Evolution of the high molecular weight glutenin loci of the A, B, D, and G genomes of wheat

Robin G. Allaby, Monica Banerjee, and Terence A. Brown

Abstract: We used PCR to obtain phylogenetically informative sequences from the high molecular weight glutenin genes of wheat. The validity of partial sequence comparisons as a means of studying glutenin phylogenetics was established by constructing neighbour-joining trees from partial alignments of 12 published glutenin allele sequences. PCR was then used to obtain 20 novel glutenin allele sequences from various *Triticum* and *Aegilops* species, including a 3000 year old preserved wheat. A neighbour-joining tree derived from all known glutenin allele sequences had eight clades, representing the eight loci from which the allele sequences were derived, and was split into two halves, one comprising alleles from the *Glu-1-1* loci and the other comprising *Glu-1-2* alleles. The topology was compatible with the postulated relationships between the A, B, D, and G genomes. The *Glu* gene duplication event was tentatively dated at 7.2–10.0 million years ago (MYA), the origin of the four genomes at 5.0–6.9 MYA, and the split between the B and G genomes at 2.5–3.5 MYA. The *Glu-B1-1* alleles in cultivated wheats fell into two subgroups that diverged 1.4–2.0 MYA, suggesting that emmer was domesticated twice. The D allele sequences were relatively diverse, indicating that the hybridization event that resulted in the hexaploid bread wheats might have occurred more than once.

Key words: ancient DNA, HMW glutenin genes, phylogenetic analysis, *Triticum*, wheat.

Résumé : L'amplification PCR a été utilisée afin d'obtenir des séquences informatives pour l'analyse phylogénétique à partir des gènes codant pour les gluténines de haut poids moléculaire chez le blé. La validité des comparaisons de séquences partielles comme outil pour l'étude de la phylogénie de gènes codant pour la gluténine a été établie en produisant des arbres phylogénétiques à partir de l'alignement de douze séquences publiées d'allèles de la gluténine. La PCR a ensuite été employée pour obtenir 20 séquences additionnelles d'allèles provenant de diverses espèces des genres *Triticum* et *Aegilops*, dont un blé conservé datant de 3000 ans. Un arbre phylogénétique dérivé des séquences de tous les allèles connus présentait huit clades représentant les huit loci desquels provenaient les allèles. De plus l'arbre était scindé en deux, une moitié comprenant les allèles des loci *Glu-1-1* et l'autre comprenant les allèles *Glu-1-2*. La topologie de l'arbre était compatible avec les relations postulées entre les génomes A, B, D et G. La date approximative de duplication du locus Glu a été estimée à 7,2–10,0 millions d'années. L'origine des quatres génomes daterait d'environ 5,0–6,9 millions d'années tandis que la séparation des génomes B et G se serait produite il y a 2,5–3,5 millions d'années. Les allèles du locus *Glu-B1-1* présents chez les blés cultivés auraient divergés il y a 1,4–2,0 millions d'années, ce qui suggère que le blé amidonnier aurait été domestiqué à deux reprises. Les séquences des allèles du génome D étaient relativement différentes, indiquant que l'événement d'hybridation ayant produit les blés hexaploïdes est peut-être survenu à plus d'une reprise.

Mots clés : ADN ancient, gènes codant pour les gluténines HMW, analyse phylogénétique, *Triticum*, blé.

[Traduit par la Rédaction]

Introduction

The high molecular weight (HMW) glutenins of wheat are members of a complex group of seed storage proteins called prolamins. There are two subfamilies of HMW glutenin subunits (x- and y-type) which are thought to have arisen from a gene duplication event that predated the divergence of the A, B, D, and G genomes of cultivated wheats. Single copies of the x- and y-type genes occur at two tightly linked loci, *Glu-1-1* and *Glu-1-2* (Thompson et al. 1983; Harberd et al. 1986) on homoeologous chromosome 1 (Payne et al. 1982). Each locus is multiallelic (Payne and Lawrence 1983) with the combination of alleles contributing to the breadmaking quality of the wheat (e.g., Dong et al. 1991).

Protein studies have shown that HMW glutenins are highly polymorphic in wild wheats (Ciaffi et al. 1993) and also display significant variation in modern cultivars (Shewry et al. 1995). The latter point is important as the majority of genetic loci in modern cultivars show little variation, a reflection of the short evolutionary history of domesticated wheat (Bell 1987) coupled with the impact of genetic erosion during the last century (Hawkes 1991). The HMW glutenin loci may therefore be useful in studies of the evolution of both wild and cultivated wheats. For example, sequence comparisons between alleles associated with the different

Corresponding Editor: G.B. Golding.

Received July 20, 1998. Accepted August 6, 1998.

R.G. Allaby, M. Banerjee, and T.A. Brown.[1] Department of Biomolecular Sciences, University of Manchester Institute of Science and Technology, Manchester, M60 1QD, U.K.

[1]Author to whom all correspondence should be addressed (e-mail: terry.brown@umist.ac.uk.).

wheat genomes (A, B, D, and G) might provide information relevant to the origin and divergence times of the chromosome sets and indicate the extent to which homogenization between different loci has occurred in polyploid wheats. Glutenin alleles might also be used to make phylogenetic comparisons between different populations of wheats, for example in biogeographical studies such as those addressing the influence of environmental factors on wild wheat genotypes (e.g., Ciaffi et al. 1993), or in studies aimed at determining the number and geographical location of the original domestication event(s).

To use the HMW glutenin loci in wheat evolutionary studies, a convenient means of obtaining phylogenetically informative nucleotide sequences from the *Glu-1-1* and *Glu-1-2* loci is first necessary. We have previously used PCRs to obtain sequences from the regions immediately upstream of the glutenin open reading frames (ORFs), in order to study the short fragments of 'ancient' DNA present in some preserved wheats (Allaby et al. 1994, 1997; Brown et al. 1998). A similar strategy has also been used to test for the presence of specified alleles in modern cultivars (e.g., D'Ovidio and Anderson 1994; D'Ovidio et al. 1994). In this paper we describe the use of this approach to obtain HMW glutenin sequences from a number of *Triticum* and *Aegilops* varieties, both modern accessions and preserved archaeological remains, and show that the data can be used to make inferences about wheat evolution.

Materials and methods

Plant material

Triticum dicoccum tragi, *T. monococcum*, *T. timopheevi*, *T. urartu*, and *Aegilops squarrosa* (nomenclature of Miller 1987) were provided as seeds by the John Innes Centre, Norwich, U.K. *T. aestivum* var. *Galahad* was provided as DNA by Dr. M. Hamer (University of Manchester, U.K.). A 3000 year old archaeological wheat (probably a mixture of *T. dicoccum* and *T. spelta*) from Assiros, Greece, excavated by K.A. Wardle (University of Birmingham, U.K.; Jones et al. 1986), was obtained as charred grains from Dr. G. Jones (University of Sheffield, U.K.). Authentication of the presence of ancient DNA in the charred grain is described elsewhere (Allaby et al. 1997).

PCR amplification and sequencing

Nucleic acid extracts were prepared from 500-mg samples as previously described (O'Donoghue et al. 1996). Extracts of the charred seeds were further purified by electroelution at 6.66 V/cm for 1 h after fractionation in a 1.3% agarose gel (Towner 1991), followed by incubation at 37°C for 2 h with 0.01 vol of 10 mg/mL (Dnase-free) RNase A and reprecipitation with ethanol. PCRs, using the "hot-start" strategy (Erlich et al. 1991), were carried out in 100-μL reaction mixes containing 10 μL extracted nucleic acid (approximately 10 ng for modern samples), 10 μL buffer (Boehringer-Mannheim), 50 μM each dNTP, 200 ng each primer and 2.5 units Taq DNA polymerase (Boehringer-Mannheim), using primers (see Fig. 2) either with or without 5′ extensions (5′-CTCTGGATCC-) containing a restriction site for cloning purposes. Cycle parameters were 2.5 min at 94°C; 30 cycles of 2 min at 60°C, 1 min at 74°C, 1 min at 94°C; 2 min at 60°C; and 9 min at 74°C. The precautions taken to minimize the risk of contaminating ancient material with modern plant DNA are described in O'Donoghue et al. (1996). PCR products (25-μL aliquots) were fractionated in 3% (w/v) NuSieve or MetaPhor agarose gels (FMC BioProducts) and cloned and sequenced as described in O'Donoghue et al. (1996).

