Meta-analysis of Polyploid Cotton QTLs Shows Unequal Contributions of Subgenomes to a Complex Network of Genes and Gene Clusters Implicated in Lint Fiber Development

Junkang Rong*, F. Alex Feltus*, Vijay N. Waghmare‡, Gary J. Pierce*, Peng W. Chee †, Xavier Draye§, Yehoshua Saranga**, Robert J. Wright††, Thea A. Wilkins‡‡, O. Lloyd May†, C. Wayne Smith‡‡, John R. Gannaway§§, Jonathan F. Wendel***, Andrew H. Paterson*1

* Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602
† Coastal Plains Experiment Station, University of Georgia, Athens, GA 30602
‡ Division of Crop Improvement, CICR, Nagpur, India
§ Université catholique de Louvain Place Croix du Sud 2/11 - 1348 Louvain-la-Neuve, Belgium
** The RH Smith Institute of Plant Science and Genetics in Agriculture, The Hebrew University of Jerusalem, Rehovot 76100, Israel.
†† Department of Plant and Soil Science, Texas Tech University, Lubbock, TX 79409
‡‡ Department of Soil and Crop Science, Texas A&M University, College Station, TX 77843
§§ Texas A&M Research and Extension Center, Lubbock TX 79401
*** Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011
Running title: Meta-analysis of cotton fiber QTLs

Key words: genetic map, candidate gene, mutant, synteny and/or colinearity

1 Communicating author: Andrew H. Paterson

Plant Genome Mapping Laboratory
University of Georgia
111 Riverbend Road, Rm 228
Athens, GA 30602
Phone 1-706-583-0162;
Fax 1-706-583-0160;
Email paterson@uga.edu
ABSTRACT

QTL mapping experiments yield heterogeneous results due to the use of different genotypes, environments, and sampling variation. Compilation of QTL mapping results yields a more complete picture of the genetic control of a trait, and reveals patterns in organization of trait variation. A total of 432 QTLs mapped in one diploid and 10 tetraploid interspecific cotton populations were aligned using a reference map and depicted in a CMap resource. Early demonstrations that genes from the non-fiber-producing diploid ancestor contribute to tetraploid lint fiber genetics gain further support from multiple populations and environments, and advanced-generation studies detecting QTLs of small phenotypic effect. Both tetraploid subgenomes contribute QTLs at largely non-homoeologous locations, suggesting divergent selection acting on many corresponding genes before and/or after polyploid formation. QTL correspondence across studies was only modest, suggesting that additional QTLs for the target traits remain to be discovered. Crosses between closely-related genotypes differing by single-gene mutants yield profoundly different QTL landscapes, suggesting that fiber variation involves a complex network of interacting genes. Members of the lint fiber development network appear clustered, with cluster members showing heterogeneous phenotypic effects. Meta-analysis linked to synteny-based and expression-based information provides clues about specific genes and families involved in QTL networks.
Most naturally occurring genetic variation in populations reflects polymorphic alleles that individually have relatively small effects but collectively result in continuous variation among members of the population. Through genetic mapping, the number and location of loci associated with complex trait variation, i.e. quantitative trait loci or QTLs, can be estimated and used to infer the genetic basis of traits that differ between varieties and/or species (Paterson et al. 1988). DNA markers linked to QTLs can also be used as diagnostic tools in the selection of desirable genotypes (marker assisted selection) and as a starting point for cloning of QTLs. For these reasons, vast numbers of QTLs representing a myriad of traits have been mapped in agronomically important crops, and also in botanical models and animals. A handful of genes underlying QTLs have been cloned (eg., Frary et al. 2000) based largely on fine mapping (Paterson et al. 1990).

A recurring complication in the use of QTL data is that different parental combinations and/or experiments conducted in different environments often result in identification of partly or wholly non-overlapping sets of QTLs. The majority of such differences in the QTL landscape are presumed to be due to environment-sensitivity of genes. The use of stringent statistical thresholds to infer QTLs while controlling experimentwise error rates (Churchill and Doerge 1994; Lander and Botstein 1989) imply that only a small fraction of these non-overlapping QTLs can be attributed to false-positive results. Small QTLs with opposite phenotypic effects might occasionally be closely linked in coupling in early-generation populations, and separated only in advanced-generation populations after additional recombination.

Comparison of multiple QTL mapping experiments by alignment to a common reference map offers a more complete picture of the genetic control of a trait than can be obtained in any
one study. One such trait, the genetic control of variation in growth and development of seedborne epidermal ‘lint’ fibers, is a natural priority in cotton genome analysis. All 50 *Gossypium* species have seedborne epidermal trichomes, often referred to as ‘fuzz’ fibers. ‘A’ genome diploid cottons are distinct from their sister ‘F’ and ‘B’ genomes in that both wild and cultivated forms have longer ‘lint’ fibers with secondary thickening, that are spinnable – this feature can thus be inferred to have evolved after the divergence of the A, B, and F genomes from a common ancestor about 5-7 million years ago (WENDEL 1989). Some view ‘lint’ as distinct from ‘fuzz’ fibers based on the secondary thickening, while others consider view the two as extremes in a continuum. However, all agree that economically-important lint fibers are much longer, reaching lengths of about 20 mm in cultivated A-genome diploids *G. herbaceum* and *G. arboreum*. Wild tetraploid cottons have fibers resembling those of wild diploids, however domestication and scientific improvement of AD tetraploids *G. hirsutum* and *G. barbadense* (from different tetraploid clades), have increased fiber length to as much as 40 mm or more, and also improved other aspects of ‘fiber quality’ (strength, fineness, and elongation) as well as higher yield, even in environments to which both have been adapted by scientific plant breeding (JIANG et al. 1998).

