Genetica (2009) 136:401–417
DOI 10.1007/s10709-008-9337-8

Linkage disequilibrium based association mapping of fiber quality traits in *G. hirsutum* L. variety germplasm

Ibrokhim Y. Abdurakhmonov · Sukumar Saha · Jonnie N. Jenkins · Zabardast T. Buriev · Shukhrat E. Shermatov · Brain E. Scheffler · Alan E. Pepper · John Z. Yu · Russell J. Kohel · Abdusattor Abdukarimov

Received: 17 August 2008 / Accepted: 17 November 2008 / Published online: 9 December 2008

Abstract Cotton is the world's leading cash crop, but it lags behind other major crops for marker-assisted breeding due to limited polymorphisms and a genetic bottleneck through historic domestication. This underlies a need for characterization, tagging, and utilization of existing natural polymorphisms in cotton germplasm collections. Here we report genetic diversity, population characteristics, the extent of linkage disequilibrium (LD), and association mapping of fiber quality traits using 202 microsatellite marker primer pairs in 335 *G. hirsutum* germplasm grown in two diverse environments, Uzbekistan and Mexico. At the significance threshold ($r^2 \geq 0.1$), a genome-wide average of LD extended up to genetic distance of 25 cM in assayed cotton variety accessions. Genome wide LD at $r^2 \geq 0.2$ was reduced to ~5–6 cM, providing evidence of the potential for association mapping of agronomically important traits in cotton. Results suggest linkage, selection, inbreeding, population stratification, and genetic drift as the potential LD-generating factors in cotton. In two environments, an average of ~20 SSR markers was associated with each main fiber quality traits using a unified mixed liner model (MLM) incorporating population structure and kinship. These MLM-derived significant associations were confirmed in general linear model and structured association test, accounting for population structure and permutation-based multiple testing. Several common markers, showing the significant associations in both Uzbekistan and Mexican environments, were determined. Between 7 and 43% of the MLM-derived significant associations were supported by a minimum Bayes factor at 'moderate to strong' and 'strong to very strong' evidence levels, suggesting their usefulness for marker-assisted breeding programs and overall effectiveness of association mapping using cotton germplasm resources.

Keywords Cotton germplasm · Genetic diversity · Fiber quality · Linkage disequilibrium (LD) · Simple sequence repeat (SSR) markers · LD-based association mapping

Electronic supplementary material The online version of this article (doi:10.1007/s10709-008-9337-8) contains supplementary material, which is available to authorized users.

I. Y. Abdurakhmonov (✉) · Z. T. Buriev · S. E. Shermatov · A. Abdukarimov
Center of Genomic Technologies, Institute of Genetics and Plant Experimental Biology, Academy of Sciences of Uzbekistan, 111226 Tashkent, Uzbekistan
e-mail: ibrokhim_a@yahoo.com; genomics@uzsci.net

S. Saha · J. N. Jenkins
Crop Science Research Laboratory, Genetics and Precision Agriculture, USDA-ARS, 8120 Highway 12E, P.O. Box 5367, Mississippi State, MS 39762, USA

B. E. Scheffler
USDA-ARS, 141 Experiment Station Road, Stoneville, MS 38776, USA

A. E. Pepper
Department of Biology, Texas A&M University, College Station, TX 77843, USA

J. Z. Yu · R. J. Kohel
Crop Germplasm Research Unit, USDA-ARS, 2881 F&B Road, College Station, TX 77845, USA

Introduction

Although wild cottons (*Gossypium* species) are perennial shrubs, cotton is largely cultivated as domesticated annual crops. Cotton is not only the most important natural textile

fiber source, but also a significant food source for humans and livestock. The worldwide economic impact of the cotton industry is estimated to ~$500 billion/year with an annual utilization of ~115-million bales or ~27-million metric tons of cotton fiber (National cotton Council, 2006, http://nationalcottoncouncil.com; Chen et al. 2007). Stagnant yield, declining fiber quality, and threat from biotic and abiotic stresses limit the profitability in world cotton production. Genetic improvement of fiber yield and quality is the primary objectives of cotton breeding programs worldwide. However, improvement of cotton fiber quality is challenging due to narrow genetic base of modern cotton cultivars (Iqbal et al. 2001; Rungis et al. 2005) and negative genetic correlation between fiber quality and yield (Culp and Lewis 1973). Both attributes indicate the great need to explore novel germplasm resources for cotton and tag the existing polymorphisms of agronomic importance.

The genomes of allotetraploid cottons (including cultivated *G. hirsutum* and *G. barbadense*) have a chromosome complement of $2n = 4X = 52$, a haploid genome size of 2,200–3,000 Mb DNA, and a total recombination length of approximately 5,200 cM (an average of 400 kb per cM) (Paterson and Smith 1999). Molecular marker technology was successfully used to create genetic linkage maps, map important agronomic quantitative trait loci (QTLs), and assess genetic diversity (Chen et al. 2007; Abdurakhmonov 2007; Zhang et al. 2008; Preetha and Raveendren 2008). A large collection of robust, portable, and PCR-based Simple sequence repeat (SSR) marker resources, were developed in multiple laboratories and made available to the cotton research community through the cotton marker database (CMD) (Blenda et al. 2006). As a result, a number of potential DNA markers, identified using QTL-mapping in bi-parental experimental populations, were made available for future breeding programs of cotton to develop superior cotton cultivars through marker assisted-selection (MAS) programs (Zhang et al. 2008).

The traditional QTL mapping using bi-parental populations is low resolution method with little allele coverage, extremely time-consuming, high-risk and expensive work—prohibitively expensive if dozens, let alone hundreds or thousands, of cotton germplasm accessions are to be examined. However, use of linkage disequilibrium (LD)-based association mapping circumvents the need for large bi-parental mapping populations by making use of information contained within the genetic recombinations that have occurred in natural populations during the course of recent evolution, and as a result of crosses performed in the course of both historical and modern breeding efforts. For example, in human, DNA markers linked to a specific locus controlling a particular genetic disease or trait will show a reduced (non-equilibrium) level of recombination with the locus controlling the trait of interest (Abdurakhmonov 2007). This LD can be detected statistically, and has been used to map and eventually clone a number of genes underlying complex genetic traits in humans (Weiss and Clark 2002; Schulze and McMahon 2002). By this method, certain alleles at a marker locus are associated with particular alleles at a linked locus affecting a trait of interest.

The extent of genome-wide LD or allelic association is the key starting point for association mapping. The extent of LD has been quantified and an association mapping has been successfully applied for many plant species (Gupta et al. 2005; Abdurakhmonov and Abdukarimov 2008). Genome-wide LD extended up to 50–250 kb in a globally derived set of 96 *Arabidopsis* accessions, with some LD blocks up to 50–100 cM in local *Arabidopsis* populations (Nordborg et al. 2002, 2005). Genome-wide LD decay was determined to occur within 200–1,500 bp in maize (Remington et al. 2001; Tenaillon et al. 2001), 10 cM in sugar cane and soybean (Flint-Garcia et al. 2003; Zhu et al. 2003), 3 cM in sugar beet (Kraft et al. 2000), 50 cM in sorghum (Hamblin et al. 2004), 10–50 cM in barley (Kraakman et al. 2004; Malysheva-Otto et al. 2006), and 10–20 cM durum wheat (Maccaferri et al. 2005). Important information from these studies was that the genome-wide extent of LD in plants could vary across genomes and between species with the examples of longer stretches of LD in local populations. Although LD-based association mapping was successfully used in many organisms, including crop germplasm resources (Gupta et al. 2005; Abdurakhmonov and Abdukarimov 2008), the serious influence of confounding population structure and relatedness of individuals are important concerns in conducting association mapping (Pritchard et al. 2000; Yu et al. 2006; Zhao et al. 2007).

In this study, we selected a large number of *G. hirsutum* (also referred to as 'Upland cotton') variety accessions from the Uzbek cotton germplasm collection and measured economically important fiber quality traits of these accessions in two distinct environments (Uzbekistan and Mexico). These accessions were genotyped with SSR markers to study the extent and distribution of diversity, population structure, kinship, and pairwise LD between SSR markers. From these data, we estimated an average extent of genome-wide LD for cotton. We also performed LD-based association mapping for the main six fiber quality traits in cotton using a unified mixed linear model (MLM) incorporating population stratification attributes as well as general linear model (GLM) and structured associations (SA) accounting for population structure and permutation-based multiple testing correction. In Uzbekistan and Mexican environment, an average of ~20 SSR markers was associated with each main fiber quality traits. Majority of these significant associations were specific to single environment, but there were several common

marker-trait associations that were significant in the both environments in MLM model. Between 7 and 43% of our MLM-derived significant associations of fiber quality traits were supported by a minimum Bayes factor at 'moderate to strong' and 'strong to very strong' evidence level (Goodman 2001), suggesting a reliable portion of associations for breeding programs. The results of this study are, to the best to our knowledge, the first report on a genome-wide LD scan and LD-based association mapping of fiber quality traits using *G. hirsutum* variety germplasm resources. The results are very useful for application of 'association study' in cotton that will accelerate development of superior cotton cultivars through MAS programs.

Materials and methods

Selection of cotton accessions

The Uzbek cotton germplasm collection constitutes more than 17,000 cotton germplasm accessions of the A- to K-genome groups from 43 cotton species that have been developed, collected, and maintained for the past century (Abdurakhmonov 2007). In that, *G. hirsutum* variety accessions comprise about 85% of the collection with broad geographic and ecotypic coverage. A total of 334 *G. hirsutum* variety accessions from Uzbek (303), Latin American (18), and Australian (13) ecotypes were selected from Uzbek cotton germplasm collection and used for genome-wide LD scans and association mapping purposes. Three other allotetraploid cotton genotypes including Texas Marker-1 (a standard genotype for *G. hirsutum*), Pima -379 (a standard genotype *G. barbadense*) and *G. tomentosum* also were included for marker genotyping (see supplemental material Table S1 for the list of accessions).

