

Organization of Repetitive Elements in the Upland Cotton Genome (*Gossypium hirsutum*)

R. J. Baker, J. L. Longmire, and R. A. Van Den Bussche

To better understand the evolutionary nature of the economically important upland cotton (*Gossypium hirsutum*), 1,719 randomly chosen recombinants from a cosmid genomic library were examined for repetitive elements. Average insert size was approximately 34 kb, which represents 2.6% to 4.4% of the haploid genome of *G. hirsutum* (depending on the estimate for haploid genome size used). Probes representative of microsatellites, tandem repeats, transposable elements, a low copy number gene (catalase), and a chloroplast gene were used. Of the four dinucleotides, (CT)_n and (GT)_n were present in 3.6% and 1% of the clones, respectively, whereas, (AT)_n hybridized to a single clone and (GC)_n, as well as the chloroplast gene, did not hybridize to any clone. The tandemly repeated rDNA cistron hybridized to 60 of the 1,719 clones. We estimate the *cop*ia-like family of retrotransposable element to be present in thousands of copies. When data from all probes are analyzed, the upland cotton genome consists of approximately 61% unique sequences and low copy number DNA, which agrees well with previously published estimates that were determined using other methods.

Upland cotton, *G. hirsutum*, is an allotetraploid ("AD-genome"; $2n-4x = 52$), cultivated species which resulted from the hybridization of a diploid African or Asian ("A-genome"; $2n = 26$) species with an American ("D-genome") species (Beasley 1940, 1942; Endrizzi et al. 1985; Skovsted 1934; Wendel 1989) and is one of the most important agricultural crops worldwide. Of the 26 currently recognized linkage groups in cotton, only 65 genes have been reported as mapped to 17 linkage groups and only 12 of these linkage groups have been associated with their respective chromosome (Endrizzi et al. 1984).

Few publications have examined the relative distribution and abundance of repetitive DNA in this economically important crop (approximately \$5.5 billion in the United States in 1990; Price et al. 1990). Walbot and Dure (1976) have divided the cotton genome into three major kinetic components (unique, middle repetitive, and highly repetitive sequences) with an interspersion of the middle repetitive sequences with the single-copy sequences constituting a large portion of the cotton genome; based on these observations, they presented a model for the number of individual sequences in each component.

Their model allows us to test their predictions concerning the organization of the cotton genome, as well as to examine the efficiency of cosmid libraries for identifying various aspects of the cotton genome. The strength of different approaches for evaluating copy number and genome organization is that when estimates are corroborating, the confidence in these estimates becomes more robust.

To test the predictions of Walbot and Dure (1976), as well as provide base line data on the nature of the cotton genome, we have isolated the most abundant repetitive elements in the upland cotton (*G. hirsutum*) genome and estimated their relative copy number and frequency of co-occurrence. In addition, we have examined the abundance of representative microsatellites, a retrotransposon, and the low copy number catalase gene.

Materials and Methods

Construction of Cosmid Genomic Library

We isolated high molecular weight genomic DNA from *G. hirsutum* leaf tissue, using a procedure modified from Hillis et al. (1990). Our procedure consisted of plac-

From the Department of Biological Sciences (Baker and Van Den Bussche) and The Museum (Baker), Texas Tech University, Lubbock, TX 79409, and the Life Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico (Longmire). We thank John Gannaway of the Texas A&M University Research Station for assistance in obtaining cotton stocks and in theoretical discussions concerning upland cotton. We thank Jonathan Wendel for supplying the *cop*ia-like clones and Michael Arnold for supplying the sunflower ribosomal probe and PCR primers Z1 and ORF106 for the chloroplast *rbc*l gene. Thanks are extended to Susan Carron, Lisa Gregory, Shelly Witte, Andy Simmons, Nancy Brown, and Laura Janecek for technical assistance. Randy Allen, Ray Jackson, Virginia Walbot, J. Gannaway, and J. Wendel critically reviewed drafts of this manuscript. This work was supported by Texas Tech University.

Journal of Heredity 1995;86:178-185; 0022-1503/95/\$5.00

ing 5–6 g freshly ground leaf tissue (ground to a fine powder in liquid nitrogen) in 10–15 ml of prewarmed (70°C) extraction buffer (5% CTAB [hexadecyltrimethylammonium bromide], 3.5 M NaCl; 0.25 M Tris-Cl, pH 8.0; 0.05 M EDTA; 0.04% β -mercaptoethanol) in a 50-ml corex tube. This solution is then incubated in a 60°C water bath for 1–1.5 h. After incubation, an equal volume of chloroform:isoamyl alcohol (Cl) is added and mixed gently on a tube rotator for 10 min. Debris is then pelleted by centrifugation for 10 min at room temperature. The aqueous phase is transferred to a fresh tube and extracted with Cl at room temperature for 2–3 min. The solution is then centrifuged again to further purify the DNA. The aqueous phase is transferred to a clean tube, and the DNA is precipitated by adding 2 vol. of absolute ETOH and inverting end over end for 3–4 min. The DNA is spooled out with a glass pipette and placed in a 1.5-ml eppendorf tube. The sample is allowed to air-dry and then is resuspended in 200–300 μ l of 1XTE (10 mM Tris-HCl, 1 mM EDTA; Sambrook et al. 1989). This method produces high molecular weight DNA that is of sufficient purity and quality to allow restriction endonuclease digestion and cloning.

After partial digestion with *Sau* 3A1 and dephosphorylation with calf intestinal alkaline phosphatase, approximately 0.5 μ g of cotton genomic DNA was ligated with 1.0 μ g of *Bam* HI cloning arms from the cosmid vector sCos-1 (Evans et al. 1989), and in vitro packaging was carried out in Giga Pack Gold extracts (Stratagene). Primary infection of *E. coli* host strain DH5 α MCR yielded 4.5×10^5 independent recombinants, giving an approximately 12-fold statistical representation of *G. hirsutum* genomic sequences. We determined average size of inserts in the cosmid vector by digesting 20 randomly chosen clones with *Eco* RI; the resulting fragments were separated in a 0.8% agarose gel. We determined fragment sizes from digitized images of photographs of the ethidium-stained gel using the computer programs Adobe Photoshop 2.0 and NCSA Gel Reader 2.02 for the Macintosh.

