

Phillip J. SanMiguel · Wusirika Ramakrishna  
Jeffrey L. Bennetzen · Carlos S. Busso  
Jorge Dubcovsky

## Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A<sup>m</sup>

Received: 8 January 2002 / Accepted: 7 March 2002 / Published online: 12 April 2002  
© Springer-Verlag 2002

**Abstract** Sequencing of a contiguous 215-kb interval of *Triticum monococcum* showed the presence of five genes in the same order as in previously sequenced colinear barley and rice BACs. Gene 2 was in the same orientation in wheat and rice but inverted in barley. Gene density in this region was 1 gene per 43 kb and the ratio of physical to genetic distance was estimated to be 2,700 kb cM<sup>-1</sup>. Twenty more-or-less intact retrotransposons were found in the intergenic regions, covering at least 70% of the sequenced region. The insertion times of 11 retrotransposons were less than 5 million years ago and were consistent with their nested structure. Five new families of retroelements and the first full-length elements for two additional retrotransposon families were discovered in this region. Significantly higher values of GC content were observed for Triticeae BACs compared with rice BACs. Relative enrichment or depletion of certain dinucleotides was observed in the comparison of introns, exons and retrotransposons. A higher proportion of transitions in CG and CNG sites that are targets for cytosine methylation was observed in retrotransposons (76%) than in introns (37%). These results showed that the wheat genome is a complex mixture of different sequence elements, but with general patterns of content and interspersions that are similar to those seen in maize and barley.

**Keywords** Wheat · Comparative genomics · Retrotransposons · Transposable elements

P.J. SanMiguel  
Purdue Genomics Core Facility, 1057 WSLR, West Lafayette,  
IN 47907-1057, USA

W. Ramakrishna · J.L. Bennetzen  
Department of Biological Sciences, Purdue University,  
West Lafayette, IN 47907-1392, USA

C.S. Busso · J. Dubcovsky (✉)  
Department of Agronomy and Range Science,  
University of California, Davis, CA 95616, USA  
e-mail: jdubcovsky@ucdavis.edu  
Tel.: +1-530-7525159, Fax: +1-530-7524361

### Introduction

The diploid wheat *Triticum monococcum* L. (2n = 14) has a genome designated A<sup>m</sup> that is closely related to the genome of *T. urartu*, the A genome donor for tetraploid and hexaploid wheats. The large genome of *T. monococcum* (1C = 5600 Mb; Bennett and Leitch 1995) is approximately 12 times larger than the genome of rice and 40 times larger than the genome of the model experimental plant, *Arabidopsis thaliana* (Bennett and Leitch 1995). One chromosome of *T. monococcum* has more DNA than the complete rice genome.

Studies based on a limited sample of completely sequenced bacterial artificial chromosome (BAC) clones from barley (Dubcovsky et al. 2001; Panstruga et al. 1998; Shirasu et al. 2000) and wheat (Wicker et al. 2001) have shown that the increase in genome size of these Triticeae species is associated with the insertion of a large number of retroelements present in a nested organization. Hence, the organization of the large wheat and barley genomes appears to be highly similar to that described for maize (SanMiguel and Bennetzen 1998; SanMiguel et al. 1996; Tikhonov et al. 1999), suggesting that this is a general feature of the large grass genomes.

All previously sequenced BACs were selected by screening BAC libraries with genes, and therefore should be considered representative of gene regions in the genome. In these examples, single genes or small islands that include a few genes were found to be separated from each other by blocks of nested repetitive elements. Despite the presence of large blocks of retroelements, gene densities in the sequenced Triticeae BACs were higher than expected from a random distribution of genes along the genome. If the total number of genes in diploid wheat is similar to the estimated 25,498 genes in *A. thaliana* (The Arabidopsis Genome Initiative 2000), and the genes are randomly distributed within the genome, then the expected average gene density would be approximately 1 gene per 220 kb. However, gene densities observed in the four Triticeae BACs sequenced so far ranged from 1 gene per 20–22 kb in barley (Dubcovsky

et al. 2001; Panstruga et al. 1998; Shirasu et al. 2000) to 1 gene per 42 kb in wheat (Wicker et al. 2001). The observed higher-than-average gene densities are similar to those calculated based on cDNA mapping of wheat deletion lines (Faris et al. 2000; Gill et al. 1996a). These results suggest that most Triticeae genes are associated with gene-rich regions that are more frequent towards the end of chromosomes.

Three clusters of markers were evident in the physical maps of the long arms of homologous chromosome group 5, surrounding fraction lengths (FL) 0.30, 0.55, and 0.75. Sixty-one percent of the markers on the long arm mapped to these three regions that encompass less than 18% of the arm (Gill et al. 1996a). The region on FL 0.75 has received particular attention because of its high gene density and because of the presence of vernalization and frost tolerance genes *Vrn-1* and *Fr-1* in wheat and barley (Dubcovsky et al. 1998; Kato et al. 1999; Sarma et al. 1998; Sutka et al. 1999). A 102-kb barley BAC selected with RFLP marker WG644 from this region was recently sequenced (Dubcovsky et al. 2001) and five genes were identified.

We present here the analysis of a 215-kb contiguous sequence of a chromosome 5 region in *T. monococcum* that contains the same five genes previously detected in the barley and rice BACs (Dubcovsky et al. 2001). These studies provide detailed insights into the structure and organization of the wheat genome and begin to explain the chromosomal changes that have occurred during the evolution of the Triticeae tribe.

## Materials and methods

### BAC selection

Probes derived from the genes detected in barley BAC 635P2 (Dubcovsky et al. 2001) were used to screen a *T. monococcum* BAC library. This library includes 276,480 BAC clones, which provides 5.6 genome equivalents (Lijavetzky et al. 1999). Positive BACs found in this library were fingerprinted with the restriction enzyme *Hind* III to determine the overlap among clones. DNA was transferred to nylon membranes and hybridized with the different genes. Since no single BAC clone was found to include the five genes, two BACs with a minimum overlap were selected for sequencing.

### Sequencing

DNA from *T. monococcum* BACs 115G01 and 116F02 was extracted using the QIAGEN Large Construct Kit and sheared with a HydroShear (Genemachines) to two different average sizes of 2 and 9 kb. The sheared fragments were converted to blunt-ended fragments with mung bean nuclease (Amersham Pharmacia Biotech), dephosphorylated with shrimp alkaline phosphatase (Roche), and "A" tails added by incubation with *Taq* polymerase in the presence of dNTPs. These fragments were inserted into the vector pCR4TOPO using the Invitrogen TA cloning kit following the instructions of the manufacturer. The resulting DNA was electroporated into DH10B electroMAX cells (Lorow and Jessee 1990; Invitrogen). Clones were picked using a Qpix colony picker (Genetix) into 384-well culture trays (Genetix) filled with 60  $\mu$ l terrific broth culture medium plus 8% glycerol. After overnight growth (12–18 h), cultures were frozen at  $-80^{\circ}\text{C}$  until needed.