Phenotypic analyses in Uzbekistan environment

Phenotypic analyses of these selected accessions were performed in field stations of the Institute of Genetics and Plants Experimental Biology (IG&PEB), Tashkent, Uzbekistan in 2003. Standard field plots and agronomic technologies were used for growing selected accessions in the Tashkent cultivation environment, which is totally irrigated. Detailed information about the Tashkent environment for specific years can be obtained from the archive of the meteorology center (available at http://meteo.infospace.ru/wcarch/html, visited May 31, 2007). Ten plants of each accession were grown and each group of accessions was allowed to self-pollinate by covering the flowers with paper bags just before the flowers opened. In late September and the beginning of October, cotton fiber samples from self-pollinated cotton bolls of each accession were harvested from field-grown plants. In brief, at least 25 fully opened self-pollinated cotton bolls were harvested from each group of accessions (pooled from ten plants per accession). Raw fiber samples were ginned manually to isolate seeds and lint. Fiber quality traits of variety accessions grown in Uzbekistan environment, such as fiber length (upper higher mean-UHM) and strength (STR), micronaire (MIC), elongation (ELO), uniformity (UI), and reflectance (Rd), were measured by High Volume Instrument (HVI) of the STARLAB, Knoxville, TN, USA.

Phenotypic analyses in Mexico environment

All selected variety cotton accessions were also grown in short-day conditions of the Cotton Winter Nursery (CWN) of USDA-ARS, in Tecoman, Mexico in 2005. Cotton accessions are grown in the CWN during the winter dry season under irrigation. Planting consisted of individual plants spaced 18 inches apart in rows space 40 inches. Seven plants of each accession were grown, and a pooled fiber sample was analyzed for fiber quality traits (UHM, STR, MIC, ELO, UI, and Rd). Fiber analysis was conducted by the cotton incorporated HVI system**.** The fiber quality measurements were obtained from a single replicate in each of the two environments due to high cost of multiple trails/replicates with such a large sample of cotton accessions. Analysis of variance (ANOVA) and phenotypic correlations between traits in the two growing-environments were performed using the Visual Statistics System (ViSta) (Young and Bann 1996).

Genotyping with SSR markers

From each accession, 8–10 young, fully expanded, leaves were collected, stored at −80°C, and genomic DNAs were isolated from the frozen leaf tissues using method of Dellaporta et al. (1983) with minor modification and optimization for frozen tissues. Prepared genomic DNAs were checked in 0.9% agarose electrophoresis and DNA concentrations were estimated based on *Hin*dIII digested λ-phage DNA. The 334 *G. hirsutum* variety accessions (Table S1) and three controls of *G. hirsutum* var. Texas Marker-1 (TM-1), *G. barbadense* var. 3–79, and *G. tomentosum* were genotyped with a select group of 202 labeled-SSR primer pairs (supplemental Table S2). These chromosome-specific primer pairs were selected using the results of different laboratories, published papers (Liu et al. 2000a; Gutierrez et al. 2002; Han et al. 2004; Lacape et al. 2005; Shen et al. 2005; Abdurakhmonov et al. 2007a), and based on informativeness relative to important QTLs and chromosome distribution. This set of 202 SSR markers distributed an average of ~10 SSRs per chromosome, spanning along a total of 3,679 cM distance (~71% of a

cotton genome coverage) or an average of ~175 cM/21 chromosomes (Table S2) (Lacape et al. 2005).

PCR-amplifications were performed in a 8 µl reaction mix containing 0.8 µl 10× PCR buffer, 0.2 µl dNTPs (10 mM each), 0.72 µl 25 mM $MgCl_2$, 0.2 µl 5 pM labeled primers (FAM, HEX, VIC, PET), 0.07 µl AmpliTaq Gold DNA polymerase (Applied Biosystems, USA), and 15 ng genomic DNA. PCR amplification was carried out using a PTC-225 DNA Engine Tetrad thermocycler (MJ Research, USA) with first denaturation at 95°C for 10 min, followed by 10 cycles of 94°C for 1 min, 60°C for 1 min (decreases of 0.5°C in each cycle) and 72°C for 2 min; 33 cycles of 94°C for 15 s, 55°C for 30 s and 72°C for 1 min. A final 6 min extension at 72°C was performed. With labeled primers (Table S2), polymorphisms among cotton accessions at amplified SSR loci were visualized in a denaturing capillary electrophoresis using an ABI 3730xl with a 96-capillary system in POP-7 polymer (Applied Biosystems, USA). PCR-products were diluted 1:30 before loading into capillaries. The size standards of LIZ 500 or ROX 500 were loaded with the diluted PCR-products according to the manufacturer's guidelines (Applied Biosystems, USA). Calling the size of amplified products was performed using GeneMapper 3.7 (Applied Biosystems, USA) as well as a visual check for band calling correctness. The SSR product sizes were also compared to the CMD panel SSR amplification product sizes where available (Blenda et al. 2006).

Molecular genetic diversity and phylogenetic analyses

Since *G. hirsutum* is an allopolyploid with reticulated germplasm resources, SSR primer pairs often yielded multiple PCR-products in our cotton accessions. There is a great risk of false allele calling for multiple-band SSR markers when wide germplasm resources with unknown pedigree information are genotyped, unless only single-band loci are selected for genotyping. Considering the cotton germplasm material used in this study were strictly self-pollinated during the past 50-years for germplasm renewing, we scored our SSR data like a dominant marker type with "1" for absent, "2" for present state, or "0" (or "?", "−999", and "−9", depending on the software requirement) for the occasional non-amplification or missing data state, taking each band as an independent marker locus with a clear size band separation (Breseghello and Sorrells 2006; Tommasini et al. 2007) to avoid assigning incorrect allelic relationships since it is the concern in association analysis (Sand 2007). The heterozygosity level of marker data was identified according to an average similarity frequency of alternative alleles (0.5 vs. 0.5 for high heterozygosity or 0.9 vs. 0.1 low heterozygosity levels) (Li et al. 2007). Allele frequencies for dominant markers were calculated using SpaGeDi software (Hardy and Vekemans 2002). The polymorphic information content (PIC) was analyzed using the PowerMarker software package (Liu and Muse 2005).

Genetic distance and phylogenetic analyses of cotton accessions were performed using Neighbor Joining (N-J) algorithms with the minimum evolution objective function (Saitou and Nei 1987) of the software package PAUP*4.0b10 (Swofford 2002). Genetic variation within and among predefined groups and pairwise F_{ST} genetic distances were measured by analysis of molecular variance (AMOVA) (Reynolds et al. 1983; Weir and Cockerham 1984; Excoffier et al. 1992) using ARLEQUIN 2.0 (Schneider et al. 2000). We also applied a Bayesian method of further partitioning genetic differentiation among population groups, which allows direct estimates of F_{ST} from dominant markers without prior knowledge of inbreeding history (Holsinger et al. 2002; Holsinger and Lewis 2003). Several runs for full, $f = 0$, $\theta = 0$ and $f =$ free models were performed using HICKORY, ver. 1.0, with the default sampling parameters (burn-in = 50,000, sample = 250,000, and thin = 50) following software guidelines (Holsinger and Lewis 2003). Although the Bayesian analysis with dominant markers revealed the estimates of inbreeding coefficients (F_{IS}) (data not shown), we did not consider or discuss the results of F_{IS} due to the biased nature of the values obtained from small within population sample sizes (Holsinger and Lewis 2003). In both AMOVA and Bayesian population differentiation analyses, the 5% minor alleles filtered SSR datasets were used.

Pairwise linkage disequilibrium and LD decay

For population structure, kinship, pairwise LD and association mapping analyses, only *G. hirsutum* genotypes were used, excluding the control *Gossypium* genotypes, *G. barbadense* and *G. tomentosum*. The genome-wide LD between pairs of SSR marker loci was studied according to Witt and Buckler (2003) using the software package TASSEL ver. 1.9.6 (http://www.maizegenetics.net). The genome-wide LD between all pairs of SSR alleles were analyzed with 'minor allele' frequencies filtered datasets where SSRs alleles with a 0.05 frequency in genotyped accessions were removed before conducting LD analyses because minor alleles are usually problematic and biased for LD estimates between pairs of loci (Mohlke et al. 2001; McRae et al. 2002). The 'minor allele' removal was performed using the TASSEL site filtration function. LD was estimated by a weighted average of squared allele-frequency correlations (r^2) between SSR loci. The significance of pairwise LD (P-values ≤ 0.005) among all possible SSR loci was evaluated using TASSEL with the rapid permutation test in 10,000 shuffles. The LD values between all pairs of SSR loci were plotted as triangle LD plots using TASSEL (http://www.maizegenetics.net/tassel) to estimate the general view

of genome-wide LD patterns and evaluate 'block-like' LD structures. Because of small sample sizes within the predefined groups, we did not evaluate a separate pairwise LD for each group to avoid biased estimates. The linkage map information for 202 chromosome-specific markers tested on the variety cotton accessions was obtained from a linkage map constructed from the interspecific BC1 cross of [(Guazuncho2 (*G. hirsutum*) × VH8-4602 (*G. barbadense*)) × Guazuncho2] (Lacape et al. 2005). The r^2 values for pairs of SSR loci were plotted as a function of map distances (cM), and LD decay (at $r^2 < 0.1$) was estimated (Witt and Buckler 2003).

Inference of population structure and Kinship

A model-based approach, implemented in the software package STRUCTURE (Pritchard et al. 2000) for dominant markers (coded as 1, −9; 2, −9), was used to identify subgroups cotton variety accessions. In the first attempt, we used both 'no-admixture' and 'admixture' co-ancestry models under independent and correlated allele frequencies using the burn-in time of 50,000 and the number of replications at 100,000 (Pritchard and Wen 2004) with the K up to 10. However, we did not determine distinct clusters and could not assign a significant number of K populations using both above models. Therefore, we used the prior population information model pre-defining accessions to a specific type of populations. For instance, cotton accessions were assigned to (1) Uzbekistan, (2) Latin American, and (3) Australian cultivated cotton varieties based on the source of origin as indicated in the germplasm collection catalog (unpublished information). We also analyzed pairwise kinship (K-matrix) for our cotton variety accessions. Pairwise kinship estimates were calculated according to Hardy and Vekemans (2002) using the software package SpaGeDi. The kinship coefficient algorithm of Hardy (2003) for dominant markers was used to obtain the pairwise kinship matrix between the studied cotton accessions.