Characterization of Repetitive Elements in the Genomic Library

A total of 1,728 independent clones from the primary library were picked, grown, and archived in 96-well microtiter plates. We used a replica plater (Sigma Chemical Co.) to inoculate nylon membranes (MSI Magna NT 0.45 micron, Biotyne B 0.45 micron) with clones from the microtiter

plates. Membranes were incubated at 37°C for 7 h on LB agar containing kanamycin (30 μ g/ml) and then transferred to LB agar containing kanamycin and chloramphenicol (170 μ g/ml; Sambrook et al. 1989) and grown overnight at 37°C. We fixed DNA onto membranes by sequentially placing them on blotting pads soaked in 0.4 M NaOH (5 min), 0.5 M Tris, and 1.5 M NaCl, pH 7.5 (5 min), and $2 \times$ SSC (5 min), followed by baking at 80°C for 1–2 h.

We hybridized membranes with a variety of radiolabeled probes to determine the distribution and relative abundance of these sequences in the cotton genome. Prior to hybridization, membranes were washed in $0.1 \times$ SSC, and 0.1% SDS for 1 h at 65°C in a shaker bath. Prehybridization was carried out at 65°C for 1 h in $6 \times$ SSC, 40% formamide (Kodak), 1% SDS, 0.005 M EDTA (pH 8.0), and 0.005 g/ml Carnation evaporated milk. Membranes were hybridized overnight at 42°C in fresh prehybridization solution containing approximately 1×10^6 cpm/ml of radioactively labeled probe. We labeled all probes with [α - 32 P]dCTP or [α - 32 P]dATP by nick translation. Unincorporated radioactivity was separated from the labeled probe by spin column chromatography (Sambrook et al. 1989). Prior to hybridization, probes were denatured for 10 min at 37°C in 0.1 M NaOH. Following hybridization, membranes were washed once for 15 min in $2 \times$ SSC, 0.1% SDS at room temperature and twice for 15 min each in $0.1 \times$ SSC, 0.1% SDS at 50°C. We autoradiographed washed membranes at -80°C using Kodak XAR-5 film and two Lightning Plus intensifying screens. For each probe, we scored each clone on a scale of 0, representing no detectable hybridization, to 3, a completely black spot on an autoradiograph representing maximum detectable hybridization.

Source of Probes

We purchased approximately 1.2 kb dinucleotide probes (GT) $_n \times$ (AC) $_n$, (CT) $_n \times$ (GA) $_n$, (AT) $_n \times$ (TA) $_n$, and (GC) $_n \times$ (CG) $_n$ from Pharmacia LKB. Hereafter, these four microsatellites are referred to as (GT) $_n$, (CT) $_n$, (AT) $_n$, and (GC) $_n$. The sunflower ribosomal gene complex (pHAR-1; supplied by Mike Arnold, University of Georgia), cotton catalase gene (Ni et al. 1990), and a portion of the *copia*-like reverse transcriptase gene (VanderWiel et al. 1993) were supplied as cloned fragments. The *copia*-like reverse transcriptase clones (A108, D104) represent two of the approximately nine families of *copia*-like retro-

transposable elements detected in the cotton genome (VanderWiel et al. 1993). We isolated all cloned fragments from their vectors by digestion with appropriate enzymes and electrophoresis in 0.8% low-melting point agarose gels. After electrophoresis, the insert was cut from the gel and purified using Prep-A-Gene (BioRad Laboratories). The purified insert was gel purified a second time to ensure that all vector DNA was removed from the sample.

Contamination of the Genomic Library With Organellar DNA

A potential source of error in our estimates for the number of unique, middle repetitive, and highly repetitive DNA of upland cotton would be if our library was contaminated with either chloroplast or mitochondrial DNA. To evaluate whether this was a source of error, we amplified an approximately 2.5-kb section of the chloroplast genome from upland cotton using primers for the ribulose-1,5-bisphosphate carboxylase gene (Z1; McIntosh et al. 1980 and ORF106 Hiratsuka et al. 1989). Conditions for amplification were as described by Arnold et al. (1991) except that 25- μ l reactions were performed. Approximately 500 ng of this PCR product were transferred to nylon membrane by slot blot transfer to be used as a positive control for hybridization experiments. We labeled 1 μ g of the 2.5-kb chloroplast fragment with α - 32 P-dCTP by a random priming reaction (Feinberg and Vogelstein 1984) and used it to probe the cotton cosmid library following the methods outlined earlier.

Characterization of Repetitive Families of Elements

We sorted families of repetitive DNA by isolating cosmid DNA after a standard miniplasmid DNA isolation procedure; they were then digested with *Eco* RI and electrophoresed in 0.8% agarose gels. DNA was transferred to nylon hybridization membrane following the method of Southern (1975) and hybridized with representative clones following the procedure outlined above for screening the genomic library.

Strengths and Limitations of the Methods

It is generally assumed that a library constructed from genomic DNA provides an accurate representation of the organism's genome from which the DNA was originally isolated. Two of the possible sources that can prevent a library from actually being representative of the organism's ge-

Table 1. Representation of nine probes in 1,719 independent clones from a cosmid library constructed from *Gossypium hirsutum* genomic DNA

Probe	No. of clones ^a		Copy no.	Inter-spersion frequency (kb)
	Negative	Positive		
(AT) _n	1,718 (100.00)	1 (0.06)	29 ± 56	
(GC) _n	1,719 (100.00)	0 (0.00)		
(CT) _n	1,657 (96.39)	62 (3.61)	1,768 ± 432	1/922
(GT) _n	1,702 (99.01)	17 (1.00)	485 ± 228	1/3,370
rDNA	1,659 (96.51)	60 (3.49)	1,714 ± 428	
<i>Copia</i> -like (A108)	1,657 (96.39)	62 (3.61)		1/839 ^b
<i>Copia</i> -like (D104)	1,666 (96.92)	53 (3.08)	1,943 ± 453	
Catalase	1,718 (99.94)	1 (0.06)	29 ± 56	
<i>G. hirsutum</i> genomic DNA	1,168 (67.95)	551 (32.05)		

^a Percentages are shown in parentheses; some values do not total 100 due to rounding error.

