Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 7409–7414, June 1999
Genetics

Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum

Alexander P. Tikhonov, Phillip J. SanMiguel, Yuko Nakajima, Nina M. Gorenstein, Jeffrey L. Bennetzen, and Zoya Avramova*

Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392

Communicated by John D. Axtell, Purdue University, West Lafayette, IN, April 29, 1999 (received for review October 12, 1998)

ABSTRACT Orthologous *adh* regions of the sorghum and maize genomes were sequenced and analyzed. Nine known or candidate genes, including *adh1*, were found in a 225-kilobase (kb) maize sequence. In a 78-kb space of sorghum, the nine homologues of the maize genes were identified in a colinear order, plus five additional genes. The major fraction of DNA in maize, occupying 166 kb (74%), is represented by 22 long terminal repeat (LTR) retrotransposons. About 6% of the sequence belongs to 33 miniature inverted-repeat transposable elements (MITEs), remnants of DNA transposons, 4 simple sequence repeats, and low-copy-number DNAs of unknown origin. In contrast, no LTR retroelements were detected in the orthologous sorghum region. The unconserved sorghum DNA is composed of 20 putative MITEs, transposon-like elements, 5 simple sequence repeats, and low-copy-number DNAs of unknown origin. No MITEs were discovered in the 166 kb of DNA occupied by the maize LTR retrotransposons. In both species, MITEs were found in the space between genes and inside introns, indicating specific insertion and/or retention for these elements. Two adjacent sorghum genes, including one gene missing in maize, had colinear homologues on *Arabidopsis* chromosome IV, suggesting two rearrangements in the sorghum and three in the maize genome in comparison to a four-gene region of *Arabidopsis*. Hence, multiple small rearrangements may be present even in largely colinear genomic regions. These studies revealed a much higher degree of diversity at a microstructural level than predicted by genetic mapping studies for closely related grass species, as well as for comparisons of monocots and dicots.

The grasses belong to a family of monocotyledonous angiosperms that are well differentiated morphologically from the other angiosperm families and have a single (monophyletic) origin. Their genome sizes, however, may vary a great deal between species. Thus, rice has an estimated genome size of 430 megabases, which is ≈11× smaller than barley, 6× smaller than maize, and 2× smaller than sorghum. These large differences in genome sizes, coupled with differences in the degree and the nature of their investigations, have obscured some common features of grass genomic design. Recent studies comparing high-density linkage maps with DNA markers revealed extensive synteny of chromosomal segments between related species (1–5). Valuable as it is, full genome recombinational mapping of DNA markers is not an efficient approach for detecting small rearrangements. Because the available high-resolution maps based on completed nucleotide sequence are largely restricted to individual genes and their proximal neighborhoods, we are left with two obvious questions that cannot be answered at a full-genome level of analysis. These questions are, will the colinearity observed at the 2- to 20-centimorgan level, the sensitivity level of standard recombinational mapping, be preserved or will it break down at a local level (5), and what will the pattern of gene distribution be, relative to the noncoding, nongene-containing portions of the chromosomes? Currently, two scenarios are considered for how genes could be distributed in complex grass genomes (6): a scattered distribution of genes among noncoding regions, leaving large distances between neighboring genes, and a clustered organization of genes, segregating large regions of transcribed DNA from surrounding large blocks of noncoding DNA. Gross analysis of both animal and plant genomes, based on density gradient centrifugation (isochore analysis), on DNA gel blot hybridization and *in situ* hybridization, support the second model: homogeneous, gene-enriched islands, >100 kilobases (kb) in size, segregated amidst a sea of repetitive DNAs (7–9). However, results from sequencing projects and studies at a molecular level have provided data on large stretches of genomic sequence, which may challenge our traditional views of genome organization. Thus, in *Drosophila*, functional genes were found not only in euchromatin, but also in the heterochromatic pericentric regions (reviewed in ref. 10), arguing that overall genome analysis cannot provide an adequate picture of genome microstructure.

Until very recently, the largest uninterrupted sequence data available from a complex grass genome constituted ≈60 kb from barley chromosome 4HL (11). It revealed the presence of three genes and one retrotransposon. However, the lack of data at a similarly detailed level from a syntenous region of a related species inevitably limited the scope of possible conclusions, particularly those concerning its evolution. The power of comparative genome analysis at a microcolinear level for gene identification and for characterization of intergene spaces in maize, sorghum, and rice has already been demonstrated (12–15). Here, we focus on a comparative analysis of orthologous *adh* regions in maize (*Zea mays*) and sorghum (*Sorghum bicolor*). Sorghum is a close relative of maize, with an ≈3.5-fold smaller genome (16). In this study, we demonstrate coexistence, in a 225-kb region of maize chromosome 1, of a mixture of gene clusters and individual genes interspersed with 14- to 70-kb blocks of highly repetitive DNA. In addition, we compared the sequences from the two grass genomes to homologous *Arabidopsis* regions available in the databases. The molecular characterization and comparison of the three plant genomes allowed us to draw a clearer picture of the structure and evolution of these regions.

MATERIALS AND METHODS

Materials. Overlapping yeast artificial chromosomes (YACs) 334B7 and 119E3, carrying a maize insert from the *adh1* region (17), were used. A maize bacterial artificial chromosome (BAC) (86A10) originating from maize line B73 was obtained from Genome Systems (St. Louis). Sorghum BAC 110K5 was obtained from the BAC library previously described (18). *Escherichia coli* strains DH5, DH10B, LE392, KW251 (Promega), and XL1 Blue (Stratagene) were used. Plasmid DNA for cloning and sequenc-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

Abbreviations: LTR, long terminal repeat; YAC, yeast artificial chromosome; BAC, bacterial artificial chromosome; EST, expressed sequence tag; SSR, simple sequence repeat; MITE, miniature inverted-repeat transposable element.