Association mapping of fiber quality traits

The MLM association test of fiber quality traits, incorporating K and Q matrices, was performed according to Yu et al. (2006) using the TASSEL software package (Bradbury et al. 2007). We also performed the GLM (Bradbury et al. 2007) and SA (Thornsberry et al. 2001) association analyses with the same data, incorporating population structure information as a covariate and using 1,000-time permutations for the correction of multiple testing. Since MLM method performs better in controlling spurious associations (Yu et al. 2006; Zhao et al. 2007), we first ranked significant association from MLM ($P \leq 0.05$) and then compared significance of these markers ($P \leq 0.05$) in the permutation-based GLM and SA association tests. We also separately tested the MLM-derived P-values for multiple testing correction using pFDR test in QVALUE program version 1.0 (Storey and Tibshirani 2003), Sidak procedure of Bonferroni adjustment (using $P = 1-(1-p)^L$, where L is a number of independent tests), and p_{ACT} method of Conneely and Boehnke (2007). Moreover, to further reliably interpret the MLM-derived significant associations, we calculated a minimum Bayes factor (BF_{min}) factor using following formula: $BF_{min} = -e*p*\ln(p)$ (Goodman 2001; Katki 2008). The MLM-derived significant associations were also compared with published literature information to judge obtained associations (see "Results" section). The 5% 'minor alleles' filtered SSR datasets were used for all association mapping models. Fiber trait data was imputed for missing data and normalized using algorithms implemented in TASSEL before conducting an association mapping.

Data availability

The SSR data genotyped in our germplasm is available to readers through the Cotton Marker Database (CMD; http://www.cottonmarker.org, verified on December 2, 2008; Blenda et al. 2006).

Results

Fiber quality properties of selected accessions in Mexican and Uzbekistan environments

Cotton accessions revealed a wide-range of phenotypic variation in fiber quality traits including fiber MIC, UHM, UI, STR, ELO, and Rd (Table 1). The variety accessions from Uzbekistan, Latin American and Australian ecotypes, while expressing a wide range of phenotypic variation for fiber quality traits within a specific environment, revealed different trait ranges for all fiber traits between two environments based on ANOVA analysis (Table 2). For instance, the trait range for MIC was varied (2.9–5.6) with a mean of 4.4 in the Uzbekistan environment whereas it varied from 2.3 to 6.5 with a mean of 4.7 in the Mexican environment (Table 1), revealing significantly ($P \leq 0.001$) different impact of specific environments to the measured fiber quality traits (Table 2). We observed significant trait correlations between the same fiber traits as well as among different fiber traits in Uzbekistan- and Mexico-grown accessions (Table 3). Results of trait correlations between the same fiber traits and correlations among different fiber traits in the two distinct environments suggest the important influence of environment on fiber development.

Table 1 Summary of the fiber quality traits from the Uzbekistan (UZB) and Mexican (MEX) environments

Traits	Summary of fiber quality traits in variety accessions						
	Number	Mean	Minimum	1st *Q*	Median	3rd *Q*	Maximum
MIC_UZB	242	4.44	2.9	4.1	4.5	4.85	5.6
UHM_UZB	242	1.06	0.82	1.03	1.07	1.11	1.24
UI_UZB	242	82.5	77.2	81.7	82.6	83.35	85.4
STR_UZB	242	28.56	17.5	26.95	28.7	30.15	36.3
ELO_UZB	242	8.23	6	7.9	8.3	8.5	9.6
RD_UZB	242	72.62	44.3	71.7	73.55	75.05	78.3
MIC_MEX	242	4.74	2.3	4.4	4.8	5.1	6.5
UHM_MEX	242	1.1	0.93	1.06	1.1	1.14	1.28
UI_MEX	242	83.9	79	82.7	83.9	84.9	88.2
STR_MEX	242	27.23	15.8	25.3	27.1	28.9	44
ELO_MEX	242	6.8	4	6.2	6.8	7.4	9.6
RD_MEX	242	74.12	0.00	74.80	76.10	77.20	80.20

MIC micronaire; *UHM* fiber length; *UI* uniformity; *STR* fiber strength; *ELO* elongation; *RD* reflectance

Table 2 Mean squares of the ANOVA of fiber quality trait measurements in two environments, Uzbekistan and Mexico

Source of variation	d. f.	MIC	UHM	UI	STR	ELO	RD
Locations	1	11.04**	0.13**	239.83**	213.82**	247.06**	271.35*
Cultivar	241	0.37**	0.01**	2.82*	11.52**	0.84**	135.08**
All sources	242	0.42**	0.01**	3.80**	12.36**	1.86**	135.04**
Error	241	0.25**	0	2.01	3.95	0.47	33.77
r^2 of the model		0.63	0.81	0.65	0.76	0.80	0.80

MIC micronaire; *UHM* fiber length; *UI* uniformity; *STR* fiber strength; *ELO* elongation; *RD* reflectance; * $P \leq 0.005$ and ** $P \leq 0.0001$

SSR marker polymorphisms

The primer pairs amplified 1,104 polymorphic SSR amplicons or alleles with an average of 5.5 SSR alleles (2–15 allele range) per primer pair. Thirty-six percent (402) of SSR alleles were unique and present in only 0.5% of the accessions. Thirty-seven percent (405) of the SSR alleles were rare and found in only 5% of the accessions genotyped. The remaining 297 (27%) SSR alleles were highly polymorphic. The overall PIC for SSRs was in the range of 0.006–0.50 with an average of 0.082. An average similar frequency of alternative alleles (refer to methodology for details) of SSR data was 0.473 for dominant (or present) allele versus 0.50 for recessive (or absent) alleles (with ~0.027 missing data), after filtering for a 5% 'minor alleles' frequency. Filtering of original SSR data set for a 5% minor allele frequency generated the SSR data set from 109 primer pairs, which covered ~55% of the cotton genome with an average of ~5.5 SSR primer pairs per each 20 chromosome spanning at an average of 142 cM distance/chromosome (Table S2).

Genetic distance estimates

The Neighbor-Joining (NJ) analysis revealed that these variety accessions have a narrow genetic base. The mean genetic distance (GD), generated by the NJ algorithm of PAUP analysis (Saitou and Nei 1987), among all Upland accessions ranged from 0.005 to 0.26 with an average of 0.12. The average GD within the *G. hirsutum* accessions of specific ecotypes (Uzbekistan, Latin American and Australian) was very close and ranged from 0.12 to 0.14. The highest GD among *G. hirsutum* varieties was observed within the Australian ecotype group (0.26). As expected, *G. hirsutum* varieties from these diverse three ecotypes was very close to TM-1 having only an average GD of 0.12–0.14. Once again, Australian varieties had the highest GD with TM-1 line (Table 4; Fig. 1).

Analysis of molecular variance (AMOVA)

To estimate genetic diversity within and among predefined groups (Uzbekistan, Latin American and Australian cotton accessions), we also analyzed Wright's F_{ST} index using

Table 3 Correlation of fiber quality traits from the Uzbekistan (Uzb) and Mexican (Mex) environments

Traits	Uzb_MIC	Uzb_UHM	Uzb_UI	Uzb_STR	Uzb_ELO	Uzb_RD	Mex_MIC	Mex_UHM	Mex_UI	Mex_STR	Mex_ELO	Mex_RD
Uzb_MIC	1.00											
Uzb_UHM	−0.04	1.00										
Uzb_UI	0.24**	0.68***	1.00									
Uzb_STR	0.30***	0.26***	0.39***	1.00								
Uzb_ELO	0.27***	0.02	0.26***	0.28***	1.00							
Uzb_RD	0.07	0.48***	0.45***	0.30***	0.11	1.00						
Mex_MIC	**0.20****	−0.25	−0.11	−0.06	0.11	−0.16*	1.00					
Mex_UHM	−0.03	**0.59*****	0.38***	0.29***	0.12	0.36***	−0.33***	1.00				
Mex_UI	0.00	0.09	**0.18***	0.14	0.23**	0.09	0.16*	0.37***	1.00			
Mex_STR	0.01	0.52***	0.41***	**0.50*****	0.09	0.37***	−0.33***	0.71***	0.10	1.00		
Mex_ELO	−0.01	0.10	0.07	−0.17*	**0.35*****	−0.02	0.13	0.08	0.33***	−0.18*	1.00	
Mex_RD	0.04	0.40***	0.39***	0.31***	0.11	**0.84*****	−0.12	0.37***	0.18**	0.39***	0.06	1.00

MIC Micronaire; *UHM* fiber length; *UI* uniformity; *STR* fiber strength; *ELO* elongation; *RD* reflectance; * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.0001$; the same trait correlations in two environments were bolded

AMOVA test. Although differentiation among groups was highly significant ($P \leq 0.0001$), 91.76% of total genetic variance was attributed to the difference within groups, and only 8.24% variance was observed among the predefined groups (Table 5). The total F_{ST} value was equal to 0.0824 ($P < 0.0001$). We observed highly significant ($P \leq 0.0001$) genetic variation of 6–10% between Uzbekistan, Latin American and Australian cotton cultivars with the highest genetic distance ($F_{ST} = 0.095$, $P \leq 0.0001$) observed between Uzbekistan and Latin American cultivar groups (Table 6).

Additionally, due to dominantly scored data, population differentiation estimates (F_{ST}) between the predefined cotton germplasm groups were also analyzed using a Bayesian approach for dominant markers without prior knowledge of inbreeding coefficients. Several Bayesian models were run and the smallest deviance information criterion (DIC = 4382.43) was observed with full model. The results of the $\theta^{(II)}$ (analogous to Weir and Cockerham's F_{ST}) test that corresponds to the amount of genetic differentiation among groups and the Bayesian analog of Nei's G_{ST} (Nei 1973) revealed the low level of population differentiation (~5%) among the predefined groups of cotton germplasm in our study (data not shown). Although the greatest proportion of genetic variance of cotton germplasm groups was attributed to within population groups, those small variations observed among predefined groups (both at the in group specific and pairwise level) were highly significant ($P \leq 0.0001$), suggesting the existence of population structure.

Assessment of population structure and kinship

Since the memberships of individuals to specific clusters (Q-matrix) and the relatedness of individuals (K-matrix) are crucial when conducting LD-based association mapping (Pritchard et al. 2000; Yu et al. 2006; Zhao et al. 2007), we measured these parameters. The model-based approach ($K = 3$) using prior population information revealed shared ancestry information among the predefined groups of Uzbekistan, Latin American, and Australian cotton accessions (Fig. 2). All 303 Uzbek variety cotton accessions were assigned to cluster 1 with the minimum probability of 0.57, out of which 264 accessions had more than a 80% probability of being in cluster 1. Thirty-six of the Uzbek cotton accessions may have shared a recent ancestry with the Australian accessions and other four Uzbek cotton varieties may share recent ancestry with the Latin American accessions. One Uzbek accession shared the recent ancestry from both Latin American and Australian clusters, suggesting the use of one or more common ancestral stocks in the Uzbek and Australian Upland cotton breeding programs (Fig. 2).

