A 3347-Locus Genetic Recombination Map of Sequence-Tagged Sites Reveals Features of Genome Organization, Transmission and Evolution of Cotton (Gossypium)

Junkang Rong,* Colette Abbey,† John E. Bowers,* Curt L. Brubaker,‡,§ Charlene Chang,† Peng W. Chee,*, Terrye A. Delmonte,† Xiaoling Ding,† Juan J. Garza,† Barry S. Marler,* Chan-hwa Park,* Gary J. Pierce,* Katy M. Rainey,* Vipin K. Rastogi,† Stefan R. Schulze,* Norma L. Trolinder,†† Jonathan F. Wendel,‡ Thea A. Wilkins,‡‡ T. Dawn Williams-Coplin,* Rod A. Wing,§§ Robert J. Wright,†,‡‡ Xinping Zhao,† Linghua Zhu† and Andrew H. Paterson*,†,[1]**

**Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602, †Department of Soil and Crop Science, Texas A&M University, College Station, Texas 77843, ‡Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa 50011, §Commonwealth Scientific and Industrial Research Organization, Canberra, Australia, **Coastal Plains Experiment Station, University of Georgia, Athens, Georgia 30602, ††Department of Plant and Soil Science, Texas Tech University, Lubbock, Texas 79409, ‡‡Department of Agronomy and Range Science, University of California, Davis, California 95616 and §§Department of Plant Sciences, University of Arizona, Tucson, Arizona 85721*

Manuscript received July 30, 2003
Accepted for publication September 23, 2003

ABSTRACT

We report genetic maps for diploid (D) and tetraploid (AtDt) Gossypium genomes composed of sequence-tagged sites (STS) that foster structural, functional, and evolutionary genomic studies. The maps include, respectively, 2584 loci at 1.72-cM (~600 kb) intervals based on 2007 probes (AtDt) and 763 loci at 1.96-cM (~500 kb) intervals detected by 662 probes (D). Both diploid and tetraploid cottons exhibit negative crossover interference; *i.e.*, double recombinants are unexpectedly abundant. We found no major structural changes between Dt and D chromosomes, but confirmed two reciprocal translocations between At chromosomes and several inversions. Concentrations of probes in corresponding regions of the various genomes may represent centromeres, while genome-specific concentrations may represent heterochromatin. Locus duplication patterns reveal all 13 expected homeologous chromosome sets and lend new support to the possibility that a more ancient polyploidization event may have predated the A-D divergence of 6–11 million years ago. Identification of SSRs within 312 RFLP sequences plus direct mapping of 124 SSRs and exploration for CAPS and SNPs illustrate the "portability" of these STS loci across populations and detection systems useful for marker-assisted improvement of the world's leading fiber crop. These data provide new insights into polyploid evolution and represent a foundation for assembly of a finished sequence of the cotton genome.

THE cotton genus, Gossypium L., is an excellent system for examining many fundamental questions relating to genome evolution, plant development, and crop productivity. Gossypium is composed of ~45 diploid and 5 allopolyploid species that occur naturally throughout the arid and semiarid regions of Africa, Australia, Central and South America, the Indian subcontinent, Arabia, the Galápagos, and Hawaii (Fryxell 1979, 1992). The diploid Gossypium species fall into eight cytological groups, or "genomes," designated A–G and K on the basis of chromosome pairing relationships (Beasley 1940; Phillips and Strickland 1966; Edwards and Mirza 1979; Endrizzi *et al.* 1985; Stewart 1994). While the diploid Gossypium all have the same chromosome number ($n = 13$), their haploid genome sizes vary from 1 to 3.5 Gb (Wendel *et al.* 2002), variation largely explicable in the best-studied cases (A *vs.* D) by different amounts of dispersed repetitive DNA (Zhao *et al.* 1998).

Allotetraploid cottons are all indigenous to the New World and unite the Old World A genome with the New World D genome in an A-genome cytoplasm (Galau and Wilkins 1989; Wendel 1989; Wendel and Albert 1992). The A- and D-genome progenitors are thought to have diverged from a common ancestor ~6–11 million years ago (MYA) and have been reunited in a common tetraploid nucleus ~1.1–1.9 MYA (Wendel 1989; Wendel and Albert 1992; Senchina *et al.* 2003; Wendel and Cronn 2003). Polyploid formation presumably involved migration of the Old World A-genome ancestor

[1]*Corresponding author:* Plant Genome Mapping Laboratory, University of Georgia, 111 Riverbend Rd., Room 228, Athens, GA 30602. E-mail: paterson@uga.edu

by saltwater dispersal, a mechanism that may contribute to the natural distributions of several members of the cotton tribe with comose seeds (Fryxell 1979). Polyploidization was followed by radiation and divergence with distinct tetraploid species now indigenous to Central America (*Gossypium hirsutum* L.), western South America (*G. barbadense* L.), northeastern Brazil (*G. mustelinum* Miers ex Watt), the Hawaiian Islands (*G. tomentosum* Nuttall ex Seemann), and the Galapagos Islands (*G. darwinii* Watt; Fryxell 1979). All tetraploid species have 26 gametic chromosomes, exhibit disomic pairing (Kimber 1961), and have similar genome sizes (Wendel *et al.* 2002) that have been variously estimated at 2.2–2.9 Gb.

Particularly interesting questions concerning Gossypium are the genetics underlying productivity and the quality of the world's leading natural fiber. World cotton commerce of ~$20 billion annually is made possible by an unusual feature of a few members of this taxon. Kim and Triplett (2001) noted that, "There are only a few cells in the plant kingdom that are as exaggerated in their size or composition as cotton fibers" and that some of these single-celled seed epidermal trichomes "may reach lengths of over 6 cm, or one-third the height of an Arabidopsis plant." (Although the quotation is correctly stated, we have been able to corroborate only reports of fiber lengths up to 5 cm in length.) Commercial cotton production is dominated by improved forms of the tetraploids *G. hirsutum* and *G. barbadense*. Wild A-genome diploids and AD-tetraploids each produce spinnable fibers that were a likely impetus for domestication (Stephens 1967; Fryxell 1979). Domesticated tetraploid cottons existed in the New World by 3500–2300 bc (Stephens and Moseley 1974) and have been widely distributed by humans throughout the world's warmer latitudes. Domesticated A-genome diploids may have existed in the Old World as early as 6000 bc (Moulherat *et al.* 2002), and *G. arboreum* remains intensively bred and cultivated in Asia.

The joining in a common nucleus of A and D genomes with very different evolutionary histories appears to have created unique avenues for response to selection. Directional selection by humans has consistently produced AD-tetraploid cottons that have yield and/or quality characteristics superior to those of A-genome diploids. Breeding of *G. hirsutum* (AD1) has emphasized maximum yield, while *G. barbadense* (AD2) is prized for fibers of superior length, strength, and fineness. Curiously, the D genome, from an ancestor that does not produce spinnable fibers (Lee 1984), contributes substantially to the fiber quality of tetraploid cottons (Jiang *et al.* 1998; Saranga *et al.* 2001; Paterson *et al.* 2002). Polyploid formation in cotton appears to have created unique avenues for response to selection not only for fiber quality but also for disease resistance (Wright *et al.* 1998), drought tolerance (Saranga *et al.* 2001), and perhaps other traits. Better understanding of the nonlinear interactions between the constituent subgenomes of cotton and other polyploids is of both basic and applied importance.

Herein, we further elucidate the structure, function, and evolution of the Gossypium genomes. We describe the first cotton map that coalesces into the expected 26 linkage groups (chromosomes), provide new insights into its transmission genetics and genome organization, and reveal new evidence suggesting that an ancient polyploidization event may have predated the A *vs.* D genome divergence of 6–11 MYA. The genetically anchored sequence-tagged sites (STS) comprising this map will foster many structural, functional, and evolutionary genomic studies relevant to development, evolution, and agriculture.

MATERIALS AND METHODS

Plant materials: The tetraploid (AtDt) map was constructed on the basis of additional probes applied to the mapping population reported by Reinisch *et al.* (1994) and is composed of 57 F_2 plants from a cross between *G. hirsutum* L. race "palmeri" and *G. barbadense* L. acc. "K101." The diploid D-genome map was based on 62 F_2 plants from a cross of *G. trilobum* (Moc. & Sesse ex. DC) Skovsted × *G. raimondii* Ulbr as described (Brubaker *et al.* 1999). DNA extraction, electrophoresis, blotting, and hybridization were performed as described by Reinisch *et al.* (1994).

To assign linkage groups to chromosomes, the hypoaneuploid stocks used by Reinisch *et al.* (1994) were supplemented with monosomic lines for *G. barbadense* chromosomes 12 and 16 and mono-telodisomic lines for the *G. barbadense* chromosome arms 16Sh, 17Sh, and 20Sh (generously contributed by David M. Stelly, curator of the Cotton Cytogenetics Collection, Texas Agricultural Experiment Station).

Probes: The probes used in this research are summarized in online supplemental Table 1 at http://www.plantgenome.uga.edu/cottonmap.htm and include genomic DNA (gDNA), cDNA, and simple-sequence repeats (SSR). Preparation and origin of most gDNA clones is described by Reinisch *et al.* (1994) for the tetraploid map and by Brubaker *et al.* (1999) for the diploid map. Most cDNA probes are derived from either a library of abscission tissue from *G. hirsutum* (prefixed Coau; provided by R. Wing) or a 7- to 10-day fiber of *G. arboreum* (prefixed Gate, Unig, and Gafb; provided by T. Wilkins and R. Wing). Probes prefixed with AEST are Arabidopsis cDNAs, generously provided by the Arabidopsis Biological Resources Center. Additional cDNA probes are from putative known-function genes (expressed sequence tags, or ESTs), disease resistance gene analogs (NBS), and ESTs relating to verticillium wilt resistance (W; generously provided by X. Fang, Chinese Academy of Agricultural Science). Most probes used in this study have been sequenced and can be found in GenBank (online supplemental Table 2 at http://www.plantgenome.uga.edu/cottonmap.htm). SSRs were developed by AMTEX (prefixed BNL, purchased from Research Genetics, Huntsville, AL), or our lab (prefixed CMS). Sequences and primers for all BNL SSR probes are listed at http//demeter.bio.bnl.gov/acecot.html. Sequences for CMS clones were also submitted to GenBank (online supplemental Table 2 at http://www.plantgenome.uga.edu/cottonmap.htm).

Map construction: Linkage groups were built using MAPMAKER/EXP 3.0 (Lander and Green 1987) on a PC largely as described by Reinisch *et al.* (1994). First, a subset of codomi-

TABLE 1

General mapping features of three subgenomes of tetraploid and diploid cotton

Subgenome	Locus number	Length (cM)	cM/locus	No. of gaps >10 cM (largest)
At	1308	2325.7	1.78	25 (14.8)
Dt	1276	2122.2	1.66	16 (19.2)
D	763	1493.3	1.96	9 (14.7)
Total	3347	5941.2	1.78	50 (19.2)

nant markers with a minimum of missing data was used to construct a framework. Linkage groups were initially assembled using the "group" command with a LOD score of 4, which is a somewhat more stringent threshold than that used for smaller genomes, but is appropriate for the ~4500-cM genome of cotton. Initial frameworks were nucleated using a small group of markers and the "compare" function. Additional markers were added into the framework with the "try" command in MapMaker and a custom computer program written in Microsoft Visual Basic (Bowers *et al.* 2003). The algorithm used by this program was to determine the genotypes of each individual for each interval between markers already placed on the map, with multiple genotypes possible for individuals in which crossovers were observed between the framework loci. In cases where genotypes were uncertain due to dominant markers or missing data, the genotype of the interval was inferred from flanking loci, assuming that the minimum number of recombinations had occurred. In search of a perfect match, an unmapped locus was tested against the possible genotypes for all intervals already on the map. If such a match was found, the marker was assigned to the appropriate interval and then the framework was recomputed. If no perfect match were found, a second pass was made looking for matches to all but one individual, followed by subsequent passes with higher numbers of nonmatching individuals. Loci from these subsequent passes were rechecked for scoring errors in the individual that did not fit the expected pattern. If the data were determined to be correct, the locus was then added to the framework map with a new recombination event not observed in the previous map. Loci mapping to the ends of the chromosomes could not be placed with this approach and had to be added to the framework manually or using Mapmaker. After the map was constructed, it was manually edited to reduce the number of recombinations by exporting the locations of crossovers observed in the map into a spreadsheet. Instances with multiple recombinations for an individual progeny plant were reordered, if possible, to reduce the total number of recombination events observed. This step involved extensive checking of the raw data (films) for errors, with the plants apparently responsible for double recombinations being rechecked. Ostensibly, codominant markers that could not be placed on the map were split into two dominant markers to attempt their mapping separately. Final map distances were computed using Kosambi (1944) centimorgans, and maps were drawn by another Visual Basic program.

Genetic interpretation of the restriction fragment length polymorphism (RFLP) phenotypes and nomenclature followed the system of Reinisch *et al.* (1994). Most names of the genetic loci were composed of the probe source followed by probe name, locus, code for the enzyme used in mapping, and finally segregation pattern. If a probe detected RFLPs at multiple loci, letters (a, b, c, . . .) were arbitrarily assigned to these loci. Six enzymes were used in mapping: E1 represents *Bam*HI; E2, *Cfo*1; E3, *Eco*RI; E4, *Eco*RV; E5, *Hin*dIII; and E6, *Xba*I. Dominant, recessive, and codominant loci were represented with D, R, and C, respectively.

DNA sequencing and sequence data processing: Plasmids were isolated using an alkaline lysis method modified for a 96-well plate format. Cycle sequencing reactions were performed using the BigDye terminator cycle sequencing kit (Applied Biosystems, Foster City, CA) and MJ Research (Watertown, MA) PTC-10 thermocyclers. Finished cycle sequencing reactions were filtered through Sephadex filter plates directly into Perkin-Elmer (Norwalk, CT) MicroAmp optical 96-well reaction plates. Sequence data processing used a pipeline anchored by a Microsoft Access relational database (RDB), which links sample and experimental information with each chromatogram file. The pipeline begins with the presequencing archival of experimental information such as sample names and specific clone attributes, which are imported into the ABI sequence plate records. Chromatogram files are archived in the RDB and then transferred to a Sun workstation for processing using Phred (Ewing *et al.* 1998) and Cross-Match. The data are then exported as simple text files in Fasta format to a PC workstation for further analysis and archival in the RDB.

RESULTS

Genetic maps: The tetraploid cotton map is composed of 2584 loci in 26 linkage groups (Table 1), a 1879-locus (nearly threefold) increase compared to our previously published map of 705 loci (Reinisch *et al.* 1994). An additional 62 loci were removed for reasons such as repeated mapping and difficulty in scoring. Despite the increased locus number, the recombinational length of the map dropped from 4675 to 4447.9 cM (Kosambi 1944), largely by virtue of a sufficiently high density of markers to distinguish scoring errors from true double recombinants. The map is based on a total of 5344 detected crossovers, which would correspond to 5370 potentially distinct map locations across the 26 chromosomes. We identified markers at 1504 (28.0%) of these locations. Overall, the average distance between consecutive loci is 1.72 cM, but the density of markers varies between chromosomes, ranging from 1.44 (LG D02) to 2.23 cM (chromosome 2). The largest gap between two adjacent loci is 19.2 cM (on chromosome 14). A total of 41 intervals in the tetraploid map remain >10 cM, with 16 in the Dt genome and 25 in the At genome. About a threefold variation in chromosome length was observed, ranging from 84 cM (chromosome 17) to 242 cM (LG D08). Detailed features of the map are in online supplemental Table 3 at http://www.plantgenome.uga.edu/cottonmap.htm.

The diploid D-genome map is composed of 763 loci detected by 662 probes and spans 1493.3 cM in 13 linkage groups (Table 1). This is a 153% increase (467 loci) over the prior version of this map, which included 306 loci. Ten loci were omitted from this map for similar reasons as the tetraploid map. Despite the significant increase in locus number, the total length of the map is essentially the same. This map is based on a total of

1929 detected crossovers corresponding to 1942 potentially distinct map locations. We have mapped markers at 505 (26.0%) of these locations. Overall, the average distance between consecutive loci is 1.96 cM, varying among chromosomes from 1.58 cM (D9) to 3.11 cM (D10). The biggest gap is 14.7 cM (D5) and only nine intervals are >10 cM. Detailed features of the map are in online supplemental Table 4 at http://www.plantgenome.uga.edu/cottonmap.htm.

