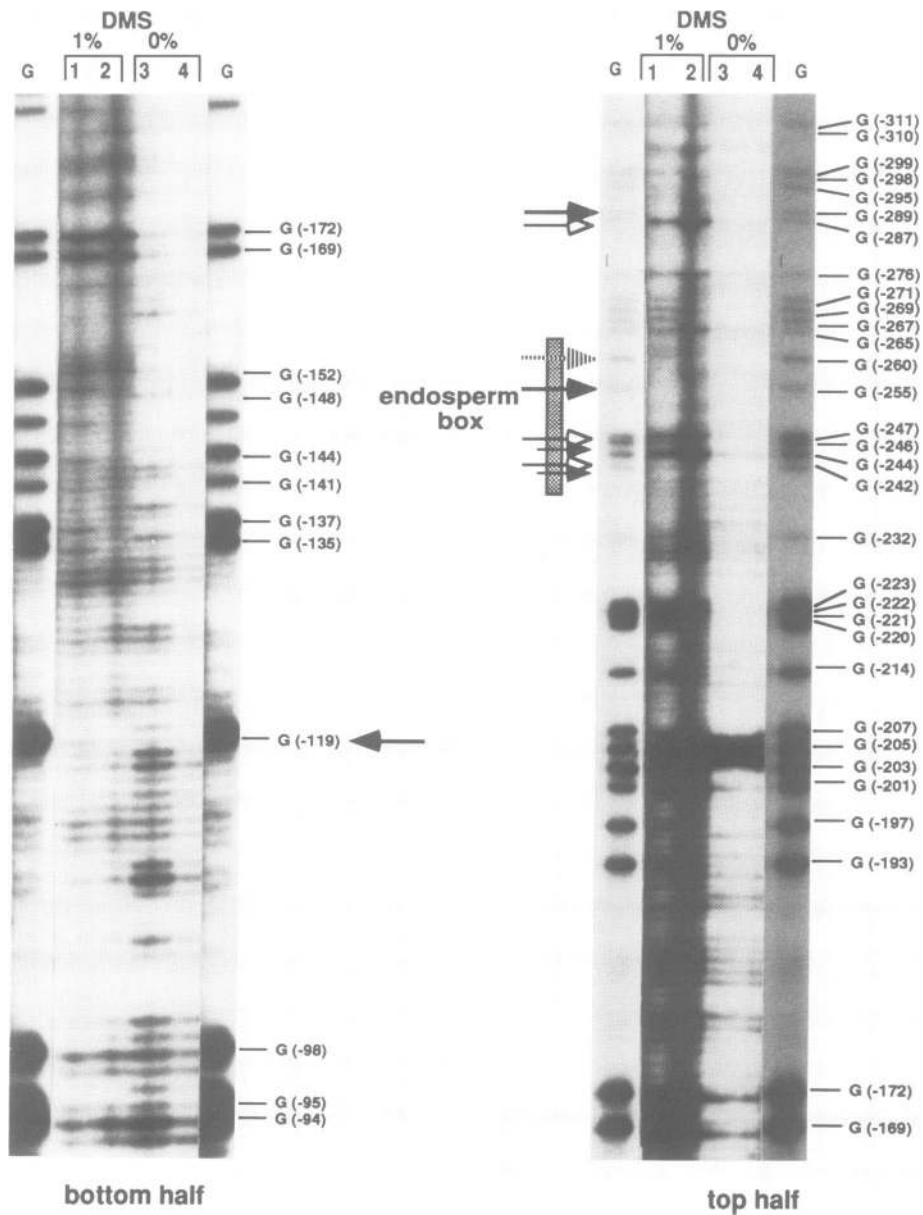

Fig. 4. *In vivo* DMS footprinting of the LMWG-1D1 5′ region in wheat endosperm by LMPCR. The primer set used was Glu3/Glu5/Glu6, thereby visualizing the top strand. Lanes 1+2 and 3+4 contain duplicates of DNA from endosperm reacted with 1% DMS or 0% DMS, respectively. The position of the G residues is marked to the right, next to the G lane. Black arrows indicate protected (hypomethylated) G residues, hollow arrows indicate hypermethylated G residues and the hatched arrow indicates a G residue whose presence may be obscured (see text). The position of the endosperm box is indicated by the hatched box.

present in *in vivo* and *in vitro* treated samples (compare lanes 1 and 2 with G). In particular the G residues adjacent to the endosperm box, G (−220 to −223) and G (−265 to −271), are clearly present in the *in vivo* treated endosperm samples. However, G (−255) is protected from DMS, as is G (−289) further upstream (black arrows). G (−260) (shaded arrow) also seems to be absent, although a rather diffuse faint band around this position might obscure the presence of this G residue. Furthermore, G residues (−242 to −247) are not equally presented in lanes G and 1. It appears that G (−247) and G (−244) are hypersensitive (hollow arrows) whereas G (−246) and G (−242) are weaker, although not completely absent (small black arrows). Also, G (−287) appears to be hypersensitive (hollow arrow). These *in vivo* footprints were obtained with primer set Glu3/Glu5/Glu6 following the original protocol of Mueller and Wold (1989). The results are summarized in Figure 5. It is interesting to note that all hypo- and hypermethylated G residues in the LMWG-1D1 promoter are at or near to putative *cis*-acting elements.