^b This apparent interspersed frequency is based on the estimated copy number of the A108 and D104 subfamilies of *Copia*-like elements.

nome are a nonrandom distribution of restriction sites for the enzyme used to digest the genomic DNA and methylation of nucleotides within the recognition site of the chosen restriction endonuclease. Whereas it is not possible to eliminate these potential sources of error in the construction of the genomic library, for the following reasons, the construction of our upland cotton cosmid library should be minimally affected by these factors, if affected at all. First, to eliminate the nonrandom distribution of restriction endonuclease recognition sites, we performed a partial digestion of the genomic DNA with *Sau* 3AI, a restriction endonuclease that has a 4-bp recognition sequence. Second, to construct the library, we chose *Sau* 3AI because it is not sensitive to methylation (Stratogene, La Jolla, California). Finally, in the construction of this cosmid library we used technology identical to that used in the construction of the human chromosome specific libraries for the Human Genome Project which have been shown to be representative of the DNA comprising each chromosome (Longmire et al. 1993).

A second cautionary note must be made in consideration of our estimates of copy number of repetitive elements. First, estimates of copy number of repeated elements are always problematic because error can come from several sources. These sources of error for direct examination of copy number experiments can be attributed to DNA loading, stringency of hybridization and washes, and the efficiency of transfer of DNA in Southern blot experiments. Moreover, estimations of copy number of repetitive elements based on C₀t analysis are subject to assumptions concerning the specific model employed.

In this study we calculated copy number by assuming that each clone con-

tained a single copy of the particular element used as a probe and then extrapolated how many copies would be present in a complete genome based on the amount of the genome the archived clones are estimated to represent. Our scoring of the hybridization based on relative intensity can bias the estimate of copy number of repetitive DNA. Highly intense hybridization (scores of 3) can be due to either a high degree of DNA similarity between the probe DNA and the insert DNA, or it can be attributed to multiple copies of a repetitive family within the cosmid. Likewise, those cosmids that gave less intense hybridization (scores of 2 and 1) can be attributed to either a short piece of a large repetitive element or divergent copies of a repetitive family (such as members of transposable element families and imperfect microsatellite repeats). The approach of screening large-insert cosmids to provide an estimate of copy number has the potential for providing an underestimate of the total number in that any recombinant cosmid could potentially contain more than a single copy of a repetitive sequence. Moreover, when genomic DNA is used as a probe, different independent clones will hybridize with varying intensity. This difference in intensity could be due to either relative copy number of the repetitive element or the relative amount of diversity seen in that family of repetitive DNA. Therefore, copy number estimates, as well as relative abundance of different classes of repetitive DNA from this approach, must be considered as minimal (in the case of copy number) or with some caution (for abundance of different classes of repetitive DNA). The advantage of using large-insert recombinants (such as cosmids), however, is that this method provides a means

to detect nonrandom associations of repetitive elements within the genome (Janecek et al. 1993).

Results

Of the 1,728 archived cosmid clones screened in this study, 1,047 did not hybridize to any of the nine probes examined. To evaluate how many of these clones did not contain cotton genomic DNA, cosmid DNA was prepared from 120 arbitrarily chosen clones of the possible 1,047 and digested with the restriction endonuclease *Eco* RI. Absence of detectable bands in one of these clones suggested that this colony did not contain a recombinant cosmid or failed to grow to a density that allowed detection of the recombinant cosmid DNA using a standard miniprep procedure. All the remaining 119 cosmid clones contained insert DNA. Thus, of the 1,047 clones that did not hybridize to any probe, we extrapolate that approximately nine of these clones did not contain inserts. This brings the total number of recombinant cosmids screened to 1,719.

The size of partially digested *G. hirsutum* DNA inserted into 20 randomly selected cosmid clones ranged from 24,400 to 45,300 bp, with a mean of 33,500 bp. Based on this mean insert size, the 1,719 cosmid clones represent 0.575×10^6 bp. Considerable variation exists among estimates of the size of the haploid cotton genome. Bennett et al. (1982) estimated that the haploid genome of upland cotton contained 3 pg DNA/cell whereas Arumunganathan and Earle (1991) estimated approximately 5.08 pg DNA/haploid cell. Given these estimates, the 1,719 archived recombinant cosmids represent 2.6% (Arumunganathan and Earle 1991) to 4.4% (Bennett et al. 1982) of the upland cotton genome (assuming haploid genome sizes of 2.2×10^9 and 1.3×10^9 bp, respectively). Of the 1,719 recombinant cosmids screened in this study, 1,047 clones (60.6%) did not hybridize to any of the nine probes examined. This is because the single (and potentially low copy number) elements are inadequately represented in the sample of DNA labeled as a probe to produce detectable hybridization (Janecek et al. 1993). Table 1 presents the number of clones that hybridized for each of the nine probes examined. Hybridization of the *G. hirsutum* library with *G. hirsutum* genomic DNA resulted in some degree of hybridization to 551 (32%) of the recombinant cosmids. Of these positive clones, 3% were

scored as maximally hybridizing (score of 3) whereas 7% and 22% were assigned scores of 2 and 1, respectively. Of the 1,719 cosmid clones, none hybridized to the cytoplasmic DNA probe from the chloroplast genome. Of the 1,719 cosmid clones examined, 60 (3.5%) hybridized to the sunflower ribosomal probe (rDNA). Of these 60 clones, 58 were scored as maximally hybridizing (score of 3), while one clone was scored as 2 and a single clone was scored as a 1.

