Sequence Organization and Conservation in *sh2/a1*-Homologous Regions of Sorghum and Rice

Mingsheng Chen, Phillip SanMiguel and Jeffrey L. Bennetzen

Genetics Program and Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907

Manuscript received July 3, 1997
Accepted for publication September 22, 1997

ABSTRACT

Previously, we have demonstrated microcolinearity of gene composition and orientation in *sh2/a1*-homologous regions of the rice, sorghum, and maize genomes. However, the *sh2* and *a1* homologues are only about 20 kb apart in both rice and sorghum, while they are separated by about 140 kb in maize. In order to further define sequence organization and conservation in *sh2/a1*-homologous regions, we have completely sequenced a 42,446-bp segment of sorghum DNA. Four genes were identified: a homologue of *sh2*, two homologues of *a1*, and a putative transcriptional regulatory gene. A solo long terminal repeat of the retroelement *Leviathan* was detected between the two *a1* homologues, and eight miniature inverted repeat transposable elements were found in this region. Comparison of the sorghum sequence with the sequence of the homologous segment from rice indicated that only the identified genes were evolutionarily conserved between these two species, which have evolved independently for over 50 million years. The introns of the *a1* homologues have evolved faster than the introns of the *sh2* homologue. The *a1* tandem duplication appears to be an ancient event that may have preceded the ancestral divergence of maize, sorghum, and rice.

MOST higher plants have complex genomes that contain large amounts of repetitive DNA. In maize, for instance, genes probably make up less than 20% of the nuclear genome. In most plants, the nature and organization of the nongenic regions remain largely unknown.

Rice, sorghum, and maize are three grass species of great agronomic value. They have very different haploid genome sizes; 430 Mbp pairs for rice, 750 Mbp pairs for sorghum, and 2500 Mbp pairs for maize (Arumuganathan and Earle 1991). Sorghum and maize are close relatives, having diverged from a common ancestor about 15–20 million years ago. Rice has undergone independent descent from sorghum and maize for over 50 million years. Rice and sorghum are true diploids, while maize is a segmental allotetraploid (Gaut and Doebley 1997). Despite their many years of independent evolution, maize, sorghum, and rice have retained a very similar gene repertoire (Bennetzen and Freeling 1993). They also have maintained extensive genetic map colinearity (Ahn *et al.* 1993; Hulbert *et al.* 1990; Moore *et al.* 1995; Whitkus *et al.* 1992).

Previously, we demonstrated microcolinearity in the *sh2/a1*-homologous regions of the rice, sorghum, and maize genomes (Chen *et al.* 1997). While the physical distance between *sh2* and *a1* is about 140 kb in maize, it is about 20 kb in both sorghum and rice. The transcriptional orientation of *sh2* and *a1* homologues is conserved among rice, sorghum, and maize. The *sh2/a1*-homologous region in rice has been completely sequenced (Chen and Bennetzen 1996). Three candidate genes were identified: Sh2, A1, and gene X.[1] Gene X appears to encode a transcription factor, as indicated by putative protein-protein interaction and zinc finger domains. Eight miniature inverted repeat transposable elements (MITEs) were identified in this region.

To characterize the sequence organization of the *sh2/a1*-homologous region in sorghum further and to investigate the nature and evolution of sequences in *sh2/a1*-homologous regions, we have completely sequenced the *sh2/a1*-homologous region in sorghum. Comparison of the sequence in these homologous regions of rice and sorghum reveals that the genes are conserved, but that the intergenic spaces are highly diverged.

MATERIALS AND METHODS

Materials: Genomic DNA was from *Sorghum bicolor* line BTx623. Sorghum bacterial artificial chromosome clones and subclones were as previously described (Chen *et al.* 1997). Restriction enzymes were from Promega (Madison, WI), New England Biolabs (Beverly, MA), and GIBCO (Grand Island, NY). DNA sequencing reagents were from Pharmacia (Piscataway, NJ).

DNA isolation and gel blot hybridization: Genomic DNA isolation, restriction enzyme digestions, agarose gel electro-

Corresponding author: Jeffrey L. Bennetzen, Department of Biological Sciences, Purdue University, West Lafayette, IN 47907
E-mail: maize@bilbo.bio.purdue.edu

[1] Gene symbols referring to known maize genes appear as lower case italics. However, there are no Sh2, A1, or X genes certified in sorghum or rice. Hence, we use only operating names (Sh2, A1-a, A1-b, and X) for these putative loci, without italics. We cannot name sorghum or rice genes with only indirect information.