Table 4 Summary of genetic distances (GD) of studied cotton accessions obtained from NJ analysis (Saitou and Nei 1987)

Accession	Overall GD		GD with *G. hirsutum* (TM-1)		GD with *G. barbadense* (3–79)		GD with *G. tomentosum*		No. of accessions
	Range	Average	Range	Average	Range	Average	Range	Average	
Variety (All)	0.005–0.26	0.12	0.1–0.23	0.13	0.40–0.47	0.43	0.35–0.44	0.39	337
Uzbekistan varieties	0.005–0.25	0.12	0.1–0.19	0.13	0.40–0.47	0.43	0.35–0.44	0.39	303
Latin American varieties	0.05–0.22	0.13	0.1–0.15	0.12	0.41–0.45	0.43	0.38–0.41	0.39	18
Australian varieties	0.02–0.26	0.14	0.1–0.23	0.14	0.41–0.45	0.43	0.37–0.42	0.38	13
TM-1	–	–	–	–	0.42	–	0.38	–	

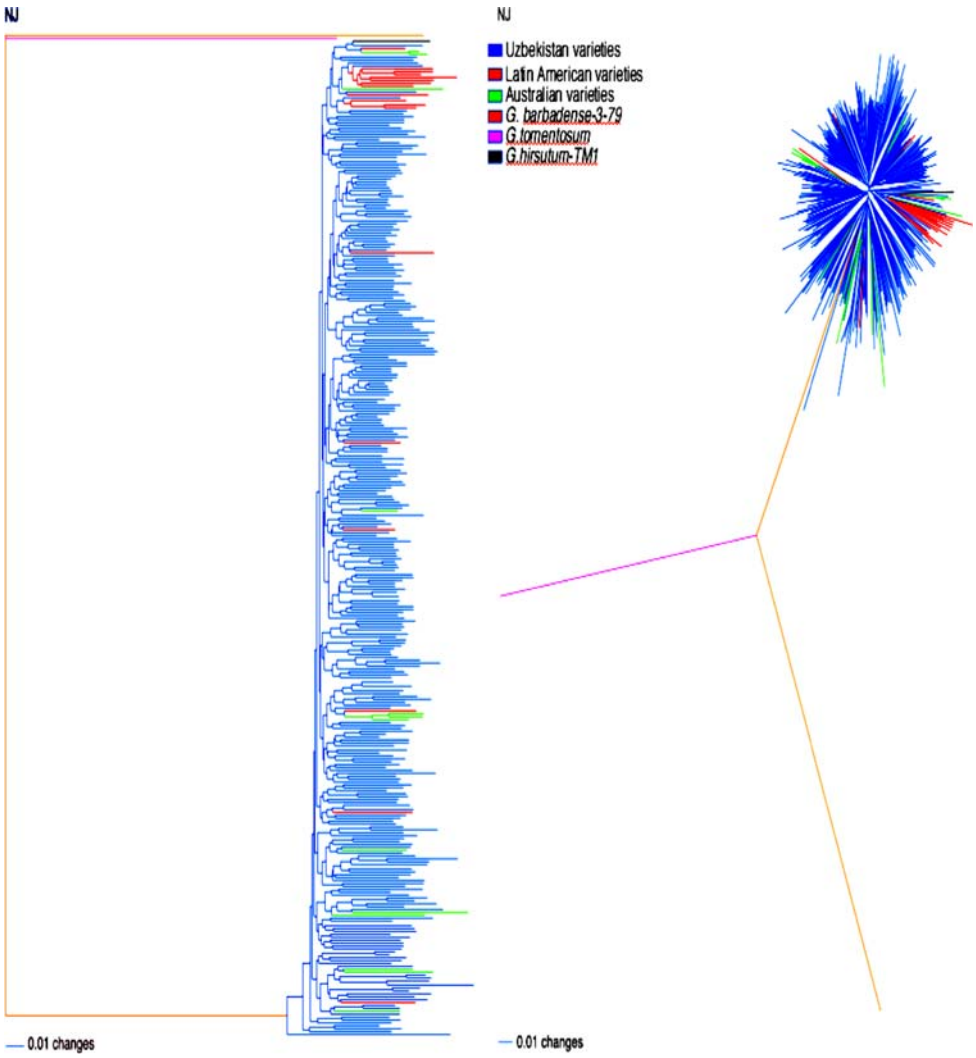

Fig. 1 Rooted and Unrooted Neighbor-joining (NJ) trees for the variety cotton accessions including 337 accessions; the control lines and ecotypes are *color* coded for simplicity. *Branch length* is shown

Table 5 Analysis of molecular variance (AMOVA)

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation	*P*-value
Among populations	2	307.252	3.84392	8.24	<0.0001
Within populations	332	14,211.733	42.80642	91.76	<0.0001
Total	334	14,518.985	46.65034		

Table 6 Pairwise and population specific F_{ST} estimated based on Weir and Cockerham approach

Groups	Uzbekistan	Latin American	Australian
Uzbekistan (304)	0.082*		
Latin American (18)	0.095*	0.084*	
Australian (13)	0.066*	0.060*	0.084*

Diagonal elements are population specific F_{ST}; below diagonal elements are pairwise F_{ST}; * Significant at $P < 0.0001$

All 18 varieties from the Latin American ecotype were assigned to the second cluster with the minimum probability of 59%. In this cluster, three cotton accessions shared recent ancestry with the Uzbek cluster, and there was a zero posterior probability that the Latin American cotton varieties had a recent ancestry with the Australian cluster. All of the Australian varieties were assigned to cluster 3 with a minimum probability score of 0.57. In this Australian cluster, two Australian cotton accessions shared a putative ancestor with the Uzbek varieties and five accessions may have shared recent ancestry with Latin American cottons.

The pairwise kinship values varied 0–0.7. We observed that the majority of the pairs of cotton accessions (55%) had zero estimated kinship values, while 22–23% of the pairs had a value of 0.05, and 20% of the pairs had a value 0.1–0.20. The remaining pairs of accessions (1–2%) had >0.25 kinship values, suggesting involvement of some common parental genotypes in the breeding history of these germplasm groups.

Pairwise linkage disequilibrium and LD decay

In a total of 335 variety accessions of *G. hirsutum* and 40,470 pairwise comparisons of 285 highly polymorphic SSR marker loci, 22, 13 and 9% of SSR marker pairs were in significant LD at $P \leq 0.05$, $P \leq 0.01$ and $P \leq 0.005$, respectively. Based on r^2 estimates, only 4% of SSR marker pairs showed significant LD at $r^2 \geq 0.05$ and ~1% of marker pairs were in LD at $r^2 \geq 0.1$. We also tried to determine the structure of haplotypic LD in the genome since a strong block-like LD structure simplifies LD mapping of complex traits (Zhang et al. 2002). Triangle plots for pairwise LD between SSR markers demonstrated significant LD blocks in the genome-wide LD analysis (Fig. S1). To identify the average genome wide sizes of these blocks, or so called a genome-wide LD decay, r^2 LD values were plotted as a function of genetic distance in cM (Fig. 3). The

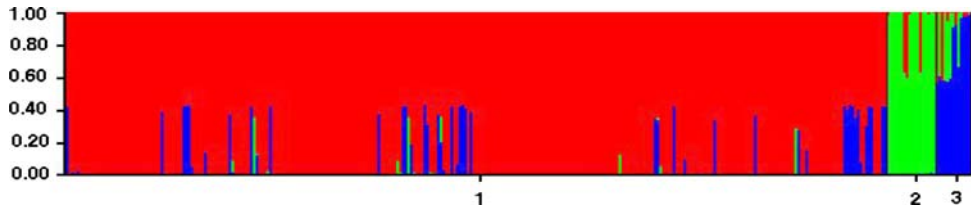

Fig. 2 The summary plots of *Q*-matrix estimates for the variety accessions: cluster 1-Uzbekistan (defined with the *red color*); cluster 2- Latin American (defined with *green color*); cluster 3-Australian ecotypes (defined with *blue color*)

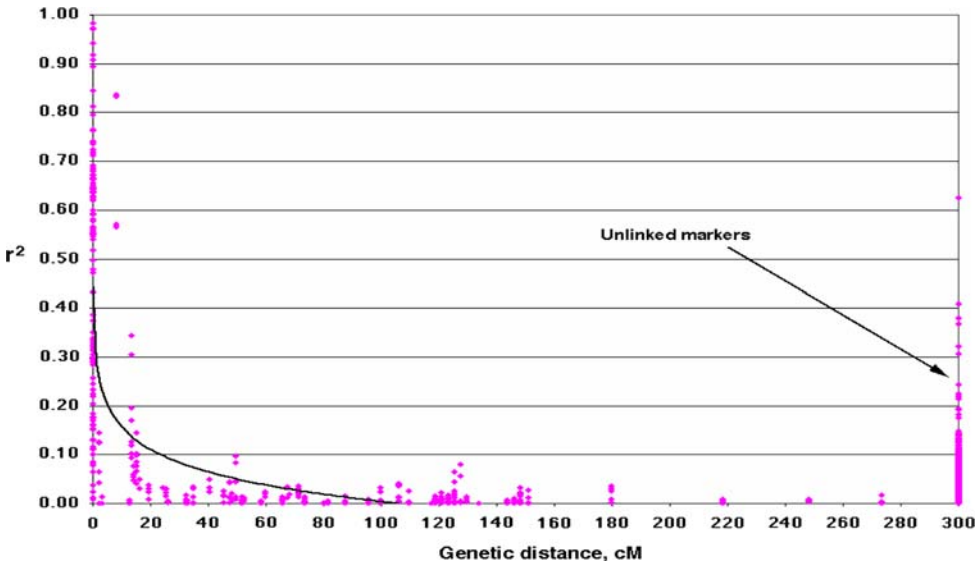

Fig. 3 LD decays within a distance. Inner fitted trend *line* is a non-linear logarithmic regression curve of r^2 on genetic distance. LD-decay is considered below $r^2 = 0.1$ threshold based on trend *line* (Witt and Buckler 2003)

significant pairwise LD ($r^2 \geq 0.1$) was observed between some SSR loci pairs within 50 cM distance. However, genome-wide averages of this pairwise LD decays within the genetic distance at 25 cM with $r^2 \geq 0.1$ (Witt and Buckler 2003), demonstrating the portion of significant LD generated by linkage. Genome wide LD at $r^2 \geq 0.2$ was reduced to ~5–6 cM revealing potential for association mapping. We observed a number of unlinked markers (located in different chromosomes) showing significant LD between pairs of loci (Fig. 3), suggesting the existence of LD generating factors other than linkage in the cotton genome. A separate chromosome-wise analysis revealed non-uniform distribution of LD in cotton chromosomes, but due to unequal number of markers assayed per each chromosome we did not present or discuss the results.