Sequence analysis

Sequences were aligned using CLUSTAL W (Thompson et al. 1994). Pairwise distances were calculated using the Kimura 2 parameter model with a transition/transversion ratio of 2.0. The ratio in the actual alignment was 2.6, but as this was an estimate based on limited data, and not very different from 2.0, we decided to use the standard ratio in our calculations. Principal coordinate analysis was carried out according to Higgins (1992). Neighbour-joining trees were constructed and bootstrapped using CLUSTAL W and topologies verified by the maximum likelihood method using the DNAML programme of PHYLIP 3.57 (Felsenstein 1989). Sites of homoplasy were identified using a compatibility analysis approach (Estabrook 1978) with DNACOMP of PHYLIP 3.57.

Results and discussion

Validity of phylogenetic analysis based on partial glutenin nucleotide sequences

Our initial objective was to devise a means of obtaining phylogenetically informative nucleotide sequences from HMW glutenin alleles. If possible, we hoped to base our studies on partial nucleotide sequences of 200–300 bp which could be acquired relatively rapidly by PCR, rather than dealing with the complications involved in sequencing complete ORFs of 1.8–3.0 kb in length. By using short sequences we could also include archaeological specimens in our study, as the fragmentary nature of ancient DNA means that it is not possible to obtain sequences of > 300 bp from preserved material (Allaby et al. 1997). In addition, we were doubtful about the veracity of phylogenetic analysis based on complete glutenin sequences, as 75–80% of each ORF comprises a central repetitive domain (Flavell et al. 1989) which is not amenable to multiple alignment.

The complete sequences of 12 glutenin alleles have so far been reported (Table 1A). To test the validity of partial sequence comparisons as a means of studying glutenin phylogenetics, we first made a multiple alignment of these 12 sequences using CLUSTAL, with modifications made by eye in an attempt to improve the alignment within the repetitive domain. This master alignment was then divided into 16 partial alignments (Fig. 1A), 161–548 bp in length, the longer segments being those in which the effective amount of sequence information was diluted by gaps or deletions in the multiple alignment. Finally, a neighbour-joining tree was constructed for each partial alignment. Ten of the partial alignments (segments 1, 2, 3, 4, 5, 6, 9, 10, 15, and 16), as well as the complete alignment (segments 1–16 combined), agreed with the tree topology shown in Fig. 1B. The trees for segments 7 and 8 placed *Glu-B1-2b* within the *Glu-D1-2* clade, but when these two segments were combined the single tree that was obtained had the topology shown in Fig 1B. Segments 11–14, when analysed individually, gave conflicting trees with low bootstrap values, but when combined produced a much stronger tree, similar to that shown in Fig. 1B, but with the *Glu-B1-1* clade shifted to a position adjacent to *Glu-B1-2b*. This may indicate that some homogenization has occurred between the *Glu-B1-1* and *Glu-B1-2* loci, but as the relevant region is within the repetitive domain (which has

Table 1. HMW glutenin alleles studied in this paper. Sequences described as "This study" have been deposited in GenBank (Accession nos. X98583–X98592, X98711–X98715, and Y12401–Y12410).

Species	Genome	Allele[a]	Reference
A. Published sequences			
T. aestivum var. *Asarce*	AABBDD	*Glu-A1-1c*	Xin et al. (1992)
T. aestivum var. *Cheyenne*	AABBDD	*Glu-A1-1b*	Anderson and Greene (1989)
		Glu-A1-2a	Forde et al. (1985)
		Glu-B1-1a	Anderson and Greene (1989)
		Glu-B1-1b	Reddy and Appels (1993)
		Glu-B1-2b	Halford et al. (1987)
		Glu-D1-1b	Anderson et al. (1989)
		Glu-D1-2a	Anderson et al. (1989)
		Glu-D1-2b	Thompson et al. (1985)
T. aestivum var. *Hope*	AABBDD	*Glu-A1-1d*	Halford et al. (1992)
T. aestivum var. *Yamhill*	AABBDD	*Glu-D1-1a*	Sugiyama et al. (1985)
Aegilops squarrosa	DD	*Glu-D1-2c*	Mackie et al. (1996)
B. Previously unpublished sequences			
T. aestivum var. *Chinese Spring*	AABBDD	*Glu-B1-1*α	A. Schlumbaum, personal communication
T. aestivum var. Galahad	AABBDD	*Glu-A1-1*α	This study
		*Glu-A1-2*α	This study
		*Glu-B1-1*β	This study
		*Glu-B1-2*α	This study
		*Glu-D1-1*α	This study
T. dicoccum tragi	AABB	*Glu-A1-1*β	This study
		*Glu-A1-2*β	This study
		*Glu-A1-2*γ	This study
		*Glu-B1-1*γ	This study
		*Glu-B1-2*β	This study
T. monococcum	AA	*Glu-A1-1*γ	This study
		*Glu-A1-2*δ	This study
T. timopheevi	AAGG	*Glu-A1-1*δ	This study
		*Glu-G1-1*α	This study
		*Glu-G1-2*α	This study
		*Glu-G1-2*β	This study
T. turgidum	AABB	*Glu-A1-1*ε	A. Schlumbaum, personal communication
		*Glu-B1-1*δ	A. Schlumbaum, personal communication
		*Glu-B1-2*γ	A. Schlumbaum, personal communication
T. urartu	AA	*Glu-A1-1*ζ	This study
		*Glu-A1-2*ε	This study
A. squarrosa	DD	*Glu-D1-1*β	This study
		*Glu-D1-2*α	This study
C. Previously unpublished sequences from archaeological wheats			
3000 year old mixed grain	AABB+	*Glu-A1-1*η	This study
	AABBDD	*Glu-A1-2*ζ	This study
		*Glu-B1-1*ε	This study
		*Glu-B1-1*ζ	This study
		*Glu-D1-1*γ	This study
5000 year old *T. aestivum*	AABBDD	*Glu-B1-1*η	A. Schlumbaum, personal communication

[a]The allele names adopted for the new sequences reported in this paper (Parts B and C of this Table) are given Greek designators (α, β, γ, etc.) to indicate that these are partial allele sequences, in contrast to the previously-published allele sequences (Part A, designators a, b, c, etc.) which are full-length.

evolved by repeat unit shuffling as well as point mutation) it is perhaps more likely that the altered tree topology is due to incorrect alignment.

Regardless of the explanation for the altered topology seen with the alignment of segments 11–14, the fact that the majority of the partial alignments gave the same tree as the complete alignment showed that comparison of these partial glutenin sequences is a valid means of studying the phylogenetics of the glutenin loci. We therefore directed our subsequent PCR and sequencing experiments at segment 1, this upstream region being chosen because the internal repetition of the glutenin open reading frames makes it difficult

to design primers that will anneal to unique sites within the coding region.

Acquisition of partial glutenin sequences from modern accessions

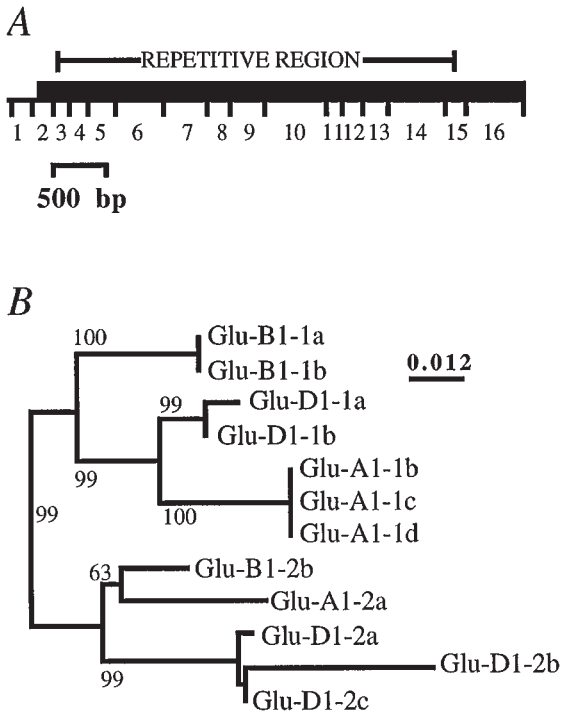
Total DNA was obtained from *Triticum dicoccum tragi*, *T. aestivum* var. *Galahad*, *T. monococcum*, *T. timopheevi*, *T. urartu*, and *Aegilops squarrosa*. For each DNA sample a PCR was carried out, the products cloned, and sequences obtained for 20–30 clones. As illustrated in Fig. 2 for *T. aestivum* var. *Galahad*, the sequences from a single accession fell into discrete groups when aligned, most sequences in a group being identical but some having minor variations which were attributed to errors introduced by the Taq polymerase during amplification. It was inferred that each group of sequences represented a single allele. The partial allele sequences obtained in this study from modern wheats and from *Aegilops squarrosa* are listed in Table 1B.