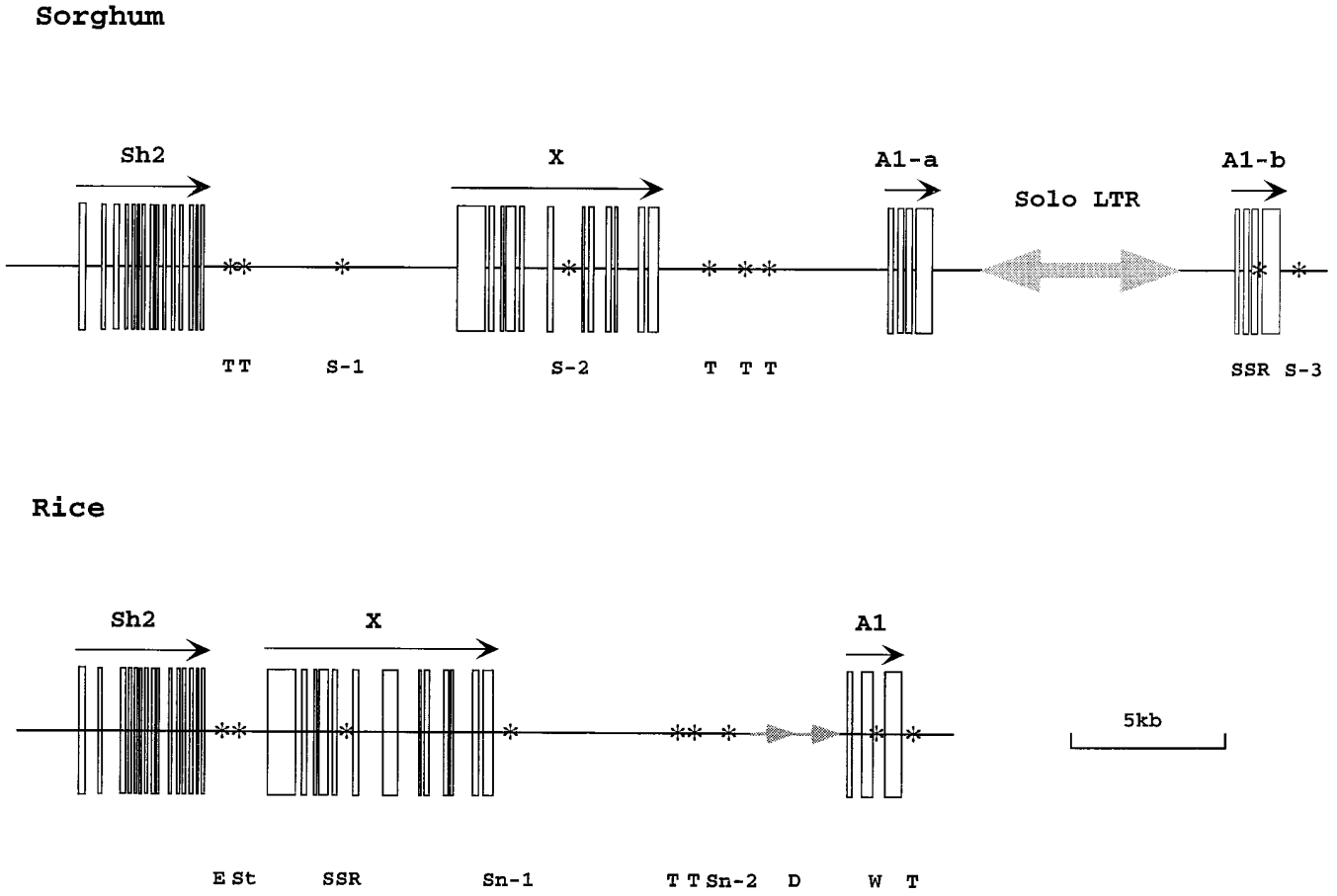

Figure 1.—Maps of *sh2*/*a1*-homologous regions of sorghum and rice. Lines represent the entire sequenced 42,446 bp of sorghum and 30,034 bp of rice (Chen and Bennetzen 1996). Empty boxes represent putative exons, and arrows above the boxes indicate their transcriptional orientation. The two shaded opposite arrows represent a solo LTR of the retroelement *Leviathan* (Bennetzen 1996) in sorghum. The two direct shaded arrows indicate a tandem duplication of 1432 bp in rice (the orientation of the arrows is arbitrary). Asterisks with letters below them indicate other sequence features. D, duplication; E, *Explorer* element; S-1, S-2, S-3, putative MITEs of sorghum; Sn-1, *Snabo-1*; Sn-2, *Snabo-2*; SSR, simple sequence repeat; St, *Stowaway* element; T, *Tourist* element; W, *Wanderer* element.

phoresis, filter replication, and hybridization were all as previously described (Chen *et al.* 1997).

DNA sequencing: A γδ transposon-facilitated DNA sequencing strategy was used (Strathmann *et al.* 1991). Cycle sequencing products were run on an express automated sequencer (ALF; Pharmacia [Piscataway, NJ]).

Sequence analysis: All sequence analysis was done using the software package from the University of Wisconsin Genetics Computer Group, Madison (Devereaux *et al.* 1984), through the Purdue University AIDS Center Laboratory for Computational Biology.

Sequence homologies were defined using the Pileup program with a penalty of 3.00 for gaps and of 0.10 for each bp of gap length. The presented nucleotide identities counted each sequence difference as a single change and each gap or insertion as a single change. We also estimated pairwise sequence divergence rates using only nongapped positions and generated highly similar results (data not shown).

RESULTS

The sequence of the *sh2*/*a1*-homologous region of sorghum: The completed sequence of the *sh2*/*a1*-homologous region in sorghum (Figure 1) is 42,446 bp (GenBank accession number AF010283). Sequence data redundancy (sum of total raw sequence data divided by length of the region sequenced) was 3.2, and all re-

TABLE 1

Comparison of the *sh2* homologues in sorghum, rice, and maize genomes

			Percent Rice Sh2	Percent Maize *sh2*
Sorghum Sh2	Nucleotide identity	Exon	83	95
		Intron	54	88
	Protein	Identity	77	91
		Similarity	87	95
Maize *sh2*	Nucleotide identity	Exon	82	
		Intron	57	
	Protein	Identity	77	
		Similarity	87	