Accurate measurement of cotton fiber quality requires complex instrumentation and therefore has been a high priority trait for which to establish diagnostic DNA markers, in view of the fact that cotton provides about 45% of the raw material for the ~$500 billion/yr worldwide textile industry. Cotton fiber quality is a complex characteristic with many components, affecting both the efficacy of processing by textile spinning machinery, as well as the comfort attributes associated with cotton fabrics. To date, more than 400 QTLs related to various aspects of fiber quality have been mapped using different kinds of populations including F2 (JIANG et al. 1998;
Meta-analysis of cotton fiber QTLs promises to contribute to our understanding of fundamental questions, as well as to expedite crop improvement. \textit{Gossypium} is an especially good model in which to advance understanding of the consequences of polyploidy in that both derived tetraploids and their ancestral diploids (or at least very close relatives) are known and remain extant, yet the polyploidization event is sufficiently old [about 1-2 million yrs: (SENCHINA \textit{et al.} 2003)] for adaptive evolution to have occurred. One particularly exciting opportunity is to shed light on why tetraploid cottons consistently have higher yield and quality than the modern descendents of their diploid progenitors, even in environments to which both have been adapted by scientific plant breeding (JIANG \textit{et al.} 1998).

Research into the genetic control of cotton fiber development may also benefit from progress in understanding the growth and development of hair-bearing epidermal cells (trichomes) in \textit{Arabidopsis}. Indeed, \textit{Gossypium} and \textit{Arabidopsis} are thought to have shared common ancestry about 83-86 million years ago (BENTON 1993), and cotton may be the best crop outside of the Brassicales in which to employ ‘translational genomics’ from \textit{Arabidopsis}.

populations. All QTLs were also projected onto a consensus map which was inferred to resemble the DNA marker arrangement of the hypothetical ancestor of the two ‘subgenomes’ of tetraploid cotton (Rong et al. 2005a). The consensus map has improved our ability to deduce cotton-Arabidopsis synteny relationships, and thus fosters study of correspondence between the cotton QTLs and fiber or trichome-related Arabidopsis genes. To encourage further utilization and online community access to these data, a CMap resource was developed and can be accessed at our website.

MATERIALS AND METHODS

**Populations:** QTL and mutant data used in this study were published previously except for those from populations segregating for n2 and im mutants (see below), as described in Table 1.

**QTL mapping of n2 and im populations:** Two new F2 populations derived from crosses between G. barbadense var. Pima S-7 and G. hirsutum cv. Texas Marker-1 (TM-1) isogenic lines with n2 and im mutants, respectively, were mapped.

A total of 124 F2 plants from each of n2 and im crosses were grown in a field at the Coastal Plain Experiment Station, Tifton, GA during the 2002 season. Five traits reflecting fiber quality parameters were determined by the Cotton Incorporated Textile Services Laboratory (Cotton Incorporated, Cary, N.C.) using the High-Volume Precision Instrument (HVI), including upper-half mean length (HVuhm), percentage short fiber by weight (HVsfc), uniformity index (HVui), fiber strength (FS), fiber elongation (ELO), and micronaire (Mic). In addition, two fiber yield components, lint percentage, and lint index, were measured as reported (Rong et al. 2005b).
DNA markers were selected from a high density reference map (RONG et al. 2004). Procedures for linkage analysis were reported previously (RONG et al. 2004). Composite interval mapping (ZENG 1994) was performed using Windows QTL Cartographer (WANG et al. 2005)(http://statgen.ncsu.edu/qtlcart/WQTLCart.htm). A stringent LR threshold of 13.8 (equivalent to LOD score 3.0) was used to declare significant QTLs in order to keep the experiment-wise likelihood of even one false positive below 5% in the large genome of cotton (JIANG et al. 1998). Effects and percent of phenotypic variance (PV) explained by single QTL (R²) were estimated with Windows QTL Cartographer 2.0 at likelihood peaks.

**Integration of QTLs from different populations:** For convenience of description, three names were used for different types of genetic maps: individual map, reference map (RONG et al. 2004) and consensus map (RONG et al. 2005a). Individual maps were the original QTL maps summarized in Table 1. To align QTLs from different individual maps, the reference map was used, from which subsets of informative markers were drawn to make the individual maps. Finally, some analyses also used a ‘consensus map’ that depicts the marker arrangement along the genome of a hypothetical common ancestor that gave rise to the diploid progenitors of polyploid cotton, inferred as described elsewhere (RONG et al. 2005a). Most of the QTL locations in reference and consensus maps were determined by BioMERCARTOR (Version 2) (ARCADE et al. 2004). QTLs in regions where the marker order was different between the individual map and reference map were plotted based on the location of the marker nearest to the likelihood peak on the reference map. The individual maps were made over several years, during which changes in the reference map occurred – marker and chromosome nomenclature was all synchronized with that of the current reference map (RONG et al. 2004). Redundant probes,
identified as detailed elsewhere (Rong et al. 2005a), were removed. Nomenclature of QTLs in the reference map was standardized to the system reported (Chee et al. 2005a; Chee et al. 2005b; Draye et al. 2005), that is, trait name first, then the name of the chromosome on which the QTL was located, followed by a number indicating the loci on this chromosome. If two QTLs were reported by (Chee et al. 2005a; Chee et al. 2005b) and (Draye et al. 2005) on a chromosome, additional QTLs for the same trait mapped in the same chromosome of other individual maps will be given a number with two digits to represent their relative locations. For example, because two QTLs for fiber length were mapped on linkage group D08 by (Chee et al. 2005b) and named FLD08.1 and FLD08.2 respectively, another QTL mapped on the same chromosome was named FLD08.15 representing the relative location of its likelihood peak between FLD08.1 and FLD08.2. The original name for each QTL used in each map, and the name in CMap of each trait as well as their category and source were summarized in Supplemental Table 9.