REAL kits (Qiagen) were used to prepare DNA minipreps from 1.3-ml cultures grown in deep 96-well plates for 14–18 h at  $37^{\circ}\text{C}$  shaking at 300 rpm. DNA was typically resuspended in 50  $\mu$ l water, with 4  $\mu$ l used for each sequencing reaction. Clones were sequenced from both directions using Big Dye Terminator chemistry (Applied Biosystems) and run on an ABI3700 capillary sequencer after terminator clean-up using Squeaky-Clean (BioRad) 96-well column plates. Base calling and quality assessment were done using PHRED (Ewing and Green 1998), assembled by PHRAP and edited with CONSED (Gordon et al. 1998). Gaps were filled by a combination of primer walking and shotgun sequencing of subclones with extremes at both sides of the sequencing gaps. Final error rate was estimated using CONSED.

### Sequence analysis

The finished sequence from the *T. monococcum* BACs was compared with the previous annotation of barley BAC 635P02 (GenBank accession number AY013246). Sequences from the *T. monococcum* BACs were compared with NCBI dbEST and non-redundant databases using BLASTN, BLASTX, and TBLASTX algorithms (Altschul et al. 1997) to search for additional genes. Transposable elements were identified using a combination of FASTA, BLAST and NETBLAST searches to the GenBank/EMBL nr database and TIGR (<http://www.tigr.org/tdb/rice/blast-search.shtml>), homology searches to known transposable elements using *cross\_match* (<http://www.phrap.org/phrap.docs/general.html>), same sequence comparisons, and orthologous sequence comparisons. The sequences of both BACs were compared to themselves (same sequence comparison) and each other (orthologous sequence comparison) using *cross\_match*. Sequence nucleotide composition was determined using COMPOSITION (GCG) and *find\_fasta* (<http://www.genome.ou.edu/informatics.html>).

A cluster analysis was performed using the long terminal repeats (LTRs) of the *Ty1-copia* retrotransposons found in this wheat contig and representative elements from the sub-groups BARE-1 (Z17327, AY013246), WIS-2 (AF339051), ANGELA (AF326781) and CLAUDIA (AF326781). Phylogenetic trees were generated from the ClustalW sequence alignments using multiple-distance based methods available in the MEGA2.1 computer software package (<http://www.megasoftware.net/>; Kumar et al. 2001). Sites containing alignment gaps were removed before the calculations using the pair-wise deletion option. Distances between each pair of LTRs were calculated and trees were constructed using the Unweighted Pair Group Method using Averages (UPGMA) and Neighbor-Joining (NJ), with the default parameters implemented in MEGA2.1. Confidence values for the nodes were calculated for the two analyses using 500 bootstraps (MEGA2.1).

The average GC contents of different BACs from wheat, barley and rice were compared using a one-way analysis of variance and orthogonal contrasts. Dinucleotide frequencies were calculated as: the (observed number – expected number)/expected number, where the expected number was calculated from the base pair composition of each sequence (proportion of bp 1 \* proportion of bp 2 \* number of bp). Comparisons of dinucleotide frequencies in exons, introns and retrotransposons were done by a one-way analysis of variance using the different genes and retrotransposons as replicates. All statistical analyses were performed using SAS 8.0 (SAS Institute 2001).

### Retrotransposon insertion times

The 5' and 3' LTRs of the complete retrotransposons were compared to date the insertion time of these elements (SanMiguel et al. 1998). The two LTRs of a single retrotransposon are identical at the time of insertion because they are mostly copied from the same template. The two LTRs were aligned, indels were eliminated, and the number of transition and transversion mutations were counted. The insertion times of retrotransposons with both LTRs were dated in a manner similar to SanMiguel et al. (1998), using

the Kimura two parameter method (K2P; Kimura 1980) and a mutation rate of  $6.5 \times 10^{-9}$  substitutions per synonymous site per year based on the *adh1* and *adh2* loci of grasses (Gaut et al. 1996).

### Restriction map

Restriction maps of 115G01 and 116F02 were constructed to experimentally validate computer sequence assembly. BACs were individually digested with 8-bp cutter restriction enzymes *AscI*, *FseI*, *NotI*, *PacI*, *PmeI*, and *SwaI* and with the different possible double digestions. The resultant fragments were analyzed by pulse field electrophoresis in 1% agarose gels (12°C, 14 h, 200 V, pulse 0.2–13 s; CHEF-DR, BIO-RAD). Gel filter replicas were hybridized with [ $\alpha$ - $^{32}$ P]-labeled clones from BAC vector pBELOBAC11 (CVU51113), and PCR-amplified segments from gene 1, gene 2 and gene 4 (RFLP marker WG644).

## Results

### Sequencing

*Triticum monococcum* BAC clones 115G01 and 116F02 overlap for a region of 20,573 bp and cover a 215,220-bp segment of the long arm of chromosome 5A<sup>m</sup>. The sequence from the two BACs was deposited as a single contiguous sequence (contig) in GenBank (accession number AF459639). Sequences from the individual transposable elements, without additional nested elements, were deposited in the GrainGenes Triticeae Repeat Sequence Database (<http://wheat.pw.usda.gov/ggpages/Repeats/>). The shotgun sequencing of 3,072 subclones in both directions provided 3,637,632 high quality bases, giving an initial 17 $\times$  coverage of the BAC contig. Finishing was accomplished with 542 sequencing reactions, including 201 reactions primed with custom oligonucleotides using, variously, dRhodamine (Applied Biosystems), Big Dye dGTP (Applied Biosystems) and ET (Amersham Pharmacia Biotech) chemistries. The construction of an additional, large-insert, shotgun library for 116F02 from which 96 clones were end-sequenced to anchor them in the existing sequence and transposon bombing (EZ::TN transposon insertion systems, Epicentre Technologies) of 15 clones were used to generate an additional 732 reads. The final error rate of the sequence estimated by CONSED was 0.05 bp per 10 kb.

### Gene density, gene order, and recombination

BAC 115G01 includes genes 1 and 2 and BAC 116F02 genes 3, 4a and 4b that were similar to the genes previously described in the colinear region in rice and barley (Dubcovsky et al. 2001). No additional genes were found in these two *T. monococcum* BACs, and the final gene density was 1 gene per 43 kb. These five genes were clustered in two groups, the first one including genes 1 and 2 and the second one including genes 3, 4a and 4b. Gene density in these two islands was significantly higher than the contig average. Gene density in the first island was 1 gene per 6.2 kb, and in the second island

1 gene per 9.6 kb. The average intergenic region in these islands was 4.7 kb.

The five genes were in the same relative order in wheat, barley and rice (gene 4b is absent in rice). Wheat gene 2 was in reverse orientation compared to barley but in the same orientation as in rice (Dubcovsky et al. 2001). As expected, the wheat genes were very similar to their barley homologues not only in their exons (93% to 99% identical at the protein level) but also in the intron regions (89%, excluding indels). Lengths of exons and introns were also similar. Only 3 of the 71 compared exons showed differences in length between the two Triticeae species and a significant correlation was observed between wheat and barley intron lengths ( $r = 0.76$ ,  $P < 0.001$ ). Conservation between wheat and barley sequences in the 5' and 3' gene regions was limited to short stretches of 150–600 bp.