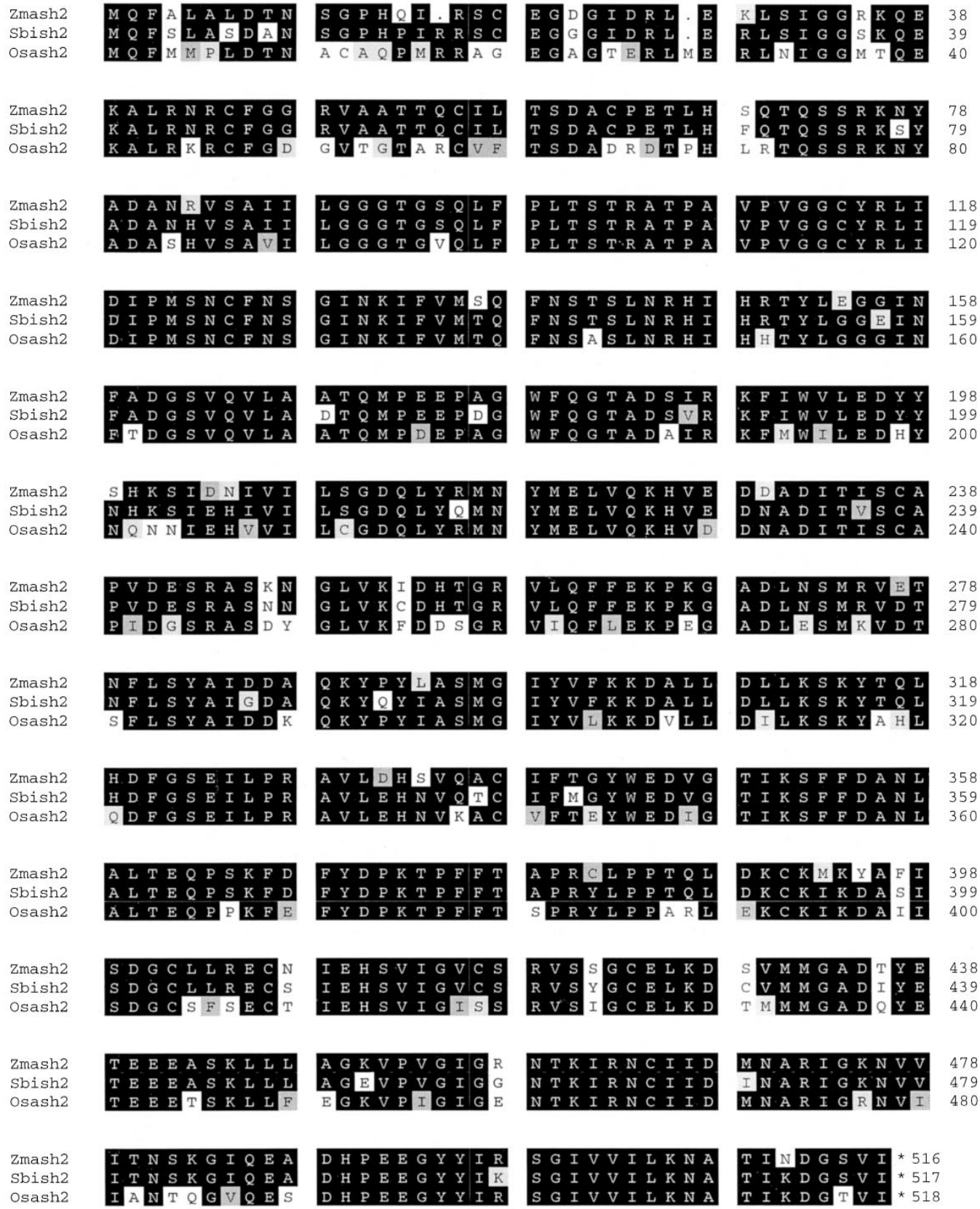

Figure 2.—Comparison of predicted amino acid sequence encoded by *sh2* homologues of sorghum, rice, and maize. Asterisks represent stop codons. Osash2, *sh2* homologue of rice (Chen and Bennetzen 1996); Sbish2, *sh2* homologue of sorghum; Zmash2, *sh2* homologue of maize (Hannah and Shaw 1992).