Association mapping of fiber quality traits

We performed association mapping of SSR loci with fiber quality (MIC, UHM, UI, STR, ELO, Rd) traits from the two environments using the MLM, GLM, and SA analyses implemented in TASSEL. In all three independent models of association analyses, out of ~285 highly polymorphic SSR loci used for association mapping, 22 (~8%) SSRs were associated with MIC, 28 (~10%) were associated with UHM, 12 (4%) were associated with UI, 24 (8%) were associated with STR, 21 (7%) were associated with ELO, and 13 (~5%) were associated with Rd traits assayed in the Uzbekistan environment. Likewise, between 5 and 9% of the SSRs showed significant associations with the one of each fiber traits assayed in the Mexican environment (Table 7). The portion of these SSR markers, associated with fiber quality traits at $P \leq 0.01$ significance level in MLM, explained between 1 and 6% phenotypic variation in Uzbekistan environment, and between 2 and 6% in the Mexican environment. The MLM test in fiber trait associations explained 7–53% phenotypic variations (data not shown) of assayed fiber traits in both environments. Although correlation of association mapping results between two diverse environments was not significant with all traits (data not shown), 3 SSRs associated with MIC, 11 SSRs associated with UHM, 7 SSRs associated with STR, and 5 SSRs associated with ELO traits were common in both Uzbekistan and Mexican environments (Tables 7, 8).

The MLM generates more accurate correlations with less inflated Type I error (Yu et al. 2006). Between 78 and 96% of the MLM-derived significant ($P \leq 0.05$) associations for fiber traits in our study were supported by 1,000-time permutated *P*-values of the GLM and SA tests at $P \leq 0.05$ (Table S3), suggesting usefulness of majority of the MLM-derived associations in two distinct environments. However, the MLM-derived significant associations are subject for multiple testing corrections (J. Yu, personal communication). We used more precise Sidak approach of Bonferroni adjustment (data not shown), p_{ACT} method (data not shown) and pFDR approach (Table S3) to adjust the MLM-derived *P*-values for multiple testing. These three methodologies performed almost similarly to adjust *P*-values of MLM. In that, most of the MLM-derived significant associations did not tolerate (at $q \leq 0.05$ level) multiple testing corrections, leaving only one SSR marker (BNL3661_199) associated with UHM in the Mexican environment, three SSR markers (BNL3661_199, CIR166_111, BNL3661_197) associated with STR in the Mexican environment, and four SSR markers (CIR166_111, JESPR65_108, BNL827_246, CIR166_114) associated with Rd in the Uzbekistan environment (see Table S3). Further, to validate the MLM-derived significant associations, we also calculated a BF_{min} from the MLM-derived *P*-values as a symmetric decreasing prior BF (Katki 2008; personal communication). Between 7 and 43% of the MLM-derived significant associations (Table 7) were supported by a BF_{min} at 'moderate to strong' ($BF_{min} \leq 0.13$) and 'strong to very strong' ($BF_{min} \leq 0.05$)

Table 7 Summary of the association mapping of fiber quality traits using MLM, GLM and SA tests in TASSEL

Traits	No. of significant associations*		Common markers***	Number of BF_{min} supported associations ($\leq$0.13)	
	Uzbekistan environment	Mexican environment		Uzbekistan environment	Mexican environment
Micronaire (MIC)	22 (12)**	16 (12)	3 (3)	8 (3)	8 (7)
Fiber length (UHM)	28 (13)	26 (12)	11 (5)	13 (8)	9 (3)
Uniformity (UI)	12 (3)	18 (9)	–	4 (1)	7 (4)
Fiber strength (STR)	24 (11)	23 (10)	7 (2)	3 (1)	9 (2)
Elongation (ELO)	21 (8)	27 (18)	5 (3)	6 (2)	5 (4)
Reflectance (Rd)	13 (6)	15 (7)	–	7 (2)	2 (1)

* Note that in this table, only a number of the MLM-derived significant associations ($P \leq 0.05$) supported by 1,000-time permutated *P*-values ($P \leq 0.05$) of the GLM and SA tests are given (see supplementary materials for TableS3); ** In parenthesis, the number of fiber related SSRs that coincided with published in literature is given; *** Number of SSR loci showing the same-trait associations (MLM, $P \leq 0.05$) in the both environments (refer to Table 8)

Table 8 SSR markers associated with the same fiber quality traits in the Mexican and Uzbekistan environments

#	Marker names	Mexican environment		Uzbekistan environment	
		F-value (MLM)	*P*-value (MLM)	*F*-value (MLM)	*P*-value (MLM)
Micronaire (MIC)					
1	BNL3255_230	11.9	0.00068$^{\ddagger gs}$	5	0.0266^{gs}
2	BNL2986_156	9.6	0.0022$^{\ddagger s}$	5.5	0.0194^{gs}
3	BNL2986_154	4.2	**0.0421**	9.4	0.0024$^{\ddagger gs}$
Length (UHM)					
1	BNL3661_199	25.5	0.0000009$^{\ddagger gs¥}$	8.7	0.0034$^{\ddagger gs}$
2	BNL3661_197	11.4	0.0009$^{\ddagger gs}$	3.9	0.0483^{gs}
3	BNL3545_187	8.3	0.0043$^{\dagger gs}$	8.1	0.0047$^{\dagger gs}$
4	BNL1122_170	7.3	0.0075$^{\dagger gs}$	8.1	0.0047$^{\dagger gs}$
5	BNL1604_116	7	0.0087$^{\dagger gs}$	11.1	0.0010$^{\ddagger gs}$
6	JESPR92_378	6.2	0.0137^{gs}	3.8	0.0510^{gs}
7	BNL2986_156	5.8	0.0167^{gs}	8.2	0.0046$^{\dagger gs}$
8	BNL3806_199	5.6	0.0191^{gs}	5.6	0.0190^{gs}
9	BNL1694_237	5.2	0.0233^{gs}	7	0.0088$^{\dagger sg}$
10	BNL4108_172	5.2	0.0237^{gs}	9.7	0.0020$^{\ddagger sg}$
11	BNL3650_337	4.7	0.0310^{gs}	8	0.0050$^{\dagger sg}$
Strength (STR)					
1	BNL3661_199	16.8	0.00006$^{\ddagger gs¥}$	4	0.0474^{gs}
2	BNL3661_197	12.3	0.0005$^{\ddagger gs¥}$	5.1	0.0243^{gs}
3	BNL2571_253	10.9	0.0011$^{\ddagger gs¥}$	4.2	0.0421^{gs}
4	BNL1122_170	8.8	0.0032$^{\ddagger gs}$	5.3	0.0219 ^{gs}
5	CIR347_246	5.1	0.0246^{gs}	4	0.0457^{gs}
6	CIR347_249	4.4	0.0374^{gs}	6.6	0.0107$^{\dagger gs}$
7	BNL3806_199	3.9	0.0494^{sg}	5.4	0.0207^{gs}
Elongation (ELO)					
1	BNL3650_337	12.5	0.0005$^{\ddagger gs}$	4.4	0.0363$^{\ddagger s}$
2	JESPR229_110	5.5	0.0200^{gs}	5	0.0261^{gs}
3	BNL3482_137	4	**0.0456**	6.7	0.0103$^{\dagger gs}$
4	BNL4108_166	4	0.0469^{g}	5.1	0.0252^{gs}
5	CIR180_220	3.9	0.0501^{gs}	6.6	0.0109$^{\dagger gs}$

‡ BF_{min} with strong to very strong evidence for association ($\leq$0.05); † BF_{min} with moderate to strong evidence for association (>0.05–0.13); ^{g} significant in GLM test after 1,000-time permutation test at $P \leq 0.05$; ^{s} significant in SA test after 1,000-time permutation test at $P \leq 0.05$; $^{¥}$ significant in pFDR test at $q \leq 0.05$. Bold-faced values significant in MLM test ($P \leq 0.05$), but were not supported by either BF_{min}, or permutated GLM and SA tests in Mexican environment; they were included because of their significance in Uzbekistan environment

evidence level (Goodman 2001; Table S3), suggesting a reliable portion of the MLM-derived significant associations.

We also compared fiber trait-associated SSR markers from our study with reported SSR markers from QTL-mapping analyses in various experimental populations (Han et al. 2004; Mei et al. 2004; Shen et al. 2005; Lacape et al. 2005; Zhang et al. 2005a; He et al. 2005, 2007; Lin et al. 2005; Wang et al. 2006; Abdurakhmonov et al. 2007a). Other published QTL-mapping studies were not compared because of the use of a different marker system. Thorough analysis of the literature data revealed that the majority of SSRs (or their alleles) showing association with the main fiber quality traits in the Mexican and Uzbekistan environments were also found to be associated with fiber quality traits in other linkage-mapping studies in cotton (Table S3; see for example Fig. S2). Twelve (54%) SSR markers associated with MIC trait, 13 (46%) SSRs associated with UHM, 3 (25%) SSRs associated with UI, 11 (45%) SSRs associated with STR, 8 (38%) SSRs associated with ELO and 6 (46%) SSRs associated with RD traits in the Uzbekistan environment coincided with earlier reported SSRs (or their alleles) linked to fiber quality (Table 7, shown in parenthesis; Fig. S1). Likewise, a number of SSR markers associated with fiber quality in the Mexican environment in our study coincided with previously

reported SSR markers of various QTL-mapping studies for fiber quality traits (Table 7). This supports the LD-based association mapping results (Kraakman et al. 2004) of our study that used diverse sets of cultivated cotton germplasm resources. The remaining SSRs are new unreported markers revealing associations with fiber quality traits in diverse set of cotton germplasm (Table S3). We also checked whether those SSRs markers showing significant associations in diverse environments are in pairwise LD or not. The comparison of LD values (r^2) and association results from the MLM ($P \leq 0.05$) revealed that 85% of the SSRs significantly associated with fiber quality traits in MLM were in significant LD with other SSR loci at $r^2 \geq 0.1$ (Fig. S2 and Fig. S3).