This pattern of hyper- and hypomethylation around the endosperm box motif was examined in tissues in which the LMWG-1D1 gene was differentially active. This would allow a correlation to be made between the observed methylation patterns and the activity of the gene. Endosperm was harvested at different times after anthesis and together with different tissues, the flag leaf (taken from the plant at 13 DAA) and the husk of the grain (comprising glume, lemma, palea and aleurone tissue), was reacted with DMS and analysed by LMPCR. The LMPCR protocol used for the subsequent analysis is modified (Rigaud *et al.*, 1991), using only two gene-specific primers.

The tissue-specific occupancy of the endosperm box of LMWG-1D1 in wheat endosperm is shown in Figure 6 for the top strand, using Glu9 and Glu10 primers. The bottom strand is not shown due to the lack of G residues within the analysed region. Endosperm was harvested at 10, 13 and

18 DAA and reacted with DMS *in vivo* as described. The state of gene expression at these stages, as derived from transgenic CAT expression experiments (Colot *et al.*, 1987) is shown above Figure 6. Also shown are the results for leaf and husk tissue where the gene is not expressed. At 10 DAA all G residues in and around the endosperm box [encompassing G (−242) to G (−260)] are still equally represented in accordance with only basal transcription levels. At 13 DAA, while all G residues are still present, G (−255) (half arrow in lane 3) is only approximately half the intensity of G (−260) as opposed to an equal intensity at 10 DAA and in the *in vitro* reacted chromosomal DNA. Also, G (−260) is represented with similar intensities in the different lanes compared with G (−271) and G (−232), suggesting that G (−255) is partially hypomethylated rather than G (−260) being hypermethylated. However, due to the lack of other nearby G residues to act as anchor points, the potential hypermethylation of G (−260) needs further investigation. At 18 DAA, where the rate of transcription is highest, not only is G (−255) protected completely, but so are the four adjacent G residues [G(−242 to −247)], as indicated by the arrows in lane 4. This shows that G residues within the GCN4/jun-like motif of the endosperm box become protected from DMS reaction during the highest rate of transcription of LMWG-1D1. Also, G (−269), G (−267) and G (−265) (see below) are protected at 18 DAA. The protection of G (−265) (indicated by a shaded arrow) is less certain due to its weaker intensity even in the *in vitro* treated samples. These residues are, however, still fully unprotected at 13 DAA when the transcription rate is above basal level but still low. Furthermore, all G residues within the endosperm box and upstream region are fully unprotected, and thus unoccupied, in leaf and husk tissues, where the LMWG-1D1 gene is not expressed.

The binding pattern at the GCN4/jun motif differs between the two experiments shown in Figures 4 and 6. While the footprinting of the endosperm box with primers Glu9 and Glu10 shows that the four G residues within the GCN4/jun motif are completely protected at 18 DAA (Figure 6), the earlier experiment (Figure 4) shows that they are only partially hyposensitive and hypersensitive. Also, in Figure 4 all G residues in the upstream GCN4 motif [G (−269, −267 and −265)] (see below) are unprotected. This could reflect differences in reactivity due to the different times after anthesis when the DMS treatment was performed. The time after anthesis could not be accurately determined to within 4 days for the earlier experiment and thus the DMS pattern observed in Figure 4 could reflect partial protection of the GCN4/jun motif.

Fig. 5. Summary of hyper- and hypomethylated sites in the 5′ region of LMWG-1D1. The arrows indicate the footprinted sites (see Figure 4). The putative *cis*-acting elements are shown in bold and the position of the primers used (Glu3, Glu5 and Glu6) is indicated. All identified footprinted G residues map to or very close to the putative *cis*-elements.
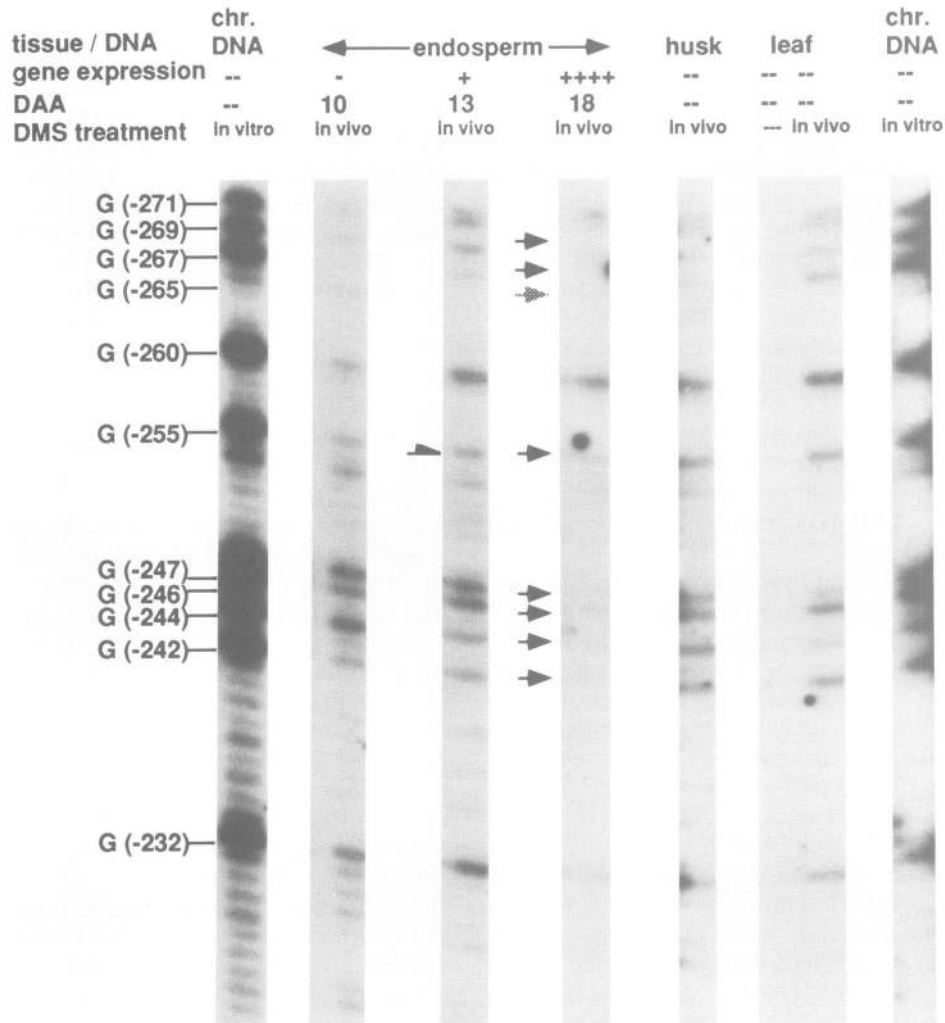
Identification of protein factors specific for the *in vivo* footprinted sites

