

Comparative sequence analysis of the *Phytochrome C* gene and its upstream region in allohexaploid wheat reveals new data on the evolution of its three constituent genomes

Katrien M. Devos^{1,4,*}, James Beales¹, Yasunari Ogihara² and Andrew N. Doust³

¹John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH, UK; ²Kihara Institute for Biological Research, Yokohama City University, 244-0813, Yokohama, Japan; ³Department of Biology, University of Missouri, St. Louis, MO, 63121, USA; ⁴Department of Crop and Soil Sciences, and Department of Plant Biology, 3111 Miller Plant Sciences Building, University of Georgia, Athens, GA, 30602, USA (*author for correspondence; e-mail kdevos@uga.edu)

Received 14 July 2004; accepted in revised form 29 April 2005

Key words: comparative sequence analysis, evolution, *Phytochrome C*, *r8s*, *Triticum aestivum*, wheat

Abstract

Bread wheat is an allohexaploid with genome composition AABBDD. *Phytochrome C* is a gene involved in photomorphogenesis that has been used extensively for phylogenetic analyses. In wheat, the *PhyC* genes are single copy in each of the three homoeologous genomes and map to orthologous positions on the long arms of the group 5 chromosomes. Comparative sequence analysis of the three homoeologous copies of the wheat *PhyC* gene and of some 5 kb of upstream region has demonstrated a high level of conservation of *PhyC*, but frequent interruption of the upstream regions by the insertion of retroelements and other repeats. One of the repeats in the region under investigation appeared to have inserted before the divergence of the diploid wheat genomes, but was degraded to the extent that similarity between the A and D copies could only be observed at the amino acid level. Evidence was found for the differential presence of a foldback element and a miniature inverted-repeat transposable element (MITE) 5' to *PhyC* in different wheat cultivars. The latter may represent the first example of an active MITE family in the wheat genome. Several conserved non-coding sequences were also identified that may represent functional regulatory elements. The level of sequence divergence (K_s) between the three wheat *PhyC* homoeologs suggests that the divergence of the diploid wheat ancestors occurred some 6.9 Mya, which is considerably earlier than the previously estimated 2.5–4.5 Mya. K_a/K_s ratios were < 0.15 indicating that all three homoeologs are under purifying selection and presumably represent functional *PhyC* genes. RT-PCR confirmed expression of the A, B and D copies. The discrepancy in evolutionary age of the wheat genomes estimated using sequences from different parts of the genome may reflect a mosaic origin of some of the Triticeae genomes.

Introduction

Bread wheat (*T. aestivum*) is an allohexaploid that arose some 8000 years ago through hybridization of the tetraploid *T. turgidum* subsp. *dicoccum* (genome composition AABB) with the diploid *Ae. tauschii* (DD). Evolutionary studies based on two nuclear

encoded genes, the plastid acetyl-CoA carboxylase (*Acc-1*) and 3-phosphoglycerate kinase (*Pgk-1*), estimated the hybridization of *T. urartu* (AA) with an unknown B genome species to have occurred less than 0.5 million years ago (Mya), and the divergence of the ancestral diploid A, B and D genomes between 2.5 and 4.5 Mya (Huang *et al.*, 2002b).

These calculations are based on a total of less than 2000 bp of sequence alignment.

To date, most comparative studies carried out at the sequence level between orthologous regions of different species have involved species that diverged earlier than 10 Mya (Chen *et al.*, 1997; Gallego *et al.*, 1998; Tikhonov *et al.*, 1999; Ilic *et al.*, 2003; Brunner *et al.*, 2003). Some data are beginning to emerge on the structural evolution of different Triticeae genomes (Wicker *et al.*, 2003; Kong *et al.*, 2004), but sequence comparisons of the A, B and D genomes within allohexaploid wheat have, so far, been limited to a few genes (Aoki *et al.*, 2002; Huang *et al.*, 2002b) and none involved intergenic DNA. The occurrence of substantial sequence divergence, in particular in the non-genic component of the wheat genome, has been suggested by several data sets. First, RFLP analyses have demonstrated the presence of high levels of intergenomic variation between the three wheat genomes. Secondly, comparative sequence analysis within the Triticeae have shown that closely related genomes differ by the presence of retroelements (Wicker *et al.*, 2003; Kong *et al.*, 2004). Lastly, there is evidence that polyploidization can lead to rapid genomic changes (Liu *et al.*, 1998a, b).

To study the comparative organization and evolution of the three wheat genomes, we embarked on an analysis of the *Phytochrome C* genes and their upstream regions in the A, B and D genomes. Phytochromes are a family of photoreceptors that mediate plant growth and development in response to varying light conditions in the red/far-red spectrum (Furuya, 1993; Smith, 1995). In their active form, phytochromes can bind to transcriptional regulators and thereby facilitate transcription of specific photo-responsive genes (Martínez-García *et al.*, 2000; Quail, 2000). In *Arabidopsis*, the phytochrome family consists of five members, *PhyA–PhyE*, while in monocots only *PhyA*, *PhyB* and *PhyC* have been identified (Clack *et al.*, 1994; Mathews and Sharrock, 1996). Their functions, in particular those of *PhyA* and *PhyB* have been well studied (reviewed in Jackson and Thomas, 1997; Deng and Quail, 1999; Smith, 2000). *PhyC* was recently shown to have a photo-sensory specificity similar to that of *PhyB* (Monte *et al.*, 2003). Based on the analysis of mutant *PhyC* alleles, Monte and colleagues suggested that *PhyC* plays a role in the perception of day length and is involved in photomorphogenesis throughout the life cycle of the plant.

Phytochrome sequences have previously been used to estimate phylogenetic relationships (Mathews and Sharrock, 1996), although these studies did not include species within the *Triticum/Aegilops* complex. Partial *PhyC* sequence information is available for some 200 species and full length *PhyC* sequences have been isolated from *Arabidopsis* (Cowl *et al.*, 1994), rice (Basu *et al.*, 2000) and sorghum (White *et al.*, 2004). The monocot *PhyC* genes consist of four exons, which, in rice, are 2065-, 817-, 294- and 238-bp-long, resulting in a protein of 1137 amino acids. We isolated *PhyC* genes from the A, B and D genomes of wheat, together with several kilobases of upstream sequence and analyzed them for structural conservation. We combined analysis of the sequence structure of the three *PhyC* genes with evidence of divergence times calculated from pairwise substitution rates and from phylogenetic analysis. Topologies and divergence times of *PhyC* and three other genes (*Acc-1*, *Pgk-1*, and *Sut1*) were compared in order to make inferences on the evolution of the three diploid genomes in bread wheat.

Materials and methods

Plant materials

Seeds from Chinese Spring nullisomic–tetrasomic (NT) (Sears, 1954) and ditelosomic (DT) lines (Sears and Sears, 1979), and from the varieties Chinese Spring, Soleil, Reward, Saitama 27 and *T. spelta* accession ‘Grey’ were obtained from the John Innes Centre germplasm collection.

Generation of a PhyC probe

Three forward and three reverse primers were designed to conserved regions of the *PhyC* sequence, identified after alignment of the rice (GenBank acc. number AB018442) and sorghum *PhyC* (acc. U56731) sequences. The primer sequences are: PCCF1: 5′-GAGATGCTCGACCTCACGC-3′; P CCF2: 5′-TATCTTGGCCTGCACTACCC-3′; P CCF3: 5′-GAAGATGCATCCACGATCTTC-3′; PCCR1: 5′-TTGCAAGAGAAA-CTCGCAAGC-3′; PCCR2: 5′-GCATCCATTTCAACATCCTCC-3′; and PCCR3: 5′-TAAGCAG-GAACCAAGATC ATTG-3′. Primer combinations

PCCF2/PCCR1 and PCCF2/PCCR2, which are located within exon 1, and PCCF3/PCCR3, which span intron 1, amplified single fragments from each of the A, B and D genomes of the hexaploid wheat variety Chinese Spring. Fragments amplified from the wheat aneuploid lines nullisomic5A–tetrasomic5B (N5AT5B; lacks chromosome 5A and has two pairs of chromosome 5B), N5BT5D (lacking chromosome 5B and carrying two pairs of 5D) and N5DT5B (lacking 5D and carrying two pairs of 5B) were cloned in the pGEM-T Easy vector (Promega). Several clones were sequenced for each of the lines.

Assignment of the chromosomal location and mapping of PhyC

A cloned PCCF2/PCCR1 fragment was used as a hybridization probe to *Hind*III digested DNA of a set of 21 NT lines and 23 DT lines of Chinese Spring wheat. Mapping was conducted in a population of 96 doubled haploid lines derived from an F₁ from the cross Chinese Spring × SQ1 (Quarrie *et al.*, 1994). Restriction fragment length polymorphism (RFLP) procedures and linkage analysis were carried out as described in Devos *et al.* (1992).

Library screening

A lambda library of the variety Soleil was provided by P. Bailey and J. Flintham, John Innes Centre, UK. The library was generated by cloning Soleil DNA, partially digested with *Mbo*I, into the vector lambda FIX II (Stratagene). The library was screened with a cloned PCCF2/PCCR1 wheat fragment. Probe labeling was carried out as described in Devos *et al.* (1992). Primary, secondary and tertiary screens and phage DNA extractions were carried out according to standard protocols (Sambrook *et al.*, 2001).