Genetic linkage between the most distal genes, gene 1 and gene 4b, was investigated in 771 F<sub>2</sub> plants (1,542 gametes) from two different crosses between diploid *T. monococcum* accessions. These two genes are separated in the *T. monococcum* contig by 175 kb of DNA. Only one recombination event was detected, representing a genetic distance of 0.065 cM (0.052–0.078, 95% confidence interval). Based on this value, the ratio of physical to genetic distance in this region can be estimated to be about 2,700 kb cM<sup>-1</sup> (2,243–3,365 kb cM<sup>-1</sup>, 95% confidence interval). Genes present in these two BACs are 0.1 cM distal to the vernalization gene *Vrn1* and none of them is a candidate for this vernalization gene. The frost tolerance gene *Fr1* was not segregating in these mapping populations and therefore its relative position to these genes is unknown.

### Intergenic regions

#### *Transposable elements nomenclature*

Given the enormous numbers of transposable elements existing in a genome and the large number of genomes that will someday be sequenced, it would be impossible to assign suitable names for each one. It is entirely possible that there are more different retroelement families within maize than there are words in the English language (SanMiguel and Bennetzen 1998). Nevertheless, applying some sort of designation seems appropriate, at least to elements that are likely to undergo further analysis.

Transposable elements found in this region were named using a family mnemonic separated with an underline character “\_” from the BAC name or GenBank accession number when the BAC name was not available. For each member of a family detected within the same BAC, consecutive numbers were added at the end of the name preceded by a hyphen, “-”, and assigned in the order of discovery. Capital letters were used to indicate families or groups of retroelements (e.g. BARE-1, WIS-2) whereas lowercase letters were used to refer to individual retroelements.

**Table 1** Retrotransposon designations, type, size, and orientation in the deposited sequence. The lengths of the 5' and 3' LTRs are indicated when present. Deletions in the LTRs were determined only when both LTRs were present

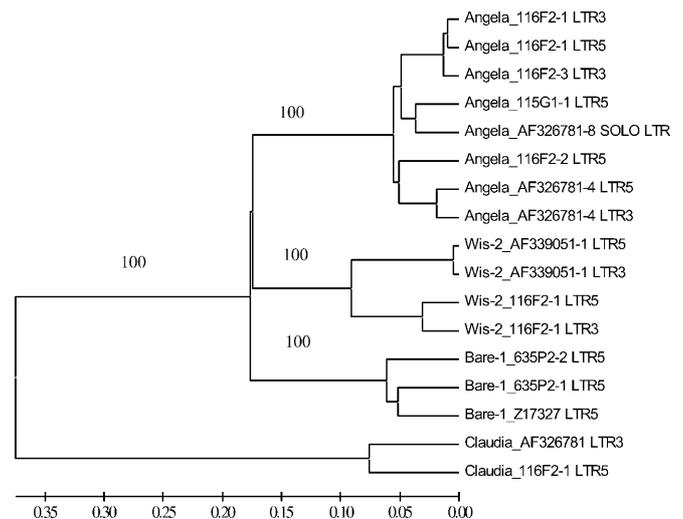
Designation	Type	Length (bp)	Strand	5' LTR size (bp)	5' LTR deletions (bp)	3' LTR size (bp)	3' LTR deletions (bp)
Angela_116F02-1	copia	8,534	+	1,737	1;1	1,738	1
Angela_116F02-2part	copia	5,943	-	1,756	-	-	-
Angela_116F02-3part	copia	2,844	+	-	-	1,740	-
Angela_115G01-1part	copia	6,420	+	1,684	-	-	-
Claudia_116F02-1part	copia	7,656	+	1,321	-	-	-
Wis_116F02-1	copia	8,766	+	1,736	-	1,730	1;1;4
Fatima_116F02-1part	gypsy	4,304	-	-	-	-	-
Latidu_116F02-1	gypsy	9,806	+	464	-	464	-
Latidu_115G01-1part	gypsy	4,827	-	-	-	463	-
Nusif_116F02-1part	gypsy	7,803	-	-	-	?	-
Sabrina_116F02-1part	gypsy	6,129	+	-	-	1,571	-
Sabrina_116F02-2	gypsy	7,370	-	1,550	18;6;1	1,559	13;3
Sabrina_115G01-1	gypsy	10,260	+	1,578	14;3;4;435	2,027	4;1;1;1
Sabrina_115G01-2	gypsy	7,530	-	1,571	1;3;1	1,573	2;1
Sukkula_116F02-1	gypsy	13,541	+	5,529	5	4,425	8;21;1,078;1
Wham_116F02-1	gypsy	9,609	-	1,415	2;3	1,402	17;1
Wham_115G01-1	gypsy	9,563	+	1,395	3;27	1,425	-
Wham_115G01-2	gypsy	11,681	-	1,415	18;2	1,418	1;1;15
Eway_115G01-1	TRIM	1,976	-	262	-	262	-
Veju_116F02-1	TRIM	2,520	-	374	-	374	-

A method for generating mnemonics from a small portion the sequence of a transposable element was developed and used in this study (details posted at [http://www.genomics.purdue.edu/~pmiguel/name\\_elements/examples](http://www.genomics.purdue.edu/~pmiguel/name_elements/examples)). In cases where a newly annotated element possessed substantial similarity to a previously described element, the name of that element was adopted as the mnemonic. The special mnemonic Wham (for wheat abundant mobile DNA) was adopted for a novel retrotransposon family that was present in a high copy number.

### *Ty1-copia transposable elements*

Six *Ty1-copia* retrotransposons were found in wheat BACs 116F02 and 115G01 (Table 1). The most common elements within this class were in the BARE-1 superfamily (Gibbon et al. 1999), with five family members. Although one of the five was a partial element and two others were truncated at the ends of this BAC contig, none were found as solo LTRs. The sixth element was similar to a retrotransposon designated Claudia found in another *T. monococcum* contig (AF326781, Wicker et al. 2001). Claudia\_116F02-1part is 3' truncated, but examination of the other partial copy of this element (AF326781) and the location of a typical tMet tRNA primer binding site (PBS) in this element allowed us to precisely identify the borders of its 5' LTR.

A cluster analysis of the LTRs of these six elements and of representative elements from the sub-groups BARE-1, WIS-2, ANGELA and CLAUDIA (see Materials and methods for accession numbers) was done (Fig. 1). The clustering analysis was used to assign mnemonics as follows. Four out of the six elements discov-



**Fig. 1** Cluster analysis of the long terminal repeats (LTRs) of *Ty1-copia* elements. UPGMA tree generated by MEGA2.1 using the pair-wise deletion option (Kumar et al. 2001). Bootstrap values for the ANGELA, BARE-1 and WIS-2 sub-groups and the BARE-1 group are indicated above the respective branches

ered in the wheat BACs 116F02 and 115G01 were clustered in a group with the two representative members of the ANGELA sub-group and were assigned the mnemonic Angela (Fig. 1). The fifth element formed a cluster with the representative of the Wis-2 sub-group and was assigned the mnemonic Wis. No elements from these wheat BACs clustered with the BARE-1 representatives. The ANGELA, WIS-2, and BARE-1 clusters had high bootstrap values (100) in both the UPGMA and Neighbor-Joining analyses. These three sub-groups formed a single major cluster with a high bootstrap value

(100) that separates the members of the BARE-1 superfamily (Gibbon et al. 1999) from the Claudia elements.

The termini of LTRs from the eight elements from the ANGELA subgroup were imperfect repeats TGAAGG-CCAACA, whereas the termini of LTRs from all the BARE-1 and WIS-2 elements were perfect repeats TGT-TGG-CCAACA. Analysis of the LTR alignments revealed 16 conserved indels that differentiate all the members of the ANGELA subgroup from all the members of the BARE-1 and WIS-2 elements.