The D genome is shorter than the tetraploid At and Dt subgenomes as a whole in genetic length (Table 1, Figure 1). Chromosome length varies over a somewhat narrower range in the diploid D genome than in the tetraploid, from 77.9 cM (D3) to 174.5 cM (D7). Markers were distributed more sparsely in the D genome than in the tetraploid genome (1.96 *vs.* 1.72 cM).

Recombinational interference: Recombinational interference was assessed by comparing the frequency of occurrence of "double crossover" genotypes (*i.e.*, aa-ab-aa; bb-ab-bb) to "adjacent crossover" genotypes (*i.e.*, aa-ab-bb; bb-ab-aa) as a function of the size of the interval that contains the two crossovers required to produce each genotype. In the absence of interference, these two different classes of genotypes would be equally probable; however, as a whole double crossovers were significantly ($P < 0.01$) more abundant than adjacent crossovers in both At and Dt subgenomes (563 *vs.* 383, 462 *vs.* 370). Virtually all of this difference was accounted for by cases in which the two loci were separated by <10 cM (Figure 2). In intervals >20 cM, the two classes of genotypes occur in a different frequency in the At and Dt subgenomes. In the At genome, the frequency of double crossovers is still generally greater than that of adjacent crossovers (except in the 20- to 29.9-cM interval) and the difference reaches significance in the 40- to 49.9- and 60- to 69.9-cM intervals. On the contrary, in the Dt genome, the two types of crossover are equally abundant in the 20- to 29.9-cM interval, and adjacent crossovers are more abundant in larger intervals.

In the diploid population, in total, double crossovers were slightly more abundant than adjacent crossovers (271 *vs.* 242), but this small difference was not statistically significant. However, over distances of 0–10 cM, the frequency of double crossovers is significantly higher than that of adjacent crossovers, consistent with the tetraploid mapping population (Figure 2). Like the Dt genome, over distances >20 cM, adjacent crossovers are generally more abundant than double crossovers, but the difference reaches significance only in the 20- to 29.9-cM interval (Figure 2).

Patterns of DNA marker locus distribution: DNA markers were unevenly distributed over the chromosomes of each map (Figure 1). To analyze this statistically, we partitioned each linkage group into intervals of 10 cM in length, except that the last interval in each group was either ≤15 or ≥5 cM to accommodate the varying lengths of the linkage groups. On the basis of the total number of loci per linkage group, the Poisson probability distribution function was applied to identify bins that contained significant ($P < 0.01$) excesses or deficiencies of various classes of markers.

A total of 65 intervals (shaded regions) comprising 49 clusters (16 composed of two consecutive intervals; see online supplemental Table 5a at http://www.plantgenome.uga.edu/cottonmap.htm) were marker rich, with one to three clusters on each chromosome except tetraploid chromosomes 1 and 25 and diploid linkage groups D3, D6, D7, D8, and D10. Many of the marker-rich regions corresponded in location between the tetraploid subgenomes and/or between the diploid and tetraploid maps. The Dt and D maps corresponded most closely: among eight concentrations of probes in the D-genome map, seven (87.5%) correspond to concentrations in the Dt map (*vs.* an overall 11.7% of intervals that are marker rich). The single D-genome marker-rich interval that is incongruous with the Dt counterpart corresponds to a marker-rich interval in the At genome. Among the 18 marker-rich regions in the Dt genome, 14 (78%) correspond to 14 marker-rich regions in the At genome (*vs.* an overall 12.7% of intervals that are marker rich). In five cases, a single At or Dt (3 and 2, respectively) marker-rich region corresponds to each of two unlinked marker-rich regions. In a single case, two genetically linked marker-rich regions that are separated by a normal region correspond on homeologous chromosomes (A01 and chromosome 18). Among the marker-rich regions that do not correspond between At and Dt, one corresponds between D and Dt, one between At and D, three are unique to Dt, and nine are unique to At.

A total of 17 intervals comprising 12 clusters were marker poor, with 9 on nine different Dt chromosomes, 3 on two different At chromosomes, and none on the D diploid (see online supplemental Table 5 at http://www.plantgenome.uga.edu/cottonmap.htm). While the 6.5% of Dt intervals that are marker poor appears significant, the 1.3% of At intervals that are marker poor is only nominally above the 1% false-positive rate. Only one case of correspondence among marker-poor intervals was found—between At chromosome 5 (cM 210–220) and Dt linkage group chromosome 22 (cM 70–80).

Distribution of dominant loci: A total of 1069 dominant loci were detected in the tetraploid map, including 520 loci segregating as the presence of alleles from *G. hirsutum* and 549 from *G. barbadense*. No cluster of dominant loci specific for one parent was located on any chromosome. Among eight main probe sources, the probes from two libraries (Coau and pVNC) show significantly more dominant alleles from *G. barbadense* than from *G. hirsutum*: 63 *vs.* 37 and 16 *vs.* 5, respectively.

SSR and single-nucleotide polymorphism (SNP)-containing loci: A total of 1749 cDNA, 710 gDNA, 124 SSR loci, and one isozyme were mapped in the tetraploid,

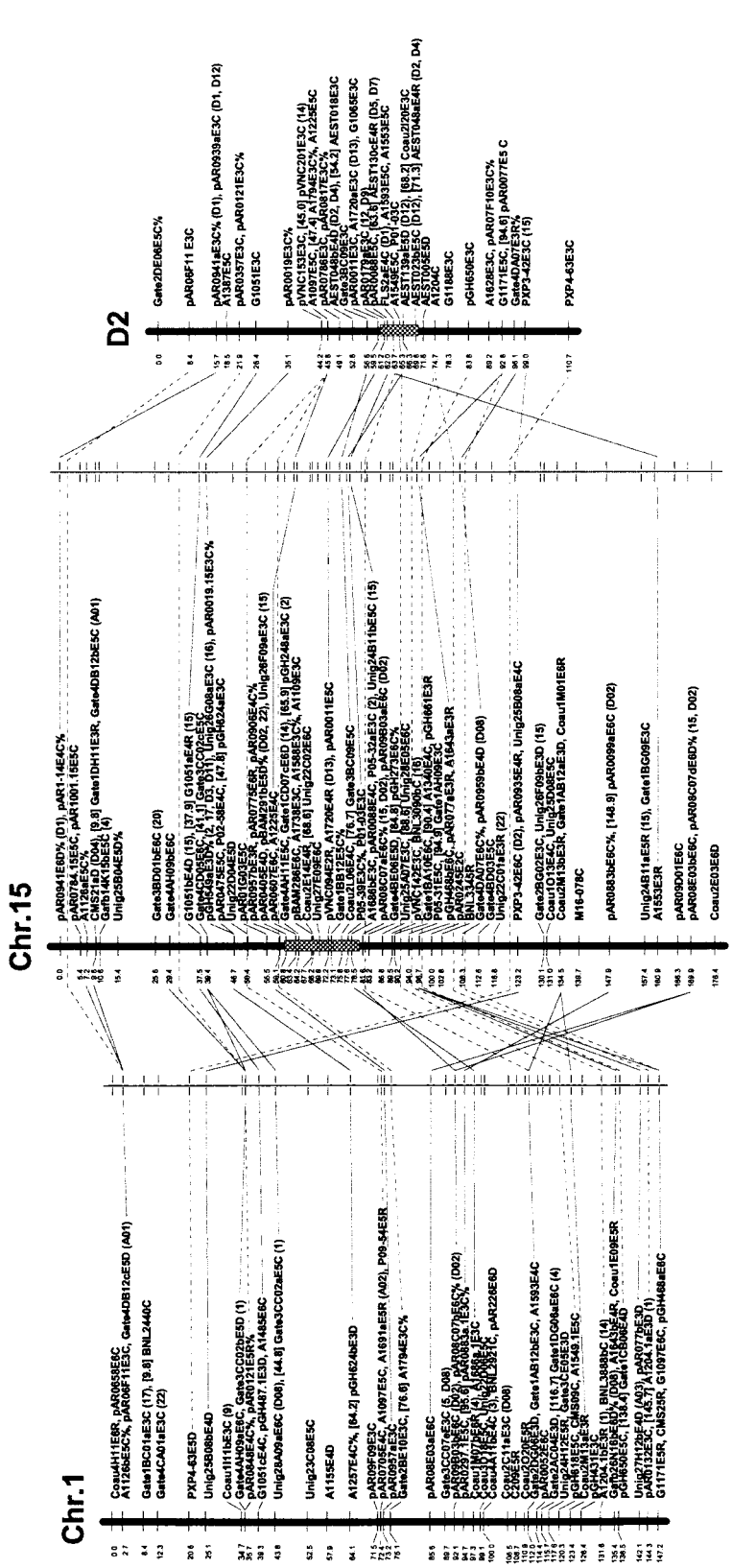

FIGURE 1.—Genetic linkage maps of tetraploid AD and diploid D genomes. Homologous and homeologous groups are arranged on the same page. Solid lines indicate At to Dt, Dt to diploid D, or At to Dt to diploid D duplications. Dashed lines indicate At to diploid D duplications. Locus names are as described in MATERIALS AND METHODS. A number and/or letter in parentheses following the loci represent the chromosomal locations of additional nonhomeologous duplications. Markers containing SSRs are shown with a percentage. Regions with significant excess or deficiency ($P < 0.01$) of loci are indicated by hatched or open boxes, respectively.

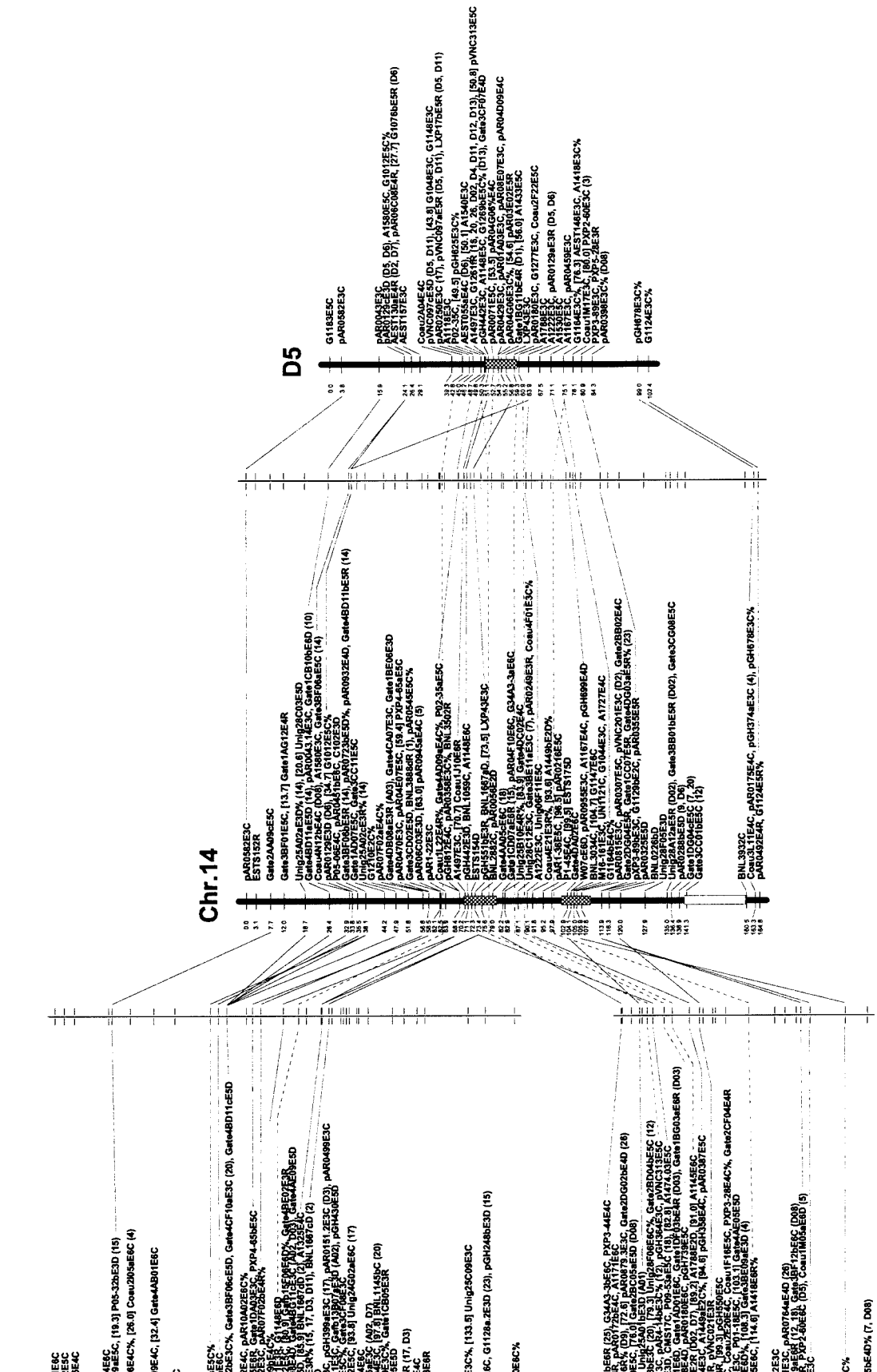

FIGURE 1.—*Continued.*

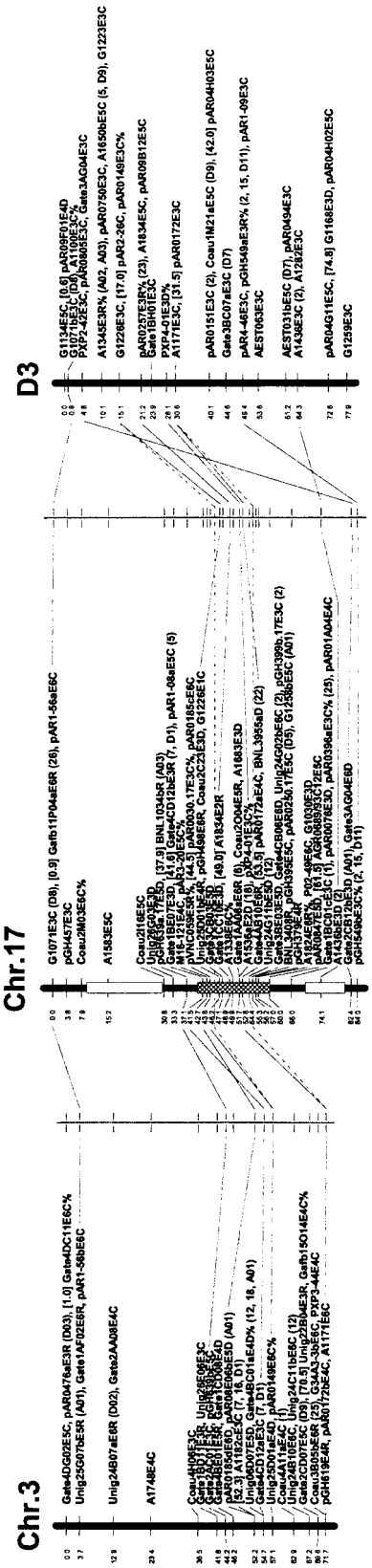

FIGURE 1.—*Continued.*

and 410 cDNA and 353 gDNA in the diploid D genome. The distribution of cDNA and gDNA loci appear to be similar in both tetraploid ($r = 0.434$) and diploid ($r = 0.505$) maps. In addition to 124 SSR loci that were mapped directly, we also identified 312 additional SSR-containing sequences (defining an SSR as six or more repeats of a dinucleotide or longer repeat units covering 15 or more base pairs; see online supplemental Table 2 at http://www.plantgenome.uga.edu/cottonmap.htm) within 1884 sequences of gDNA and cDNA probes mapped as RFLPs. The SSR-containing loci were distributed across the map in a pattern similar to that of the entire probe set ($r = 0.547$).