A TAC library of the hexaploid wheat variety Chinese Spring was screened with primer pairs PCCF2/PCCR2 and PCCF3/PCCR3. A description of the TAC library and screening methods are given in Liu *et al.* (2000).

Restriction mapping

Phage DNA was digested with the restriction enzymes *Not*I, *Sal*I, *Hind*III, *Sst*I and with the enzyme combinations *Not*I/*Hind*III and *Hind*III/

*Sal*I. TAC DNA was digested with *Hind*III, *Eco*RI, *Eco*RV, *Dra*I, *Pst*I, *Not*I, *Xho*I and with *Hind*III in combination with each of these enzymes. Fragments were separated on 0.8% agarose gels, transferred to nylon filters and hybridized using a cloned PCCF2/PCCR1 fragment as probe. Probe labeling, hybridization and washes were carried out as described in Devos *et al.* (1992).

DNA sequencing

On the basis of the restriction maps, a 2.8 kb *Hind*III/*Not*I fragment (the *Not*I site is located in the vector polylinker) carrying the PCCF2/PCCR1 sequence and an upstream 4 kb *Hind*III fragment were subcloned from the A-genome phage. A 6.5 kb *Hind*III/*Sal*I subclone, carrying the PCCF2/PCCR1 sequence was obtained from the B-genome phage. Three fragments, covering a total of some 8 kb, were subcloned from the D-genome TAC clone. All fragments were cloned into the pBluescript II SK M13(+) plasmid vector (Stratagene). Subclones were sequenced by primer walking using the Big Dye Terminator v.3 kit and separation of the fragments on an ABI 3700 (Applied Biosystems). Once sequence information was obtained from one of the genomes, it was possible to design primers that amplified homologous fragments from the other two wheat genomes. Consequently, some of the sequences, including the 3' ends of the *PhyC-5A* and *PhyC-5B* genes that were missing from the phage clones, were obtained using PCR product as template. To sequence the *PhyC-5A* region from the varieties Chinese Spring, Reward, Saitama 27 and from *T. spelta* acc. 'Grey', and the *PhyC-5D* region of Soleil, a series of overlapping primer pairs covering the entire gene were designed against the 5A Soleil and 5D Chinese Spring template, respectively. The PCR products were cloned in the pGEM-T Easy vector (Promega) and sequenced. At least two clones from independent PCR reactions were sequenced to eliminate PCR errors from the sequence. In case of a discrepancy, the ambiguous sequence was compared with available sequence from one or both of the other wheat genomes and a conserved base rather than a SNP was accepted as the correct sequence. If the discrepancy could not be resolved by sequence comparison, a third clone generated from yet another independent PCR was sequenced.

Sequence annotation

The *PhyC* start and stop codons, and the intron–exon splice sites were located by comparing the structure of the wheat *PhyC* genes with that of rice. The upstream intergenic region was annotated by BLASTN analysis to the Triticeae repeats database TREP (<http://wheat.pw.usda.gov/ggpages/ITMI/Repeats/index.shtml>) and to the ‘nr’ section of Genbank, and by BLASTX analysis against the ‘nr Peptide Sequence Section’ of Genbank. Conserved motifs present in the 5′ upstream regions of the wheat and rice (acc. AF141942) *PhyC* genes were identified using the program VISTA (Dubchak *et al.*, 2000). Criteria used were a minimum of 70% identity over the four aligned sequences in a 20 bp window.

RT-PCR

To test whether all three *PhyC* homoeologs are transcribed, RT-PCR was carried out using forward primer PCCF23 (5′-AGGTTTCGTCCTTCCCTC-3′) and reverse primer PCCR18 (5′-ATAGGGGGTATGAGCTCATTG-3′) on first-strand cDNA synthesized from RNA extracted from 4-week-old Reward seedlings. RNA was extracted using the RNeasy Plant Mini kit (Qiagen) and first strand synthesis was conducted using an oligo(dT) primer and Superscript II reverse transcriptase (RT). The PCCF23/PCCR18 primer pair spans intron 1, hence RT-PCR products are 109 bp shorter than products amplified from any contaminating DNA that might be present in the RNA preparations. PCR products were subsequently digested with three restriction enzymes, *BtsI*, *EcoRV* and *NsiI*, separated on a 5% denaturing polyacrylamide gel and visualized by silver staining.

Phylogenetic analysis

Representative sequences of *PhyC* coding regions from across the major grass groups (see Supplementary Material) were analyzed with PAUP 4.0b10 (Swofford, 2001), using a maximum likelihood (ML) heuristic search with 10 random addition sequences, tree bisection reconnection branch swapping, and steepest descent activated. A reduced data set (see Supplementary Material), using exemplar taxa from each clade produced in

the first analysis was then constructed and compared with the full data set to ensure that topological relationships remained the same. The correct likelihood model for both data sets was calculated using MrModeltest (Nylander, 2002). A likelihood ratio test was used to compare alternative phylogenies, constrained by the assumption of substitution rate constancy (‘molecular clock’), against the unconstrained phylogeny (Felsenstein, 2004). 1000 ML bootstraps of each data set were analyzed, using the ‘fast’ heuristic bootstrap method in PAUP 4.0b10.

The *Acc-1* and *Pgk-1* data sets from Huang *et al.* (2002a) were also analyzed by ML, with both full and reduced data sets (data sets are given in Supplementary Material). Only coding regions were used as non-coding regions could not be unambiguously aligned. A fourth data set consisted of the coding region of sucrose transporter genes (*Sut1*) from the A, B and D genomes of wheat (acc. AF408842, AF408843 and AF408844; Aoki *et al.*, 2002), barley (acc. AJ272309), maize (acc. AB008464) and rice (acc. AF280050). In all data sets, analyses included appropriate outgroup taxa (*PhyC*: *Joinvillea*; *Acc-1*: *Brassica napus* [cytosolic *Acc*], *Lolium rigidum* [cytosolic *Acc-2*]; *Pgk-1*: *Arabidopsis thaliana*; *Sut1*: *Oryza sativa* [*Sut3*]).

Calculating divergence times

Divergence times (*T*) were calculated by two different methods. One uses the same methods of pairwise sequence comparisons as previous analyses of divergence times in the Pooid grasses (Huang *et al.*, 2002a; Huang *et al.*, 2002b), and assumes a molecular clock. The pairwise sequence comparisons estimate the absolute rate of synonymous substitutions per site per year (*k*) between two sequences whose divergence time can be calibrated by fossils or geological events, using the formula $k = K_s/2T$, where K_s is the estimated number of synonymous substitutions per site between orthologous sequences and *T* is the time elapsed since the two sequences diverged. Absolute divergence times were calibrated with the timing of 60 My for the divergence of the Pooid and Panicoid grasses (50–70 My; Wolfe *et al.*, 1989). From the absolute rate *k*, it is then possible to compare K_s values for the divergence of other pairs of sequences and thus date them relative to the

calibrated substitution rate. The same method was used to date a duplication present in the *PhyC-5A* upstream region. K_s values were estimated using the Nei-Gojobori method and the Jukes-Cantor correction as implemented in the program MEGA2 (Kumar *et al.*, 2001). For non-genic regions, overall nucleotide substitutions (K) were used in the equation, and were estimated using the Kimura 2-parameter method in MEGA2.

Divergence times were also calculated using a ML tree-based approach (*r8s* version 1.50, Sanderson, 2002; Sanderson, 2003), in order to compare differences given by the various genes for the divergence dates of particular nodes in the trees. The same fossil dating was used in this method as in the pairwise sequence comparisons. Divergence times were calculated for both the reduced *PhyC* coding data set and for the coding sequences for *Acc-1* and *Sut1*, in order to compare the divergence times of designated clades for each of the genes. *Pgk-1* was not used because the nodes on the tree constructed from the coding regions of this gene that were to be compared for divergence times had no support. The taxa used in each analysis are given in Supplementary Material.

Trees were analyzed in *r8s* using the semi-parametric penalized likelihood (PL) method (Sanderson, 2002). The method allows branches to vary in substitution rate but imposes a smoothing parameter which limits how quickly rates can vary. The best value for the smoothing parameter was estimated in each data set by cross-validation using multiple values for the parameter. High values of the smoothing parameter approximate analyses that assume a molecular clock while low values allow rates to vary without constraint (Sanderson, 2003).

The optimal ML trees estimated in PAUP were saved as rooted trees with branch lengths, with outgroups being deleted before analysis of the trees in *r8s*. The truncated-Newton algorithm was used for the analyses, and all age estimations were started three times with different starting positions to reduce the possibility of being caught in sub-optimum local optima. Analyses used a divergence time for the Panicoid (*Zea*, *Sorghum*)–Poooid (*Triticum*, *Secale*, *Hordeum*) split of 60 Mya as was also done for the pairwise nucleotide substitution comparisons.

Distributions of dates for each node in the phylogeny were calculated by producing ML trees

from 1000 bootstrapped data sets (generated by the Seqboot program in Phylip [Felsenstein, 1993]), constrained by the topology of the ML tree from the original data set. *R8s* was then used to calculate divergence times for designated clades in each tree. These times were collated at the end of the analysis and 95% confidence intervals calculated for each clade.