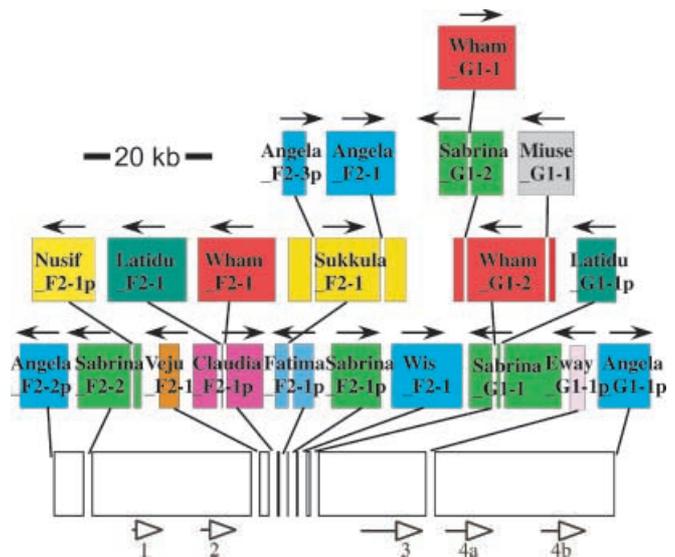
### Ty3-gypsy transposable elements

Twelve Ty3-gypsy retrotransposons were found in wheat BACs 116F02 and 115G01 (Table 1). Four of these elements belong to the SABRINA family, including the first full-length element from this family. Earlier, Sabrina elements were characterized as partial elements in barley (Shirasu et al. 2000). Another full-length representative of the SUKKULA family, previously characterized as solo LTRs in barley (Shirasu et al. 2000), was discovered in these wheat BACs.

In addition to the elements related to previously described Ty3-gypsy retrotransposons, six elements showed no significant DNA similarity with known Ty3-gypsy elements and were assigned the new mnemonics Wham (three elements), Latidu (two elements), and Nusif (one element, Table 1). The termini of LTRs of the three Wham elements were perfect inverted repeats TGATA-TATCA, whereas the two Latidu elements had imperfect repeats TGAGA-TCACA terminating their LTRs.

The Nusif element has a deletion that appears to have removed its 5' LTR without deleting any of the Sabrina\_116F02-2 element into which it is inserted. The length of Nusif\_116F02-1's 3' LTR cannot be ascertained. Two potential polypurine tracts that may be adjacent to the start of a 3' LTR can be located. These sequences, GGGGGCTACTGTC and GAGACGGAGATGTC, would place the start of the 3' LTR at 6,831 or 7,029 bases from the beginning of the element, respectively.

Fatima\_116F02-1part appears to be both 3' and 5' truncated. However, it was possible to locate its borders using sequence similarity to the Fatima element discovered by Wicker et al. (2001). In addition to elements depicted in Fig. 2, two other Ty3-gypsy were identified in wheat BACs 116F02 and 115G01 by homology to known elements but provided no means of defining their borders. The 1,851 bp immediately downstream of Angela\_116F02-2part showed good homology (67% protein similarity) with the 5' half of the GAG-POL precursor of a rice RIRE2 retrotransposon (AP002484) up to the stop codon. The location of Angela\_116F02\_2part relative to this RIRE2-like element suggests that it inserted within the RIRE2-like element. No significant BLASTX or BLASTN hits were detected for the 1,790 bp between the RIRE2-like element and the complete Sabrina\_116F02-2 element inserted downstream. It



**Fig. 2** Organization of retroelements. Distribution of colored horizontal bars represent retroelements; open horizontal bars represent host sequence into which retroelements have inserted. Diagonal lines extending from a retroelement into the space between bars below it are used to represent the order and position of insertion of the retroelements. Genes are denoted with numbered arrows. A duplicated gene is designated “4a” and “4b”. Arrows indicate the direction of transcription of genes and retroelements. Members of the same retroelement family share the same fill color. BARE-1 superfamily members, including Wis and Angela, are allotted a single color. A single non-LTR retroelement (retroposon) is designated with the fill color gray. Retroelement designations are explained in the text. Here they have been shortened such that “F2” denotes a retroelement originating or largely residing on the BAC 116F02. G1 is the designation for 115G01 (p part)

is possible that this region contains the 5' LTR of the RIRE2-like element.

The 5,456 bp between Sabrina\_116F02-2 and the start of gene 1 includes a 1,254-bp region with 44% protein similarity to another putative RIRE2 rice retroelement (AP002746) and 165 bp similar to a rice Retosat 2 retrotransposon (AF 111709). These similarities suggest the presence of an ancient transposable element that has been modified by sequence drift to a point beyond unambiguous recognition.

The 20 retrotransposons listed in Table 1 make up 70% of this BAC contig (Fig. 2).

### TRIM elements

Two thousand base pairs downstream of gene 2 there is a retrotransposon that was designated Veju\_116F02-1 that shows 88% similarity with a similar element in *Secale cereale* (SCAF000227). Its diminutive size (2.5 kb) suggests that it is a member of a sub-category of retrotransposons recently named “Terminal-repeat retrotransposons in miniature” (TRIM; Witte et al. 2001). Another member of this sub-category, Eway\_115G01-1, is inserted between genes 3 and 4a. These two elements share no nucleotide similarity.

**Table 2** Miniature terminal-inverted repeat transposable elements (MITEs): size of the complete elements and the tandem inverted repeats (TIR), and sequence of the target site duplications (TSD)

Location	Size (bp)	TIR	TSD
Gene 1, intron 4	38	3	TA
Gene 1, intron 4	81	7	TA
Gene 1, intron 5	17	5 <sup>a</sup>	TA
Gene 2, intron 5	99	9	TA
Gene 3, intron 9	65	11 <sup>a</sup>	TA
Gene 4A, intron 10	123	23	TA
Between genes 4A–4B	175	3?	TAA
Between genes 4A–4B	123	6?	TC/TG?
Gene 4B, intron 4	64	7	TTG
RIRE2 Gag-Pol	226	13	TC

<sup>a</sup> Imperfect

### Retroposons

Only one LINE retroposon was found in this region and was designated Miuse\_115G01–1. Possessing no LTRs, it would normally be difficult to specify its borders. However, because Miuse\_115G01–1 inserted in the 5' LTR of Wham\_115G01–2, comparison of its putative target site duplications to the empty site in the 3' LTR allows its borders to be easily delineated.