CMap display of genetic maps: CMap v0.16 was downloaded from the Generic Components for Model Organism Database project (GMOD; http://www.gmod.org). Genetic mapping data were prepared in spreadsheets and imported into CMap.

RESULTS

A CMap visualization tool for meta-analysis: To streamline comparisons of QTL and other genomic data among chromosomes, subgenomes, populations, and syntenic locations in Arabidopsis, we developed a cotton CMap resource (publicly available at www.plantgenome.uga.edu/cmap). Using BioMERCARTOR (Arcaide et al. 2004), a total of
432 QTLs from 11 individual maps (Table 1) were aligned to the reference map (RONG et al. 2004), and to the consensus map representing inferred gene (marker) arrangements along the chromosomes of a common ancestor of the A and D genomes (RONG et al. 2005a). Among them, 224 QTLs from five populations were related to fiber quality, 64 from four populations were related to yield components, 63 from two populations (one tetraploid and one diploid A genome) were related to flower morphology, 33 from one population were related to drought response, 32 from two populations were related to leaf morphology, 8 from 4 populations were related to trichome density, 7 from 4 populations were related to bacterial blight resistance, and one from one population was related to earliness. Available from the CMap resource are individual QTL maps, the reference map and the consensus map, plus relationships of cotton chromosomes to inferred gene orders for a hypothetical common ancestor of cotton and Arabidopsis (RONG et al. 2005a) based on correspondence inferred using CSII (LEVINE 2002) and FISH (CALABRESE et al. 2003).

In an example of the visualizations possible with this resource (Figure 1), a QTL on one chromosome (Chr.16) of an individual population (n2) is aligned to the corresponding ‘chromosome’ of the consensus map, also indicating corresponding regions of Arabidopsis and showing all QTLs known from all populations for the region shown. Thus, the CMap resource expedites the comparison of QTLs between different homoeologous chromosomes as well as between cotton QTLs and Arabidopsis orthologs in the syntenic regions.

**Cotton fiber QTLs are enriched in the D subgenome:** More QTLs were detected in the tetraploid D (hereafter Dt) than in the tetraploid A (hereafter At) subgenome, specifically 221 vs 184 (Table 1). Of special interest were QTLs related to lint fiber quality characteristics, in that
the At progenitor produces spinnable fiber but the Dt progenitor does not. We focused this analysis on fiber quality characteristics because of their relatively high heritability (May and Wofford 2000) and economic importance. QTLs for fiber yield components from a small number of studies are also in the CMap database, but in view of the complexity and low heritability of this trait we considered that more extensive testing would be needed before a meaningful meta-analysis could be performed on these.

Based on three well-studied crosses for which both parents are elite cultivars or breeding lines (Table 2), the Dt subgenome contained 112 lint fiber related QTLs versus 84 in the At (Tables 2 and 3), a marginally significant difference (p=0.0455). The total of 196 genes and QTLs mapped in these elite crosses were further classified into the following trait categories: elongation (EL), fiber color (FC), fiber fineness (FF), fiber length (FL), fiber strength (FS), fiber uniformity (FU), micronaire (MIC) and short finer content (SF) (Table 3). All traits except FC and FL have more QTLs on Dt than on At, but the major contributors to the subgenome bias are fiber strength (FS) and micronaire (MIC), for which the Dt subgenome contains significantly more QTLs than does the At subgenome (p=0.041, p=0.005, respectively; Table 3). FF is controlled by the largest number of loci (57), followed by FL and EL (34, 33). SF and FC are controlled by the fewest loci, 8 and 11 respectively.

Cotton fiber QTLs are clustered: QTLs were non-randomly distributed across chromosomes and chromosomal regions (Supplemental table 1). Chromosome length (from the genetic map, in centiMorgans) had a significant relationship with the number of total QTLs and fiber QTLs in the Dt subgenome (r=0.62, p<0.01; and r=0.51, p<0.05, respectively), but not in the At (r=-0.11, r=0.15). As a result, although the Dt genome as a whole carries more QTLs for
fiber quality than does the At genome, the distributions of QTLs between different pairs of homoeologs varied widely. Seven Dt chromosomes or linkage groups (Chr. 14, 17, 20, 22, 23; L.G. D03 and D08) had more fiber-related QTLs than their homoeologous partners, with the largest difference between linkage group D08 (21 QTLs) and Chr.4/5 (8 QTLs). Five At chromosomes or linkage groups (Chr. 1, 6, 7; L.G. A01 and A03) have more fiber QTLs than their homoeologous counterparts, with the largest difference between Chr.06. (the shortest At chromosome, 9 QTLs), and Chr.25 (4 QTLs).

When each chromosome was divided into bins of 20 cM and the center of the QTL likelihood interval used as the QTL location, fiber-related QTLs were found to be clustered in some chromosomal regions (Fig. 2; Supplemental Table 2). Based on the total of 235 bins across the cotton genome and 196 fiber QTLs, the Poisson probability distribution function indicated that the presence of three or more QTLs in a bin was significantly (P<0.05) higher than the random expectation. A total of 87 (44.5% of) QTLs were located in 24 such bins (Supplemental Table 2). The 3 bins with the largest numbers of fiber QTLs (5 or 6) were on Chr.17 and linkage group D08 (2 bins). In no cases were statistically-significant QTL clusters found in corresponding regions of homoeologous chromosomes. These clusters are further elucidated below as examples of how integration of positional, comparative, expression and functional data might narrow the lists of candidate genes that may represent the QTLs.