Acquisition of partial glutenin sequences from 3000 year old archaeological wheat samples

We have previously shown that the PCR primers used in this study can be used to amplify ancient DNA from the charred remains of a 2000 year old spelt wheat (Allaby et al. 1994, 1997). To determine if the phylogenetic analysis of glutenin loci can be extended to archaeological wheats, we amplified DNA from a second charred wheat, a 3000 year old sample of mixed grain from Assiros, Greece (Jones et al. 1986). Nineteen clones were sequenced, but these sequences did not fall into groups as seen with the modern specimens. In fact, the 19 sequences were all different. Although archaeological wheats are probably not genetically homogeneous, it seemed unlikely that the sample of approximately 20 charred grains used to prepare the DNA extract would contain 19 or more different glutenin alleles. Furthermore, the 19 sequences did not fall into discrete clusters when examined by principal coordinate analysis (Fig. 3), unlike the sequences of the modern alleles (previously published and determined in this project) which did form clusters because of sequence relationships displayed by the alleles affiliated with a single locus. This observation suggested that the sequences obtained from the charred grains did not represent genuine glutenin alleles. One possibility is that these sequences were chimera that arose by PCR jumping (Pääbo et al. 1990; Goloubinoff et al. 1993), an amplification artefact expected with ancient DNA preparations in which the preserved molecules are highly fragmented (e.g., Pääbo 1989; Rollo et al. 1991; Handt et al. 1994). The multiple alignment of the charred wheat sequences showed that segments were indeed shared as would be expected if PCR jumping had occurred. To reconstruct the original allele sequences from which the chimera had been derived, we devised a method similar to that described by Eickbush and Burke (1985). First, the multiple alignment of the chimeric sequences was examined by eye and each nucleotide position identified as "non-informative" or "informative," a non-informative position being one at which the same base occurred in each sequence, and an informative position being one at which at least two of the sequences were different. From the identities of the bases at the informative positions within each of the chimera, it was possible to break each sequence into a series of segments (Fig. 4) and to allocate each segment to one of five groups. This rearrangement could be done in such a way that the segments within any one group formed an overlapping full length sequence, the overlapping regions between segments taken from different chimera displaying no nucleotide differences except for a few anomalies (marked in Fig. 4) that were ascribed to PCR artefacts. The five resulting reconstructed sequences were inferred to be the partial sequences of five alleles present in the charred wheat sample (Table 1C; sequences shown in Fig. 5). These allele sequences clustered with modern sequences in a principal coordinate analysis suggesting that the reconstructions had been carried out correctly. The reconstruction methodology was further supported by a paper exercise in which the partial allele sequences from *T. dicoccum tragi* were randomly fragmented, assembled into chimera by simulated PCR jumping, and then successfully reconstructed.

These results extend our previous observation that glutenin loci can be amplified from at least some charred archaeological wheats (Allaby et al. 1994) and indicate that meaningful genetic analysis of ancient specimens might be possible.

Fig. 1. Phylogenetic analysis of 12 partial glutenin sequences. (A) Schematic representation of a HMW glutenin allele indicating the positions of the 16 segments on which the analysis was based. The black box indicates the open reading frame and the position of the repetitive domain is indicated. (B) Neighbour-joining tree obtained from the multiple alignment of segment 1 of the 12 published glutenin alleles. Multiple alignments of segments 1, 2, 3, 4, 5, 6, 7+8, 9, 10, 15, and 16, as well the complete sequence (segments 1–16 combined), all gave trees with topologies in agreement with the one shown. The bootstrap values that are shown were obtained by constructing 100 replicate trees. See Table 1A for the sources of the allele sequences.