For a sampling of simple-sequence repeat motifs, $(GA)_{15}$, $(CA)_{15}$, $(AT)_{15}$, and $(A)_{30}$, we roughly estimated the number of copies present in the cotton genome that are composed of 10 or more repeat units (and thus tend to be more polymorphic than shorter arrays). Specifically, we end-labeled synthetic oligonucleotides with ^{32}P and screened ~20,400 λ genomic clones (~11.3% of the genome) from a partial *Sau*3AI library of *G. barbadense* "K101" with average insert length of 15 kb (Table 2), using a hybridization stringency that should detect only arrays of 10 or more repeat units (adjusted to accommodate different binding affinities of oligonucleotides; Table 2). Subcloning and sequencing of a sampling (CMS loci) showed that arrays of 9–29 units were detected. Assuming that each positive clone contained only one simple-sequence repeat array, the number of such long arrays in the genome was estimated as 1800, 538, 4500, and 9000 for the four dinucleotides, respectively. On the basis of a genome size of 2700 Mb, one of these four arrays should occur an average of once in 170 kb. Considering only the GA and CA elements, which are more reliable in PCR amplification, we find one array per 1155 kb (or ~2 cM).

We also explored levels and patterns of SNP variability in a sampling of genetically mapped sequence-tagged sites among a small selection of tetraploid cottons. For diploids, PCR amplification products might be sequenced directly; however, for a recently formed polyploid such as cotton, most PCR amplification products are mixtures of sequences from two or more divergent loci and require further purification before sequencing. To accommodate this, after PCR amplification of each of four genotypes (*G. barbadense* accession "K101" and cultivar Pima S6 and *G. hirsutum* race palmeri unnamed accession and cultivar "Acala Maxxa") with each of 24 primer pairs (individually, using conditions determined to be optimal for each primer pair), we diluted the samples to equimolar concentrations, then pooled the 24 samples for each genotype, purified with Sephadex G-50 columns, and cloned using the pGEM-T Easy vector system (Promega, Madison, WI) following the manufacturer's protocols. From the pooled cloned products, a total of 768 clones (192 per genotype, or ~5× the coverage of each homeologous locus) were sequenced from one direction, with 664 (86.5%) sequences, or an aver-