**Homology and homoeology**: Integration of QTLs from different populations into a common map facilitates exploration of their allelic and homoeologous relationships, albeit at a level of resolution limited by comparative marker densities, variation in recombination rates in different crosses, variation in gene densities across the genome, and other factors. QTLs for the same trait
from different crosses, that had a likelihood peak within the same 20 cM bin, were considered potential alleles (albeit requiring further study to confirm). Considering the different fiber-related traits individually, with only a maximum of 57 QTLs per trait (FF), the occurrence of even 2 QTLs in the same bin was statistically unlikely to occur by chance (at \( p < 0.05 \)). A total of 4 such clusters were found, each comprising 2 QTLs for FF (3 cases) and EL (1 case) (Supplemental Table 3). The 8 QTLs comprise 4.1% of the total. All of these clustered QTLs were located in fiber QTL-rich bins identified above (bin 5 of Chr.02, bin 2 of Chr.12, bin 6 of Chr.15 and bin 8 of D08), although these bins collectively comprise only 3.6% of the genome. The two EL QTLs in bin 8 of D08 were detected in the same population (ELD08.3 and ELD08.4). However, the members of the 3 pairs of FF QTLs, respectively, were detected in different experiments.

Likewise, QTLs for the same trait were considered to be potentially homoeologous if their likelihood peaks fell within pairs of bins inferred to be homoeologous based on non-random alignments of duplicated DNA markers (Rong et al. 2004). Based on these criteria, a total of 41 (20.9%) QTLs from 6 fiber traits, were found to be potentially homoeologous (Supplemental Table 4). Fiber fineness was unusual in that 24 (42.1%) of 57 QTLs were potentially homoeologous. Figure 3 displays a middle region of consensus chromosome C12 with two potentially homoeologous loci for EL, that were originally mapped on Chr.05 by Paterson (EL05.1) (Paterson et al. 2003) and Chr.22 by Chee (EL22.3) (Chee et al. 2005a). This is an example of homoeologous loci for a fiber related trait detected in different experiments.

**Impact of mutants in fiber development on the QTL landscape:** We postulated that the QTL landscape might be altered by nonlinear interactions among unlinked genes in the fiber
development program. Most QTL mapping, however, has only modest power to resolve such epistasis. As a test case we explored how different discrete single-gene mutations alter the spectrum of QTLs explaining variation in a trait. Specifically, we compared the QTL maps resulting from crosses of one common parent, Pima S7, to two closely-related genetic stocks containing different discrete mutations implicated in lint fiber development.

\( n^2 \) is a recessive mutant expressing completely or partially naked seed, previously assigned to Chr.26 using aneuploid stocks (PERCY and KOHEL 1999), but recently suggested based on DNA markers to map to its homoeolog, Chr.12 (RONG et al. 2005b). A total of 242 probes detecting 370 loci mapped to 41 linkage groups covering 3116.2 cM of the \( n^2 \) population, at an average spacing of 8.42 cM.

\( im \) is a recessive mutant causing immature fiber, characterized by reduced secondary cell wall development and fiber weight reduction of 40% or more (KOHEL and MCMECHL 1990), assigned to Chr.03 based on aneuploid stocks (KOHEL et al. 2002). A total of 233 probes detecting 363 loci mapped to 37 linkage groups in the \( im \) population, at an average spacing of 9.6 cM. Both \( n^2 \) and \( im \) maps have similar marker orders and recombination distances, and are available in the CMap database.

Nine traits, including two lint yield components, seed weight, and six fiber quality components, were measured (see Methods). While all traits varied over a considerable range, no significant difference between the \( n^2 \) and \( im \) mapping populations was found for any measured trait (Supplemental Table 5).

A total of 22 QTLs were detected on 14 chromosomes in the \( n^2 \) population (Supplemental Table 6a). More than half of the QTLs were on only 4 chromosomes, with 4 on Chr.12, 3 on its homoeolog (Chr.26), 3 on Chr.17 and 2 on linkage group D08. Ten other chromosomes each
contained single QTLs (Supplemental Table 6a). Two QTLs for lint index each explain a remarkably high percentage of phenotypic variation (66.5%), with other QTLs ranging from 10.4% to 25.3% of phenotypic variance explained.

In the \textit{im} population, only 17 QTLs were detected for the same traits measured in \textit{n2}. One fiber strength QTL was detected in both populations (FS02.1 and FS02.2) in bin 5 of Chr.02. The other 16 QTLs were all located on different chromosomes than their \textit{n2} counterparts (Supplemental Table 6b). For example, no QTLs were detected in the \textit{im} population on chr.12, 17 and 26, which each had 3-4 QTLs in \textit{n2}. Chr.02 and A02, which show one or zero QTLs in the \textit{n2} population, had four and three QTLs in the \textit{im} population, respectively. One large-effect QTL (explaining 46.7% of phenotypic variation) was detected on linkage group D02 for FUD02.1(HVui) in the \textit{im} population, but no corresponding QTL was found in \textit{n2}. Likewise, the strong QTLs for lint index in the \textit{n2} population showed no corresponding effects in the \textit{im} population, although a relatively weak QTL was detected for this trait on Chr.14 of \textit{im}. While the two study populations were only of sufficient size to detect the largest-effect segregating QTLs, the use of stringent significance thresholds supports the validity of the QTLs that were claimed, and even these relatively large-effect QTLs did not correspond in the two populations. Since these two populations involved very similar genetic backgrounds tested in the same environment, the non-overlapping QTLs detected suggest that the differing qualitative mutations cause different pathways and/or networks to be limiting factors of fiber quality.