The four dinucleotide microsatellites used in this study varied greatly in their representation among the 1,719 clones examined. Dinucleotides (CT)_n and (GT)_n hybridized to 62 (3.6%) and 17 (1.0%) clones, respectively. For the 62 (CT)_n, 40 clones (64.52%) were scored as maximally hybridizing, 12 (19.35%) were assigned a score of 2, and 10 (16.13%) were scored as 1. For the 17 (GT)_n, eight (44.44%) were scored as maximally hybridizing, five (27.78%) were scored as 1, and four were scored as 2. The (AT)_n probe produced detectable hybridization to a single recombinant cosmid, and this hybridization was scored as a 3. No hybridization was detected with the dinucleotide probe (GC)_n. Previous reviewers expressed concern as to whether the probe was efficiently labeled in these experiments. Therefore, hybridization of the cotton cosmid library was performed along with slot blots containing 500 ng of the (GC)_n and (AT)_n microsatellites. In these experiments, the positive controls gave very intense hybridization, but there was no detectable hybridization with the (GC)_n probe.

The two *copia*-like reverse transcriptase clones (VanderWiel et al. 1993) hybridized to 62 (3.6%) and 53 (3.1%) of the clones examined. Of the 62 clones hybridizing to the reverse transcriptase portion of *copia*-like clone A108, 20 (32.3%) were assigned a score of 3, 18 (29.0%) were assigned a score of 2, and 24 (38.7%) were scored as 1. When the reverse transcriptase portion of *copia*-like clone D104 was used as a probe, of the 53 only eight clones were assigned a score of 3 (15.1%), 21 clones were assigned a score of 2 (39.6%), and the remaining 24 clones (43.4%) produced minimal levels of hybridization (scored as 1). Finally, a single clone hybridized with the cotton catalase cDNA probe, and this clone was assigned a score of 3 (maximum intensity).

Co-occurrence between pairs of probes hybridizing on single cosmids was detected at levels expected based on their representation in individual clones (Table 2).

Table 2. Pairwise comparisons of the seven probes detected in 1,719 cosmid clones from the genomic library of *Gossypium hirsutum*

Probe	(CT) _n	(GT) _n	rDNA	<i>Copia</i> -like A108	<i>Copia</i> -like D104	Catalase	<i>G. hirsutum</i> genomic DNA
(CT) _n	—	0.65 (0.20)	2.15 (2.15)	2.22 (1.43)	1.90 (0.43)	0.04 (0.04)	19.77 (12.58)*
(GT) _n	1	—	0.62 (0.62)	0.65 (0.65)	0.55 (0.55)	0.01 (0.01)	5.74 (1.31)
rDNA	0	0	—	2.15 (0.62)	1.84 (1.84)	0.03 (0.03)	19.13 (83.10)*
<i>Copia</i> -like A108	4	0	1	—	1.9 (1,023.58)*	0.04 (0.04)	19.77 (12.58)*
<i>Copia</i> -like D104	1	0	0	46 +	—	0.03 (0.03)	16.90 (16.90)*
Catalase	0	0	0	0	0	—	0.32 (0.32)
<i>G. hirsutum</i> genomic DNA	4	3	59 +	4	0	0	—

Two probes—(GC)_n and (AT)_n—that failed to hybridize to any library clones or a single clone are not included. Above the diagonal: the expected number for each pair of probes based on total occurrence in the 1,719 clones examined. Chi-square values are shown in parentheses, significant chi-square ($P < .05$) values are denoted by an asterisk. Below the diagonal: observed numbers for each pair of probes; plus sign indicates pairs of probes observed more often than expected; minus sign indicates pairs of probes observed less than expected.

However, five pairs of probes were detected to co-occur either significantly greater than or less than expected, based on their individual representation. The three pairs of probes co-occurring significantly less than expected were (CT)_n/*G. hirsutum* genomic DNA, *copia*-like A108/*G. hirsutum* genomic DNA, and *copia*-like D104/*G. hirsutum* genomic DNA; however, rDNA/*G. hirsutum* genomic DNA and the two *copia*-like reverse transcriptase clones co-occurred significantly more frequently than expected.

To determine the number of different repetitive families represented by the recombinant cosmids hybridizing to *G. hirsutum* genomic DNA, we first examined the 41 clones that were assigned a score of 3 (that class giving maximal hybridization) but that did not hybridize to probes for microsatellites, *copia*-like elements, or the catalase gene. Upon digestion with *Eco*RI, all of these clones produced the characteristic banding pattern of the rDNA cistron (bands approximately 3.1 and 2.9 kb) and the approximately 6.7-kb sCos-1 vector. Probing these clones with the sunflower rDNA probe confirmed that all 41 of these clones contained the rDNA cistron. Likewise, when representatives of the remaining 113 clones that were assigned a score of 2 and did not hybridize to probes for microsatellites, *copia*-like elements or rDNA were used as probes, all 113 clones and essentially all bands hybridized. These 113 clones appear to represent a second family of repetitive DNA.

Discussion

Repetitive elements in the genome vary tremendously among species, and the up-

land cotton genome appears to be among those genomes containing the fewest families and relative abundance of tandemly repeated DNAs. Based on C_t analysis, Walbot and Dure (1976) concluded that the cotton genome is comprised of approximately 60.5% unique sequence DNA, 27% middle repetitive sequence DNA, and the remaining approximately 12.5% highly repetitive sequence DNA. They produced a model for the organization of the cotton genome which predicts that there exists 1.5×10^5 unique sequence elements and 1,184 middle repetitive sequences. They did not provide an estimate for the number of highly repetitive and palindromic sequences because their data for these components of the genome were not amenable to such computations.