The divergence dates for particular nodes gained from the different genes can be compared in a variety of ways, although it is not straightforward to construct an appropriate test to ascertain whether the dates are significantly different. As a first approximation, confidence intervals for a node calculated from the distribution of dates produced by bootstrapping were compared. We also compared pairs of dates for a node derived from different genes using a resampling approach (Simon, 1995), to test the null hypothesis that the divergence dates for each gene were sampled from the same underlying age distribution. Divergence dates for a pair of genes to be compared were pooled, and 10 000 resampled pairs of dates were obtained from the combined distribution. The differences between the sampled dates were collated, and this distribution was used to test whether the measured difference between the original divergence dates given by different genes was more extreme than the middle 95% of the resampled distribution.

Results

Chromosomal location of the PhyC gene in wheat

Nullisomic–tetrasomic and ditelosomic analysis established that the *PhyC* gene was single copy in the wheat genome and was located on the long arms of the homoeologous group 5 chromosomes. Mapping refined the location of the 5A locus, which was polymorphic in a Chinese Spring X SQ1 doubled haploid population, to the distal region of the long arm, some 90 cM from the centromere. The *PhyC* gene was shown to cosegregate with the vernalization response locus, *Vrn-A1*. Analysis in *T. monococcum*, however, has shown that *PhyC* does not underlie the *Vrn-1* locus, but is located some 300 kb from *APETALA1*, the gene conferring vernalization response at *Vrn-A^m1* (Yan *et al.*, 2003).

Identification of wheat genomic clones carrying PhyC

Screening of the Chinese Spring TAC library with primer combinations PCCF2/PCCR2 and PCCF3/PCCR3 identified three positive clones. All three clones were detected with both primer sets. The Soleil lambda library, screened with a cloned PCCF2/PCCR1 wheat fragment, yielded three positive clones. *Hind*III fingerprints of the three TAC and three lambda clones suggested that the clones originated from three different wheat genomes. A comparison of the patterns obtained after hybridization of the PCCF2/PCCR2 fragment to *Hind*III digested DNA of each of the clones and of the group 5 nullisomic-tetrasomic lines revealed that two lambda clones were derived from the A genome and one from the B genome. All TAC clones were derived from the D genome. One representative clone from each of the three genomes was selected for sequencing.

Sequence analysis of PhyC-5A, PhyC-5B and PhyC-5D

A total of 9533 bp was sequenced of Soleil *PhyC-5A*, including some 5 kb upstream of the ATG start codon. 5525 bp of sequence was obtained for Soleil *PhyC-5B* and 7992 bp of sequence for Chinese Spring *PhyC-5D*. In addition, some 5500 bp, covering the *PhyC-5A* gene and 1400 bp of upstream sequence, were obtained from the varieties Chinese Spring, Saitama 27, Reward and *T. spelta* acc. 'Grey'. Sequences have been deposited in GenBank

under accession numbers AY672994 (*PhyC-5A* – *T. spelta* acc. 'Grey'), AY672995 (*PhyC-5A* – Chinese Spring), AY672996 (*PhyC-5A* – Reward), AY672997 (*PhyC-5A* – Saitama 27); AY672998 (*PhyC-5A* – Soleil), AY672999 (*PhyC-5B* – Soleil), AY673000 (*PhyC-5D* – Soleil), AY673001 and AY673002 (*PhyC-5D* – Chinese Spring). During the course of our experiments, a *PhyC-5A* sequence from the variety CPAN 1676 was submitted to Genbank by R. Kulshreshtha (acc. AJ295224). The sequence differed from that of Soleil by nine base substitutions. Seven of the differences, six of which were in adjacent codons, led to four amino acid differences. Two of these were amino acids that were otherwise conserved in phytochrome sequences across dicots and monocots. They are likely to represent sequencing errors in the CPAN sequence and therefore this sequence was not included in our analysis. Some 5500 bp of sequence was also obtained from the Soleil *PhyC-5D* locus. An alignment of the wheat PHYC proteins can be found as Supplementary Material.

Soleil *PhyC-5A* has an open reading frame of 4196 bp and contains three introns of 109 bp, 322 bp and 346 bp (Figure 1A). The introns in *PhyC-5B* and *PhyC-5D* are 107, 317 and 363 bp, and 110, 317 and 355 bp, respectively. The rates of synonymous (K_s) and non-synonymous substitutions (K_a) between *PhyC-5A-PhyC-5B*, *PhyC-5A-PhyC-5D* and *PhyC-5B-PhyC-5D* are given in Table 1. Most of the intergenomic variation is present in the introns and in the 3rd codon position. In all cases, K_a/K_s values were <0.15 . This indicates that *PhyC* is under strong selection

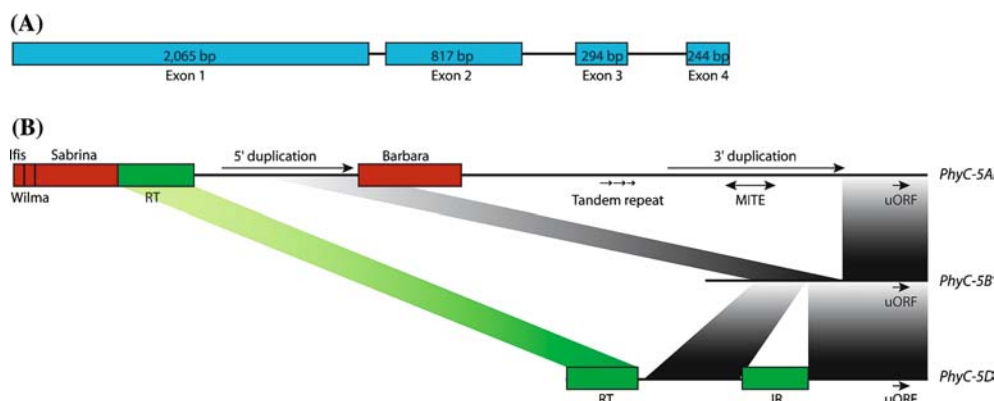


Figure 1. Structure of the *PhyC* genes in wheat (A) and of their upstream regions (B). RT indicates similarity to a reverse transcriptase; IR stands for inverted repeat. Connecting bars between *PhyC-5A*, *PhyC-5B* and *PhyC-5D* indicate regions of similarity.

Table 1. K_s values and standard errors for pairwise comparisons of the Soleil *PhyC-5A*, *PhyC-5B* and *PhyC-5D* genes.

	<i>PhyC-5A</i>			<i>PhyC-5B</i>		
	Introns	Exons (K_s)	Exons (K_a)	Introns	Exons (K_s)	Exons (K_a)
<i>PhyC-5B</i>	0.051 (± 0.008)	0.064 (± 0.009)	0.007 (± 0.002)			
<i>PhyC-5D</i>	0.030 (± 0.006)	0.052 (± 0.008)	0.005 (± 0.001)	0.048 (± 0.008)	0.053 (± 0.008)	0.007 (± 0.002)

for retained function. Intervarietal variation in the coding region of *PhyC-5A* was limited to 10 single nucleotide polymorphisms (SNPs) across the sample of five varieties. Five of the SNPs resulted in amino acid changes, two of which characterized *T. spelta* acc. 'Grey', two were specific to Saitama 27 and one to Soleil. The introns of the *PhyC-5A* genes were identical with the exception of a 1-bp deletion in Soleil in intron 2. No differences were observed between the coding regions of the Soleil and Chinese Spring *PhyC-5D* genes.

The region 5' to PhyC

PhyC-5A

The 5300 bp upstream of the *PhyC-5A* ATG start codon consist mainly of repetitive DNA (Figure 1B). The first 650 bp comprises remnants of long terminal repeats (LTRs) of the retroelements Ifis, Wilma and Sabrina inserted into each other into a nested fashion. The sequence adjacent to Sabrina has homology at the amino acid level to a putative non-LTR retroelement reverse transcriptase. An interesting feature of the region is the presence of an 800 bp duplication (Figure 1B). The duplicated copies are separated by a truncated retroelement 'Barbara' (620 bp), 797 bp sequence of unspecified nature and 2.6 units of an 85 bp tandem repeat. The duplicated region ends 504 bp upstream of the *PhyC-5A* start codon. A miniature inverted-repeat transposable element (MITE) belonging to the *Tourist* family is present in the 3' duplication. In the varieties Chinese Spring and Reward, MITEs that are 98% homologous have been inserted in orthologous positions at both duplicated loci.

We were unable to establish the transcription start site of *PhyC-5A*. 5' RACE PCR failed, presumably due to the high GC-content in the 5' UTR and no full-length wheat cDNAs for *PhyC* are present in Genbank. In rice, the *PhyC* 5'-UTR comprises 715 bp (Basu *et al.*, 2000), some 580 bp of which show homology to wheat. It is thus

possible that the *PhyC-5A* promoter in wheat extends into the duplicated region. One hundred and forty eight basepairs upstream of the ATG start codon, and thus likely located within the *PhyC-5A* 5'-UTR, a second start codon was found. This upstream open reading frame (uORF) encodes a 28 amino acid peptide.

PhyC-5B

The 1123 bp immediately upstream of *PhyC-5B* are highly similar to the A genome sequence (Figure 1B). Similarly to *PhyC-5A*, *PhyC-5B* contains an uORF of 28 amino acids. The region of similarity is not continuous. Interruptions of colinearity were caused by the insertion of retroelement Barbara and other repeats in the A genome. The most 5' 102 bp of the 1225 bp sequenced lacks detectable similarity both with the 5A and 5D sequence. BLAST searches carried out at the nucleotide as well as at the amino acid level also failed to identify any similarity with Genbank entries. The nature of this region remains undefined.