Although only one correspondence (FS02.1 and FS02.2) was found in fiber QTLs between the \textit{n2} and \textit{im} populations, considerable correspondence was found between the \textit{n2} or \textit{im} fiber QTLs and those detected in the three non-mutant populations introduced above. Seven QTLs found in the mutant populations map to bins containing QTLs for the same fiber traits from other
populations, increasing the number of non-random QTL clusters from 4 (based only on the 3 non-mutant populations) to 12 (including the QTLs from the mutant populations) (Supplemental Table 3). Clusters had 2 QTLs each, for EL (3 clusters), FF (3), FL (2), FS (3) and SF (1) and comprised 10.9% (24 out of 224) of all fiber QTLs. This correspondence to QTLs from other larger populations provides further support for the validity of at least a subset of the inferred QTLs.

The QTLs for different fiber characters located on the same chromosome were generally clustered both in \( n^2 \) and \( im \). For example, three QTLs on Chr.26 of \( n^2 \) were all in bin 5. Four QTLs on Chr.02 of \( im \) were centered in bins 4 and 5 and 2 more in bin 2 of A02. Many of the QTLs, scattered on different bins of Chr.12 in \( n^2 \), or different chromosomes, were clustered with QTLs detected in the elite populations. In total, inclusion of mutant population data increased QTL clusters to 27 bins of 105 QTLs, versus 24 bins of 87 QTLs based on the elite populations alone, with 6 bins containing 5-7 fiber QTLs on chr. 1, 2, 16, 17 and D08 (2 bins). As found in the three non-mutant populations, no statistically-significant QTL clusters occurred in corresponding regions of the homoeologous chromosomes.

**Relationships between QTLs and fiber-related cDNAs/genes:** A ‘consensus map’ that depicts the inferred marker arrangement along the genome of the common ancestor that gave rise to the diploid progenitors of tetraploid cotton (Rong et al. 2005a) sets the stage for exploring relationships between cotton QTLs and genes from taxa such as *Arabidopsis* that diverged from cotton prior to polyploid formation. All QTLs from the reference map were plotted to corresponding locations on the consensus map. In parallel, we determined the locations in the *Arabidopsis* genome of a total of 203 candidate genes representing up to the four best matches.
(at E < 10^{-10}) for each of 78 cDNAs/genes known from prior published work to be preferentially expressed during cotton lint fiber growth and development (Supplemental Table 7, which includes their protein family affiliation, GenBank accession number, expression pattern if known, and citations of published descriptions) and 28 Arabidopsis genes known to be critical for trichome and/or fiber development (Supplemental Table 8). The Arabidopsis orthologs were plotted to their locations in an inferred ancestral gene order that mitigates the effects of a whole-genome duplication in Arabidopsis since its divergence from cotton (i.e., alpha duplicated segments --). Established syntenic relationships between cotton and Arabidopsis (Rong et al. 2005a) were used to determine correspondence between the QTLs and candidate genes. QTL locations were plotted based on the midpoint of the QTL likelihood interval.

We found evidence of a general association between concentrations of candidate genes, and cotton fiber-related QTLs. Based on synteny inferred using CrimeStat II (Rong et al. 2005a), a significant correlation (r = 0.260, N=167, p = 0.0003), was found between QTL number in cotton and candidate gene number in Arabidopsis. When we considered only the smaller number (largely a subset) of regions of synteny inferred using FISH (Rong et al. 2005a), the correlation was weaker (r = 0.072, N=189) and fell short of significance (p=0.164). In both CSII and FISH-based models, the correlations between numbers of fiber QTLs and all (not just candidate) Arabidopsis genes in the syntenic regions detected by CrimestatII or FISH were not significant, suggesting that the relationship is specific to the population of candidate genes. Figure 3 presents an example in which four fiber-related QTLs on a region of consensus chromosome. C12 ranging from 54.1 to 73.1 cM showed synteny with Arabidopsis α 4, 10 and 14 duplicated segments. At least 6 Arabidopsis sequences in this region matched two fiber related cDNAs
(CesA1 and GhExp1) and one Arabidopsis trichome gene (TTG2). Numbers and names of all Arabidopsis orthologs in the syntenic regions can be found in CMap.

DISCUSSION

The subgenome from the non-fiber-producing ancestor plays a large role in the genetic control of fiber growth and development in polyploid cotton: Tetraploid cottons have higher lint fiber yield and quality than the modern descendents of their diploid progenitors, even in environments to which both have been adapted by plant breeding (Jiang et al. 1998). While the genetic potential for ‘lint’ fiber development appears to have been transmitted to tetraploid cottons by their A genome progenitor, early demonstrations that Dt-subgenome loci have been recruited to contribute further to lint fiber genetics in tetraploids (Jiang et al. 1998) gain further support from the more extensive data now available from larger numbers of QTLs in additional independent populations that include a higher degree of replication across environments (Paterson et al. 2003), and advanced-generation studies able to detect QTLs of smaller phenotypic effect (Chee et al. 2005a; Chee et al. 2005b; Draye et al. 2005; Lacapect al. 2005). This finding implicates the Dt-genome (from the non-fiber producing ancestor) in evolution of the transgressive fiber quality and yield of polyploid cottons relative to their diploid progenitors.

A wealth of research has revealed that gene copies from genome duplications (polyploidizations) experience different fates during their evolution including gene loss, subfunctionalization and neofunctionalization (Lynch and Force 2000; Rastogi and Liberles 2005). This raises the possibility that differential evolution of homoeologous fiber-related genes
duplicated by polyploid formation (CRONN et al. 1999), is partly responsible for modern cotton fiber quality. Randomly-chosen genes show similar evolutionary rates in the two genomes before and after polyploid formation. However, lack of knowledge of the gene network involved in lint fiber morphogenesis has precluded study of its specific constituents – perhaps the Dt alleles of key lint fiber-related genes have evolved more quickly. In other words, selection may have played a different role in genetic changes of homoeologous genes in At and Dt genomes, respectively, after reunion of these two genomes about 1-2 mya ago.