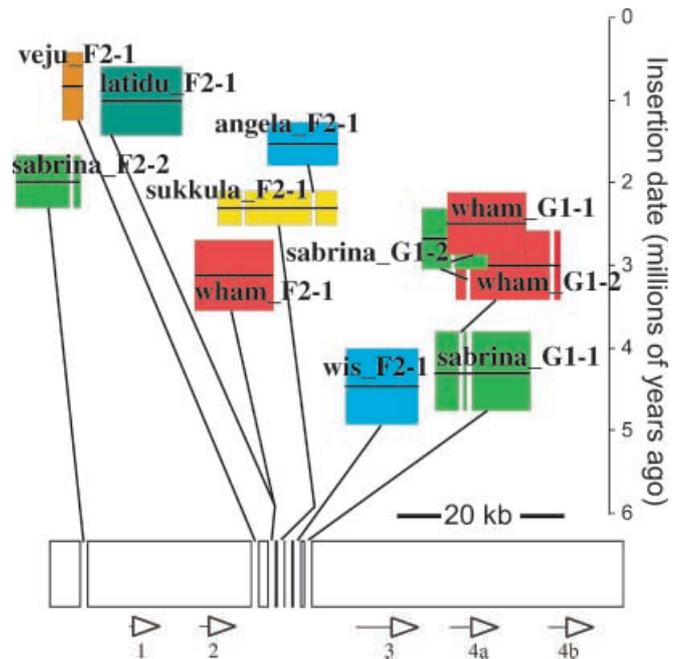
### MITEs and other terminal inverted repeat elements

MITEs are miniature terminal-inverted repeat transposable elements (Bureau and Wessler 1992, 1994). Most MITE elements were found in the gene regions (Table 2). However, a 231-bp MITE flanked by TC host duplications and with an 11-bp terminal inverted repeat (TIR) was found inserted within the RIRE2-like element. In the gene regions, nine MITEs were identified. The detection of the MITE insertions was facilitated by comparison with the homologous barley sequence (Dubcovsky et al. 2001). MITEs appear as clear interruption in the colinearity flanked by the target site duplications (TSD). No other terminal inverted repeat elements were discovered in this region.

### Dating retrotransposon insertions

The transposable elements present in this region are organized in a nested structure that includes four layers of nested insertions. The position in the nested structure provides a relative estimate of their insertion times.

More precise estimates of insertion time were possible for 11 elements, that have intact LTRs at both their 5' and 3' ends (Fig. 3, Table 3.) One retrotransposon, Eway\_115G01–1, had identical LTRs suggesting a very recent insertion. Four of the 11 datable elements involved insertions into other datable elements. In all four cases, differences in insertion ages corresponded with insertion order. That is, younger elements were found to be inserted into older ones.



**Fig. 3** Insertion times of retroelements. Diagonal lines, arrows, open and colored bars are as described in the legend to Fig. 2. Black horizontal lines inside boxes indicate estimated insertion date with one standard deviation above and below this date indicated by the height of the box. Retrotransposons which could not be dated are omitted from this figure. In cases where a dated retrotransposon had inserted into a retrotransposon that could not be dated a bend was added to the diagonal line to denote this

**Table 3** Time of insertion of retrotransposon in millions of years (MYA; with their standard deviation, SD; Tv number of transversions, Tr number of transitions). The proportion of transitions in CG-CNG sites (%TrCmet) was calculated as the number of transitions in CG-CNG divided by the total number of transitions

Designation	MYA	SD	Tv	Tr	%TrCmet
Veju_116F02–1	0.83	0.42	1	5	(50)
Latidu_116F02–1	1.01	0.41	1	5	(20)
Angela_116F02–1	1.53	0.26	8	26	88
Sabrina_116F02–2	1.99	0.32	15	25	80
Sukkula_116F02–1	2.30	0.20	37	92	76
Wham_115G01–1	2.48	0.38	15	27	70
Sabrina_115G01–2	2.67	0.37	16	37	81
Wham_115G01–2	3.01	0.42	16	38	76
Wham_116F02–1	3.12	0.43	16	39	82
Sabrina_115G01–1	4.29	0.47	30	54	67
Wis_116F02–1	4.47	0.46	28	68	75

The insertion ages presented in Table 3 for these 11 elements are consistent with the absence of these elements in the homologous region in barley (Dubcovsky et al. 2001), whose ancestor diverged from the wheat lineage between 10 and 14 million years ago (Wolfe et al. 1989). None of the other retroelements described above was present in the homologous region in barley, and none of the retroelements described in barley were found in the homologous wheat BACs. This suggests that all these elements were inserted in the last 10–14 million years.

**Table 4** Overall GC content of wheat compared with barley and rice. BACs including orthologous genes in wheat, barley and rice are indicated in *bold*

Species	Accession	Chromosome	Reference	Total bp	GC%
Wheat	AF326781	1	Wicker et al. (2001)	211,001	45.6
Wheat	<b>AF459639</b>	5	This manuscript	215,241	45.4
Wheat	NA	1	Lagudah et al. (2001)	128,505	46.3
Wheat	NA	Genomic	Lagudah et al. (2001)	1,020,000	45.7
<b>Barley</b>	<b>AY013246</b>	5	Dubcovsky et al. (2001)	102,433	44.2
Barley	AF254799	2	Shirasu et al. (2000)	65,979	45.6
Barley	Y14573	4	Panstruga et al. (1998)	59,748	46.0
Rice	AC084762	3	Unpublished	148,021	43.6
Rice	AC090882	3	Unpublished	158,838	41.9
Rice	AC091811	3	Unpublished	139,872	44.5
Rice	AC091494	3	Unpublished	150,663	44.5
Rice	AC092263	3	Unpublished	164,179	43.6
<b>Rice</b>	<b>AY013245</b>	3	Dubcovsky et al. (2001)	50,000	42.7
Rice	AC093093	10	Unpublished	141,050	43.0
Rice	AC087182	10	Unpublished	148,814	44.1
Rice	AC068923	10	Unpublished	136,120	43.9
Rice	AC078894	10	Unpublished	140,623	43.5
Rice	AC077693	10	Unpublished	187,207	44.2
Rice	AC084405	Unknown	Unpublished	166,451	44.1

An additional comparison of the wheat and barley homologous intergenic regions was performed after removal of all the non-orthologous retroelements present in both BACs. A 752-bp fragment of barley BAC 635P2 (bases 96,356–97,228) is 84% identical in sequence to a 750-bp segment of this wheat BAC contig (bases 206,989–207,852) when aligned with the program *cross\_match*. This level of similarity is consistent with the 89% identity (excluding indels) found between the wheat and barley introns from the five colinear genes present in this region. These conserved sequences are 85% and 84% (barley) and 88% and 89% (wheat) identical to disjoint segments of the Barbara element reported in GenBank accession AF326781 (Wicker et al. 2001).

#### Nucleotide composition

##### GC composition

The overall GC content of this contig was 45.4%, with higher values in exons (GC =52%±1% SE) and lower values in introns (GC =40%±2% SE). The 500 bp upstream of the 5' end of the five genes showed significantly higher ( $P < 0.05$ ) levels of GC content (GC =58%±5% SE) than the 500 bp downstream of the 3' end (GC =43%±2% SE). The overall nucleotide composition was similar to that observed in two other *T. monococcum* BACs (Lagudah et al. 2001; Wicker et al. 2001) and in the sequence of one megabase of nebulized genomic DNA from Chinese Spring, a cultivar of the hexaploid wheat *T. aestivum* (Lagudah et al. 2001; Table 4).

