

EST mining for structure and expression of genes in the region of the wheat high-molecular-weight glutenin loci

O.D. Anderson

Abstract: An in-depth analysis was carried out with expressed sequence tags (ESTs) for genes in and near the HMW-GS loci. Considerations for using ESTs are discussed, including the occurrence of chimeric and aberrant HMW-GS ESTs. Complete gene sequences demonstrated the feasibility of constructing accurate full-length coding regions from EST assemblies and found, or supported, errors in several previously reported HMW-GS gene sequences. New complete HMW-GS gene sequences are reported for the cultivars Chinese Spring and Glenlea. The Ay subunit gene, which is considered null in cultivated wheats, was shown to transcribe in at least two germplasms. Analyses support the conclusion that of the five known genes within this genomic region, the two HMW-GS genes and the globulin gene are highly expressed. The other two genes, encoding a receptor kinase and a protein kinase, have one and no identifiable wheat EST, respectively, although ESTs are found for the orthologous genes in barley. The ESTs of all five genes within the HMW-GS region are either definitely associated with the endosperm or possibly originate from imbibed seed, suggesting the four distinct gene classes in this region are part of a seed or endosperm chromatin domain. EST resources were also used to determine relative abundance of ESTs for all classes of wheat prolamines and indicated differential levels of expression both among germplasms and among the three genomes of hexaploid wheats.

Key words: wheat, glutenin, genes, expressed sequence tags, *Triticum aestivum*.

Résumé : Une analyse approfondie a été réalisée sur les étiquettes de séquences exprimées (EST) provenant des gènes situés dans ou à proximité des locus HMW-GS. Les raisons motivant l'emploi des EST sont discutées, incluant la présence d'EST chimériques et aberrants parmi les EST des locus HMW-GS. Des séquences nucléotidiques complètes ont montré la possibilité d'assembler des régions codantes complètes à partir d'assemblages d'EST et ont permis de trouver ou de confirmer des erreurs au sein de plusieurs des séquences rapportées antérieurement. De nouvelles séquences nucléotidiques complètes des gènes HMW-GS sont rapportées pour les cultivars Chinese Spring et Glenlea. Il a été montré que le gène Ay du locus HMW-GS, considéré comme nul chez les blés cultivés, était transcrit chez aux moins deux germoplasmes. Les analyses confirment que parmi les cinq gènes situés dans cette région du génome, les deux gènes HMW-GS et celui codant pour la globuline sont fortement exprimés. Les deux autres gènes, codant pour une récepteur kinase et une protéine kinase, présentent respectivement un et aucun EST chez le blé, bien que des EST aient été trouvés pour leur orthologues chez l'orge. Les EST des cinq gènes de la région HMW-GS sont soit définitivement associés à l'albumen ou proviennent possiblement de graines imbibées. Cela suggère que les quatre classes de gènes représentés au sein de cette région font partie d'un domaine de la chromatine qui est spécifique de la graine ou de l'albumen. Les banques d'EST ont également été employées pour déterminer l'abondance relative des EST pour toutes les classes de prolamines du blé. Ces analyses ont indiqué des niveaux différents d'expression tant parmi les lignées que parmi les trois génomes des blés hexaploïdes.

Mots-clés : blé, gluténine, gènes, étiquettes de séquences exprimées, *Triticum aestivum*.

[Traduit par la Rédaction]

Introduction

The wheat seed proteins are composed mainly of prolamines: proteins with high proline plus glutamine composition that have the distinctive physicochemical property of being soluble in dilute alcohols (Osborne 1907). While prolamines are found in many monocots, their percentage of the total seed protein composition varies from relatively minor

components in oats to the majority portion of seed proteins in the tribe Triticeae. In wheat, prolamines account for about 70% of the seed protein (Osborne 1907) and are historically divided into 2 classes, glutenins and gliadins (Shewry and Halford 2002). One subclass of the glutenins, the high-molecular-weight glutenin subunits (HMW-GS), comprise approximately 10% of the seed protein (2% of the total seed mass) but are the most important seed protein com-

Received 22 December 2008. Accepted 22 April 2009. Published on the NRC Research Press Web site at genome.nrc.ca on 7 August 2009.

Corresponding Editor: P. Gulick.

O.D. Anderson. Genomics and Gene Discovery Research Unit, Western Regional Research Center, Agricultural Research Service, US Department of Agriculture, 800 Buchanan Street, Albany, CA 94710, USA (e-mail: olin.anderson@ars.usda.gov).

ponent because of their unique and pivotal role in wheat dough viscoelasticity (MacRitchie et al. 1990). Because of their importance, the genes for the HMW-GS have been the most studied of all the wheat genes, and the region of the wheat genome containing those genes has been studied to a greater degree than any other site in the wheat genome (Gu et al. 2006). Two tightly linked HMW-GS genes form the compound *Glu-1* locus in each wheat genome, and encode the x- and y-type HMW-GS. Thus, there are 6 HMW-GS genes in hexaploid bread wheat — a gene family much smaller than the large multigene families typical of other prolamine classes (e.g., the 100+ genes of the α -gliadin family; Anderson et al. 1997).

The main features of the HMW-GS are a large repetitive domain composed of tandem 6, 9, and 15 amino acid motifs, and short non-repetitive domains at the N- and C-termini (Anderson and Greene 1989; Shewry and Halford 2002). There are sufficient differences in the repetitive domains and termini of the HMW-GS and their genes to allow unambiguous gene and genome assignments for expressed sequence tag (EST) and genomic DNA sequences.

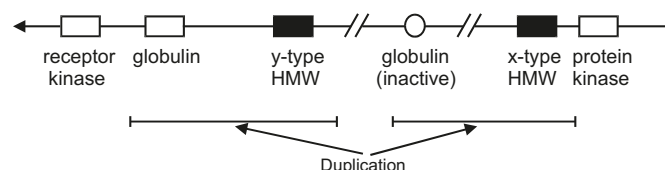
The *Glu-1* locus has been isolated and sequenced from a range of Triticeae germplasms including the D genome of diploid *Triticum tauschii*, donor of the D genome of hexaploid wheat (Anderson et al. 2003); the A and B genomes of tetraploid *Triticum dicoccoides* (Kong et al. 2003; Gu et al. 2004); the A, B, and D genomes of hexaploid *Triticum aestivum* (Gu et al. 2006); and the orthologous D-hordein locus of barley (Gu et al. 2003). The complement and order on the chromosome of 6 potential genes around the *Glu-1* loci are conserved in all wheat genomes studied thus far and are shown in Fig. 1: genes for a receptor kinase, globulin, y-type HMW-GS, globulin remnant, x-type HMW-GS, and protein kinase.

Structural genome studies are relatively straightforward compared with analyses of gene transcription. For example, although it is known that certain HMW-GS alleles are more correlated with quality characteristics than other alleles (Békés et al. 2006), theories on the molecular basis of the difference have focused on the allele sequences, while relatively little attention has been given to the possibility that differential rates of mRNA production are sources of quality differences. Among the different approaches to assessing wheat transcriptional activity, an under-utilized resource is the over one million ESTs in GenBank. Under specific conditions, ESTs can reflect the mRNA titer and thus can be a reasonable measure of gene activity — the necessary conditions being that ESTs are assignable to unique gene sequences to avoid paralogs confounding the analyses, and that sufficient ESTs exist for target genes to allow statistically significant comparative conclusions.

In addition to yielding information on gene transcription, EST analyses can contribute to structural analysis by assembling sequences of genes not previously cloned. If sufficient numbers of overlapping ESTs are available from unique genes, then there is no discrepancy between assembling a sequence from subclones of a cloned gene or from ESTs as long as there is reasonable assurance that the assembled ESTs represent a single gene.

In the present report, the wheat EST collection was mined for information on gene structure and relative transcriptional

Fig. 1. Gene order around the *Glu-1* loci. The conserved structure in the genomic region of the *Glu-1* loci is diagrammed. HMW-GS genes are indicated by black boxes and other genes by white boxes as labeled. The white circle is a globulin gene remnant. Variable gene spacings due to transposon clusters are shown by paired slashes. An ancestral duplication is indicated. The arrowhead pointing to the left indicates that additional genes, not shown, are known in at least one wheat genome. The single slash at the right end of the chromosome sequence indicates that nothing is known of genes in that direction.



activity for genes in the region of the wheat *Glu-1* locus. New complete gene sequences are reported, along with corrections to previously published sequences, transcription of what had been considered null HMW-GS genes, evidence of differential expression of HMW-GS and neighboring genes, differences in relative transcription among germplasms, and comparison of EST titers for all prolamine classes across different germplasms. These analyses both add to knowledge of this important wheat genomic region and point to future research perhaps not previously appreciated.

Materials and methods

Total EST analyses were accomplished using direct querying to GenBank (<http://www.ncbi.nlm.nih.gov>). For individual germplasm queries, EST entries were downloaded from GenBank and placed within separate databases at the “wEST” site (<http://wheat.pw.usda.gov/wEST>) associated with the GrainGenes database (<http://wheat.pw.usda.gov/GG2/index.shtml>). Within the wEST site and GrainGenes are facilities for downloading ESTs from specific wheat germplasms and blast querying those germplasm-specific ESTs. Empirical tests were used to identify BLAST cut-off expectation values for genuine orthologous sequences. ESTs were assigned to specific wheat prolamine classes based on the best BLAST match. Assignments to the HMW-GS class were unambiguous, since the sequence of this class is significantly different from the sequences of other classes. The gliadins and low-molecular-weight glutenin subunits (LMW-GS) have sequences similar enough to each other that some uncertainty can exist with ESTs of short length and (or) poorer quality sequence. There was usually a range of BLAST matches with borderline scores between those ESTs with definitive assignments and ESTs from other prolamine classes. In these cases, 10–30 ESTs at this boundary were manually inspected to confirm assignments. For the HMW-GS, differentiation of ESTs among the 6 genes is unambiguous except for very short ESTs, but in the current analysis all ESTs of less than 200 base pairs were excluded. Sequences for BLAST probes to GenBank are as follows: α -gliadin, U08287; γ -gliadin, AF234646; ω -gliadin, AF280605; LMW-GS, U86028; HMW-GS, X12929; receptor kinase, globulin, and protein kinase sequences from *T. tauschii* BAC DQ537337.

ESTs were assembled into contigs using known gene sequences, when available, as scaffolds for the assemblies. ESTs were assembled using Clustal V and Clustal W implementations in the MegAlign and SeqMan software modules of the Lasergene II package (DNASTAR, Inc.), followed by visual inspection of alignments, manual trimming of terminal vector sequences, and alignment adjustments when necessary. Chimeric sequences were identified either by individually analyzing ESTs that joined alignments but included a sequence region that totally diverged from the rest of the alignment, or by identifying ESTs that did not join contigs but that BLAST confirmed contained target sequences. Such chimeric sequences were counted in EST totals only when the sequence origin could be clearly identified.