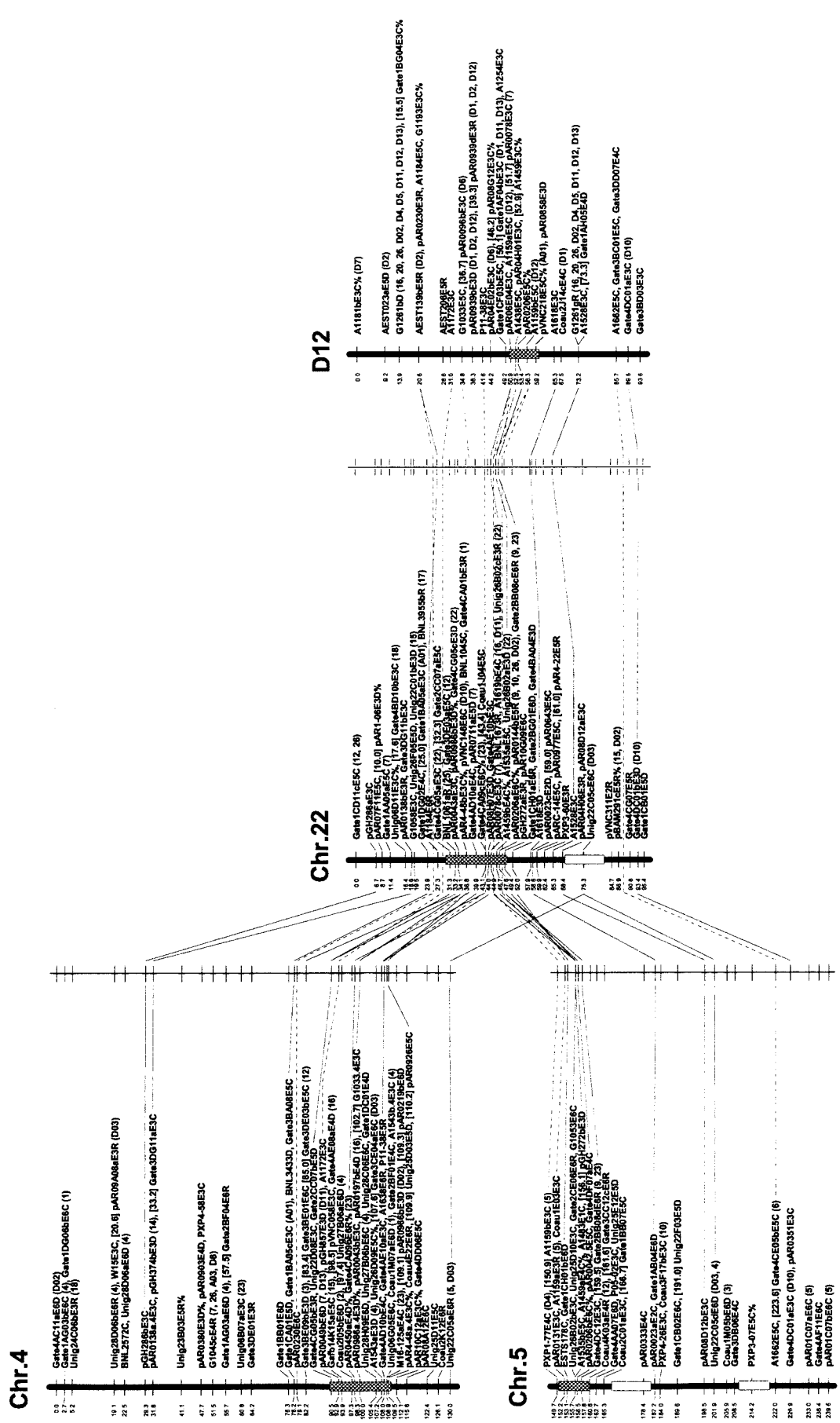

FIGURE 1.—*Continued.*

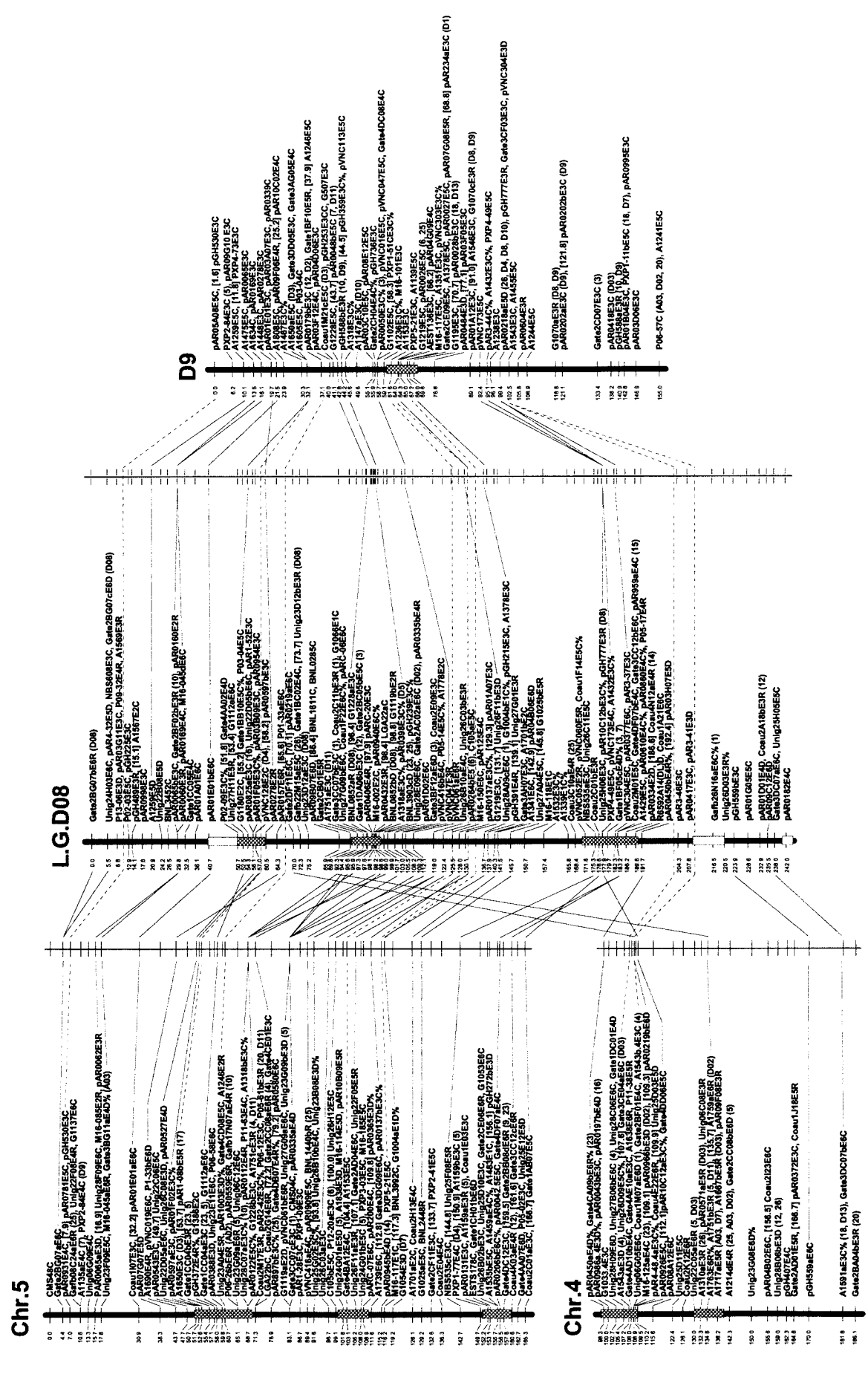

FIGURE 1.—*Continued.*

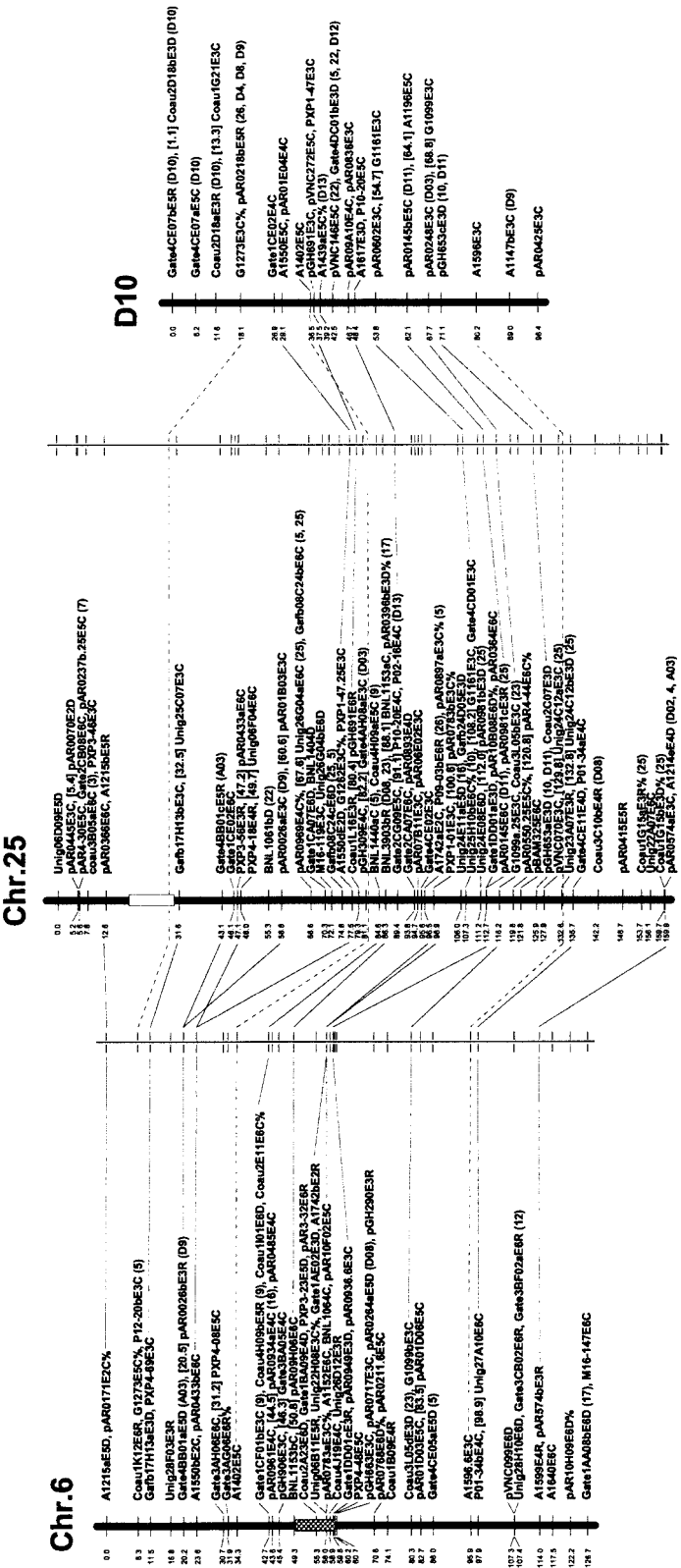

FIGURE 1.—*Continued.*

FIGURE 1.—*Continued.*

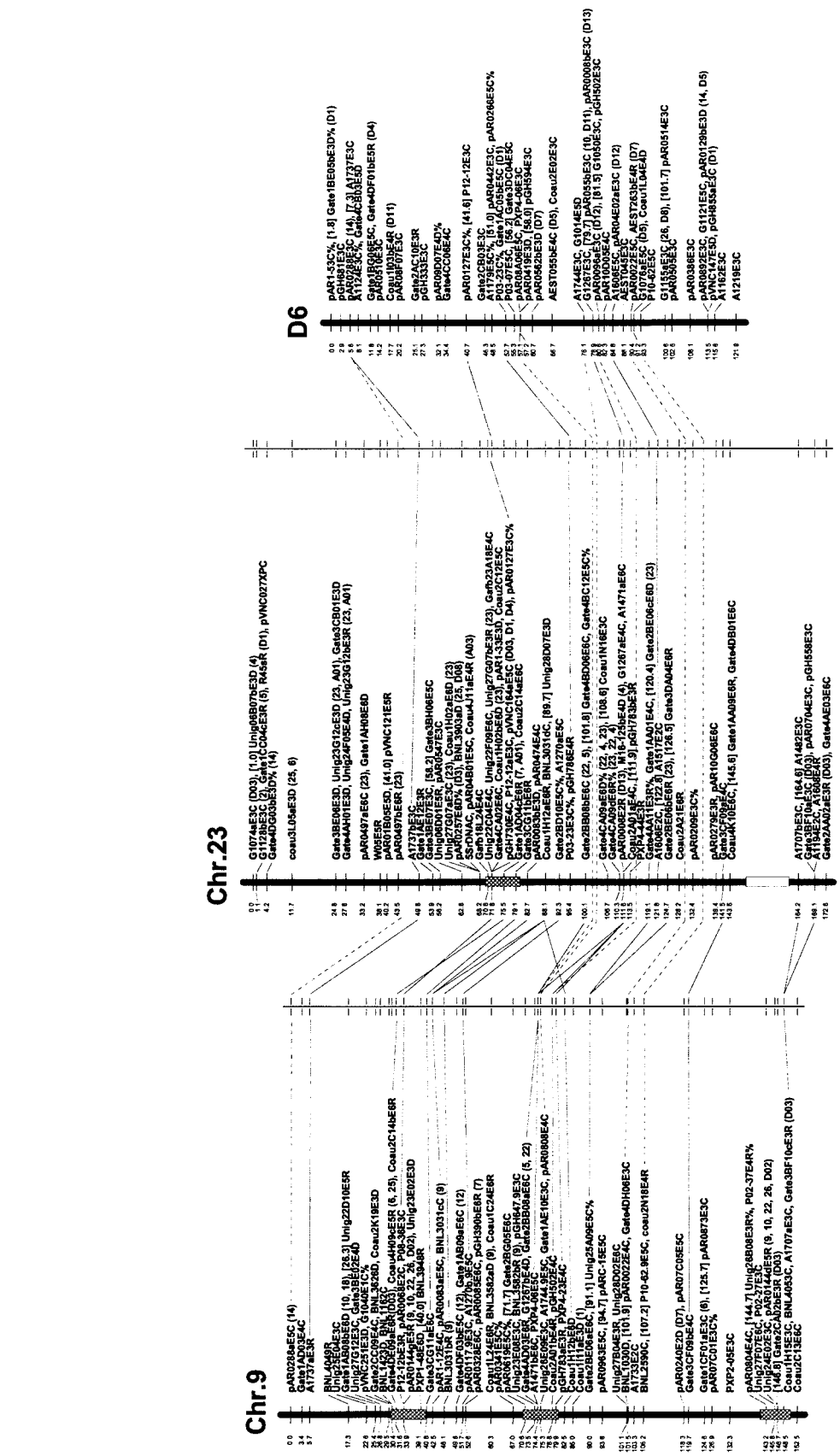

FIGURE 1.—*Continued.*

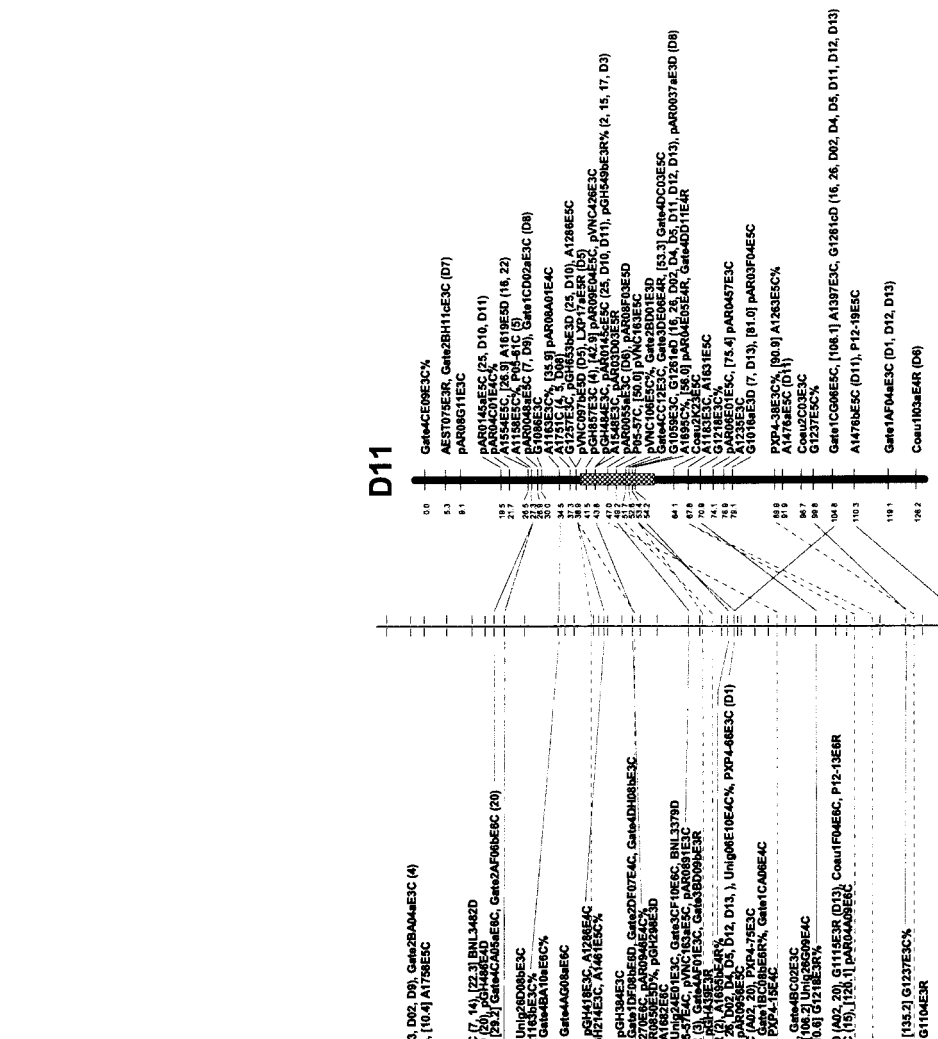

FIGURE 1.—*Continued.*

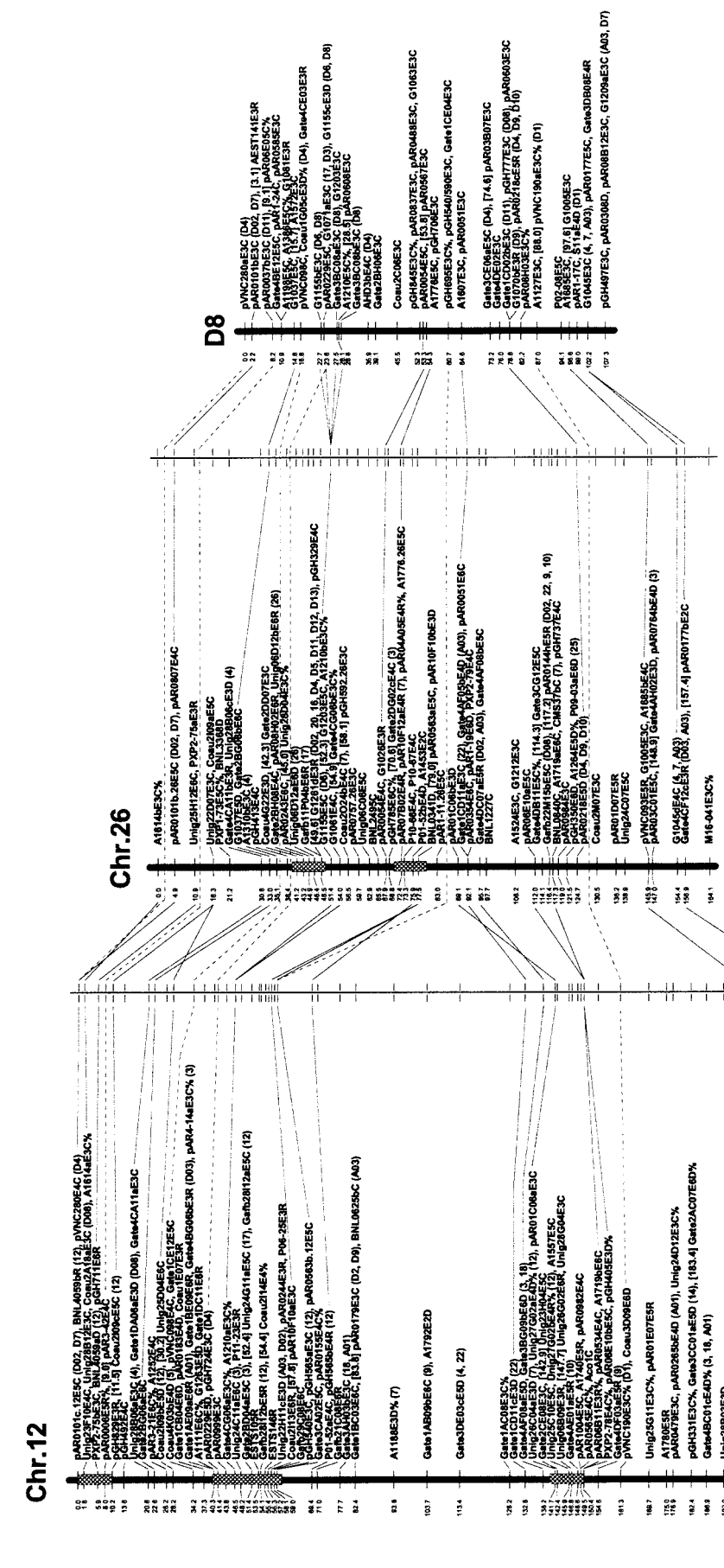

FIGURE 1.—*Continued.*

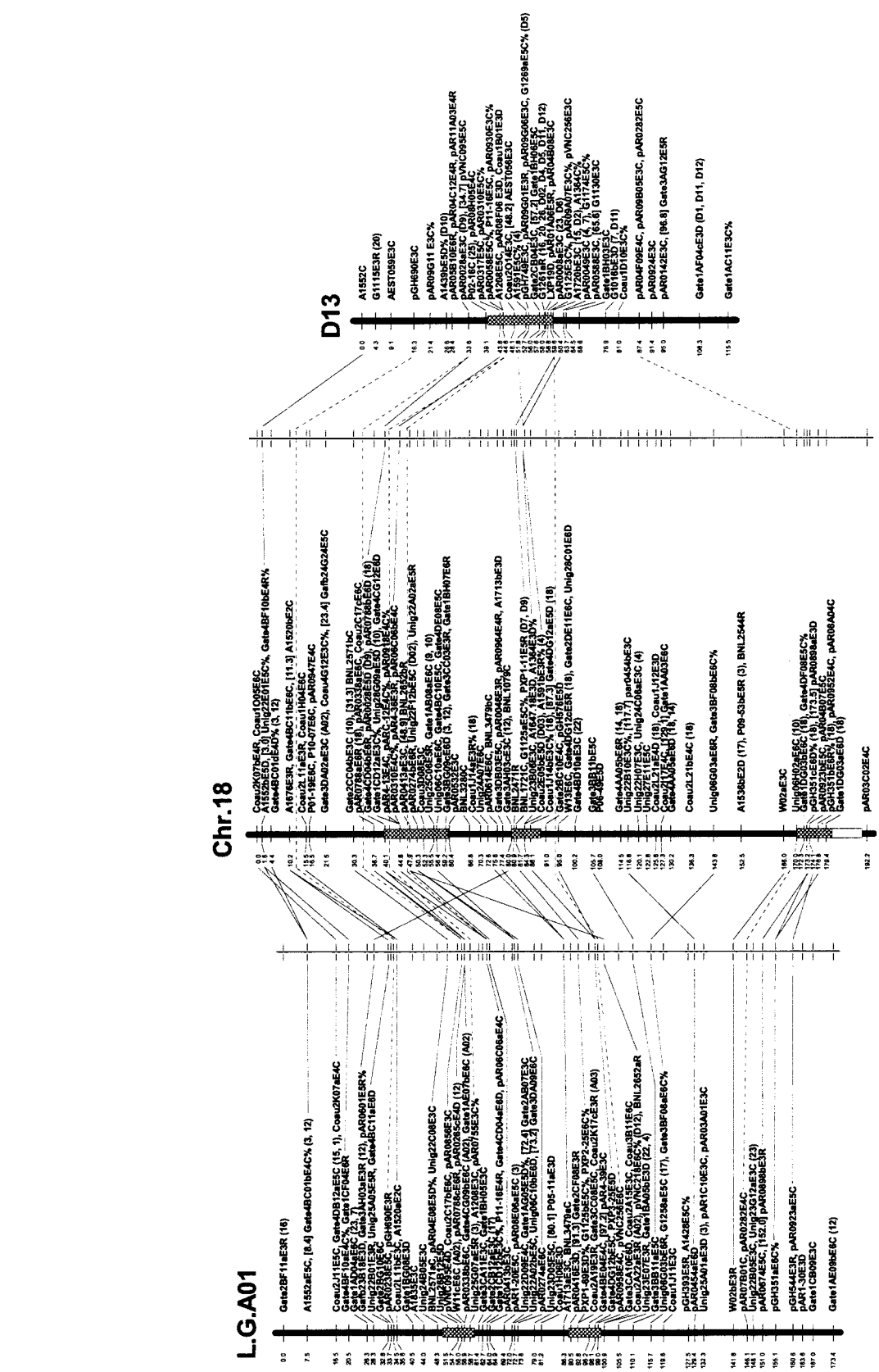

FIGURE 1.—*Continued.*

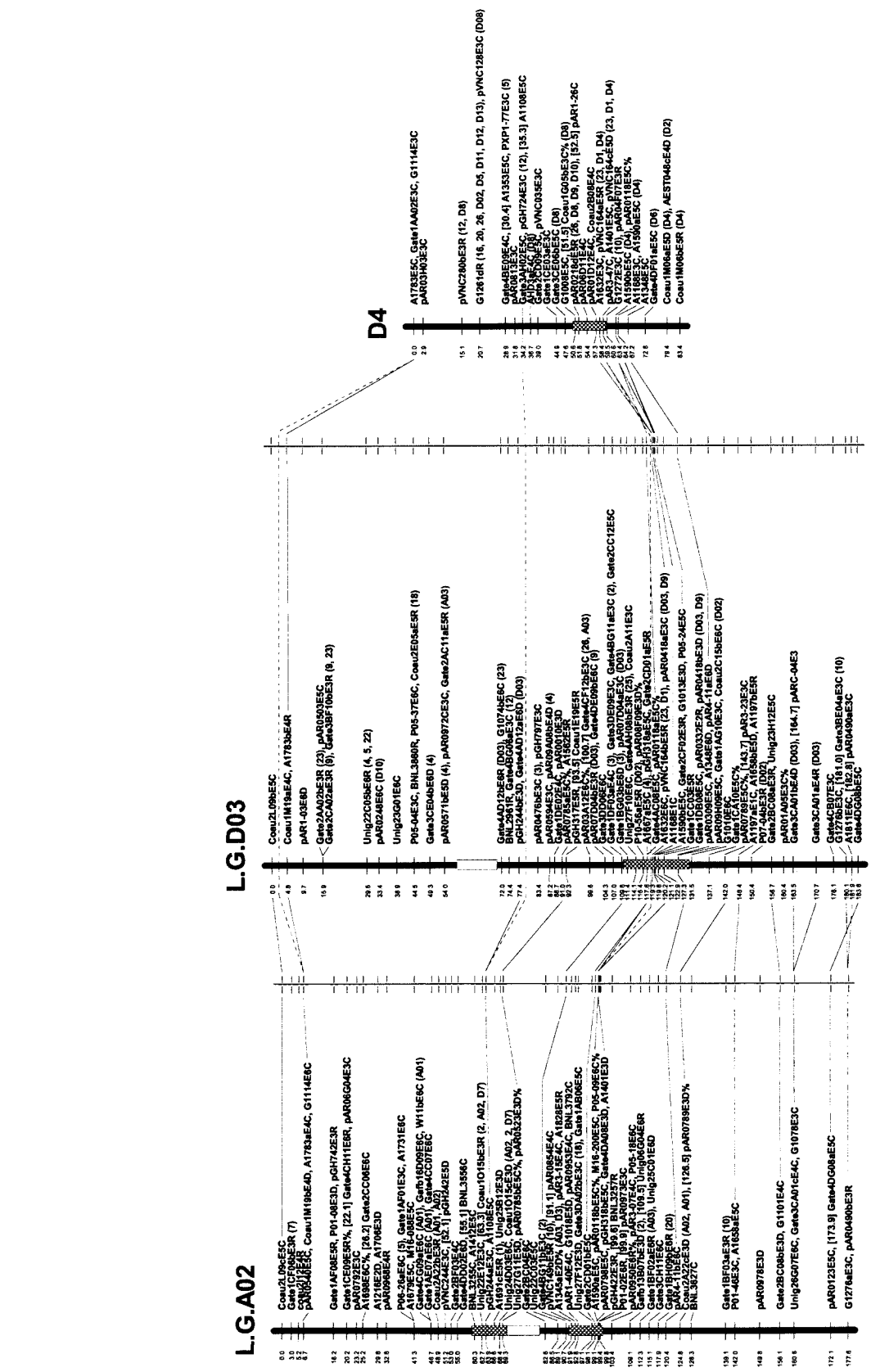

FIGURE 1.—*Continued.*

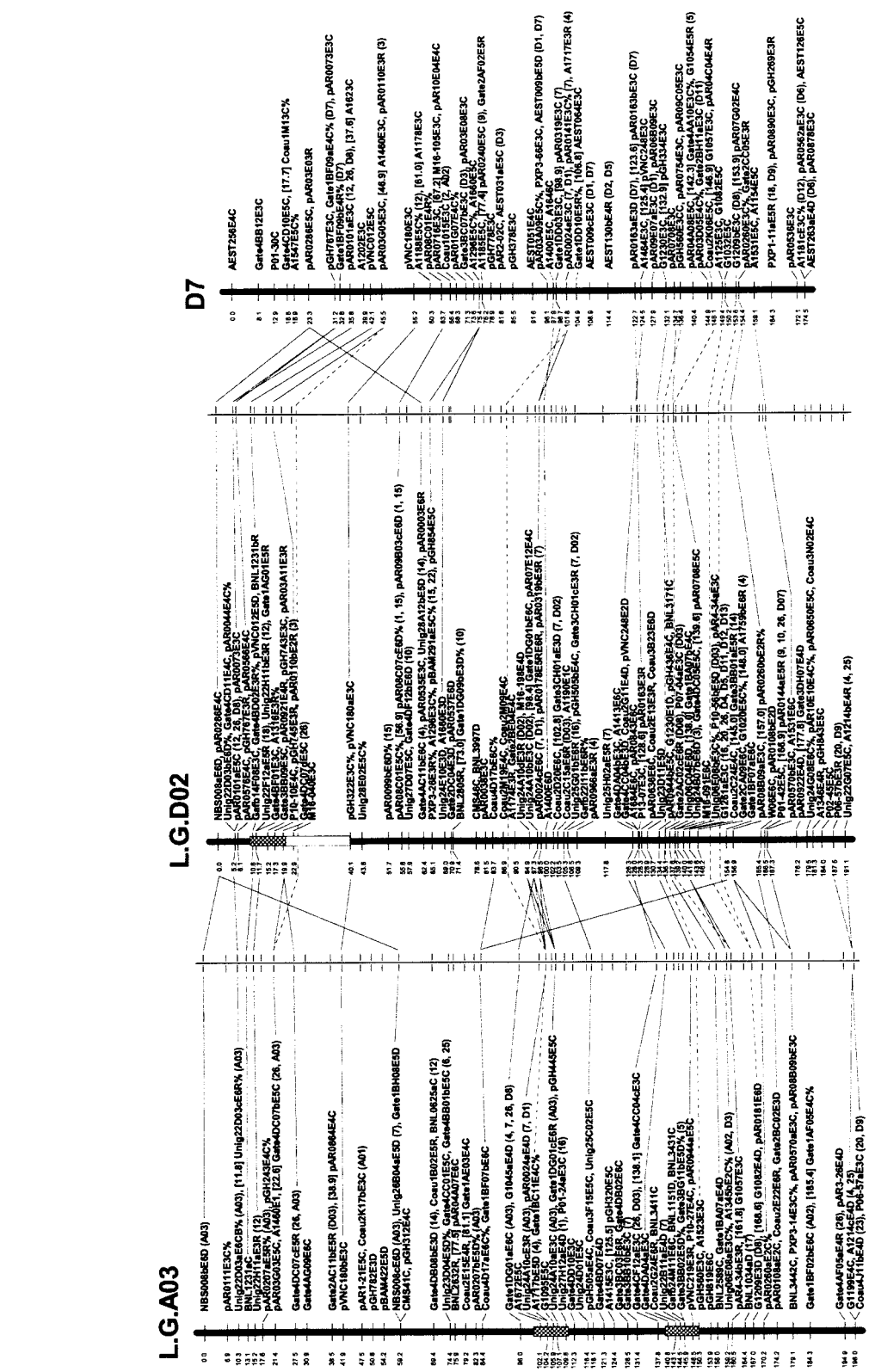

FIGURE 1.—*Continued.*

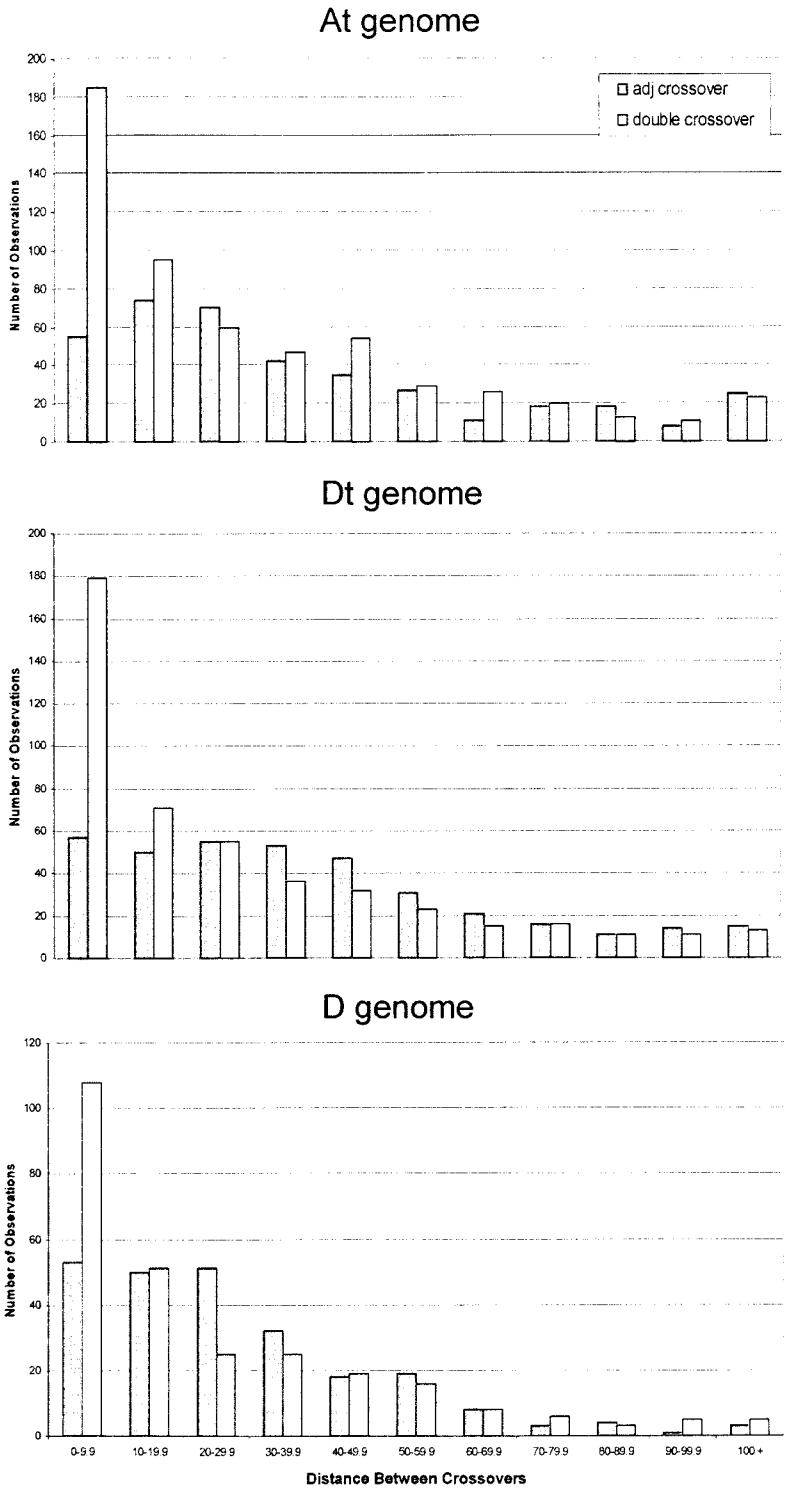

FIGURE 2.—Frequency of adjacent crossovers (two crossovers occurring on different chromosomes within the same individual) *vs.* double crossovers (two crossovers occurring on the same chromosome in the same individual) in intervals at 10-cM increments.

age of ~3 per homeologous locus in each genotype, meeting our (Q20) quality criteria. All primer sets were represented by at least one sequence, but with a range of 1–65 clones per primer pair (not shown). Hence, the savings in cost and labor associated with cloning and sequencing of individual amplifications from each genotype was partly sacrificed to uneven sampling of different loci.

Comprehensive analysis was performed on a subset of 12 primer pairs (Table 3) spanning 5409 nucleotides (nt) in each of the two subgenomes, which had relatively complete representation among the genotypes. For 10 of the 12 primer pairs, sufficient information was available to infer the subgenomic affinity of SNPs (on the basis of correspondence to diploid sequences): 6 of these 10 pairs showed SNPs or indels that distinguished

TABLE 2

Occurrence of selected simple-sequence repeats in the cotton genome

Oligo[a]	No. of positive clones	Estimated no. of arrays in genome
$(A)_{30}$	1,020	9,000
$(AT)_{15}$	510	4,500
$(GA)_{15}$	204	1,800
$(CA)_{15}$	61	538
Total	1,795	15,838

[a] Arrays $(A)_{30}$ or $(AT)_{15}$: hybridization at 40°, 6× SSC; wash two times at 25°, 6× SSC and then one time at 40°, 6× SSC. Arrays $(GA)_{15}$ or $(CA)_{15}$: hybridization at 65°, 6× SSC; wash two times at 25°, 6× SSC and then one time at 65°, 6× SSC.

between subgenomes. The minimal rate of variation per nucleotide of 1.06% between subgenomes (estimated considering indels as a single mutational event regardless of their length) was substantially greater than the level of variation between species (0.35%) or genotypes within species (0.37 and 0.14%). The vast majority (72.3%) of mutational events were SNPs, although a total of 12 indels were found, ranging from 1 to 37 nt in length.

Assignment of linkage groups and morphological markers to subgenomes and chromosomes: Coalescence of the map into 26 linkage groups (from 41; Reinisch *et al.* 1994), equal to the gametic chromosome number, has enhanced and extended the association of linkage groups with subgenomes and cytologically distinguishable chromosomes. Compared to the prior map (Reinisch *et al.* 1994), the following concatemers have been formed: LG A07 was linked to chromosome 5, LG A08 to chromosome 9, LG U03 to LG A01, LG U04 to LG A03, LG U07 to LG A06, LG U09 to chromosome 14, LG D10 to LG D01, LG U06 and LG U05 to chromosome 18, and LG U01 to LG D02. Thirteen linkage groups were found to be At subgenome and 13 Dt subgenome, as described (Reinisch *et al.* 1994).

A total of 20 cotton chromosomes have now been identified, adding to our prior work (Reinisch *et al.* 1994): 2 on the basis of the availability of additional cytological stocks and DNA probes (chromosomes 12 and 16) and 2 on the basis of inferences from the identities of homeologous linkage groups. The identity of chromosome 12 was also confirmed by recent mapping of a fiber mutant (*N1*; J. Rong and A. H. Paterson, unpublished data) previously reported to locate on this chromosome (Endrizzi *et al.* 1985). Several duplicated DNA markers identify our former LG A05 (Reinisch *et al.* 1994) to be the homeolog of chromosome 1; morphological data (Endrizzi *et al.* 1985) suggest that the homeolog of chromosome 16 is chromosome 7, which replaces our earlier designation (LG A05). This inference has recently been supported by marker screening on a chromosome 7 hypoaneuploid stock (Lacape *et*

TABLE 3

Summary of interspecific, intraspecific, and subgenomic SNPs discovered

		Rate of variation per nucleotide (%)						
STS	Total nt[a] analyzed	Homeologous (subgenomic)	Intraspecific *G. barbadense* (K101 *vs.* Pima)	Intraspecific *G. hirsutum* (Palmeri *vs.* Maxxa)	Interspecific (*G. hirsutum vs. barbadense)*	Variation due to SNPs	Variation due to indels	Length of indels (nt)
A1107	420	0.95	0.48	0	0.48	100	0	—
A1110	550	0	0	0	0	0	0	—
A1126	285	1.05	0	0	0	66.7	33.3	37
A1131	502	1.59	0	0.2	0	66.7	33.3	1, 2, and 3
A1148	505	0	0	0	0.2	100	0	—
A1152	422	0	0	0	0	0	0	—
A1167	450	0.89	0.44	0	0.44	66.6	33.3	1 and 4
A1174	470	n.d	0	0	1.27	83.3	16.7	7
A1364	480	0	1.04	1.04	1.04	100	0	—
PAR0060	480	2.5	2.08	0	0	90.9	9.1	3 and 22
PAR0827	385	3.64	0	0.26	0.52	93.8	6.2	11
PCK	460	n.d	0.43	0.22	0.22	100	0	6 and 18
Average % of variation		1.06	0.37	0.14	0.35	72.3	11.9	