The *in vivo* footprinting data revealed that three sites within and around the endosperm box of the LMWG-1D1 promoter were protected from methylation specifically in the endosperm. Although this hypomethylation could be the result of a number of mechanisms, the most likely one involves the specific binding of proteins at these sites. The proteins would be tightly bound at their cognate sequences and thus protect any G residues from DMS methylation. A comparison of the sequences within the protected regions revealed that whereas site I [TGTAAAGT (protected Gs in bold)] is distinct; the other two sites, II (GGTGAGTCAT)

Fig. 6. The endosperm box becomes occupied during grain ripening in a tissue-specific manner only in the endosperm. Endosperm was harvested 10, 13 and 18 DAA, reacted with 1% DMS and analysed using LMPCR with primers Glu9 and Glu10, visualizing the top strand. Leaf and husk tissues were treated under the same conditions. The position of each G residue is marked on the left. The black full arrows indicate fully protected G residues at 18 DAA, while the half arrow indicates a partially protected G (−255) at 13 DAA. The grey arrow [G (−265)] indicates that this residue is under-represented even in *in vitro* treated DNA, so that it is not clear whether this residue is protected at 18 DAA. No protected G residues can be seen in DNA from DMS-treated leaf and husk tissues.

and III (TGTGTGACAT), are very similar and may be regarded as two copies of the GCN4/jun-like motif. It was thus reasonable to expect at least two different activities would bind sites I and II,III. A close inspection of the sequences around the endosperm box in the promoter of LMWG-1D1 shows that not only the GCN4 motif but also the endosperm motif are duplicated. The second copy of the endosperm motif also contained a DMS footprint (TGTAAAAGTG, Figure 4). Therefore the extended region containing two copies of each motif can be viewed either as an extended endosperm box or as a duplicated box.

In order to assay directly for the endosperm box-specific binding proteins postulated above, an electrophoretic mobility shift assay (EMSA) was employed. As the duplicated motifs differ slightly in their exact sequence, the extended region comprising nucleotides −302 to −232 was used as a probe (Figure 7A) to screen heparin−agarose column fractions of a wheat endosperm nuclear protein extract prepared from endosperm tissue ~15 ± 3 days after anthesis. Unfractionated nuclear extracts contained a plethora

of binding activities which obscured specific binding activities. To differentiate between proteins binding to the endosperm and GCN4/jun motifs, mutant probes were made so that all G residues which had been shown to be either hypo- or hypermethylated in the *in vivo* footprinting experiments were substituted with T residues. Three mutant probes were used where either both copies of each motif separately (oligos II and III, Figure 7A) or combined (oligo IV, Figure 7A) were substituted as indicated in Figure 7A. Three fractions were identified, designated A (binding activity eluting from the column at 100 mM KCl), B (275 mM KCl) and C (650 mM KCl), that contain proteins which bound specifically to the extended wild-type endosperm-box probe. The precise binding sites of proteins within these fractions were identified using the mutated endosperm-box oligonucleotides (Figure 7B and C). Fractions A, B and C all contained activities that bound oligos I and II, but not III or IV, indicating that these activities specifically bound the endosperm motifs, requiring G nucleotides for binding at the positions exactly specified by the *in vivo* footprinting

A.