Newly reported nucleotide sequence data are available in the Third Party Annotation (TPA) section of the DDBJ/EMBL/GenBank database under the TPA accession Nos. BK006459 (Dy12 of *T. aestivum* 'Chinese Spring'), BK006460 (Dx2 of Chinese Spring), BK006773 (Bx7 of Chinese Spring), and BK006458 (Dx5 of *T. aestivum* 'Glenlea'). The ESTs used in the assemblies are given in the TPA entries. For information about the TPA database, see <http://www.ncbi.nlm.nih.gov/Genbank/TPA.html>.

For easier reading, the common names for HMW-GS will be used to refer to the HMW-GS genes and subunits in place of formal genetic nomenclatures, i.e., "Bx7 subunit" instead of "*Glu-B1-1a* subunit" and "Bx7 gene" instead of "*Glu-B1-1a* gene". In addition, it will be considered understood that a reference such as "Bx" will refer to a version of the DNA or protein sequence of the *Glu-B1* gene without reference to a specific allele or subunit.

Results and discussion

The HMW-GS gene region

The organization of the genes within the genomic regions containing the HMW-GS genes is shown in Fig. 1. The two HMW-GS genes arose from a tandem duplication of a DNA segment containing a HMW-GS gene and a seed-specific globulin gene (Kong et al. 2003). The globulin gene between the two HMW-GS genes has been inactivated in all known sequences because of extensive deletions, insertions, and sequence rearrangements (Gu et al. 2006). Flanking the duplication are receptor kinase and protein kinase genes, and the pre-duplication order of genes is conserved in barley (Gu et al. 2003), i.e., receptor kinase, globulin, D-hordein (HMW-GS ortholog), and protein kinase.

The utility and caveats of utilizing ESTs for gene analysis

Determination of a DNA sequence generally involves multiple reads of sequence across all regions of the target DNA. Errors in individual sequence reads are cancelled out by taking the consensus sequence from aligning sufficient reads. The same strategy can be used with EST collections to reconstruct original gene coding sequences, with the limitation that sufficient EST sequences are necessary to deter-

mine the original contiguous sequence with multiple reads across all regions of the sequence. In addition, the sequences must be distinct enough to rule out assembly artifacts from closely related paralogous and orthologous genes.

There are also 3 basic assumptions or requirements when using EST numbers as reflections of relative gene activity in hexaploid wheats, particularly when comparing orthologs. One assumption is that the EST representation is effectively random, since all the genes exist within the same hexaploid nuclei; i.e., there is no bias of cDNA recovery from mRNA among the 3 wheat genomes. The second assumption is that there is no bias in the cloning and sequencing significant enough to invalidate the analyses. Finally, for transcription comparisons, the wheat EST collections must come from non-normalized cDNA libraries.

In addition, it is important to be aware of known artifacts of ESTs such as chimeric clones composed of sequences generated from 2 or more original mRNAs. Such chimeras can be generated at several points in the cloning process, including during first-strand cDNA synthesis when the growing DNA chain transcribes from more than one mRNA molecule, during ligation into the vector when multiple cDNA molecules fuse, and by recombination in a bacterial host cell transformed with multiple cDNA molecules. Although it is believed (and assumed in the present report) that these potential sources of artifacts account for observed chimeric sequences, note that the existence of *in vivo* chimeric mRNA molecules cannot be ruled out.

Chimeric EST sequences are most easily identified by abrupt shifts of similarity from one reference sequence to another. Chimeras will tend to terminate contig assemblies; for example, if the 5' end of a chimera joins a HMW-GS contig, but the 3' end originates from a different gene, then additional HMW-GS ESTs may not be added at the 3' end of the contig, depending on order of assembly and algorithm. Estimates in the present analysis are that about 3%–4% of ESTs involving HMW-GS sequences are chimeric. If sufficient overlapping sequence reads are available, such chimeric sequences become obvious. In such cases, chimeric ESTs were removed from assemblies, or the non-HMW-GS portions were removed to allow the ESTs to form reliable contigs. Diagrammatic examples of typical chimeric ESTs involving HMW-GS sequences are shown in Fig. S1.¹ Additional artifacts of ESTs are deletions or duplications of sequence from a single original mRNA molecule that occur either during copying from a source mRNA or during replication within host bacteria. Such deletions or duplications of sequence are the most common artifacts with HMW-GS ESTs and occur especially in repetitive sequence regions of all prolamine ESTs. These repetitive domain EST sequence deletions or tandem duplications are directly visible in assemblies if the deletion or duplication is small, but manual realignment is necessary for larger events. In all cases of wheat prolamine EST chimeras or ESTs with internal deletions or duplications, the ESTs were unique; i.e., in no case did multiple independent ESTs possess the same sequence

¹Supplementary data for this article are available on the journal Web site (<http://genome.nrc.ca>) or may be purchased from the Depository of Unpublished Data, Document Delivery, CISTI, National Research Council Canada, Building M-55, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada. DUD 3988. For more information on obtaining material refer to <http://cisti-icist.nrc-cnrc.gc.ca/eng/ibp/cisti/collection/unpublished-data.html>.

anomaly. In a few cases where ESTs had similar anomalies, the ESTs originated as 5' and 3' sequences from the same cDNA clone; for example, the ESTs BJ234294 and BJ240159 are from the same cDNA clone and both contain an 81 bp deletion of 3 tandem repeat motifs.

Although Nagy et al. (2004, 2005) have reported chimeric gliadin-LMW-GS genes using PCR amplification of sequences from genomic DNA, in the present study the uniqueness of all anomalous EST sequences argues against both bona fide active chimeric genes and any significant contribution to dough properties even if such chimeric genes do exist in the wheat genome.

Reconstructing known gene sequences from ESTs

To test the use of ESTs to assemble original gene sequences, EST-assembled consensus sequences were compared with previously published HMW-GS gene sequences. The large repetitive domain of the HMW-GS gene sequences provides an especially rigorous test of gene assemblies from ESTs. The known gene sequences were from *T. aestivum* cultivars where one or more HMW-GS genes have previously been reported and where numerous EST sequences exist, i.e., the Dx2 and Dy12 genes of Chinese Spring, all 6 genes of Cheyenne, and the Bx7 gene of Glenlea. The comparison also serves the complementary function of checking the accuracy of the originally reported sequences.

Chinese Spring

The complete Chinese Spring Dx2 coding sequence was derived from an assembly of 106 ESTs with a minimum coverage of 4 ESTs at each base position. In Fig. 2, the derived amino acid sequence is compared with the published Chinese Spring Dx2 amino acid sequence (X03346; Sugiyama et al. 1985), Dx5 amino acid sequences from Cheyenne (Anderson et al. 1989) and *T. aestivum* 'Renan' (Gu et al. 2006), and the derived amino acid sequence from a Dx5 EST assembly from Glenlea. The Chinese Spring Dx2 EST consensus DNA sequence is identical to X03346 except for an additional CCG codon encoding proline at amino acid residue position 81 of the EST-encoded consensus subunit (arrowhead in Fig. 2). This additional codon is supported by 23 ESTs at the site of the sequence difference, is identical at this position to 3 other Dx sequences as indicated in Fig. 2, and is the consensus of ESTs produced by 2 independent laboratories. The 2 tandem CCG triplets at this position suggest an error in the original reading of the X03346 sequence. The revised Chinese Spring Dx2 sequence is available from the GenBank TPA database under accession No. BK006460.

The Chinese Spring Dy12 EST consensus sequence was derived from the assembly of 76 ESTs with a minimum coverage of 4 ESTs at all positions of the sequence and agrees with the published Dy12 sequence (X03041; Thompson et al. 1985) except at 2 sites (Fig. S2).¹ First, there was a synonymous base difference at position 87 of the coding region that was supported by 16 ESTs and the Dy10 sequence (Anderson et al. 1989). Second, there was a GGACAG sequence at position 1425 of the originally reported DNA sequence that would encode additional glycine and glutamine residues at position 478 of the subunit. The EST consensus sequence reported here was supported by 14 ESTs, the Dy10 sequence (Fig. S2), and a Dy12 sequence from *T. aestivum* 'Dong-

nong7742' (AY486484; not shown). As suggested for Dx2, the 2 tandem GGACAG sequences in the originally reported Dy12 sequence are likely due to an error in the sequence determination. The revised Chinese Spring Dy12 sequence is available from the GenBank TPA database under accession No. BK006459.

Cheyenne

EST assemblies were compared with our previously reported HMW-GS gene sequences from Cheyenne. Full-length Cheyenne sequences could not be derived from available ESTs because there were gaps in the central portion of the repetitive domains. However, there were sufficient ESTs for reliable comparison of some 5' and most 3' portions of several sequences. Differences between the reported and EST-generated sequences were confirmed by re-reading the original autoradiograms for the sequences reported for the Dx5 (X12928; Anderson et al. 1989) and Bx7 (X13927; Anderson and Greene 1989) genes. The errors were corrected in GenBank and included one synonymous correction and one correction for a single misread repeat motif for the Dx5 gene (X12928), and a single synonymous correction for the Bx7 gene (X13927). No differences were found between ESTs and reported sequences for the other 3 active Cheyenne HMW-GS genes (Ax2*, By9, Dy10; not shown).

Glenlea

Glenlea contains 2 Bx7 genes of identical sequence resulting from a tandem duplication (Ragupathy et al. 2008). The large number of Glenlea HMW-GS ESTs was sufficient for reconstruction of the Bx7 gene sequence; 109 ESTs were used, with a minimum of 7 ESTs at all base positions. The Glenlea EST set is characterized by a higher incidence of variant base reads than other EST sets. Examination of the alignments showed that all these uncertain assignments occurred at C positions in 3' to 5' sequence reads; individual Glenlea ESTs often had a T at these positions or occasionally an A or G (not shown). When the bases were assigned by the majority base in the alignment at each position, an exact reconstruction of the reported Glenlea Bx7 sequence was obtained except for the uncertainty between C and T at 6 positions (not shown).

Summary

The differences between EST-derived HMW-GS sequences and the reported gene sequences in Chinese Spring, Cheyenne, and Glenlea could be due to either errors in sequencing or differences in germplasms. The possibility that different germplasm sources were used is ruled out for the Cheyenne sequences, since the same source of seed was used for both the original genes and the ESTs, but the Chinese Spring genomic and EST sequences were generated in 1 and 2 different laboratories, respectively, for each Chinese Spring gene. Although the differences from published sequences are minor, the Cheyenne and Chinese Spring HMW-GS genes are commonly used as reference HMW-GS sequences and, as such, the accuracy of their sequences is essential.