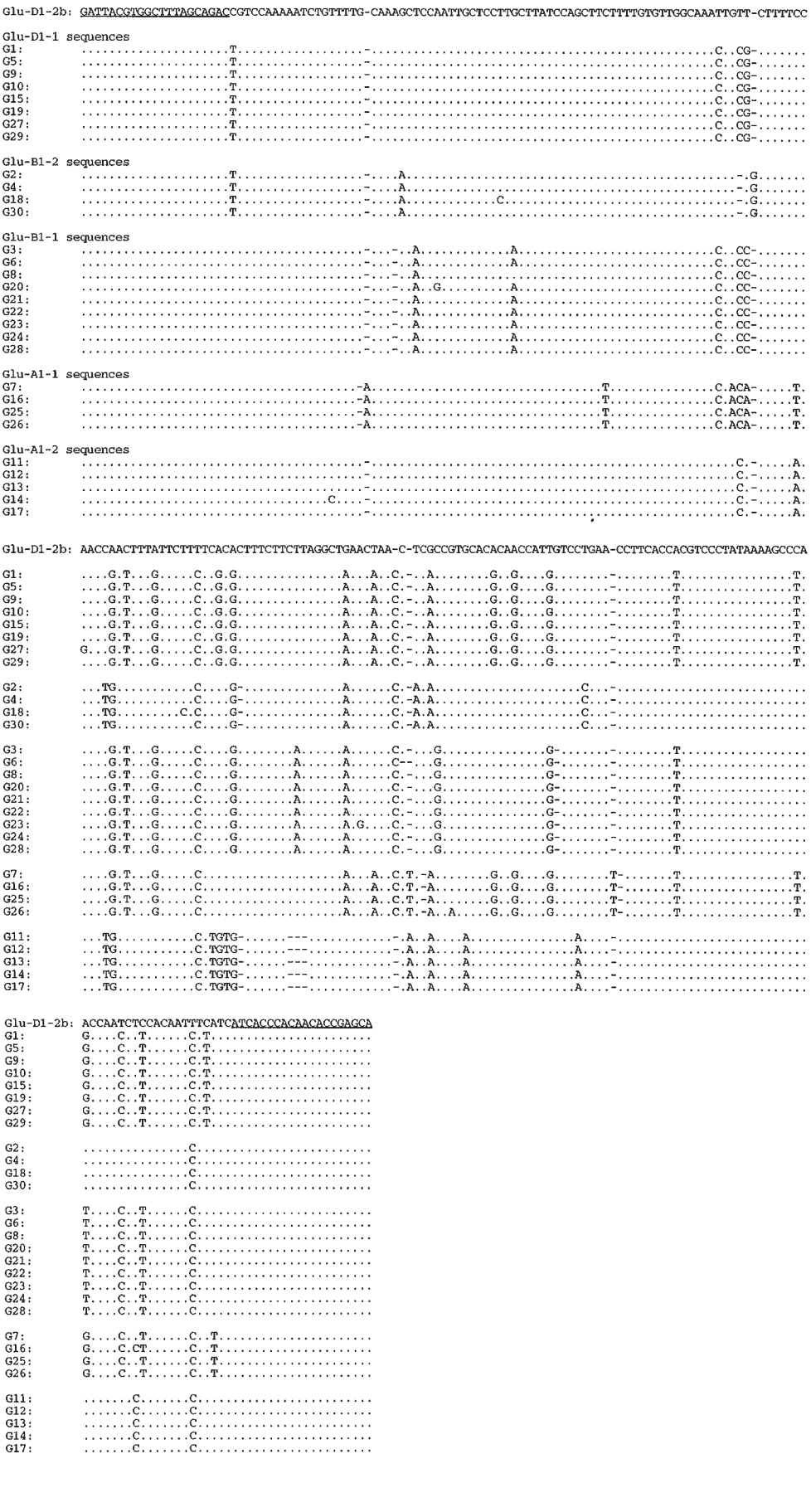


Fig. 2. Sequences of cloned PCR products obtained from *T. aestivum* var. *Galahad*. The amplified region is 241–243 bp and is located immediately upstream of the glutenin ORF (the last nucleotide of the downstream primer is 39–40 bp upstream of the initiation codon). The *Glu-D1-2b* allele is used as the reference sequence. G1 to G30 are the sequences of the PCR products, shown as discrete groups of highly similar sequences; each group is believed to represent a single allele. The locus to which each allele group belongs is indicated. Dots indicate identity with the *Glu-D1-2b* allele, dashes indicate deletion/insertion points. The primer annealing sites are underlined. Only five allele groups are seen as no sequence corresponding to a *Glu-D1-2* allele was obtained. This may have been simply due to chance, or might indicate that the sequence of the *Glu-D1-2* allele in this cultivar has an alteration in one or both of the primer annealing sites, preventing amplification.

Table 2. Mean-corrected pairwise distances (K) between loci.

	Glu-D1-1	*Glu-G1-1*	*Glu-B1-1*	*Glu-A1-2*	*Glu-G1-2*	*Glu-B1-2*	*Glu-D1-2*
Glu-A1-1	0.0472	0.0713	0.0859	0.1385	0.1339	0.1196	0.1156
Glu-D1-1	—	0.0616	0.0625	0.1079	0.1033	0.0851	0.0980
Glu-G1-1		—	0.0352	0.1098	0.0930	0.0862	0.0827
Glu-B1-1			—	0.1077	0.0931	0.0842	0.0816
Glu-A1-2				—	0.0622	0.0642	0.0657
Glu-G1-2					—	0.0360	0.0438
Glu-B1-2						—	0.0494

Phylogenetic analysis of partial glutenin sequences

The 42 partial glutenin sequences listed in Table 1 are aligned in Fig. 5. Compatibility analysis showed that there were 12 instances of incompatibility (boxed in Fig. 5) to the resulting neighbour-joining tree (Fig. 6). The distribution and nature of these incompatible sites gave no indication of major gene conversion events and were therefore treated as instances of homoplasy with the lineages assumed to be independent. The topologies obtained by the neighbour-joining and maximum likelihood methods were in agreement.

In the neighbour-joining tree (Fig. 6), eight clades were clearly formed representing the eight loci from which the allele sequences were derived. The tree divided into two halves, comprising the *Glu-1-1* orthologs at the top and the *Glu-1-2* orthologs at the bottom, rooted at the presumed gene duplication event. This topology suggests that the two lineages have evolved independently since the duplication. The branching order of the *Glu-1-2* alleles was only weakly supported by bootstrapping, possibly a result of the homoplasy between the *Glu-1-1* and *Glu-1-2* lineages, but was equivalent to the branching order obtained from the analysis of the 16 partial segments of the glutenin genes (Fig. 1B) and we therefore have confidence in it.

Mean pairwise differences between the clades are shown in Table 2. The two loci displaying the highest mean pairwise difference were *Glu-A1-1* and *Glu-A1-2*. The latter locus is not expressed in hexaploid wheats, presumably due to the presence of an internal stop codon (Forde et al. 1985) and hence is a pseudogene in these plants. Pseudogenes are under no functional constraint and are expected to change relatively rapidly (Li and Graur 1991) but in fact the branch lengths of the *Glu-A1-2* lineage relative to the central root of the tree where the gene duplication is presumed to have taken place are shorter than those of the *Glu-A1-1* lineage, implying a slightly higher fixation rate of mutations in *Glu-A1-1* compared to *Glu-A1-2*. Possibly this discrepancy is related in some way to the striking conservation of the *Glu-A1-1* lineage in the cultivated AABB tetraploids (*T. dicoc-*

Fig. 3. Principal coordinate plots of sequence data from (A) the amplified regions of modern allele sequences (published and determined in this project), and (B) the PCR product sequences obtained from the 3000 year old charred wheat. Principal coordinate analysis is a multivariate distance analysis representing high dimensional data in three dimensional space (Higgins 1992). Alleles associated with a particular locus cluster closely in space, as can be seen in (A). The chimeric nature of the sequences obtained from the archaeological grain (B) is evident from the lack of ordered clustering.

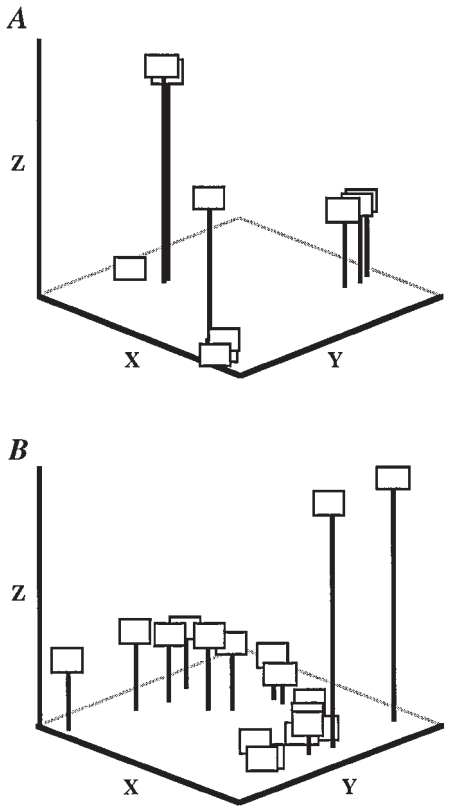


Fig. 4. Sequences of cloned PCR products (A1–A19) obtained from the 3000 year old charred wheat. The deduced segmental structure of each sequence is shown, the numbers indicating the group affiliation of each segment. For the method used to deduce the segment affiliations, see the text and also Eickbush and Burke (1985). The boxed nucleotides are anomalous sites that are ascribed to PCR artefacts. Some small regions (e.g., at the 5′-end of sequence A5) could not be assigned to a single group because of limitations in the distribution of the informative sites used in the analysis. The complete sequences of the five alleles reconstructed from this multiple alignment are given in Fig. 5: allele 1 is *Glu-B1-1ζ*; 2, *Glu-D1-1γ*; 3, *Glu-A1-1η*; 4, *Glu-A1-2ζ*; 5, *Glu-B1-1ε*.