Wild type - EM+GCN4+ (oligonucleotide I)

```
      •  •                              •• •          •        •• • •
CTC GGT TGT AAA AGT GAT ACT ATC TTG ATA AGT GTG TGA CAT GTA AAG TTA ATA AGG TGA GTC ATA TAT AGC
GAG CCA ACA TTT TCA CTA TGA TAG AAC TAT TCA CAC ACT GTA CAT TTC AAT TAT TCC ACT CAG TAT ATA TCG
```
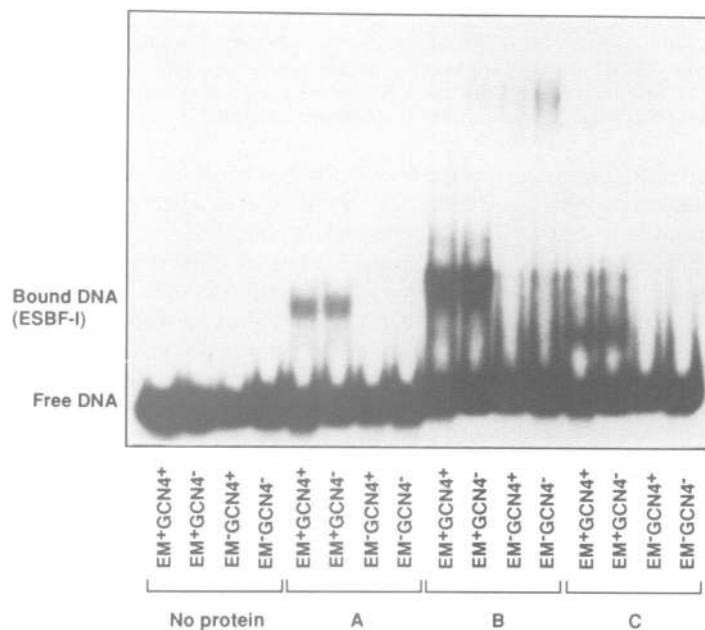
GCN4 motifs mutated - EM+GCN4- (oligonucleotide II)

```
CTC GGT TGT AAA AGT GAT ACT ATC TTG ATA AGT tTt TtA aAT GTA AAG TTA ATA Att TtA tTa ATA TAT AGC
GAG CCA ACA TTT TCA CTA TGA TAG AAC TAT TCA aAa AaT tTA CAT TTC AAT TAT Taa AaT aAt TAT ATA TCG
```

Endosperm motifs mutated - EB-GCN4+ (oligonucleotide III)

```
CTC GGT TGT AAA AtT tAT ACT ATC TTG ATA AGT GTG TGA CAT GTA AAt TTA ATA AGG TGA GTC ATA TAT AGC
GAG CCA ACA TTT TaA aTA TGA TAG AAC TAT TCA CAC ACT GTA CAT TTa AAT TAT TCC ACT CAG TAT ATA TCG
```

Double mutant - EB-GCN4- (oligonucleotide IV)

```
CTC GGT TGT AAA AtT tAT ACT ATC TTG ATA AGT tTt TtA aAT GTA AAt TTA ATA Att TtA tTa ATA TAT AGC
GAG CCA ACA TTT TaA aTA TGA TAG AAC TAT TCA aAa AaT tTA CAT TTa AAT TAT Taa AaT aAt TAT ATA TCG
```

B.

C.

Bound DNA (ESBF II)

Bound DNA (ESBF-I)

Free DNA

EM+GCN4+ EM+GCN4- EM-GCN4+ EM-GCN4- EM+GCN4+ EM+GCN4- EM-GCN4+ EM-GCN4- EM+GCN4+ EM+GCN4- EM-GCN4+ EM-GCN4- EM+GCN4+ EM+GCN4- EM-GCN4+ EM-GCN4-

EM+GCN4+ EM+GCN4- EM-GCN4+ EM-GCN4-   Probe

No protein   A   B   C   C   Protein fraction
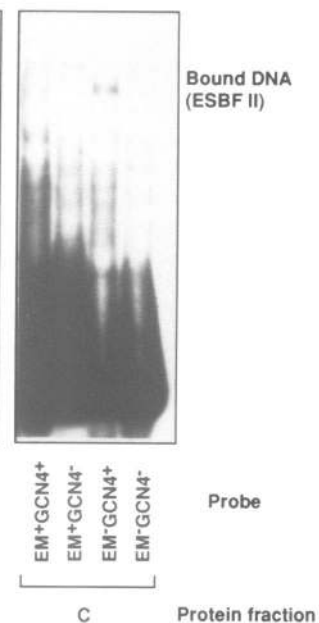
Fig. 7. *In vitro* interaction of wheat endosperm nuclear factors with the endosperm box. A. DNA sequences of oligonucleotides used as probes in EMSA analysis. Nucleotides altered in mutant probes are indicated in lower case. B and C. EMSA analysis of wheat endosperm nuclear protein fractions partially purified by heparin–agarose column chromatography (fractions labelled A, B and C). B. Identification of factors in fractions A, B and C that bind the endosperm motif (collectively called ESBF-I). C. Identification of a factor in fraction C that binds the GCN4-motif (ESBF-II). The positions of protein-bound DNA and free DNA are indicated.

data. As well as containing an endosperm motif-binding activity, fraction C also contained a less abundant activity that bound oligos I and III, but not II or IV, indicating that it interacted specifically with the GCN4/jun motifs.

Discussion

We have shown that is possible to footprint putative *trans*-acting factor interactions with DNA in intact plant organs and dissected tissues. In particular, the LMPCR technology allows one to use <1 μg chromosomal DNA, thus potentially allowing analysis of individual organs or small amounts of dissected tissue. This is the first time that the feasibility of *in vivo* footprinting of intact plant organs has been demonstrated, although protein–DNA interactions on the light-responsive chalcone synthase promoter from parsley and the alcohol dehydrogenase gene from maize and

Arabidopsis have been demonstrated previously by *in vivo* footprinting of undifferentiated suspension cultured cells (Ferl and Nick, 1987; Schulze-Lefert *et al.*, 1989; McKendree *et al.*, 1990). This approach should therefore prove to be particularly valuable for studying the regulation of gene expression in higher plants, because differentiated plant cell types cannot yet be maintained as cell lines.