In summary, the sequences derived from ESTs were comparable to, and in some cases more accurate than, previously reported sequences for specific HMW-GS genes, and are a

Fig. 2. Dx subunit sequences. The derived amino acid sequence from the Chinese Spring Dx2 EST consensus sequence is compared with those from the published Dx2 sequence (X03346; Sugiyama et al. 1985), Dx5 sequences (Anderson et al. 1989; Gu et al. 2006), and the Dx5 sequence assembled from ESTs of Glenlea. The site of the additional codon in the Dx2 EST assembly is marked at the position of the encoded proline by an arrowhead. The asterisk indicates the cysteine residue considered unique for the Dx5 subunit. Shaded residues differ among the sequences. X's denote uncertain positions in the Glenlea Dx5 EST consensus.

	20	40	60	80	100	
1	MAKRLVLFVAVVVALV	LTVAEGEASEQLQCERELQELQERELKACQVMDQQLRDISPECHPVVSPVAGQYEQIVVPPKGGSFYPGETTTPPQQLQR				Dx2-CS-ESTs
1	MAKRLVLFVAVVVALV	LTVAEGEASEQLQCERELQELQERELKACQVMDQQLRDISPECHPVVSPVAGQYEQIVVPPKGGSFYPGETTTPPQQLQR				Dx2-CS-X03346
1	MAKRLVLFVAVVVALV	LTVAEGEASEQLQCERELQELQERELKACQVMDQQLRDISPECHPVVSPVAGQYEQIVVPPKGGSFYPGETTTPPQQLQR				Dx5-Cheyenne
1	MAKRLVLFVAVVVALV	LTVAEGEASEQLQCERELQELQERELKACQVMDQQLRDISPECHPVVSPVAGQYEQIVVPPKGGSFYPGETTTPPQQLQR				Dx5-Renan
1	MAKRLVLFVAVVVALV	LTVAEGEASEQLQCERELQELQERELKACQVMDQQLRDISPECHPVVSPVAGQYEQIVVPPKGGSFYPGETTTPPQQLQR				Dx5-Glenlea-ESTs
	120	140	160	180	200	
101	IFWGI	PALLKRYYP	SVTSPQ	QVSYYPGQAS	PQRPGQGQ	Dx2-CS-ESTs
100	IFWGI	PALLKRYYP	SVTSPQ	QVSYYPGQAS	PQRPGQGQ	Dx2-CS-X03346
101	IFWGI	PALLKRYYP	SVTSPQ	QVSYYPGQAS	PQRPGQGQ	Dx5-Cheyenne
101	IFWGI	PALLKRYYP	SVTSPQ	QVSYYPGQAS	PQRPGQGQ	Dx5-Renan
101	IFWGI	PALLKRYYP	SVTSPQ	QVSYYPGQAS	PQRPGQGQ	Dx5-Glenlea-ESTs
	220	240	260	280	300	
201	RQPGQGQ	PGYYPTSSQLQ	PGQLQQPAQGGQ	GGQPGQGQ	GGQGGQ	Dx2-CS-ESTs
200	RQPGQGQ	PGYYPTSSQLQ	PGQLQQPAQGGQ	GGQPGQGQ	GGQGGQ	Dx2-CS-X03346
195	QPGQGQ	PGYYPTSSQLQ	PGQLQQPAQGGQ	GGQPGQGQ	GGQGGQ	Dx5-Cheyenne
195	QPGQGQ	PGYYPTSSQLQ	PGQLQQPAQGGQ	GGQPGQGQ	GGQGGQ	Dx5-Renan
195	QPGQGQ	PGYYPTSSQLQ	PGQLQQPAQGGQ	GGQPGQGQ	GGQGGQ	Dx5-Glenlea-ESTs
	320	340	360	380	400	
301	QQLGQGQ	SGYYPTSPQ	PGQGQPGQLQQPAQGGQ	PGQGQGQGG		

reliable source for original gene sequences if sufficient high-quality ESTs are available.

New HMW-GS sequences

To determine whether new HMW-GS sequences could be assembled from ESTs, the EST collections for Chinese Spring and Glenlea were analyzed; no other wheat cultivars had sufficient ESTs to attempt the reconstruction of previously unknown full-length sequences.

Chinese Spring

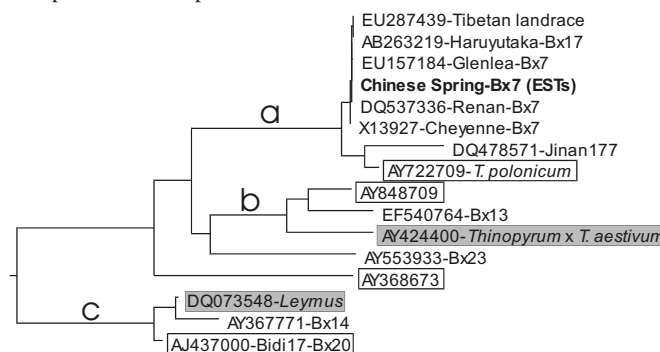
The hexaploid wheat cultivar Chinese Spring is the accepted reference germplasm for wheat research. It would be useful to know the sequences of its complete set of HMW-GS genes, but thus far only the Dx and Dy genes have been reported (Sugiyama et al. 1985; Thompson et al. 1985). The full-length coding sequence of the Chinese Spring Bx7 gene was assembled from 99 ESTs, with at least 6 ESTs covering all portions of the sequence. The consensus is identical to the

Glenlea Bx7 gene and differs from the Renan and Cheyenne Bx7 genes only by the presence of a single additional repeat motif (encoding the amino acid sequence PGQGQQ and supported by 46 ESTs) in the Chinese Spring and Glenlea Bx sequences at position 774 (Fig. S3).¹ This difference is shared by a number of other germplasms, including tetraploid wheat and *Leymus* (Fig. S3), and presumably indicates a deletion in the common ancestor gene in lineages missing the repeat. Further confirming the Chinese Spring Bx7 gene assembly, tryptic peptides of the derived amino acid sequence match exactly the sequences of the unordered Chinese Spring Bx7 subunit peptides (not shown) determined by Alberghina et al. (2005). As the reference germplasm, the Chinese Spring Bx7 gene would be designated as the *Glu-B1-1a* allele at the complex B-genome HMW-GS locus (McIntosh et al. 2003). The Chinese Spring Bx7 sequence is available from the GenBank TPA database under accession No. BK006773.

To place the Bx genes in an evolutionary context, all known *Triticum* full-length Bx subunit plus signal peptide sequences were aligned (Fig. S3) and the resulting phylogenetic tree is shown in Fig. 3. Several observations are worth noting. First, the identical Chinese Spring and Glenlea sequences were not expected, since Chinese Spring is considered more distantly related to most modern cultivars than those cultivars are to one another. Second, note that the Chinese Spring–Glenlea and Cheyenne–Renan Bx subunit pairs are all classified as subunit 7 (Bx7) in spite of the repeat motif difference. It is known that the HMW glutenin subunits migrate anomalously in SDS–PAGE gels, with apparent molecular weights as much as 50% higher than standard molecular weight markers (Shewry and Halford 2002). The HMW-GS classification system is based on relative gel mobility; in the original classifications, the most slowly migrating subunit was designated subunit 1, and other subunits were numbered sequentially with faster migration. This initial system has had to adjust to a continuing series of identified gradations in migration, leading to subtle, and increasingly awkward, nomenclatures such as Dx2, Dx2.1, Dx2.2 and Ax2, Ax2*, etc. (McIntosh et al. 2003). Even in cases where migration differences have been observed, the nomenclature of subunits has remained identical (e.g., Bx7 subunits of Cheyenne and Chinese Spring; D'Ovidio et al. 1997). In such cases the nomenclature must acknowledge known differences. In addition, an inability to distinguish subunit migration differences is not sufficient to assign the same subunit nomenclature if unique identifiers are needed. At a minimum, any subunit classification should include the germplasm source of the HMW-GS sequence, and with sequencing technology advancing rapidly, the actual DNA sequence may become standard for HMW-GS classification. Finally, the phylogenetic tree exhibits distinct branches (a, b, and c in Fig. 3) that each contain HMW-GS sequences from both hexaploid and tetraploid wheats. In addition, branch c contains a sequence (DQ073548) from the more distantly related Triticeae genus *Leymus*, and branch b contains a sequence of uncertain origin, since it comes from a *Thinopyrum* × *T. aestivum* cross. Implications for understanding wheat evolution are discussed below.

The By sequence would complete the set of active Chinese Spring HMW-GS genes, but only 24 mainly 3' Chinese Spring By ESTs were found in the public EST database —

Fig. 3. Phylogenetic tree of wheat Bx subunit sequences. Full-length derived Bx subunit sequences and the EST consensus of Chinese Spring were aligned with Clustal V (Fig. S3) using the Dx5 coding sequence as the outgroup. Tetraploid *T. turgidum* sequences are boxed. Sequences from *Leymus* and a *Thinopyrum* × *T. aestivum* cross are shaded. Sequences are named by GenBank accession, cultivar or genus or species, and subunit classification (when subunit is assigned). Branches a, b, and c each contain both hexaploid and tetraploid members.



not sufficient to assemble a full-length sequence (assembly not shown). Of the two “inactive” Chinese Spring HMW-GS genes, no Ax ESTs were detected and the 2 Ay ESTs are discussed below.

Glenlea

For Glenlea, only the Bx gene has been previously reported (Ragupathy et al. 2008). The Glenlea Dx5 gene sequence was found to be represented by 72 ESTs that cover the entire coding sequence to a minimum depth of 3 or more ESTs. The derived Glenlea Dx5 amino acid sequence is compared with the Cheyenne and Renan Dx5 and Chinese Spring Dx2 sequences in Fig. 2 and is confirmed to contain the cysteine residue at position 118 that is unique to Dx5 subunits and speculated to be related to higher dough quality owing to increased potential for gluten polymer cross-linking (Shewry and Halford 2002). The Glenlea Dx5 sequence is most similar to the Renan Dx5 sequence, and those two sequences share 4 single base differences and 1 repeat difference compared with the Cheyenne Dx5 sequence — again showing the problem in naming subunits by protein gel migration when clear sequence differences exist.

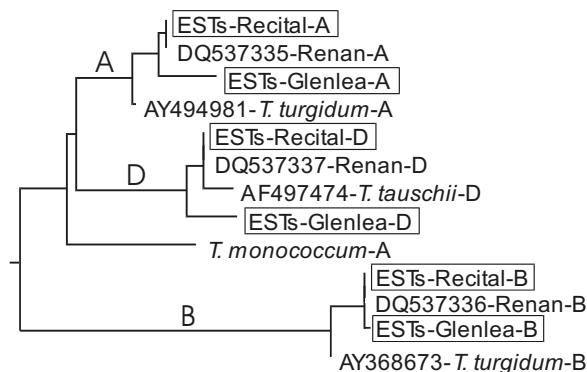
The Glenlea Dy10 gene sequence is represented by 79 ESTs that cover the entire coding sequence to a minimum depth of 5 ESTs. Five positions could not be determined — all exhibiting the previously discussed anomaly of 3' to 5' sequence reads — but only 3 of those led to uncertainties in amino acid sequence. Otherwise, the Glenlea and Cheyenne Dy10 genes are identical (not shown).

