Theor Appl Genet (2012) 124:1201–1214
DOI 10.1007/s00122-011-1780-8

ORIGINAL PAPER

Development and mapping of SNP assays in allotetraploid cotton

Robert L. Byers · David B. Harker · Scott M. Yourstone · Peter J. Maughan · Joshua A. Udall

Received: 3 August 2011 / Accepted: 22 December 2011 / Published online: 18 January 2012

Abstract A narrow germplasm base and a complex allotetraploid genome have made the discovery of single nucleotide polymorphism (SNP) markers difficult in cotton (*Gossypium hirsutum*). To generate sequence for SNP discovery, we conducted a genome reduction experiment (*Eco*RI, *Baf*I double digest, followed by adapter ligation, biotin–streptavidin purification, and agarose gel separation) on two accessions of *G. hirsutum* and two accessions of *G. barbadense*. From the genome reduction experiment, a total of 2.04 million genomic sequence reads were assembled into contigs with an N_{50} of 508 bp and analyzed for SNPs. A previously generated assembly of expressed sequence tags (ESTs) provided an additional source for SNP discovery. Using highly conservative parameters (minimum coverage of 8× at each SNP and 20% minor allele frequency), a total of 11,834 and 1,679 non-genic SNPs were identified between accessions of *G. hirsutum* and *G. barbadense* in genome reduction assemblies, respectively. An additional 4,327 genic SNPs were also identified between accessions of *G. hirsutum* in the EST assembly. KBioscience KASPar assays were designed for a portion of the intra-specific *G. hirsutum* SNPs. From 704 non-genic and 348 genic markers developed, a total of 367 (267 non-genic, 100 genic) mapped in a segregating F_2 population (Acala Maxxa × TX2094) using the Fluidigm EP1 system. A *G. hirsutum* genetic linkage map of 1,688 cM was constructed based entirely on these new SNP markers. Of the genic-based SNPs, we were able to identify within which genome ('A' or 'D') each SNP resided using diploid species sequence data. Genetic maps generated by these newly identified markers are being used to locate quantitative, economically important regions within the cotton genome.

Communicated by A. Schulman.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1780-8) contains supplementary material, which is available to authorized users.

R. L. Byers · D. B. Harker · S. M. Yourstone · P. J. Maughan · J. A. Udall (✉)
Department of Plant and Wildlife Sciences, Brigham Young University, Provo, UT 84602, USA
e-mail: jaudall@byu.edu

Introduction

High throughput DNA sequencing technology facilitates the rapid discovery of large numbers of single nucleotide polymorphism (SNP) markers at relatively low cost compared to other traditional approaches. Recently, a few different strategies employing high throughput sequencing have reported identifying large numbers of SNP markers (Barbazuk et al. 2007; Van Tassell et al. 2008), including some in organisms with little previous molecular research (Maughan et al. 2010; Bundock et al. 2009; Baird et al. 2008) as well as organisms with little genetic variation such as cotton (Udall et al. 2006). These strategies utilize transcriptome sequencing, gene-enriched sequencing using methylation sensitive digestion, or sequencing of reduced representational libraries (RRL), also termed genome reduction.

Some of these strategies target genic regions while others target both genic and non-genic sequences. Genic strategies primarily target transcribed sequences through the development of Expressed Sequence Tags (ESTs). Expressed sequences may contain limited amounts of SNPs due to purifying selection of genic regions. In addition, the location of SNPs discovered in ESTs is limited to the

transcribed regions of the genome. Genic regions can be indirectly targeted and transcription biases can be avoided using methylation sensitive digestion of genomic DNA combined with sequencing. Gene-enriched sequencing using methylation sensitive digestion and sequencing of RRLs are similar techniques but vary in digestion specificity resulting in subtly different distributions of sequence types. RRLs of genomic DNA provide a method to isolate small, equivalent portions of the genome from two or more individuals regardless of gene expression or methylation state (Van Tassell et al. 2008; Wiedmann et al. 2008). Maughan et al. (2009) developed a genome reduction methodology that is based on restriction site conservation (GR-RSC) in which double-digested DNA is selectively amplified and size-selected on an agarose gel. GR-RSC libraries contain only a fraction of the entire genome and allow for identification of mostly non-genic SNPs while providing low sequence bias and fairly uniform genomic distribution (Maughan et al. 2009). This GR-RSC strategy could be used to identify non-genic molecular markers in the tetraploid cotton genome that are (1) well distributed throughout the genome, (2) variable between domesticated and undomesticated germplasm and (3) likely neutral with respect to natural (or artificial) selection.

Cotton is a major world agricultural crop, estimated at ~115 million bales (USDA 2011). In the United States, cotton fiber and seed by-product revenue accounts for an estimated five billion dollars annually (Wallace et al. 2009). *Gossypium hirsutum* (Upland cotton) and *G. barbadense* (Pima cotton) represent 96.7 and 3.3% of the total cotton fiber produced in the United States (USDA 2011). Both *G. hirsutum* and *G. barbadense* are allotetraploid ($2n = 4x = 52$) species of cotton and are composed of a A_T (~1,700 Mb) and a D_T (~900 Mb) genome. Polyploidy combined with a narrow germplasm base have hindered the development of SNP-based marker assays in cotton. SNP-based molecular markers offer the possibility of constructing dense genetic maps as well as facilitating map-based gene cloning efforts and haplotype-based association studies. In cotton, the most extensive work to date on SNP development reported the characterization and mapping of 270 SNPs based on EST sequencing (Van Deynze et al. 2009).

Here we report the discovery and application of thousands of SNPs in the allotetraploid genome of cotton. Our efforts of SNP discovery and application were circumscribed by four main objectives: (1) utilization of the GR-RSC methodology to identify the first large-scale set of SNP markers in cotton, (2) conversion of several hundred putative SNPs into functional SNP genotyping assays using KASPar genotyping chemistry, (3) evaluate the utility of these functional SNPs across a broad panel of domesticate and wild cotton accessions, and (4) develop the first genetic linkage map of *G. hirsutum* based solely of SNP markers.

Materials and methods

Plant materials

Four accessions were used for marker discovery in cotton. These accessions represent domestic and wild accessions of two species of allotetraploid cotton: *G. hirsutum* (Acala Maxxa and TX2094) and *G. barbadense* (Pima-S6 and K101). These accessions were selected for their agricultural significance (Brubaker and Wendel 1994) and historical relevance with regard to previous studies (Hovav et al. 2008; Rapp et al. 2009). The allotetraploid genome of cotton contains two genomes, A_T and D_T where the 'T' subscript indicates the genome in a tetraploid nucleus (Wendel and Cronn 2003). SNPs identified in the EST dataset were based on sequence data from the same *G. hirsutum* accessions (Acala Maxxa and TX2094) and additional diploid sequences from *G. arboreum* (A_2 genome) and *G. raimondii* (D_5 genome).

Additional plant materials include an F_2 population and a diversity panel of *G. hirsutum*. An F_2 population of 174 individuals was derived from a cross of the *G. hirsutum* parents Acala Maxxa × TX2094. A diversity panel of 48 accessions was created to represent the extant genetic diversity within *G. hirsutum* (Wendel et al. 1992). This panel includes representative domesticated accessions from the Mississippi delta, High Plains, and Eastern and Western United States. A broad representation of landraces and wild accessions was included to evaluate introgression potential of SNP markers with exotic germplasm.

DNA extraction

Separate DNA extractions were performed for all samples using freeze-dried leaf tissue. Extractions for GR-RSC sequencing and F_2 genotyping were performed using a cetyltrimethylammonium bromide (CTAB) extraction procedure scaled for 1.7 mL extractions (Kidwell and Osborn 1992). DNA of the diversity panel was extracted using the Qiagen DNeasy kit (Qiagen, Valencia, CA). Extracted DNA was suspended in DNase-free water and quantified using a NanoDrop spectrophotometer (ND 1000, NanoDrop Technologies Inc., Montchanin, DE).