*To whom reprint requests should be addressed. e-mail: zavramov@bilbo.bio.purdue.edu.

ing was prepared according to Del Sal *et al.* (19). λ phage DNA was prepared according to Sambrook *et al.* (20).

DNA Manipulation and Gel Blot Analysis. All standard recombinant DNA procedures were conducted according to enzyme manufacturer's recommendations or as described (19, 20). Reverse Southern hybridizations were performed by digesting BAC or YAC subclones with alternative sets of restriction enzymes to generate overlapping fragments. Gels were blotted onto nylon filters and were blotted and probed with radioactively labeled total genomic DNA of sorghum (21).

Nucleotide Sequencing and Computer Analysis. DNA sequencing was conducted by the *Tn*1000 strategy (22), modified by using a rifampicin resistant derivative of *E. coli* strain DH5 as a recipient and XL1 Blue as a donor strain. Programs used for sequence analysis were the GCG package V.8 and V.9 (23), BLAST (24), AAT at http://genome.cs.mtu.edu/aat/aatdoc.html (25), and REPEATMASKER (http://ftp.genome.washington.edu/RM/RepeatMasker.html).

The nucleotide sequences were intensively analyzed for the presence of direct and/or inverted repeats, retroelements, and/or other repetitive DNA by six different programs: BLAST, COMPARE, FASTA, REPEAT, STEMLOOP, and REPEATMASKER. Direct comparison of maize and sorghum sequences and sorghum and *Arabidopsis* sequences was conducted by using the program COMPARE. Graphical outputs of the comparisons were obtained by the DOTPLOT program. Insertions/deletions were identified by the GAP program, with gap weight 100 and length weight 0.

RESULTS

The Sorghum Region: Composition and Organization

A portion of sorghum BAC 110K5, containing an ortholog of the maize *adh1* gene, was sequenced. A contiguous 78-kb nucleotide sequence was generated with an average redundancy of 5.8 (GenBank accession no. AF124045). The search for genes was carried out by applying four criteria: (*i*) comparison to GenBank databases; (*ii*) comparison to cDNA sequences generated in the lab or expressed sequence tag (EST) databases; (*iii*) a search for exon/intron structure, as defined by the internet-based gene prediction programs and by a comparison (gapped alignment) to ESTs; and (*iv*) homology to sequences in the orthologous maize and *Arabidopsis* regions.

Fourteen gene candidates, including the *adh* gene, were identified. They were assigned numeric names in the order of their appearance on the BAC (Fig. 1). Only one gene, 110K5.3, corresponds to a gene with a known function, the pollen specific *adh* gene, homologous to maize *adh1*. Six gene candidates exhibited varying degrees of homology to known or putative proteins. These genes are 110K5.1, similar to *Mesocricetus auratus* guanine nucleotide-binding protein ($3.1e^{-36}$, U13152); 110K5.6, homologous to *Arabidopsis thaliana* carbohydrate kinase ($3e^{-07}$, 91N13T7); 110K5.7, homologous to the human cyclinH ($2e^{-12}$, P51946); 110K5.8, homologous to the yeast small GTP-binding protein ($4e^{-42}$, 010190); 110K5.9, homologous to the small nuclear ribonucleoprotein Prp4p of *A. thaliana* ($2e^{-26}$, 022212YG62); and 110K5.14, homologous to various plant chloroplast ATPase synthase Δ subunits: tobacco ($9.1e^{-12}$, P32980), pea ($9.2e^{-10}$, 002758), and spinach ($2.5e^{-11}$, 170098).

The remaining eight genes were predicted on the basis of homology to various plant ESTs and cDNAs. Thus, comparison to maize cDNAs (AF124736, AF124737, AF124738, AF124739, AF124740) outlined putative exons for four gene candidates: 110K5.1, 110K5.2, 110K5.4, and 110K5.8. Two adjacent maize ESTs (AA979993 and AA979792) with high homologies ($2.6e^{-109}$ and $1e^{-64}$, respectively), suggested a poly(A) tail region and putative exons for 110K5.12 gene. An ORF of 1,027 bp was predicted. An observation, important for our conclusions later, was the homology found between two maize cDNAs (AF124737, AF124739) and candidate gene 110K5.8 because the respective maize homologue is one of the genes missing from the colinear maize contig. Putative exons for two more gene-candidates (110K5.5 and 110K5.4) were outlined based on homologies found to *A. thaliana* sequences from BAC T10M13 (ATAF001308).

A rice cDNA (DDBJ RICS1659a) is homologous to three putative terminal exons of 110K5.11 gene, designated *u22* in an earlier study (12). Another rice cDNA (28880) displays some homology ($4e^{-20}$) over a 200-bp region of what could be a sorghum gene 110K5.13. The homologous maize region is missing, and it is difficult to recognize other features of a potential gene. This sequence, however, might contain a separate gene