In this study, 60.9% of the 1,719 clones did not hybridize to any of the probes examined. This group of clones would represent the unique to low copy DNA or extremely divergent members of repetitive families such as *Copia*-like elements, and our estimation of this class of DNA is not different from the estimate of 60.5% by Walbot and Dure (1976). When *G. hirsutum* DNA was radioactively labeled and probed back to the library, 32% of the cosmid clones resulted in some degree of hybridization. This would represent the highly repetitive and middle repetitive classes of DNA. If we assume those clones scored as producing maximal hybridization represent the highly repetitive class of Walbot and Dure (1976), we estimate that this component makes up approximately 3% of the repetitive DNA, whereas Walbot and Dure (1976) estimated this component to

comprise approximately 12.5% of the repetitive DNA. Finally, 7% of the repetitive DNA was scored as a 2 (indicating intermediate levels of hybridization), whereas 22% of the repetitive DNA was assigned a value of 1. If we assume these two categories represent middle repetitive DNA, and even if we add those clones that hybridized to the dinucleotide and *copia*-like probes, then this value for middle repetitive DNA is higher but still not remarkably different from the 27% value calculated by Walbot and Dure (1976) for this class of repetitive DNA. The presence of dinucleotide microsatellites and *copia*-like elements would occupy only a small portion of a 33-kb insert so that their presence would not greatly affect the estimation of middle repetitive DNA present in the cotton genome. Some discrepancies between our estimates for the relative abundance of each class of repetitive DNA and those of Walbot and Dure (1976) may be explained by (1) the subjectivity of assigning the level of hybridization in our study to classes 0, 1, 2, and 3, without quantifying the actual amount of hybridization; (2) the difference in hybridization intensity among clones in our study that may relate not only to copy number but also to the degree of divergence among members of a family of repetitive elements; and (3) the different estimates of the haploid genome size used by Walbot and Dure (0.8 pg/haploid cell) and this study (3 pg to 5.08 pg/haploid cell). However, even with these caveats, there is considerable concordance between these two studies concerning the relative abundance of the three classes of DNA (unique, middle repetitive, and highly repetitive).

The 20 randomly selected recombinant cosmids from the upland cotton library that were analyzed contained inserts ranging from 24,400 bp to 45,300 bp, with a mean insert size of 33,500 bp. Additionally, because the sCos-1 vector will not accept foreign DNA smaller than approximately 25 kb (Evans et al. 1989), all cosmids will contain inserts larger than the size classes examined by Walbot and Dure (1976). Therefore, the use of cosmid libraries allows us to make some general tests of the interspersed patterns predicted by Walbot and Dure (1976). They estimated interspersed patterns for these various classes of DNA and suggested that the genome should be arranged in the following four groups: (1) 25% of the repetitive and 36% of the unique DNA should be arranged as 1,250 nucleotide repetitive element plus 1,800 nucleotide nonrepetitive

element; (2) 5% of the repetitive and 16% of the unique DNA should be arranged as 1,250 nucleotide repetitive element and 4,000 nucleotide nonrepetitive element; (3) 8% of the unique DNA should be arranged as nonrepetitive DNA 6,000 nucleotides from the nearest repetitive DNA; and (4) 8% of the genome should be palindromic and highly repetitive DNA reassociated by $C_{0,t} = 0.1$. Calculations based on the cosmid library disagree with these predictions. When the 113 class-2 repetitive DNA clones are digested with *Eco* RI and probed with radioactively labeled upland cotton genomic DNA, nearly every band from each clone produces a detectable signal. Additionally, 61% of the 1,719 cosmid clones, containing an average insert size of approximately 34 kb, did not hybridize to radioactively labeled genomic DNA. Finally, few of the 1,719 cosmids hybridized to more than a single probe. Those clones that did hybridize to two probes were either cases in which the two *copia*-like reverse transcriptase clones were compared or when the rDNA and *G. hirsutum* genomic DNA were compared (Table 2). As explained below, these cases of co-occurrence greater than expected are probably due to either significant DNA similarity between the two *copia*-like reverse transcriptase probes (VanderWiel et al. 1993) or the high copy number of rDNA cistrons in the cotton genomic DNA and not the presence of multiple families of repetitive DNA within a single recombinant cosmid.

We interpret these data to suggest that from an organizational standpoint, the repetitive DNA appears to be more closely grouped and that stretches of single-copy DNA are larger in the upland cotton genome than hypothesized by Walbot and Dure (1976). If the estimations of Walbot and Dure (1976) that unique DNA was separated from the nearest repetitive DNA by approximately 6 kb had been accurate, we would expect to see very few of the 1,719 cosmid clones not hybridizing to any of the repetitive probes used in this study, given that the average cosmid insert size is 34 kb.

Characterization of Repetitive DNA Families in Upland Cotton

Ribosomal DNA. Walbot and Dure (1976) hybridized 125 I-labeled cytoplasmic rDNA to whole-cell DNA on filters and in solution and concluded that there were approximately 300 to 350 copies of the rRNA cistron per haploid genome and suggested that along with the smaller genome size

(0.8 pg DNA/cell), *G. hirsutum* had a reduced number of rDNA cistrons than reported for other higher plant species. They further suggested that the wide range of multiplicity of the rDNA cistron reported for other higher plant species (1,580 to 27,000) may be due in part to the use of polyploid seed tissue for these determinations.

Probing the *G. hirsutum* library with the sunflower rDNA cistron (pHAR-1) revealed hybridization to 60 independent cosmid clones. Based on the number of clones that hybridized to the rDNA probe, and an examination of 3.5% of the cotton genome (average of the 2.6% to 4.4% estimates of genome size), we estimate there are approximately 1,714 copies of the rDNA gene complex in the cotton genome (Table 1). However, it is probable that this represents an underestimation of the actual number of rDNA genes in the upland cotton genome. The reason for this possible underestimation is that when the 60 clones containing the rDNA genes were digested with *Eco* RI, electrophoresed, and visualized by staining with ethidium bromide, the predicted insert size from each of these clones appeared to be smaller than could be cloned into the sCos-1 vector and still maintain a viable cosmid (Evans et al. 1989). However, if multiple copies of the rDNA cistron were tandemly arranged, this would produce an adequate size for cosmid viability. Moreover, when the recombinant cosmids containing the rDNA cistron were hybridized with genomic DNA, only those bands that also hybridize with the sunflower rDNA probe show detectable hybridization, providing further evidence that these clones only contain rDNA cistrons. We interpret these results to indicate that these cosmids represent multiple copies of the tandemly repeated rDNA gene complex. If this observation is correct, then the actual number of copies of the rDNA gene complex in the upland cotton genome would be significantly higher than our estimate.