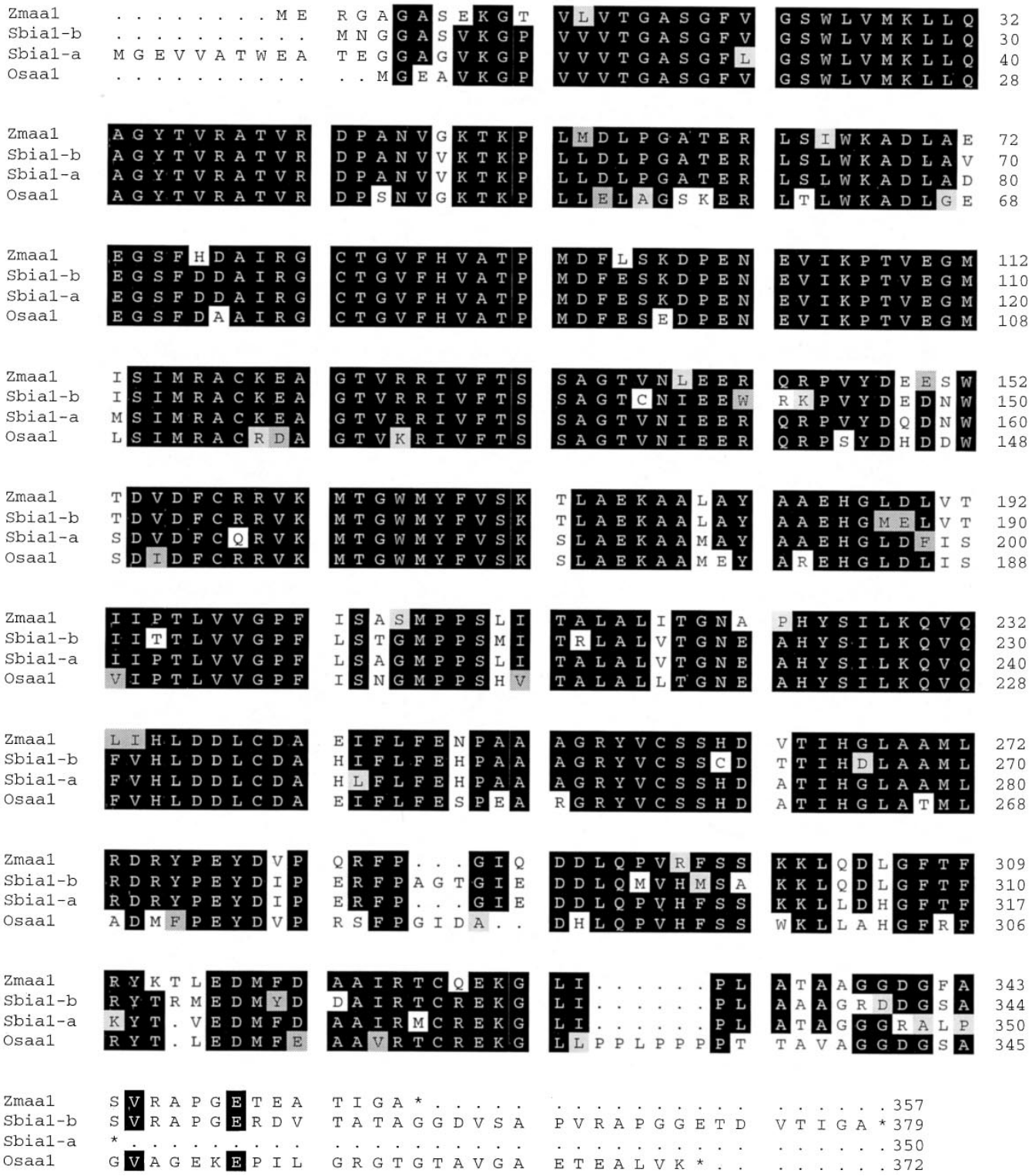

Figure 3.—Comparison of predicted amino acid sequence encoded by *a1* homologues of sorghum, rice, and maize. Asterisks represent stop codons. Dotted lines represent gaps. Sbia1-a, A1-a homologue of sorghum; Sbia1-b, A1-b homologue of sorghum; Osaa1, *a1* homologue of rice (Chen and Bennetzen 1996); Zmaa1, *a1* homologue of maize (Schwarz-Sommer *et al.* 1987).

gions were sequenced on both strands.

On the basis of comparisons to the maize and rice sequences, this region contains four genes; an *sh2* homologue, a gene homologous to gene X of rice (Chen and Bennetzen 1996), and two *a1* homologues. The *sh2* homologue in sorghum appears to have 15 protein-encoding exons. The two *a1* homologues have four exons each. The distance between the 3′ end of the *sh2*

homologue and the 5′ end of the closer *a1* homologue is 22,025 bp. Gene X was identified between the *sh2* homologue and the nearer *a1* homologue. It appears to have 12 protein-encoding exons. A solo long terminal repeat (LTR) of the retroelement *Leviathan* (Bennetzen 1996) was identified between the two *a1* homologues. Five *Tourists* were observed and three putative MITEs were also detected. A simple sequence repeat (SSR) with 21 ATs is present in the third intron of the A1-b homologue.

Comparison of *sh2* homologues in sorghum, rice, and maize: The *sh2* homologue in sorghum apparently encodes 517 amino acids, compared to 516 and 518 amino acids in maize and rice, respectively. Sequence identity between the *sh2* homologues in sorghum, rice, and maize is listed in Table 1. An amino acid sequence comparison is shown in Figure 2.

As expected, protein-encoding portions of exons are significantly more conserved than introns within these exons. The *sh2* homologue in sorghum is more homologous to the *sh2* homologue in maize than to the *sh2* homologue in rice for both exons and introns. The degree of homology between the *sh2* homologues of sorghum and rice is very similar to the degree of homology between the *sh2* homologues of maize and rice.

Comparison of *a1* homologues in sorghum, rice, and maize: The A1-a and A1-b genes in sorghum apparently encode peptides of 350 and 379 amino acids, respectively, compared to orthologous 357 and 372 amino acid peptides in maize and rice. These substantial differences in predicted protein size are primarily due to variations in stop codon location at the C terminus of each putative peptide (Figure 3).

Sequence comparisons of the *a1* homologues in the sorghum, rice, and maize genomes are displayed in Table 2. An amino acid sequence alignment is shown in Figure 3. As found previously, exons are more conserved than introns. The sequence homology is higher between sorghum and maize than between sorghum and rice or maize and rice.

Each of the A1-a and A1-b sorghum genes has four exons, the same number as in maize, but one more than in rice. The *Wanderer* element located in the second intron of the rice *a1* homologue is not present in either sorghum *a1* homologue.

Gene X between the *sh2* and *a1* homologues: Gene X was identified between the *sh2* and *a1* homologues by comparison to the sequences of the *sh2*/*a1*-homologous region in rice. Gene X has 12 exons in sorghum. It is predicted to encode an 895 amino acid protein, compared to 1070 amino acids in rice gene X. The putative zinc finger motif identified in the 3′ end of the rice version is not present in the sorghum sequence. At the nucleotide level, gene X of sorghum and gene X of rice share, 82% and 57%, respectively, identity of exons and introns. At the protein level, they are 79% identical and 85% similar (Figure 4). The SSRs in the fifth intron of gene X in rice are not present in the gene X sequence in sorghum. A putative MITE is positioned in the sixth intron of sorghum gene X, but it is absent in the rice gene X sequence.

Sequence comparison of *sh2/a1*-homologous regions in the rice and sorghum genomes: We compared the entire sequenced regions in rice and sorghum to define other conserved components in the *sh2*/*a1*-homologous regions. A window size of 50 and a stringency of 40 were used to reveal homologies ≥80% (Figure 5). By this criterion, only the protein-encoding portions of the *sh2* homologues, gene X, and the *a1* homologues were detected.