Sequences with subgenomic affinity cannot be positively identified, due to either a high divergence between the A-genome probe sequence and the PCR-amplified tetraploid genome sequences or no diploid genome sequence being available (PCK primers were designed from a *G. barbadense* cDNA clone sequence that preferentially expressed in fiber tissues). Sequences are available from GenBank, accession nos. CC737118–CC737319.

[a] The total nucleotides analyzed may be different from the PCR fragment length because of poor sequence quality, particularly in reads beyond 500 nucleotides.

al. 2002). Independent data have shown that the homeolog of chromosome 17 is chromosome 3 (Lacape *et al.* 2002), which replaces our earlier designation (LG A06).

A few incongruities between various data types suggest the need for reassignment of several chromosome names. First, the identification of chromosome 12 herein and elsewhere required us to revisit the identity of its inferred homeolog. The chromosome that we now (see above) know to be chromosome 12 had previously been shown to be homeologous to what we thought was chromosome 22 (Reinisch *et al.* 1994); however, classical data strongly suggest homeology of chromosomes 12 and 26. The incongruity was traced to an error in a 1992 lab notebook; three loci previously thought to be diagnostic of chromosome 22 were actually diagnostic of chromosome 26, reconciling our data with the classical inference. Classical cytological data (Endrizzi *et al.* 1985) show that chromosome 22 is homeologous to chromosome 4; consequently, our prior LG D07 is now inferred to be chromosome 22, an inference supported by recent mapping of the *Li1* fiber mutation (J. Rong and A. Paterson, unpublished data) previously assigned to this chromosome (Endrizzi *et al.* 1985). Finally, Reinisch *et al.* (1994) referred to classical assignments of homeology on the basis of data from Endrizzi *et al.* (1984) and unfortunately overlooked Endrizzi *et al.* (1985), which contained additional data. Hence, it was not noted that chromosome 10 is homeologous to chromosome 20 (Endrizzi *et al.* 1985); on the basis of only a single diagnostic marker, we had tenuously inferred the homeolog of chromosome 5 to be chromosome 20 (Reinisch *et al.* 1994). Reexamination of the single diagnostic marker, detected by a multicopy probe, raised doubts. Further, a new marker (pAR0827) on LG D04, the homeolog of chromosome 10, showed the *G. barbadense* allele in a cytogenetic stock monosomic for chromosome 20. Consequently, we renamed our former LG D04 chromosome 20 and changed our former chromosome 20 to a new designation, LG D08.

While our chromosomal assignments are closely congruent with one recent study (Lacape *et al.* 2002), some conflicts with a second study may reflect the complexity associated with the use of hypoaneuploid stocks, even in a recently formed polyploid such as cotton. Zhang *et al.* (2002) derived new hypoaneuploids by crossing a local genotype (Hai 7124) to hypoaneuploids obtained from the same source as those that we used directly (D. Stelly, Texas A&M University). On the basis of their new hypoaneuploids, Zhang *et al.* (2002) reported identification of one previously unidentified chromosome (chromosome 11). Among eight genetically mapped SSR loci that can be compared, chromosome 11 of Zhang *et al.* (2002) corresponds to our LG D08. However, chromosome 11 is an A-subgenome chromosome, and LG D08 has been firmly assigned to the D-subgenome on the basis of seven allo-allelic loci with no conflicting data [Reinisch *et al.* (1994), who noted that at the time, it was erroneously named chromosome 20]. We are retaining our nomenclature for LG D08 at this time. Several other chromosome assignments by Zhang *et al.* (2002) conflict with both ours and those of Lacape *et al.* (2002). Homeologous linkage groups named chromosome 5 and chromosome 18 by Zhang *et al.* (2002) correspond to our homeologous LGs D02 and A03 on the basis of 2 of 6 and 7 of 10 comparable SSR loci, respectively. While our data strongly support the homeology of these two groups, the chromosomal assignments of Zhang *et al.* (2002) conflict with firm subgenomic assignments (Reinisch *et al.* 1994) for each of the two linkage groups. We and others (Lacape *et al.* 2002) have each identified different, nonhomeologous linkage groups to be chromosome 5 and chromosome 18, which are each of the correct subgenomic affinity, albeit noting that our chromosome 5 assignment is based on only a single marker. While we remain cautious about our assignment of chromosome 5, on the basis of the incongruous subgenomic assignment of the Zhang *et al.* (2002) designation, we have not revised our nomenclature. Chromosome 20 of Zhang *et al.* (2002) corresponds to our chromosome 18 on the basis of 6 of 10 comparable loci, with 2 of the incongruous loci mapping to the homeolog (our LG A01). While this invokes no conflicts with subgenome assignment (both are D), our identification of chromosome 18 is supported by both RFLP and morphological marker data. Further, our chromosome 20 assignment is supported by homeology, RFLP and subgenomic data, and an independent study (Lacape *et al.* 2002); therefore, pending additional information, we are retaining this assignment. Finally, chromosome 3 of Zhang *et al.* (2002) corresponds to our chromosome 14; however, this is based on only 2 loci, and the two chromosomes are partly homeologous, so this may simply represent different homeologous polymorphisms.