Although our estimate for copy number of the rDNA cistron is greater than that predicted by Walbot and Dure (1976), the primary difference between these values is attributable to the different estimates used for the size of the haploid genome. When these values are adjusted for the different genome size estimates, copy number estimates of the rDNA cistron in cotton are essentially identical and indicate that, relative to other plants with similar sized genomes, upland cotton has a reduced number of rDNA sites.

The remaining 113 clones that hybridized to cotton genomic DNA but did not hybridize to any of the other probes used in this study appear to represent a second family of repetitive DNA. Not only do all clones cross hybridize when representatives of these clones were used as a probe, but all also cross hybridize when various *Eco* RI bands were isolated from individual clones and used as probes. These data are interpreted to suggest that this group of 113 clones represents a second family of repetitive DNA which occurs in the upland cotton genome as tandemly repeated units whose total size is at least as great as the size of the insert (24 to 45 kb) in the recombinant cosmid.

Microsatellites. Microsatellites are relatively short (<100-bp) runs of tandemly repeated DNA with repeat lengths of 6 bp or less (Stallings 1992). These microsatellites have been shown to be useful genetic markers because the number of repeats within a specific cluster is often highly variable and because these variations can easily be analyzed using the polymerase chain reaction (Litt 1991; Weber 1990a,b). Characterization of the relative abundance of microsatellites in the cotton genome could provide valuable genetic markers for mapping the cotton genome or for the identification of the numerous available cultivars of upland cotton. The four dinucleotides used in this study represent all possible combinations of two different dinucleotides. The four dinucleotide microsatellites varied greatly in their representation among the 1,719 clones examined. Dinucleotides (CT)_n and (GT)_n each hybridized to 3.6% and 1% of the clones that were screened, (AT)_n hybridized to a single cosmid, but (GC)_n did not hybridize to any of the clones screened. Assuming that we examined approximately 3.5% of the cotton genome (averaging the estimates of genome size), the upland cotton genome contains approximately 1,768 (CT)_n and 485 (GT)_n repetitive sequences (Table 1). Based on the single hybridizing recombinant cosmid to the (AT)_n repeat probe, we would predict that the upland cotton genome contains approximately 29 copies of this repetitive sequence; however, due to the small number of hybridizing clones, considerable error surrounds this estimate (Table 1).

As discussed by Rafalski and Tingey (1993), the only published data on the relative frequency of occurrence of dinucleotide microsatellites in plants is based on a search of the sequences in the major databases. This search revealed that the di-

nucleotide (AT)_n was the most abundant of the sequences examined, which is in contrast to our results. However, our knowledge of the relative abundance of microsatellite repeats in plants is not sufficient to allow for generalizations concerning their organization and distribution in the genome (Rafalski and Tingey 1993).

Whereas data on the abundance of dinucleotide repeats are lacking in plants, data from mammalian systems do exist, to allow a direct comparison between their frequency in cotton and mammalian genomes. Using an identical approach as this study, Janecek et al. (1993) examined the four dinucleotide repeats in a genomic library of the white-footed mouse (*Peromyscus leucopus*) and found that there were approximately 75,000 (GT)_n and 51,000 (CT)_n repetitive sequences, whereas the dinucleotides (AT)_n and (GC)_n each hybridized to a single clone. We have also examined the frequency of these four dinucleotides in the New World bat, *Macrotus waterhousii* (Van Den Bussche RA, Longmire JL, and Baker RJ, in preparation), an organism that relative to other mammals has a reduced genome size (Baker et al. 1992; Burton et al. 1989). We found that even though there is a reduction in the relative number of each of these dinucleotides in *M. waterhousii* compared to *P. leucopus*, the relative order of abundance in *M. waterhousii* is the same as detected by Janecek et al. (1993).

Based on a computer search of published human and rat sequences in the major data bases, Beckmann and Weber (1992) found the relative abundance of these four dinucleotides were the same as detected in *P. leucopus* (Janecek et al. 1993) and *M. waterhousii* (Van Den Bussche et al., in preparation), in that (GT)_n was the most abundant closely followed by (CT)_n, whereas (AT)_n and (GC)_n were infrequently detected. Beckmann and Weber (1992) further extrapolated from their data that there would exist 100,000 total (GT)_n sequences in the human genome, which they suggest agrees well with estimates based on hybridization data (Hamada et al. 1982; Litt and Luty 1989; Stallings et al. 1991; Sun et al. 1984; Tautz and Renz 1984).

An additional method of examining the relative abundance of dinucleotide repeats, which takes into account the genome size, is the apparent interspersed frequency. As with the estimates of relative abundance, considerable variation exists for interspersed frequencies among organisms. Interspersed frequencies of

(GT)_n have been estimated as one (GT)_n repeat every 106 kb in *Macrotus* (Van Den Bussche et al., in preparation), 54 kb in humans (Moyzis et al. 1989), 40 kb in *Peromyscus* (Janecek et al. 1993), 21 kb in *Rattus* (Stallings et al. 1991), and 18 kb in *Mus* (Stallings et al. 1991). For upland cotton, the apparent interspersed frequency of the dinucleotide (GT)_n is one repeat cluster every 3,370 kb.

Although fewer genomes have been examined for the dinucleotide repeat (CT)_n, patterns similar to (GT)_n, albeit slightly reduced, have been documented in mammals. For example, the apparent interspersed frequency of (CT)_n is one repeat sequence every 115 kb and 59 kb in the genome of *Macrotus* (Van Den Bussche et al., in preparation) and *Peromyscus* (Janecek et al. 1993), respectively. Within the upland cotton genome, the apparent interspersed frequency of (CT)_n microsatellites is one repeat cluster every 922 kb.

When the cotton dinucleotide repeat data are interpreted in light of these other studies, it appears that the cotton genome has a greatly reduced number of these microsatellites relative to the few well-studied mammalian cases. The cotton genome is also unique among these studies in that (CT)_n is more common than (GT)_n.