Tandem duplication of the *a1* homologue in sorghum: The *a1* homologue was directly duplicated in the *sh2*/*a1*-homologous region of sorghum. The distance between the putative stop codon of A1-a and the presumptive start codon of A1-b is 9756 bp. To characterize the features of the duplication, we made a comparison between these two *a1* homologues. A window size of 50 and a stringency of 33.3 (66.7% homology) were used to analyze the sequences (Figure 6). The comparison indicated that only the *a1* coding regions are highly homologous. The 5′ end and the 3′ end are not homologous, except that there are an 89-bp stretch with 72% identity on the 5′ end of each *a1* homologue (201–289 bp upstream of the putative A1-a start codon and 281–357 bp upstream of the apparent A1-b start codon) and a 94-bp stretch with 80% identity on the 3′ end of each *a1* homologue (82–175 bp downstream of the putative A1-a stop codon and 184–270 bp downstream of the presumptive A1-b stop codon). Further analysis revealed that the 5′ and 3′ end homologies are conserved in the respective 5′ and 3′ termini of the *a1* gene in maize. Putative TATA and CAAT boxes are located within the 5′ end homologies, and potential poly-

TABLE 2

Comparison of the *a1* homologues in rice, maize, and sorghum genomes

			Percent Rice A1	Percent Maize a1	Percent Sorghum A1-b
Sorghum A1-a	Nucleotide identity	Exon	84	91	92
		Intron	52	55	62
	Protein	Identity	80	85	87
		Similarity	89	92	92
Sorghum A1-b	Nucleotide identity	Exon	78	90	
		Intron	54	69	
	Protein	Identity	72	83	
		Similarity	81	90	
Maize *a1*	Nucleotide identity	Exon	82		
		Intron	52		
	Protein	Identity	75		
		Similarity	83		

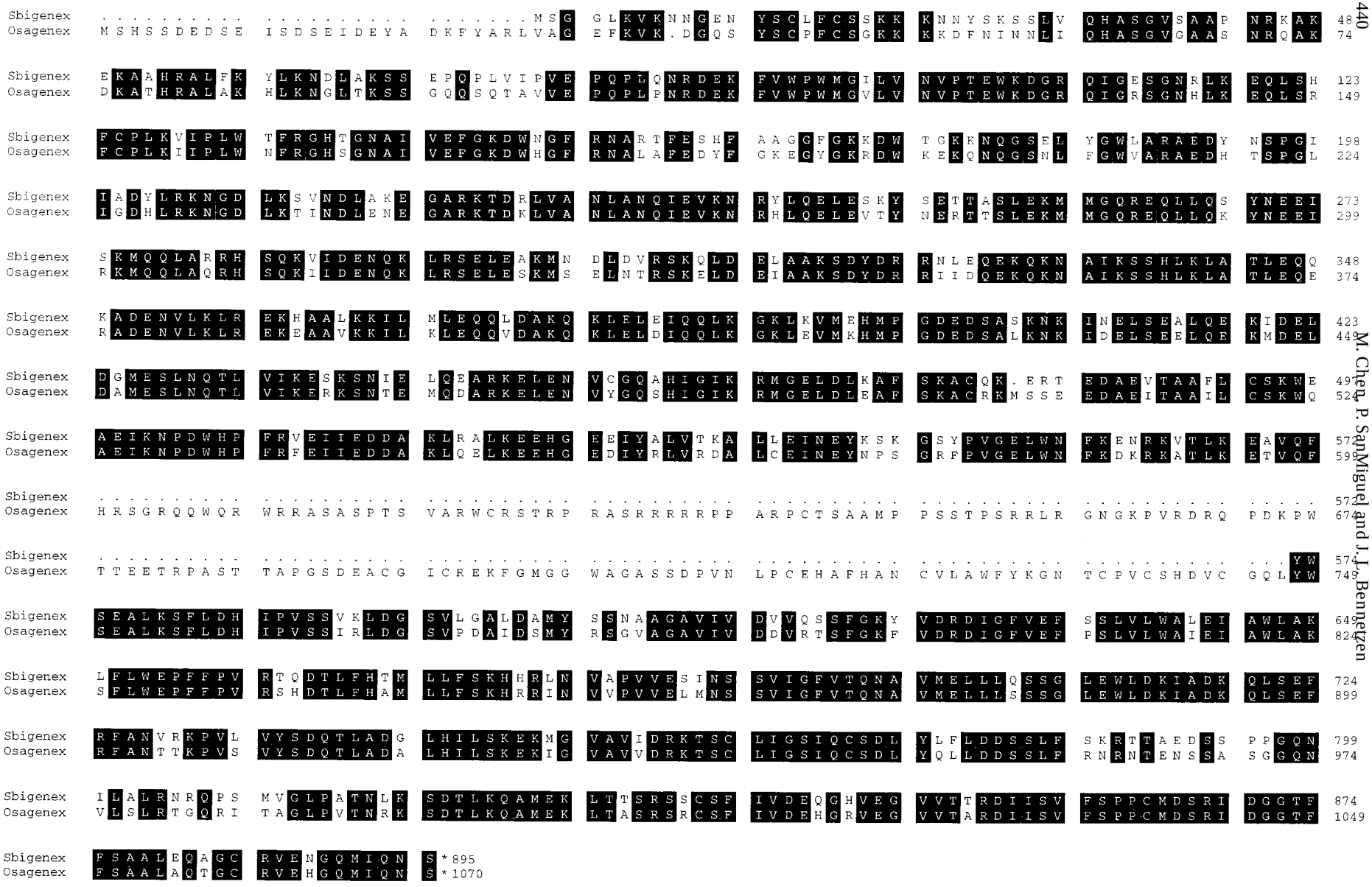

Figure 4.—Comparison of predicted amino acid sequence encoded by gene X orthologues of sorghum and rice. Asterisks indicate stop codons. Dotted lines represent gaps. The gap in position 572–573 in Sbigenex represents the putative zinc finger motif missing in the sorghum gene X homologue. Osagenex, gene X homologue of rice (Chen and Bennetzen 1996); Sbigenex, gene X homologue of sorghum.