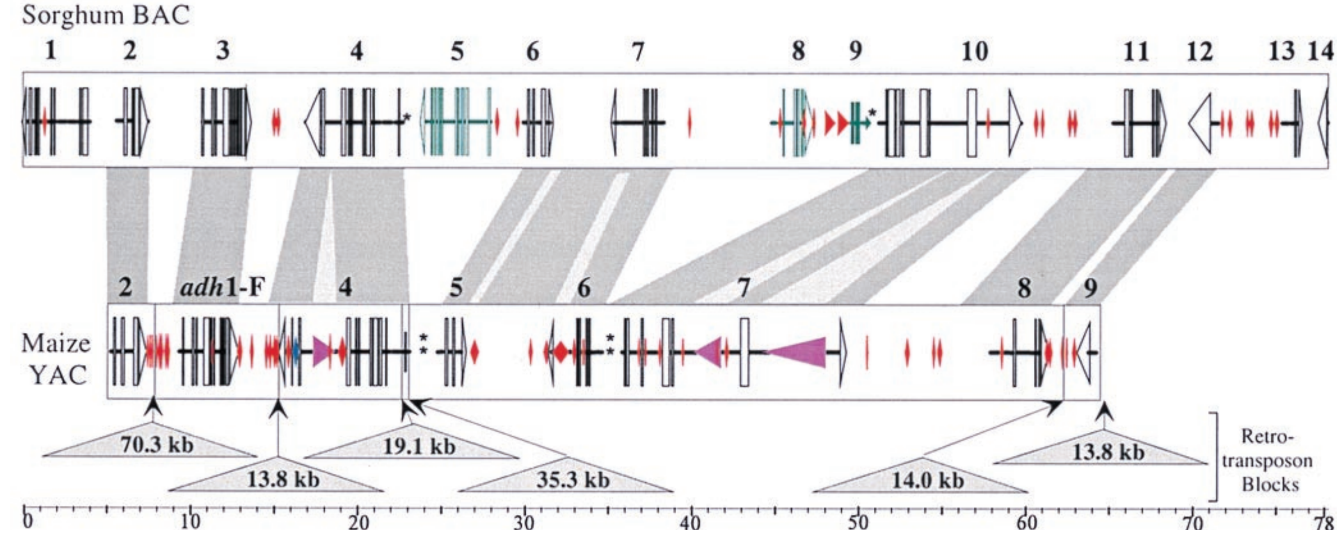

FIG. 1. Schematic representation of the structures of orthologous *adh* regions in maize and sorghum. The upper bar represents sorghum BAC 110K5. Putative genes are numbered by the order of their appearance on the contig, and the putative exons/introns are shown by boxes and connecting lines, respectively. The open triangles mark the end and the direction of transcription for the established cases. The red diamonds show the location and the size of MITEs and other transposon-like elements. The genes shown in blue are genes, lying among orthologous genes but missing from the maize contig. The stars reflect the location of simple repetitive DNAs. The composition of the maize region, derived from YAC 334B7, is shown below. Putative genes, exon/introns, and inserted small elements are marked as above. The purple triangles show the location, the size, and the orientation of transcription of the putative LINEs. The triangles show the location and sites of insertion of the retroelement blocks, and the numbers indicate the block sizes. The shaded regions connecting the maize and the sorghum contigs outline the regions of sequence homology between the two species. The lighter strips within correspond to stretches of interrupted homology, usually associated with insertion of small mobile elements. Both contigs are shown in scale.

because it would have a direction of transcription opposite to both flanking 110K5.12 and 110K5.14 genes.

Analysis of reiteration frequencies of the various classes of DNA on the sorghum *adh* BAC (determined by reverse Southern hybridization) suggested that the region was composed mainly of low-copy-number and some middle repetitive DNAs (data not shown). The nucleotide sequences were analyzed for the presence of direct and/or inverted repeats, retroelements, and other repetitive DNA by six different programs (see *Materials and Methods*).

No obvious long terminal repeat (LTR)-retroelements were revealed. However, numerous transposon-like elements and simple sequence repeats (SSRs) were identified (Fig. 1). Seventeen of the putative transposon-like elements were found in the intergenic space, and three were found in introns. For 12 of them, homologies to the *Tourist* elements found in the *sh2-a1* region of sorghum (15) were revealed. Eleven of them are in the spacer regions between genes, and one is inside an intron of the 110K5.8 gene. One *Tourist* element is found inserted within another *Tourist*, in the spacer between *adh* and the 110K5.4 gene. Two direct moderately repetitive sequences, 516 bp and 634 bp long, displaying >70% homology to each other, were found between 110K5.8 and 110K5.9 genes. SSRs are represented by two $(AT)_{25}$ and $(AT)_{67}$ stretches and by three islands of $(GGA)_{46}$, $(CGG)_{30}$, and $(GAA)_{26}$ repeats. In summary, ≈66 kb (85%) of the 78.2 kb of contiguous sorghum DNA around the pollen *adh* gene represent the gene-containing fraction, containing 14 genes. The remaining 12 kb (15%) belong to miniature inverted-repeat transposable elements (MITEs), SSRs, and DNA lacking an identified origin.

Analysis of the Maize Region

A contiguous series of restriction fragments (contig), covering 258 kb from YAC 334B7, containing the maize *adh1*-F gene (26), has been the focus of our studies. Previous work has established the overall structure of the contig, the location of the only known gene (*adh1*-F), the spatial segregation of the low-copy-number regions from highly reiterated DNA, and the nature and organization of the dispersed repetitive DNAs (26–28). Studies based on a cross-referencing approach, using cross-hybridization with the orthologous sorghum region, identified *adh1* and several other potential genic regions (12). To determine the nature and the microstructure of these possible genic regions, all low-copy-number regions were sequenced with a higher redundancy whereas regions consisting of retroelements of repetitive nature were sequenced with lower redundancy (average redundancy of 4.8). The available sequence data, therefore, represent 217.9 kb of maize chromosome 1, spanning a region of ≈225 kb. The only gaps in the sequence, corresponding to a total of 7.4 kb, were located within highly repetitive retroelements (in the internal part of *Grande*-zm1, *Ji*-1, and *Ji*-6 and inside *Kake*-1) (Fig. 3). The sequences of some retrotransposon LTRs in this region were reported previously (27, 29). All sequences from the maize contig have GenBank accession no. AF123535.