DNA sequencing

Genome complexity was reduced using the GR-RSC method as described by Maughan et al. (2009). Briefly, total genomic DNA was double digested to completion using *EcoR*I and *Bfa*I endo-restriction nucleases. *Bfa*I and *EcoR*I site-specific adapters were ligated to the digested fragment's sticky ends. *EcoR*I adapters included a biotin end which allowed *EcoR*I cut fragments (the less frequent

cut site) to be selected for using a biotin–streptavidin magnetic bead separation, reducing genomic complexity by about 90%. Resulting fragments were then PCR amplified with adaptor specific primers that also contained multiplex identifier (MID) barcodes to allow for sample multiplexing on the Roche 454 pyrosequencing platform. Genomes were further reduced by selecting genomic fragments from the PCR reaction in the range of 450–600 bp via agarose gel separation. Samples were sequenced using Titanium reagents on the Roche 454 Genome Sequencer FLX at the BYU DNA Sequencing Center. Separate genome reductions were performed for each accession in the GR-RSC experiment (Acala Maxxa, TX2094, Pima-S6, K101).

Genomic fragment assembly

GR-RSC sequence reads were grouped into separate files based on their MID barcodes. Newbler de novo assembler v2.3 was used to create all of the GR-RSC sequence assemblies. Stringent assembly parameters of 97% sequence identity and 100 bp minimum overlap were used to minimize co-assembly of A_T and D_T homoeologous sequences. Combined GR-RSC assemblies were created to identify SNPs within *G. hirsutum* (between Acala Maxxa and TX2094), within *G. barbadense* (between Pima-S6 and K101), and between the two species (*G. hirsutum* and *G. barbadense*). Since less sequencing was performed on the *G. barbadense* accessions, a subset of *G. hirsutum* reads, referred to as "reduced *G. hirsutum*" hereafter, was created and used to form a fourth combined assembly. This "reduced *G. hirsutum*" assembly consisted of a random subset of reads, comparable both in number of reads and total bases to the *G. barbadense* assembly. The reduced assembly eliminated assembly size bias and allowed for direct comparison of results between the *G. hirsutum* and *G. barbadense* assemblies. Separate assemblies were also created for each of the four accessions.

To remove repetitive sequences from our analysis, the combined assemblies were run through RepeatMasker (Smit et al. 1996–2010). Categorized repeats for *Gossypium* (Grover, personal communication) were included along with the *Arabidopsis* repeats in our RepeatMasker database. All contigs which contained repetitive fragments were excluded from the assemblies used for SNP discovery. Contigs were also screened by reference mapping consensus sequences to the chloroplast genomes of the *G. hirsutum* and *G. barbadense* and the mitochondrial genome of *Arabidopsis* using 454 gsMapper (v2.3) prior to SNP analysis.

Identification of SNPs and microsatellites

Accurate identification of SNPs (i.e. polymorphisms that occurred in only one of the genomes) in tetraploid sequence data was a challenge due to the potential co-assembly of homoeologs. If homoeologs were co-assembled, single nucleotide differences between the A_T and D_T genomes could have been confounded with SNP of a single locus. Unless specifically indicated by italics, the term SNP in this study refers only to allelic single nucleotide differences and not differences that distinguish the two resident genomes of the allotetraploid (*alias* genome-specific SNPs). Whether homoeologs co-assembled or assembled separately depends on homoeolog sequence divergence and the strictness of the assembly parameters (Fig. 1). Separate assembly of homoeologs provided the simplest means for identification of genome-specific SNPs. In both the GR-RSC and EST assemblies, identification of these genome-specific SNPs in separately assembled homoeologs was performed using SNP_Finder (Maughan et al. 2009). Putative SNPs were identified if the following conditions were met: (1) coverage depth at the SNP was $\geq$8; (2) the minor allele represented at least 20% of the alleles observed; and (3) 90% of the alleles from a specific MID barcode were identical.

Unique to the EST dataset, SNPs were also identified in contigs containing co-assembled homoeologs and their A_2 and D_5 diploid homologs. Coverage of at least 12× was required at the SNP with a minimum of 3× coverage from each of the diploid and tetraploid accessions. Both diploids and one of the tetraploid accessions had to be identical (i.e. homozygous) for the major allele while the other tetraploid accession segregated for the major and minor alleles (i.e. heterozygous). Finally, a minimum of 2× coverage was required for the minor allele (Fig. 2). Flanking sequences and SNPs have been deposited in NCBI dbSNP under the handle UDALL_LAB (Batch A, #1051857; Batch B, #1051858; Batch C, #1051850).

Potential SSRs were identified in assemblies of the *G. hirsutum* and *G. barbadense* using MISA v.1.0 (http://pgrc.ipk-gatersleben.de/misa) with a unit size/minimum number of repeats threshold of 2/6, 3/5, 4/5, 5/5, 6/5 and a maximal number of bases interrupting 2 SSRs in a compound microsatellite of 100. Mono-repeats were not reported because 454 homopolymer sequencing errors would be confounded with SSR loci.

SNP assay design

The KASPar (KBioscience Ltd., Hoddesdon, UK) assay was used to convert a portion of identified SNPs and estimate a conversion rate of putative SNPs to functional assays. KASPar assays were developed to target 1,052 genome-specific SNPs identified between accessions of *G. hirsutum* (Acala Maxxa and TX2094; Supplemental Table 1). All assay primer sets were designed using PrimerPicker (KBioscience 2009) with default parameters.