The Glenlea Ax2* and By8 gene sequences are represented by 39 and 20 ESTs, respectively. However, both assemblies lack the central portion of the repetitive domain and there are too few ESTs covering significant flanking sequences to those gaps to form either complete coding sequences or reliable consensus.

Globulin

More support for the reliability of EST gene assemblies comes from the active globulin gene shown in Fig. 1 — the

Fig. 4. Assignments of globulin EST consensus sequences to specific genomes. Globulin coding sequences from Renan (Gu et al. 2006) were compared using Clustal V (Fig. S4) and the relationships are shown by phylogenetic tree. The orthologous barley globulin sequence was used as the outgroup. Reported or assigned genomes are indicated by the letter following the cultivar or species name. Genomes are indicated on major branches (A, B, D). EST consensus sequences are boxed. The *T. monococcum* sequence is unpublished (Y. Gu, personal communication, 2008).



third, and only other, gene within the known HMW-GS region to have major representation in Triticeae EST collections. The globulin gene and derived protein sequences can be reliably assembled only for the hexaploid wheat cultivars Recital and Glenlea; the other cultivars have too few globulin ESTs for either full-length assemblies or conclusive genome assignments. The relationship of these 6 globulin protein sequences to each other, to 3 sequences of known genome assignment (from Renan), and to sequences from tetraploid *Triticum turgidum* and diploid *Triticum monococcum* is shown in Fig. 4 and is based on the alignment in Fig. S4.¹ The Glenlea sequences are of sufficient quality to unambiguously assign the gene sequences and their member ESTs to genomes, but the number of uncertain base assignments (see above) prevents completely accurate comparisons with allelic sequences — thus, the apparent variation in Glenlea amino acid sequences in Fig. 4 is mainly due to the number of uncertain base positions. The Recital globulin sequences are identical to the genome sequences previously reported for Renan (Gu et al. 2006), indicating the close relationship of these two major cultivars grown in France. The A- and D-genome globulins are clearly more related to each other than to the B-genome sequence, in agreement with other observations such as the greater synteny between the A and D genomes than to the B genome (Akhunov et al. 2003) and the closer relationship of the A- and D-genome HMW-GS sequences to each other than to the B-genome sequences (Anderson and Greene 1989; Gu et al. 2004). The branching of the *T. monococcum* sequence before the common A- and D-genome ancestor occurs with both the DNA (not shown) and derived amino acid sequences. While the number of differences among sequences is small, this tree supports other findings that the A^m genome is significantly divergent from the cultivated polyploid A genomes, which are believed to be descended from the wild diploid *Triticum urartu* (Dvořák et al. 1993). Taken together, these findings may indicate that the split between the A and A^m

genomes is closer than previously thought to the split between D-genome and common A-genome ancestors.

Kinases

In contrast to the HMW-GS and globulin genes, the two kinase genes within the *Glu-1* regions (Fig. 1) are apparently not highly expressed, at least as reflected by the current EST collection. Only 1 wheat EST (CJ652079; Chinese Spring) matches the genomic receptor kinase sequence near the HMW-GS loci. This EST matches best to the D-genome sequence when compared with the 3 orthologous receptor kinase sequences of Renan (Gu et al. 2006), having 4 mismatches to the D-genome sequence versus 14 and 39 mismatches to the A- and B-genome sequences, respectively (Fig. 5). The 4 D-genome mismatches can be attributed to expected uncertainties in any single EST sequence and allelic differences between Chinese Spring and Renan. Three cDNA sequences (EF394367, EF394368, EF394369) from *T. aestivum* and *T. turgidum* were reported after isolation using primers derived from the receptor kinase sequence adjacent to the HMW-GS loci (Niu et al. 2006). Although those 3 cDNA sequences are very similar to the HMW-GS region's receptor kinase gene, they are separate receptor kinase genes more distantly related to the 3 orthologous HMW-GS receptor kinase genes and to EST CJ652079 (not shown).

More support for the activity of the receptor kinase gene comes from barley ESTs BQ466381 and CA015492, which match (Fig. S5) the orthologous barley receptor kinase gene (Gu et al. 2003).¹ The mRNAs for the wheat and barley receptor kinase ESTs originated from endosperm and imbibed seed, respectively. There are currently over 1.6 million wheat and barley ESTs, with a heavy weighting to seed-related tissues. The low representation of these receptor kinase sequences in the EST collection indicates that the gene is expressed either at low levels or under specific conditions not well sampled in current EST collections, and demonstrates an inherent limitation in using ESTs for transcription studies.

The sequences of the protein kinase genes 3' to the wheat x-type HMW-GS and barley D-hordein genes were used to search for related ESTs. No matching wheat ESTs were found, but 4 barley ESTs (EX574932, EX572933, EX580979, EX580984) matched the barley orthologous protein kinase gene, as shown in Fig. 6. The 4 barley ESTs match the barley BAC sequence AY268139 over a continuous 2223 bp, except for gaps in the alignment that suggest the ESTs span 5 exons and 4 introns (junctions indicated by asterisks in Fig. 6). The few mismatches are attributed to single-EST read errors and germplasm differences — the BAC sequence is from *Hordeum vulgare* 'Morex', while the ESTs are from *H. vulgare* 'Barke'. The tissue source of the original barley mRNAs is not known, since the cDNA libraries were from a pool of 15 tissues and conditions which include germinating seed 12 and 62 h after imbibition. The possibility that the protein kinase ESTs originate from the imbibed tissues is intriguing, given the origin of the barley receptor kinase ESTs.

In summary, of the 5 genes shown in Fig. 1, 3 are relatively highly expressed and are specific to the wheat endosperm (2 HMW-GS genes and 1 globulin gene). The 2

Fig. 5. Single EST match to the wheat receptor kinase gene near the *Glu-1* loci. EST CJ652079 is aligned with the relevant portions of the three orthologous receptor kinase sequences from the A, B, and D genomes of Renan (Gu et al. 2006). Mismatches to the EST sequence are shaded. Dots at ends of the three genomic sequences indicate sequences extending beyond the EST ends.

	10	20	30	40	50	60	70	80	
1	-----	-----	-----	-----	-----	-----	-----	-----	CJ652079
1	TGCTGCAGTTAGCCAAGGAGGT	CAGAAAATCCCAATCTCTTGACTCTGTAGCCTGCAGCTCATTATTTCTACTTCA							Renan-D
1	...TGCTGCAGTTAGCCAAGGAGGT	CAGAAAATCCCAATCTCTTGACTCTGTAGCCTGCAGCTCATTATTTCTACTTCA							Renan-A
1	...TGCTGCAGTTAGCCAAGGAGGT	CAGAAAATCCCAATCTCTTGACTCTGTAGCCTGCAGCTCATTATTTCTACTTCA							Renan-B
	90	100	110	120	130	140	150	160	
81	-----	-----	-----	-----	-----	-----	-----	-----	CJ652079
81	AATACAGTCAACATCCACTTTGGCCAATATTTGATGGGATTCGGCCACGAAATCTATTCATGGAGAGAGAGATGTTTC								Renan-D
81	AATACAGTCAACATCCACTTTGGCCAATATTTGATGGGATTCGGCCACGAAATCTATTCATGGAGAGAGAGATGTTTC								Renan-A
81	AGTACAGTCAACATCCACTTTGGCCAATATTTGATGGGATTCGGCCACGAAATCTATTCATGGAGAGAGATGTTTC								Renan-B
	170	180	190	200	210	220	230	240	
161	-----	-----	-----	-----	-----	-----	-----	-----	CJ652079
161	AAGAGAAGATATGTTTGACAAGGAGGCGAGGAATCTGGCCCTCAAATTTGTTGTAGAATACACTGAACCTTTTTCAGGTTAG								Renan-D
161	AAGAGAAGATATGTTTGACAAGGAGGCGAGGAATCTGGCCCTCAAATTTGTTGTAGAATACACTGAACCTTTTTCAGGTTAG								Renan-A
161	GAGAAGATATGTTTGACAAGGAGGCGAGGAATCTGGCCCTCAAATTTGTTGTAGAATACACTGAACCTTTTTCAGGTTAG								Renan-B
	250	260	270	280	290	300	310	320	
241	-----	-----	-----	-----	-----	-----	-----	-----	CJ652079
241	GAAGTATAGAGCCAATATCTTGTGGTAGAGAGCCTGACAGTTGGTTTGACCCGAAATTGAGGTATTCAAGTGACGACATA								Renan-D
241	GAAGTATAGAGCCAATATCTTGTGGTAGAGAGCCTGACAGTTGGTTTGACCCGAAATTGAGGTATTCAAGTGACGACATA								Renan-A
241	GAAGTATAGAGCCAATATCTTGTGGTAGAGAGCCTGACAGTTGGTTTGACCCGAAATTGAGGTATTCAAGTGACGACATA								Renan-B
	330	340	350	360	370	380	390	400	
321	-----	-----	-----	-----	-----	-----	-----	-----	CJ652079
321	TTAAATAACACTGGAGGGATCAAACCTTGTAGGTTATTGGCTGCTAGATTGAGATTGAGGTTGGTAAGCTTAGACAA								Renan-D
321	TTAAATAACACTGGAGGGATCAAACCTTGTAGGTTATTGGCTGCTAGATTGAGATTGAGGTTGGTAAGCTTAGACAA								Renan-A
321	TTAAATAACACTGGAGGGATCAAACCTTGTAGGTTATTGGCTGCTAGATTGAGATTGAGGTTGGTAAGCTTAGACAA								Renan-B
	410	420	430	440	450	460	470	480	
401	-----	-----	-----	-----	-----	-----	-----	-----	CJ652079
401	AGCTGGTGAACATGTCCGCTCATAATATTTCCACCCATGTTCAAATCATTCAAGTGTCTGCAAAATGCGGAGCCATGGTG								Renan-D
401	AGCTGGTGAACATGTCCGCTCATAATATTTCCACCCATGTTCAAATCATTCAAGTGTCTGCAAAATGCGGAGCCATGGTG								Renan-A
401	AGCTGGTGAACATGTCCGCTCATAATATTTCCACCCATGTTCAAATCATTCAAGTGTCTGCAAAATGCGGAGCCATGGTG								Renan-B
	490	500	510	520	530	540	550	560	
481	-----	-----	-----	-----	-----	-----	-----	-----	CJ652079
481	GTATTTGCCCATGACATGGTTCTTTACTATACTGAAGAGGGTGACTGTTGCAAGATCTGCAAAAGAGGAGGAATGGTA								Renan-D
481	GTATTTGCCCATGACATGGTTCTTTACTATACTGAAGAGGGTGACTGTTGCAAGATCTGCAAAAGAGGAGGAATGGTA								Renan-A
481	GTATTTGCCCATGACATGGTTCTTTACTATACTGAAGAGGGTGACTGTTGCAAGATCTGCAAAAGAGGAGGAATGGTA								Renan-B
	570	580	590	600	610	620	630		
561	-----	-----	-----	-----	-----	-----	-----	-----	CJ652079
561	CCCGAGATATTGTTGCTGCCAATAGCCAGAACCAACAGCTTTGACAGGTTTCCCATGGCGGGGAGGATGGC...								Renan-D
561	CCCGAGATATTGTTGCTGCCAATAGCCAGAACCAACAGCTTTGACAGGTTTCCCATGGCGGGGAGGATGGC...								Renan-A
561	CCCGAGATATTGTTGCTGCCAATAGCCAGAACCAACAGCTTTGACAGGTTTCCCATGGCGGGGAGGATGGC...								Renan-B