**Meta-analysis reveals a complex QTL landscape:** Although the sets of QTLs found for the same fiber-related trait in different experiments showed non-random correspondence, we were surprised that correspondence was only observed for about 10% of the total QTL set. Comparison in the same environment of two new mapping populations which shared one identical parent, Pima S-7, and had closely-related alternative parents that differed in carrying specific fiber mutants ($n^2$ and $im$, respectively), showed virtually no common QTLs. However, the validity of many QTLs in each of these populations is supported by their correspondence to QTLs found in other populations. This suggests that lint fiber development may involve a complex gene network in which perturbations at one point have widespread consequences. The discrete mutants studied may impose particularly large perturbations – if small differences in QTL alleles among populations also cause the same phenomenon, then the generally low level of correspondence among QTLs found in different populations might be explained.

These findings have several implications. In terms of basic genetics, the heterogeneity in QTLs revealed by the different experiments reiterates the need for study of a broad sampling of germplasm, in a wide range of environments, to gain a representative picture of the true genetic
complexity of a trait in a particular taxon. From an applied standpoint, the data highlight the need to validate QTLs specifically in each genetic background in which they are to be deployed (for example in mainstream breeding).

Members of the lint fiber development network appear clustered, but cluster members show heterogeneous phenotypic effects: Several lines of evidence suggest non-random distribution of genes implicated in lint fiber development. First, 105 (46.9% of) lint fiber-related QTLs were concentrated in 27 bins comprising 12.1% of the genome. Further, the corresponding regions of the *Arabidopsis* genome also contain non-random concentrations of genes that are implicated in fiber/trichome development based on either expression patterns or functional information.

Most of these clusters show heterogeneous phenotypic consequences. In only a small subset of cases (12) did more than one QTL affect the same phenotype. Moreover, in many cases, the genes were detected in different experimental populations, precluding pleiotropy as a general explanation (although it may contribute in some cases).

A tantalizing hypothesis is that concentrations of cotton lint fiber QTLs may represent groups of coordinately-regulated genes, and/or groups of small gene families that have undergone proximal duplication followed by sub- or neo-functionalization. Progress in physical mapping of several members of the *Gossypium* genus may permit this hypothesis to be tested in the near future.

Meta-analysis linked to synteny-based and expression-based information provides clues about specific genes and families involved in QTL networks: Integration of QTLs from
individual maps using a common reference map will be helpful to infer possible positional relationships to QTL locations on other maps that share some DNA markers with our reference map (LACAPE et al. 2005; LACAPE et al. 2003; HAN et al. 2006; SHEN et al. 2005). For example, a QTL governing fiber elongation (ELO) (LACAPE et al. 2005) near marker G1058 in Chr.22 corresponds closely to a QTL for the same trait (EL22.1) on the reference map, suggesting that these two QTLs may be the same. Similarly, if new QTLs were plotted in maps which share common markers with this high density map, it can quickly be determined if they might correspond to previously identified QTLs.

The integration of QTLs, discrete mutants and regions syntenic with Arabidopsis in a common map can facilitate the exploitation of Arabidopsis sequence information in further study and cloning of fiber genes. Burgeoning information about differential expression patterns of cotton genes, together with extensive knowledge of the molecular control of epidermal trichome development in Arabidopsis, each provide valuable clues about possible cotton genes in the fiber development network. The locations of cotton fiber QTLs, and information about synteny between cotton and Arabidopsis, make it possible to begin to search for intersections among these various data types that point to specific genes as candidates worthy of further functional testing to determine if they are directly responsible for genetic variation in cotton fiber development.

For example, the collective data in clusters of QTLs suggest numerous candidate genes and possible approaches for advancing progress toward identification of the causal gene(s), as follows:

The cluster with 5 fiber QTLs in Chr.17 was in a region ranging from 2.4 to 11.4 cM on bin 1. These QTLs each are based on different measurements of fiber quality, and reported in two
different studies [FS17.1 and FC17.1: (PATERNON et al. 2003) and FL17.1, FU17.1 and MIC17.1: (CHEE et al. 2005a; CHEE et al. 2005b; DRAYE et al. 2005)]. The region harboring these QTLs showed synteny (identified by both CrimestatII and FISH) with Arabidopsis duplications α03 and α21. Gene At4g18780 in α21 is CelA1 (E=0) and two Arabidopsis genes (At1g12560 and At1g62980) in α03 are members of the expansin gene family (best matching AtEXP7, E< 1E-154 and AtEXP18, E< 6E-156) involved in Arabidopsis trichome development.

In linkage group D08, 11 of 22 fiber QTLs were clustered in bins 3 and 8. QTLs in bin 3 are ELD08.2, FLD08.1(HVsl2.5,HVuhm,Lw), FUD08.1(HVui), SFD08.1(HVsfc, SFCn), FSD08.2(STR) and FFD08.2, and distributed from 40.3-53.1 cM, a marker sparse segment. Except for FSD08.2 (STR) reported by (JIANG et al. 1998), all others were detected in advanced backcross populations (CHEE et al. 2005a; CHEE et al. 2005b; DRAYE et al. 2005). No QTLs were found in the homoeologous region on Chr.05. These QTLs were located in a region (D05.153) showing synteny with a segment of Arabidopsis duplication α5 where four candidate genes were found including paralogs (At1g80350 and At1g80360, E=8E-118 and 1.4E-56, respectively) of fra2, and homologs (At1g15690 and At1g79840, E=4.8E-97 and 0) of proton-translocating pyrophosphatase (Ppase) and a homeobox protein (GhHox1). Another 5 fiber related QTLs on D08 were located in a region of bin 8 from 140.3-157.4 cM. All five QTLs were found in advanced backcross populations (CHEE et al. 2005a; CHEE et al. 2005b; DRAYE et al. 2005), including three fiber length loci, FLD08.2 (HVsl2.5, HVuhm, Lw), FUD08.2 (HVui) and SFD08.2 (HVsfc), and two fiber elongation loci, ELD08.3 and ELD08.4. The region showed synteny with part of α7 (D07.160) and one of the Arabidopsis genes, At4g09820, is the paralog of EGL3 (E< 1.5E-38).
ACKNOWLEDGEMENTS