Discussion

We explored the Uzbek cotton germplasm collection, one of the largest, with a broad geographic and genetic diversity coverage that provided us an opportunity to identify the extent of a genome-wide LD and apply an 'association-mapping' approach in cotton. The application of LD-based association mapping for cotton facilitates comprehensive utilization of natural genetic diversity conserved within cotton germplasm collections worldwide (Abdurakhmonov 2007), as the case with in other plant germplasm resources (Rafalski and Morgante 2004). Assessing a moderately large number of cotton accessions in our study, we observed a wide range of diversity in fiber quality traits within specific environments and between environments demonstrating (1) existence of potential genetic variation for primary fiber quality traits within *G. hirsutum* germplasm that is useful for future breeding programs; and (2) the influence of environmental factors in the development of fiber quality traits. Fiber quality trait correlations within the Uzbekistan and Mexican environments, as well as between these two diverse environments suggested different performance of the same cotton cultivars and genetic make-up in the specific environment (s) that should be taken into account when breeding for these complex traits (Zhang et al. 2005b). It should be mentioned that we used only two environments fiber trait data in our study due to funding constraints of growing of these accessions in multiple environments in both Uzbekistan and Mexico. This is the first time effort in phenotypic measurements of fiber quality traits of Upland cotton accessions in such moderately large samples.

The genetic distance analysis based on SSR markers further revealed the narrow genetic clustering all of variety accessions in our study at molecular level. A number of studies on the genetic diversity of *Gossypium* species revealed a low level of genetic diversity within Upland cotton germplasms inferred from isozymes (Wendel et al. 1992), random amplified polymorphisms (RAPDs; Tatineni et al. 1996; Iqbal et al. 1997, 2001), restricted fragment length polymorphisms (RFLPs; Brubaker and Wendel 1994), amplified fragment length polymorphisms (AFLPs; Pillay and Myers 1999; Abdalla et al. 2001) and SSRs (Liu et al. 2000b; Gutierrez et al. 2002; Zhang et al. 2005b; Rungis et al. 2005; Lacape et al. 2007; Abdurakhmonov et al. 2007b). Our results obtained from phylogenetic analysis of an Upland cotton variety germplasm from diverse ecotypes/breeding programs (Uzbekistan, Australia, Latin America) further confirms the narrow genetic base of Upland cotton variety accessions and supports an evidence for the occurrence of a genetic 'bottleneck' during domestication events of the Upland cotton with rigorous selection for early maturity (Iqbal et al. 2001).

The extent of genome-wide LD and potential cause of LD in cotton genome

We scored our SSR data as a dominant marker type. Although a dominant type of coding has limited statistical power compared to co-dominant markers in population-based analyses due to missing heterozygote information, previous studies suggested that it can be successfully applied to in genetics analyses (Pritchard et al. 2000; Hollingsworth and Ennos 2004; Hardy 2003), including LD quantification and association mapping (Hansen et al. 2001; Kraakman et al. 2004, 2006; Malosetti et al. 2007; Iwata et al. 2007; Tommasini et al. 2007; Li et al. 2007; Abdurakhmonov and Abdukarimov 2008) with the use of a large number of loci and samples. Accordingly, the moderately large sample size and the large number of heterozygous SSR loci used in this study should give unbiased estimates of the genetic distance, genome-wide LD, population structure, and kinship.

The amplified SSR fragments per primer pairs in the cotton accessions (2–15 marker/primer pairs) were comparable with those reported in literature (Liu et al. 2000b; Lacape et al. 2007). The percentage of SSR loci pairs in LD observed in cotton was comparable with reports in maize (10%) (Remington et al. 2001), and sorghum (8.7%) (Hamblin et al. 2004), yet it was comparatively lower than that obtained in other studies. In different maize population groups, 49–56% of the SSR pairs were in significant LD (Stich et al. 2005, 2006). Also a high percentage of SSR pairs in LD was reported for population groups of cultivated barley (45–100%) germplasm (Kraakman et al. 2004; Malysheva-Otto et al. 2006) as well as for the drum wheat elite germplasm (52–86%) (Maccaferri et al. 2005). The low percentage of pairwise LD between SSR loci in cotton could also be associated with higher recombination rate observed in allopolyploid cottons (Brubaker et al. 1999) as

well as mutation, and experimental hybridizations undertaken in the recent breeding history of Upland cotton germplasm. This is the first reported insight into a genome-wide LD level for cotton measured with SSR markers.

The size of LD blocks in plants is largely influenced by a recombination rate, mating system (selfing vs. outcrossing), genetic isolation, population subdivision and admixture, selection, mutation, and effective population sizes (Tenaillon et al. 2001; Ching et al. 2002; Liu et al. 2003; Rafalski and Morgante 2004). The genome-wide averages of LD block size for cotton was comparable with LD decay estimates reported in some local *Arabidopsis* populations, sugar cane, sorghum, barley, durum wheat, and grapevine (Nordborg et al. 2002; Flint-Garcia et al. 2003; Kraakman et al. 2004; Malysheva-Otto et al. 2006; Maccaferri et al. 2005; Barnaud et al. 2006), yet was larger than those reported in *Arabidopsis*, maize, sugar beet, and wheat (Kraft et al. 2000; Remington et al. 2001; Breseghello and Sorrells 2006). It should also be noted that our estimate of genome-wide averages for the extent of LD in cotton may not represent LD patterns of specific genomic regions or population groups. Each of these specific regions or populations of interest might require a quantification of LD patterns for successful association mapping.

The decay of LD with the genetic distance demonstrates that linkage is the main factor in conserving LD between SSR loci in cotton that is useful for a genome-wide association mapping (Stich et al. 2005, 2006). However, we observed a number of unlinked markers showing significant LD ($r^2 \geq 0.1$) between pairs of SSR loci, revealing the existence of other LD generating factors than linkage in the cotton genome (Nordborg et al. 2002; Stich et al. 2006). One of those factors is selection since we observed a number of unlinked marker pairs in significant LD. However, this might also be the result of co-selection of loci during breeding for multiple traits (co-adapted genes) that is common in cotton breeding programs worldwide. Relatedness is another factor that might also generate LD between unlinked loci pairs when predominant parents exist in germplasm groups. In cotton accessions tested about 22% accession pairs had ≥10% kinship values, suggesting the potential effect of relatedness in observed LD. The other factors such as genetic drift or bottlenecks might have also generated LD in the *G. hirsutum* cotton genome (Huttley et al. 1999; Stich et al. 2005). There is an evidence of a genetic bottleneck occurring in the domestication of Upland varieties because of rigorous selection for early maturity (Lewis 1962; Iqbal et al. 2001). Although we did not separately estimate a pairwise LD level of specific predefined groups due to small sample sizes and dominantly scored SSR data, theoretically, population stratification is the one of the main forces generating LD (Stich et al. 2005, 2006). LD generated by selection, population stratification, and genetic drift might be theoretically useful for association mapping (less number of markers are needed) in a specific population groups and situations, yet it tends to reveal spurious marker-trait associations (Pritchard et al. 2000; Stich et al. 2005, 2006). This underlies the necessity for serious consideration of population structure and relatedness to perform population-based association mapping in cotton germplasm resources, at least, in our samples.

Prospects of association mapping in cotton and its application in mapping of fiber quality traits

In contrast to the human genome, where a very high density of molecular markers is needed for association mapping in the majority of cases (Kruglyak 1999), the cotton genome may require significantly fewer numbers of markers for effective LD–mapping of complex traits, which is also the case reported for other crops (Kraakman et al. 2004; Barnaud et al. 2006). Considering the tetraploid cotton genome with a total recombinational length of about 5,200 cM and an average 400 kb per cM (Paterson and Smith 1999), the high threshold ($r^2 \geq 0.2$) LD block sizes of ~5–6 cM distance in cotton is large enough to conduct an association mapping of complex traits that would require a maximum of ~1,000 polymorphic markers for successful association mapping. This number could be cut down to ~200–250 markers if the extent of LD were considered at the $r^2 \geq 0.1$ threshold, which is extended up to a 25 cM distance. However, this conclusion is preliminary due to the polyploid nature of the Upland cotton genome, dominantly coded SSR data, and our inability to identify the linkage phase of marker data in our study. This needs to be further explored.

Because of the existing population stratification attributes and unbalanced number of accessions in our cotton germplasm groups, we applied the MLM approach of Yu et al. (2006) considering both population structure (Q) and kinship (K) to eliminate possible spurious associations. We identified a number of SSR markers significantly associated with the main fiber quality traits of cotton in the two diverse environments. Nevertheless, majority of the MLM-derived significant associations of fiber quality traits in both environments did not tolerate to multiple testing corrections using different statistical methodologies, suggesting only a few markers could represent reliable associations. The results of correction for multiple testing, however, could be misleading due to (1) known conservativeness of multiple testing adjustment methodologies such as Bonferroni corrections and (2) unknown influence of other *P*-value adjustment methods used to the model based MLM approach. In that, results have already been corrected to minimize the spurious associations, which

represent less inflated Type I error (Yu et al. 2006; Zhao et al. 2007). Perhaps, a modified statistical approach (e.g. Bonferroni adjustment with a reduced number of independent tests) should be applied to adjust MLM *P*-values, which requires further studies (Dr. J. Yu, personal communication). Hence, to reliably interpret the MLM-derived significant associations in our study, we used minimum BF estimation for the MLM association results because "even without formal Bayesian analysis, the use of minimum Bayes factor may provide an antidote for the worst inferential misdeeds" (Goodman 2001). Although it is not a perfect, minimum BF estimates over *P*-values of the MLM approach in our study, may help to understand overall impact of our associations (Katki 2008). Accordingly, at least, a portion of significant MLM-associations in our study, which were supported by BF_{min} estimate at 'moderate to strong' and 'strong to very strong' level, should be meaningful correlations.