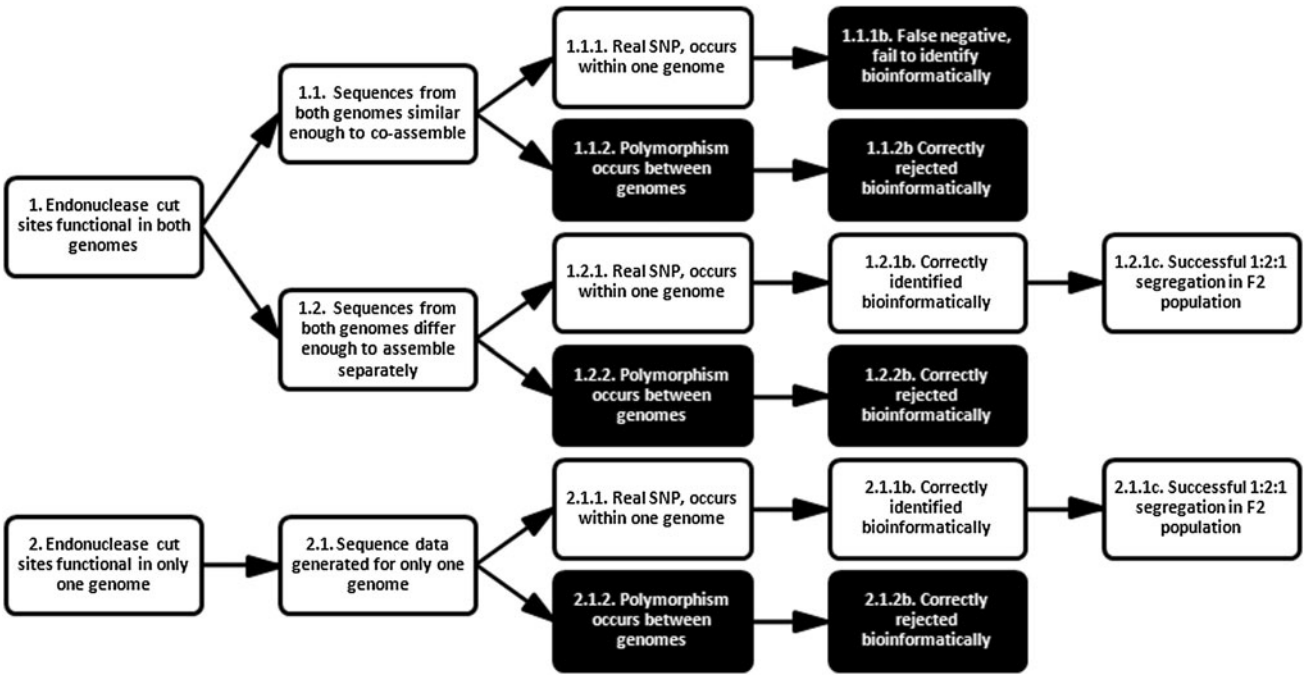


Fig. 1 SNP discovery flowchart for GR-RSC in allotetraploid cotton. A number of different SNP identification situations can occur depending on whether the endonuclease cut sites are present in one (flow 2) or both (flow 1) genomes, whether a homoeologous sequences co-assemble (flow 1.1) or assemble separately (flows 1.2 and 2.1), and whether the SNP occurred within one genome (flows 1.1.1, 1.2.1, and 2.1.1) or between the A_T and D_T homoeologs (flows 1.1.2, 1.2.2, and 2.1.2). The conservative strategy fails to identify some real SNPs, but in all cases rejects false SNPs created by assembly of homoeologous sequences from different genomes (*both highlighted in black*). SNPs identified in the GR-RSC assemblies fall into two categories: (1) SNPs derived in locations where endonuclease cut sites are conserved in both genomes and A_T and D_T sequences differ enough to cause separate assembly of homoeologs (flow 1.2.1) and (2) SNPs derived in sequences where endonuclease cut sites are only conserved in the genome in which the SNP exists (flow 2.1.1)

Of the 1,052 assays, 704 were designed to target SNPs from the GR-RSC *G. hirsutum* assembly while the remaining 348 were designed to target *G. hirsutum* SNPs located in EST sequences.

Because diploid sequence data from related species existed, two different strategies were employed in the development of the 348 EST SNP assays. In the first strategy, 192 of the assays were intended to amplify a single locus in a single genome with coincidental amplification of the non-target genome as background 'noise'. In many of these SNP assays the resident genome was identified using diploid sequence information, hereafter referred to as genome-identified (GI) SNP assays. In contigs where homoeologs co-assembled, diploid sequence data were used as a reference to categorize tetraploid reads by genome (A_T or D_T) as indicated by genome distinguishing SNPs (polymorphisms which differed between genomes, but were identical between accessions) occurring in the same tetraploid read. Based on this categorization, the base identity of the minor allele identified the genome of the SNP assay (e.g. the major allele was found in both A_T reads and D_T reads but the minor allele was only found in D_T reads, thus the resident genome of the SNP was D_T). In contigs where homoeologs separately assembled, co-assembly of diploid reads identified the resident genome of the SNP (e.g. only A_2 reads resided in the contig, thus the resident genome of the contig and SNP was A_T). While only 192 assays were designed using this strategy, the putative genome for many thousands of SNPs was identified in the EST assembly (Flagel et al. 2011).

In the second strategy, SNP assay design was 'improved' for the remaining 156 assays by adding base-pair mismatches between the assay's primers and the sequence of the homoeologous non-target loci so that the assays specifically targeted the resident genome of the SNP. Design of these 156 genome-targeted (GT) SNP assays (A_T or D_T targeted) was only possible for a limited number of loci because (1) assay design required that contigs contained homoeologous, co-assembled loci and (2) SNP assays needed to have closely positioned genome distinguishing SNPs such that the KSAPar assay primers would overlap these genome distinguishing bases (i.e. distinguishing polymorphisms between genomes, but identical between accessions, Fig. 3). Thus, primer sequences for these GT SNP assays were designed to specifically match only one of the two cotton genomic sequences. The intent of the genome-targeted primer design was to create genome-specific amplification (or sufficient amplification bias) such that only one genome

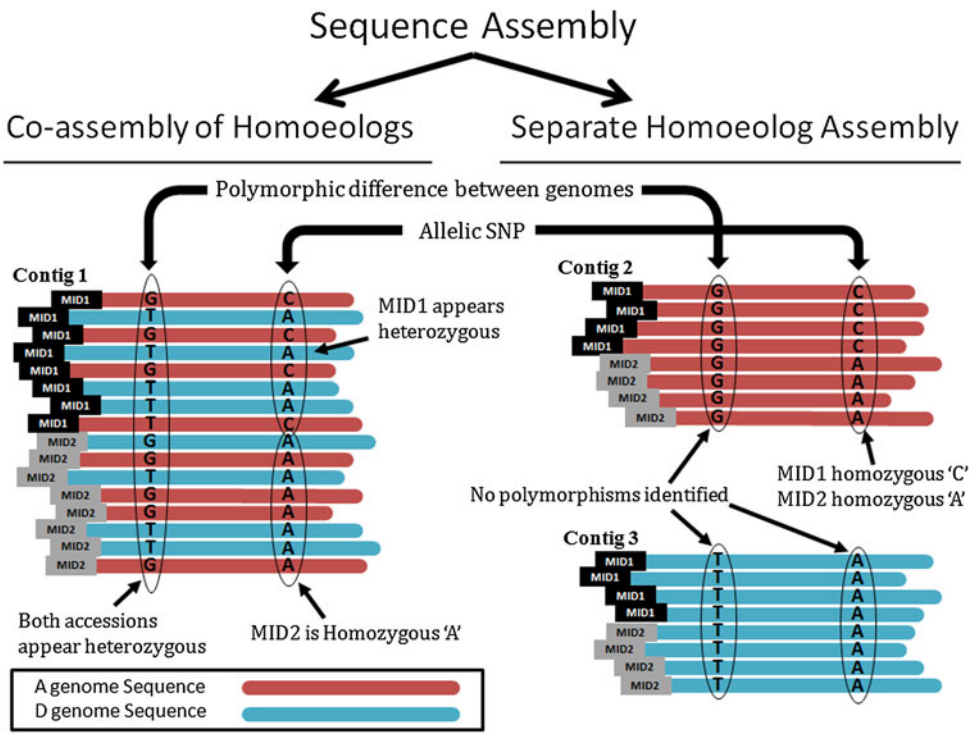


Fig. 2 Allotetraploid SNP identification. Co-assembly and separate assembly of homoeologs each require a unique strategy for identifying SNPs. In each case, a unique pattern distinguishes allelic SNPs from other types of polymorphisms. In assemblies of separate homoeologs, each of the individuals appears homozygous and the SNP segregates between them (Contig 2). In co-assembly of homoeologs, one individual appears homozygous while the other appears heterozygous (Contig 1). The observed pattern for separately assembled homoeologs that have one homozygous individual and one heterozygous individual is identical to the observed pattern of co-assembled homoeologs that have homozygous segregating individuals. As a result, SNPs cannot be identified when homoeologs co-assemble unless enough genome-specific SNPs are present in the sequences to separate reads by genome

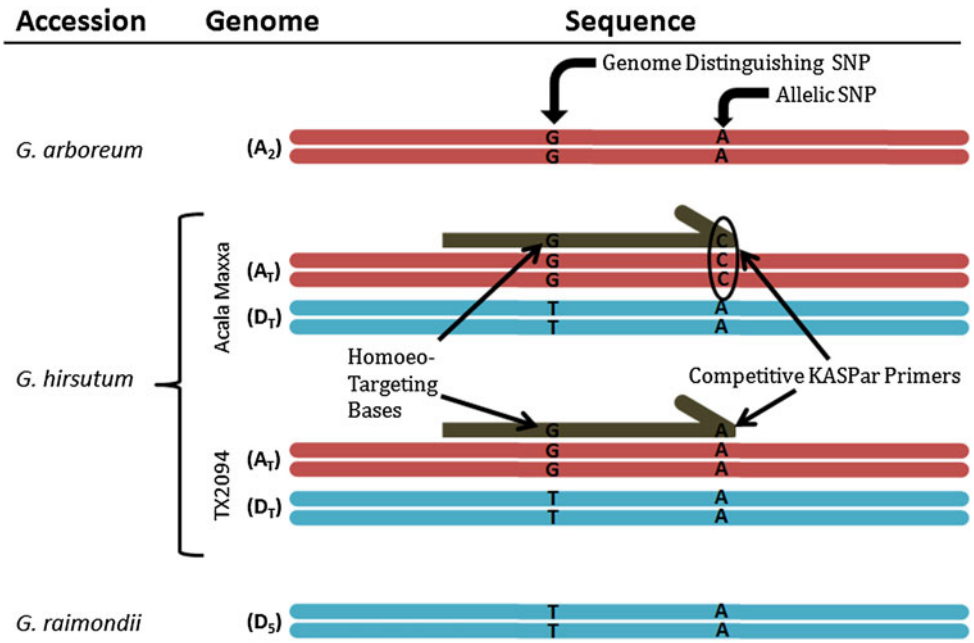


Fig. 3 Marker design to directly target a single genome. In the EST SNP assays designed to amplify only one genome, allelic SNPs were targeted if they had nearby genome distinguishing SNP(s). The intent was to develop a genome-specific PCR assay that would only target the genome in which the SNP resided. It was hoped this would reduce interference from amplification of the non-resident genome and improve the conversion rate from putative SNPs to functional markers