Transposable elements. VanderWiel et al. (1993) performed a phylogenetic analysis of a portion of the reverse transcriptase gene of *copia* -like elements present in the genome of upland cotton and concluded that the upland cotton genome may contain as many as nine families of *copia* -like elements. Although VanderWiel et al. (1993) did not perform formal copy-number experiments, they estimated that *copia* -like retrotransposons are present in the genome of upland cotton in hundreds if not thousands of copies. Based on screening the upland cotton cosmid library with representatives of two of the nine families, we estimate the number of *copia* -like retrotransposons in *G. hirsutum* to be in the thousands. Hybridization of the cosmid library with the *copia* -like reverse transcriptase clone A108 produced detectable hybridization to 62 of the 1,719 clones, while hybridization with a *copia* -like reverse transcriptase probe from another family, D104, resulted in some level of hybridization to 53 of the 1,719 clones. Based on these values we estimate these families to be present in approximately 1,400–2,387 and 1,200–2,308 copies, respectively depending on how much of the cotton genome our 1,719 archived clones represent. These values probably repre-

sent overestimates of the number of *copia*-like retrotransposons in that these two probes hybridize to the same recombinant cosmids with a significantly higher co-occurrence than would be expected based on their individual frequency of occurrence in the library. Additionally, because VanderWiel et al. (1993) have shown significant levels of DNA similarity between these two families, a more conservative estimate for the copy number of *copia*-like retrotransposable elements would be to base it on only those clones that hybridized to either of the two probes and counting those clones that hybridized to both probes only once. In this case there were 68 clones that hybridized to *copia*-like reverse transcriptase. This would indicate that approximately 1,900 copies of *copia*-like retrotransposable elements of the A108 and D104 type exist in the genome (Table 1).

Low copy genes. The upland cotton genome contains two copies of the catalase gene (Percy and Wendel 1990). When a cDNA probe from the catalase gene of cotton was used as a probe to the cosmid library, a single clone hybridized, and this hybridization was scored as maximal. Additionally, the clone containing this catalase gene did not hybridize to any of the other probes in this study. This is interpreted as further evidence that the unique and low copy DNA is separated from repetitive elements by distances greater than that previously estimated (Walbot and Dure 1976).

Representation of Organellar DNA in the Library

Although the *rbcl* chloroplast gene did not hybridize to any of the 1,719 clones, it is still possible that the chloroplast genome is represented in the library because the chloroplast genome is sufficiently large that a cosmid insert could be present without including the region covered by the chloroplast *rbcl* probe. However, if it is assumed that chloroplast DNA would be randomly inserted into cosmids, it is not highly probable that chloroplast DNA is represented in many of the cosmid clones of which the library consists. No tests were performed to test for the presence of mitochondrial DNA in the library. As chloroplast DNA is much more abundant in preparations of cellular DNA than is mitochondrial DNA, we think that it is highly improbable that organellar DNA has significantly biased our calculations of the abundance of repetitive elements in the cotton nuclear genome.

Organization of the Upland Cotton Genome

Walbot and Dure (1976) stated that cotton contains over 1,000 middle repetitive DNA families, each occurring on the average over 100 times per genome. Based on our results, the number of middle repetitive families in cotton would be much smaller than this estimate of 1,000. We conclude that the highly repetitive and middle repetitive families in cotton would consist of at least rDNA, the dinucleotide repeats (CT)_n and (GT)_n, a *copia*-like retrotransposable element family, and, finally, another tandemly repeated family present in 113 of the 1,719 cosmid clones sampled.

In general, estimates for the percentage of the upland cotton genome comprising the unique, middle repetitive, and highly repetitive components are in agreement with estimates based on C_t analysis (Walbot and Dure 1976). Some of the discrepancies that exist between estimates for the middle and highly repetitive classes of DNA, as well as the total number of families, found in the upland cotton genome may be related to the different approaches used in this study and that of Walbot and Dure (1976). Probing cosmid libraries may bias the results toward identifying those repetitive families that are under concerted evolution. For instance, although we estimated the *copia*-like retrotransposon to be present in approximately 1,500 to 2,600 copies, only nine clones that hybridized with *G. hirsutum* genomic DNA contained the *copia*-like family. This resulted in these probes (the two *copia*-like reverse transcriptase probes and *G. hirsutum* genomic DNA) occurring significantly less frequently than expected based on their relative frequency of occurrence (Table 2). It may be that enough variation exists in this family of repetitive DNA that under the stringency conditions we employed few of the members of the other seven families of *copia*-like elements were detected (VanderWiel et al. 1993).

It is not clear how much of a bias is actually introduced into our estimates, but based on hybridization of the cosmid library with *G. hirsutum* genomic DNA, our estimates for the number of unique, middle repetitive, and highly repetitive DNA are in good agreement with that of Walbot and Dure (1976). If our approach failed to detect a significant number of families of repetitive DNA, then these values would also change. Probably the estimates that would be most affected by this bias would concern the interspersed pattern of repetitive elements within the unique se-

quences. Correcting for underestimations of interspersed repetitive elements in the cotton genome would bring our estimates of interspersed patterns closer to those described by Walbot and Dure (1976), but this correction would also result in a greater divergence of our estimates of the relative kinetic groups from that described by Walbot and Dure (1976).