Additionally, more support and validity of the MLM-derived significant associations in our study come from following other evidences: (1) the majority of the significant associations found using the MLM approach were also supported by the GLM and SA tests, accounting for population structure and multiple test issues. (2) The MLM correlation identified several common SSR markers showing significant association (supported also by both GLM and SA) with the fiber quality traits (MIC, UHM, STR, ELO) in both the Uzbekistan and Mexican environments implying not only association of these markers with structural fiber gene(s), but also correctness of the model that validates the association results (Table 8). Furthermore, (3) the majority of SSRs associated with the main fiber quality traits in our study coincided with reported fiber quality trait-associated SSRs from QTL-mapping studies in various experimental populations, which is useful in the judging of the association results (Kraakman et al. 2004). At the same time, we detected additional, new, and unreported SSR markers associated with fiber quality traits in our cotton germplasm. The common SSRs showing significant associations in both environments are potential markers for effective MAS programs in cotton. In contrast, SSR markers detecting significant associations only in one specific environment (in the case with UI and Rd traits) suggested the importance of environment on these two traits, and the environment specific expression of QTLs responsible for these traits that needs to be considered in future MAS programs.

Currently, a very few SSR markers are efficiently used in MAS programs in cotton because majority of marker information has been derived from populations of bi-parental crosses with limited genetic background, covering a few meiotic events since experimental hybridization. Initial association study in cotton (Kantartzi and Stewart 2008) reported several SSR markers associated with fiber quality traits in diploid cotton germplasm. Compared to our study, those markers were obtained from association analysis of significantly less number of cotton germplasm resources. SSR markers associated with the main fiber quality traits in our study were detected across the genetic background of several hundred diverse cotton accessions from different geographic locations, enabling us to utilize multiple meiotic events occurred since history of germplasm development. Also these markers provided a broad scan view of variation associated with important QTL regions of the cotton genome because these markers were pre-selected from previously published reports as a prior linked markers with important fiber QTLs that increase the power of association mapping (Ball 2005). Hence, these markers should have more potential to be efficient markers for MAS programs and provide very valuable information useful for genetic improvement of Upland cotton.

Taken together, the moderately large extent of LD in the cotton genome illustrates the significant potential for LD-based association mapping of complex traits in cotton with a relatively small number of markers. Conversely, the mapping resolution may be limited, in particular, with breeding germplasm. The potential LD generating forces in cotton, discussed herein, suggest that population structure and relatedness should be taken into serious account to perform unbiased population-based association mapping in cotton germplasm resources. The first insights of LD-based association mapping using MLM approach, reported herein, highlight the potential and feasibility of this approach in cotton. A number of SSR markers associated with fiber quality traits, discovered using *G. hirsutum* germplasm resources in the two very diverse environments, will not only be useful for marker-assisted breeding programs in the development of cultivars with superior fiber qualities, but will also provide insights on the environment-specific functions of genes controlling the fiber development.

Acknowledgments We acknowledge the Science and Technology Center of Ukraine for the project coordination, and the technical assistance of project participants of P120/P120A. We also thank the administration of Academy of Sciences of Uzbekistan for their continual support of the research efforts. We thank Drs. A. Abdullaev (Uzbek cotton germplasm curator), and S. M. Rizaeva for valuable suggestions on germplasm selection and phenotypic analyses in the Uzbekistan environment. We thank Ms Linda Ballard, Genomics laboratory, USDA ARS at Mississippi for useful suggestions during manuscript edition.

Funding This project was funded by the Office of International Research Programs (OIRP) of United States Department of Agriculture (USDA) in the frame of USDA-Former Soviet Union (FSU) cooperation programs with the research grant of P120/P120A to IYA, RJK, JZY, SS, and AEP.

Disclaimer Mention of trademark or proprietary product does not constitute a guarantee or warranty of the product by the United States Department of Agriculture and does not imply its approval to the exclusion of other products that may also be suitable.

References

Abdalla AM, Reddy OUK, El-Zik KM, Pepper AE (2001) Genetic diversity and relationships of diploid and tetraploid cottons revealed using AFLP. Theor Appl Genet 102:222–229. doi:10.1007/s001220051639

Abdurakhmonov IY (2007) Exploiting genetic diversity. In: Ethridge D (ed) Plenary presentations and papers. Proceedings of World Cotton Research Conference-4. Lubbock, TX, USA, 10–14 Sept 2007 p 2153

Abdurakhmonov IY, Abdukarimov A (2008) Application of association mapping to understanding the genetic diversity of plant germplasm resources. Int J Plant Genomics 2008:574927

Abdurakhmonov IY, Buriev ZT, Saha S, Pepper AE, Musaev JA, Almatov A et al (2007a) Microsatellite markers associated with lint percentage trait in cotton, *Gossypium hirsutum*. Euphytica 156:141–156. doi:10.1007/s10681-007-9361-2

Abdurakhmonov IY, Kushanov FN, Djaniqulov N, Buriev ZT, Pepper AE, Fayzieva N et al (2007b) The role of induced mutation in conversion of photoperiod dependence in cotton. J Hered 98:258–266. doi:10.1093/jhered/esm007

Ball RD (2005) Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. Genetics 170:859–873. doi:10.1534/genetics.103.024752

Barnaud AT, Lacombe , Doligez A (2006) Linkage disequilibrium in cultivated grapevine, *Vitis vinifera* L. Theor Appl Genet 112:708–716. doi:10.1007/s00122-005-0174-1

Blenda A, Scheffler J, Scheffler B, Palmer M, Lacape JM, Yu JZ et al (2006) CMD: a cotton microsatellite database resource for *Gossypium* genomics. BMC Genomics 7:132. doi:10.1186/1471-2164-7-132

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. doi:10.1093/bioinformatics/btm308

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165–1177. doi:10.1534/genetics.105.044586

Brubaker CL, Wendel JF (1994) Re-evaluating the origin of domesticated cotton (*Gossypium hirsutum*, Malvaceae) using nuclear restriction fragment length polymorphism (RFLP). Am J Bot 81:1309–1326. doi:10.2307/2445407

Brubaker CL, Paterson AH, Wendel JF (1999) Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome 42:184–203. doi:10.1139/gen-42-2-184

Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang T, Guo W et al (2007) Toward sequencing cotton (*Gossypium*) genomes. Plant Physiol 145:1303–1310. doi:10.1104/pp.107.107672

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S et al (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:1–14. doi:10.1186/1471-2156-3-19

Conneely KN, Boehnke M (2007) So many correlated tests, so little time! Rapid adjustment of *P* values for multiple correlated tests. Am J Hum Genet 81:1158–1168. doi:10.1086/522036

Culp TW, Lewis CF (1973) Breeding methods for improving yield and fiber quality of upland cotton (*Gossypium hirsutum*). Crop Sci 13:686–689

Dellaporta SL, Wood J, Hicks JP (1983) A plant DNA minipreparation: version II. Plant Mol Biol Rep 1:19–21. doi:10.1007/BF02712670

Excoffier L, Smouse P, Quattro J (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491

Flint-Garcia SA, Thornsberry JM, Buckler ESIV (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374. doi:10.1146/annurev.arplant.54.031902.134907

Goodman SN (2001) Of *P*-values and Bayes: a modest proposal. Epidemiology 12:295–297. doi:10.1097/00001648-200105000-00006

Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol 57:461–485. doi:10.1007/s11103-005-0257-z

Gutierrez OA, Basu S, Saha S, Jenkins JN, Shoemaker DB, Cheatham CL et al (2002) Genetic distance among selected cotton genotypes and its relationship with F2 performance. Crop Sci 42:1841–1847

Hamblin MT, Mitchell SE, White GM, Gallego J, Kukatla R, Wing RA et al (2004) Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum* bicolor. Genetics 167:471–483. doi:10.1534/genetics.167.1.471

Han ZG, Guo W, Song XL, Zhang T (2004) Genetic mapping of EST-derived microsatellites from the diploid *Gossypium arboreum* in allotetraploid cotton. Mol Genet Genomics 272:308–327. doi:10.1007/s00438-004-1059-8

Hansen M, Kraft T, Ganestam S, Sall S, Nilsson NO (2001) Linkage disequilibrium mapping of the bolting gene in sea beet using AFLP markers. Genet Res 77:s61–s66. doi:10.1017/S0016672300004857

Hardy OJ (2003) Estimation of pairwise relatedness between individuals and characterization of isolation by distance processes using dominant genetic markers. Mol Ecol 12:1577–1588. doi:10.1046/j.1365-294X.2003.01835.x

Hardy OJ, Vekemans X (2002) SpaGeDi: a versatile computer program to analyze spatial genetic structure at the individual or population levels. Mol Ecol Notes 2:618–620. doi:10.1046/j.1471-8286.2002.00305.x

He DH, Zhong-Xu L, Zhang XL, Nie YC, Guo XP, Feng CD et al (2005) Mapping QTLs of traits contributing to yield and analysis of genetic effects in tetraploid cotton. Euphytica 144:141–149. doi:10.1007/s10681-005-5297-6

He DH, Lin ZX, Zhang XL, Nie YC, Guo XP, Feng CD et al (2007) QTL mapping for economic traits based on a dense genetic map of cotton with PCR-based markers using the interspecific cross of *Gossypium hirsutum* × *Gossypium barbadense*. Euphytica 153:181–197. doi:10.1007/s10681-006-9254-9

Hollingsworth PM, Ennos RA (2004) Neighbor joining trees, dominant markers and population genetic structure. Heredity 92:490–498. doi:10.1038/sj.hdy.6800445

Holsinger KE, Lewis PO (2003) HICKORY: a package for analysis of population genetic data, version 1.0. Department of Ecology and evolutionary Biology, University of Connecticut, Storrs, Connecticut, USA

Holsinger KE, Lewis PO, Dey DK (2002) A Bayesian approach to inferring population structure from dominant markers. Mol Ecol 11:1157–1164. doi:10.1046/j.1365-294X.2002.01512.x

Huttley GA, Smith MW, Carrington M, O'Brien SJ (1999) A scan for linkage disequilibrium across the human genome. Genetics 152:1711–1722

Iqbal MJ, Aziz N, Saeed NA, Zafar Y, Malik KA (1997) Genetic diversity evaluation of some elite cotton varieties by RAPD

analysis. Theor Appl Genet 94:139–144. doi:10.1007/s001220050392

Iqbal J, Reddy OUK, El-Zik KM, Pepper AE (2001) A genetic bottleneck in the 'evolution under domestication' of Upland cotton *Gossypium hirsutum* L. examined using DNA fingerprinting. Theor Appl Genet 103:547–554. doi:10.1007/PL00002908

