Pericentromeric Regions of Soybean (Glycine max L. Merr.) Chromosomes Consist of Retroelements and Tandemly Repeated DNA and Are Structurally and Evolutionarily Labile

Jer-Young Lin,* Barbara Hass Jacobus,* Phillip SanMiguel,† Jason G. Walling,* Yinan Yuan,* Randy C. Shoemaker,† Nevin D. Young‡ and Scott A. Jackson*§

*Department of Agronomy, Purdue University, West Lafayette, Indiana 47907, †Purdue University Genomics Core, Department of Horticulture, Purdue University, West Lafayette, Indiana 47907, ‡USDA-ARS-CICGR and Department of Agronomy, Iowa State University, Ames, Iowa 50011 and §Department of Plant Pathology, University of Minnesota, Saint Paul, Minnesota 55108

Manuscript received February 3, 2005
Accepted for publication April 1, 2005

ABSTRACT

Little is known about the physical makeup of heterochromatin in the soybean (Glycine max L. Merr.) genome. Using DNA sequencing and molecular cytogenetics, an initial analysis of the repetitive fraction of the soybean genome is presented. BAC 076J21, derived from linkage group L, has sequences conserved in the pericentromeric heterochromatin of all 20 chromosomes. FISH analysis of this BAC and three subclones on pachytene chromosomes revealed relatively strict partitioning of the heterochromatic and euchromatic regions. Sequence analysis showed that this BAC consists primarily of repetitive sequences such as a 102-bp tandem repeat with sequence identity to a previously characterized ~120-bp repeat (STR120). Fragments of Calypso-like retroelements, a recently inserted SIRE1 element, and a SIRE1 solo LTR were present within this BAC. Some of these elements and are not conserved outside of G. max and G. soja, a close relative of soybean, except for STR102, which hybridized to a restriction fragment from G. latifolia. These data present a picture of the repetitive fraction of the soybean genome that is highly concentrated in the pericentromeric regions, consisting of rapidly evolving tandem repeats with interspersed retroelements.

Our knowledge of the structural makeup of the soybean (Glycine max L. Merr.) genome is superficial. The genome size of soybean is 1100 Mb (Arumuganathan and Earle 1991), the chromosome number 2n = 40, and the repetitive fraction, based on Cot analyses, ranges between 40 and 60% (Goldberg 1978; Gurley et al. 1979). Despite more than a decade of genomics, we still know little about the DNA composition of the repetitive fraction, the distribution of genes relative to repeats, the molecular structure of the heterochromatic/euchromatic regions, and how duplicated regions of the genome have evolved structurally.

It has long been suspected that the soybean genome has undergone multiple rounds of duplication as evidenced by the number of RFLP fragments in mapping experiments (Shoemaker et al. 1996), sequence analysis of expressed sequence tags (ESTs) (Schlueter et al. 2004), and the construction of a bacterial artificial chromosome (BAC)-based physical map (Wu et al. 2004). Analysis of ESTs has shown that there have been at least two rounds of duplication, ~15 and 44 MYA (Schlueter et al. 2004). Despite the relatively large genome size, little has been published about the repetitive fraction of the soybean genome. A few tandem repeats (SB92, Vaheidian et al. 1995; STR120, Morgante et al. 1997) and retroelements (Laten and Morris 1993; Graham et al. 2002) have been described, but the sequence composition and chromosomal distribution for much of the repetitive DNA that accounts for 40–60% of the soybean genome remains unknown.

In another legume, Medicago truncatula, cytological evidence has shown that chromosome arms are almost exclusively euchromatic and that the majority of the heterochromatin (repetitive sequences) is found in pericentromeric regions (Kulikova et al. 2001). This same study indicated that genes in M. truncatula are overwhelmingly localized to the euchromatic arms. In many cereals such as maize, wheat, and barley, repetitive sequences are dispersed throughout the chromosomes and there is little evidence of demarcated euchromatic and heterochromatic regions (Mroczek and Dawe 2003). Previous cytogenetic analysis of soybean pachytene chromosomes has shown that ~36% of the physical length is heterochromatic and that most of this is pericentromeric or localized to a few highly heterochromatic short arms (Singh and Hymowitz 1988).