In addition to those described above, several additional morphological traits were mapped to chromosomes. On the basis of previously described populations (Wright 1999) and χ^2 tests using stringent statistical criteria ($P > 0.001$, to accommodate the large genome of cotton), segregation of petal color (the classical Y1 gene; Hutchinson and Silow 1939) was associated with the interval between *pAR4-39* and *pAR338b* on linkage group A01 (Empire B3 cross: $P = 0.00019$), consistent with its prior assignment to the A subgenome (Endrizzi *et al.* 1984). Anther color was associated with the *A1535b–pXP2-41* interval on chromosome 5 in two populations (Empire B2, B3 × Pima S6 crosses: $P = 0.00026$ and 0.00012, respectively) and with *pAR0711* on either chromosome 7 (46.0 cM) or chromosome 22 (39.9 cM) in the S295 × Pima S6 cross only ($P = 0.00012$). Petal spot mapped to two regions, including the *pGH377–pAR449* interval of chromosome 1 ($P = 0.00036$) and the *A1182b-pAR888* region of chromosome 7 ($P = 0.0003$), possibly representing two of the

three reported classical mutations for the trait (Harland 1929; Stephens 1974; Rhyne and Carter 1991). The mapping of pubescence (Wright *et al.* 1999) and leaf morphology (Jiang *et al.* 2000b) are described elsewhere.

Patterns of DNA marker locus duplication: Among the 2007 different probes mapped in the tetraploid cotton genome, a total of 427 detected 2 loci, 58 detected 3 loci, 10 detected 4 loci, and 1 detected 6 loci. Duplicated loci are not randomly distributed (Figure 3; see also clickable annotated version at http://www.plantgenome.uga.edu/cottonmap.htm). In tetraploid cotton, loci on a particular At-subgenome chromosome almost invariably showed duplications that were concentrated on 1 or 2 Dt-subgenome chromosomes. For example, 21 of 43 probes mapped on chromosome 1 have counterparts on chromosome 15. The duplicated loci of the remaining 22 probes were distributed across 13 different chromosomes with a maximum of only 4 on any single chromosome. In total, 10 Dt-genome chromosomes have a much higher frequency of duplicate loci (in all cases numbering 15 or more) on one of the At-genome chromosomes than on any other chromosomes. The remaining 3 Dt-genome chromosomes (chromosome 14, chromosome 22, and LG D08) have a higher frequency of duplicated loci (6 or more) on 2 At-genome chromosomes than on any other chromosomes. On the basis of the Poisson distribution, 5 or more loci duplicated on a pair of chromosomes deviates significantly from a random distribution ($P = 0.03$). In most cases, the linear order of the duplicated loci on the pairs of chromosomes is quite consistent (Figures 1 and 3), except for the subset in which we deduced structural changes (see below). Putatively homeologous chromosomes (or segments) are presented with putatively homeologous loci connected by lines (Figure 1).

To investigate the correspondence of diploid and tetraploid chromosomes, 257 common probes were mapped on both populations. For any single linkage group in the D diploid map, corresponding loci were heavily concentrated on two homeologous chromosomes of tetraploid cotton (Figures 1 and 4; see also a clickable annotated version at http://www.plantgenome.uga.edu/cottonmap.htm). For example, among the 40 loci mapped on diploid chromosome D1, counterparts of 29 were located on homeologous tetraploid chromosomes 7 (16 loci) and 16 (13 loci). The remaining 11 loci were distributed over nine other chromosomes, with no chromosome having >2 loci. The linear order of the common probes is similar between tetraploid and diploid chromosomes. Putatively corresponding diploid and tetraploid chromosomes (or segments) are presented with putatively corresponding loci connected by lines (Figure 1).

Comparative transmission and organization of homologous and homeologous groups: Chromosomes of the At subgenome averaged ~10% longer than those of the Dt subgenome, and ~55% longer than those of the D diploid (Table 2). Since very similar numbers of markers mapped to the At and Dt subgenomes, the shorter recombinational length may explain the slightly higher marker density on the Dt genome than on the At genome (1.65 *vs.* 1.77 cM/locus). All D chromosomes are shorter than their corresponding At homeologs and eight are shorter than their Dt homeologs, with the remaining five (D1, D3, D7, D11, and D12) roughly equivalent to their Dt homeologs (chromosome 16, chromosome 17, LG D02, chromosome 20, and chromosome 22; Figure 1). Five At chromosomes (chromosome 3, chromosome 4, chromosome 7, chromosome 10, and chromosome 12) were much longer than homeologous Dt chromosomes. Four At chromosomes (chromosome 5, LG A01, LG A02, and LG A03) are close in length to their homeologous Dt counterparts and only four At chromosomes were shorter than Dt chromosomes.

Chromosome structural changes: Above it was shown that three Dt-genome chromosomes (chromosome 14, chromosome 22, and LG D08) have homeologous relationships with two At chromosomes. Nonoverlapping sets of loci on chromosome 22 and LG D08 have counterparts on different regions of chromosome 4 and chromosome 5, respectively. By contrast, chromosome 22 and LG D08 have a simple one-to-one homeologous relationship with their corresponding diploid D chromosomes, D12 and D9. This is consistent with the finding (Brubaker *et al.* 1999) that chromosomes 4 and 5 have undergone a reciprocal translocation. Additional markers now available more precisely delineate the breakpoint to a region that corresponds to a marker-rich interval likely to represent the centromere of homeologous chromosome 22. Across most pairs of homeologous chromosomes, the linear order of loci was substantially conserved (Figures 1, 3, and 4), although there is also clear evidence for localized inversions. See online supplemental Table 6 at http://www.plantgenome.uga.edu/cottonmap.htm for detailed information about inversion. Most locus order differences are due to reversals of neighboring markers explicable by inversion. Many apparent inversions involved only two neighboring loci and in our small mapping population (57 plants) may be due to occasional missing data or scoring errors. However, some putatively orthologous loci mapped to distal locations that cannot be easily explained by inversion. For example, Unig25B08 mapped to locations of 25.1 cM on chromosome 1 and 123.2 cM on chromosome 15, which was 93.8 cM away from the next common marker, Gate4AH09bE6C. Such differences may be the result of ancient duplication accompanied by structural changes or by proximal duplication together with failure to find the true ortholog due to either lack of polymorphism or its deletion.

Duplication within individual subgenomes: In addition to loci duplicated on homeologous chromosomes, many loci were duplicated on chromosomes that were nonho-

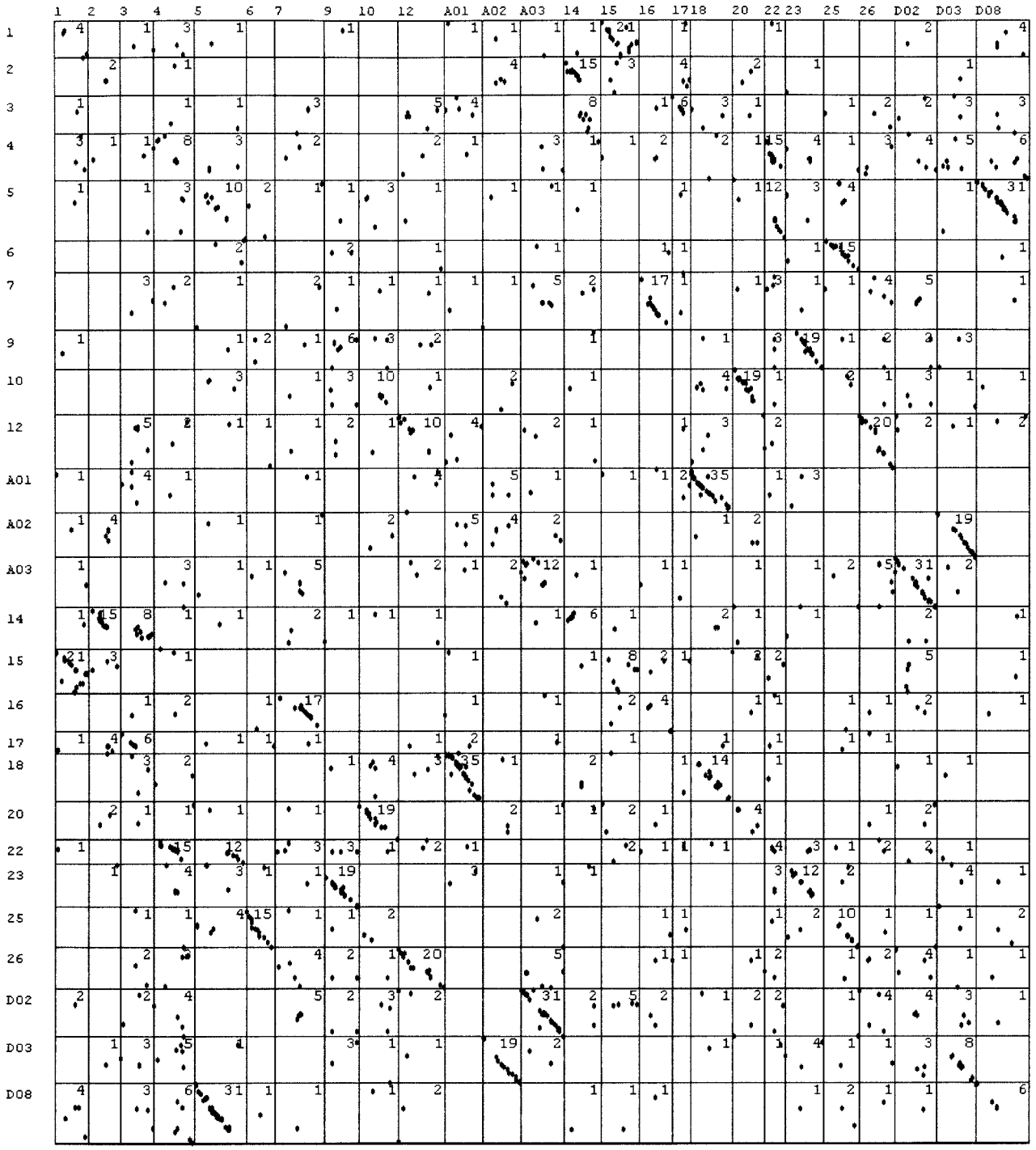

FIGURE 3.—Oxford plot showing At-Dt genome correspondence. Each dot represents a probe that is mapped multiple times in tetraploid cotton, with *x*- and *y*-axes representing chromosomal locations. The number at the top right in each cell represents the total number of corresponding loci for the pair of chromosomes. Some points represent multiple hits with the same genetic location. For further detail and a listing of data points in each cell, see http://www.plantgenome.uga.edu.

meologous and/or within the same subgenome. Among 679 pairs of duplicate loci detected with 499 probes in the tetraploid map (Figure 3), 289 pairs were on putatively homeologous chromosomes. Among the 389 pairs on nonhomeologous chromosomes, 160 pairs involved both subgenomes, 120 were on pairs of At chromosomes only, and 109 were on pairs of Dt chromosomes only. Unlike the duplicate loci on homeologous chromosomes, the nonhomeologous duplicate loci were scattered over many chromosomes. For example, 67 probes on LG A03 detected duplicated loci on 18 chromosomes, with 31 on the homeologous chromosome LG

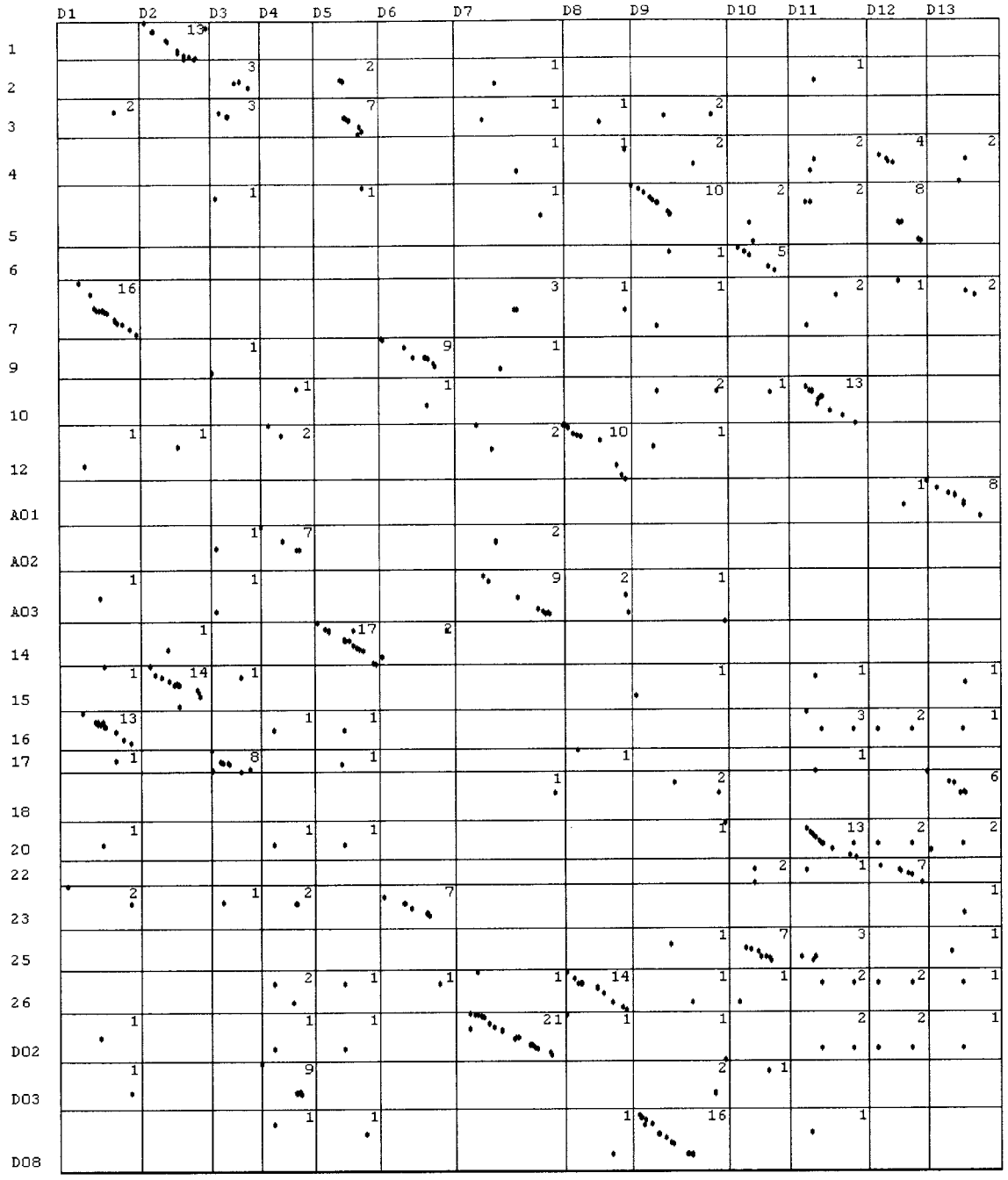

FIGURE 4.—Oxford plot showing tetraploid (AtDt)-diploid (D) genome correspondence. Each dot represents a probe that is mapped in both tetraploid and diploid cotton genomes, with *x*- and *y*-axes representing chromosomal locations. The number at the top right in each cell represents the total number of corresponding loci for the pair of chromosomes. Some points represent multiple hits with the same genetic location. For further detail and a listing of data points in each cell, see http://www.plantgenome.uga.edu.