kinase genes are now known to be active in at least some Triticeae at apparently lower expression levels; the few known matching ESTs are from imbibed seed and endosperm (receptor kinase) and a pool of tissues that includes imbibed seeds (protein kinase). It seems unlikely that 4 distinctive gene classes whose expression is restricted to seed-associated tissues would accidentally occur in tandem in the genome unless they are part of a higher-order chromatin domain, in this case a seed- or endosperm-specific domain. Future work can determine whether the expression pattern of the 2 kinase genes is restricted to the endosperm and whether other genes in the immediate flanking region in Fig. 1 show seed- or endosperm-specific expression.

Whenever a presumptive gene sequence is missing or under-represented in EST collections, there are several possible explanations: the lack of ESTs could be an artifact of cloning (some sequences are less clonable or even unclonable), the level of transcription could be so low that it is difficult to detect without very extensive EST sequencing, the specific tissue(s) in which the gene is active may not have been sampled to generate the ESTs, or the gene may re-

spond to specific external stimuli not applied prior to mRNA preparations. Finally, a combination of the above may apply: if the receptor kinase and protein kinase genes flanking the *Glu-1* loci are expressed at relatively low levels in seed endosperm before or during imbibition, then such low representation within the large wheat EST collection would be explained by insufficient depth and breadth of endosperm EST sequencing, since no current wheat ESTs were generated from imbibed tissue.

Transcription of null alleles

Although the hexaploid bread wheats contain 6 HMW-GS genes, almost all cultivars express only 4 or 5 of those genes; the Ay gene is always considered null in major cultivated wheats, and sometimes also the Ax gene (Payne et al. 1981). Known examples of active Ay genes are mainly in diploid and tetraploid wheats (Sun et al. 2004), and the few cases where cultivated hexaploid wheats express the Ay gene are hypothesized to have occurred through introgression from wild *T. urartu* or *T. dicoccoides* during breeding programs (Margiotta et al. 1996).

Fig. 6. Barley protein kinase ESTs. Four barley ESTs are aligned to the protein kinase region of barley BAC AY268139. Dashes are used in alignment gaps. Shaded bases indicate sequence differences. Dots at ends of the BAC sequence indicate genomic sequences extending beyond the EST sequence ends. Asterisks indicate suggested intron–exon boundaries.

The “null” categorization of lack of expression of specific HMW-GS genes in cultivars has been mainly based on the absence of subunits as detected by protein gel electrophoresis. Previous reports of apparent transcriptional inactivity of Ay genes in Chinese Spring (Thompson et al. 1983) and Cheyenne (Forde et al. 1985) relied on detection of the subunit or methods unlikely to detect low levels of mRNA. A stricter conclusion from such reports is that the relative number of transcripts is too low to easily detect Ay and (or) Ax subunits or mRNA in the mature wheat endosperm. However, an examination of wheat ESTs shows that the Ay gene is being actively transcribed to mRNA in at least some germplasms. From about 1000 wheat HMW-GS ESTs in GenBank, 4 were identified as originating from Ay genes. Figure 7 compares 2 Chinese Spring ESTs (BJ233120 and BE590573) with the 3 y-type HMW-GS sequences from Cheyenne (Forde et al. 1985; Anderson and Greene 1989; Anderson et al. 1989). These 2 ESTs mismatch the Cheyenne Ay sequence at only 2 positions (boxed bases in Fig. 7) but mismatch the By and Dy genes at 31 positions each (shaded bases in Fig. 7), clearly identifying these 2 ESTs as originating from the Ay gene. These Chinese Spring EST sequences were compared with sequences from Cheyenne and not directly with the Chinese Spring Ay gene sequence because the available Chinese Spring Ay sequence (X05995 and X05996) overlaps only at the 3' end of the Chinese Spring ESTs (arrowhead in Fig. 7). However, this overlap allowed the reconstruction of the full-length Chinese Spring Ay sequence by adding the EST sequences (BJ233120 and BE590573) from their start codons to the 5' end of the reported partial sequences (X05995 and X05996).

The Chinese Spring Ay gene has been identified as containing an 8 kb transposon insertion and at least 2 stop codons within the coding region (Harberd et al. 1987). In addition to these structural features of the Ay gene coding region in Chinese Spring, Halford et al. (1989) reported the apparent lack of transcription function of the Ay promoter in a transgenic tobacco assay and speculated that a subset of 10 specific single-base changes in the first 280 bp upstream of the transcription start site were the cause of the inactivity. However, sequence analysis of the 5' non-coding promoter sequences of inactive and active Ay genes found no sequence differences (D'Ovidio et al. 1996). Assuming the Chinese Spring Ay mRNAs would translate into a polypeptide truncated by the first known stop codon (not shown), an abnormal HMW-GS polypeptide of 271 amino acids would terminate at approximately the middle of the repetitive domain; in comparison, the Cheyenne Ay gene would encode for a 318 amino acid polypeptide prior to its premature stop codon.

Since most HMW-GS ESTs cluster at the 3' end of known sequences, the fact that the only 2 Chinese Spring Ay ESTs are 5' suggests a possible lack of 3' matching mRNA sequences. The complete structure of the originating mRNAs is not known, but transcription would create RNAs synthesized into the 8 kb transposon insertion and likely terminat-

ing somewhere within the insertion sequence, and thus not including 3' HMW-GS sequences.

In addition to the 2 Chinese Spring Ay ESTs, 2 Glenlea ESTs are identified as Ay sequences. EST BQ244559 aligns to y-type 3' sequences as shown in Fig. S6A and mismatches the Cheyenne Ay sequence at 4 positions (boxed bases in Fig. S6A), versus 56 positions for the By gene and 45 positions for the Dy gene. In addition, the alignment shows that this EST and the Ay sequence have common gaps when compared with the By and Dy sequences, thus confirming BQ244559 as originating from the Glenlea Ay gene. The second Glenlea Ay EST, BQ241779, is aligned in Fig. S6B to the 3 Cheyenne y-type sequences and mismatches the Ay gene at 10 positions and the By and Dy sequences at 34 positions each. The By sequence is even more diverged, since from downstream of position 367 the By 3' untranslated sequences diverge completely from the Ay and Dy sequences (Anderson et al. 2002). The Glenlea EST sequences have some ambiguity at cytosine positions, as previously noted, and 7 of the 10 mismatches to the Cheyenne Ay sequence occur at cytosine positions in the Cheyenne sequence.

Nonsense-mediated mRNA decay is an incompletely understood mechanism by which cells identify and degrade mRNAs with premature stop codons, aberrant splicings, transposon insertions, and frame shifts (Mitrovich and Anderson 2005). From EST analyses it cannot be determined whether the low number of Ay ESTs results from lower rates of transcription or from faster turnover of the mRNA, but the lack of obvious promoter sequence differences between functioning Ay genes and apparent nulls (D'Ovidio et al. 1996) suggests nonsense-mediated mRNA decay as the primary factor. Any truncated Ay polypeptides that are produced would function as gluten chain terminators, since models predict that only one N-terminal cysteine residue would be available for intermolecular cross-links (Shewry et al. 1992, 2003). Whatever the basis of the low mRNA titer, the Ay gene contributes little to the seed HMW-GS protein complement and likely has negligible impact on wheat quality.

Phylogenetic context and implications of novel HMW-GS sequences

Specific levels of polyploidy could originate from single or multiple events (Soltis and Soltis 1999); the latter can be distinguished by multiple sequence patterns occurring vertically at different ploidy levels. The possibility of multiple polyploidization events leading to hexaploid *T. aestivum* has been suggested in reports such as studies of the Bx gene sequences (Ragupathy et al. 2008), studies of relative nucleotide substitution rates in the *Glu-1* region of the tetraploid A and B genomes (Gu et al. 2006), and findings of multiple shared alleles at orthologous regions of the D genome of hexaploid wheat and the D-genome donor, *T. tauschii* (Caldwell et al. 2004). Similar analyses with the Ay gene offer a slightly different perspective, since Ay genes are contribu-