PhyC-5D

A total of 2070 bp of sequence information was obtained upstream of *PhyC-5D* in the variety Chinese Spring. The uORF in *PhyC-5D* encoded 31 amino acids, and the stop codon was located 55 bp upstream of the *PhyC-5D* initiation codon. Sequence similarity between A and D sequences extended over the same region as the A-B similarity and was interrupted by the A genome repeats and by an inverted repeat inserted into the 5D sequence. The repeat was not fully sequenced but the available sequence showed high similarity to a foldback element identified in *T. monococcum* BAC 231A16. The termini of the foldback element were 21 bp direct repeats (consensus sequence GTCAAATTTGTCAAATTTGAC), which themselves formed a hairpin structure. This foldback element was absent from Soleil, but part of the direct repeat was still

present. In *T. monococcum*, the foldback element was flanked by a 9 bp host duplication. No obvious host duplication was observed in the 5D element. Nevertheless, the terminal sequence of the element and the presence of a 9 bp target site duplication in *T. monococcum* suggest that the repeat is most likely a non-autonomous *Mu*-like element. The most 5' D genome region sequenced showed similarity at the amino acid level to the same non-LTR retrotransposon reverse transcriptase as the sequence adjacent to retroelement Sabrina in 5A. However, little similarity was observed at the DNA sequence level between these two regions.

Expression analysis

RT-PCR gave a single amplification product of 1180 bp. Because primer pair PCCF23/PCCR18 amplifies fragments of identical length from all three wheat genomes, RT-PCR products were digested with *BtsI*, *EcoRV* and *NsiI*. Following *BtsI* digestion, fragments of 1180 (uncut), 659 and 519 bp were obtained. *EcoRV* digestion generated fragments of 1180 (uncut), 823 and 355 bp, and *NsiI* digestion generated fragments of 657, 521, 462 and 195 bp. As the A genome sequence contains a single *EcoRV* site and two *NsiI* sites, the B genome sequence a single *NsiI* site, and the D genome sequence single sites for all three enzymes, we can conclude that the 462 and 195 bp *NsiI* products are amplified from cDNA transcribed from the A genome, the 1180 bp uncut fragment provides evidence for expression of *PhyC-5B*, and the 659 and 519 bp *BtsI* fragments for expression of *PhyC-5D* (Figure 2).

Phylogenetic relationships

PhyC marginally failed the assumption of a molecular clock either with ($P = 0.03$) or without ($P = 0.025$) the outgroup *Joinvillea* excluded from the analysis. However, *PhyC* was clocklike when the data was analyzed using the Jukes-Cantor model, which assumes equal probability of base changes (Felsenstein, 2004), and is the model that corresponds to the assumptions of the pairwise substitution approach for analyzing divergence times. *Sut1* marginally failed the assumption of a molecular clock with the outgroup sequence of the rice *Sut3* gene ($P = 0.03$), but was clocklike when

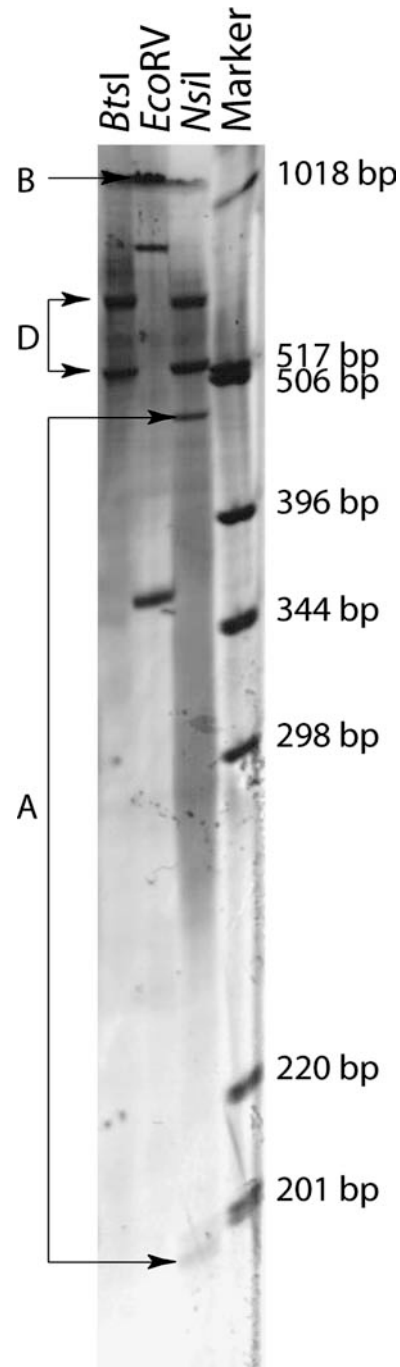


Figure 2. Wheat *PhyC* RT-PCR products, digested with *BtsI*, *EcoRV* and *NsiI* and separated on a 5% denaturing polyacrylamide gel. The 659 and 519 bp *BtsI* fragments, the uncut (1180 bp) *EcoRV* fragment, and the 462 and 195 bp *NsiI* fragments demonstrate expression of *PhyC-5D*, *PhyC-5B* and *PhyC-5A*, respectively.

this sequence was excluded. *Acc-1* satisfied the assumption of a molecular clock with or without outgroups, whereas *Pgk-1* was not clocklike in either case ($P < 0.001$).

The topology of the *PhyC* coding region tree had wheat genomes A and D as sister taxa, followed by wheat genome B, *Hordeum*, *Oryza*, and *Sorghum* (Figure 3). Of these, only the Triticeae (*Hordeum* and the three wheat genomes) were well supported as a clade (100% bootstrap support). The Triticeae plus *Oryza* had weak support (75%) as did the clade of the A and D wheat genomes (69%). The clade of the three wheat genomes was essentially not supported (56%).

The topology of the *Acc-1* coding region tree showed the three wheat genomes as being equally related to each other, followed by *Secale* and the S genome from *Aegilops sharonensis*, then *Hordeum* (Figure 3). *Oryza* and *Zea* form a clade of their own, although this is not supported. There is weak bootstrap support for the Triticeae (*Hordeum*, *Secale* and wheat; 75%) and for *Secale* and wheat (65%).

The topology of the *Pgk-1* coding region has the A genome sister to a clade of the B and D genomes (Figure 3). Like the *Acc-1* tree, the *Pgk-1* tree has *Zea* and *Oryza* forming a clade, rather than the Triticeae and *Oryza*, as in *PhyC*. The only bootstrap support for the topology formed from the *Pgk-1* coding region is for the grasses as a whole (60%), and for *Zea* and *Oryza* (99%), indicating that the sequences used were poor in phylogenetic information. No further analyses were conducted on this gene.

The topology of the *Sut1* coding region tree placed the wheat A and B genomes as sister taxa (83%), followed by the wheat D genome (90%), and *Hordeum* (95%) (Figure 3). This clade was sister to a clade comprising *Zea* and *Oryza*. These two taxa were supported as a clade (89%).

Divergence times

The preferred model of sequence evolution for *PhyC* (general time reversible + gamma [Felsenstein, 2004]), resulted in the *PhyC* data set being marginally non-clocklike. However, a simpler model of sequence evolution, with DNA base substitution rates being held equal (Jukes-Cantor) found the *PhyC* data set to be clocklike. As these are the same assumptions as in the pairwise

substitution rate method used by previous authors (Huang *et al.*, 2002a; Huang *et al.*, 2002b), we felt justified in exploring the pairwise substitution rate method to analyze divergence times for *PhyC*. To establish a molecular clock, we compared the *PhyC* coding sequences of wheat (*Triticum*; this study) and barley (*Hordeum*; acc. AF406643) with sorghum (acc. U56731). Based on the number of synonymous substitutions between the Triticeae (averaged over A, B, D and H) and sorghum *PhyC* sequences and using 60 My as the divergence time for the Pooideae (includes wheat, barley and rye) and Panicoideae (includes maize and sorghum) lineages (50–70 Mya; Wolfe *et al.*, 1989), the average substitution rate per synonymous site per year was estimated to be 4.4×10^{-9} . Assuming that the *PhyC* genes have evolved at the same rate in the different lineages, we can use the value of 4.4×10^{-9} to calculate the divergence of the diploid ancestors of the A, B and D genomes of hexaploid bread wheat. These data suggest that the three wheat genomes diverged some 6.4 ± 0.9 Mya. The *Triticum–Hordeum* divergence was estimated to have occurred 14.8 ± 1.9 Mya (Table 2).

The same pairwise methodology was used to analyze divergence times with *Acc-1*. The clock was calibrated on the average number of synonymous substitutions between the Triticeae wheat (acc. AF343519, AF343510 and AF343496) and barley (acc. AF343509), and the Panicoid maize *Acc-1* (*Zea*; AF342954) coding regions and an absolute age of 60 My for their divergence, leading to a k -value of 3.5×10^{-9} . Based on this k value, the divergence of the A, B and D genomes was estimated to have taken place 1.8 ± 1.3 Mya and the *Triticum–Hordeum* divergence 10.4 ± 3.1 Mya (Table 2).