We are attempting to determine the molecular orga-
ization of the soybean genome by defining the types of sequences that are repetitive within the soybean genome and concurrently determining the chromosomal location of these sequences. We began by using fluorescence in situ hybridization (FISH) to map genetically anchored BACs to molecular linkage groups (PAGEL et al. 2004). With a series of anchored BACs spanning MLG I, we further sought to determine the molecular organization of this chromosome by integrating BACs and regions of the linkage map to the chromosomal map. We hypothesized that the centromeric regions would be composed of a series of highly repetitive DNA sequences, including tandem repeats, interspersed with retroelements, as observed in other plant species (reviewed in JIANG et al. 2003), and that the heterochromatic regions would be delimited from euchromatin, reflecting the organization of chromatin in M. truncatula (KULIKOVA et al. 2001).

**MATERIALS AND METHODS**

**FISH and fiber-FISH:** Mitotic chromosomes were prepared by acetoarmine squashes using the meristematic portion of hydroxyquinoline-treated root tips. Pachytene chromosomes were prepared using squashes of anthers from flowers fixed in 3:1 ethanol to gallic acid acetic. Slides were screened using a phase-contrast microscope and kept at −80°C until used for FISH. Nuclei were isolated for fiber-FISH following established protocols (ZHONG et al. 1996) except that a 22-μm filtration was added.

For FISH, ~1 μg of plasmid (or BAC) DNA was labeled with either biotin-UTP or digoxigenin-UTP (Hoffman-La Roche) using nick translation. Chromosomes were hybridized following previously published protocols (JIANG et al. 1996b) except that detection was with AlexaFluor 488 streptavidin (Molecular Probes, Eugene, OR) for biotin and mouse antidigoxigenin (Hoffman-La Roche) followed by AlexaFluor 568 anti-mouse (Molecular Probes) for digoxigenin-labeled probes. Digital mapping followed previously published protocols (JACKSON et al. 1999). Grayscale digital images were captured using either an Olympus BX60 with an Hamamatsu Orca ER CCD camera controlled with MetaMorph (Universal Imaging, West Chester, PA) or a Nikon E400 with an Optronics MagnaFire CCD camera controlled by ImagePro (Media Cybernetics). Images were further analyzed using MetaMorph and final publication images were prepared using Adobe Photoshop v7.0 for Macintosh.

**DNA isolation and Southern analysis:** G. max cv. Resnik 2000 seeds were provided by Niels Nielsen [U. S. Department of Agriculture (USDA)-Agricultural Research Service (ARS), Purdue University]. Seeds for all other Glycine species were provided by the USDA Soybean Germplasm Collection, USDA-ARS, University of Illinois, Champaign-Urbana (Table 1).

Plant genomic DNA was extracted from young, frozen leaf tissue using a standard CTAB extraction protocol. BAC DNA was extracted using a QIAGEN (Chatsworth, CA) large-construct kit, with the following modification to the kit protocol during the final precipitation step (QIAGEN protocol step 17), 4 μl of n-ethylglycine were added to the isopropanol to aid in DNA precipitation.

Two micrograms of DNA from BACs 009M21 and 076J21 (Iowa State University, Marek and Shoemaker 1997) and 123007 (University of Minnesota, DANES et al. 1998) were digested in a 37°C water bath overnight with 30 units HindIII restriction enzyme (New England Biolabs, Beverly, MA) and separated on an 1% agarose gel. For plant species, 1 μg of plant genomic DNA from each species was restriction digested with 6 units HindIII, 5 units HpaII, or 6 units MspI (New England Biolabs) in a 37°C water bath overnight and separated on a 0.8% agarose gel. DNA from the gels was blotted onto Zeta-PROBE GT genomic tested blotting membrane (Bio-Rad, Hercules, CA).

The membrane was prehybridized for at least 30 min in Church hybridization buffer (1% BSA/1 mM, EDTA/10% SDS/0.5 M sodium phosphate) at 58°C. Probes were prepared using the Rediprime II random prime labeling system (Amersham Biosciences). Before the probes were used for hybridization, they were purified using the QIAquick nucleotide removal kit (QIAGEN). The probe was hybridized to the membrane at 58°C overnight. After hybridization, the membrane was washed in 1.5× SSC/0.1% SDS for 30 min at 58°C, then in 1× SSC/0.1% SDS for 30 min. The membrane was exposed to autoradiography film overnight at −80°C. Alternatively, the membrane was exposed overnight to a Fuji BAS-MS imaging plate and digitally scanned using a Fuji FLA-5000 Bio Imaging Analyzer.