(e.g. A_T) would amplify in the KASPar reaction while the non-targeted homoeologous genome sequence (e.g. D_T) would not amplify (or amplify less) due to base mismatches at the primer binding site. A series of custom Perl scripts was used to analyze the assembly, create A_T and D_T consensus sequences, identify SNPs that were different between the

two genomes, design the primers with PrimerPicker, and replace one or more bases in the primers generated by PrimerPicker with bases that were genome specific to create primers that had strong binding preferences.

Subsequent genetic mapping of SNP assays from both design methods determined how accurately a single locus in a single genome could be targeted. Agreement of multiple predictive markers in linkage (e.g. five linked, targeted assays, all of which predict the A_T genome) was used as an indication of success.

Genotyping and genetic mapping

Assay screening and genotyping were performed on two different platforms. Initially, a small set of genomic SNPs (20) was validated using traditional KASPar with a 384-well plate reader. Subsequent, large-scale screening and genotyping of SNPs were then performed on Fluidigm 96.96 Dynamic Arrays using the genotyping EP1 System (San Francisco, CA). Fluorescence intensity was measured with the PHERAstar plus (BMG LABTECH, Durham, NC) microplate reader or the EP1 (Fluidigm Corp, San Francisco, CA) reader and plotted in two axes. Genotypic calls from PHERAstar measurements were made in KlusterCaller (KBioscience Ltd., Hoddesdon, UK) while genotypes based on EP1 measurements were made using the Fluidigm SNP Genotyping Analysis (Fluidigm 2011) program.

All functional SNP assays were used to genotype the F_2 population and 277 co-dominant assays between Acala Maxxa and TX2094 were used to genotype the 48 accessions of the *G. hirsutum* diversity panel. All genotype calls were manually checked for accuracy and ambiguous data points that failed to cluster were scored as missing data. A genetic map was constructed using regression mapping in JoinMap4 (Van Ooijen 2006). Markers which had greater than 30% of their genotypic data missing were excluded during the mapping process. A minimum LOD threshold of 5.0 was used and linkage distances were corrected using the Kosambi mapping function.

Results

Sequencing and assembly of GR-RSC reads

A total of 577 Gb of Roche 454 sequencing data were generated from the GR-RSC libraries of the *G. hirsutum* and *G. barbadense* accessions. Nearly a half (44%) Pico Titer Plate (PTP) of 454 titanium sequence data was allocated for each of the *G. hirsutum* samples (220 Mb of anticipated sequence) while only a quarter of a PTP was allocated for each of the *G. barbadense* samples (125 Mb of anticipated sequence). Actual sequencing results were slightly lower than expected with *G. hirsutum* samples yielding ~178 Mb each and *G. barbadense* samples yielding ~95 Mb each. *G. hirsutum* and *G. barbadense* samples represented 65.2 and 34.8% of the total sequencing, respectively. Theoretical calculations predicted that the GR-RSC libraries would represent 0.73% (17.5 Mb) of the tetraploid cotton genome (2.4 Gb). Actual assembly sizes ranged from a low of 6.39 Mb for the Pima-S6 assembly to a high of 40.3 Mb for the inter-specific combined assembly (*G. hirsutum* vs. *G. barbadense*). Sequence coverage of 12.6 and 7.1× was anticipated for each of the *G. hirsutum* and *G. barbadense* samples, respectively, based on the expected size of the 'reduced' cotton genome and planned amounts of sequencing. Actual sequence coverage in the assemblies ranged from 6.1× in the K101 accession assembly to 8.6× in the inter-specific combined assembly (Table 1).

The sequence data were assembled to form multiple GR-RSC assemblies (Table 1). The *G. hirsutum* assembly (Acala Maxxa and TX2094) resulted in 79,953 contigs with an N_{50} contig length of 516 bp while the *G. barbadense* assembly (Pima-S6 and K101) resulted in 51,307 contigs with and N_{50} contig length of 491 bp. Comparing the *G. barbadense* assembly with the results of the reduced *G. hirsutum* assembly, the reduced *G. hirsutum* assembly formed slightly more contigs (55,160) with an N_{50} contig length of 513 bp. The combined, inter-specific assembly

Table 1 Summary of GR-RSC sequence assemblies

Species	Accession(s)	Reads (k)	Bases (Mb)	Assembled bases (Mb)	Assembly length (Mb)	Average coverage	Total contigs
G. hirsutum	Acala Maxxa	617	183	94 (51.5%)	14.5	6.51	40,035
G. hirsutum	TX2094	588	173	85 (48.9%)	13.4	6.33	37,793
G. barbadense	Pima-S6	358	92	40 (43.5%)	6.39	6.26	19,513
G. barbadense	K101	373	98	41 (42.4%)	6.79	6.11	20,963
G. hirsutum assembly	Acala Maxxa, TX2094	1,310	387	218 (56.2%)	28.4	7.67	79,953
Reduced *G. hirsutum* assembly	Acala Maxxa, TX2094	839	248	127 (51.1%)	13.6	9.31	55,160
G. barbadense assembly	Pima-S6, K101	836	221	116 (52.6%)	16.6	7.01	51,307
Inter-specific assembly	Maxxa, TX2094, S6, K101	2,042	577	346 (59.9%)	40.3	8.58	112,506

(*G. hirsutum* vs. *G. barbadense*) resulted in 112,506 contigs from 1.25 million reads with an N_{50} contig length of 508 bp. The percent of bases that assembled ranged from 51.1% in the reduced *G. hirsutum* assembly to 59.9% in the inter-specific assembly. Assemblies with a greater number of input reads had greater percentages of bases incorporated into their alignments. Read depth between the two accessions within an assembly was compared and most contigs in the combined assembly were found to contain reads from both accessions (e.g. *G. hirsutum* assembly, Supplemental Fig. 1), suggesting that the genome reduction was successful in isolating homologous regions from the sampled accessions.

GR-RCS SNP discovery

The combined GR-RSC assemblies allowed for the identification of SNPs between accessions (Table 2). Within the intra-specific comparisons of *G. hirsutum* (Acala Maxxa and TX2094) and *G. barbadense* (Pima-S6 and K101), 11,834 and 1,679 SNPs were identified in 6,467 and 965 contigs, respectively. Contigs containing SNPs averaged between 1.74 and 1.83 SNPs per contig (Supplemental Fig. 2). Because of SNP detection method, most of these SNPs were likely detected in contigs where homoeologs had separately assembled. Comparing the reduced *G. hirsutum* and *G. barbadense* assemblies (equal size datasets), *G. hirsutum* still contained 2.4 times as many SNPs as the *G. barbadense* assembly (4,045 vs. 1,679). This difference is likely a reflection of the larger genetic distance between Acala Maxxa and TX2094 than between K101 and Pima-S6 (Percy and Wendel 1990; Wendel et al. 1992). In the inter-specific assembly, 29,066 SNPs were identified between *G. hirsutum* and *G. barbadense* within 14,905 contigs with an average of 1.95 SNPs per contig. Larger assemblies had a larger portion of contigs containing SNPs (Fig. 4). The average coverage of SNPs in assemblies ranged from 9.3× (*G. barbadense*) to 11.0× (inter-specific), with the most common coverage always at 8× (the minimum threshold for identification: Fig. 5). Since the GR-RSC process selected equivalent portions of the cotton genome, any two or more accessions could have also been jointly assembled to identify different, putative SNPs (e.g. Acala Maxxa and K101), but they are not reported here.