The *Sut1* gene was clocklike in analyses that excluded the outgroup sequence *Oryza sativa Sut3*. For *Sut1*, the clock was also calibrated on the

Table 2. Divergence times with standard errors as calculated by the pairwise comparison method given a divergence age of 60 Mya for the divergence of the Poooid and Panicoid grasses (i.e. between the Triticeae genomes and *Zea* or *Sorghum*).

Divergence	<i>PhyC</i>	<i>Acc-1</i>	<i>Sut1</i>
Wheat A–Wheat D	5.9 ± 0.9	1.7 ± 1.3	8.0 ± 1.5
Wheat A–Wheat B	7.3 ± 1.0	0.9 ± 0.9	5.6 ± 1.3
Wheat B–Wheat D	6.0 ± 0.9	2.7 ± 1.6	6.4 ± 1.4
Wheat– <i>Hordeum</i>	14.8 ± 1.9	10.4 ± 3.1	15.9 ± 2.2

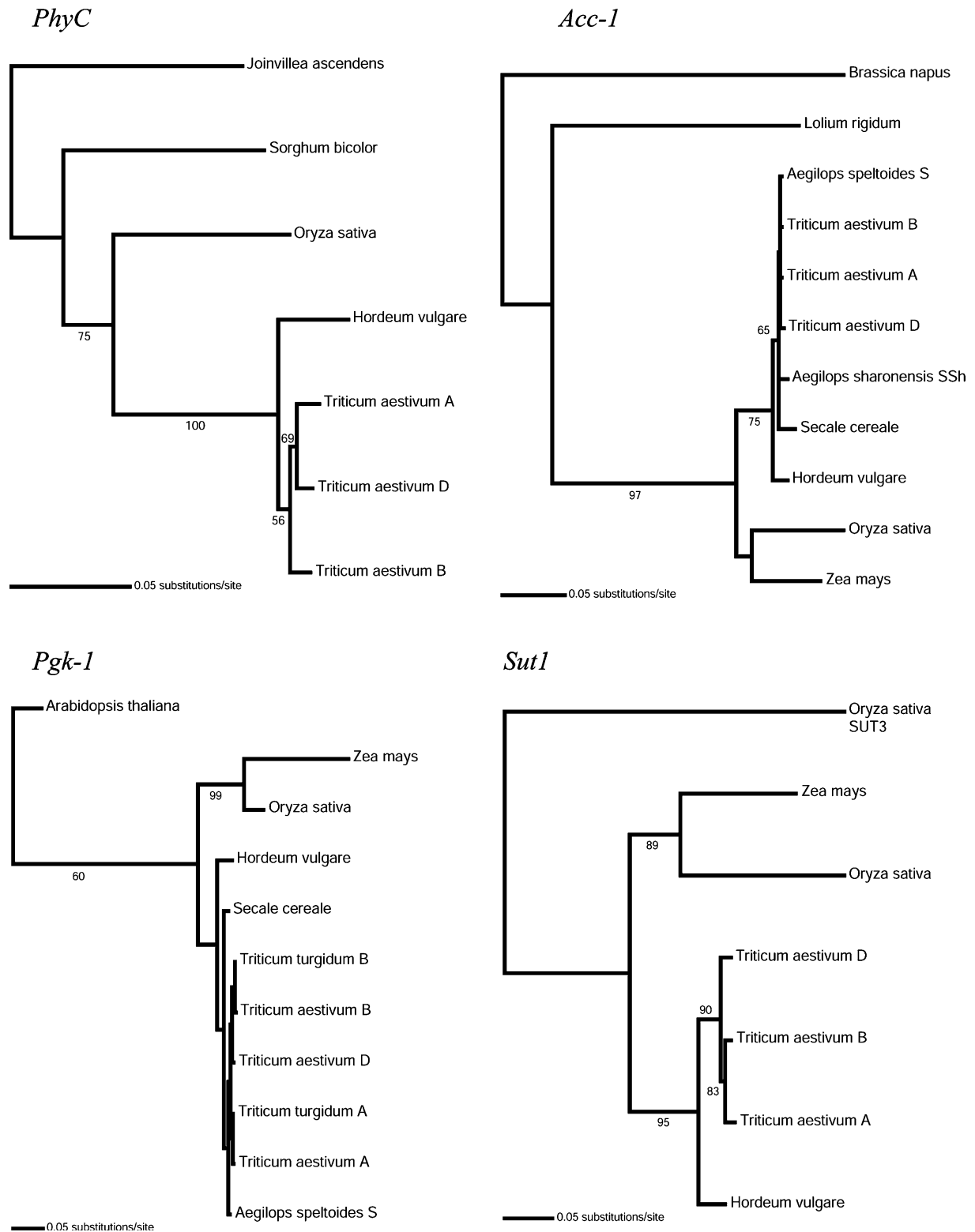


Figure 3. ML topologies for *PhyC*, *Acc-1*, *Pgk-1* and *Sut1* coding regions. Bootstrap support values are shown for those branches with over 70% support.

Table 3. Divergence times [and means (in bold) and 95% confidence intervals of bootstrapped age distributions] of various grass lineages for *PhyC*, *Acc-1* and *Sut1*, as calculated by the semi-parametric PL method in *r8s*, given a divergence age of 60 Mya for the divergence of the Pooid and Panicoid grasses.

Divergence	<i>PhyC</i>	<i>Acc-1</i>	<i>Sut1</i>
Wheat A–Wheat D	5.75 [3.49– 5.42 –8.49]	2.15 [0.54– 1.48 –4.75]	6.52 [4.13– 7.24 –12.63]
Wheat A–Wheat B	8.19 [5.11– 7.35 –10.83]	2.15 [0.54– 1.48 –4.75]	4.15 [2.44– 4.65 –8.71]
Wheat B–Wheat D	8.19 [5.11– 7.35 –10.83]	2.15 [0.54– 1.48 –4.75]	6.52 [4.13– 7.24 –12.63]
Wheat– <i>Hordeum</i>	12.34 [8.44– 11.65 –15.25]	13.86 [6.45– 14.37 –25.81]	18.93 [13.29– 20.03 –30.04]

average number of synonymous substitutions between the Triticeae genomes A (acc. AF408842), B (acc. AF408843), D (acc. AF408844), H (acc. AJ272309) and the Panicoid maize genome (acc. AB008464). A k -value of 4.8×10^{-9} provided an estimated average divergence time for the wheat progenitor genomes of 6.7 ± 1.4 Mya and for *Triticum–Hordeum* of 15.9 ± 2.2 Mya (Table 2).

In a second method, we used ML analyses of coding regions of the genes via the *r8s* program to estimate the ages of divergences in the grass phylogeny, again using an absolute age of 60 My for the divergence of the Panicoid grasses from the Pooid grasses (Table 3). The age of the *Hordeum–Triticum/Secale* divergence varies from 12.3 Mya for *PhyC* to 18.9 Mya for *Sut1*, and the age of the divergence of the three wheat genomes ranges from 2.2 Mya for *Acc-1* to 8.2 Mya for *PhyC*. Confidence intervals for the ages were calculated from 1000 bootstrapped data sets constrained to find the same phylogeny as the real data set. These overlapped in every comparison except between the A and B wheat genomes for *PhyC* and *Acc-1*. The shape of the bootstrapped distributions were skewed, resulting in the mean of the bootstrapped distribution for any particular node not being coincident with the actual divergence age for that node (Table 3).

The differences in divergence ages given by the various genes at particular nodes were tested for significance by comparing them against a null distribution of age differences. This was calculated by resampling pairs of random ages from the combined bootstrapped age distributions for that pair of genes, and taking the differences between the values for each pair. A difference in divergence ages was recognized as significant if the size of the difference lay outside the central 95% of the null distribution. Using this criterion we found that there were no pairs of divergence ages that were significantly different (Table 4).

Discussion

Evolution of the *PhyC* genes

The low K_a/K_s values for *PhyC-5A*, *5B* and *5D* indicate that the genes are under strong purifying selection. This is true for both the N-terminal photosensory domain and for the C-terminal signal transduction domain and suggests that missense mutations in either domain may greatly affect protein function. RT-PCR has demonstrated that all three genes are expressed. The average number of nucleotide substitutions

Table 4. Difference and significance of divergence dates given by different genes. All differences were non-significant at $P < 0.05$.

Divergence	Age difference (My)	95% confidence interval for age differences
<i>Acc-1–PhyC</i> , Wheat A–Wheat B	–6.04	–7.22–+7.25
<i>Acc-1–PhyC</i> , Wheat A–Wheat D	–3.60	–5.04–+5.09
<i>Acc-1–PhyC</i> , Wheat B–Wheat D	–6.04	–7.22–+7.25
<i>Sut1–PhyC</i> , Wheat A–Wheat B	–4.04	–5.74–+5.62
<i>Sut1–PhyC</i> , Wheat A–Wheat D	0.77	–5.69–+5.83
<i>Sut1–PhyC</i> , Wheat B–Wheat D	–1.67	–5.38–+5.28
<i>Acc-1–Sut1</i> , Wheat A–Wheat B	–2.0	–5.18–+5.10
<i>Acc-1–Sut1</i> , Wheat A–Wheat D	–4.37	–8.12–+8.05
<i>Acc-1–Sut1</i> , Wheat B–Wheat D	–4.37	–8.12–+8.05
<i>Acc-1–PhyC</i> , Wheat A– <i>Hordeum</i>	1.52	–11.03–+11.12
<i>Acc-1–Sut1</i> , Wheat A– <i>Hordeum</i>	–5.07	–14.07–+14.83
<i>Sut1–PhyC</i> , Wheat A– <i>Hordeum</i>	6.59	–15.0–+14.82

between the *PhyC-5A* coding regions of the five varieties analyzed is 4.0 ± 1.2 , with the largest number of differences being observed between Soleil and *T. spelta* (four synonymous and three non-synonymous substitutions) and the highest similarity being present between Reward and Chinese Spring (1 synonymous substitution, no amino acid differences). All differences are present in exon 1. Each of the five varieties had a unique haplotype, which is not surprising since the varieties were selected to be unrelated by pedigree.