Structure of the Low-Copy-Number Gene-Containing Fraction

Nine putative genes were identified and designated in the order of their appearance on the YAC. Gene 334B7.1, located at the left end of the YAC, has been sequenced only partially and will not be discussed further. 334B7.2 is the first gene at the 5′ end of the contig, located 70.3 kb upstream of *adh1.* It is homologous to the sorghum gene 110K5.2 from the colinear region. A typical gene structure was established for 334B7.2: a putative TATA box and 5 exons, four of which are homologous to cloned cDNA (AF124736). There is a 700-bp ORF in the same orientation as *adh1*, and the respective homologous regions in the two species are of similar length, ≈2.4 kb.

334B7.3. In the numeric system adopted in this study, this corresponds to the *adh1* gene. It is homologous over a region of 3.8 kb to the sorghum region 110K5.3, which contains the *adh1* ortholog.

334B7.4. The closest downstream neighbor to *adh1* is separated by a 13.8-kb block of retrotransposons. Comparison to the maize cDNA (AF124740), to ESTs from various sources (maize, *A. thaliana*, *Caenorhabditis elegans*, human, W49897, N65123, 2104528, 2088768, HSPD06991, respectively), and to the colinear sorghum and *Arabidopsis* (ATAF001308) sequences revealed eight putative exons.

A very small region, 421 bp of low-copy-number DNA, is located 19.1 kb downstream, separated by a block of repetitive DNA. It is interesting to note that, in a previous study, by using cross-hybridization as a tool for gene identification, we were able to locate this small, low-copy-number fragment embedded in a sea of highly repetitive DNA (12). Although the nature and the origin of this sequence was completely obscure at that time, its high level of conservation between maize and sorghum suggested that it might bear some function (12). Here, analysis at a sequence level and comparison to the homologous sorghum 110K5.4 gene helped us recognize it as a region containing the putative first exon of the 334B7.4 gene. In maize, it has been separated from the rest of the gene by the insertion of two unrelated retroelements, *Huck*-2 and *Fourf*, displacing it by ≈19 kb. A third retroelement, *Milt*, has inserted into the 3′ untranslated end of the gene, followed by a more recent insertion of *Opie*-2 into *Milt* (29). As a result, the genomic space occupied by portions of this gene is scattered over 42 kb. The eight predicted exons have uninterrupted ORFs with the opposite orientation from *adh1*, but it remains to be tested whether this gene is expressed. A 1.13-kb LINE was found within a putative intron.

334B7.5, 334B7.6, 334B7.7, and 334B7.8. Proceeding further to the right (Fig. 1), after a 35.3-kb block of highly repetitive DNA, comes a large region, 39.2 kb, of low-copy-number DNA. This is the largest segment in the maize genome, observed by us so far, that lacks a detected LTR-retrotransposon. Analysis of its sequence suggested the presence of at least four genes. Because no cDNAs or ESTs were available to identify putative transcription starts or stop codons for 334B7.5 and 334B7.6, the first two genes in the group, the boundaries of their space were defined by the beginning and by the end of homology to the corresponding sorghum region. As shown by the shaded areas between the two contigs (Fig. 1), the homology stretches for ≈10 kb. Comparison at the predicted amino acid level found homology for 334B7.5 to *A. thaliana* carbohydrate kinase cDNA, as found and described above for the sorghum 110K5.6 gene.

Homology to sorghum (110K5.7 gene) and to the human cyclin H protein ($2e^{-12}$, P51946), determined the position of the putative maize gene 334B7.6. It is interesting to point out that, between the regions occupied by these putative genes, there is ≈5 kb of sequence homology between sorghum and maize, interrupted only by inserted elements. At this stage, we cannot decide what the region between the two putative genes contains, but its conservation indicates that it might have biological relevance. Whether it contains regions belonging to the already identified genes, or a different gene, remains to be established.

Surprisingly, there are only 200 bp of unconserved sequence separating 334B7.6 from 334B7.7, apparent genes, to be transcribed in opposite directions. This fact was unexpected because it provides a first example of two closely positioned but otherwise unrelated genes in maize. In contrast, the respective sorghum homologous genes, 110K5.7 and 110K5.10, are 12 kb away from each other. The nature of this intervening sorghum DNA will be discussed in detail below.

The space occupied by the maize gene-candidate, 334B7.7, spans ≈15.1 kb. Its homology with the sorghum 110K5.10 gene is interrupted only at the sites of inserted elements, two of which are LINEs (related to *Colonist1* and *Colonist2* elements, ZMU90128). A putative TATA box and ATG codon were located. Five putative exons were predicted by BLASTX, and gapped alignment of sequences with homology to the putative gene 11 on *A. thaliana* BAC T5I7 (2642163). Comparative analysis with several cDNAs isolated from a maize seedlings library provided further evidence that this region encompasses a gene (data not shown).

The last of the predicted genes in this 39-kb low-copy-number DNA stretch, 334B7.8, is 6.1 kb downstream of 334B7.7, occupying 4.8 kb of genomic space. Earlier, it was predicted that this region might contain a gene because a homologous rice EST (DDBJ RICS1659a) was found (12). Here, we managed to identify the putative TATA-box, a start codon, a poly(A) signal, four putative exons, and an ORF in the same orientation as *adh*.

334B7.9. About 14 kb downstream of 334B7.8 is the site of the last predicted gene in the sequenced maize region. It is homologous to the sorghum 110K5.12 sequence and to two maize cDNAs (AA979993 and AA979792) with 96 and 93% homology, respectively.