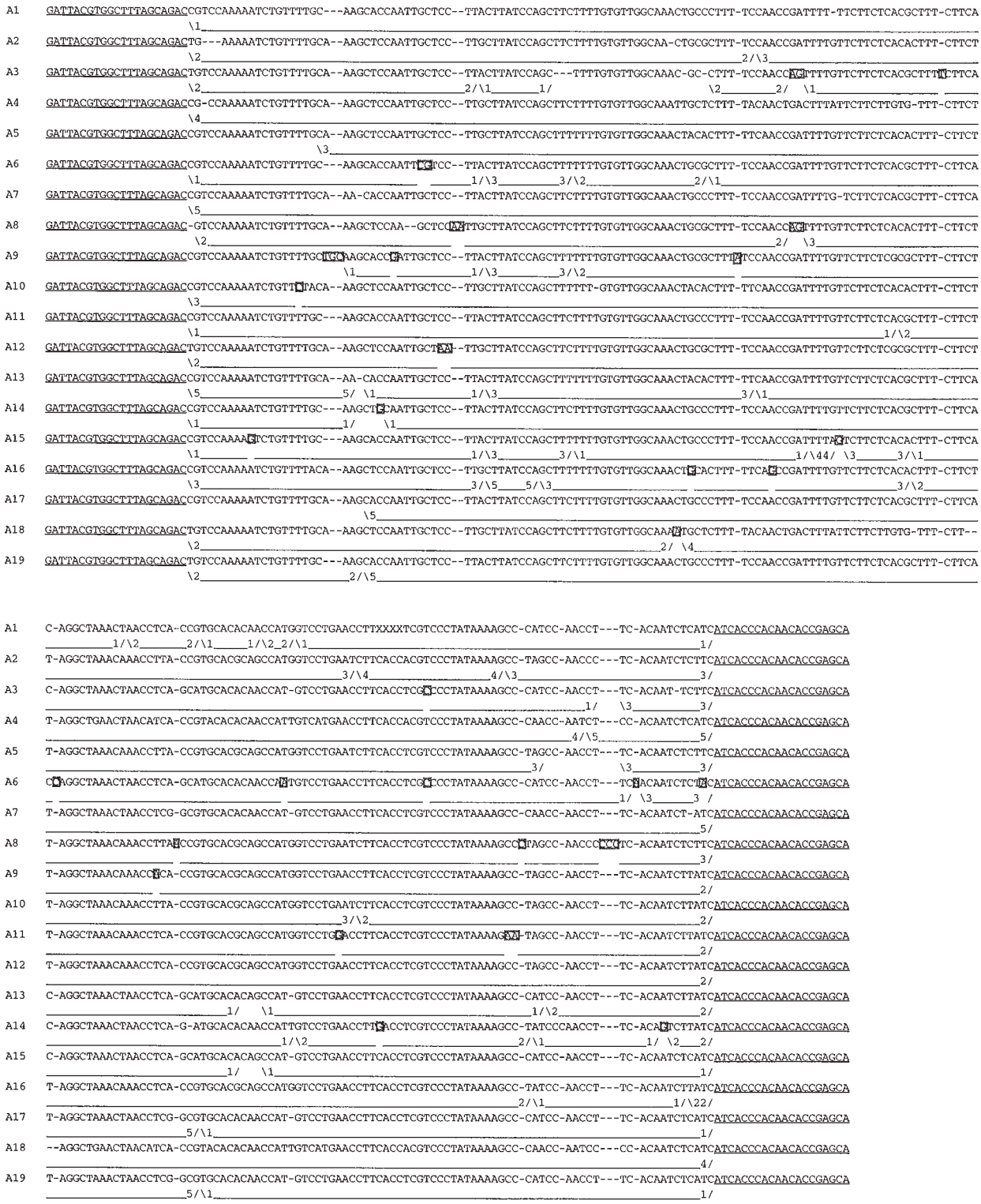


Fig. 5. Multiple alignment of the 42 partial glutenin sequences listed in Table 1. The individual sequences are placed in similarity groups. Dots indicate identity with the *Glu-D1-2b* allele, dashes indicate deletion or insertion points, and the primer annealing sites are underlined. The boxes indicate the 12 instances of homoplasy revealed by compatibility analysis.

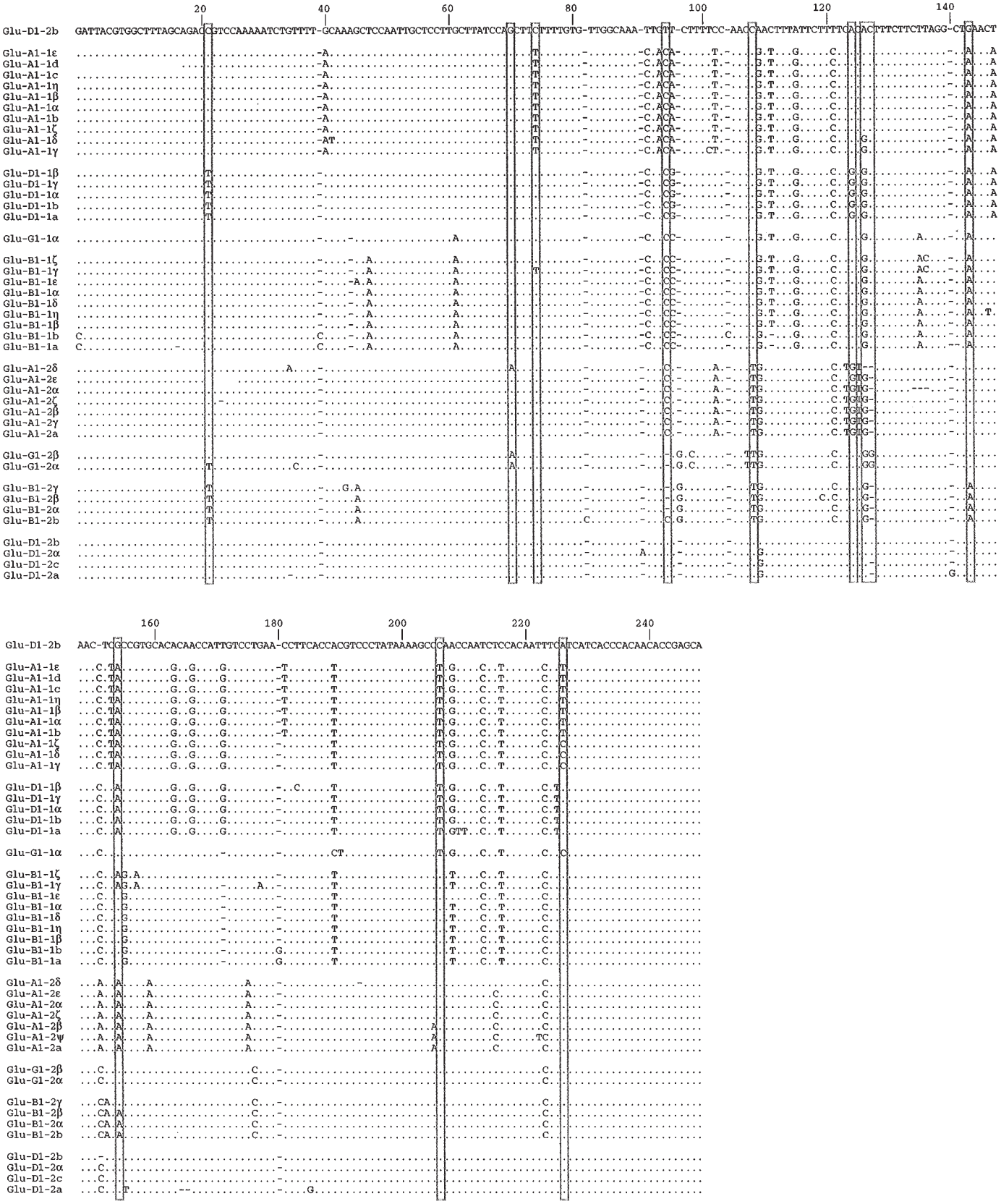
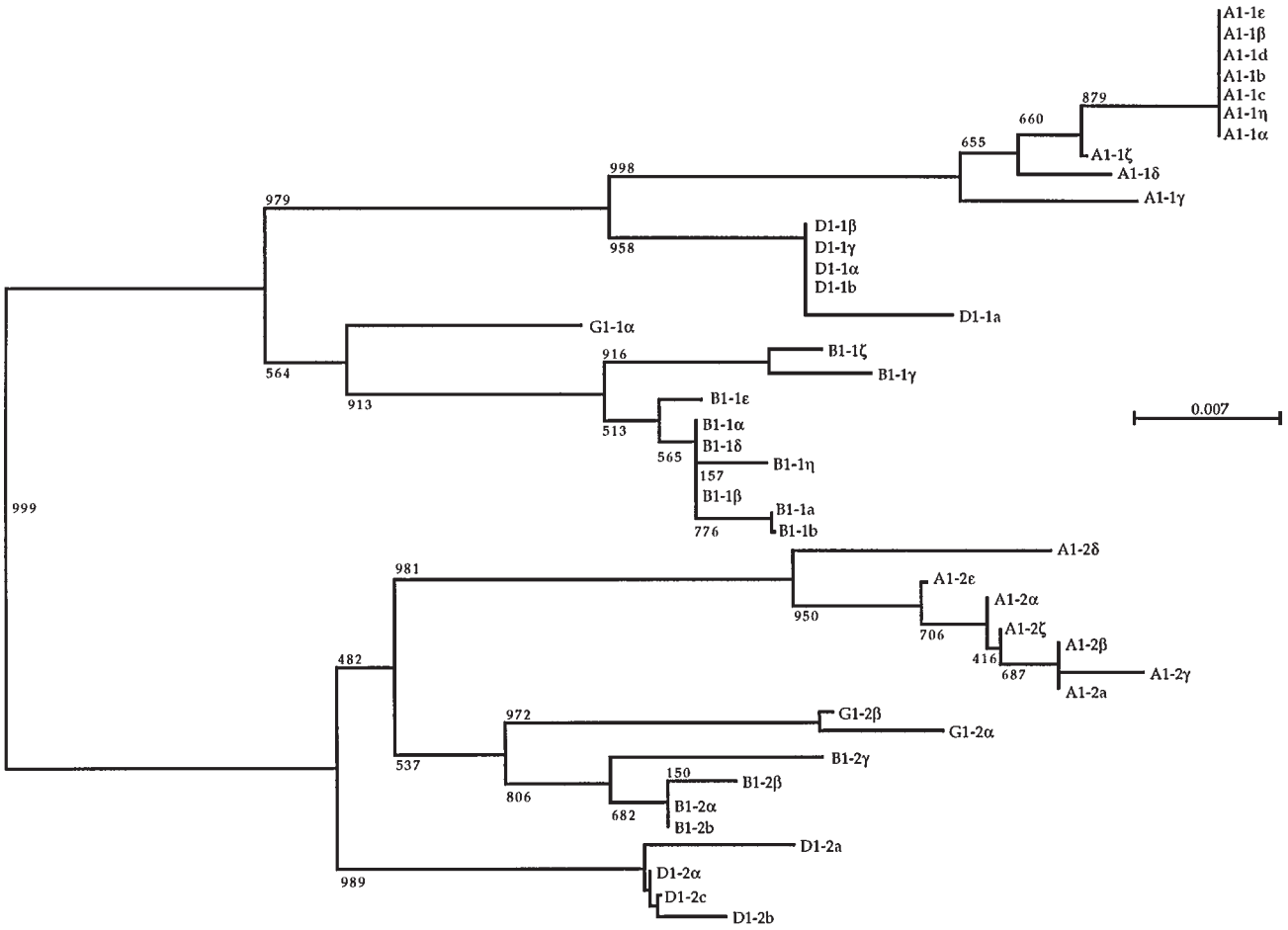