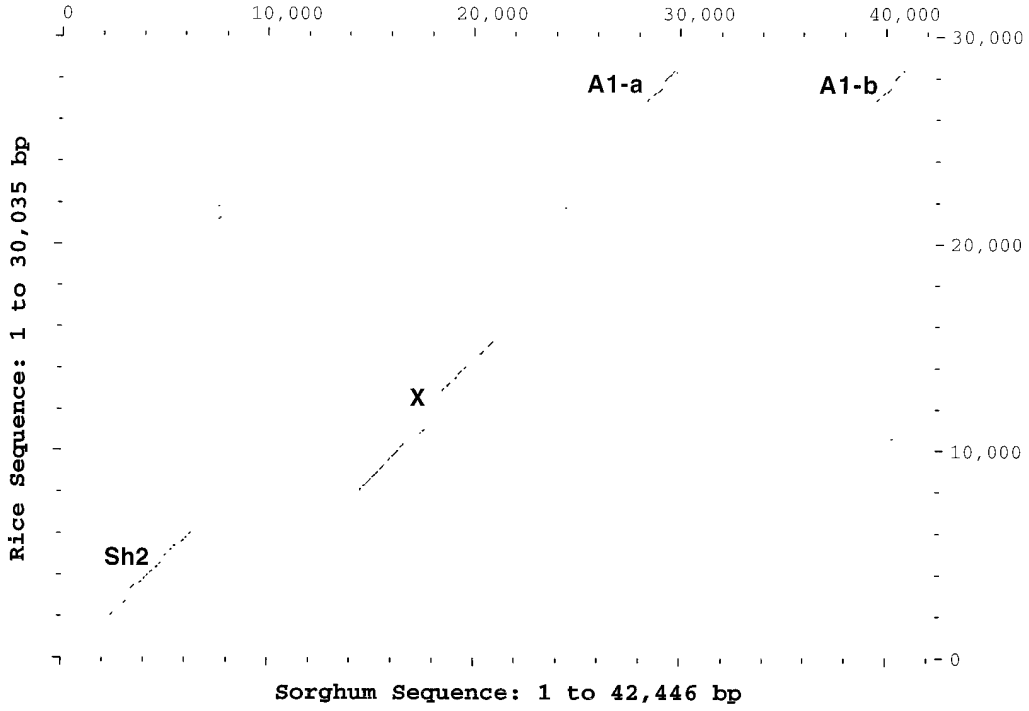

Figure 5.—DNA sequence comparison of *sh2*/*a1*-homologous regions of rice and sorghum.

adenylation signals are located within the 3′ end homologies.

A solo LTR between the two *a1* homologues: A solo LTR of the retroelement *Leviathan* (Bennetzen 1996) was detected between the two *a1* homologues (Figure 1). It is 4479 bp long and flanked by a 5-bp CTACA target site duplication. This solo LTR shares 53% identity with the 5′ LTR of the *Leviathan* element described previously (Bennetzen 1996). The solo LTR in our clone might have resulted from an unequal recombination event between the two LTRs of a complete *Leviathan* during the cloning process. To investigate this possibility, we used a 2.5-kb *Kpn*I/*Pme*I single-copy DNA fragment between the solo LTR and A1-b as a probe in gel blot hybridization experiments. For every restriction enzyme used (*Bam*HI, *Eco*RI, *Eco*RV, *Hin*dIII, *Sac*I, and *Xho*I), the hybridization pattern was the same for the sorghum genomic DNA and the sorghum BAC DNA (data not shown). It appears that this solo LTR is present, in this form, between the *a1* homologues in the sorghum genome.

MITEs in the sorghum *sh2/a1* region: Five *Tourist* transposable elements were identified in this region. Two are 3′ to the *sh2* homologue, while three are between gene X and A1-a (Figure 1). Three putative MITEs were observed (S-1, S-2, and S-3 in Figure 1). S-1 is located between the *sh2* homologue and gene X. It is 678 bp long, with perfect, 43-bp terminal inverted repeats.

S-2 is positioned in the sixth intron of gene X. It has two homologues in the EMBL/GenBank database. One is in the second intron of the nucleic-acid-binding protein gene of maize (Cook and Walder 1992), while the other is in the 5′ end of the 27-kD zein gene of maize (Das *et al.* 1991). S-2 and the element in the nucleic-acid-binding protein gene have 70% homology over 124 bp. S-2 and the element in the 27-kD zein gene have 61% homology over 136 bp. S-2 can form a fold-back structure with 70% complementarity over 103 bp.

S-3 is positioned in the 3′ end of A1-b. It has one homologue in the GenBank database, which is located in the 3′ end of the *ltp2* (lipid transfer protein) gene of sorghum (Pelese-Siebenbourg *et al.* 1994). They share 78% identity over 147 bp. S-3 can form a snap-back structure with 78% complementarity over 98 bp.

Target site duplications were not identified for S-1, S-2, or S-3. Terminal repeats (either inverted or direct) were also not identified for S-2 or S-3.

DISCUSSION

We have completely sequenced the *sh2*/*a1*-homologous region in sorghum. Mapping studies (A. Melake-Berhan and J. L. Bennetzen, unpublished results) position these two loci in an orthologous location to their homologues in maize. Analysis of sequence data identified four putative genes: a homologue of maize *sh2*, a homologue of rice gene X, and two homologues of maize *a1*. The order and orientation of the genes are conserved among sorghum and rice, but whether gene X exists in the orthologous position of maize is not known. Sequence comparison of this region in rice and sorghum revealed that only these genes are conserved. All the other sequence features (such as MITEs, SSRs, and other noncoding regions) are not recognizably conserved.