Organization of the Repetitive DNA on the Maize Contig

Earlier, a model for the organization of the repetitive DNA in the maize *adh1* region was proposed based on the order of 37 classes of repeats of the contig and the diagnostic sequencing of putative LTRs (27). Nineteen nested LTR retroelements and two solo LTRs were identified (27). Subsequently, as a result of complete sequencing of most retroelement blocks, regions of ambiguity have been clarified: a retroelement from the *Cinful* family, *Cinful*-2, and an older element, *Tekay*, have been located immediately upstream of *adh1* (29). *Rle*, an even older retroelement, has provided an insertion site for ≈70 kb of retrotransposons, occupying a block at the 5′ end of the contig and 3′ the 334b7.2 gene (Fig. 3).

Overall, the LTR-retroelements comprise 166.4 kb (74%) of the available sequence. The high- and middle-repetitive retroelements are arranged in six blocks of different sizes on the contig: 70.3 kb, 13.8 kb, 19.1 kb, 35.3 kb, 14.0 kb, and 13.9 kb (Fig. 1). The first block of 70.3 kb (with internal gaps in our sequence data of 6.8 kb) covers all of the intergenic space between 334b7.2 and *adh1* genes. The next two blocks are found inside the space occupied by the 334b7.4 gene, and the 35.3-kb block separates the 334b7.4 and 334b7.5 genes. The fifth block of retroelements, containing nested *Reina* and *Cinful*-1 retrotransposons, is inserted between two genes (334B7.8 and 334B7.9), placing them 14 kb apart. The last retroelement block contains nested *Kake*-1 and *Kake*-2 retrotransposons, covering a region of 13.9 kb.

The whole region has been analyzed for the presence and the distribution of mobile DNAs and simple sequence repeats. Because of the overwhelming amount of data and the complexity of such an analysis, at this stage, MITEs were identified by homology to already existing MITE sequences in the databases or by the presence of DNA sequences flanked by defined repeats and host insertion duplications (30, 31). Most of these putative MITEs were found within low-copy-number sequences, as short inserts interrupting the homology between maize and sorghum. Not a single MITE was identified within the space occupied by the retroelements.

A total of 33 DNA transposon-like elements were recognized to this end: For 8, homology to the *Tourist* elements was discovered; for 3, homology to the *Castaway* family was discovered (30). Two *Tourist* elements and a *Tourist*/*Castaway* inserted into each other are located 5′ of the *adh1* gene. A set of three inserted MITEs is found 3′ of the 334B7.8 gene.

Immediately 5′ to the 334B7.2 gene is a 50-bp sequence with 86% identity to the *Sleepy* transposon of maize (ZMU28041), followed by two closely positioned, almost perfect, 44/46 bp, direct repeats. A 230-bp remnant of a maize *Ds* insertion element (X51632) is found between the 334B7.7 and 334B7.8 genes.

Finally, ≈200 bp of simple repeats (CGG and TGG) are found in the space 5′ to the 334B7.5 gene; several (CT)n clusters and a 28-bp stretch of uninterrupted Cs are at the immediate 5′ end of 334B7.7, and a 62 bp-long SSR is found between 334B7.7 and 334B7.8. In summary, the 225.5 kb of the maize sequence around the *adh1* gene is composed of ≈74% LTR-retroelements and 20% genic DNA. The remaining 6% is made up of MITEs, insertion elements, SSRs, and DNA lacking identified origin.

Comparison of the *adh* Regions in Maize and Sorghum

Nature of the Unconserved Sorghum Sequences. Direct comparison of the two colinear regions revealed a patchy pattern of homology (Fig. 2). The diagonal lines, indicating regions of homology, correspond to the gene-coding sequences and their immediate flanking regions whereas the gaps between the diagonal lines represent unconserved sequences. Examining the plot, it is seen that unconserved regions belong to two major categories: missing genic sequences and intergenic spacers.

Three of the fourteen putative genes in sorghum, 110K5.1, 110K5.13, and 110K5.14, are beyond the sequenced regions of the maize contig analyzed here. Three other genes, 110K5.5, 110K5.8, and 110K5.9, were not found in the maize region, although they lie among orthologous genes (Fig. 2). The possibility that the maize homologues were artifactually deleted during the isolation of YAC 334B7 was tested. A second YAC, 119E3, and maize BAC 86A10, covering the controversial area, were hybridized to a probe homologous to the maize 334B7.5 gene, as a positive control, and to two probes recognizing 110K5.5 and 110K5.8 genes from the 4-kb and 12-kb regions missing in maize, respectively. The results indicated that maize homologues of the sorghum genes were not present on either of the new clones (data not shown). These genes, however, are present elsewhere in the genome because they hybridized to the sorghum probes in gel blot hybridizations to total maize genomic DNA (data not shown).

Because the expected loci for both missing genes would have been flanking the maize 334B7.5 and 334B7.6 genes, special attention was devoted to the DNA flanking them in maize and the respective DNA in sorghum. The 1.4 kb between 110K5.4 and 110K5.5 in sorghum displayed no obvious features other than a 20-bp stretch of alternating (A)n and (T)n. In the respective maize region, the DNA between the two genes (over the gap of the missing homologue of 110K5.5) is a 2.1-kb sequence rich in (CGG)n and (TGG)n repeats. A BLAST search found homologies to human fragile DNAs (AF012603 and U48436).