Fig. 6. Neighbour-joining tree constructed from the multiple alignment shown in Fig 5. The bootstrap values that are shown were obtained by constructing 1000 replicate trees. In the absence of a suitable outgroup, the two paralogous groups in the tree are used as outgroups for one another (e.g., Iwabe et al. 1989).



cum tragi and *T. turgidum*) and AABBDD hexaploids (*T. aestivum* varieties *Asarce*, *Cheyenne*, *Galahad*, and *Hope*). There are no nucleotide differences in this subgroup, and an identical sequence was also obtained from the 3000 year old archaeological assemblage. If the prevalence of this allele in cultivated wheats indicates that it has been subject to artificial selection pressure then this pressure is presumably not still acting, as the locus is not expressed in all modern bread wheats. It is active in *Cheyenne* (Harberd et al. 1986), produces a truncated protein in *Asarce* (Xin et al. 1992), and is not expressed at all in *Chinese Spring* (Thompson et al. 1983).

Compatibility of the phylogenetic tree with the origins of the wheat genomes

Support for the veracity of the glutenin phylogenetic tree is provided by the compatibility between its topology and the postulated origins of the A, B, D, and G wheat genomes. For example, the positions in the *Glu-A1-1* and *Glu-A1-2* clades of the *T. urartu* and *T. monococcum* sequences, with *T. urartu* adjacent to the AABB and AABBDD sequences and *T. monococcum* relatively distant, is in agreement with the identification of *T. urartu* rather than *T. monococcum* as the A genome donor for these tetraploid and hexaploid wheats (Dvorák et al. 1993). Similarly, the relative positioning of the *Glu-A1-1* sequences from *T. urartu*, *T. monococcum*, and *T. timopheevi* is consistent with *T. monococcum* being the A genome donor for the AAGG tetraploid *T. timopheevi* (Dvorák et al. 1993): the presence of a G at position 128 in the *Glu-A1-1*γ *and Glu-A1-1*δ sequences, but A at this position in *Glu-A1-1*ζ, is particularly suggestive in this respect.

The B and G genomes were both donated by members of the *Sitopsis* section of the *Aegilops* genus. *A. speltoides* is generally accepted as the G genome donor (Gill and Appels 1988), but the origin of B is less clear, this genome having traits in common with both *A. speltoides* and other members of the *Sitopsis* section (Gill and Chen 1987; Miller 1987; Daud and Gustafson 1996). The positioning within the glutenin tree of the G alleles from *T. timopheevi* in outgroup positions relative to the *Glu-B1-1* and *Glu-B1-2* clades reflects this interpretation of the B and G genomes as being closely related, but distinct. The D genome of AABBDD hexaploids was donated by *A. squarrosa* (Miller 1987), and as expected the *Glu-D1-1* and *Glu-D1-2* sequences from this grass display close similarity with the D genome sequences from hexaploids.

Inferences regarding the evolution of the HMW glutenin loci

The topology of the glutenin tree allows inferences to be made about the time of duplication of the *Glu-1-1* and *Glu-1-2* loci and the divergence times of the A, B, D, and G

genomes. We used a relative rate test (Li 1997) to examine the variance of evolutionary rate between gene lineages (Table 3). This test showed that overall there was no significant variance between lineages with the exception of *Glu-A1-1*. On the basis of this, we concluded that it was valid to use the average genetic distances between clades for approximate dating purposes. In attempting this analysis we would ideally have used a substitution rate specific for the 5′-flanking regions of plant nuclear genes, but unfortunately such a figure has not been reported. As a best alternative we have used the rate of 5.1–7.1 × 10^{-9} substitutions/site/year calculated by Wolfe et al. (1989) for synonymous substitutions in the coding regions of plant nuclear genes. It has been suggested that a synonymous substitution rate is likely to be slightly higher than the substitution rate for 5′-flanking regions (Li and Graur 1991), so the divergence times that we derive from the glutenin tree should be treated as low estimates.

Two divergence times can be calculated from components of the glutenin tree that display a reasonable degree of robustness. The first of these is the time of the gene duplication event. The average of the mean pairwise differences for all combinations of *Glu-1-1* and *Glu-1-2* clades (Table 2) is 0.1025, implying a gene duplication date of 7.2–10.0 million years ago (MYA).

The second "robust" calculation concerns the time of origin of the A, B, D, and G genomes. The average of the mean pairwise differences for all combinations of *Glu-1-1* clades (*Glu-A1-1*/*Glu-B1-1*, *Glu-A1-1*/*Glu-B1-1*, etc) is 0.0703 giving a date for the origin of the four genomes of 5.0–6.9 MYA. In the *Glu-1-2* lineage there is some doubt about the positioning of the *Glu-B1-2*/*Glu-G1-2* clades relative to the origin, so the estimate can be based only on the mean pairwise difference between the *Glu-A1-2* and *Glu-D1-2* clades, which at 0.0657 gives a date of 4.6–6.4 MYA, in good agreement with the figure calculated from *Glu-1-1*.

Other calculations of divergence times are less reliable without the corrective effect of averaging the rates of lineages. However, the *Glu-1-1* and *Glu-1-2* lineages agree on the split of the B and G genomes (2.5–3.5 MYA). Sasanuma et al. (1996) used RFLP analysis to divide the *Aegilops* genus into two sections, one containing *A. speltoides* (the donor of the wheat G genome) and the other containing *A. searsii*, *A. longissima*, *A. sharonensis*, and *A. bicornis* (one of which might be the B genome donor). They dated the divergence time for these two sections at 1.4–8.4 MYA, but the upper limit that they used for the substitution rate (3.0 × 10^{-8} substitutions/site/year) was derived from the split between spinach and *Silene* which, being based on suspect paleontological data, may be too high (Wolfe et al. 1989). Recalibration of the RFLP data with the upper substitution rate used in this paper gives an adjusted figure of 6.0–8.4 MYA, which is less similar to the figure for the B/G split calculated from the glutenin tree.