Published by NRC Research Press

Fig. 7. Chinese Spring Ay ESTs. Two Chinese Spring ESTs are aligned to the three Cheyenne y-type HMW-GS sequences. Shaded bases indicate differences between Ay and By or Dy sequences. Boxed bases are two single-base sequence differences between the Chinese Spring ESTs and the Cheyenne Ay gene and a single base unique for the Cheyenne Ay gene. An arrowhead marks the 5' end of the known Chinese Spring Ay genomic sequence (X05995). Start codons are underlined. Dots indicate genomic sequences continuing upstream or downstream from EST ends.

```

1  ATGGCTAAGCGGTGGTCTCTTTGCGACAGTAGTCAATTTGGCCCTGTGCTCTCACCGTCTGAAGGTGAGGCCCTTAAGCAACTACAGTCGCGAGCTCCAGGAGTTTCG BJ233120
1  GCTAAGCGGTGGTCTCTTTGCGACAGTAGTCAATTTGGCCCTGTGCTCTCACCGTCTGAAGGTGAGGCCCTTAAGCAACTACAGTCGCGAGCTCCAGGAGTTTCG BE590573
1  ...ATGGCTAAGCGGTGGTCTCTTTGCGACAGTAGTCAATTTGGCCCTGTGCTCTCACCGTCTGAAGGTGAGGCCCTTAAGCAACTACAGTCGCGAGCTCCAGGAGTTTCG Ay
1  ...ATGGCTAAGCGGTGGTCTCTTTGCGACAGTAGTCAATTTGGCCCTGTGCTCTCACCGTCTGAAGGTGAGGCCCTTAAGCAACTACAGTCGCGAGCTCCAGGAGTTTCG By9
1  ...ATGGCTAAGCGGTGGTCTCTTTGCGACAGTAGTCAATTTGGCCCTGTGCTCTCACCGTCTGAAGGTGAGGCCCTTAAGCAACTACAGTCGCGAGCTCCAGGAGTTTCG Dy10
121 CTTGAGGCATGCCGCTGGTCTGTGGACCAACAGTTGGCCGCTGCGGCTGCCATGGAGCACGGGGCTCCAGATCGGTCGCCAGCTCCAGATATTAGTCGCAAGTGTGCCCCGTC BJ233120
118 CTTGAGGCATGCCGCTGGTCTGTGGACCAACAGTTGGCCGCTGCGGCTGCCATGGAGCACGGGGCTCCAGATCGGTCGCCAGCTCCAGATATTAGTCGCAAGTGTGCCCCGTC BE590573
121 CTTGAGGCATGCCGCTGGTCTGTGGACCAACAGTTGGCCGCTGCGGCTGCCATGGAGCACGGGGCTCCAGATCGGTCGCCAGCTCCAGATATTAGTCGCAAGTGTGCCCCGTC Ay
121 CTTGAGGCATGCCGCTGGTCTGTGGACCAACAGTTGGCCGCTGCGGCTGCCATGGAGCACGGGGCTCCAGATCGGTCGCCAGCTCCAGATATTAGTCGCAAGTGTGCCCCGTC By9
121 CTTGAGGCATGCCGCTGGTCTGTGGACCAACAGTTGGCCGCTGCGGCTGCCATGGAGCACGGGGCTCCAGATCGGTCGCCAGCTCCAGATATTAGTCGCAAGTGTGCCCCGTC Dy10
241 GCCCTCAGCCAAAGTCGCAAGAGCAATATGGGCAACCCGCGTCCGCCCAAGGGCGGATCTTCTACCATCGGAGCACCGGCTCTGAGCAACTCCCAACAGGAATATTTGGGGGAACA BJ233120
238 GCCCTCAGCCAAAGTCGCAAGAGCAATATGGGCAACCCGCGTCCGCCCAAGGGCGGATCTTCTACCATCGGAGCACCGGCTCTGAGCAACTCCCAACAGGAATATTTGGGGGAACA BE590573
241 GCCCTCAGCCAAAGTCGCAAGAGCAATATGGGCAACCCGCGTCCGCCCAAGGGCGGATCTTCTACCATCGGAGCACCGGCTCTGAGCAACTCCCAACAGGAATATTTGGGGGAACA Ay
241 GCCCTCAGCCAAAGTCGCAAGAGCAATATGGGCAACCCGCGTCCGCCCAAGGGCGGATCTTCTACCATCGGAGCACCGGCTCTGAGCAACTCCCAACAGGAATATTTGGGGGAACA By9
241 GCCCTCAGCCAAAGTCGCAAGAGCAATATGGGCAACCCGCGTCCGCCCAAGGGCGGATCTTCTACCATCGGAGCACCGGCTCTGAGCAACTCCCAACAGGAATATTTGGGGGAACA Dy10
361 TCTTCACAAACAGTACAAGGGTATTACCAAGTG BJ233120
358 TCTTCACAAACAGTACAAGGGTATTACCAAGTGTAATATCTCTCAGCAGGGGTCTATATTCAGGCCAAGCTTCTCCACACAGCCCA BE590573
361 TCTTCACAAACAGTACAAGGGTATTACCAAGTGTAATATCTCTCAGCAGGGGTCTATATTCAGGCCAAGCTTCTCCACACAGCCAGGA... Ay
361 TCTTCACAAACAGTACAAGGGTATTACCAAGTGTAATATCTCTCAGCAGGGGTCTATATTCAGGCCAAGCTTCTCCACACAGCCAGGA... By9
361 TCTTCACAAACAGTACAAGGGTATTACCAAGTGTAATATCTCTCAGCAGGGGTCTATATTCAGGCCAAGCTTCTCCACACAGCCAGGA... Dy10

```

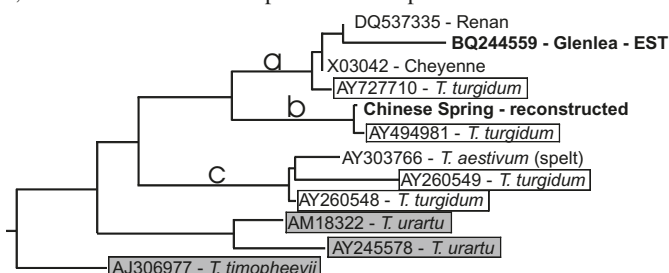
ting less to seed protein composition than other HMW-GS genes and thus are potentially under less selective pressure. We had previously reported the same transposon insertion into the Ay gene of *T. turgidum* 'Langdon' as in the partial Ay gene sequence of Chinese Spring (Gu et al. 2004). A full-length Chinese Spring Ay coding sequence was reconstructed from 2 clones bordering the transposon insertion site (X05995 and X05996) and the 2 Chinese Spring Ay ESTs. This reconstructed Chinese Spring sequence and the partial Glenlea sequence were used in a phylogenetic analysis with all available Ay sequences to place the new sequences in context with previously reported complete Ay sequences (Fig. 8). The relationships show at least 3 potential tetraploid to hexaploid lineages of Ay sequences, with each branch (a, b, c in Fig. 8) containing at least one hexaploid and one tetraploid member, i.e., one branch with 3 major bread wheat cultivars (Cheyenne, Glenlea, Renan), one with Chinese Spring, and the third with a spelt (*T. aestivum*). More distantly related branches contain the 2 diploid *T. urartu* sequences and further yet is a sequence from tetraploid *Triticum timopheevii* (AG genomes).

The phylogenetic analyses of both the Bx (Fig. 3) and Ay (Fig. 8) sequences show multiple branches containing both tetraploid and hexaploid wheats. Butow et al. (2004) reported a similar Bx marker distribution in tetraploid and hexaploid wheats in a study using PCR assays and HPLC. In Fig. 3, the Bx sequences cluster into 3 main branches (a, b, c), with the AY368673 sequence having a mixture of sequence features (Fig. S3) common to branches b and c. The possibility of significant gene flow into polyploid wheats is suggested by the clustering pattern of Bx genes from tetraploid wheat, hexaploid wheat, and wildrye (*Leymus*) (branch c in Fig. 3). The origin of the nuclear genome of *Leymus* is not clear. The *Leymus* Ns genome is believed to be contributed by *Psathyrostachys*, but the Xm genome is of unrecognized origin. It has also been suggested that *Leymus* may be a segmental autopolyploid of 2 *Psathyrostachys* species (Wu et al. 2003). These hypotheses do not account for a *Leymus* Bx-type protein that differs by only 3 amino acids from a hexaploid Bx14 gene (Fig. S3). Although the number of examples is limited, the Bx and Ay data together suggest either a large number of individual polyploidization events or significant introgression of diploid or tetraploid genes into hexaploid wheat in breeding programs.

Comparative transcript titers from across the HMW-GS gene region

Although there is a large volume of literature on the structure of the wheat prolamine genes and the distributions of alleles, there is little quantitative information on the relative expression of the 5 major wheat prolamine classes, i.e., the high- and low-molecular-weight glutenins and the α -, γ -, and ω -gliadins. In the current report, the relative distribution of ESTs is examined first for all prolamine classes and then for the 2 paralogous HMW-GS genes and the adjacent globulin gene in each genome among 7 cultivars. The basic assumption is that the relative distribution of ESTs is a measure of the relative titers of different mRNAs extracted from wheat tissues, then converted to cDNA clones and randomly sequenced. For this assumption to be valid, the cDNA libraries cannot have been normalized.

Fig. 8. Phylogenetic tree of wheat Ay coding sequences. Coding sequences of full-length Ay genes, the consensus of Glenlea ESTs, and the reconstructed Chinese Spring sequence were aligned with Clustal V using the Dy10 coding sequence as the outgroup. Tetraploid *T. turgidum* sequences are boxed. Tetraploid *T. timopheevii* and diploid *T. urartu* sequences are shaded. Sequences are named by cultivar or species and GenBank accession number. Branches a, b, and c contain both hexaploid and tetraploid members.



The one million *T. aestivum* ESTs in GenBank were examined for wheat prolamine sequences. Over 11 000 ESTs of specific prolamine classes were identified, as shown in the first row of Table 1. The distribution is generally consistent with our understanding of the composition of the wheat seed prolamine complement; i.e., the α -gliadins tend to be the most abundant prolamine type, followed by the γ -gliadins, LMW and HMW glutenins, and ω -gliadins. This distribution also generally agrees with estimates of gene copy numbers for the specific prolamine classes, although exact gene numbers vary greatly among germplasms for the gliadins and LMW glutenins (Shewry and Halford 2002; Shewry et al. 2003).

To compare relative EST distributions among different germplasms, ESTs were parsed from GenBank based on germplasm origin. Sufficient numbers of ESTs were not available from any diploid or tetraploid wheat, but 7 hexaploid cultivars had enough prolamine ESTs for comparisons (Table 1). These 7 hexaploid cultivars account for 67% of the total wheat prolamine ESTs — the remaining 33% are divided among germplasms of uncertain identity and germplasms with too few ESTs to allow comparisons. The Recital results are included to show the leveling effects of normalizing (P. Leroy, personal communication 2008). The relative distributions among the 7 hexaploid germplasms are graphed in Fig. S7 and show a varied distribution of prolamine ESTs.¹ The most prominent EST class is the α - or γ -gliadins, depending on the specific germplasm; i.e., either the α - (Cheyenne, Glenlea, Mercia) or γ -gliadins (Butte 86, Chinese Spring, Wyuna) predominate. The relative abundance of these 2 classes varies from over 3 $\times$ as many α -gliadin ESTs as γ -gliadin ESTs in Glenlea to over twice as many γ -gliadin ESTs as α -gliadin ESTs in Wyuna. The ω -gliadin ESTs are by far the smallest component for all the germplasms. The prolamine class with the least quantitative variability is the LMW glutenins, while the α - and γ -gliadins and HMW glutenins have the highest variability. If the unnormalized EST totals of Table 1 are used to perform a χ^2 test on the hypothesis that the cultivar and EST class distributions are independent, then this hypothesis is rejected, since the probability that the EST distribution is by chance is zero ($\chi^2 = 644$ at 20 df). To reiterate, the distribution of

ESTs among the 5 prolamine classes is dependent on the specific germplasm from which they originate.