D02, and the remaining 36 scattered over 17 chromosomes. However, close study of nonhomeologous duplications reveals that 6 were duplicated on LG A03, 5 on chromosome 7, and 5 on chromosome 26. On the basis of the Poisson distribution test, these numbers are significantly higher than expected to occur by chance. On the tetraploid map, 19 pairs of chromosomes shared 4 or more common loci, including 6 on At, 10 on both At and Dt, and 3 on Dt. In addition, 10 individual chromosomes contained 4 or more pairs of duplicate probes (Figure 4).

A total of 82 probes (12.4%) detected 137 pairs of duplicate loci in the diploid D genome (Figure 5; see also clickable annotated version at http://www.plantgenome.

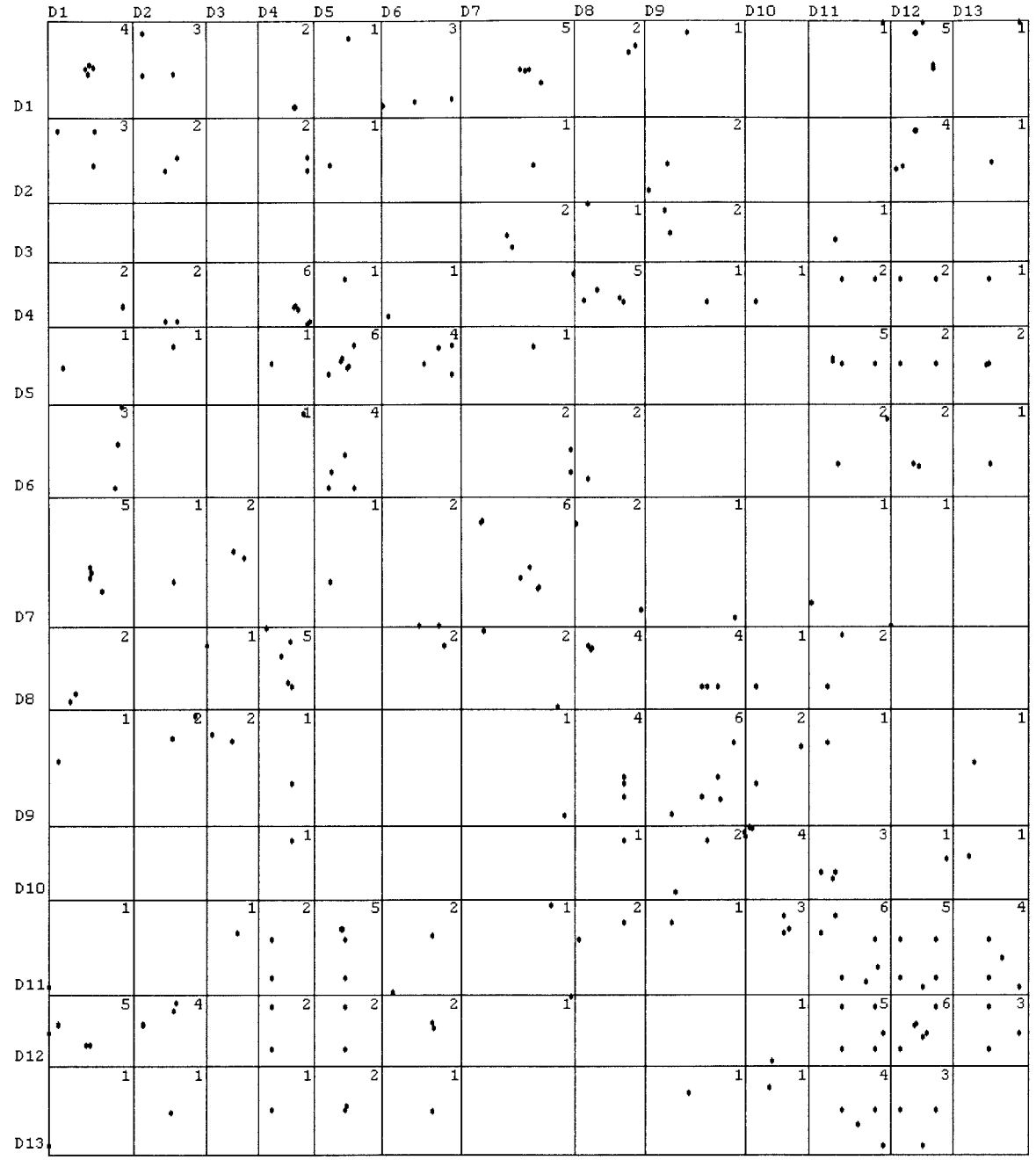

Figure 5.—Oxford plot showing intragenomic correspondence in the D genome. Each dot represents a probe that is mapped multiple times in the diploid D genome, with the *x*- and *y*-axes representing chromosomal locations. The number at the top right in each cell represents the total number of corresponding loci for the pair of chromosomes. Some points represent multiple hits within the same genetic location. For further detail and a listing of data points in each cell, see http://www.plantgenome.uga.edu.

uga.edu/cottonmap.htm) despite that genome having a smaller total number of mapped loci (763) than the tetraploid. The frequency of mapped duplicated loci is higher in the D diploid map than in either the At or the Dt subgenomes, due to a generally higher level of DNA polymorphism in the D diploid cross (Brubaker *et al.* 1999).

The D-genome duplicate loci, like the intrasubgenomic duplications in the tetraploid, were not randomly distributed (Figure 5). Nine pairs of chromosomes each share four or five common probes, significantly higher than the random expectation. In addition, another 29 pairs of chromosomes contain two or three common markers each, most of which are closely distributed on the chromosomes. As a result, each chromosome can be divided into several blocks according to the clustering of

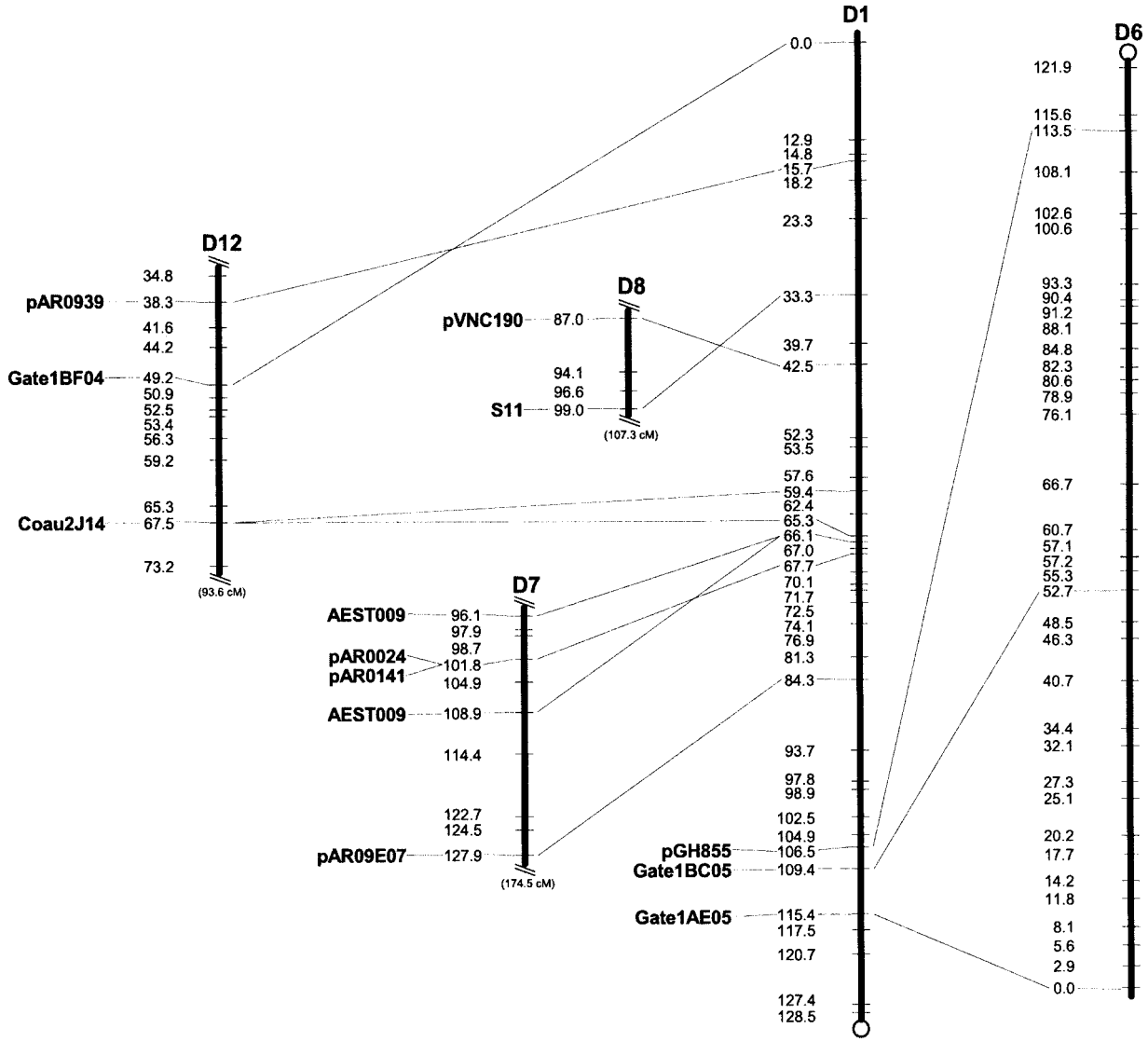

FIGURE 6.—Examples of possible ancient homeologous relationships between different segments of the diploid D genome. Duplicated loci are connected with lines.

duplicate loci. Frequently, the blocks on one chromosome show a syntenic relationship with blocks on other chromosomes. For example, D1 can be divided into four regions, the top one corresponding with the middle region of D12, the middle one with the middle regions of D8 and D7, and the bottom one with D6 (Figure 6). Although a clear syntenic relationship was observed in these cases, the linear order of the duplicated loci between homeologous blocks was not identical, especially regarding the top region of D1. In this case, the differences in linear order can be explained by the inversion of the top region (Gate1BF04-Coau2J14) of this chromosome.

Conservation of intragenomic duplication also can be detected in some regions of the tetraploid genome. A typical example was the high-marker-density region on the corresponding regions of chromosome 7, LG A03, and D02 (Figure 7). In diploid cotton, four probes (AEST009, pAR0141, pAR0024, and pAR09E07) located on the region of D1 spanning 18.2 cM detected five loci on D7 spanning 31.8 cM. In the corresponding region of the tetraploid map, the counterparts of four probes (pAR0024, pAR0319, Gate3CH01, and Unig25H02) from 98.5 to 117.8 cM of LG D02 were concentrated in the region from 102.4 to 119.8 cM of chromosome 7. Four probes (G1045, pAR0024, Gate3BB10, and Unig22B11) mapped on LG A03 from 96.0 to 144.5 cM were also located in this region of chromosome 7, except for Unig26B04 that was 36.8 cM away from the closest duplicate marker on LG A03 and 57.5 cM on chromosome 7. All this information shows that these high-marker-density regions are homeologous fragments among the chromosomes D1, D7, 7, A03, and D02, and they retain much similarity since divergence from the same ancestor.

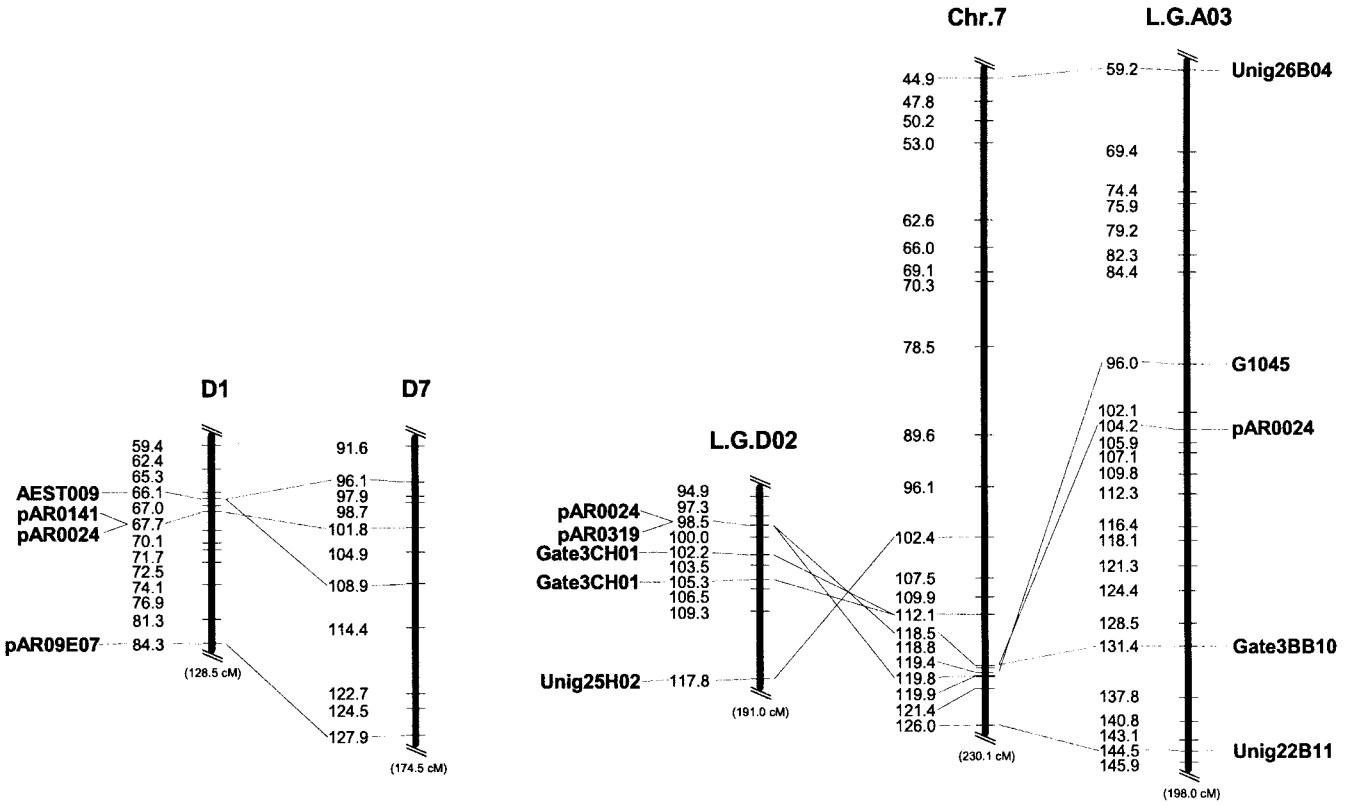

FIGURE 7.—Conservation of duplicated segments among chromosomes of D1, D7, chromosome 7, D02, and A03. Duplicated loci were connected with lines.

Some probes have duplicated loci on the same chromosome. In the tetraploid map, 73 probes, or 11.0% of the total number of duplicated probes, were duplicated on the same chromosome, with as many as seven duplications on a single chromosome (chromosome 18). In the diploid map, 26 (18.8%) of the duplicated probes show two or more loci on the same chromosome, with up to three on a single chromosome. In both cases, the level of intrachromosomal duplication is higher than the random expectations of 4 and 8.25% for the AD and Dt genomes, respectively. Many of these duplications were present in tandem in all three genomes, suggesting an origin that predates the divergence of the A and D genomes.

The frequency of duplicated loci generally increased with locus number per chromosome ($r = 0.829$) and linkage group length ($r = 0.811$). A few exceptions to this trend were found. For example, chromosome 22, the second shortest chromosome (63 loci, 95.4 cM) in the tetraploid genome, has the highest frequency (27, or 42.9%) of duplicate loci, with corresponding loci distributed over 16 nonhomeologous chromosomes, not including 27 on homeologous chromosomes 4 and 5.