No significant differences were observed between the GC content values in wheat and the GC content reported for three barley BACs ( $P = 0.42$ , Table 4) but the combined Triticeae values (GC =45.5%±0.3% SE) were significantly higher than those observed for a set of 12 com-

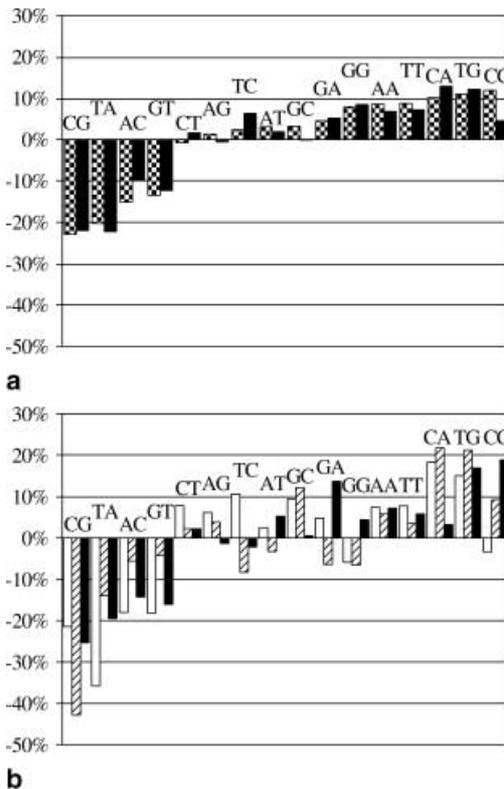
pletely sequenced rice BACs (GC =43.6%±0.2% SE,  $P = 0.0001$ ; Table 4).

The overall GC content in the Triticeae species is more than 20% higher than the overall GC content observed in the five chromosomes of *Arabidopsis* (The Arabidopsis Initiative 2000). This large overall difference is also reflected in the GC content of the genes. The four *Arabidopsis* genes that are most similar to the ones present on the wheat BACs (Dubcovsky et al. 2001) have a GC content (43.9%±0.6% SE) that is consistent with the average for 25,498 *Arabidopsis* genes (44.1%, The Arabidopsis Genome Initiative 2000) but that is significantly smaller ( $P < 0.0001$ ) than the GC content in the grass genes (GC =51.4%±0.6% SE).

##### Dinucleotide composition

Genomic dinucleotide frequencies were determined for the contig on chromosome 5A<sup>m</sup> presented here and for a contig on chromosome 1A<sup>m</sup> reported before (Wicker et al. 2001; Fig. 4A). In both regions, dinucleotide composition is not random. Certain dinucleotides (CG, TA, AC, and GT) appear at a lower frequency than would be expected by random chance, whereas others (GA, GG, AA, TT, CA, TG and CC) appear at higher than expected frequencies. A similar pattern has been reported before for vertebrate and invertebrate genomic sequences (Nussinov 1984).

Dinucleotide frequency in exons, introns and retroelements were also compared (Fig. 4B). Dinucleotide composition of retroelements largely mirrors that of the contig as a whole, a result that was expected because retroelements represent more than 70% of the contig. Exons follow the same distribution as retroelements although, in a number of cases, their skew is greater than that of the contig as a whole. This is particularly evident for the dinucleotide TA that showed a significantly lower frequency in exons than in retroelements ( $P = 0.007$ ) or introns



**Fig. 4** Dinucleotide composition. Dinucleotides frequencies were calculated as (observed – expected)/expected %. **A** Comparison between two *Triticum monococcum* contigs of more than 200 kb each (checkered bars BAC contig 116F02-115G01, shaded bars GenBank AF326781; Wicker et al. 2001). **B** Components of BAC contig 116F02-115G01 (open bars exons, hatched bars introns, filled bars retro-elements)

( $P = 0.003$ ; Fig. 4B). The lower frequency of TA in the coding regions may be related to the fact that TA is the first dinucleotide of the stop codons TAA and TAG (Arques and Michel 1995). Another difference was observed for the CG dinucleotide, which showed a larger depletion in introns compared to exons ( $P = 0.03$ ) or retroelements ( $P = 0.02$ ; Fig. 4B).

#### Transition rates in CG and CNG sites

The number of transitional changes that occurred in the CG and CNG sites ( $N^{\circ}\text{Cmet}$ ), that are the targets for cytosine methylation in plants (Gruenbaum et al. 1981), were compared with the number of transitions that occurred in other positions using the parameter %TrCmet ( $N^{\circ}\text{Cmet}/\text{total number of transitions} \times 100$ ).

Comparisons of the 5' and 3' LTRs of 11 retrotransposons showed that 76% of the 413 transitional mutations observed in these retroelements were in sites that are targets for cytosine methylation. The %TrCmet values ranged from 67% in Sabrina\_G1-1 to 88% in Angela\_F2-1 (Veju\_116F02-1 and Latidu\_116F02-1 were excluded from the analysis because they had only two and five transitions, respectively, Table 3).

A completely different result was obtained when the same comparison was performed for the introns of the five genes present in this BAC. Introns from wheat and barley were first aligned, and regions with indels were eliminated (13,886 aligned bp from the five genes). Only 37% of the 877 transitions were observed in the CG and CNG sites, with values ranging from 33% for the introns of gene 1 to 43% for the introns of gene 4b. A similar result was observed in the comparison of sorghum and maize introns for the *adh1* gene (%TrCmet = 35%). Therefore, the %TrCmet in the introns was less than half of the %TrCmet in the retrotransposons. If the differences in %TrCmet found in this study prove to be consistent between methylated and unmethylated regions, the characterization of this parameter in comparative sequence analyses may be a useful tool to infer the degree of methylation of different genomic regions.

## Discussion

### The relationship between genetic and physical distances

Cytogenetic studies using C-banding have shown that meiotic recombination was largely absent in proximal regions of the wheat chromosomes and that its general frequency increased exponentially with distance from the centromere (Dvorak and Chen 1984; Lukaszewski and Curtis 1993). Studies using deletion lines have shown a pattern similar to that observed in the cytogenetic studies. For example, recombination in the distal 25% of the long arm of chromosomes 5A, 5B and 5D was more than 20 times more frequent than recombination in the proximal 25% of the same chromosomes (Gill et al. 1996a).

The region studied here is located between the break-points in deletion lines 5AL-6 (FL 0.68) and 5AL-17 (FL 0.78), in a more proximal location than regions used before to estimate ratios between physical and genetic distances in wheat. Stein et al. (2000) reported a 1,400 kb  $\text{cM}^{-1}$  ratio for a 350-kb contig located in the telomeric region of chromosome 1A (6,240 gametes analyzed). Smaller ratios of 20–270 kb  $\text{cM}^{-1}$  were reported for the homologous region on chromosome 1D, but were calculated based on a small population of 58  $F_2$  plants (Spielmeyer et al. 2000). As predicted by the cytogenetic studies (Dvorak and Chen 1984; Lukaszewski and Curtis 1993), the rate of physical distance per unit of recombination was higher in the more proximal region analyzed in this study. The ratio of 2,700 kb  $\text{cM}^{-1}$  found between gene 1 and gene 4b was close to the average genome-wide estimate of 3,000 kb  $\text{cM}^{-1}$  (Bennett and Smith 1991).

It has been suggested that high recombination regions are associated with regions with high gene density (Gill et al. 1996a, b). Sequencing of 211 kb from a 350-kb contig from chromosome 1A<sup>m</sup> (Stein et al. 2000; Wicker et al. 2001) disclosed a gene density of 1 gene per 42 kb, a value that is almost identical to the 1 gene per 43 kb re-

ported in this study. In spite of the identical gene density, the proximal chromosome 5A contig showed a twofold higher physical distance per unit of recombination than the distal 1A contig. This suggests that gene density may not be the only factor determining the ratio between genetic and physical distances and that the relative position of the contig within the chromosome may also play a role in the determination of this ratio. Alternatively, as noted in maize (Xu et al. 1995) and other eukaryotes, some genes and/or regions of genes may be much more recombinationally active than others.