Tetraploid and hexaploid wheat were formed approximately 0.5 Mya and 8000 years ago, respectively (Huang *et al.*, 2002b). In current-day tetraploid and hexaploid wheats, the action of the pairing control gene *Ph1* restricts pairing to homologs (Wall *et al.*, 1971). Our data show that at least some form of pairing control must have been present also in the early polyploids. Indeed, if intergenomic recombination had taken place in the early stages of the polyploidization process, the sequence differences between the wheat genomes would have reflected the onset of restricted pairing between the homoeologous genomes, which would have been at most 0.5 Mya. The divergence date for the A, B and D genomes of some 6.9 Mya (6.4 Mya using pairwise comparisons, 7.4 Mya using the semi-parametric PL method in *r8s*) clearly demonstrates that no recombination or gene conversion has taken place between the *PhyC* homoeologs since the formation of hexaploid bread wheat 8000 years ago (Huang *et al.*, 2002b) or its tetraploid progenitor 0.5 Mya (Huang *et al.*, 2002b).

The estimated time of divergence of the A, B and D ancestors based on the *PhyC* sequences is considerably earlier than the 2.5–4.5 Mya established based on the extent of intergenomic differences observed in the *Acc-1* and *Pgk-1* genes by Huang *et al.* (2002b). To ensure that the divergence times were calculated in the same way, the *Acc-1* and *Pgk-1* data used by Huang and colleagues were reanalyzed, but only coding regions were used as the introns could not be aligned unambiguously between the different grass orthologs. Following the phylogenetic analysis, *Pgk-1* was dropped from the study because of the lack of support for any of the clades produced in the phylogenetic analysis as well as its highly significant deviation from clocklike evolution. The

values obtained for *Acc-1* in our analyses (Tables 2 and 3) differ from those by Huang *et al.* (2002b) because we only used synonymous changes in coding regions. This was because the non-coding regions could not be aligned across the grass taxa used. Nevertheless, the *Acc-1* genes indicate substantially later divergence (approximately 2 Mya) of the diploid wheat progenitors than the *PhyC* sequence (around 6.9 Mya). *Sut1* genes, on the other hand, suggest divergence ages for the wheat ancestral genomes that are close to those obtained based on *PhyC* (around 6.2 Mya).

The results using the pairwise substitution method and *r8s* analyses show very similar divergence ages. However, the confidence intervals calculated by bootstrapping the data are wide and overlapping, except for the age of the divergence of the wheat A and B genomes estimated with *PhyC* and *Acc-1*. Testing the difference in ages against a null distribution of age differences derived from resampling the bootstrapped distributions revealed that there were no significant age differences. This implies that the differences in ages observed are due to sampling error, as well as reflecting the low information content available in the genes sampled. This is also evident from the low bootstrap support of many of the nodes that we attempted to date. Certainly there is more information in the coding regions of *PhyC* and *Sut1* than in *Acc-1*, and it may be that the dates derived from these two genes may be more robust than those derived from *Acc-1*.

One of the main problems in accurately dating clade divergences in the grasses is that sequence information is limited to just a few exemplar taxa. The lack of adequate sampling of other taxa (e.g. only one Panicoid and one Erhartoid grass were used) resulted in the topologies that were produced deviating from that in the most recent large scale molecular systematic studies (Barker *et al.*, 2001). In the Grass Phylogeny Working Group (GPWG) study, using seven molecular and one morphological data set, a clade of the Erhartoid and Panicoid grasses (as exemplified by *Oryza* and either *Zea* or *Sorghum*) is never found, yet we found it in all trees except that of *PhyC*. The relationship between *Hordeum*, *Secale*, and the wheat genomes conform to the GPWG tree. The relationships between the three wheat genomes are variable as has also been observed in previous work (Kellogg *et al.*, 1996; Mason-Gamer *et al.*, 1998). This may

reflect the lack of sampling, especially outside the wheat genomes.

Another possibility which may explain some of the discrepancy in results between the ages estimated from the different genes and the variable topologies for the wheat genomes is the previous reports in the Triticeae of different portions of the genome having different evolutionary histories. Phylogenetic trees based on different genes, although generally consistent in overall topography, display significantly different relationships for Triticeae species (Kellogg *et al.*, 1996; Mason-Gamer *et al.*, 1998). The *Acc-1* genes are located on the wheat group 2 chromosomes, *Sut1* on the group 1 chromosomes and *PhyC* on the group 5 chromosomes. The discrepancy between our data and those obtained by Huang *et al.* (2002b) may be yet another example of the mosaic nature of some Triticeae genomes.

Comparative analysis of the homoeologous 5A, 5B and 5D regions upstream of PhyC

Sequence conservation between the A, B and D genomes upstream of the *PhyC* initiation codon is limited to a region of some 1100 bp (Figure 1B). Colinearity over this region is not continuous, but is interrupted by insertion and duplication events in the 5A and 5D homoeologs. The repeats are not conserved between the genomes, suggesting that they inserted after the divergence of the diploid ancestors of hexaploid wheat. A possible exception may be the regions identified in the 5A and 5D sequences that showed similarity to the same putative non-LTR retrotransposon reverse transcriptase identified in *Arabidopsis*. If these sequences indeed have the same ancestral origin, they have degenerated to the point where similarity between the A and D sequences is no longer detectable at the nucleotide level. The four LTR retroelements identified in the A genome have also been severely truncated. These elements could not be dated using the methodology described by SanMiguel *et al.* (1998), but comparative data shows that at least one element, Barbara, inserted after the divergence of the diploid wheat ancestral genomes, implying that the element is relatively young (< 6 My). The relatively fast degradation of this retroelement conforms to previous observations that the repeat fraction of a genome is

very fluid. Studies in *Arabidopsis* and rice have demonstrated that LTR-retroelements have a half-life of less than 5 million years. Mechanisms responsible for their removal are mainly unequal homologous recombination and illegitimate recombination (Devos *et al.*, 2002; Ma *et al.*, 2004). Although this is the first comparative analysis at the sequence level between the A, B and D genomes of bread wheat, the differential insertion of repeats in the three genomes is expected to be a common feature. Analyses of the regions around the low molecular weight glutenin loci in the closely related A^m genome of *T. monococcum* and the A genome of tetraploid wheat (Wicker *et al.*, 2003) and of the high molecular weight glutenin region in the D genome of *T. tauschii* and the B genome of tetraploid wheat (Kong *et al.*, 2004) have also uncovered genome-specific insertion of repeats. The presence of genome-specific repeats is also consistent with the fact that levels of intergenomic restriction fragment length polymorphism (RFLP) are close to 100% and detectable with multiple enzyme combination (K.M. Devos, unpublished observations). Most likely, the differential presence of repeats in the three wheat genomes lies at the basis of the RFLP variation.

We have no data to establish whether the retroelements are differentially present in the A genome of our sample of five varieties. We do know, however, that the inverted repeat that is inserted in CS 5D is largely absent in Soleil. The presence of part of the direct repeat termini of the foldback element in Soleil indicates that the element may have been present in the D genome progenitor that gave rise to hexaploid wheat but was subsequently deleted in the line that led to Soleil. Intervarietal differences in the organization of repeats have been observed in maize (Fu and Dooner, 2002). In contrast to wheat, maize has high levels of intervarietal RFLP variation. Intervarietal differences in the organization of repeats, although clearly present in wheat genomes, are expected to occur at much lower frequency than in maize.

The upstream regions were also searched for the presence of conserved non-coding sequences (CNSs) which, potentially, may represent regulatory elements. In closely related genomes such as wheat, there is some 30% chance that a stretch of 20 bp has remained conserved in sequence without

selective pressure even after 7 My of divergence (calculated according to Kaplinsky *et al.*, 2002). Therefore, we looked for CNSs that were conserved in both wheat and rice in the region upstream of *PhyC*. Using the criterion of >70% identity in a 20 bp window established by Guo and Moose (2003) for finding significant conserved non-coding sequences (CNSs) in cereal genomes, six CNSs were identified that were conserved across the wheat and rice upstream regions (see Supplementary Material). All motifs identified were located 3' of the presumed TATA box in rice and are thus likely present in the 5'-UTR. The CNS located most closely to the translation start site corresponded to the uORF. A search of plantCARE (<http://intra.psb.ugent.be:8080/PlantCARE/>), a database for *cis*-acting plant regulatory elements, revealed that three out of the six CNSs contained putative regulatory elements. Two CNSs contained a GAG motif and the most conserved CNS, which was 82% identical over 33 bp, contained both an I-box and an ATCT motif (see Supplementary Material). All three motifs have been implicated in light-response. This suggests that *PhyC* mRNA abundance might be light-regulated by a post-transcriptional process. The presence of an uORF in the 5'-UTR may be a further indication that the *PhyC* genes in wheat are under translational control. Eukaryotic ribosomes can reinitiate translation at downstream ATG codons, although this ability is affected by the length of the uORF with a cut-off length of around 30 amino acids and diminishes with decreasing intercistronic distances (Kozak, 1987; Luukkonen *et al.*, 1995). In rice, the peptide encoded by the uORF comprises only four amino acids. In wheat, however, the peptides are 28–31 amino acids long and the uORF stop codons are in close proximity (55–61 bp) to the ATG initiation codon. It has been shown that the presence of a 38 codon long uORF that ended 62 basepairs upstream of the *Lc* initiation codon in maize reduced *Lc* expression 25- to 30-fold (Damiani, Jr. and Wessler, 1993). Furthermore, the repressing effect depended greatly on the nucleotide sequence of the uORF. It is thus conceivable that a non-synonymous nucleotide substitution (arginine/proline replacement) present in the penultimate codon of the uORF in the variety Reward represents an intervarietal difference in the regulation of *PhyC-5A*.