**BAC DNA sequencing and analysis:** BAC 076J21 was sheared and shotgun cloned as previously described (SANMIGUEL et al. 2002) (bankit659574, GenBank no. AY748457). Sequence coverage of 11× was generated by sequencing 1152 of these clones with both T3 and T7 primers. These sequence reads were assembled using the Phred/Phrap script of the phred/phrap/consed package (Ewing et al. 1998; GORDON et al. 1998). Further analysis was done using the Genetics Computer Group (GCG) package (Accelrys), the Artemis viewer (RUTHERFORD et al. 2000), Dotter (SONNHAMMER and DURRN 1995), and the Tandem Repeat Finder (BENSON 1999). Sequence comparisons were made to the February 15, 2004 v. 140.0 release of GenBank. For EST comparison, the NCBI Blast server was used with default settings and the top 100 are reported. Soybean genome shotgun sequences (GSS) were gen-

---

**TABLE 1**

<table>
<thead>
<tr>
<th>Species</th>
<th>PI no. or cultivar</th>
<th>Lane</th>
</tr>
</thead>
<tbody>
<tr>
<td>G. tabacina</td>
<td>509494</td>
<td>1</td>
</tr>
<tr>
<td>G. falcata</td>
<td>612231</td>
<td>2</td>
</tr>
<tr>
<td>G. angyrea</td>
<td>595792</td>
<td>3</td>
</tr>
<tr>
<td>G. pindanica</td>
<td>595818</td>
<td>4</td>
</tr>
<tr>
<td>G. clandestina</td>
<td>440961</td>
<td>5</td>
</tr>
<tr>
<td>G. canesens</td>
<td>440933</td>
<td>6</td>
</tr>
<tr>
<td>G. latrobeana</td>
<td>505184</td>
<td>7</td>
</tr>
<tr>
<td>G. stenophita</td>
<td>546981</td>
<td>8</td>
</tr>
<tr>
<td>G. curvata</td>
<td>505164</td>
<td>9</td>
</tr>
<tr>
<td>G. tomentella</td>
<td>441006</td>
<td>10</td>
</tr>
<tr>
<td>G. cyrtaloba</td>
<td>373993</td>
<td>11</td>
</tr>
<tr>
<td>G. rubiginosa</td>
<td>591588</td>
<td>12</td>
</tr>
<tr>
<td>G. pescadrensis</td>
<td>505195</td>
<td>13</td>
</tr>
<tr>
<td>G. latifolia</td>
<td>321393</td>
<td>14</td>
</tr>
<tr>
<td>G. arenaria</td>
<td>505204</td>
<td>15</td>
</tr>
<tr>
<td>G. soja</td>
<td>507457</td>
<td>16</td>
</tr>
<tr>
<td>Vigna radiata</td>
<td>cv. nm92</td>
<td>17</td>
</tr>
<tr>
<td>G. max</td>
<td>cv. Resnik 2000</td>
<td>18</td>
</tr>
</tbody>
</table>
gerated from 4-kb randomly sheared genomic fragments cloned into ToPo (Invitrogen, San Diego) and sequenced using T3/ T7 primers (GenBank CL867099–CL868434).

RESULTS

Identification of two BAC clones derived from centromeric heterochromatin: Two soybean BAC libraries were screened with SSR and RFLP markers to derive a set of genetically anchored BAC clones (Marek et al. 2001). These anchored BAC clones were subsequently used for FISH to integrate the genetic and chromosome-based maps (Pagel et al. 2004). Two BAC clones from MLG L, 076J21 (position 34.6 cM) and 09M21 (position 32.4 cM), both selected with SSR markers, were found to map to pericentromeric regions of all 40 soybean chromosomes (Figure 1a). It was not clear at the resolution of mitotic chromosomes if these BACs were derived from either centromeric or pericentromeric regions or if they were found at interstitial heterochromatic locations (Figure 1a, inset); therefore, further molecular characterization was undertaken.