Maize and sorghum are close relatives despite their

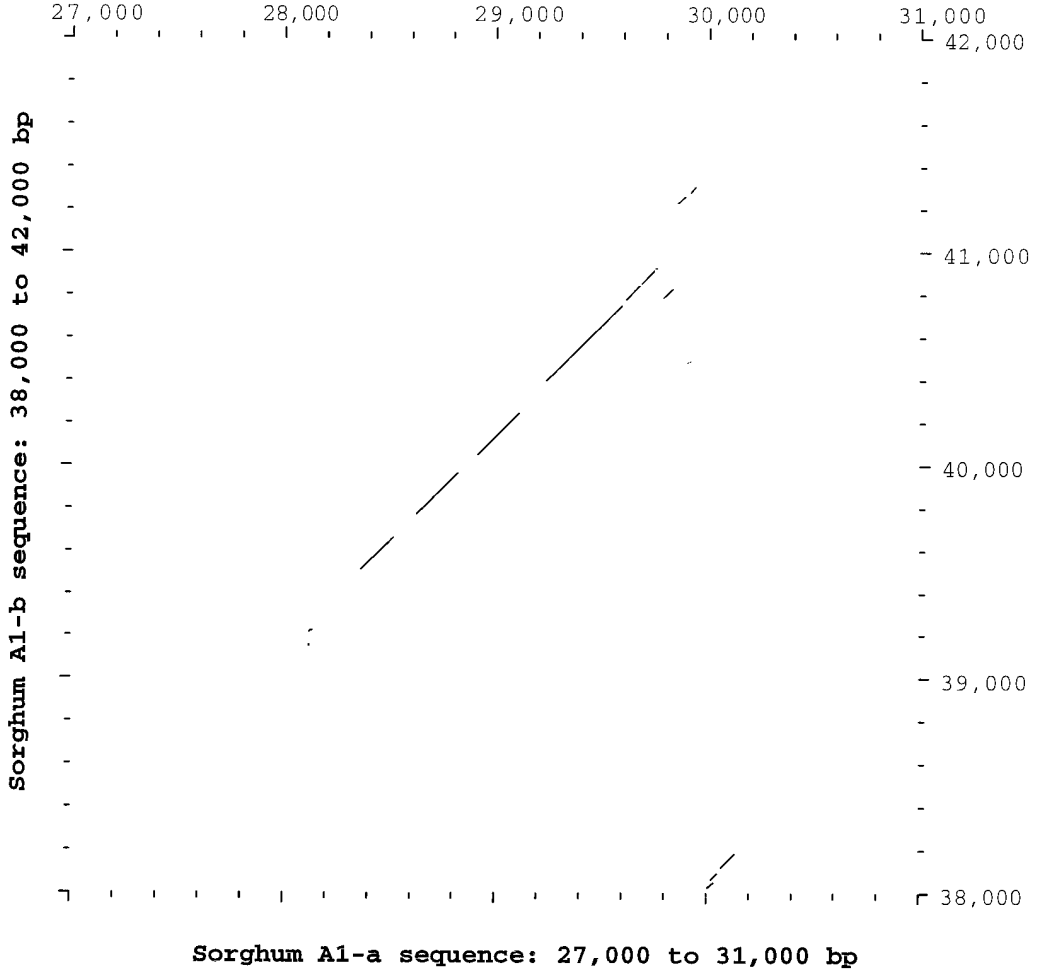

Figure 6.—DNA sequence comparison of the tandem duplication of *a1* homologue in sorghum.

different genome sizes. They have retained the same gene content and gene order (Whitkus *et al.* 1992), although maize has a 3.5-fold larger genome (Arumuganathan and Earle 1991). In the *Adh1* region of maize, intergenic regions are composed of a small set of nested retroelements that make up over 50% of the maize genome (SanMiguel *et al.* 1996). In the 42-kb region that we have sequenced in sorghum, only a 4.5-kb solo LTR of the retroelement *Leviathan* was identified, accounting for less than 15% of this region. In maize, the physical distance between *sh2* and *a1* is 140 kb (Civardi *et al.* 1994), whereas in sorghum it is only 22 kb. Consequently, nested retrotransposons between the genes in maize probably account for the majority of the over sixfold size difference in this region.

Comparison of the sorghum and maize sequences reveals that the introns in the *sh2* homologue are more conserved than the introns of the *a1* homologues. The noncoding regions of these genes apparently have different rates of divergence during evolution, despite their tight linkage. We know that the *sh2* homologue in rice is transcriptionally active because exons of the *sh2* genomic sequence were 100% identical to a cDNA sequence from rice (Chen and Bennetzen 1996). It is not clear whether the *a1* homologues in sorghum and rice are expressed. However, the high homology shared with the maize *a1* homologue, especially within the exons, suggests that they are active. The reason for the different rates of divergence for introns of closely linked loci and their possible implications in evolution remain unknown.

Gene X was first identified in the rice *sh2*/*a1*-homologous region (Chen and Bennetzen 1996). It is transcriptionally active in rice and encodes a putative transcription factor. By comparison to the rice sequence, we have identified a gene X homologue in this region in sorghum. Surprisingly, the putative zinc finger motif is missing in the sorghum sequences. This putative zinc finger motif belongs to the ring finger family and might participate in protein-protein interactions (Berg and Shi 1996). The implications of the absence of the zinc finger motif on the function of sorghum gene X remain to be investigated, but the absence of this motif suggests that the sorghum gene may have fewer protein interactions. We have identified an expressed gene X relative in maize by searching a cDNA database (B. Bowen, personal communication), which revealed a maize endosperm-specific homologue. Therefore, gene

X homologues exist in the sorghum, rice, and maize genomes. They are located between the *sh2* and *a1* homologues in rice and sorghum and may also be in this orthologous position in maize.