The other two tightly linked sorghum genes, 110K5.8 and 110K5.9, whose homologues are absent from the maize region, are separated from a neighboring (110K5.7) gene by 6 kb of DNA for which no unusual features were recognized. At the other end, in 0.6 kb of spacer DNA separating 110K5.9 from 110K5.10, an

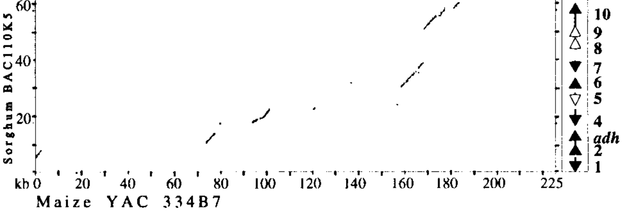

FIG. 2. DOTPLOT homology comparisons for the colinear maize and sorghum regions. The location of the high- and middle-copy-number retroelement blocks is shown by the shaded and dotted boxes, respectively. The open boxes with the arrows show the putative maize genes identified on the contig. On the vertical line, the sorghum BAC is shown. The diagonals, reflecting homologous regions, coincide with the regions taken by genes. The sorghum genes 110K5.5, 110K5.8, and 110K5.9 missing from maize are shown by lighter arrows on the sorghum BAC.

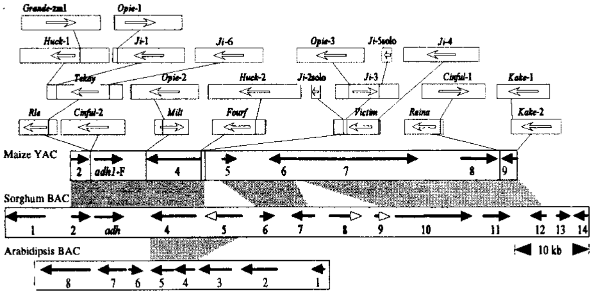

FIG. 3. A simplified schematic comparison between the contigs of maize (YAC 334B7), sorghum (BAC 110K5), and *Arabidopsis* (BAC T10M13). The arrows show the location and the orientation of the putative genes. The shaded regions outline the regions of colinearity between the three species.

SSR with $(CGG)_{10}$ repeats was found. In maize, a short 200-bp spacer separates the neighboring 334B7.6 and 334B7.7 genes. This linking DNA contains two $(CT)_{5–9}$ bracketing a 28-bp stretch of uninterrupted Cs.

In summary, 66.4 kb of sorghum sequence is colinear to 211.7 kb of maize in the orthologous *adh* regions. The overall amount of conserved DNA between the two species is 22% for maize and 57% for sorghum whereas the unconserved fractions represent 78% and 43% of the maize and sorghum DNAs, respectively. The majority (74%) of the maize DNA is composed of LTR-retroelements in the intergenic spaces whereas not a single LTR-retrotransposon was detected in the sorghum region.

Comparisons to the *Arabidopsis* Genome

The missing genes in maize raise a question of whether they have been deleted from the colinear maize chromosome or whether they inserted into the sorghum genome at their present position after the two species diverged, some 15–20 million years ago (32). A possible answer for one of the genes, 110K5.5, is suggested by the finding that two genes, homologous to adjacent sorghum genes, 110K5.4 and 110K5.5, were found next to each other in *Arabidopsis* BAC T10M13 (AF001308) (Fig. 3). Their homology to the respective *Arabidopsis* T10M13.5 and T10M13.4 genes, their tight linkage, and their conserved orientation all suggest that these genes were linked in the ancestor of monocots and dicots. This fact suggests that the maize homologue of 110K5.5 was deleted from the *adh1* region after the separation of maize and sorghum.

All sequence data generated from this study were compared with the sequence information available for *Arabidopsis.* No other colinear regions were identified, although syntenous genes, homologous to the sorghum 110K5.8 and 110K5.10 sequences, were found on *Arabidopsis* chromosome II (AC002510 and AC003000). In the case of *Arabidopsis*, however, the homologous sequences are on two nonoverlapping BACs, separated by at least 100 kb.

DISCUSSION

The importance of small reference genomes and the power of comparative genomics for gene discovery and for evolutionary studies have been demonstrated for bacteria (34, 35), animals (36, 37), and plants (5, 38, 39). The obvious advantages of the smaller and simpler sorghum genome, the fact that differences in genome size do not correlate with morphological and physiological complexity of the organisms (16, 21, 40), as well as the observed similarity and colinearity between the maize and sorghum genomes (1, 12, 41) suggested sorghum as an important model organism for the characterization of the maize genome. Subsequently, the high level of gene homology and synteny found between rice and sorghum (13–15), rice and wheat (38), and rice and barley (42), coupled with the small size of the rice genome, provoked the idea that description of the rice genome alone will indicate the gene content and the key aspects of genome organization for all grasses (42).

The results of this study provide an example of the extent to which such assumptions may be correct. Despite a 3.5-fold difference in total genome size (2,500 megabases for maize and 750 megabases for sorghum) (16), the two grass genomes have a similar set of unique sequences (1, 41) and a general colinearity that, at least in the orthologous genomic regions of the *adh* loci, is well conserved (12).

The major unconserved fraction belongs to the intergenic DNA that, in maize, is represented by 22 nested LTR-retroelements, 33 MITEs, SSRs, and DNA of an unidentified origin. A remarkable distinction between these genomes is that no LTR-retrotransposons was found in the 78-kb continuous sequence in sorghum.

A notable feature of the distribution of MITEs is that they are found within introns or in proximal regions flanking genes, as observed earlier for various grass genes (reviewed in ref. 30). Our data demonstrate that, within 166 kb of DNA composed of LTR-retrotransposons, not a single MITE was found. Hence, MITEs must be unable to insert and/or be retained in these methylated, presumably heterochromatic regions (43). Alternatively, the MITEs may have arrived before the appearance of the retroelements, ≈2–6 million years ago (29). A gene-specific insertion preference of MITEs is similar to that of other maize inverted-repeat transposable elements, like *Mutator* (44).