BACs 09M21 and 076J21 have sequences in common but are not entirely redundant: It was not clear if these two BACs differed in DNA content as they both colocalized to entire centromeric regions. Three approaches were undertaken to test whether these two BACs had DNA sequences in common. First, FISH of BACs 076J21 (red) and 09M21 (green) on extended genomic fibers (fiber-FISH) of soybean showed that the signals from the BACs were not entirely coincident (Figure 1b). BAC 076J21 had long stretches of hybridization signals that did not overlap with any hybridization signal from BAC 09M21 (Figure 1b, inset). Second, both BACs were digested with HinDIII, gel blotted, and reciprocally hybridized with the other BAC. These reciprocal Southern analyses showed that although some restriction fragments did hybridize to the other BAC, others did not (Figure 2). Third, we performed ~1.2× draft coverage sequencing of BAC 09M21, which has a ~40-kb insert (data not shown), and compared it to an 11× coverage sequence of 076J21. This revealed that ~29% of the sequences from 09M21 had strong matches (E < e−32) to 076J21 (sequence results below) and that most of the matches were in the SIRE1 element and calypso-like retroelements; none of the sequences from 09M21 had significant matches to either STR102 or STR120 (described below). Even though these BACs hybridized to the same chromosomal region, DNA sequencing, fiber-FISH, and Southern analyses show that they are not entirely redundant in sequence composition.

Organization of DNA sequences within BAC 076J21: To more fully understand the DNA sequence composition, we sequenced BAC 076J21 to 11× coverage using a shotgun approach. Due to the highly repetitive nature of this BAC, assembly of the DNA sequence reads was difficult. However, it was possible to collapse 10 sequence contigs into one 79,623-bp scaffold (Figure 3a), and the order and orientation of the contigs with respect to one another was inferred by forward/reverse sequence reads from clones putatively spanning gaps between the contigs. The full set of contigs comprised 104,573 bases with phred/phrap-generated quality scores >20, ~32 kb smaller than the PFGE size estimate of 136 kb of the BAC clone (data not shown). A caveat to this assembly is that sequence scaffolds of BACs comprised primarily of repetitive DNA elements are often difficult to assemble and may contain errors.

FISH mapping directly on the BAC plasmid (digital mapping, Jackson et al. 1999) was employed to assess the overall accuracy of the assembly. Three STR102 clusters (102-bp tandem repeat) are apparent in 076J21 (Figure 3b), but only two are present in the final sequence scaffold. Quantitative analysis of the digital mapping data revealed that the STR102 repeat accounts for 25.5% (SD 4.1). This would correspond to 35 kb of sequence but only 19 kb are present in sequence contigs. Further, only 193 of 1152 shotgun subclones yielded sequence containing STR102 repeats.

Sequence analysis of these contigs revealed a number of features (Figure 3a). One was a SIRE1 element (Latin and Morris 1993) inserted into a tract of STR102 repeats. The 5’ LTR of the SIRE1 element contained a sequence gap as it spanned two contigs within the scaffold. A SIRE1 solo LTR with 99.8% sequence identity to the 5’ LTR of the full element was also found inserted directly into a block of tandem repeats (Figure 3a). The SIRE1 element and the SIRE1 solo LTR were found in identical positions in two STR102 repeats. Several regions with homology to Calypso-like retroelements (Wright and Voytas 2002) were found scattered across the BAC (Figure 3a).

The 102-bp tandem repeat (named soybean tandem repeat 102, STR102) had 82.6% sequence identity (determined using BestFit of GCG) to the previously described 120-bp tandem repeat STR120 (Figure 3c, gi1147200) (Morgante et al. 1997). Two STR102 representatives were aligned to STR120 using ClustalW (Figure 3c). This analysis showed that the ~102-bp monomers had regions of sequence identity with STR120 with two gaps (15 and 7 nt) in the alignment. Other tandem repeats (ranging from 5 to 191 bp with a minimum copy number of 5) were found within this BAC using the Tandem Repeat Finder program (Benson 1999), none of which were as frequent as STR102 but some of which can be seen as smaller blocks in the Dotter generated dot plot (Figure 3d) (Sonnhammer and Durbin 1995).