We have characterized a direct tandem duplication of *a1* homologues in sorghum. The exons of these genes are highly homologous, but the 5′ and 3′ noncoding regions are not conserved. This tandem duplication presumably arose by unequal recombination, but the duplicated region has apparently undergone extensive subsequent divergence. The similarity in sequence between A1-a and A1-b is not much greater than their homology to the *a1* locus of maize, suggesting that these duplicated genes have diverged for about 15–20 million years. The timing of the duplication event could have been more ancient, however, because subsequent unequal conversions might have permitted a reasonable duration of concerted evolution (Elder and Turner 1995). Laughnan (1952, 1955) demonstrated that some *a1* alleles in maize also exist as directly duplicated tandems. We have recently observed that the *a1* tandem duplication is present in the *sh2*/*a1*-homologous region in rice, where the two *a1* homologues are separated by about 5 kb (M. Chen and J. L. Bennetzen, unpublished data). These results suggest that the *a1* tandem duplication may have occurred before the divergence of maize, rice, and sorghum from a common ancestor. Alternatively, independent tandem duplication of *a1* orthologues must be a relatively frequent event.

We thank J. Wendel for helpful comments on this manuscript and S. Frank for technical assistance. This research was supported by United States Department of Agriculture grant #94-37300-0299.

LITERATURE CITED

Ahn, S., J. A. Anderson, M. E. Sorrells and S. D. Tanksley, 1993 Homoeologous relationships of rice, wheat and maize chromosomes. Mol. Gen. Genet. **241:** 483–490.

Arumuganathan, K., and E. D. Earle, 1991 Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. **9:** 208–218.

Bennetzen, J. L., 1996 The contributions of retroelements to plant genome organization, function and evolution. Trends Micro. **4:** 347–353.

Bennetzen, J. L., and M. Freeling, 1993 Grasses as a single genetic system: genome composition, colinearity and compatibility. Trends Genet. **9:** 259–261.

Berg, J. M., and Y. Shi, 1996 The galvanization of biology: a growing appreciation for the roles of zinc. Science **271:** 1081–1085.

Chen, M., and J. L. Bennetzen, 1996 Sequence composition and organization in the *sh2*/*a1*-homologous region of rice. Plant Mol. Biol. **32:** 999–1001.

Chen, M., P. SanMiguel, A. C. De Oliveira, S.-S. Woo, H. Zhang *et al.*, 1997 Microcolinearity in the *sh2*-homologous regions of the maize, rice and sorghum genomes. Proc. Natl. Acad. Sci. USA **94:** 3431–3435.

Civardi, L., Y. Xia, K. J. Edwards, P. S. Schnable and B. J. Nikolau, 1994 The relationship between genetic and physical distances in the cloned *a1-sh2* interval of the *Zea mays L.* genome. Proc. Natl. Acad. Sci. USA **91:** 8268–8272.

Cook, W. B., and J. C. Walker, 1992 Identification of a maize nucleic acid-binding protein (NBP) belonging to a family of nuclear-encoded chloroplast proteins. Nucleic Acids Res. **20:** 359–364.

Das, O. P., E. Poliak, K. Ward and J. Messing, 1991 A new allele of the duplicated 27KD zein locus of maize generated by homologous recombination. Nucleic Acids Res. **19:** 3325–3330.

Devereaux, J., P. Haeberli and O. Smithies, 1984 A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12:** 387–395.

Elder, J. F., and B. J. Turner, 1995 Concerted evolution of repetitive DNA sequences in eukaryotes. Quart. Rev. Biol. **70:** 297–320.

Gaut, B. S., and J. F. Doebley, 1997 DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci. USA **94:** 6809–6814.

Hannah, L. C., and J. R. Shaw, 1992 Genomic nucleotide sequence of a wild-type *shrunken-2* allele of *Zea mays*. Plant Physiol. **98:** 1214–1216.

Hulbert, S. H., T. E. Richter, J. D. Axtell and J. L. Bennetzen, 1990 Genetic mapping and characterization of sorghum and related crops by means of maize DNA probes. Proc. Natl. Acad. Sci. USA **87:** 4251–4255.

Laughnan, J. R., 1952 The action of allelic forms of the gene *A* in maize: IV. On the compound nature of A^b and the occurrence and action of its A^d derivatives. Genetics **37:** 375–395.

Laughnan, J. R., 1955 Structural and functional aspects of the A^b complexes in maize: I. Evidence for structural and functional variability among complexes of different geographic origin. Proc. Natl. Acad. Sci. USA **41:** 78–84.

Moore, G., K. M. Devos, Z. Wang and M. D. Gale, 1995 Grasses, line up and form a circle. Curr. Biol. **5:** 737–739.

Pelese-Siebenbourg, F., C. Caelles, J. C. Kader, M. Delseny and P. Puigdomenech, 1994 A pair of genes coding for lipid-transfer proteins in *Sorghum vulgare*. Gene **148:** 305–308.

Sanmiguel, P., A. Tikhonov, Y.-K. Jin, N. Motchoulskaia, D. Zakharov *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765–768.

Schwarz-Sommer, Z., N. Shepherd, E. Tacke, A. Gierl, W. Rohde *et al.*, 1987 Influence of transposable elements on the structure and function of the *A1* gene of *Zea mays*. EMBO J. **6:** 287–294.

Strathmann, M., B. A. Hamilton, C. A. Mayeda, M. I. Simon, E. M. Meyerowitz *et al.*, 1991 Transposon-facilitated DNA sequencing. Proc. Natl. Acad. Sci. USA **88:** 1247–1250.

Whitkus, R., J. Doebley and M. Lee, 1992 Comparative genome mapping of sorghum and maize. Genetics **132:** 1119–1130.

Communicating editor: J. A. Birchler