The duplication upstream of PhyC-5A

To evaluate whether the 800 bp duplication observed some 500 bp upstream of the *PhyC-5A* initiation codon (Figure 1B) arose before or after the divergence of the diploid ancestral wheat genomes, an attempt was made to date the duplication. Based on the number of nucleotide substitutions between the duplicated regions and a molecular clock of 4.4×10^{-9} substitutions per synonymous site per year calculated for the *PhyC* gene, the duplication was estimated to have taken place 7.3 ± 1.1 Mya. This value falls within the 95% confidence interval for the divergence age of the diploid ancestral wheat genomes calculated on the basis of the number of synonymous substitutions differentiating *PhyC-5A*, *PhyC-5B* and *PhyC-5D*. No conclusions can therefore be drawn as to whether the duplication is ancestral to or occurred shortly after the divergence of the A, B and D genomes. It should be noted that the distribution of nucleotide substitutions between the two copies is not uniform across the region. A higher number of substitutions was observed in the most 5' 200 bp and particularly between basepairs 100 and 200. This could indicate that part of the duplication has undergone conversion. The overall outcome of a conversion event would be an underestimation of the age of the duplication. On the other hand, one or both copies of the duplication may be under reduced selective constraint and therefore evolve faster in which case the 7.3 My would be an overestimation.

The sequence data provide some indication that the duplication is specific to the A genome. Only one copy of this region is present in the B and D genome sequence. Of course, it is possible that the upstream copy is not covered by our sequence data. However, if the putative non-LTR retrotransposon reverse transcriptases observed in the A and D genomes have the same ancestral origin, then this sequence should delineate the duplicated region at the 5' end. Consequently, the presence of only one copy of the duplication 3' of the reverse transcriptase sequence in the D genome sequence suggests that the duplication occurred after the divergence of the ancestral A and D genomes. Most likely, the duplication arose in tandem and the two copies were subsequently separated by the

insertion of transposons, including the LTR retroelement Barbara. In *T. spelta* acc. 'Grey', only the 5' copy of the duplication is present. Presumably, deletion of the 3' copy and any intervening DNA occurred through unequal homologous recombination.

The presence of a MITE in the 3' copy of the duplication in Soleil but not in the 5' copy initially suggested that the MITE insertion had occurred after the duplication. However, PCR amplification from the variety Chinese Spring using primer pairs specific to each of the duplicated copies and sequencing of the amplification products demonstrated the presence of the same MITE in identical positions in both paralogs. Surprisingly, the 5' and 3' paralogous MITEs were 98% identical. This was observed not only in Chinese Spring, but also in Reward, a second variety in which we sequenced across the paralogous MITEs. The high sequence similarity of the MITEs may be explained by homogenization of the two copies through a conversion event. This is a possible scenario as the nucleotide substitution rate is not uniform across the duplication, but higher 5' of the MITE insertion site. In the variety Soleil, only the 3' copy of the MITE is present. If the MITE indeed inserted before the duplication event, the rationale is that the MITE located in the 5' copy of the duplication must have excised. MITE excision has been demonstrated in rice, but was shown to be mostly imprecise and leave different types of excision footprints (Nakazaki *et al.*, 2003; Kikuchi *et al.*, 2003). No excision footprint was observed in Soleil.

An equally if not more likely scenario is that the MITEs are of recent origin and have inserted independently in the same location in both copies of the duplication. MITEs are preferentially associated with genes and have a target site preference (Wessler *et al.*, 1995). Independent insertions at the same locus have previously been observed for other transposable elements, including *Mu* and *PIF* elements (Hardeman and Chandler, 1993; Walker *et al.*, 1997; Zhang *et al.*, 2001). *PIF*, a DNA element family related to *Tourist* MITEs has been shown to be inserted repeatedly at the same location in the *R* locus in maize (Walker *et al.*, 1997; Zhang *et al.*, 2001).

Our data may represent the first example of an active MITE family in wheat. A search of Genbank revealed that this MITE was found in *T. aestivum*, *T. tauschii* and *H. vulgare*. This suggests that this MITE family dates back to at

least before the divergence of the *Triticum* and *Hordeum* lineages some 10–19 Mya. To our knowledge, this is the first report of the presence of the same MITE in more distantly related species.

Conclusions

Comparisons of *PhyC* genes and their upstream region in the A, B and D homoeologs of wheat have shown that, while the genes are conserved, the non-genic regions have diverged by the insertion of retroelements and other repeats. K_a/K_s values were low, suggesting that all three *PhyC* homoeologs were functional. RT-PCR confirmed that *PhyC-5A*, *PhyC-5B* and *PhyC-5D* were expressed. Our study highlights the need for a larger sample of genes for dating the evolutionary divergence of the wheat genomes. The intergenomic K_s values of the *PhyC* and *Sut1* genes were considerably higher than those calculated for the *Acc-1* genes. This suggests that the previously published value for the divergence of the ancestral wheat genomes of 2.5–4.5 Mya may be an underestimation and that the true value may lie closer to 7 Mya. These results were confirmed using *r8s*, although resampling tests and examination of bootstrapped derived confidence intervals revealed that there was relatively little power in the sequence information available to distinguish between the divergence dates given by different genes. More genes from more species need to be analyzed to evaluate whether these differences in evolutionary rate are due to sampling error or are related to chromosomal positions. The latter possibility echoes previous suggestions that some parts of the genome may have originated through introgression.

Acknowledgements

This manuscript is largely an output from the project PMC68 funded by Syngenta. We thank Elizabeth Kellogg, University of Missouri – St. Louis, for critical reading of the manuscript.

References

- Aoki, N., Whitfield, P., Hoeren, F., Scofield, G., Newell, K., Patrick, J., Offler, C., Clarke, B., Rahman, S. and Furbank,