The SNP frequency, calculated as the number of SNPs in assembly divided by length of assembly, ranged from 0.0001 in the intra-specific assembly of *G. barbadense* to 0.00067 in the inter-specific assembly of *G. hirsutum* and *G. barbadense* (Table 2). These observed frequencies were not unexpected and reflect the narrower genetic base of the intra-specific *G. barbadense* comparison and higher genetic diversity of the inter-specific comparison. We note that the frequencies reported here are most likely underestimates due to conservative nature of SNP identification parameters.

Transition mutations (A ⇔ G, or T ⇔ C) are defined as a change from a purine to a purine or a pyrimidine to a pyrimidine, while transversion mutations (e.g. A ⇔ T, A ⇔ C, G ⇔ T, G ⇔ C) are defined as a change from a purine to a pyrimidine or a pyrimidine to a purine. Nucleotide transitions naturally account for the majority of observed SNPs and are thought to be driven by hypermutability effects of CpG di-nucleotide sites or deamination of methyl cytosine and entropy constraints (Li 1997). In all four combined GR-RSC assemblies, transitions were the most common SNP type, with transition-to-transversion ratios of 2.3:1. These ratios are similar to those recently found in human, maize, and amaranth (Maughan et al. 2009; Morton et al. 2006; Zhang and Zhao 2004).

EST SNP discovery

A de novo assembly of ESTs that included Acala Maxxa, TX2094, *G. arboreum* (A_2 genome) and *G. raimondii* (D_5

Table 2 Summary of GR-RSC SNP discovery

Category	Assembly	Accessions	SNPs	Contigs with SNPs	SNPs per contig	SNP frequency
By individual	*G. hirsutum*	Acala Maxxa	45,590	8,660	5.26	0.003149
	G. hirsutum	TX2094	42,166	8,201	5.14	0.003156
	G. barbadense	Pima-S6	26,662	4,934	5.4	0.004174
	G. barbadense	K101	29,420	5,455	5.39	0.004335
Between accessions	*G. hirsutum*	Acala Maxxa and TX2094	11,834	6,469	1.83	0.00039
	Reduced *G. hirsutum*	Acala Maxxa and TX2094	4,045	2,176	1.86	0.0002
	G. barbadense	Pima-S6 and K101	1,679	965	1.74	0.0001
	Inter-specific assembly	(Maxxa, TX2094) and (S6, K101)	29,066	14,905	1.95	0.00067
All SNPs	Inter-specific assembly	Maxxa, TX2094, S6, K101	151,712	39,396	3.85	0.003517

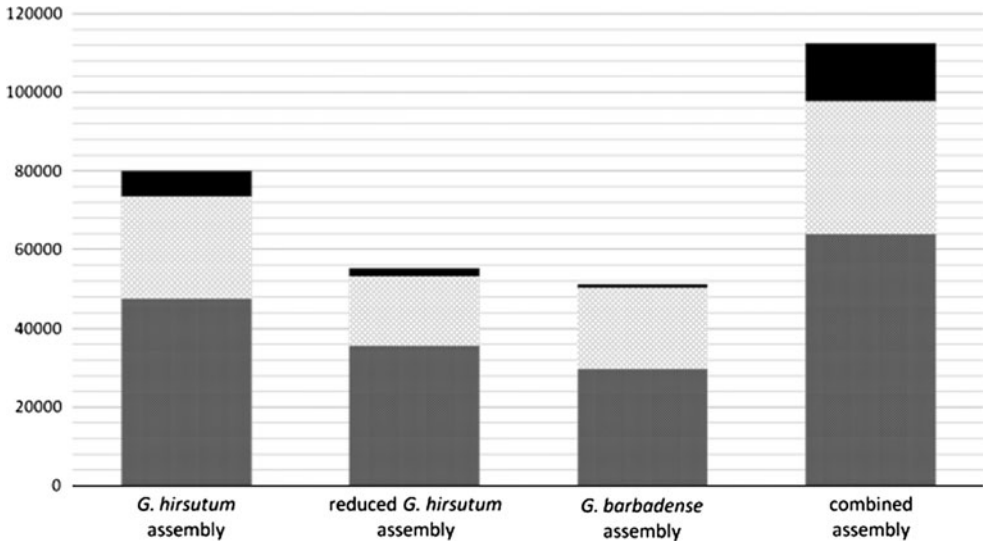
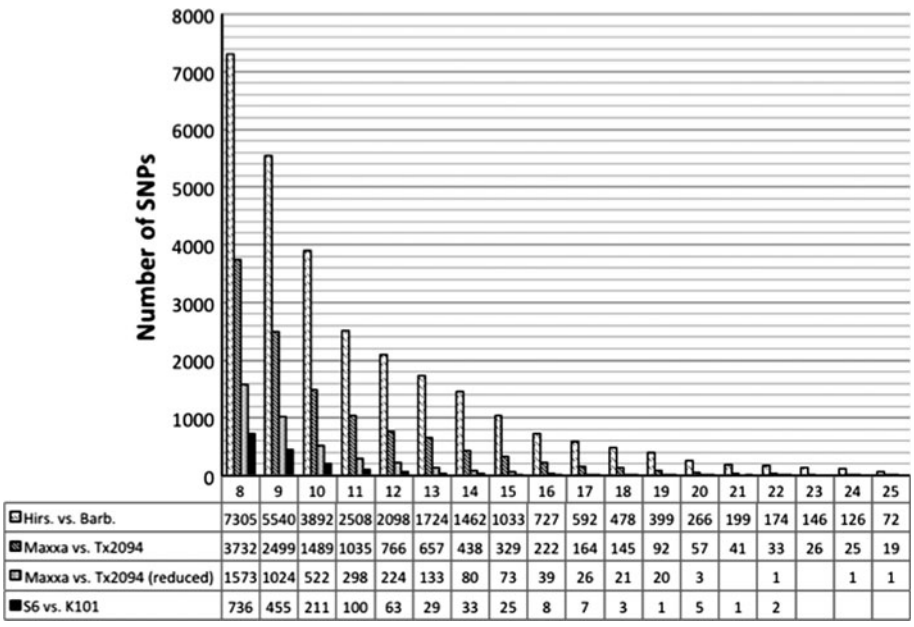


Fig. 4 Distribution of contigs in the GR-RSC assemblies. Each *column* represents one of the 4 combined GR-RSC assemblies. The *bottom* of each *column* represents the portion of contigs that did not meet minimum SNP requirements due to lack of sequence coverage from one or both accessions. The *middle* of each column represents the portion of contigs that met minimum SNP requirements, but contained no SNPs. The *top* of each column represents the portion of contigs that contained SNPs. In each of the four assemblies, the proportion of contigs with SNPs increases with assembly size



	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Hirs. vs. Barb.	7305	5540	3892	2508	2098	1724	1462	1033	727	592	478	399	266	199	174	146	126	72
Maxxa vs. Tx2094	3732	2499	1489	1035	766	657	438	329	222	164	145	92	57	41	33	26	25	19
Maxxa vs. Tx2094 (reduced)	1573	1024	522	298	224	133	80	73	39	26	21	20	3		1		1	1
S6 vs. K101	736	455	211	100	63	29	33	25	8	7	3	1	5	1	2			

Fig. 5 Distribution of SNPs by sequence coverage in the GR-RSC assemblies. *Columns* represent the number of SNPs in each assembly at a given sequence coverage. The chart displays SNPs in the range from 8× to 25× coverage. This range has been selected because 8× was used at the minimum coverage required and coverage above 25× becomes less informative. Across all levels of coverage the highest and lowest numbers of SNPs were found in the combined and *G. barbadense* assemblies, respectively. Across all assemblies, the number of SNPs was exponentially decays as coverage increases

genome) sequences provided a basis for SNP discovery in coding regions (Flagel et al. 2011). The joint assembly of diploid (A_2 and D_5) and tetraploid ESTs allowed for identification of genome-specific SNPs in contigs of both separate and co-assembled homoeologs. A total of 3,319 SNPs were identified between Acala Maxxa and TX2094 in contigs where homoeologs did not co-assemble. In contigs of co-assembled homoeologs, 1,009 SNPs were identified between Acala Maxxa and TX2094.