The HMW-GS genome region as yet remains unique in wheat because the extensive structural studies allow identification of genome-specific sequences for all 5 genes in this region (Fig. 1). In addition, the HMW-GS are unique among the wheat prolamines by being encoded by a small conserved family of genes, 2 per genome, in contrast to the gliadin groups with higher and highly variable gene copy numbers. The distribution of ESTs matching the x- and y-type HMW-GS gene sequences and the globulin gene sequence was determined from the publicly available wheat ESTs for 6 cultivars. Results are given in Table 2. For the globulin gene, only EST totals for Glenlea were significant enough to use in further analyses.

Since values of less than 5 are not recommended for analysis by a χ^2 test, only Glenlea, Chinese Spring, and Cheyenne have sufficient ESTs distributed among the 4–5 active HMW-GS genes to exhibit significant differences in distribution (Fig. 9). The distribution of ESTs among the different HMW-GS genes within each cultivar in Table 2 indicates significant variability in gene activity among the active genes. To confirm this, the results in Table 2 were analyzed by the χ^2 test for goodness of fit individually for Glenlea, Chinese Spring, and Cheyenne and yielded p values sufficiently low to reject the hypothesis that EST distributions across the HMW-GS genes are random for each of these 3 cultivars. A similar result was obtained from the total ESTs for the 4 active HMW-GS genes common to all 3 cultivars (Bx, By, Dx, Dy) (Table 2). Next, the distribution of HMW-GS transcripts among the 3 germplasms was tested for the null hypothesis that the distribution of ESTs is independent of the germplasm, using the χ^2 test of association for the 4 active HMW-GS genes. The result was a χ^2 score of 11.1 with $p = 0.084$ at $df = 6$. This provides suggestive evidence against the null hypothesis, i.e., weak support for the idea that the distribution of ESTs across active genes is dependent on the germplasm. Additional data with more ESTs and germplasms are needed to support the significance of the suggested dependence.

The most common ESTs are for the 2 D-genome (Dx, Dy) HMW-GS genes and the Bx gene (Fig. 9). The Ax and By genes are present at lower levels in most cultivars, and the Ay gene is represented by only 4 ESTs (as described above). The apparently lower expression from the A genome is mainly explained by the nearly universally silenced state of the Ay gene in cultivated wheats and the occasionally silenced Ax gene, such as for Chinese Spring, which contributes about one-third of the HMW-GS ESTs. The two D-genome HMW-GS genes are expressed at similar levels to each other, whereas Bx ESTs are 3–5.5 $\times$ more abundant than By ESTs in the 3 germplasms. This preponderance of Bx transcript over By transcript is unexplained. It may be that the By promoter possesses a quantitative defect that has been tolerated either because the By gene is a member of a gene family with similar functions or because the Bx and By genes are inherited together and the Bx gene provides some level of compensation.

Glenlea is known to have 2 copies of the Bx7 gene resulting from a tandem duplication of a DNA segment containing the Bx and protein kinase genes (Ragupathy et al. 2008).

Table 1. Distribution of *T. aestivum* prolamine ESTs among prolamine classes (α -, γ -, and ω -gliadins and high- and low-molecular-weight glutenins).

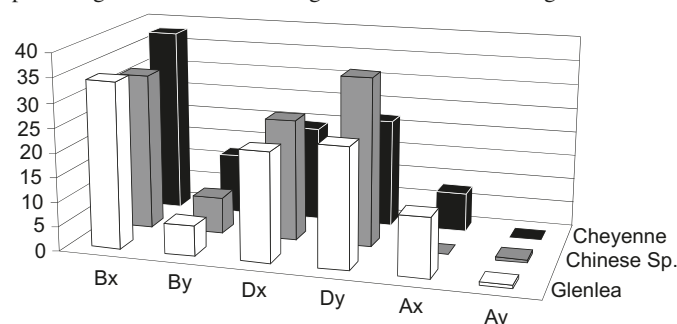
	α -gliadin	γ -gliadin	ω -gliadin	LMW glutenin	HMW glutenin	Total ESTs
GenBank	4605	2584	433	1978	1748	11348
By cultivar						
Butte 86	137	156	16	99	29	437
Cheyenne	262	153	15	145	117	692
Chinese Spring	868	921	215	543	303	2850
Glenlea	939	277	18	412	333	1979
Mercia	171	91	6	90	24	382
Recital ^a	(152)	(102)	(129)	(198)	(91)	(672)
Wyuna	135	279	16	105	38	573
Total	2512	1877	286	1394	844	6913

^aRecital cDNA libraries were normalized and are shown in parentheses and not counted in prolamine class totals.

Table 2. EST distribution for the two HMW glutenin genes and the globulin gene for each of the three genomes of six hexaploid bread wheat cultivars.

Gene	Glenlea	Chinese Spring	Cheyenne	Butte 86	Mercia	Wyuna	Total ESTs
Ay	2	2	0	0	0	0	4
By	20	23	14	1	3	4	65
Ax	39	0	9	8	0	11	67
Bx	111	99	43	7	2	6	268
Dy	79	106	25	5	10	8	233
Dx	72	76	22	9	8	8	195
Glo-A2	16	2	2	2	1	0	23
Glo-B2	27	1	2	1	0	0	31
Glo-D2	8	2	7	2	1	0	20
Chi-square ^a	78.8	55.8	31.1	—	—	—	124
Degrees of freedom	4	3	4	—	—	—	3
<i>p</i> value	<0.0001	<0.0001	<0.0001	—	—	—	<0.0001

^aGoodness-of-fit chi-square values determined for individual cultivars and total of all cultivars for the Ax, Bx, By, Dx, and Dy genes for Glenlea and Cheyenne, and for the Bx, By, Dx, and Dy genes for Chinese Spring and for the three cultivar totals.

Fig. 9. Distribution of HMW glutenin ESTs among three cultivars. The EST totals for three cultivars from Table 2 are plotted as the percentage of the total HMW glutenin ESTs for each gene.

Similarly duplicated Bx genes have been reported for a number of germplasms, but likely originated from a single original duplication event (D'Ovidio et al. 1997; Ragupathy et al. 2008). There is no evidence of a duplicated Bx gene in Cheyenne and Chinese Spring (unpublished data), and no other duplications of specific HMW-GS genes are known at this time. For Glenlea, the Bx ESTs comprise about 34% of the total HMW-GS ESTs — close to the 39% reported in a protein densitometry study for the Glenlea Bx subunit as a percentage of the total HMW-GS subunits (Marchylo et al. 1992). If there are twice as many Bx7 genes in Glenlea,

why is the Bx EST fraction not higher than in Cheyenne and Chinese Spring? D'Ovidio et al. (1997), in reporting the Bx duplication in *T. aestivum* 'Red River 68', noted that the coincidence of Bx gene duplication and higher Bx subunit composition is strong evidence, but not proof, that both copies of the duplicated gene are active. It also does not prove that both copies are as active as the original gene. Attempts to quantify HMW-GS composition have not shown conclusive evidence for Bx compositions consistent with doubling of Bx HMW-GS in specific cultivars (Marchylo et al. 1992).

The similarity of D-genome x- and y-type EST numbers is also consistent with a relatively uniform rate of mRNA generation between these 2 HMW-GS genes. Since the Dy coding region is shorter than the Dx coding region, slightly more Dy mRNAs should be generated, assuming equal rates of mRNA synthesis and minimal lag time in transcription initiation. Thus, using an estimated 2700 bp for a Dx mRNA and 2100 bp for a Dy mRNA, there should be a Dy/Dx ratio of approximately 1.29 — close to the actual ratio of 1.19 for the sums of 6 cultivars shown in Table 2.

The relatively balanced EST ratio from the D genome is interesting because of the findings of Békés et al. (1993); their in vitro studies of HMW-GS effects on dough mixing properties showed that the strongest effects were found with approximately equal amounts of the x- and y-type D-

genome subunits. Similarly, a balance of the 2 D-genome subunits was associated with positive mixing characteristics in transgenic wheat (Butow et al. 2003). The possibly significant quality effects of even relatively small differences in D-genome subunits requires further analysis, since quality differences associated with the poor bread-making quality pair Dx2 + Dy12 (as in Chinese Spring) versus the good-quality pair Dx5 + Dy10 (as in Glenlea and Cheyenne) are usually hypothesized to be caused by the small structural differences in the Dx5 and (or) Dy10 alleles (Flavell et al. 1989; Kasarda 1999). However, it is possible that the Dx/Dy ratio could also contribute to the observed quality differences. If this is the case, then the lesser variability in the x/y ratio encoded by the D genome compared with the B genome could be due to some level of selective advantage to maintaining a balanced D-genome x/y ratio in breeding programs selecting for bread-making quality.

Acknowledgements

Thanks to Ann Blechl and Don Kasarda for comments and suggestions.