One of the causes of differential reactivity of G residues towards DMS is the specific binding of proteins close to these residues. To confirm this possibility, gel retardation experiments using probes with those G residues altered to T residues were carried out. The results obtained indicated that one class of endosperm nuclear protein bound to the endosperm motif and another to the GCN4/jun motif, thereby validating the observations made in the *in vivo* footprinting experiments.
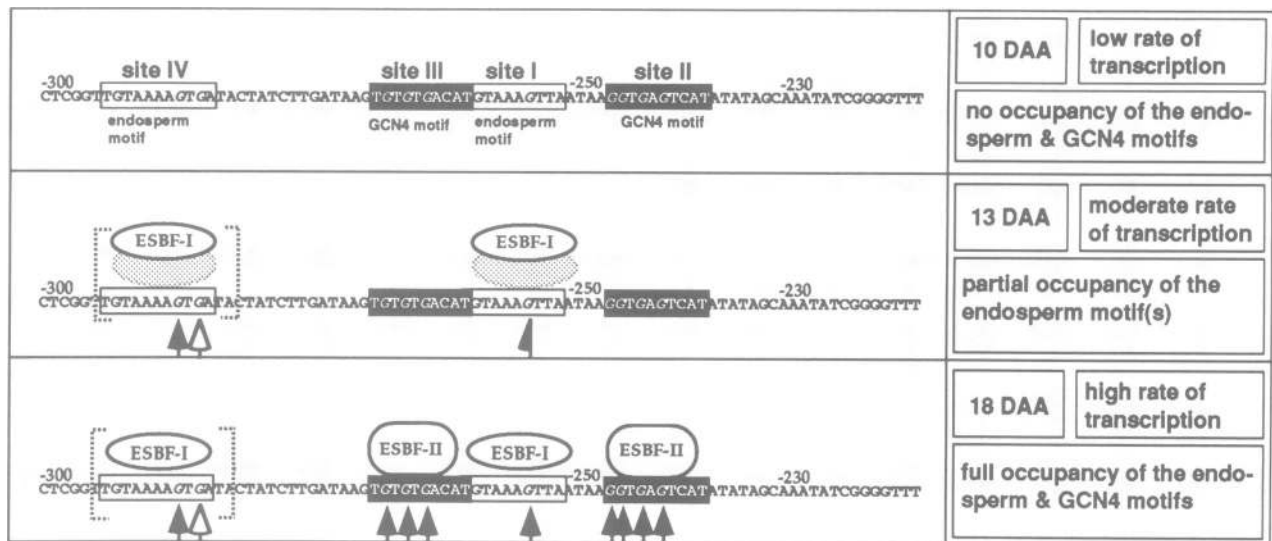
Fig. 8. A model for the function of the endosperm region in prolamin genes. The endosperm box, present in two adjacent copies, contains two *cis*-acting elements to which two specific *trans*-acting factors, ESBF-I and ESBF-II, bind sequentially in order to confer high level endosperm-specific gene expression. The boundaries shown here for each putative *trans*-acting factor are largely arbitrary, due to the limitations of DMS as a footprinting reagent. Site IV is shown in brackets because this footprint was observed in a different experiment.

This allows us to propose a model, outlined in Figure 8, where two separate factors bind to the extended endosperm box sequentially. The first factor, called endosperm box factor I (ESBF-I) binds to the endosperm motif in the endosperm box around 13 DAA, raising the level of transcription. This protein may be responsible for the endosperm-specific expression of prolamin genes in wheat and other cereals. At the highest rate of transcription around 18 DAA, a second factor, endosperm box factor II (ESBF-II) is bound at the GCN4/jun motif, and it is likely that the binding of this factor(s) confers the high rate of transcription. This is supported by the observation, in *opaque-2* maize mutants lacking the opaque-2 transcription factor, of markedly reduced but still endosperm-specific expression of 22 kDa zeins (Pedersen *et al.*, 1980; Kodrzycki *et al.*, 1989; Lohmer *et al.*, 1991). This differential reduction of 22 kDa zein transcription as compared with 19 kDa zeins (see Introduction) arises because only 22 kDa zein genes contain the GCN4/jun motif as well as the endosperm motif. While binding of a putative transcription factor like ESBF-I probably contributes to endosperm-specific transcription of both classes of zein genes, in *opaque-2* mutants the opaque-2 transcription factor is non-functional and thus the level of transcription of the 22 kDa zein appears much more reduced than that of the 19 kDa zein. It has recently been shown by Holdsworth *et al.* (1992a) that maize opaque-2 binds site II in the endosperm box and activates transcription in yeast, supporting the view that there is a wheat version of the maize factor opaque-2 that functions in similar ways. It is possible that the two proteins interact with each other and that binding of ESBF-I is necessary for binding of ESBF-II.