### Retrotransposon invasion

In the 240-kb interval of the maize genome around the *adh1* gene (SanMiguel et al. 1996; Tikhonov et al. 1999), 16 of the 23 retrotransposons possessed both LTRs, thereby allowing these insertions to be dated. All but two of these insertions had occurred within the last 3 million years, suggesting a recent and sudden increase in retrotransposon activity in this lineage (SanMiguel et al. 1998).

The *T. monococcum* BAC contig studied in this project is similar in retrotransposon composition to the maize *adh1* region (SanMiguel et al. 1996; Tikhonov et al. 1999). In the maize *adh1* region, 23 retrotransposons and 3 LINEs comprised about 71% of the entire region. Here, 20 retrotransposons and 1 retrotransposon accounted for about 70% of the total sequence. However retrotransposon insertion times in *T. monococcum* were more evenly distributed throughout the last 4.5 million years than in maize.

Deletions are common in both species. In maize, 3 out of the 23 retrotransposons were missing one or both LTRs and two additional retrotransposons were solo LTRs. Here (ignoring the 2 elements that were truncated by cloning sites at the ends of the BACs), 6 out of the 18 retrotransposons had lost one or both LTRs. Elements in both species show evidence of internal deletions, especially obvious by comparison of 5' and 3' LTRs of the same element (Table 1; SanMiguel and Bennetzen 1998).

No element older than 6 million years could be dated in either case. This suggests either that elements over 6 million years old are not commonly intact enough to date, or that the retrotransposon amplification in both species is a recent phenomenon. The extensive deletions observed in many of the retrotransposons present in this region indicate that there are active deletion mechanisms that have the potential to eliminate any similarity in the intergenic regions of grass species that diverged more than 10 million years ago. The presence of a tattered portion of a Barbara element in a likely orthologous position in both wheat and barley suggests, at least in the Triticeae, that some retrotransposon amplification occurred prior to the divergence of these two lineages some 10–14 million years ago (Wolfe et al. 1989). This hypothesis is also supported by the presence of very large chromosomes in all the members of the grass tribes Triti-

ceae, Poeae, Aveneae, Bromaeae, and Hainardieae that share the basic chromosome number  $n = 7$ . These two characteristics are considered derived in the grass family (Clayton and Renvoize 1986) and suggest that a large amplification of the genome size occurred during the origin of this group of grasses.

Studies in the maize *adh1* region, in this BAC contig and in another *T. monococcum* BAC contig (Wicker et al. 2001) are consistent with a model of grass genomes arising from a relatively compact ancestral genome that was swelled to varying degrees in descendent species. Further, retrotransposon amplification is likely to be the force responsible for most grass genome size increases (Bennetzen 2000). Because some grass genomes remained small (for example, rice) while others increased drastically, control of retrotransposon amplification and persistence seem obvious mechanisms for control of genome sizes over time. Deleting a retrotransposon or enough of one to destroy its ability to transpose and thus replicate could be a method of control. But other mechanisms of retrotransposon suppression are also known (Hirochika et al. 2000). Methylation, aside from any role it may have in silencing transcription of repetitive DNA, may promote increased mutation rates that can render a retrotransposon defective (SanMiguel et al. 1998; Singer et al. 1995).

The retrotransposons we have studied in wheat and in the maize *adh1* region, though plentiful in number, seem not to have inserted into genes. If host genomes are capable of forcing the majority of retrotransposons to insert into other repetitive DNA, a genome will preserve the function of its own genes and also inactivate the target retrotransposons. Bennetzen (2000) has proposed that this is a selectable trait that should evolve as an intrinsic property of high-copy-number retrotransposons. By specifically targeting other retrotransposons, especially their LTRs, a retrotransposon can acquire any possible beneficial regulatory properties of the element at the site of the insertion and can inactivate this competitor for retrotransposition. In the large genome plants that have been investigated by sample BAC sequencing so far, namely barley, maize and wheat, it is clear that the retrotransposons are the quantitatively dominant genome constituents. Their great variety and abundance suggests that LTR-retrotransposons might best be viewed as a collection of parasitic or mutualistic competitors within the nuclear ecosystem. Given the effects that these elements can have on adjacent genes and genomic segments after they insert (Kumar and Bennetzen 1999), their study can provide enormous insights into the nature and rate of gene and genome evolution.

### Nucleotide composition

The higher GC content observed in the Triticeae BACs compared with the rice BACs seems to be associated with the frequent insertion of large retrotransposons. These elements account for most of the DNA present in

the Triticeae BACs and their base pair composition largely determines the BAC base pair composition. This hypothesis was supported by the comparison of the GC content of the Triticeae and rice BACs with that from a few maize (AF391808, AF090447, AF123535) and sorghum (AF061282, AF124045) BACs. The GC content from these maize BACs, that also have a high proportion of retroelements, did not differ significantly from the GC content observed in the Triticeae BACs, but showed significantly higher GC content than the BACs from rice or sorghum ( $P < 0.01$ ). The analyzed BACs from the small genomes of rice and sorghum did not differ significantly in GC content.

Barakat and coworkers (1997) postulated the existence of a “gene space” with a particular GC composition based on preparative centrifugation in  $\text{Cs}_2\text{SO}_4$  (Barakat et al. 1997). This study shows that the GC content of wheat BACs 116F02 and 115G1 (A =26.9%, T =27.8%, C =22.6%, and G =22.7%) selected with gene probes and including two gene islands does not differ significantly from the GC content of the 1 Mb of nebulized genomic DNA from hexaploid wheat (A =27.2%, T =27.1%, C =22.7%, and G =23.0%; Lagudah et al. 2001). This result suggests that the base pair composition of the gene-rich regions is not significantly different from the average genome base pair composition in the Triticeae.

An unexpected result from the analysis of the dinucleotide frequencies was the discovery that introns from the five genes present in this region were significantly more depleted in CG dinucleotides than retrotransposons. This difference was not related to the differences in base pair composition in these two regions because the expected CG frequencies were calculated from the actual base pair composition of each intron and retrotransposon. CG dinucleotides (and CNG sites) are expected to be cytosine-methylated in the retrotransposons (Gruenbaum et al. 1981; SanMiguel et al. 1998) and this is probably the cause of the significantly higher proportion of transitions observed in CG and CNG sites in the retrotransposons (%TrCmet =76%) compared with introns (%TrCmet =37%). A similar higher rate of transitions in CG and CNG sites was observed in maize (SanMiguel et al. 1998). Based on this observation, a higher depletion of CG sites was expected in the retrotransposons compared to the introns.

However, depletion of CG sites would not tend to accumulate through generations of retrotransposons because, after having suffered a given number of mutations, retrotransposons should become incapable of transposition. Therefore, new insertions do not initially exhibit many of the dinucleotide specificities that can accumulate in non-functional elements. In addition, many of the retrotransposons found in this contig were of recent origin and did not have enough time to suffer a significant depletion of CG dinucleotides by the higher rate of transitions. The increased number of transitions in CG sites represented only 1% to 9% of the total CG sites present in the LTRs of the dated retrotransposons. When the CG

depletion values from the five introns were compared to the values from the nine retrotransposons with insertion times higher than 1.5 million years ago, the differences were no longer significant ( $P = 0.08$ ). Alternatively, selection forces may determine the higher depletion of CG pairs in the introns relative to the retro-transposons.