We thank numerous colleagues for valuable assistance and suggestions, and the United States - Israel Binational Agricultural Research and Development Fund (#US-2506-94R to AHP and YS), the BOYSCAST program of the DST, India (to VNW), USDA National Research Initiative (02-01412 to AHP), the National Science Foundation Plant Genome Research Program (DBI-9872630, DBI-0211700 to AHP and JFW), the Fonds National belge pour la Recherche Scientifique (Research associate grant to XD), the Texas and Georgia Agricultural Experiment Stations, Texas Higher Education Coordinating Board, Cotton Incorporated, and USDA-IFAFS (00-52100-9685 to AHP, PWC, JRG, OLM, CWS) for financial support.

LITERATURE CITED


PATERSON, A. H., E. S. LANDER, J. D. HEWITT, S. PETERSON, S. E. LINCOLN et al., 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335: 721-726.


WANG, S., C. J. BASTEN and Z. B. ZENG, 2005 Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC.


TABLE 1 Crosses and QTLs of different populations and their file names at the CMap website (www.plantgenome.uga.edu/cmap)

<table>
<thead>
<tr>
<th>Cross*</th>
<th>Population (# individuals)</th>
<th>Category**</th>
<th>CMap db identifiers</th>
<th>Total QTLs</th>
<th>QTLs on At***</th>
<th>QTLs on A</th>
<th>QTLs on Dt***</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pima S-7(GB) x im (GH)</td>
<td>F2 (124)</td>
<td>EL,FF,FL,FS, FU,YC</td>
<td>im</td>
<td>17</td>
<td>10</td>
<td>7</td>
<td>Herein</td>
<td></td>
</tr>
<tr>
<td>Tamcot 2111(GH) X Pima S6 (GB)</td>
<td>BC3F2 (24 families, 33-191 plants/family)</td>
<td>EL,FF,FL,FS, FU</td>
<td>ABHB</td>
<td>108</td>
<td>45</td>
<td>63</td>
<td>Chee et al. (2005); Draye et al. (2005)</td>
<td></td>
</tr>
<tr>
<td>CAMD-E (GH) x Sea Island Seaberry (GB)</td>
<td>F2 (271)</td>
<td>EL,FF,FL,FS,YC</td>
<td>HBGF</td>
<td>14</td>
<td>4</td>
<td>10</td>
<td>Jiang et al. (1998)</td>
<td></td>
</tr>
<tr>
<td>Pima S-7(GB) x n2 (GH)</td>
<td>F2 (124)</td>
<td>EL,FF,FL,FS, FU,YC</td>
<td>n2</td>
<td>22</td>
<td>11</td>
<td>11</td>
<td>Herein</td>
<td></td>
</tr>
<tr>
<td>Siv’on (GH) x F-177(GB)</td>
<td>F2 (406); F3 (208)</td>
<td>EL,FC,FF,FL,FS,FU, DP,YC</td>
<td>HBPS</td>
<td>79</td>
<td>36</td>
<td>43</td>
<td>Paterson et al. (2003)</td>
<td></td>
</tr>
<tr>
<td>Deltapine 61(GH) x Sea Island Seaberry (GB)</td>
<td>F2 (180)</td>
<td>LM</td>
<td>HBJS</td>
<td>21</td>
<td>7</td>
<td>14</td>
<td>Saranga et al. (2004)</td>
<td></td>
</tr>
<tr>
<td>TMS-22 (GH) x acc. WT936 (GT)</td>
<td>F2 (82)</td>
<td>LM</td>
<td>HT</td>
<td>11</td>
<td>7</td>
<td>4</td>
<td>Jiang et al. (2000)</td>
<td></td>
</tr>
<tr>
<td>Pima S-7(GB) x Empire B2b6, B2, B3, S295(GH)</td>
<td>F2 (119-150)</td>
<td>BB</td>
<td>BHW</td>
<td>7</td>
<td>1</td>
<td>6</td>
<td>Wright et al. (1998)</td>
<td></td>
</tr>
<tr>
<td>Pima S-7(GB) x Fb(1GH)</td>
<td>F2 (140)</td>
<td>YC</td>
<td>Fbl</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>Wright et al. (1999)</td>
<td></td>
</tr>
<tr>
<td>Pima S-7(GB) x N1(1GH)</td>
<td>F2 (143)</td>
<td>YC</td>
<td>N1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Rong et al. (2005)</td>
<td></td>
</tr>
<tr>
<td>Total ( Tetraploid)</td>
<td></td>
<td></td>
<td></td>
<td>405</td>
<td>184</td>
<td>221</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acc. PI529740 (GA) x Acc. PI529670 (GHe)</td>
<td>F2 (167)</td>
<td>FM</td>
<td>A</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>Desai et al. (unpublished)</td>
<td></td>
</tr>
</tbody>
</table>

* All populations are crosses between *G. hirsutum* and *G. barbadense* except for “A” (between *G. arboreum* acc. PI529740 and *G. herbaceum* acc. PI529670) and HT (between *G. hirsutum* acc. TMS-22 and *G. tomentosum* acc. WT936).