SNP assay development

A total of 1,052 SNPs were selected for SNP assay development from the SNPs identified in the GR-RSC and EST datasets. The assays were based on the KBiosciences KASPar genotyping chemistry and were tested for a Mendelian segregation ratio using an F_2 mapping population derived from a cross of Acala Maxxa × TX2094 (Additional File 1). Of the 704 SNPs derived from the

GR-RSC assembly, 252 (35.8%) amplified and segregated as expected for an F_2 population (1:2:1 or 3:1; $p > 0.05$) (Fig. 6). Of the 252 GR-RSC assays, 130 (51.6%) were co-dominant (1:2:1 segregation ratio), while the remaining 122 (48.4%) assays were dominant (3:1 segregation ratio). In dominant SNP assays, the heterozygote cluster was indistinguishable from one of the homozygous clusters

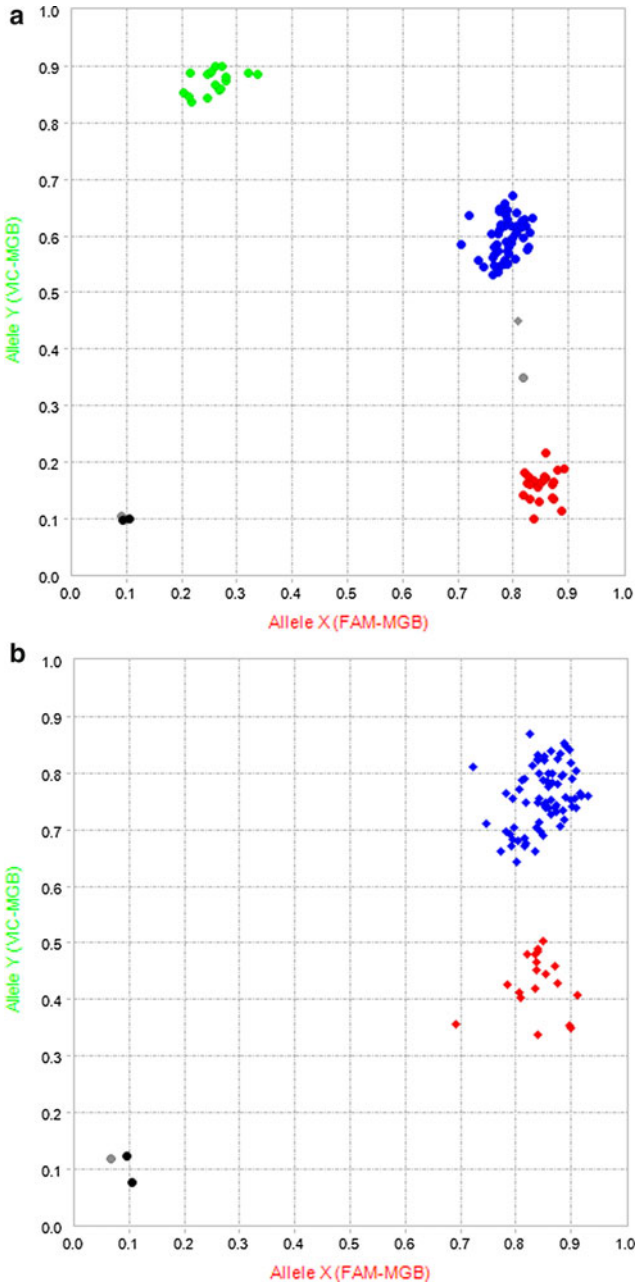


Fig. 6 F_2 genotyping plots from the Fluidigm SNP Genotyping Analysis software. Fluorescence values obtained using Kbioscience KASPar genotyping assays with the Fluidigm EP1 system. *Y*-axis represents VIC fluorescence intensity, *x*-axis represents FAM fluorescence intensity. Both intensity values normalized by ROX fluorescence. Displayed are 88 F_2 individuals and 8 controls genotyped by **a** a co-dominant marker and **b** a dominant marker

(Fig. 6). The distinction between co-dominant and dominant loci was determined by F_2 segregation patterns and parental genotypes (Acala Maxxa, TX2094 and their F_1 progeny). The parental samples consistently produced distinct genotypes in homozygous clusters among all SNP assays and the F_1 consistently produced genotypes in heterozygous or dominant clusters among all SNP assays.

The EST-based assay conversion rates were similar to the GR-RSC assay conversion rate. Of the two types of EST SNP assays, 156 GT SNP assays and 192 GI SNP assays, 50 (32.1%) and 59 (30.7%) met a χ^2 test for 1:2:1 or 3:1 segregation, respectively. Of the remaining 691 assays which did not segregate as expected for an F_2 population, the vast majority (86%) failed to amplify or separate into clusters while the remainder (14%) formed clusters, but the clusters did not conform to a 1:2:1 or 3:1 Mendelian pattern of inheritance, though they were used for genetic mapping (below). Some of these non-conforming assays may actually represent functional SNP assay that are simply linked to strongly skewed genomic regions (segregation distortion) in this F_2 population. Skewness of molecular markers has been attributed to chromosomal regions containing possible gametophytic or zygotic viability factors (Lu et al. 2002; Zamir and Tadmor 1986) and/or underlying genetic factors (i.e., quantitative trait loci) conferring a selective advantage for the particular growing conditions used to produce the mapping population.

SNP assay utility

To characterize the applied potential of these SNP assays in cotton breeding, the SNP assays were screened in a panel of 48 diverse *G. hirsutum* accessions (Supplemental Table 2). Several observations can be made from the observed genotypic patterns. First, of the 48 accessions genotyped no two individuals shared the same genotype across all assays (277 co-dominant SNP assays). Second, several accessions shared many wild alleles with TX2094 (the wild parent of the F_2 mapping population), with the most similar individual, TX2090, sharing 80.0% of its alleles with TX2090. Third, comparison of domesticated accessions to Acala Maxxa confirmed that domesticated accessions had nearly all alleles common with Acala Maxxa (of all domesticated accessions genotyped, no individual had more than 6.14% of its alleles different from Acala Maxxa and when considering all domesticated accessions together, only 17.7% of the 277 assays exhibited any TX2094 allele). An average heterozygosity of 2.43% was observed across all SNP assays with the highest heterozygosity of any assay being 15.2%. Of the 277 assays tested, 259 (93.5%) had a minor allele frequency of greater than 10% and 188 (67.9%) had a minor allele frequency of greater than 20%.

The results of the GT and GI SNP assays in the diversity panel of *G. hirsutum* were further inspected. 25 A-genome assays and 23 D-genome assays were included in the screening of 277 total assays. Across all accessions, A-genome assays identified 34.0% wild alleles and the D-genome assays identified 34.7% wild alleles. In the domesticated accessions, A-genome assays identified 7.1% wild alleles and D-genome assays identified 3.0% wild alleles. These results suggest that wild alleles are equally represented in both A- and D-genomes across the panel of other landraces and primitive cultivars. These assays also suggested a slight bias of wild alleles in the A-genome of cultivated cotton compared to the D-genome, though the limited number of assays detecting any wild alleles (9 A-genome and 7 D-genome assays total) in cultivated cotton prevented any broader assertions.