Iwata H, Uga Y, Yoshioka Y, Ebana K, Hayashi T (2007) Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms. Theor Appl Genet 114:1437–1449. doi:10.1007/s00122-007-0529-x

Kantartzi SK, Stewart JM (2008) Association analysis of fibre traits in *Gossypium arboreum* accessions. Plant Breed 127:173–179. doi:10.1111/j.1439-0523.2008.01490.x

Katki HA (2008) Invited commentary: evidence-based evaluation of *P* values and Bayes factors. Am J Epidemiol 268:384–388. doi:10.1093/aje/kwn148

Kraakman ATW, Niks RE, Van den Berg PMMM, Stam P, Van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168:435–446. doi:10.1534/genetics.104.026831

Kraakman ATW, Martinez F, Mussiraliev B, van Eeuwijk FA, Niks RE (2006) Linkage disequilibrium mapping of morphological, resistance and other agronomically relevant traits in modern spring barley cultivars. Mol Breed 17:41–58. doi:10.1007/s11032-005-1119-8

Kraft TM, Hansen , Nilsson NO (2000) Linkage disequilibrium and fingerprinting in sugar beet. Theor Appl Genet 101:323–326. doi:10.1007/s001220051486

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144. doi:10.1038/9642

Lacape JM, Nguyen TB, Courtois B, Belot JL, Giband M, Gourlot JP et al (2005) QTL analysis of cotton fiber quality using multiple *Gossypium hirsutum* × *Gossypium barbadense* backcross generations. Crop Sci 45:123–140

Lacape JM, Dessauw D, Rajab M, Noyer JL, Hau B (2007) Microsatellite diversity in tetraploid *Gossypium* germplasm: assembling a highly informative genotyping set of cotton SSRs. Mol Breed 19:45–58. doi:10.1007/s11032-006-9042-1

Lewis H (1962) Catastrophic selection as a factor in speciation. Evol Int J Org Evol 16:257–271. doi:10.2307/2406275

Li Y, Li Y, Han K, Wang Z, Hou W, Zeng Y et al (2007) Estimation of multilocus linkage disequilibria in diploid populations with dominant markers. Genetics 176:1811–1821. doi:10.1534/genetics.106.068890

Lin Z, He D, Zhang X, Nie Y, Guo X, Feng C et al (2005) Linkage map construction and mapping QTL for cotton fiber quality using SRAP, SSR and RAPD. Plant Breed 124:180–187. doi:10.1111/j.1439-0523.2004.01039.x

Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129. doi:10.1093/bioinformatics/bti282

Liu S, Saha S, Stelly D, Burr B, Cantrell RG (2000a) Chromosomal assignment of microsatellite loci in cotton. J Hered 91:326–332. doi:10.1093/jhered/91.4.326

Liu S, Cantrell RG, McCarty JCJ, Stewart JM (2000b) Simple sequence repeat-based assessment of genetic diversity in cotton race stock accessions. Crop Sci 40:1459–1469

Liu KJ, Goodman M, Muse S, Smith JS, Buckler E, Doebley J (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:2117–2128

Maccaferri M, Sanguineti MC, Noli E, Tuberosa R (2005) Population structure and long-range linkage disequilibrium in a drum wheat elite collection. Mol Breed 15:271–289. doi:10.1007/s11032-004-7012-z

Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. Genetics 175:879–889. doi:10.1534/genetics.105.054932

Malysheva-Otto LV, Ganal MW, Roder MS (2006) Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). BMC Genet 7:6. doi:10.1186/1471-2156-7-6

McRae AF, McEwan JC, Dodds KG, Wilson T, Crawford AM, Slate J (2002) Linkage disequilibrium in domestic sheep. Genetics 160:1113–1122

Mei M, Syed NH, Gao W, Thaxton PM, Smith CW, Stelly DM et al (2004) Genetic mapping and QTL analysis of fiber-related traits in cotton (*Gossypium*). Theor Appl Genet 108:280–291. doi:10.1007/s00122-003-1433-7

Mohlke KL, Lange EM, Valle TT, Ghosh S, Magnuson VL, Silander K et al (2001) Linkage disequilibrium between microsatellite markers extends beyond 1 cM on chromosome 20 in Finns. Genome Res 11:1221–1226. doi:10.1101/gr.173201

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323. doi:10.1073/pnas.70.12.3321

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hanblad J et al (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 30:90–193. doi:10.1038/ng813

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H et al (2005) The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol 3:el96. doi:10.1371/journal.pbio.0030196

Paterson AH, Smith RH (1999) Future horizons: biotechnology of cotton improvement. In: Smith CW, Cothren JT (eds) Cotton: origin, history, technology, and production. Wiley, New York, pp 415–432

Pillay M, Myers GO (1999) Genetic diversity in cotton assessed by variation in ribosomal RNA genes and AFLP markers. Crop Sci 39:1881–1886

Preetha S, Raveendren TS (2008) Molecular marker technology in cotton. Biotechnol Mol Biol Rev 3:032–045

Pritchard JK, Wen W (2004) Documentation for Structure software. The University of Chicago Press, Chicago

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multi-locus genotype data. Genetics 155:945–959

Rafalski A, Morgante M (2004) Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet 20:103–111. doi:10.1016/j.tig.2003.12.002

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J et al (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci USA 98:11479–11484. doi:10.1073/pnas.201394398

Reynolds J, Weir BS, Cockerham CC (1983) Estimation for the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:7767–7779

Rungis D, Llewellyn D, Dennis ES, Lyon BR (2005) Simple sequence repeat (SSR) markers reveal low levels of polymorphism between cotton (*Gossypium hirsutum* L.) cultivars. J Agric Res 56:301–307. doi:10.1071/AR04190

Saitou M, Nei N (1987) The neighbor joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sand PG (2007) A lesson not learned: allele misassignment. Behav Brain Funct 3:65. doi:10.1186/1744-9081-3-65

Schneider S, Roessll D, Excoffier L (2000) ARLEQUIN: a software for population genetics data analysis, version 2.0. Genetics and Biometry Laboratory, Department of Anthropology, Geneva, Switzerland, University of Geneva

Schulze TG, McMahon FJ (2002) Genetic association mapping at the crossroad: which test and why? Overview and practical guidelines. Am J Med Genet 114:1–11. doi:10.1002/ajmg.10042

Shen X, Guo W, Zhu X, Yuan Y, Yu JZ, Kohel RJ et al (2005) Molecular mapping of QTLs for fiber qualities in three diverse lines in Upland cotton using SSR markers. Mol Breed 15:169–181. doi:10.1007/s11032-004-4731-0

Stich B, Melchinger AE, Frisch M, Maurer HP, Heckenberger M, Reif JC (2005) Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theor Appl Genet 111:723–730. doi:10.1007/s00122-005-2057-x

Stich B, Maurer HP, Melchinger AE, Frisch M, Heckenberger M, van der Voort JR et al (2006) Comparison of linkage disequilibrium in elite European maize inbred lines using AFLP and SSR markers. Mol Breed 17:217–226. doi:10.1007/s11032-005-5296-2

Storey JD, Tibshirani R (2003) Statistical significance for genome-wide experiments. Proc Natl Acad Sci USA 100:9440–9445. doi:10.1073/pnas.1530509100

Swofford DL (2002) Phylogenetic analysis using parsimony (*and other methods). Sinauer, Sunderland

Tatineni V, Canlrell RG, Davis DD (1996) Genetic diversity in elite cotton germplasm determined by morphological characteristics and RAPD. Crop Sci 36:186–192

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS et al (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Xea mays* ssp. mays L.). Proc Natl Acad Sci USA 98:9161–9166. doi:10.1073/pnas.151244298

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) Dwarf 8 polymorphisms associate with variation in flowering time. Nat Genet 28:286–289. doi:10.1038/90135

Tommasini L, Schnurbusch T, Fossati D, Mascher F, Keller B (2007) Association mapping of *Stagonospora nodorum* blotch resistance in modern European winter wheat varieties. Theor Appl Genet 115:697–708. doi:10.1007/s00122-007-0601-6

Wang B, Guo W, Zhu X, Wu Y, Huang N, Zhang T (2006) QTL mapping of fiber quality in an elite hybrid derived-RIL population of Upland cotton. Euphytica 152:367–378. doi:10.1007/s10681-006-9224-2

Weir BS, Cockerham CC (1984) Estimating *F* statistics for the analysis of population structure. Evol Int J Org Evol 38:1358–1370. doi:10.2307/2408641

Weiss KM, Clark AG (2002) Linkage disequilibrium and mapping of human traits. Trends Genet 18:19–24. doi:10.1016/S0168-9525(01)02550-1

Wendel JF, Brubaker CL, Percival AE (1992) Genetic diversity in *Gossypium hirsutum* and the origin of upland cotton. Am J Bot 79:1291–1310.

Witt SR, Buckler ES (2003) Using natural allelic diversity to evaluate gene function. Methods Mol Biol 236:123–139

Young FW, Bann CM (1996) Data analyses with ViSta. In: Fox J, Stine R (eds) Statistical computing environments for social research. Sage Publications, California, pp 207–235

Yu J, Pressoir G, Briggs WH, Vroh BI, Yamasaki M, Doebley JF et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208. doi:10.1038/ng1702

Zhang K, Deng M, Chen T, Waterman MS, Sun F (2002) A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci USA 99:7335–7339. doi:10.1073/pnas.102186799

Zhang ZS, Xiao YH, Luo M, Li XB, Luo XY, Hou L et al (2005a) Construction of a genetic linkage map and QTL analysis of fiber-related traits in upland cotton (*Gossypium hirsutum* L.). Euphytica 144:91–99. doi:10.1007/s10681-005-4629-x

Zhang J, Lu Y, Cantrell RG, Hughs E (2005b) Molecular marker diversity and field performance in commercial cotton cultivars evaluated in the Southwestern USA. Crop Sci 45:1483–1490. doi:10.2135/cropsci2004.0581

Zhang HB, Li Y, Wang B, Chee PW (2008) Recent advances in cotton genomics. Int J Plant Genomics 2008:742304

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C et al (2007) An *Arabidopsis* example of association mapping in structured samples. PLoS Genet 31:e4. doi:10.1371/journal.pgen.0030004

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR et al (2003) Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134