While we have shown that the endosperm box and endosperm region are involved in endosperm-specific high level expression, it is most likely that other *cis*-sequences are also involved. In particular, our analysis of the LMWG-1D1 promoter from −90 to −300 has revealed two additional footprints, both of which reside within potential *cis*-elements. The footprint around −289 is within a second copy of the endosperm motif, so it appears that two copies

of the endosperm box containing the endosperm and GCN4 motifs lie immediately adjacent to each other. Both endosperm motifs contain hypomethylated Gs at the same position and binding of ESBF-I is completely abolished *in vitro* by changing only the unreactive G residues, further indicating that ESBF-I binds to both endosperm motifs. A third copy of the endosperm box resides further upstream, at −513 to −540; this may also be involved in endosperm-specific expression but it has not yet been footprinted. The footprint at −119 is immediately adjacent to a CCAAT motif, which has been shown to be important for the general transcription machinery in mammalian genes and in the expression of genes in terminally differentiated cell types (Bienz and Pelham, 1986; Cao *et al.*, 1991; Gai *et al.*, 1992). This footprint suggests that a protein binds to this element in the LMWG-1D1 promoter, although binding could occur downstream of G (−119) because the DMS footprinting only identifies G-residues involved in binding. This is of particular interest because involvment of the CCAAT motif in plant gene expression has not yet been demonstrated. While many plant genes do not contain a CCAAT motif in their proximal promoter, this seems to be another shared feature among seed storage protein genes (Thompson and Larkins, 1989).

The methodology used here will be useful for investigating and understanding the regulation of gene expression in higher plants. These investigations need not be confined to the endosperm tissue, because we have shown that the gene of interest can be footprinted also in leaf and husk tissues. It is therefore likely that most plant tissues could be used for DMS *in vivo* footprinting.

Materials and methods

Materials

A plasmid containing the proximal promoter of the low molecular weight glutenin gene LMWG-1D1 from position −326 to +30 in pUC19, called p326VC, was provided by Vincent Colot. p326VC was linearized with *Hind*III, which cuts in the polylinker upstream of the promoter, prior to use in LMPCR and chemical sequencing. The linearized plasmid was called

p326H. Restriction enzymes were obtained from BCL, NBL and Pharmacia. All routine molecular biology procedures were performed as described in Sambrook *et al.* (1989).

Chromosomal DNA preparation

The method used is a modification of the method of Sørensen (Saluz and Jost, 1990). The tissue (endosperm, leaf and husk) was ground to a fine powder in liquid nitrogen and 1−5 ml of extraction buffer [10 mM Tris−HCl pH 9.5, 350 mM EDTA pH 8.0, 50 mM EGTA pH 8.0, 1% SDS, 2 mg/ml proteinase K (BCL)] was added with thorough mixing to wet all the tissue. Samples were incubated at 55°C overnight, followed by a brief spin to pellet starch and cell debris. The supernatant was extracted with phenol at least three times, followed by one phenol−chloroform−isoamylalcohol (24:24:1) extraction and one chloroform−isoamylalcohol extraction. The aqueous phase was then diluted 3-fold with water and precipitated by addition of 0.1 vol 2.5 M NaOAc pH 5.2 and 1 vol propan-2-ol. The DNA pellet was resuspended in TE pH 7.5 and reprecipitated. The final pellet was rinsed in 70% ethanol, briefly air-dried and resuspended in 20−100 µl TE pH 7.5 overnight at 4°C. Integrity of the obtained chromosomal DNA was checked by agarose gel electrophoresis and the concentration determined by fluorimetry. DNA samples were cut with *Hind*III prior to LMPCR.

Ligation mediated PCR

The procedure used follows the protocol of Mueller and Wold (1989) and Rigaud *et al.* (1991) with the following modifications. All reactions were carried out in a Techne PHC-3 PCR machine in microtitre plates. For first strand synthesis, primer hybridization was carried out using the program 95°C 2 min→50−60°C 30 min→20°C. It was essential to layer mineral oil on top. In our experience this does not interfere with the subsequent enzymatic steps when pipetting the enzymes and buffers through the oil. For strand extension, 0.5 µl T7 DNA polymerase (10 U/µl, Pharmacia) and 8.5 µl extension buffer (20 mM MgCl₂, 20 mM DTT, 200 µM dNTPs) were added and incubation proceeded at 40°C 10 min→65°C 10 min→15°C. For linker ligation 51 µl consisting of 6 µl 310 mM Tris−HCl pH 7.5, 20 µl ligation buffer A (17.5 mM MgCl₂, 42 mM DTT, 125 µg/ml BSA), 17 µl ligation buffer B (14.7 mM MgCl₂, 29 mM DTT, 4.4 mM ATP, 74 µg/ml BSA) and 5 µl annealed linker (15 µM in 250 mM Tris−HCl pH 7.5) and 3 µl T4 DNA ligase (1 U/µl) were added and the ligation was carried out at 15°C overnight, followed by heat inactivation (70°C for 10 min). All samples were subsequently precipitated with 0.1 vol 2.5 M NaOAc, 1 vol propan-2-ol after addition of 10 µg glycogen to ensure efficient precipitation of all samples. Samples were rinsed in 70% ethanol and resuspended in 20 µl H₂O. Samples were transferred into wells of a new microtitre plate and mixed with 27 µl PCR mix, containing 5 µl 10× PCR buffer (100 mM Tris−HCl pH 8.9, 500 mM KCl, 0.01% gelatin, 30 mM MgCl₂), 5 µl 2 mM dNTPs (ultrapure, Pharmacia), 0.5 µl *Taq* DNA polymerase (BCL) but no primers. Samples were denatured at 94°C for 2.5 min, after which the second or third gene specific primer and linker primer Ib were added while pausing the first cycle at 85°C. The cycling was then continued using the program 30×[94°C 40 s → 59°C 2 min → 76°C 3 min]. After completion of 30 cycles, 25 µl of each sample were used for the final extension cycling, while the rest was stored at −20°C for reamplification. To visualize the amplified sequence-specific fragments 25 µl of each PCR reaction were mixed with 5 µl 2 mM dNTPs, 2.5 µl 10×PCR buffer, 0.5 µl *Taq* DNA polymerase, 16 µl H₂O and 1 µl end-labelled second or third gene specific primer (50−500 000 d.p.m/µl). Samples were then cycled as follows: (i) 94°C 2.5 min, (ii) 9×[94°C 40 s → 59°C 3 min → 76°C 5 min]. Samples were cooled on ice, extracted once with phenol−chloroform−isoamylalcohol and precipitated with propan-2-ol as described above. After rinsing in 70% ethanol and briefly air drying the samples, they were resuspended in 10 µl sequencing gel loading buffer and 1−2 µl were loaded on a 6% sequencing gel. Gels were fixed, dried under vacuum and exposed to film for between 2 h and 5 days. The primers used for (LM)PCR were: Glu1, 5′-GTTCCAAACTCGGTTG-3′ (PCR upstream primer); Glu3, 5′-GCCCAGGAACCACCTCTTTTCCTT-CTTTGC-3′ (first gene specific primer for top strand); Glu5, 5′-CCTTC-TTTGCATCAGCTTTAATAATTC-3′ (second gene specific primer for top strand); Glu6, 5′-GCATCAGCTTTAATAATTCTGTTTGGC-3′ (third gene specific primer for top strand); Glu7, 5′-CTTAGTCATACACAA-GTTTTGCC-3′ (second gene specific primer for bottom strand); Glu8, 5′-GATTGCAAACTTAGTCATACACAAG-3′ (first gene specific primer for bottom strand); Glu9, 5′-GTTGTGCATACGATCACACACAAA-GTAC-3′ (first gene specific primer for top strand, second pair); Glu10, 5′-CGATCACACACACAAAGTACAGAAACCCC-3′ (second gene specific primer for top strand, second pair); linker primer Ib, 5′-GCAA-