It was speculated that, after "subtracting" the unconserved fraction, exact colinearity of the gene-containing portions of the two genomes might be observed. However, three genes were found to be deleted from the maize continuum. The true absence of these genes from the maize chromosomal *adh1*-F region was confirmed by testing an additional maize YAC and a maize BAC that cover the investigated area. Homologues of these sorghum genes are present in the maize genome, however, because they hybridized to the sorghum probes in gel hybridizations to maize genomic DNA. In addition, two maize cDNAs were found homologous to the sorghum 110K5.8 gene, supporting a conclusion that the respective maize gene(s) are present, and active, elsewhere in the maize genome.

These results suggest that, even in largely colinear genomic regions, multiple small rearrangements may be present. In maize, such events might be tolerated because of its tetraploid nature (45, 46) and may reflect the "fluidity" of the maize genome caused by the retroelements (47) or other mobile DNAs. Therefore, despite the general principle of similarity and colinearity of the grass genomes, sequence analysis at a microstructural level may reveal significant divergence. For instance, we have found a complete lack of colinearity between the region carrying the *adh1* orthologous locus of rice to the *adh1* region of maize (Y.N., P.J.S., A.P.T., Z.A., H. Zhang, R. Wing, and J.L.B., unpublished work). Although this result provides valuable information about the evolution of the region in parallel with speciation, it also illustrates the risk of making general conclusions based solely on overall genome analysis.

Although a two-gene colinearity was found between sorghum genes 110K5.4–110K5.5 and *Arabidopsis* genes T10M13.5-T10M13.4, the two flanking *Arabidopsis* genes, T10M13.3 and T10M13.6, are not linked to each other in either maize or sorghum. This indicates a minimum of two chromosomal rearrangements in a stretch of four genes. Two other tightly linked sorghum genes, 110K5.8 and 110K5.10, have homologues located on *Arabidopsis* chromosome II. In the case of *Arabidopsis*, however, the genes are not tightly linked because they were discovered on different BACs, >100 kb apart. This result indicates that, although occasional two-gene, or somewhat larger, conserved linkages will be observed between monocots and dicots, long conserved segments may be rare. This is a surprising result. If, as it appears, long sequence blocks approaching whole chromosome or chromosome arms have been conserved in the 50–100 million years since the grasses diverged from each other, then why have there been so many rearrangements in the approximately 200 million years since monocots and dicots diverged? Because intradicot genome colinearity appears to be less than that for the grasses, it is possible that most of these rearrangements occurred early and often in the dicot descendants of the primordial angiosperm. Alternatively, the lineage that has given rise to *Arabidopsis* may have undergone an exceptionally large number of rearrangements in recent evolutionary time.

A major observation of our studies is the pattern of gene arrangement in the related sorghum and maize genomes. Thus, nine genes in a 225-kb region provide an average gene density of about one gene per 25 kb in maize. In sorghum, the density we observe is one gene per 5.6 kb. Assuming that there are ≈30,000 genes in sorghum (genome size of 750 megabase pairs) and ≈50,000 genes in maize (genome size of 2,500 megabase pairs), the expected average gene density would be one gene per 25 kb in sorghum and one gene per 50 kb of maize. The total gene numbers here are crude, of course, but are based on a predicted 25,000 or so genes for *Arabidopsis* (33). Sorghum maps as a diploid with a reasonable amount of segmental duplication (1, 3) whereas maize is an ancient tetraploid (32), suggesting respective gene numbers of 30,000 and 50,000. Therefore, the experimentally determined distances indicate a much denser packing of genes than expected, especially for the sorghum region. In sorghum, genes are evenly distributed along the 78 kb studied. A similar pattern was observed earlier for the genes in the *sh2-a1* region (13–15). This pattern also would fit into a gene-cluster model because these regions might actually represent isolated, gene-enriched segments of sorghum chromosomes. Similarly, the organization of 60 kb of a barley chromosome agreed with a gene-clustering model (11).

Contrary to previous observations, however, that genes in maize are usually separated from regions containing highly repetitive DNAs and rarely mixed within them (7–9), our results illustrate two patterns of gene distribution: individual genes amidst a sea of repetitive DNA, like 334B7.2, *adh1*, 334B7.4, and 334B7.9, separated by sizable blocks of retroelements, and clustered genes, like 334B7.5, 334B7.6, 334B7.7, and 334B7.8, occupying a space uninterrupted by highly repetitive DNAs. If the structural organization of the *adh1* region is a faithful representation of the general pattern of DNA organization in most of the maize genome, as suggested earlier (48), then different species might display different patterns. For large genomes, containing massive amounts of retrotransposon DNA, this study illustrates how important it is to know that large repetitive DNA blocks are not necessarily void of genes and that functional genes may be found interspersed within the repetitive DNAs.

The authors are indebted to Dr. Ben Bowen (Pioneer Hi-Bred) for providing EST Database access and some of the cDNA clones used for establishing relevant gene structures. This work was partially supported by Pioneer Hi-Bred CRA Program, by U.S. Department of Agriculture Grant 98-35300-6167 to Z.A., and by U.S. Department of Agriculture Grants 94-37300-0299 and 97-35300-4594 to J.L.B.