Dot-plot analysis of 076J21 showed that except for the SIRE1 element much of this BAC is duplicated internally (Figure 3d). For instance, the first ~14 kb, before the first STR102 cluster, is duplicated several times from 48 kb to the end of the BAC. The LTRs of the SIRE1
Fluorescence in situ hybridization analysis of BACs and subclones to chromosomes and extended DNA fibers of soybean. (a) FISH of BAC 076J21 (green) to mitotic chromosomes (red). (Inset) Three chromosomes from another preparation showing, from left to right: DAPI stained (black and white), 076J21 (green), 09M21 (red), and merged. (b) Fiber-FISH of BACs 076J21 (red) and 09M21 (green) to DNA fibers of soybean showing little overlap in FISH signal. (Inset) An image of a single fiber with the two color channels shown separately. (c–f) FISH to pachytene chromosomes of soybean with subclones of BAC 076J21. Arrows indicate heterochromatic regions and arrowheads indicate centromeric heterochromatin. (c) DAPI-stained chromosomes. (d) STR102 [2_P01]. (e) SIRE1 [1_L22] and Calypso 5-1 [1_E15] pooled. (f) Merged image. (g) Fiber-FISH analysis of STR102 [2_P01] on extended DNA fibers of soybean showing long interrupted arrays of STR102 [2_P01]. Line with arrows indicates a 435.6-kb cluster of repeats.

The assembled contigs for BAC 076J21 were used to query 1454 soybean GSSs derived from paired reads of clones with ~3-kb inserts. These GSS sequences represent 1.29 Mb or ~0.1% of the 1110 Mb soybean genome. RepeatMasker/Crossmatch (http://www.repeatmasker.org/) was used with STR102 and SIRE1 to estimate the frequency of each in the GSS data set as 0.7 and 0.6%, respectively. This is also shown in Figure 3a where the frequency of BLASTN hits of 076J21 to the GSS data set is plotted along the length of the BAC. Eight paired GSS sequences were almost entirely copies of STR102, indicating that these four clones may be composed primarily of STR102.

The 076J21 sequence was used to query the entire GenBank EST collection using an e-value cutoff of 4.0 × 10⁻⁸. A single soybean EST (gi22524207) had up to 98% sequence identity to the STR102 repeat. The SIRE1 element had sequence similarity to ESTs in the LTRs and one EST showed similarity to parts of the internal regions. Of the top 100 matches, 50 were derived from a soybean root hair subtracted cDNA library (gmrhRww).
Neither of the retroelements had long fiber-FISH signals indicative of tandem repeats (data not shown); rather, the fiber-FISH signals were dispersed. On pachytene chromosomes subclones containing SIRE1 (1.L22) and Calypso 5-1 (E15) were pooled for FISH and were found to localize to pericentromeric heterochromatin and heterochromatic knobs on euchromatic arms (Figure 1, e and f).

### Conservation and methylation status of sequences derived from 076J21:

Centromeric sequences from several other plant species have been isolated previously and, in the case of the cereals, a centromere-specific retrotransposon is conserved in both sequence and chromosomal locations across the cereal family (Aragon-Alcaide et al. 1996; Jiang et al. 1996a; Presting et al. 1998). Using a hybridization-based assay, we tested the conservation of both BACs (09M21 and 0076J21) and the three 076J21 subclones across a set of evolutionarily related legume species. When the two BACs were used as probes, they were conserved only in *G. max* and the closely related and sexually compatible *G. soja*, both of which are annuals (Figure 4, a and b). However, when the STR102 repeat (subclone 2_P01) was used as a probe on the blots containing the Glycine annuals and perennials, it hybridized to a fragment in *G. latifolia* (Figure 4c). The methylation status of the three subclones was tested using the methyl-cytosine-sensitive isoschizomer restriction enzymes *Msp*I and *Hpa*II. Both enzymes cut asymmetrically at 5’-CCGG-3’; however, when this sequence is CpG methylated, *Hpa*II will not cut, whereas *Msp*I will. In the case of Calypso 4-1 or 5-1, these sequences were methylated in both *G. max* and *G. soja* (Figure 4d). The STR102 repeat does not have a CCGG restriction site, but surrounding sequences do appear to be methylated on the basis of the hybridization of the SR102-containing clone 2_P01 to the blot (Figure 4d). The SIRE1 element (1.L22) did not appear to cut with either enzyme so its methylation status could not be determined using this approach.