Table 3. Relative rate tests.

	Glu-A1-1	*Glu-D1-1*	*Glu-B1-1*	*Glu-G1-1*
A. *Glu-1-1* lineages				
Glu-A1-1	—	0.0176	0.0340	0.0329
Glu-D1-1	0.0217	—	0.0164	0.0153
Glu-B1-1	0.0288	0.0245	—	0.0011
Glu-G1-1	0.0264	0.0249	0.0182	—
	Glu-A1-2	*Glu-D1-2*	*Glu-B1-2*	*Glu-G1-2*
B. *Glu-1-2* lineages				
Glu-A1-2	—	0.0099	0.0228	0.0046
Glu-D1-2	0.0253	—	0.0129	0.0053
Glu-B1-2	0.0250	0.0219	—	0.0182
Glu-G1-2	0.0247	0.0225	0.0188	—

Note: Differences (*d*) in K values were calculated between lineages and an outgroup (*Glu-D1-2* in the case of the *Glu-1-1* lineages, and *Glu-D1-1* for *Glu-1-2* lineages). These values are shown above the diagonal. Values for twice the standard error (S.E.) of *d* are shown below the diagonal. Rates of nucleotide change are significantly different between lineages if $d > 2$ (S.E.) (Li 1997).

Implications for the origin of cultivated wheats

An interesting feature of the glutenin tree is the apparent split of the *Glu-B1-1* clade into two subgroups, one comprising allele sequences from the 3000 year old archaeological wheat (*Glu-B1-1*ζ) and *T. dicoccum tragi* (*Glu-B1-1*γ), the second including seven sequences (*Glu-B1-1*ε, *Glu-B1-1*α, *Glu-B1-1*δ, *Glu-B1-1*η, *Glu-B1-1*β, *Glu-B1-1a*, *Glu-B1-1b*). The bootstrap value for the branchpoint between these two subgroups is 913/1000, suggesting that the split is real. The estimated divergence time for these two subgroups is 1.4–2.0 MYA, an unexpected finding as all the wheats in the two subgroups are cultivated varieties and presumably trace their origins to the first domestication of wheat some 10 000 years ago. There are three possible explanations for this observation. The first is that emmer wheat, *T. dicoccum* (from which the hexaploids are derived), was domesticated twice, the two domestication events involving wild progenitors with different genotypes. This interpretation is contrary to current thinking in archaeobotany which states that domestication of each of the founder crops of Southwest Asian agriculture, with the possible exception of barley, was a unique event (Zohary 1996). It is, however, generally accepted that with a few exceptions (notably einkorn, Heun et al. 1997) the existing evidence regarding the numbers of domestications is weak. The second possibility is that emmer was domesticated only once but that the original domesticates included plants with different *Glu-B1-1* genotypes. This would require that the original cultivated emmers were taken from a wild population that included both *Glu-B1-1* subgroups, and that representatives of both genotypes acquired mutations leading to modified seed dispersal (the underlying basis to domestication) during the period when the cultivated crop was becoming domesticated. Whether or not this scenario is possible depends on the harvesting techniques employed by the first farmers, which are not understood (Hillman and Davies 1990). The third explanation is that emmer wheat was domesticated once, but subsequently acquired a second *Glu-B1-1* subgroup by introgression from a wild emmer. The likelihood that this hypothesis is correct is reduced by the fact that emmer wheat is predominantly self-pollinating, and by archaeological and ecological considerations which suggest that significant introgression of wild genes into cultivated wheats did not occur (Flannery 1969).

The glutenin tree also has interesting implications for the origins of the modern bread wheats. These AABBDD wheats are thought to have arisen by hybridization between

T. dicoccum (AABB) and *A. squarrosa* (DD), the event presumably occurring since the beginning of wheat cultivation as the natural ranges of the two species do not overlap. This recent origin of the AABBDD wheats (not earlier than 9000 years ago) is reflected by the topology of the *Glu-D1-1* clade, which is comprised of very similar sequences, but not by the topology within *Glu-D1-2*, which displays the greatest intracladal differences of all eight glutenin clades. It seems unlikely that the heterogeneity within *Glu-D1-2* could have evolved in just 9000 years, suggesting that the hybridization leading to the AABBDD hexaploids might have occurred on multiple occasions. Interestingly, we have also observed that AABBDD hexaploids display a relatively high degree of heterogeneity in the rDNA loci on the B chromosome set (unpublished data).

Conclusions

We have shown that partial sequences of the HMW glutenin loci can be obtained by PCR followed by sequencing of cloned PCR products. This procedure is applicable not only to modern specimens but also to those archaeological specimens from which ancient DNA can be extracted. Although the amplified region is relatively short, it can provide phylogenetic data relating to the evolution of these loci and can also be used to make inferences about the origins and divergence times of the individual wheat genomes.

Acknowledgements

We thank Angela Schlumbaum (University of Basel) for allowing us to use her unpublished glutenin sequences. We also thank Terry Miller (John Innes Centre), Marcus Hamer (University of Manchester), and Glynis Jones (University of Sheffield) for providing wheat specimens, and Des Higgins (Trinity College, Dublin) for advice regarding phylogenetic analysis. This work was supported in part by a grant from the Ancient Biomolecules Initiative of the Natural Environment Research Council and by a studentship from the same source awarded to M.B.

References

Allaby, R.G., Jones, M.K., and Brown, T.A. 1994. DNA in charred wheat grains from the Iron Age hillfort at Danebury, England. Antiquity, **68**: 126–132.

Allaby, R.G., O'Donoghue, K., Sallares, R., Jones, M.K., and Brown, T.A. 1997. Evidence for the survival of ancient DNA in charred wheat seeds from European archaeological sites. Ancient Biomol. **1**: 119–129.

Anderson, O.D., and Greene, F.C. 1989. The characterization and comparative analysis of high-molecular-weight glutenin genes from genomes A and B of a hexaploid bread wheat. Theor. Appl. Genet. **77**: 689–700.

Anderson, O.D., Greene, F.C., Yip, R.E., Halford, N.G., Shewry, P.R., and Malpica-Romera, J.-M. 1989. Nucleotide sequences of two high-molecular-weight glutenin genes from the D-genome of a hexaploid bread wheat, *Triticum aestivum* L. cv Cheyenne. Nucleic Acids Res. **17**: 461–462.

Bell, G.D.H. 1987. The history of wheat cultivation. *In* Wheat breeding — its scientific basis. *Edited by* F.G.H. Lupton. Chapman and Hall, London. pp. 31–49.

Brown, T.A., Allaby, R.G., Sallares, R., and Jones, G. 1998. Ancient DNA in charred grains: Taxonomic identification of mixed and single grains. Ancient Biomol. **2**: 185–193.

Ciaffi, M., Lafiandra, D., Porceddu, E., and Benedettelli, S. 1993. Storage protein variation in wild emmer (*Triticum turgidum* ssp. *dicoccoides*) from Jordan and Turkey. II. Patterns of allele distribution. Theor. Appl. Genet. **86**: 518–525.

Daud, H.M., and Gustafson, J.P. 1996. Molecular evidence for *Triticum speltoides* as a B-genome progenitor of wheat (*Triticum aestivum*). Genome, **39**: 543–548.

Dong, H., Cox, T.S., Sears, R.G., and Lookhart, G.L. 1991. High molecular weight glutenin genes: Effects on quality in wheat. Crop Sci. **31**: 974–979.

D'Ovidio, R., and Anderson, O.D. 1994. PCR analysis to distinguish between alleles of a member of a multigene family correlated with bread-making quality. Theor. Appl. Genet. **88**: 759–763.

D'Ovidio, R., Porceddu, E., and Lafiandra D. 1994. PCR analysis of genes encoding allelic variants of high-molecular-weight glutenin subunits at the *Glu-D1* locus. Theor. Appl. Genet. **88**: 175–180.

Dvorák, J., Di Terlizzi, P., Zhang, H.-B., and Resta, P. 1993. The evolution of polyploid wheats: Identification of the A genome donor species. Genome, **36**: 21–31.