**Acknowledgements** The authors thank P. Parker, D. Murphy and J. Emberton for excellent technical assistance and M. Helguera and L. Yan for providing the clones used to select these BACs. This research was supported by a grant from the NSF PGRP (no. 9975793) and USDA NRI (2000-01678).

## References

- Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
- Arques DG, Michel CJ (1995) Analytical solutions of the sinucleotide probability after and before random mutations. *J Theor Biol* 175:533–544
- Barakat A, Carels N, Bernardi G (1997) The distribution of genes in the genomes of Gramineae. *Proc Natl Acad Sci USA* 94:6857–6861
- Bennett MD, Leitch IJ (1995) Nuclear DNA amounts in Angiosperms. *Ann Bot* 76:113–176
- Bennett MD, Smith JB (1991) Nuclear DNA amounts in Angiosperms. *Philos Trans R Soc Lond Ser B* 334:309–345
- Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. *Plant Mol Biol* 42:251–269
- Bureau TE, Wessler SR (1992) Tourist – a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell* 4:1283–1294
- Bureau TE, Wessler SR (1994) Stowaway – a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* 6:907–916
- Clayton WD, Renvoize SA (1986) *Genera Graminum. Grasses of the world*. Royal Botanic Gardens Kew, London
- Dubcovsky J, Lijavetzky D, Appendino L, Tranquilli G (1998) Comparative RFLP mapping of *Triticum monococcum* genes controlling vernalization requirement. *Theor Appl Genet* 97:968–975
- Dubcovsky J, Ramakrishna W, SanMiguel P, Busso C, Yan L, Shiloff B, Bennetzen J (2001) Comparative sequence analysis of colinear barley and rice BACs. *Plant Physiol* 125:1342–1353
- Dvorak J, Chen KC (1984) Distribution of nonstructural variation between wheat cultivars along chromosome arm 6Bp: evidence from the linkage map and physical map of the arm. *Genetics* 106:325–333
- Ewing B, Green P (1998) Base-calling of automated sequencer traces using PHRED. II. Error probabilities. *Genome Res* 8:186–194
- Faris JD, Haen KM, Gill BS (2000) Saturation mapping of a gene-rich recombination hot spot region in wheat. *Genetics* 154:823–835
- Gaut BS, Morton BR, McCaig BC, Clegg MT (1996) Substitution rate comparisons between grasses and palms – synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *Rbcl*. *Proc Natl Acad Sci USA* 93:10274–10279
- Gill KS, Gill BS, Endo TR, Boyko EV (1996a) Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. *Genetics* 143:1001–1012
- Gill KS, Gill BS, Endo TR, Taylor T (1996b) Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. *Genetics* 144:1883–1891

- Gordon D, Abajian C, Green P (1998) CONSED: a graphical tool for sequencing finishing. *Genome Res* 8:195–202
- Gribbon BM, Pearce SR, Kalendar R, Schulman AH, Paulin L, Jack P, Kumar A, Flavell AJ (1999) Phylogeny and transpositional activity of Ty1-copia group retrotransposons in cereal genomes. *Mol Gen Genet* 261:883–891
- Gruenbaum Y, Naveh-Many T, Cedar H, Razin A (1981) Sequence specificity of methylation in higher plant DNA. *Nature* 292:860–862
- Hirochika H, Okamoto H, Kakutani T (2000) Silencing of retrotransposons in *Arabidopsis* and reactivation by the *ddm1* mutation. *Plant Cell* 12:357–368
- Kato K, Miura H, Sawada S (1999) Comparative mapping of the wheat *Vrn-1* region with the rice *Hd-6* region. *Genome* 42:204–209
- Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16:111–120
- Kumar A, Bennetzen JL (1999) Plant retrotransposons. *Annu Rev Genet* 33:479–532
- Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Arizona State University, Tempe, Ariz.
- Lagudah E, Dubcovsky J, Powell W (2001) Wheat genomics. *Plant Physiol Biochem* 39:335–344
- Lijavetzky D, Muzzi G, Wicker T, Keller B, Wing R, Dubcovsky J (1999) Construction and characterization of a bacterial artificial chromosome (BAC) library for the A genome of wheat. *Genome* 42:1176–1182
- Lorow D, Jessee J (1990) Max efficiency DH10B: a host for cloning methylated DNA. *Focus* 12:19
- Lukaszewski AJ, Curtis CA (1993) Physical distribution of recombination in B-genome chromosomes of tetraploid wheat. *Theor Appl Genet* 84:121–127
- Nussinov R (1984) Doublet frequencies in evolutionary distinct groups. *Nucleic Acids Res* 12:1749–1763
- Panstruga R, Buschges R, Piffanelli P, SchulzeLefert P (1998) A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. *Nucleic Acids Res* 26:1056–1062
- SanMiguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann Bot* 82:37–44
- SanMiguel P, Tikhonov A, Jin Y-K, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274:765–768
- SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. *Nat Genet* 20:43–45
- Sarma RN, Gill BS, Sasaki T, Galiba G, Sutka J, Laurie DA, Snape JW (1998) Comparative mapping of the wheat chromosome 5A *Vrn-A1* region with rice and its relationship to QTL for flowering time. *Theor Appl Genet* 97:103–109
- SAS Institute (2001) SAS user's guide, version 8. SAS Institute, Cary, N.Y.
- Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res* 10:908–915
- Singer MJ, Marcotte BA, Selker EU (1995) DNA methylation associated with repeat-induced point mutation in *Neurospora Crassa*. *Mol Cell Biol* 15:5586–5597
- Spielmeier W, Moullet O, Laroche A, Lagudah ES (2000) Highly recombinogenic regions at seed storage protein loci on chromosome 1DS of *Aegilops tauschii*, the D-genome donor of wheat. *Genetics* 155:361–367
- Stein N, Feuillet C, Wicker T, Schlagenhauf E, Keller B (2000) Subgenome chromosome walking in wheat: a 450-kb physical contig in *Triticum monococcum* L. spans the Lr10 resistance locus in hexaploid wheat (*Triticum aestivum* L.). *Proc Natl Acad Sci USA* 97:13436–13441
- Sutka J, Galiba G, Vagujfalvi A, Gill BS, Snape JW (1999) Physical mapping of the *Vrn-A1* and *Frl* genes on chromosome 5A of wheat using deletion lines. *Theor Appl Genet* 99:199–202
- The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
- Tikhonov AP, SanMiguel PJ, Nakajima Y, Gorenstein NM, Bennetzen JL, Avramova Z (1999) Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. *Proc Natl Acad Sci USA* 96:7409–7414
- Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J* 26:307–316
- Witte CP, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc Natl Acad Sci USA* 98:13778–13783
- Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH (1989) Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proc Natl Acad Sci USA* 86:6201–6205
- Xu XJ, Hsia AP, Zhang L, Nikolau BJ, Schnable PS (1995) Meiotic recombination break points resolve at high rates at the 5' end of a maize coding sequence. *Plant Cell* 7:2151–2152