1. Hulbert, S. H., Richter, T. E., Axtell, J. D. & Bennetzen, J. L. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 4251–4255.
2. Ahn, S. & Tanksley, S. D. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7980–7984.
3. Moore, G., Foote, T., Helentjaris, T., Devos, K., Kurata, N. & Gale, M. (1995) *Trends Genet.* **11,** 81–82.
4. Devos, K. M. & Gale, M. D. (1997) *Plant Mol. Biol.* **35,** 3–15.
5. Bennetzen, J. L. & Freeling, M. (1997) *Genome Res.* **7,** 301–306.
6. Moore, G., Gale, M., Kurata, N. & Flavell, R. (1993) *Bio/Technology* **11,** 584–489.
7. Bernardi, G. & Bernardi, G. (1986) *J. Mol. Evol.* **24,** 1–11.
8. Barakat, A., Carels, N. & Bernardi, G. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 6857–6861.
9. Schmidt, T. & Heslop-Harrison, J. S. (1998) *Trends Plant Sci.* **3,** 195–199.
10. Wakimoto, B. T. (1998) *Cell* **93,** 321–324.
11. Panstruga, R., Busches, R., Piffanelli, P. & Schulze-Lefert, P. (1998) *Nucleic Acids Res.* **26,** 1056–1062.
12. Avramova, Z., Tikhonov, A., SanMiguel, P., Jin, Y. K., Liu, C., Woo, S. S., Wing, R. A. & Bennetzen, J. L. (1996) *Plant J.* **10,** 1163–1168.
13. Chen, M., SanMiguel, P., de Oliveira, A. C., Woo, S. S., Zhang, H., Wing, R. A. & Bennetzen, J. L. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 3431–3435.
14. Chen, M. S. & Bennetzen, J. L. (1996) *Plant Mol. Biol.* **32,** 999–1001.
15. Chen, M. S., SanMiguel, P. & Bennetzen, J. L. (1998) *Genetics* **148,** 435–443.
16. Laurie, D. A. & Bennett, M. D. (1985) *Heredity* **55,** 307–313.
17. Edwards, K. J., Thompson, H., Edwards, D., de Saizieu, A., Sparks, C., Thompson, J. A., Greenland, A. J., Eyers, M. & Schuch, W. (1992) *Plant Mol. Biol.* **19,** 299–308.
18. Woo, S. S., Jiang, J., Gill, B. S., Paterson, A. H. & Wing, R. A. (1994) *Nucleic Acids Res.* **22,** 4922–4931.
19. Del Sal, G., Manfioletti, G. & Scneider, C. (1988) *Nucleic Acids Res.* **16,** 9878.
20. Sambrook, J., Maniatis, T. & Fritsch, E. F. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY).
21. Bennetzen, J. L., Schrick, K., Springer, P. S., Brown, W. E. & SanMiguel, P. (1994) *Genome* **37,** 565–576.
22. Strathmann, M., Hamilton, B. A., Mayeda, C. A., Simon, M. I., Meyerowitz, E. M. & Palazzolo, M. J. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 1247–1250.
23. Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12,** 387–395.
24. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
25. Huang, X., Adams, M. D., Zhou, H. & Kerlavage, A. R. (1997) *Genomics* **46,** 37–45.
26. Springer, P. S., Edwards, K. J. & Bennetzen, J. L. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 863–867.
27. SanMiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z. & Bennetzen, J. L. (1996) *Science* **274,** 765–768.
28. Avramova, Z., SanMiguel, P., Georgieva, E. & Bennetzen, J. L. (1995) *Plant Cell* **7,** 1667–1680.
29. SanMiguel, P., Gaut, B. S, Tikhonov, A., Nakajima, Y. & Bennetzen, J. L., (1998) *Nat. Genet.* **20,** 43–45.
30. White, S. E., Habera, L. F. & Wessler, S. R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 11792–11796.
31. Song, W.-Y., Pi, L.-Y., Bureau, T. E. & Ronald, P. C. (1998) *Mol. Gen. Genet.* **258,** 449–456.
32. Gaut, B. S. & Doebley, J. F. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 6809–6814.
33. Bevan, M., Bancroft, I., Bent, E., Love, K., Goodman, H., Dean, C., Bergkamp, R., Dirkse, W., Van Staveren, M., Stiekema, W., *et al.* (1998) *Nature (London)* **391,** 485–488.
34. Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W. S., Borodovsky, M., Rudd, K. E. & Koonin, E. V. (1996) *Curr. Biol.* **6,** 279–291.
35. Koonin, E. V., Mushegian, A. R., Galperin, M. Y. & Walker, D. R. (1997) *Mol. Microbiol.* **25,** 619–637.
36. Elgar, G., Sandford, R., Aparicio, S., Macrae, A., Venkatesh, B. & Brenner, S. (1996) *Trends Genet.* **12,** 145–150.
37. Koop, B. F. & Hood, L. (1994) *Nat. Genet.* **7,** 48–53.
38. Dunford, R. P., Kurata, N., Laurie, D. A., Money, T. A., Minobe, Y. & Moore, G. (1995) *Nucleic Acids Res.* **23,** 2724–2728.
39. Guimaraes, C. T., Sills, G. R. & Sobral, B. W. S. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 14261–14266.
40. SanMiguel, P. & Bennetzen, J. L. (1998) *Ann. Bot. (London)* **82,** 37–44.
41. Berhan, A. M., Hulbert, S. H., Butler, L. G. & Bennetzen, J. L. (1993) *Theor. Appl. Genet.* **86,** 598–604.
42. Kilian, A., Kudrna, D. A., Kleinhofs, A., Yano, M., Kurata, N., Steffenson, B. & Sasaki, T. (1995) *Nucleic Acids Res.* **14,** 2729–2733.
43. Bennetzen, J. L., Schrick, K. M., Springer, P. S., Brown, W. E. & SanMiguel, P. (1994) *Genome* **37,** 565–576.
44. Cresse, A., Hulbert, S., Brown, W., Lucas, J. & Bennetzen, J. B (1995) *Genetics* **140,** 315–324.
45. Soltis, D. E. & Soltis, P. S. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 8089–8091.
46. Song, K., Lu, P., Tang, K. & Osborn, T. C. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 7719–7723.
47. Voytas, D. F. (1996) *Science* **274,** 257–261.
48. Edwards, K. J., Veuskens, J., Rawles, H., Daly, A. & Bennetzen, J. L (1996) *Genome* **39,** 811–817.