Genetic mapping of SNP assays

A genetic map was constructed based on an F_2 of *G. hirsutum* ($2n = 4x = 52$) population to further validate the SNPs discovered in this study and demonstrate genome targeting of EST SNP assays. The genetic map was created using 267 GR-RSC SNP assays and 100 EST SNP assays (367 total) for which genotypic data were available (Fig. 7). Of the 367 markers, 346 formed 38 linkage groups ($n = 26$ *G. hirsutum*) with an average of 9.1 markers per linkage group and a total genetic distance of 1,688 cM. The longest linkage group was 136.2 cM while the average length of linkage groups was 44.4 cM. Linkage groups contained between 2 and 18 SNP assays each and the average distance between SNPs was 5.48 cM. A total of 41 SNP assays were identified as skewed ($p < 0.05$) in the map while 7 were extremely skewed ($p < 0.001$). Skewed SNP assays were found in 17 of the of the map's linkage groups, while 5 linkage groups had extremely skewed assays (Fig. 7). Linkage groups 15 and 21 exhibited the largest numbers of skewed assays.

The resident genome of most EST SNP assays was identified a priori (GI) or was identified a priori and targeted (GT) during assay development. 100 of 348 EST SNP assays were placed in the genetic map. 81 of these assays had an a priori identification of their resident genome. Of these 81 assays, at least one was found in 32 (84%) of the 38 linkage groups, while at least two were found in 25 (66%) of the 38 linkage groups (Fig. 7). 74 of 81 assays (91%) resided in linkage groups with at least one other GI or GT assay. To determine whether the resident genome of these SNPs was accurately identified, linkage groups with multiple GI and GT SNP assays were examined for genome consensus. Seventy (94%) of the 74 assays that resided in linkage groups with at least one other GI/GT assay agreed with the consensus for the target genome. Of the 25 linkage groups with two or more GI/GT SNP assays, 21 (84%) perfectly agreed with their genome identification (Fig. 7). Of the four linkage groups with assays that disagreed, each case consisted of only two GI/GT SNP assays. Thus, of the 38 linkage groups in the map, 28 (74%) of these can be putatively assigned a genome based on these predictive SNP assays. These assignments suggest that 12 linkage groups (#1, 2, 5, 6, 7, 9, 18, 19, 21, 23, 28, and 30) are representative of the D_T genome while 16 (#3, 4, 10, 11, 12, 13, 14, 15, 16, 17, 22, 24, 25, 29, 36, and 38) are representative of the A_T genome.

Fig. 7 Genetic map of *G. hirsutum*. A 1,688 cM map constructed from an intra-specific *G. hirsutum* (Acala Maxxa × TX2094) F_2 population of 174 individuals. 346 markers based on newly discovered SNPs form 38 linkage groups. The average distance between markers is 5.48 cM. The average length of a linkage group is 44.4 cM with the longest linkage group being 136.2 cM. Distances shown in centiMorgans (cM) and corrected with Kosambi mapping function. *Red* and *blue* highlighted marker had their resident genome bioinformatically predicted prior to mapping and *colors* indicate a prediction of the 'D' or 'A' genome, respectively. *Marker is skewed ($p = 0.05$), **marker is skewed ($p = 0.01$), ***marker is skewed ($p = 0.001$)

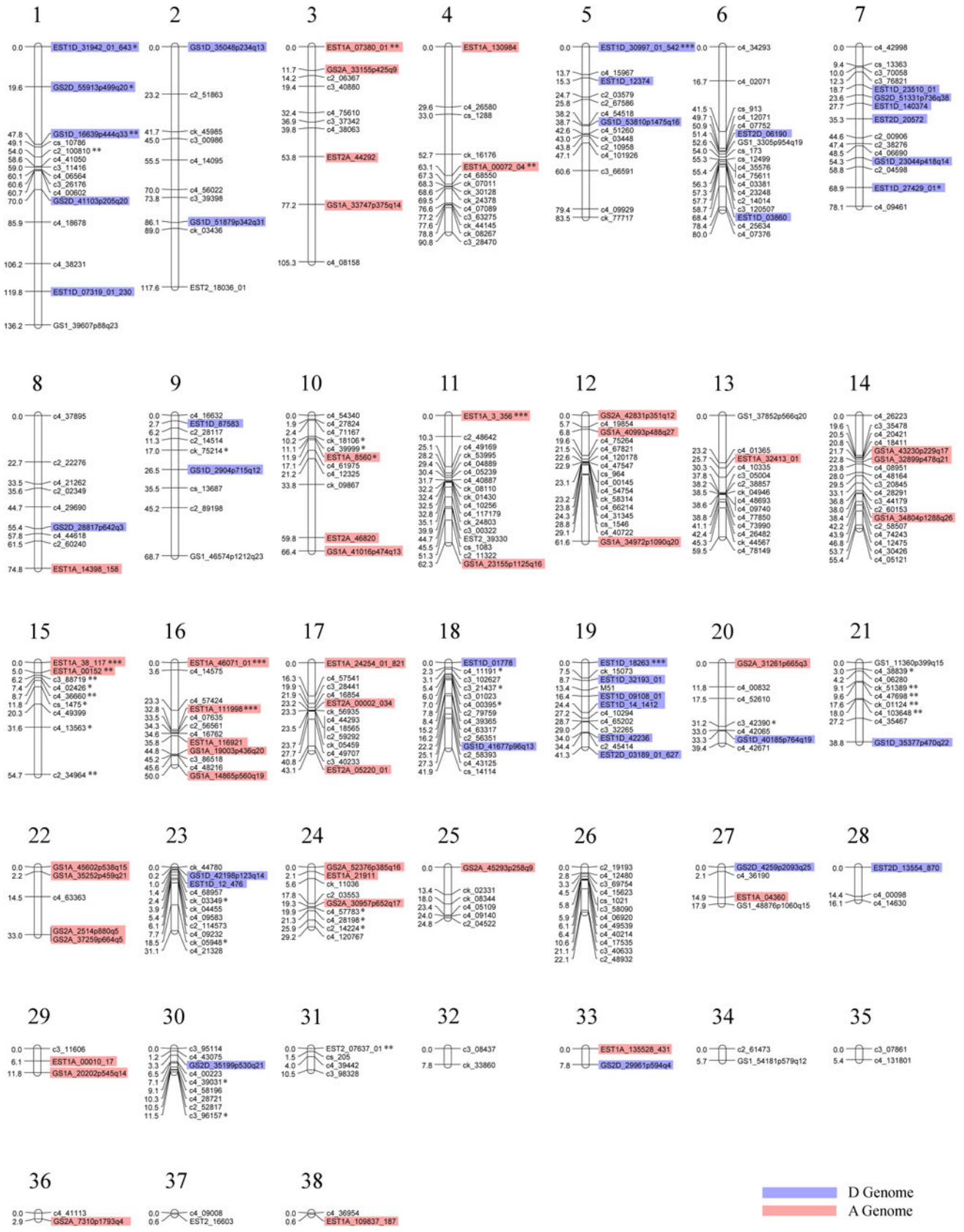
SSR discovery

In the GR-RSC sequence assemblies potential SSR markers were identified using the MISA v.1.0 Perl script (http://pgrc.ipk-gatersleben.de/misa) (Supplemental Fig. 3). The AT/TA class was the most abundant, similar to SSR abundance in other species (Varshney et al. 2002). Di-nucleotide repeats were the most common followed by tri-nucleotide repeats and the frequency of each repeat decreased as repeat length increased. As expected the number of detected repeats identified was also correlated with size of assembly. The assembly of both *G. hirsutum* and *G. barbadense* together contained the most SSRs, in part because it contained the highest number of reads. We report this discovery of additional SSR markers for cotton because SSRs continue to be broadly used in cotton research (Zhang et al. 2011; Gutiérrez et al. 2010; Lacape et al. 2009; Zhang et al. 2009; Lin et al. 2009; Rong et al. 2007).

Discussion

SNP discovery and mapping

A narrow germplasm base coupled with the complexity of a tetraploid genome presented a significant challenge in identifying and developing functional SNP assays in cotton. Despite the difficulties, we successfully identified genome-specific SNP markers (validated by Mendelian


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
D Genome
A Genome

segregation patterns in an F_2 population) from both the GR-RSC and EST approaches and have shown that genome specificity (A_T or D_T) of EST SNP assays could be determined a priori via the inclusion of A_2 and D_5 diploid sequences in the EST assemblies. The SNPs identified in this study have a transition/transversion ratio similar to other plant genomes and we have shown that 361 of these SNPs to exhibit normal Mendelian inheritance expected in a segregating F_2 population.