TCATTTGAGAGATCTGAATTC-3′; linker primer II, 5′-GAATTCAG-ATC-3′.

The positions of all primers used are shown in Figure 1. All primers were synthesized on a Pharmacia LKB Gene Assembler Plus and purified on 16% sequencing gels prior to use. The linker primers were annealed as described elsewhere (Mueller and Wold, 1989). The concentration of the annealed linker was 15 µM.

DMS treatment of tissues

All reactions were performed using 0.1−1% dimethylsulfate (DMS) in MS salts (Murashige and Skoog, 1962). The lower DMS concentration resulted in a larger size distribution of PCR amplified fragments when analysed on agarose gels prior to extending the end-labelled primer, but gave only very faint signals on the autoradiograph after the last extension step. Thus we chose 1% DMS for most experiments resulting in proportionally higher signal intensities, but adequate size distribution.

Single leaves from 'Chinese Spring' wheat were harvested between 13 and 18 DAA, immersed in 3 ml DMS solution and vacuum infiltrated for 5 min at room temperature. For treatment of endosperm, the endosperm tissue was either squeezed out from the grain or isolated by peeling off all protective layers including the aleurone and removing the embryo. The endosperm thus isolated was immersed in DMS solution for 5 min at room temperature. For samples where the squeezed endosperm was used, the remainder of the grain, the husk, was also reacted with DMS under the same conditions. After treatment all samples were washed thoroughly at least three times with MS salts (without DMS) and frozen in liquid nitrogen, and the chromosomal DNA was prepared as described above.

Chemical sequencing

Maxam−Gilbert sequencing was carried out using the reactions specific for G (DMS, Aldrich), AG (sodium formate), TC (hydrazine, Eastman Kodak) and AC (NaOH) following the modified protocol by Bencini *et al.* (1984) and Sambrook *et al.* (1989, pp. 13.96−13.97). For sequencing of plasmid DNA 15 µg of salmon sperm DNA were included in all reactions.

Preparation of endosperm nuclear extracts and electrophoretic mobility shift assays

Nuclear protein from wheat endosperm (~15 DAA) was prepared as previously described (Jackson, 1990), and was fractionated by passage over a heparin−agarose column. Protein was eluted with a linear gradient of 40−1200 mM KCl. Fractions were assayed for DNA binding activity by EMSA as described below, and positive fractions containing the same activities were pooled. Aliquots from heparin−agarose columns were used to define the precise binding sites of the endosperm factors in EMSAs as described previously (Holdsworth *et al.*, 1992). Assays included radiolabelled oligonucleotide DNA probe (DNA sequence as shown in Figure 7, isolated from a plasmid as a *Hind*III−*Bam*HI fragment, ~300 pM), 1000 ng poly(dAdT·dAdT) unspecific competitor DNA and 2.5 µl protein fraction. EMSA reactions were carried out for 20 min at 25°C in a volume of 10 µl containing 25 mM HEPES (pH 7.5), 10% glycerol, 1 mM EDTA. Reactions with fraction A also contained 10 mM MgCl₂, which was necessary for the activity of the endosperm-box binding factor in this fraction. Protein−DNA complexes were resolved by electrophoresis on native 4% polyacrylamide gels before autoradiography.