### DISCUSSION

Repetitive sequences can account for a major portion of eukaryotic genomes. Although often referred to as “junk” DNA, repetitive sequences are known to function in the organization of telomeres (Blackburn and Hall 1978) and centromeres (reviewed in Jiang et al. 2003) and may be involved in chromosome packaging, thereby regulating gene expression (Stam et al. 2002). Ribosomal clusters are another example of tandemly repeated but functional DNA. In soybean, estimates of the repetitive fraction range from 40 to 60% on the basis of DNA:DNA renaturation experiments (Goldberg 1978; Gurley et al. 1979). In maize, retroelements often insert within other retroelements, leading to “nested transposons” (SanMigue 1996) that separate “islands” of genic or low-copy sequences; in Arabidopsis, the majority of...
the repetitive sequences are localized to centromeric and pericentromeric regions of the genome (Arabidopsis Genome Initiative 2000). The genome of *M. truncatula*, a legume, has demarcated euchromatic/heterochromatic regions as shown by FISH mapping (Kulikova et al. 2001).

Two BACs anchored to MLG L were found to hybridize to the pericentromeric regions of all 20 pairs of soybean chromosomes. This indicated that sequences within these BACs are (1) repetitive within the soybean genome and (2) conserved in the chromosomal location. To follow up on this observation and to further characterize the physical makeup of the soybean genome, these BACs were molecularly analyzed using DNA sequencing, Southern analysis, and FISH to pachytene chromosomes and extended DNA fibers. These analyses allowed us to determine the distribution, DNA sequence composition, conservation, and methylation status of sequences in this BAC.

FISH of these BACs to pachytene chromosomes more
TABLE 2
Annotation of FGENESH prediction coding sequences on BAC 76J21

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coordinates</th>
<th>No. of hits to soybean ESTs ($E &lt; e^{-10}$)</th>
<th>TBLASTN*</th>
<th>BLASTP UniProt ($E &lt; 0.05$)</th>
<th>Annotation $^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>279..677</td>
<td>4</td>
<td>—</td>
<td>—</td>
<td>Unknown</td>
</tr>
<tr>
<td>2</td>
<td>1856..3554</td>
<td>13</td>
<td>Calypso 4-1</td>
<td>Gag/pol polyprotein</td>
<td>Retroelement</td>
</tr>
<tr>
<td>3</td>
<td>6982..8974</td>
<td>10</td>
<td>—</td>
<td>—</td>
<td>Unknown</td>
</tr>
<tr>
<td>4</td>
<td>29856..3585</td>
<td>15</td>
<td>SIRE1</td>
<td>Gag-pol polyprotein</td>
<td>LTR retrotransposon</td>
</tr>
<tr>
<td>5</td>
<td>34791..50620</td>
<td>2</td>
<td>SIRE1</td>
<td>Envelope-like protein</td>
<td>LTR retrotransposon</td>
</tr>
<tr>
<td>6</td>
<td>41539..42707</td>
<td>3</td>
<td>STR120</td>
<td>—</td>
<td>Tandem</td>
</tr>
<tr>
<td>7</td>
<td>47926..4794</td>
<td>9</td>
<td>SCBI</td>
<td>Hypothetical protein</td>
<td>LTR retrotransposon</td>
</tr>
<tr>
<td>8</td>
<td>51974..53860</td>
<td>14</td>
<td>SCBI</td>
<td>Hypothetical protein</td>
<td>LTR retrotransposon</td>
</tr>
<tr>
<td>9</td>
<td>54862..55980</td>
<td>4</td>
<td>Calypso 4-1</td>
<td>Gag/pol polyprotein</td>
<td>Retroelement</td>
</tr>
<tr>
<td>10</td>
<td>57411..57809</td>
<td>6</td>
<td>—</td>
<td>—</td>
<td>Unknown</td>
</tr>
<tr>
<td>11</td>
<td>63592..64086</td>
<td>19</td>
<td>AOX</td>
<td>—</td>
<td>Hypothetical</td>
</tr>
<tr>
<td>12</td>
<td>66348..66722</td>
<td>2</td>
<td>MDH1</td>
<td>—</td>
<td>Retroelement</td>
</tr>
<tr>
<td>13</td>
<td>70800..71918</td>
<td>4</td>
<td>Calypso 4-1</td>
<td>Gag/pol polyprotein</td>
<td>Retroelement</td>
</tr>
<tr>
<td>14</td>
<td>73861..74259</td>
<td>6</td>
<td>—</td>
<td>—</td>
<td>Unknown</td>
</tr>
<tr>
<td>15</td>
<td>78107..78931</td>
<td>3</td>
<td>MDH1</td>
<td>—</td>
<td>Retroelement</td>
</tr>
</tbody>
</table>

* Top hits are shown to the GenBank nucleotide database.