Eickbush, T.H., and Burke, W.D. 1985. Silkmoth chorion gene families contain patchwork patterns of sequence homology. Proc. Natl. Acad. Sci. U.S.A. **82**: 2814–2818.

Erlich, H.A., Gelfand, D., and Sninsky, J.J. 1991. Recent advances in the polymerase chain reaction. Science, **252**: 1643–1651.

Estabrook, G.F. 1978. Some concepts for the estimation of evolutionary relationships in systematic botany. Syst. Bot. **3**: 146–158.

Felsenstein, J. 1989. PHYLIP — Phylogeny Inference Package (Version 3.20). Cladistics, **5**: 164–166.

Flannery, K. 1969. Origins and ecological effects of early domestication in Iran and the Near East. *In* The domestication and exploitation of plants and animals. *Edited by* P.J. Ucko and G.W. Dimbleby. Duckworth, London. pp. 73–100.

Flavell, R.B., Goldsborough, A.P., Robert, L.S., Schnick, D., and Thompson, R.D. 1989. Genetic variation in wheat HMW glutenin subunits and the molecular basis of bread-making quality. Biotechnology, **7**: 1281–1285.

Forde, J., Malpica, J.-M., Halford, N.G., Shewry, P.R., Anderson, O.D., Greene, F.C., and Miflin, B.F. 1985. Nucleotide sequence of a HMW glutenin subunit gene located on chromosome 1A of wheat (*Triticum aestivum* L.). Nucleic Acids Res. **13**: 6817–6832.

Gill, B.S., and Chen, P.D. 1987. Role of cytoplasm specific introgression in the evolution of the polyploid wheats. Proc. Natl. Acad. Sci. U.S.A. **84**: 6800–6804.

Gill, B.S., and Appels, R. 1988. Relationships between *Nor*-loci from different *Triticeae* species. Plant Syst. Evol. **160**: 77–89.

Goloubinoff, P., Pääbo, S., and Wilson, A. 1993. Evolution of maize inferred from sequence diversity of an *adh2* gene segment from archaeological specimens. Proc. Natl. Acad. Sci. U.S.A. **90**: 1997–2001.

Halford, N.G., Forde, J., Anderson, O.D., Greene, F.C., and Shewry, P.R. 1987. The nucleotide and deduced amino acid sequence of an HMW glutenin subunit gene from chromosome 1B of bread wheat (*Triticum aestivum* L.) and comparison with those of genes from chromosomes 1A and 1D. Theor. Appl. Genet. **75**: 17–126.

Halford, N.G., Tatham, A.S., Sui, E., Daroda, L., Dreyer, T., and Shewry, P.R. 1992. Identification of a novel ß-turn-rich repeat

motif in the D hordeins of barley. Biochim. Biophys. Acta, **1122**: 118–122.

Handt, O., Höss, M., Krings, M., and Pääbo, S. 1994. Ancient DNA: Methodological challenges. Experientia, **50**: 524–529.

Harberd, N.P., Bartels, D., and Thompson, R.D. 1986. DNA restriction-fragment variation in the gene family encoding high molecular weight (HMW) glutenin subunits of wheat. Biochem. Genet. **24**: 579–596.

Hawkes J.G. 1991. The importance of genetic resources in plant breeding. *In* Genetic conseravtion of world crop plants. *Edited by* J.G. Hawkes. Linnean Society, London. pp. 3–10.

Heun, M., Schäfer-Pregl, R., Klawan, D., Castagna, R., Accerbi, M., Borghi, B., and Salamini, F. 1997. Site of einkorn wheat domestication identified by DNA fingerprinting. Science, **278**: 1312–1314.

Higgins, D.G. 1992. Sequence ordinations: A multivariate analysis approach to analysing large sequence data sets. CABIOS, **8**: 15–22.

Hillman, G.C., and Davis, M.S. 1990. Measured domestication rates in wild wheats and barley under primitive cultivation, and their archaeological implications. J. World Prehist. **4**: 157–222.

Iwabe, N., Kuma, K.-I., Hasegawa, M., Osawa, S., and Miyata, T. 1989. Evolutionary relationships of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. U.S.A. **86**: 9355–9359.

Jones, G., Wardle, K.A., and Wardle, D. 1986. Crop storage at Assiros. Sci. Amer. **254**(3): 95–103.

Li, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Mass. pp. 216–218.

Li, W.-H., and Graur, D. 1991. Fundamentals of molecular evolution. Sinauer, Sunderland, Mass. pp. 72–73.

Mackie, A.M., Sharp, P.J., and Lagudah, E.S. 1996. The nucleotide and derived amino acid sequence of a HMW-glutenin gene from *Triticum tauschii* and comparison with those from the D-genome of bread wheat. J. Cereal Sci. **24**: 73–78.

Miller, T.E. 1987. Systematics and evolution. *In* Wheat breeding — its scientific basis. *Edited by* F.G.H. Lupton. Chapman and Hall, London. pp. 1–30.

O'Donoghue, K., Clapham, A., Evershed, R.P., and Brown, T.A. 1996. Remarkable preservation of biomolecules in ancient radish seeds. Proc. R. Soc. Lond. Ser. B, **263**: 541–547.

Pääbo, S. 1989. Ancient DNA: Extraction, characterization, molecular cloning and enzymatic amplification. Proc. Natl. Acad. Sci. U.S.A. **86**: 1939–1943.

Pääbo, S., Irwin, D.M., and Wilson, A.C. 1990. DNA damage promotes jumping between templates during enzymatic amplification. J. Biol. Chem. **265**: 4718–4721.

Payne, P.I., Holt, L.M., Worland, A.J., and Law, C.N. 1982. Structural and genetic studies on the high-molecular-weight subunits of wheat glutenin. 3. Telocentric mapping of the subunit genes on the long arms of the homeologous group 1 chromosomes. Theor. Appl. Genet. **63**: 129–138.

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

Reddy, P., and Appels, R. 1993. Analysis of a genomic DNA segment carrying the wheat high-molecular-weight (HMW) glutenin Bx17 subunit and its use as an RFLP marker. Theor. Appl. Genet. **85**: 616–624.

Rollo, F., Venanzi, F. M., and Amici, A. 1991. Nucleic acids in mummified plant seeds: Biochemistry and molecular genetics of pre-Columbian maize. Genet. Res. **58**: 193–201.

Sasanuma, T., Miyashita, N.T., and Tsunewaki, K. 1996. Wheat phylogeny determined by RFLP analysis of nuclear DNA. Intra- and interspecific variations in five *Aegilops* Sitopsis species. Theor. Appl. Genet. **92**: 928–934.

Shewry, P.R., Tatham, A.S., Baro, F., Barcelo, P., and Lazeri, P. 1995. Biotechnology of breadmaking: Unravelling and manipulating the multi-protein gluten complex. Biotechnology, **13**: 1185–1190.

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

Thompson, R.D., Bartels, D., Harberd, N.P., and Flavell, R.B. 1983. Characterisation of the multigene family coding for HMW glutenin subunits in wheat using cDNA clones. Theor. Appl. Genet. **67**: 87–96.

Thompson, R.D., Bartels, D., and Harberd, N.P. 1985. Nucleotide sequence of a gene from chromosome 1D of wheat encoding a HMW-glutenin subunit. Nucleic Acids Res. **13**: 6833–6846.

Thompson, J.D., Higgins, D., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**: 4673–4680.

Towner, P. 1991. Recovery of DNA from electrophoresis gels. *In* Essential molecular biology: A practical approach. *Edited by* T.A. Brown. IRL Press and Oxford University Press, Oxford. pp.127–141.

Wolfe, K., Sharp, P.M., and Li, W.-H. 1989. Rates of synonymous substitution in plant nuclear genes. J. Mol. Evol. **29**: 208–211.

Xin, H., Huang, W., Gu, Q., and Sun, C. 1992. Nucleotide sequence of a high-molecular-weight glutenin gene in wheat. Acta Bot. Sin. **34**: 729–735.

Zohary, D. 1996. The mode of domestication of the founder crops of Southwest Asian agriculture. *In* The origins and spread of agriculture and pastoralism in Eurasia. *Edited by* D.R. Harris. UCL Press, London. pp. 142–158.