Marker sequence annotation: Multiple local alignment searches using the programs blastn and tblastx were used for sequence annotation against publicly available databases of the National Center for Biotechnology Information (NCBI) as of April 20, 2003. The default matrix BLOSUM 62 and a cutoff of 1×10^{-6} were used in all BLAST searches. The NCBI database was subdivided into several taxon-specific groups to allow for the efficient determination of not only the best overall match, but also the best match among closely related species, excluding unannotated EST and genomic survey sequence database entries. Additional analyses included the use of hidden Markov models to classify sequence data by protein sequence signature. The program InterProScan (ZDOBNOV *et al.* 2002) was used to search and compare the translated cotton sequences against several protein databases (Pfam, SMART, and ProDom) and genome ontology (ASHBURNER *et al.* 2001) numbers for each of these classifications were obtained. The results of these analyses revealed that the 1983 cotton query sequences (including 1884 sequences of probes mapped as RFLPs and 99 sequences mapped as SSRs) could be classified into 515 distinct protein families and 138 different functional molecular groups. The most common molecular functional groups were ATP binding (GO: 0005524; 68 hits) and the structural constituent of ribosome (GO: 0003735; 48 hits). Results of sequence similarity analyses are available at http://www.plantgenome.uga.edu/cottonmap.htm.

DISCUSSION

This genetically anchored set of sequence-tagged sites, composed of 3347 loci (2584 on the tetraploid map and 763 on the diploid map), provides transferable genetic markers suitable for a wide range of investigations in structural, functional, and evolutionary genomics. A genetically anchored STS framework provides a foundation for physical mapping and ultimately for assembling a robust finished sequence of the cotton genome. The present tetraploid map equates to an average interval of ~606 kb between genetic markers on the basis of a consensus estimate of genome size of ~2700 Mbp. The map is presently being used to anchor high-coverage bacterial artificial chromosome (BAC) libraries for *G. hirsutum*, *G. barbadense*, and *G. raimondii* on the basis of hybridizing genetically mapped probes to the BACs (see http://www.plantgenome.uga.edu). The *G. raimondii* BAC library is slated for fingerprinting to 10× depth, permitting the resulting "contigs" to be extended further. By selective BAC end sequencing, a robust, genetically anchored physical map is expected to coalesce.

Although the map was mostly created using the RFLP method and has been applied to several goals by this technology (*e.g.*, Jiang *et al.* 1998, 2000a,b; Wright *et al.* 1998, 1999; Jiang and Chee *et al.* 2000; Saranga *et al.* 2001; Paterson *et al.* 2002), STS markers are amenable to a variety of detection systems, increasing the value of these loci and reducing the costs associated with their wider utilization. Survey of the cotton genome with a few SSR motifs shows that there appears to be an ample store of additional SSR loci to be characterized. A total of 124 simple-sequence repeat loci were mapped directly, and another 312 mapped RFLP loci contain SSRs of either six or more dinucleotide repeats or 15 or more base pairs of longer repeat units. However, the incremental advantage of SSRs in cotton appears to be smaller than that in many taxa. In a small sampling of 10 *G. hirsutum* cultivars from diverse U.S. production regions, variation was detected by 6 (46%) of 13 CMS-SSRs and by 11 (28%) of 39 randomly chosen RFLPs (data not shown). SSRs revealed an average of 1.64 different DNA fingerprints per primer pair, while RFLPs yielded an average of 1.28 different DNA fingerprints per probe.

Rates of intraspecific DNA sequence variation within amplicons derived from the genetically mapped sequence-tagged sites are somewhat higher than the (typically ~0.1%) rates found in human or other taxa, indicating that SNP discovery and application in cotton is feasible. Several approaches can be used to discover and detect SNPs in the STS loci. The scope of SNP detection systems has been extensively reviewed elsewhere. Many labs with a need for small numbers of specific markers may find the CAPS method attractive (Konieczny and Ausubel 1993). A small sampling of 39 primer pairs derived from EST sequences deposited in GenBank (designated EST) were empirically screened by digestion with restriction enzymes, identifying nine mappable loci (P. Chee, J. Rong and A. H. Paterson, unpublished results). Constraints due to the modest levels of polymorphism in cotton can be countered partly by the availability of many closely linked STSs to evaluate.

Nonrandom patterns of DNA marker distribution provide clues regarding interesting and important features of cotton genome organization. On most chromosomes, at least one significant concentration of loci occurs, possibly corresponding to the centromeric regions. Virtually all marker-rich regions corresponded between the D and Dt genomes, and most also corresponded with the At genome, suggesting that these may be the locations of many of the cotton centromeres. In several cases, the breakpoints of structural rearrangements between the A and D subgenomes locate squarely in these regions, consistent with the widespread observation that chromosomal inversion breakpoints often lie at or near centromeres. A total of three marker-rich regions are unique to Dt and nine are unique to At, generally consistent with the much larger quantity of repetitive DNA in the A genome (Zhao *et al.* 1998); we are investigating this relationship further by evaluating the BAC hybridization patterns of mapped RFLPs and members of abundant repetitive DNA families. Marker-poor regions showed little correspondence and in the A genome occurred only at the false-positive level, but did seem to be real in the Dt genome. These clues await more information about cotton genome organization to unravel their significance, if any.

Patterns of DNA marker duplication in the tetraploid cotton genome are especially important. The detailed delineation of homeologous relationships among the chromosomes of tetraploid cotton and between the chromosomes of tetraploids and diploids is important from both a basic and an applied standpoint. Tetraploid cotton, containing At and Dt subgenomes, was derived from a naturally occurring cross between two diploids with A and D genomes, respectively, ~1–2 million years ago (Wendel 1989; Senchina *et al.* 2003; Wendel and Cronn 2003). We found many duplicated loci (42.6% of the total duplicated probes) distributed on pairs of chromosomes in similar linear orders, strongly supporting the inference that these pairs of chromosomes are derived from common ancestors. On the basis of distribution of duplicated loci, all 13 expected homeologous groups have been identified. The counterparts of most probes (67.6%, or 278/411) on the diploid map are also mapped on tetraploid cotton, and their arrangements further support the assemblages of tetraploid homeologous groups. Comparison of diploid and tetraploid linkage maps reveal close correspondence of marker order among Dt and D chromosomes, with most exceptions explicable by inversion. However, three Dt or D chromosomes (chromosome 14 or D5, chromosome 22 or D12, LG D08 or D9) each have syntenic relationships with two At chromosomes, as a result of reciprocal translocation

between chromosomes 4 and 5 and between chromosomes 2 and 3 after divergence of At and Dt. These structural changes must have occurred since the formation of New World tetraploid cotton, because the sytenic and linear order is the same between diploid A genome and diploid D genome (Brubaker *et al.* 1999).

In addition to the homeologous duplications, many sequences are also duplicated within each subgenome. While the observed proximal (intrachromosomal) bias of intragenomic duplication suggests that some of it may be accounted for by retrotransposition, we suggest that the majority of intragenomic duplication may have another basis. In the diploid D genome, 12.4% of probes detected multiple polymorphic loci, which appear to fall into syntenic and even colinear regions. This finding, together with the recent report of intragenomic similarities in Giemsa-banding patterns (Muravenko *et al.* 1998), transforms prior hints of an ancient chromosomal duplication that predated the A-D divergence into genetic mapping evidence and also supports the classical notion that present-day cotton may be derived from a putative ancestor containing six or seven chromosomes. Although a high frequency of duplicated loci was also detected within each subgenome of tetraploid cotton, and some syntenic regions between nonhomeologous chromosomes were found (Figures 3 and 7), in most cases the patterns were not as clear as in the D genome, probably reflecting lower polymorphism rates that permitted mapping of duplicated loci in the tetraploid less frequently. Early hints suggest that chromatin duplication may have been important to the evolution of agriculturally productive cotton (Jiang *et al.* 1998; Saranga *et al.* 2001), a topic of high priority for further study.

We thank many colleagues in the Paterson lab for technical and moral support and David Stelly for helpful comments. This work was funded by the U.S. Department of Agriculture (91-37300-6570; 97-35300-5305), the National Science Foundation (DBI-9872630; 0211700), Georgia and Texas Cotton Commissions, and Georgia and Texas Agricultural Experiment Stations.

LITERATURE CITED

Ashburner, M., C. A. Ball, J. A. Blake, H. Butler, J. M. Cherry *et al.*, 2001 Creating the gene ontology resource: design and implementation. Genome Res. **11:** 1425–1433.

Beasley, J. O., 1940 The production of polyploids in *Gossypium*. J. Hered. **31:** 39–48.

Bowers, J. E., C. Abbey, S. Anderson, C. Chang, X. Draye *et al.*, 2003 A high-density genetic recombination map of sequence-tagged sites for Sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics **165:** 367–386.

Brubaker, C. L., A. H. Paterson and J. F. Wendel, 1999 Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome **42:** 184–203.

Edwards, G. A., and M. A. Mirza, 1979 Genomes of the Australian wild species of cotton. II. The designation of a new G genome for *Gossypium bickii*. Can. J. Genet. Cytol. **21:** 367–372.

Endrizzi, J., E. L. Turcotte and R. J. Kohel, 1984 Qualitative genetics, cytology, and cytogenetics, pp. 81–129 in *Cotton*, edited by R. J. Kohel and C. F. Lewis. ASA/CSSA/SSSA, Madison, WI.

Endrizzi, J., E. L. Turcotte and R. J. Kohel, 1985 Genetics, cytology, and evolution of Gossypium. Adv. Genet. **23:** 272–375.

Ewing, B., L. Hillier, M. C. Wendl and P. Green, 1998 Base-calling of automated sequencer traces using *Phred*. I. Accuracy assessment. Genome Res. **8:** 175–185.

Fryxell, P. A., 1979 *The Natural History of the Cotton Tribe*. Texas A&M University Press, College Station, TX.

Fryxell, P. A., 1992 A revised taxonomic interpretation of Gossypium L. (Malvaceae). Rheedea **2:** 108–165.

Galau, G. A., and T. A. Wilkins, 1989 Alloplasmic male-sterility in AD allotetraploid *Gossypium hirsutum* upon replacement of its resident cytoplasm with that of D-species *Gossypium harknessii*. Theor. Appl. Genet. **78:** 23–30.

Harland, S. C., 1929 The genetics of cotton. I. The inheritance of petal spot in New World cottons. J. Genet. **20:** 365–385.

Hutchinson, J. B., and R. A. Silow, 1939 Gene symbols for use in cotton genetics. J. Hered. **30:** 461–464.

Jiang, C., R. Wright, K. El-Zik and A. Paterson, 1998 Polyploid formation created unique avenues for response to selection in *Gossypium* (cotton). Proc. Natl. Acad. Sci. USA **95:** 4419–4424.

Jiang, C., P. Chee, X. Draye, P. Morrell, C. Smith *et al.*, 2000a Multi-locus interactions restrict gene flow in advanced-generation interspecific populations of polyploid *Gossypium* (cotton). Evolution **54:** 798–814.

Jiang, C., R. Wright, S. Woo, T. Delmonte and A. Paterson, 2000b QTL analysis of leaf morphology in tetraploid *Gossypium* (cotton). Theor. Appl. Genet. **100:** 409–418.

Kim, J. K., and B. A. Triplett, 2001 Cotton fiber growth in planta and *in vitro*. Models for plant cell elongation and cell wall biogenesis. Plant Physiol. **127:** 1361–1366.

Kimber, G., 1961 Basis of the diploid-like meiotic behavior of polyploid cotton. Nature **191:** 98–99.

Konieczny, A., and F. M. Ausubel, 1993 A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. Plant J. **4:** 403–410.

Kosambi, D., 1944 The estimation of map distance from recombination values. Ann. Eugen. **12:** 172–175.

Lacape, J., M. Giband, T. Nguyen, B. Courtois and B. Hau, 2002 Overview of activities and major achievements in molecular genetics at CIRAD/France. Cotton Sci. **14** (Suppl.): 14.

Lander, E. S., and P. Green, 1987 Construction of multilocus genetic-linkage maps in humans. Proc. Natl. Acad. Sci. USA **84:** 2363–2367.

Lee, J., 1984 Cotton as a world crop, pp. 6–24 in *Cotton*, edited by R. J. Kohel and C. F. Lewis. ASA/CSSA/SSSA, Madison, WI.

Moulherat, C., M. Tengberg, J.-F. Haquet and B. Mille, 2002 First evidence of cotton at neolithic Mehrgarh, Pakistan: analysis of mineralized fibres from a copper bead. J. Archaeol. Sci. **29:** 1393–1401.

Muravenko, O. V., A. R. Fedotov, E. O. Punina, L. I. Fedorova, V. G. Grif *et al.*, 1998 Comparison of chromosome BrdU-Hoechst-Giemsa banding patterns of the A(1) and (AD)(2) genomes of cotton. Genome **41:** 616–625.

Paterson, A., J. Estill, J. Rong, D. Williams and B. Marler, 2002 Toward a genetically-anchored physical map of the cotton genomes. Cotton Sci. **14** (Suppl.): 31.

Phillips, L. L., and M. A. Strickland, 1966 The cytology of a hybrid between *Gossypium hirsutum* and *G. longicalyx*. Can. J. Genet. Cytol. **8:** 91–95.

Reinisch, A., J.-M. Dong, C. Brubaker, D. Stelly, J. Wendel *et al.*, 1994 A detailed RFLP map of cotton (*Gossypium hirsutum* × *G. barbadense*): chromosome organization and evolution in a disomic polyploid genome. Genetics **138:** 829–847.

Rhyne, C. L., and J. C. Carter, 1991 Anthocyaninless cotton, pp. 540–541 in *Proceedings of the Beltwide Cotton Products Research Conference*, edited by D. Richter. National Cotton Council of America, Memphis, TN.

Saranga, Y., M. Menz, C. Jiang, R. Wright, D. Yakir *et al.*, 2001 Genomic dissection of genotype × environment interactions conferring adaptation of cotton to arid conditions. Genome Res. **11:** 1988–1995.

Senchina, D. S., I. Alvarez, R. C. Cronn, B. Liu, J. Rong *et al.*, 2003 Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. Mol. Biol. Evol. **20:** 633–643.

Stephens, S. G., 1967 Evolution under domestication of the new world cottons (*Gossypium* spp.). Cienc. Cult. **19:** 118–134.

Stephens, S. G., 1974 Geographic and taxonomic distribution of anthocyaninless genes in New World cottons. J. Genet. **61:** 128–141.

Stephens, S. G., and M. E. Moseley, 1974 Early domesticated cottons from archaeological sites in central coastal Peru. Am. Antiq. **39:** 109–122.

Stewart, J. McD., 1994 Potential for crop improvement with exotic germplasm and genetic engineering, pp. 13–327 in *Challenging the Future: Proceedings of the World Cotton Research Conference-1*, Brisbane, Australia, February 14–17, edited by G. A. Constable and N. W. Forrester. CSIRO, Melbourne.

Wendel, J. F., 1989 New world cottons contain Old World cytoplasm. Proc. Natl. Acad. Sci. USA **86:** 4132–4136.

Wendel, J. F., and V. A. Albert, 1992 Phylogenetics of the Cotton genus (Gossypium): character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and biogeographic implications. Syst. Bot. **17:** 115–143.

Wendel, J. F. and R. C. Cronn, 2003 Polyploidy and the evolutionary history of cotton. Adv. Agron. **78:** 139–186.

Wendel, J. F., R. C. Cronn, J. S. Johnston and H. J. Price, 2002 Feast and famine in plant genomes. Genetica **115:** 37–47.

Wright, R. J., 1999 Toward high resolution mapping of genes conferring resistance in cotton to the bacterial blight pathogen. Ph.D. Thesis, Texas A&M University, College Station, TX.

Wright, R., P. Thaxton, K. El-Zik and A. Paterson, 1998 D-subgenome bias of Xcm resistance genes in tetraploid Gossypium (cotton) suggests that polyploid formation has created novel avenues for evolution. Genetics **149:** 1987–1996.

Wright, R., P. Thaxton, A. Paterson and K. El-Zik, 1999 Molecular mapping of genes affecting pubescence of cotton. J. Hered. **90:** 215–219.

Zdobnov, E. M., and R. Apweiler, 2002 InterProScan: an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17:** 847–848.

Zhang, J., W. Guo and T. Zhang, 2002 Construction of molecular linkage map of cultivated allotetraploid cotton (*Gossypium hirsutum* L. X *G. barbadense* L.) with SSR and RAPD Markers. Cotton Sci. **14** (Suppl.): 16.

Zhao, X., Y. Si, R. E. Hanson, C. F. Crane, H. J. Price *et al.*, 1998 Dispersed repetitive DNA has spread to new genomes since polyploidy formation in cotton. Genome Res. **8:** 479–492.

Communicating editor: J. A. Birchler