$^b$ “Unknown” indicates EST matches with no GenBank or UniProt hits; “hypothetical” indicates EST hit and GenBank hit.

finely determined the chromosome distribution of sequences from BAC 076J21. It was evident from DAPI staining that many of the chromosomal arms of soybean are euchromatic, confirming previous observations (Singh and Hymowitz 1988). This indicates that, for some chromosomes, the majority of the heterochromatin is likely to be confined to the pericentromeric regions. FISH analysis showed that BAC 076J21 and several of the subclones from this BAC localized to either side of the primary constriction (centromere). However, it is possible that there are homologous sequences within the centromeres that, due to chromosomal packaging, are unavailable as hybridization targets. A similar phenomenon, where centromeric sequences did not hybridize to FISH probes on meiotic chromosome preparations, was seen in potato (J. Jiang, personal communication).

The distribution of repeats on either side of a centromere was not even and sequences were occasionally found in heterochromatic regions outside of pericentromeric heterochromatin. In Arabidopsis, heterochromatic knobs containing pericentromeric sequences have been found physically disassociated from the centromeric regions, such as that seen on the short arm of chromosome 4 (Fransz et al. 2000). Although soybean pachytene chromosomes appear to be generally euchromatic, knob-like regions of heterochromatin are found in the euchromatic arms and some of the sequences in BAC 076J21 hybridize to these regions.

Sequence analysis of BAC 076J21 showed the presence of a 102-bp tandem repeat (STR102), fragments of Calypso-like elements, a SIRE1 element, and a SIRE1 solo LTR. Tandem repeats are a common motif of higher eukaryotic centromeric/pericentromeric regions. The 180-bp PAL1 repeat of Arabidopsis (Martinez-Zapater et al. 1986), the 155-bp CentO repeat of rice (Cheng et al. 2002), the 137-bp pSau3A9 repeat of sorghum, and the 156-bp CentC of maize (Ananiev et al. 1998) are examples of centromere-specific tandem repeats (also reviewed in Houben and Schubert 2003). Given the commonality of the tandem repeat feature at centromeric regions, it is thought that these approximately nucleosomal-length repeats may play a role in organizing centromere-specific nucleosomes (Nagaki et al. 2003; Black et al. 2004).

Very few tandem repeats have been reported for G. max apart from STR120 (Morgante et al. 1997); SB92, a 92-bp tandem repeat (Vaahedian et al. 1995); and now STR102, a ~102-bp repeat with 82.6% sequence similarity to STR120. FISH analysis of the STR102-containing clone 2_P01 showed that this repeat is almost exclusively seen in potato (J. Jiang, personal communication).

The STR102 repeat was detected in clusters up to 435.6 kb in length, although longer arrays may exist. A similar organization of centromeric tandem repeats has been reported for rice (Cheng et al. 2002) and maize (Ananiev et al. 1998). The STR102 sequence was conserved in G. latifolia outside of G. max and G. soja; however, only one restriction fragment showed weak hybridization in G. latifolia, so it is possible that there was only limited sequence identity between STR102 and
sequence(s) in *G. latifolia*. This is not unexpected since tandem repeats, such as the α-satellite of primates, have been found to evolve very rapidly (Waye and Willard 1989).

Retroelements are another common feature of eukaryotic genomes (reviewed in Bennetzen 1996). In maize, retroelements often insert within other retroelements leading to nested transposons (SanMiguel et al. 1996). A few retroelements have been described from the soybean genome, the most unusual of which is the SIRE1 element (Laten and Morris 1993), which falls into the *copia/Ty1* family (Laten et al. 1998). Laten et al. (2003) found that the flanking sequences of 3 of 10 SIRE1 insertions were repetitive, belonging to either Ty3-gypsy or other repetitive families. In BAC 076J21, the SIRE1 element was inserted into a long array of STR102 repeats and BAC 09M21 also had sequences with significant sequence alignments to the SIRE1 element. This indicates that the SIRE1 element may preferentially insert into heterochromatic and/or pericentromeric regions; alternatively, insertions into gene-rich euchromatic regions may be selected against. This is further corroborated by our FISH results showing hybridization to obvious heterochromatic regions.