Acknowledgements

References

Bartels,D. and Thompson,R.D. (1986) *Plant Sci.*, **46**, 117−125.
Bencini,D.A., O'Donovan,G.A. and Wild,J.R. (1984) *BioTechniques*, **2**, 22−27.
Bienz,M. and Pelham,H.R.B. (1986) *Cell*, **45**, 753−760.
Cao,Z., Umek,R.M. and McKnight,S.L. (1991) *Genes Dev.*, **5**, 1538−1552.

Colot,V. (1990) In Setlow,J.K. (ed.), *Genetic Engineering*. Plenum Press, New York, Vol. 12, pp. 225–242.

Colot,V., Robert,L.S., Kavanagh,T.A., Bevan,M.W. and Thompson,R.D. (1987) *EMBO J.*, **6**, 3559–3564.

Colot,V., Bartels,D., Thompson,R. and Flavell,R. (1989) *Mol. Gen. Genet.*, **216**, 81–90.

Ferl,R.J. and Nick,H.S. (1987) *J. Biol. Chem.*, **262**, 7947–7950.

Forde,B.G., Kreis,M., Baharmian,M.B., Matthews,J.A. and Miflin,B.J. (1981) *Nucleic Acids Res.*, **9**, 6689–6707.

Forde,B.G., Heyworth,A., Pywell,J. and Kreis,M. (1985) *Nucleic Acids Res.*, **13**, 7327–7339.

Gai,X.X., Lipson,K.E. and Prystowsky,T.U.B. (1992) *Nucleic Acids Res.*, **20**, 601–606.

Hammond-Kosack,M.C.U., Dobrinski,B., Lurz,R., Docherty,K. and Kilpatrick,M.W. (1992) *Nucleic Acids Res.*, **20**, 231–236.

Harberd,N.P., Bartels,D. and Thompson,R.D. (1985) *Mol. Gen. Genet.*, **198**, 234–242.

Hartings,H., Maddaloni,M., Lazzaroni,N., Di Fonzo,N., Motto,M., Salamini,F. and Thompson,R. (1989) *EMBO J.*, **8**, 2798–2801.

Hartings,H., Lazzaroni,N., Marsan,P.A., Araguay,A., Thompson,R., Salamini,F., Di Fonzo,N., Palau,J. and Motto,M. (1990) *Plant Mol. Biol.*, **14**, 1031–1040.

Hill,D.E., Hope,I.A., Macke,J.P. and Struhl,K. (1986) *Science*, **234**, 451–457.

Holdsworth,M.J., Grierson,C.S., Schuch,W. and Bevan,M.W. (1992) *Plant Mol. Biol.*, **19**, 455–464.

Jackson,S. (1990) Ph.D. Thesis, Cambridge University.

Kodrzycki,R., Boston,R.S. and Larkins,B.A. (1989) *Plant Cell*, **1**, 105–114.

Kreis,M., Shewry,P.R., Forde,B.G., Rahman,S., Bahramian,M.B. and Miflin,B.J. (1984) *Biochem. Genet.*, **22**, 231–255.

Kridl,J.C., Vieira,J., Rubinstein,I. and Messing,J. (1984) *Gene*, **28**, 113–118.

Lohmer,S., Maddaloni,M., Motto,M., Di Fonzo,N., Hartings,H., Salamini,F. and Thompson,R.D. (1991) *EMBO J.*, **10**, 617–624.

McKendree,W.L., Paul,A.-L., DeLisle,A.J. and Ferl,R.J. (1990) *Plant Cell*, **2**, 207–214.

Mitchell,P.J., Wang,C. and Tjian,R. (1987) *Cell*, **50**, 847–861.

Mueller,P.R. and Wold,B. (1989) *Science*, **246**, 780–786.

Mueller,P.R. and Wold,B. (1991) *Methods*, **2**, 20–31.

Murashige,T. and Skoog (1962) *Physiol. Plant*, **15**, 473–497.

Pedersen,K., Bloom,K.S., Anderson,D.V., Glover,D.V. and Larkins,B.A. (1980) *Biochemistry*, **19**, 1644–1650.

Piette,J., Hirai,S.-I. and Yaniv,M. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 3401–3405.

Ransone,L.J. and Verma,M.I. (1990) *Annu. Rev. Cell Biol.*, **6**, 539–557.

Rigaud,G., Roux,J., Pictet,R. and Grange,T. (1991) *Cell*, **67**, 977–986.

Saluz.,H.P. and Jost,J.P. (1990) In Azzi,A., Polak,J.M. and Saluz,H.P. (eds), *BioMethods*. Birkhäuser Verlag, Basel, Boston, Berlin, pp. 35–38.

Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schmidt,R.J., Ketudat,M., Aukerman,M.J. and Hoschek,G. (1992) *Plant Cell*, **4**, 689–700.

Schulze-Lefert,P., Becker-André,M., Schulz,W., Hahlbrock,K. and Dangl,J.L. (1989) *Plant Cell*, **1**, 707–714.

Soave,C. and Salamini,F. (1984) *Phil. Trans. R. Soc. Lond. B*, **304**, 341–344.

Sørensen,M.B. (1992) *Proc. Natl. Acad. Sci. USA*, in press.

Sørensen,M.B., Cameron-Mills,V. and Brandt,A. (1989) *Mol. Gen. Genet.*, **217**, 195–201.

Thomas,M.S. and Flavell,R.B. (1990) *Plant Cell*, **2**, 1171–1180.

Thompson,G.A. and Larkins,B.A. (1989) *BioEssays*, **10**, 108–113.