References

- Arnold ML, Buckner CM, and Robinson JJ, 1991. Pollen-mediated introgression and hybrid speciation in Louisiana irises. *Proc Natl Acad Sci USA* 88:1398-1402.
- Arumunganathan K and Earle RD, 1991. Nuclear DNA content of some important plant species. *Plant Mol Biol Rep* 9:208-218.
- Baker RJ, Maltbie M, Owen JG, Hamilton MJ, and Bradley RD, 1992. Reduced number of ribosomal sites in bats: evidence for a mechanism to contain genome size. *J Mamm* 73:847-858.
- Beasley JO, 1940. The origin of American tetraploid *Gossypium* species. *Am Nat* 74:285-286.
- Beasley JO, 1942. Meiotic chromosome behavior in species, species hybrids, haploids and induced polyploids of *Gossypium*. *Genetics* 27:25-54.
- Beckman JS and Weber JL, 1992. Survey of human and rat microsatellites. *Genomics* 12:627-631.
- Bennett MD, Smith JB, and Heslop-Harrison JS, 1982. Nuclear DNA amounts in angiosperms. *Proc R Soc Lond B* 216:179-199.
- Burton DW, Bickham JW, and Genoways HH, 1989. Flow-cytometric analyses of nuclear DNA content in four families of Neotropical bats. *Evolution* 43:756-765.
- Endrizzi JE, Turcotte EL, and Kohel RL, 1984. Qualitative genetics, cytology, and cytogenetics. In: *Cotton* (Kohel RJ and Lewis CF, eds). Madison, Wisconsin: American Society of Agronomy; Adv Agron 24:81-129.
- Endrizzi JE, Turcotte EL, and Kohel RL, 1985. Genetics, cytology, and evolution of *Gossypium*. *Adv Genet* 23: 271-375.
- Evans GA, Lewis K, and Rothberg BE, 1989. High efficiency vectors for cosmid microcloning and genomic analysis. *Gene* 79:9-20.
- Feinberg AP and Vogelstein B, 1984. A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal Biochem* 137:266-267.
- Hamada H, Petrino MG, and Kakunaga T, 1982. A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. *Proc Natl Acad Sci USA* 79:6465-6469.
- Hillis DM, Larson A, Davis SK, and Zimmer EA, 1990. Nucleic acids 3: Sequencing. In: *Molecular systematics* (Hillis DM and Moritz C, eds). Sunderland, Massachusetts: Sinauer; 318-370.
- Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun C-R, Meng B-Y, Li Y-Q, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, and Sugiura M, 1989. The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct trnA genes accounts for a major plastid DNA inversion during the evolution of the cereals. *Mol Gen Genet* 217:185-194.
- Janecek LL, Longmire JL, Wichman HA, and Baker RJ, 1993. Organization of repetitive elements in the genome of the white-footed mouse, *Peromyscus leucopus*. *Mamm Genome* 4:374-381.
- Litt M, 1991. PCR of TG microsatellites. In: *PCR: a practical approach* (McPherson MC, Quirke P, and Taylor, eds). New York: Oxford University Press; 85-99.
- Litt M and Luty JA, 1989. A hypervariable microsatellite

- revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* 4:397-401.
- Longmire JL, Brown NC, Meinke LJ, Campbell ML, Albright KL, Fawcett JJ, Campbell EW, Moyzis RK, Hildebrand CE, Evans GA, and Deaven LL, 1993. Construction and characterization of partial digest libraries made from flow-sorted human chromosome 16. *Genet Anal. Tech and Appl* 10:69-76.
- McIntosh L, Poulsen C. and Bogorad L, 1980. Chloroplast gene sequence for the large subunit of ribulose biphosphatecarboxylase of maize. *Nature* 288:556-560.
- Moyzis RK, Torney DC, Meyne J, Buckingham JM, Wu JR, Burk C, Sirotkin KM, and Goad WB, 1989. The distribution of interspersed repetitive DNA sequences in the human genome. *Genomics* 4:273-289.
- Ni W, Turley RB, and Trelease RN, 1990. Characterization of a cDNA encoding cottonseed catalase. *Biochem Biophys Acta* 1049:219-222.
- Percy RG and Wendel JF, 1990. Allozyme evidence for the origin and diversification of *Gossypium barbadense* L. *Theor Appl Genet* 79:529-542.
- Price HJ, Stelly DM, McKnight TD, Scheuring CF, Raska D, Michaelson MJ, and Bergey D, 1990. Molecular cytogenetic mapping of a nucleolar organizer region in cotton. *J Hered* 81:365-370.
- Rafalski JA and Tingey SV, 1993. Genetic diagnostics in plant breeding: RAPDs, microsatellites and machines. *TIG* 9:275-280.
- Sambrook J, Fritsch EF, and Maniatis T, 1989. *Molecular cloning: a laboratory manual*, 2nd ed. Plainview, New York: Cold Spring Harbor Laboratory Press.
- Skovsted A, 1934. Cytological studies of cotton: 2. Two interspecific hybrids between Asiatic and New World cotton. *J Genet* 28:407-424.
- Southern EM, 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* 98:503-517.
- Stallings RL, 1992. CpG suppression in vertebrate genomes does not account for the rarity of (CpG)_n microsatellite repeats. *Genomics* 17:1520-1521.
- Stallings RL, Ford AF, Nelson D, Torney DC, Hildebrand CE, and Moyzis RK, 1991. Evolution and distribution of (GT)_n repetitive sequences in mammalian genomes. *Genomics* 10:807-815.
- Sun L, Paulson KE, Schmid CW, Kadyk L, and Leinwand L, 1984. Non-*Alu* family interspersed repeats in human DNA and their transcriptional activity. *Nucleic Acids Res* 12:2669-2690.
- Tautz D and Renz M, 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res* 12:4127-4138.
- VanderWiel PL, Voytas DF, and Wendel JF, 1993. *Copia*-like retrotransposable element evolution in diploid and polyploid cotton (*Gossypium* L.). *J Mol Evol* 36:429-447.
- Walbot V and Dure III LS, 1976. Developmental biochemistry of cotton seed embryogenesis and germination: 7. Characterization of the cotton genome. *J Mol Biol* 101:503-536.
- Weber JL, 1990a. Human DNA polymorphisms and methods of analysis. *Curr Opin Biotechnol* 1:166-171.
- Weber JL, 1990b. Human DNA polymorphisms based on length variations in simple sequence tandem repeats. In: *Genomic analysis series*, vol. 1 (Tilgham S and Davies S, eds). Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press: 159-181.
- Wendel JF, 1989. New World tetraploid cottons contain Old World cytoplasm. *Proc Natl Acad Sci USA* 86:4132-4136.

Received October 26, 1993

Accepted November 10, 1994