A SIRE1 solo LTR was also found in a tract of STR102 repeats with the same 5-bp insertion duplication as the SIRE1 element. There are two possible explanations for this. Two SIRE1 elements inserted into same tract of STR102 repeats and BAC 09M21 also had sequences with significant sequence alignments to the SIRE1 element. This indicates that the SIRE1 element may preferentially insert into heterochromatic and/or pericentromeric regions; alternatively, insertions into gene-rich euchromatic regions may be selected against. This is further corroborated by our FISH results showing hybridization to obvious heterochromatic regions.

A SIRE1 solo LTR was also found in a tract of STR102 repeats with the same 5-bp insertion duplication as the SIRE1 element. There are two possible explanations for this. Two SIRE1 elements inserted into same tract of STR102 repeats and BAC 09M21 also had sequences with significant sequence alignments to the SIRE1 element. This indicates that the SIRE1 element may preferentially insert into heterochromatic and/or pericentromeric regions; alternatively, insertions into gene-rich euchromatic regions may be selected against. This is further corroborated by our FISH results showing hybridization to obvious heterochromatic regions.

A SIRE1 solo LTR was also found in a tract of STR102 repeats with the same 5-bp insertion duplication as the SIRE1 element. There are two possible explanations for this. Two SIRE1 elements inserted into same tract of STR102 repeats and BAC 09M21 also had sequences with significant sequence alignments to the SIRE1 element. This indicates that the SIRE1 element may preferentially insert into heterochromatic and/or pericentromeric regions; alternatively, insertions into gene-rich euchromatic regions may be selected against. This is further corroborated by our FISH results showing hybridization to obvious heterochromatic regions.

A SIRE1 solo LTR was also found in a tract of STR102 repeats with the same 5-bp insertion duplication as the SIRE1 element. There are two possible explanations for this. Two SIRE1 elements inserted into same tract of STR102 repeats and BAC 09M21 also had sequences with significant sequence alignments to the SIRE1 element. This indicates that the SIRE1 element may preferentially insert into heterochromatic and/or pericentromeric regions; alternatively, insertions into gene-rich euchromatic regions may be selected against. This is further corroborated by our FISH results showing hybridization to obvious heterochromatic regions.
recombination was found to be the primary causative mechanism (Devos et al. 2002). Pairing and recombination among LTRs within an element would leave one LTR while removing the rest of the retrotransposon. The 5′ LTR from the SIRE1 is 99.8% identical over 1130 bases to the solo LTR, indicating that this is likely a recent insertion within the last 30,000 years (H. Laten, personal communication) and further confirming a previous suggestion that this element may still be active within the G. max genome (Laten et al. 2003).

The Calypso-like elements that are dispersed across 076J21 were fragments of either Calypso 4-1 or Calypso 5-1. The Calypso 5-1 fragments had homology to the Pol domain and the Calypso 4-1 fragments had homology to the first few kilobases of the elements that are not soybean chromosomes. One of the major sequence copies of 076J21 have homology to the Pol MLG L that are partially redundant and contain sequences. How these sequences function to organize DNA into functional chromosome units is still a mystery, but the organization suggests that sequencing of the euchromatic regions of the soybean genome may be a tractable approach to recover many of the genes of soybean.

We found two BAC clones, 09M21 and 076J21, from MLG L that are partially redundant and contain sequences found in the pericentromeric regions of all soybean chromosomes. One of the major sequence constituents of 076J21, tandem repeat STR102, was found at all or most major heterochromatic blocks in soybean. The repeat sequences present in 076J21 are not generally found outside of the two annuals G. max and G. soja, indicating that these are fast-evolving sequences.

We thank the Purdue University Agricultural Experiment Station and the United Soybean Board for generous support of this research, Jingjing Jiang (University of Wisconsin-Madison) for advice and comments, and Laura Marek (Iowa State University) for the initial screening of the BAC library. We thank Jeff Doyle (Cornell University) for thoughtful advice in choosing various legume species for analysis of sequence conservation.

LITERATURE CITED


Ewing, B., L. Hillier, M. Wendl and P. Green, 1998 Basecalling of...


MARTINEZ, M., I. JURMAN, L. SHI, T. ZHU, P. KIM et al., 1997 The *STR120* satellite DNA of soybean: organization, evolution and chromosome specificity. Chromosoma Res. 5: 363–373.


Communicating editor: A. H. PATTERSON