References

- Akhunov, E.D., Akhunova, A.R., Linkiewicz, A.M., Dubcovsky, J., Hummel, D., Lazo, G., et al. 2003. Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. *Proc. Natl. Acad. Sci. U.S.A.* **100**(19): 10836–10841. doi:10.1073/pnas.1934431100. PMID:12960374.
- Alberghina, G., Cozzolino, R., Fisichella, S., Garozzo, D., and Savarino, A. 2005. Proteomics of gluten: mapping of the 1Bx7 glutenin subunit in Chinese Spring cultivar by matrix-assisted laser desorption/ionization. *Rapid Commun. Mass Spectrom.* **19**(14): 2069–2074. doi:10.1002/rcm.2032. PMID:15988718.
- 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**(5): 689–700. doi:10.1007/BF00261246.
- Anderson, O.D., Greene, F.C., Yip, R.E., Halford, N.G., Shewry, P.R., and Malpica-Romero, J.M. 1989. Nucleotide sequences of the two high-molecular-weight glutenin genes from the D-genome of a hexaploid bread wheat, *Triticum aestivum* L. cv Cheyenne. *Nucleic Acids Res.* **17**(1): 461–462. doi:10.1093/nar/17.1.461. PMID:2563152.
- Anderson, O.D., Litts, J.C., and Greene, F.C. 1997. The α -gliadin gene family. I. Characterization of ten new wheat α -gliadin gene clones, evidence for limited sequence conservation of flanking DNA, and Southern analysis of the gene family. *Theor. Appl. Genet.* **95**: 50–58. doi:10.1007/s001220050531.
- Anderson, O.D., Larka, L., Christoffers, M.J., McCue, K.F., and Gustafson, J.P. 2002. Comparison of orthologous and paralogous DNA flanking the wheat high molecular weight glutenin genes: sequence conservation and divergence, transposon distribution, and matrix-attachment regions. *Genome*, **45**(2): 367–380. doi:10.1139/g01-137. PMID:11962634.
- Anderson, O.D., Rausch, C., Moullet, O., and Lagudah, E.S. 2003. The wheat D-genome HMW-glutenin locus: BAC sequencing, gene distribution, and retrotransposon clusters. *Funct. Integr. Genomics*, **3**(1–2): 56–68. doi:10.1007/s10142-002-0069-z. PMID:12590343.
- Békés, F., Anderson, O.D., Gras, P.W., Gupta, R.B., Tam, A., Wrigley, C.W., and Appels, R. 1993. The contributions to mixing properties of 1D HMW glutenin subunits expressed in a bacterial system. In *Proceedings of the Royal Australian Chemical Institute Symposium on Improvement of Cereal Quality by Genetic Engineering*, Sydney, Australia, 12–16 September 1993. Edited by R.J. Henry and J.A. Ronalds. Plenum Press, New York. pp. 97–193.
- Békés, F., Kemény, S., and Morell, M. 2006. An integrated approach to predicting end-product quality of wheat. *Eur. J. Agron.* **25**(2): 155–162. doi:10.1016/j.eja.2006.04.008.
- Butow, B.J., Tatham, A.S., Savage, A.W.J., Gilbert, S.M., Shewry, P.R., Solomon, R.G., and Bekes, F. 2003. Creating a balance—the incorporation of a HMW glutenin subunit into transgenic wheat lines. *J. Cereal Sci.* **38**(2): 181–187. doi:10.1016/S0733-5210(03)00035-3.
- Butow, B.J., Gale, K.R., Ikea, J., Juhász, A., Bedő, Z., Tamás, L., and Gianibelli, M.C. 2004. Dissemination of the highly expressed Bx7 glutenin subunit (*Glu-B1a* allele) in wheat as revealed by novel PCR markers and RP-HPLC. *Theor. Appl. Genet.* **109**(7): 1525–1535. doi:10.1007/s00122-004-1776-8. PMID:15340686.
- Caldwell, K.S., Dvorak, J., Lagudah, E.S., Akhunov, E., Luo, M.C., Wolters, P., and Powell, W. 2004. Sequence polymorphism in polyploid wheat and their D-genome diploid ancestor. *Genetics*, **167**(2): 941–947. doi:10.1534/genetics.103.016303. PMID:15238542.
- D'Ovidio, R., Masci, S., and Porceddu, E. 1996. Sequence analysis of the 5' non-coding regions of active and inactive 1Ay HMW glutenin gene from wild and cultivated wheats. *Plant Sci.* **114**(1): 61–69. doi:10.1016/0168-9452(95)04302-0.
- D'Ovidio, R., Masci, S., Porceddu, E., I., and Kasarda, D. 1997. Duplication of the high molecular weight glutenin subunit gene in bread wheat (*Triticum aestivum* L.) cultivar 'Red River 68'. *Plant Breed.* **116**(6): 525–531. doi:10.1111/j.1439-0523.1997.tb02184.x.
- Dvořák, 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**(1): 21–31. doi:10.1139/g93-004. PMID:18469969.
- Flavell, R.B., Goldsbrough, A.P., 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. doi:10.1038/nbt1289-1281.
- Forde, J., Malpica, J.-M., Halford, N.G., Shewry, P.R., Anderson, O.D., Greene, F.C., and Mifflin, B.J. 1985. The nucleotide sequence of a HMW glutenin subunit gene located on chromosome 1A of wheat (*Triticum aestivum* L.). *Nucleic Acids Res.* **13**(19): 6817–6832. doi:10.1093/nar/13.19.6817. PMID:2997729.
- Gu, Y.Q., Anderson, O.D., Londeoré, C.F., Kong, X., Chibbar, R.N., and Lazo, G.R. 2003. Structural organization of the barley D-hordein locus in comparison with its orthologous regions of wheat genomes. *Genome*, **46**(6): 1084–1097. doi:10.1139/g03-071. PMID:14663527.
- Gu, Y.Q., Coleman-Derr, D., Kong, X., and Anderson, O.D. 2004. Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four Triticeae genomes. *Plant Physiol.* **135**(1): 459–470. doi:10.1104/pp.103.038083. PMID:15122014.
- Gu, Y.Q., Salse, J., Coleman-Derr, D., Dupin, A., Crossman, C., Lazo, G.R., et al. 2006. Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. *Genetics*, **174**(3): 1493–1504. doi:10.1534/genetics.106.060756. PMID:17028342.
- Halford, N.G., Forde, J., Shewry, P.R., and Kreis, M. 1989. Functional analysis of the upstream regions of a silent and an ex-

- pressed member of a family of wheat seed protein genes in transgenic tobacco. *Plant Sci.* **62**(2): 207–216. doi:10.1016/0168-9452(89)90083-6.
- Harberd, N.P., Flavell, R.B., and Thompson, R.D. 1987. Identification of a transposon-like insertion in a *Glu-1* allele of wheat. *Mol. Gen. Genet.* **209**(2): 326–332. doi:10.1007/BF00329661. PMID:17191344.
- Kasarda, D.D. 1999. Glutenin polymers: the *in vitro* to *in vivo* transition. *Cereal Foods World*, **44**: 566–571.
- Kong, X.Y., Gu, Y.Q., You, F.M., Dubcovsky, J., and Anderson, O.D. 2003. Dynamics of the evolution of orthologous and paralogous portions of a complex locus region in two genomes of allopolyploid wheat. *Plant Mol. Biol.* **54**(1): 55–69. doi:10.1023/B:PLAN.0000028768.21587.dc.
- MacRitchie, F., du Cros, D.L., and Wrigley, C.W. 1990. Flour polypeptides related to wheat quality. *Adv. Cereal Sci. Technol.* **10**: 79–145.
- Marchylo, B.A., Lukow, O.M., and Kruger, J.E. 1992. Quantitative variation in high molecular weight glutenin subunit 7 in some Canadian wheats. *J. Cereal Sci.* **15**: 29–37.
- Margiotta, B., Urbano, M., Colaprico, G., Johansson, E., Buonocore, F., D'Ovidio, R., and Lafiandra, D. 1996. Detection of y-type subunit at the *Glu-A1* locus in some Swedish bread wheat lines. *J. Cereal Sci.* **23**(3): 203–211. doi:10.1006/jcsc.1996.0021.
- McIntosh, R.A., Yamazaki, Y., Devos, K.M., Dubcovsky, J., Rogers, W.J., and Appels, R. 2003. Catalogue of gene symbols for wheat [online]. Available from <http://wheat.pw.usda.gov/ggpages/wgc/2003/>.
- Mitrovich, Q.M., and Anderson, P. 2005. mRNA surveillance of expressed pseudogenes in *C. elegans*. *Curr. Biol.* **15**(10): 963–967. doi:10.1016/j.cub.2005.04.055. PMID:15916954.
- Nagy, I.J., Takács, I., Juhász, A., Tamás, L., and Bedő, Z. 2004. Identification and characterization of new chimeric storage protein genes from an old Hungarian wheat variety. In *The gluten proteins*. Edited by D. Lafiandra, S. Masci, and R. D'Ovidio. The Royal Society of Chemistry, Cambridge. pp. 66–69.
- Nagy, I.J., Takács, I., Juhász, A., Tamás, L., and Bedő, Z. 2005. Identification of a new class of recombinant prolamin genes in wheat. *Genome*, **48**(5): 840–847. doi:10.1139/g05-042. PMID:16391690.
- Niu, J., Liu, R., Wang, L., and Lei, Z. 2006. Cloning of Xa21-like protein kinase genes and its sequence comparison at the orthologous loci in *Triticum*. *Henan Agricultural University Journal*, **41**: 475–479.
- Osborne, T.B. 1907. The proteins of the wheat kernel. Carnegie Institute, Washington.
- Payne, P.I., Holt, L.M., and Law, C.N. 1981. Structural and genetic studies on the high-molecular-weight subunits of wheat glutenin. Part 1: Allelic variation in subunits amongst varieties of wheat (*Triticum aestivum*). *Theor. Appl. Genet.* **60**(4): 229–236. doi:10.1007/BF02342544.
- Ragupathy, R., Naeem, H.A., Reimer, E., Lukow, O.M., Sapirstein, H.D., and Cloutier, S. 2008. Evolutionary origin of the segmental duplication encompassing the wheat *GLU-B1* locus encoding the overexpressed Bx7 (Bx7^{OE}) high molecular weight glutenin subunit. *Theor. Appl. Genet.* **116**(2): 283–296. doi:10.1007/s00122-007-0666-2. PMID:17985111.
- Shewry, P.R., and Halford, N.G. 2002. Cereal seed storage proteins: structures, properties and role in grain utilization. *J. Exp. Bot.* **53**(370): 947–958. doi:10.1093/jexbot/53.370.947. PMID:11912237.
- Shewry, P.R., Halford, N.G., and Tatham, A.S. 1992. The high molecular weight subunits of wheat glutenin. *J. Cereal Sci.* **15**: 105–120.
- Shewry, P.R., Halford, N.G., and Lafiandra, D. 2003. Genetics of wheat gluten proteins. *Adv. Genet.* **49**: 111–184. doi:10.1016/S0065-2660(03)01003-4. PMID:12779252.
- Soltis, D.E., and Soltis, P.S. 1999. Polyploidy: recurrent formation and genome evolution. *Trends Ecol. Evol.* **14**(9): 348–352. doi:10.1016/S0169-5347(99)01638-9. PMID:10441308.
- Sugiyama, T., Rafalski, A., Peterson, D., and Söll, D. 1985. A wheat HMW glutenin subunit gene reveals a highly repeated structure. *Nucleic Acids Res.* **13**(24): 8729–8737. doi:10.1093/nar/13.24.8729. PMID:3001648.
- Sun, M., Yan, Y., Jiang, Y., Xiao, Y., Hu, Y., Cai, M., et al. 2004. Molecular cloning and comparative analysis of a y-type inactive HMW glutenin subunit gene from cultivated emmer wheat (*Triticum dicoccum* L.). *Hereditas*, **141**(1): 46–54. doi:10.1111/j.1601-5223.2004.01835.x. PMID:15383071.
- Thompson, R.D., Bartels, D., Harberd, N.P., and Flavell, R.B. 1983. Characterization of the multigene family coding for HMW glutenin subunits in wheat using cDNA clones. *Theor. Appl. Genet.* **67**(1): 87–96. doi:10.1007/BF00303930.
- 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**(19): 6833–6846. doi:10.1093/nar/13.19.6833. PMID:3840588.
- Wu, X., Larson, S.R., Hu, Z., Palazzo, A.J., Jones, T.A., Wang, R.R.-C., et al. 2003. Molecular genetic linkage maps for allotetraploid *Leymus* wildryes (Gramineae: Triticeae). *Genome*, **46**(4): 627–646. doi:10.1139/g03-048. PMID:12897871.

Copyright of *Genome* is the property of NRC Research Press and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.