** EL-elongation; FF-fineness; FL-length; FS-strength; FU- uniformity; FC-color; YC-yield component; FM-flower morphology; LM-leaf morphology; DP-drought response; BB- bacterial resistance; TR-trichome.

*** At, Dt: Tetraploid chromosomes derived from A-genome and D-genome diploid progenitors, respectively.
TABLE 2  Chi-squared test of fiber related QTLs distributed on At and Dt subgenome in three non-mutant populations

<table>
<thead>
<tr>
<th>Population*</th>
<th>QTL</th>
<th>At**</th>
<th>Dt**</th>
<th>Goodness fit to 1:1</th>
<th>Reference</th>
</tr>
</thead>
</table>
| CAMD-E (GH) x Sea Island Seaberry (GB) | 9   | 3    | 6    | 0.317               | Jiang et al. (1998)***(
| Siv'on (GH) x F-177(GB) | 79  | 36   | 43   | 0.431               | Paterson et al. (2003) |
| Tamcot 2111(GH) X Pima S6 (GB) | 108 | 45   | 63   | 0.083               | Chee et al. (2005); Draye et al. (2005) |
| Total       | 196 | 84   | 112  | 0.045               |           |

* GH-G. hirsutum; GB-G. barbadense  
** At, Dt: Tetraploid chromosomes derived from A-genome and D-genome diploid progenitors, respectively.
*** Four QTLs for yield components and one for earliness not included, as discussed in text.
TABLE 3  Chi-squared test of the QTL distribution between tetraploid At and Dt subgenomes for different fiber traits

<table>
<thead>
<tr>
<th>Trait*</th>
<th>QTL</th>
<th>At**</th>
<th>Dt***</th>
<th>Goodness of fit to 1:1</th>
</tr>
</thead>
<tbody>
<tr>
<td>EL</td>
<td>33</td>
<td>15</td>
<td>18</td>
<td>0.6015</td>
</tr>
<tr>
<td>FC</td>
<td>11</td>
<td>7</td>
<td>4</td>
<td>0.3657</td>
</tr>
<tr>
<td>FF</td>
<td>57</td>
<td>25</td>
<td>32</td>
<td>0.3538</td>
</tr>
<tr>
<td>FL</td>
<td>34</td>
<td>17</td>
<td>17</td>
<td>1.0000</td>
</tr>
<tr>
<td>FS</td>
<td>24</td>
<td>7</td>
<td>17</td>
<td>0.0412</td>
</tr>
<tr>
<td>FU</td>
<td>21</td>
<td>10</td>
<td>11</td>
<td>0.8273</td>
</tr>
<tr>
<td>MIC</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>0.0047</td>
</tr>
<tr>
<td>SF</td>
<td>8</td>
<td>3</td>
<td>5</td>
<td>0.4795</td>
</tr>
<tr>
<td>Total</td>
<td>196</td>
<td>84</td>
<td>112</td>
<td>0.0455</td>
</tr>
</tbody>
</table>

* EL-elongation; FF-fineness; FL-length; FS-strength; FU- uniformity; FC-color; YC-yield component; FM-flower morphology; LM-leaf morphology; DP-drought response; BB- bacterial resistance; TR-trichome.

** At, Dt: Tetraploid chromosomes derived from A-genome and D-genome diploid progenitors, respectively.
FIGURE LEGENDS

FIGURE 1. A CMap display of the comparison between an individual map and a consensus map. The left map is chromosome 16 from the n2 population including one QTL (MIC) shown on the left side of the map. The right map is consensus chromosomes C4, reconstructed from the At and Dt maps as described (Rong et al. 2005a). To the immediate right of this map, blocks of inferred synteny with Arabidopsis are shown. D (or Ds)## indicates correspondence to the accordingly-numbered Arabidopsis alpha-duplicated segment (Bowers et al. 2003), and the number after the dot indicates the span of this particular segment of correspondence in the alpha-duplicated segment (which can be visualized at our CMAP website, also see Rong et al. 2005a for further details). To the right of blocks of inferred synteny, QTL likelihood intervals are plotted, labeled as described in the text. Where appropriate, parenthetical information indicates experimental treatments (for example, ww = well-watered and bt = both treatments: Paterson et al. 2003), or alternate name of the trait if different in prior publications are presented in parenthesis after each QTL name. Detailed explanation of these descriptions can be found in the references cited.

FIGURE 2. Number and distribution of genetic markers and fiber-related QTLs in 20 cM bins along chromosomes 1 and 2 of a reference genetic map (Rong et al. 2004). Empty
bars represent marker number and solid bars are 2-LOD likelihood intervals for fiber QTLs.

FIGURE 3. Genetic map and *Arabidopsis* syntenic regions of a segment of cotton consensus chromosome C12 (from 54.1 to 73.1 cM) showing the locations of fiber related QTLs, *Arabidopsis* orthologs of cotton fiber related cDNAs, and *Arabidopsis* genes functioning in trichome development. Vertical bars on the left side of C12 represent the locations of QTLs for fiber elongation (Paterson *et al.* 2003), fiber length (Chee *et al.* 2005b), and mass of seed cotton (log transformed) (Jiang *et al.* 1998). Solid lines link putative orthologs between cotton and *Arabidopsis* which were used to infer correspondence in this region. The numbers on the left side of the *Arabidopsis* segments are consecutive numbers assigned to genes in each α duplicated segment, provided to permit one to know (by subtraction) the number of intervening genes between two anchor points. Values in parentheses to the right of *Arabidopsis* gene designations indicate gene identifier and e-value for match to a candidate gene identified based on expression pattern (as detailed in Supplemental Table 7) or a role in *Arabidopsis* trichome development (Supplemental Table 8).
Figure 1
Figure 2
Figure 3