- R.T. 2002. Three sucrose transporter genes are expressed in the developing grain of hexaploid wheat *Plant Mol. Biol.* 50: 453–462.
- Barker, N.P., Clark, L.G., Davis, J.I., Duvall, M.R., Guala, G.F., Hsiao, C., Kellogg, E.A., Linder, H.P., Mason-Gamer, R.J., Mathews, S.Y., Simmons, M.P., Soreng, R.J. and Spangler, R.E. 2001. Phylogeny and subfamilial classification of the grasses (Poaceae) *Ann. Missouri Bot. Garden* 88: 373–457.
- Basu, D., Dehesh, K., Schneider-Poetsch, H.-J., Harrington, S., McCouch, S.R. and Quail, P.H. 2000. Rice *PHYC* gene: structure, expression, map position and evolution *Plant Mol. Biol.* 44: 27–42.
- Brunner, S., Keller, B. and Feuillet, C. 2003. A large rearrangement involving genes and low-copy DNA interrupts the microcollinearity between rice and barley at the *Rph7* locus *Genetics* 164: 673–683.
- Chen, M., SanMiguel, P., de Oliveira, A.C., Woo, S.S., Zhang, H., Wing, R.A. and Bennetzen, J.L. 1997. Microcollinearity in *sh2*-homologous regions of the maize, rice and sorghum genomes *Proc. Natl. Acad. Sci.* 94: 3431–3435.
- Clack, T.S., Mathews, S. and Sharrock, R.A. 1994. The phytochrome apoprotein family in *Arabidopsis* is encoded by five genes: the sequences and expression of *PHYD* and *PHYE* *Plant Mol. Biol.* 25: 413–427.
- Cowl, J.S., Hartley, N., Xie, D.X., Whitelam, G.C., Murphy, G.P. and Harberd, N.P. 1994. The *PHYC* gene of *Arabidopsis* *Plant Physiol.* 106: 813–814.
- Damiani, R.D. Jr. and Wessler, S.R. 1993. An upstream open reading frame represses expression of *Lc*, a member of the *R/B* family of maize transcriptional activators *Proc. Natl. Acad. Sci.* 90: 8244–8248.
- Deng, X.W. and Quail, P.H. 1999. Signalling in light-controlled development *Sem. Cell Dev. Biol.* 10: 121–129.
- Devos, K.M., Atkinson, M.D., Chinoy, C.N., Liu, C. and Gale, M.D. 1992. RFLP based genetic map of the homologous group 3 chromosomes of wheat and rye *Theor. Appl. Genet.* 83: 931–939.
- Devos, K.M., Brown, J.K.M. and Bennetzen, J.L. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis* *Genome Res.* 12: 1075–1079.
- Dubchak, I., Brudno, M., Loots, G.G., Mayor, C., Pachter, L., Rubin, E.M. and Frazer, K.A. 2000. Active conservation of noncoding sequences revealed by 3-way species comparisons *Genome Res.* 10: 1304–1306.
- Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.
- Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, Massachusetts, pp. 156–158, 309–310.
- Fu, H. and Dooner, H.K. 2002. Intraspecific violation of genetic colinearity and its implications in maize *Proc. Natl. Acad. Sci.* 99: 9573–9578.
- Furuya, M. 1993. Phytochromes – Their molecular-species, gene families, and functions *Annu Rev Plant Physiol. Plant Mol. Biol.* 44: 617–645.
- Gallego, F., Feuillet, C., Messmer, M., Penger, A., Graner, A., Yano, M., Sasaki, T. and Keller, B. 1998. Comparative mapping of the two wheat leaf rust resistance loci *Lr1* and *Lr10* in rice and barley *Genome* 41: 328–336.
- Guo, H. and Moose, S.P. 2003. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution *Plant Cell* 15: 1143–1158.
- Hardeman, K.J. and Chandler, V.L. 1993. Two maize genes are each targeted predominantly by distinct classes of *Mu* elements *Genetics* 135: 1141–1150.
- Huang, S., Sirikhachornkit, A., Faris, J.D., Su, X., Gill, B.S., Haselkorn, R. and Gornicki, P. 2002a. Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses *Plant Mol. Biol.* 48: 805–820.
- Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B.S., Haselkorn, R. and Gornicki, P. 2002b. Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat *Proc. Natl. Acad. Sci.* 99: 8133–8138.
- Ilic, K., SanMiguel, P.J. and Bennetzen, J.L. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes *Proc. Natl. Acad. Sci.* 100: 12265–12270.
- Jackson, S. and Thomas, B. 1997. Photoreceptors and signals in the photoperiodic control of development *Plant Cell Environ.* 20: 790–795.
- Kaplinsky, N.J., Braun, D.M., Penterman, J., Goff, S.A. and Freeling, M. 2002. Utility and distribution of conserved coding sequences in the grasses *Proc. Natl. Acad. Sci.* 99: 6147–6151.
- Kellogg, E.A., Appels, R. and Mason-Gamer, R.J. 1996. When genes tell different stories: the diploid genera of Triticeae (Gramineae) *Syst. Bot.* 21: 1–17.
- Kikuchi, K., Terauchi, K., Wada, M. and Hirano, H.Y. 2003. The plant MITE mPing is mobilized in anther culture *Nature* 421: 167–170.
- Kong, X.-Y., Gu, Y.Q., You, F.M., Dubcovsky, J. and Anderson, O.D. 2004. Dynamics of the evolution of orthologous and paralogous portions of a complex locus region in two genomes of allopolyploid wheat *Plant Mol. Biol.* 54: 55–69.
- Kozak, M. 1987. Effects of intercistronic length on the efficiency of reinitiation by eukaryotic ribosomes *Mol. Cell Biol.* 7: 3438–3445.
- Kumar, S., Tamura, K., Jakobsen, I.B. and Nei, M. 2001. MEGA2: molecular evolutionary genetics analysis software *Bioinformatics* 17: 1244–1245.
- Liu, B., Vega, J.M. and Feldman, M. 1998a. Rapid genomic changes in newly synthesized amphiploids of *Triticum* and *Aegilops*.II. Changes in low-copy coding DNA sequences *Genome* 41: 535–542.
- Liu, B., Vega, J.M., Segal, G., Abbo, S., Rodova, M. and Feldman, M. 1998b. Rapid genomic changes in newly synthesized amphiploids in *Triticum* and *Aegilops*.I. Changes in low-copy noncoding DNA sequences *Genome* 41: 272–277.
- Liu, Y.G., Nagaki, K., Fujita, M., Kawaura, K., Uozumi, M. and Ogihara, Y. 2000. Development of an efficient maintenance and screening system for large-insert genomic DNA libraries of hexaploid wheat in a transformation-competent artificial chromosome (TAC) vector *Plant J* 23: 687–695.
- Luukkonen, B.G., Tan, W. and Schwartz, S. 1995. Efficiency of translation on human immunodeficiency virus type 1 mRNAs is determined by the length of the upstream open reading frame and by intercistronic distance *J. Virol.* 69: 4086–4094.
- Ma, J., Devos, K.M. and Bennetzen, J.L. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice *Genome Res.* 14: 860–869.

- Martínez-García, J.F., Huq, E. and Quail, P.H. 2000. Direct targeting of light signals to a promoter element-bound transcription factor *Science* 288: 859–863.
- Mason-Gamer, R.J., Weil, C.F. and Kellogg, E.A. 1998. Granule-bound starch synthase: structure, function and phylogenetic utility *Mol. Biol. Evol.* 15: 1658–1673.
- Mathews, S. and Sharrock, R.A. 1996. The phytochrome gene family in grasses (*Poaceae*): a phylogeny and evidence that grasses have a subset of the loci found in dicot angiosperms *Mol. Biol. Evol.* 13: 1141–1150.
- Monte, E., Alonse, J.M., Ecker, J.R., Zhang, Y., Li, X., Yound, J., Austin-Phillips, S. and Quail, P.H. 2003. Isolation and characterization of *phyC* mutants in arabidopsis reveals complex crosstalk between phytochrome signaling pathways *Plant Cell* 15: 1962–1980.
- Nakazaki, T., Okumoto, Y., Horibata, A., Yamahira, S., Teraishi, M., Nishida, H., Inoue, H. and Tanisaka, T. 2003. Mobilization of a transposon in the rice genome *Nature* 421: 170–172.
- Nylander, J.A. 2002. Mrmodeltest version 1.1b. Program distributed by the author. Department of Systematic Zoology, EBC, Uppsala University, Sweden. (<http://www.ebc.uu.se/systzoo/staff/nylander.html>).
- Quail, P.H. 2000. Phytochrome interacting factors *Sem. Cell Dev. Biol.* 11: 457–466.
- Quarrie, S.A., Gulli, M., Calestani, C., Steed, A. and Marmioli, N. 1994. Location of a gene regulating drought-induced abscisic acid production on the long arm of chromosome 5A of wheat *Theor. Appl. Genet.* 89: 794–800.
- Sambrook, J., Russell, D.W. and Sambrook, J. 2001. *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor.
- Sanderson, M.J. 2002. r8s, version 1.50, User's manual. (<http://ginger.ucdavis.edu/r8s>).
- Sanderson, M.J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock *Bioinformatics* 19: 301–302.
- SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize : dating the strata *Nature Genet.* 20: 43–45.
- Sears, E.R. 1954. The aneuploids of common wheat *Mo. Agric. Exp. Stn. Res. Bull.* 572: 1–59.
- Sears, E.R. and Sears, L.M.S. 1979. *Proc 5th Int. Wheat Genet. Symp.*, The Indian Society of Genetics & Plant Breeding, Indian Agricultural Research Institute, New Delhi.
- Simon, J.L. 1995. Resampling, the new statistics. Resampling Stats, Inc., Arlington, Virginia, pp. 1–208.
- Smith, H. 1995. Physiological and ecological function within the phytochrome family *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 46: 289–315.
- Smith, H. 2000. Phytochromes and light signal perception by plants – an emerging synthesis *Nature* 407: 585–591.
- Swofford, D.L. 2001. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Massachusetts.
- Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y., Gorenstein, N.M., Bennetzen, J.L. and Avramova, Z. 1999. Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum *Proc. Natl. Acad. Sci.* 96: 7409–7414.
- Walker, E.L., Eggleston, W.B., Demopoulos, D., Kermicle, J. and Dellaporta, S.L. 1997. Insertions of a novel class of transposable elements with a strong target site preference at the *r* locus of maize *Genetics* 146: 681–693.
- Wall, A.M., Riley, R. and Gale, M.D. 1971. The position of a locus of chromosome 5B in *Triticum aestivum* affecting homoeologous meiotic pairing *Genet Res* 18: 329–339.
- Wessler, S.R., Bureau, T.E. and White, S.E. 1995. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes *Curr. Opin. Genet. Dev.* 5: 814–821.
- White, G.M., Hamblin, M.T. and Kresovich, S. 2004. Molecular evolution of the phytochrome gene family in sorghum: changing rates of synonymous and replacement evolution *Mol. Biol. Evol.* 21: 716–723.
- Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E., Liu, Z.-D., Dubcovsky, J. and Keller, B. 2003. Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat *Plant Cell* 15: 1186–1197.
- Wolfe, K.H., Gouy, M., Yang, Y.-W., Sharp, P.M. and Li, W.-H. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data *Proc. Natl. Acad. Sci.* 86: 6201–6205.
- Yan, L., Loukoianov, A., Tranquilli, G., Helguera, M., Fahima, T. and Dubcovsky, J. 2003. Positional cloning of the wheat vernalization gene *VRN1* *Proc. Natl. Acad. Sci.* 100: 6263–6268.
- Zhang, X., Feschotte, C., Zhang, Q., Jiang, N., Eggleston, W.B. and Wessler, S.R. 2001. *P* instability factor: an active maize transposon system associated with the amplification of *Tourist*-like MITEs and a new superfamily of transposases *Proc Nat Acad Sci* 98: 12572–12577.