In addition to Mendelian segregation patterns, SNP assays based on these putative SNPs have been used to create an intra-specific map of *G. hirsutum* from a large segregating F_2 population. The map covers 1,688 cM (37.5%) of the approximate 4,500 cM (Rong et al. 2004; Reinisch et al. 1994) recombination length of allotetraploid cotton. While this is not the largest intra-specific map to date in terms of cM, it is comparable to previous intra-specific maps (Zhang et al. 2009; Lin et al. 2009; Ulloa et al. 2002; Shen et al. 2005) and is the first map to be constructed in cotton exclusively with SNP-based markers. We have not attempted to associate linkage groups with specific chromosomes in this map, but the anticipated release of the diploid cotton genome sequences (*G. arboreum* and *G. raimondii*) within the next year, should allow us to unambiguously assign SNP loci to particular chromosomes. Considering the conversion rate of putative GR-RSC SNPs to function KASPar SNP assays (35.8%), we estimate that of the 11,834 SNPs we have identified within *G. hirsutum* in this study, 4,237 are expected to yield functional SNP assays. With additional assay development, these markers could provide the means to establish the first high-density linkage map of *G. hirsutum* based solely on SNP loci. SNP assays are an ideal marker choice as they represent the highest resolution molecular marker possible and are highly amenable to genotyping automation.

Previous work suggests that GR-RSC markers are evenly distributed along chromosomes (Maughan et al. 2009). The even distribution of GR-RSC markers is of particular interest as it has recently become apparent that many agronomically important genes are controlled by regulatory sequences located in non-genic portions of the genome (Elshire et al. 2011). Thus, our development of SNP assays has targeted both genic and non-genic portions of the cotton genome. Specifically, GR-RSC SNP assays have been shown to also access pericentric and centromeric regions of the genome in *Arabidopsis* (Maughan et al. 2010). In maize, approximately 21% of genes lie in pericentric regions but most of the recombination occurs outside of these regions (Gore et al. 2009). If gene distribution within the cotton genome proves to be similar to maize, GR-RSC SNP assays may prove a valuable complement to previously identified molecular markers.

Homoeolog specific markers

We attempted to target alleles in only one of the two genomes resident in the tetraploid nucleus through two different methods of SNP assay design. The first and simplest method for identifying SNPs in a tetraploid is to force the separate assembly of homoeologs (i.e., only sequences from the A_T genome assemble together and only sequences from the D_T genome assemble together) through the utilization of strict assembly parameters. Loose assembly parameters (default) lead to co-assembly of homoeologous sequences that confound the identification of true SNPs. Neither strict nor loose assembly parameters produced ideal assemblies for all genome reduction fragments as the amount of sequence divergence in the selected fragments was locally constrained. The set of strict assembly parameters used 97% sequence identity and 100 bp minimum overlap to force sequences from each genome to assemble separately (i.e. genome-specific contigs). In addition to these parameters, our conservative SNP identification method (8× coverage, 90% identity and 20% minor allele frequency) only considered a subset of all SNPs in the dataset in which we had high confidence. Co-assembly of highly similar homoeologous sequences also likely occurred even in this strict assembly but this type of contig was ignored during SNP discovery in the GR-RSC assemblies. These contigs were ignored because accurate identification of SNPs without diploid reference sequences was impossible. Without the diploid reference sequences, we were unable to distinguish between a SNP in a co-assembly of homoeologs and a heterozygous locus in a separate assembly of homoeologous sequences (Fig. 2). Thus, only SNP loci were used that were homozygous in Acala Maxxa, homozygous in TX2094 and had a different nucleotide between the two accessions.

In contrast, the EST dataset provided sufficient A_2 and D_5 diploid sequence data to create genomic sequence references for each of the tetraploid genomes, thus allowing us to assign specific tetraploid reads to the A_T or D_T genome. Individual reads within a co-assembled tetraploid contig were assigned to either the A_T or D_T genome by genome distinguishing SNPs matching bases in either the A_2 or D_5 diploid sequences (Fig. 2). Both the observance of expected Mendelian segregation ratios and the successful prediction of resident genomes (A_T or D_T) for greater than 94% of the GT/GI SNP assays supports the conclusion that tetraploid reads were correctly assigned to genomes using A_2 and D_5 diploid reference sequences. In a few cases, designed GT/GI assays failed to indicate a consensus genome for their linkage group. Possible explanations for these disagreements include bioinformatic errors due to paralogous assemblies, differences between the diploid A_2 and D_5 genomes and the tetraploid A_T and

D_T genomes, or poorly mapped linkage groups containing markers from both genomes. As far as we know, this was the first report of large-scale design of genome-specific SNP markers in a polyploid plant.

Consideration for SNP assay development and utilization in cotton

While bioinformatic filters can identify thousands of putative SNPs, often only a subset can be successfully converted to functional marker assays due to the (1) simultaneous assay targeting of duplicate loci (paralogs or homoeologs), (2) local nucleotide limitations of primer design near the SNP, (3) proximity of the SNP to repetitive elements such as transposons, and (4) initial identification of false SNPs owing to sequencing errors and/or poor assembly. Our conversion rate of SNP assays was lower than initially anticipated. In amaranth, a diploid species, a conversion rate of nearly 70% was observed using a GR-RSC-based SNP discovery method (Maughan et al. 2009). The GR-RSC and EST SNP identification methods in this study had conversion rates of 35.8 and 31.3%, respectively. The difference between these two conversion rates is likely a difference in ploidy levels between the two species. In cotton, many of the 'failed' assays could be amplifying or partially amplifying segregating loci on both resident genomes resulting in uninterpretable cluster patterns. In designing the EST SNP assays, 156 of the assays were specifically chosen at SNP loci where the flanking sequences had diverged between the A_T and D_T genomes. These SNP assays were developed to test whether a design of genome specific primers could improve marker success rate. We observed similar conversion rates between the GR-RSC markers and both types of EST markers, suggesting that regardless of the source of the putative SNPs (EST or GR-RSC) or genome specificity of the KASP primers only subtle improvement in SNP assay conversion rates may be achieved in a polyploid genome.

We characterized these SNP assays in a diverse germplasm panel of *G. hirsutum* to ascertain their broader utility for trait introgression via marker assisted selection analysis of the germplasm panel on a selection of our SNP markers showed that Acala Maxxa and TX2094 were characteristics of domestic and wild varieties, respectively, and that few wild alleles exist in cultivated varieties of cotton. It also demonstrated that the narrow germplasm base of cotton could be broadened dramatically via the introgression of wild alleles into the cultivated cotton germplasm. We expect the putative SNPs identified within *G. barbadense* (nearly 1,700) to possess similar utility in expanding the germplasm base of *G. barbadense*.

Conclusions

We report the discovery of over 151,000 putative SNPs in non-transcribed sequences of allotetraploid cotton. These polymorphisms were identified using a GR-RSC technique combined with 454 FLX high throughput sequencing. These SNPs represent both intra- and inter-specific SNPs identified in accessions of *G. hirsutum* and *G. barbadense*. We also identified 4,327 SNPs from a recent assembly of cotton ESTs. For many EST-based SNPs, we identified its resident genome ('A_T' or 'D_T') using diploid genome sequence data. Of these putative SNPs, we developed 1,052 KASPar-based SNP marker assays and evaluated the broad utility of 277 of them using a diverse panel of *G. hirsutum* accessions. Finally, we constructed the first genetic linkage map of *G. hirsutum* based entirely on 346 SNP markers. Hundreds of putative microsatellites were also identified.

Acknowledgments We thank Cotton Incorporated, the National Science Foundation Plant Genome Program, and BYU Mentored Environment grants for their generous support. We thank Jonathan Wendel and Armel Salmon for construction of the cotton diversity panel and its corresponding DNA samples. We also thank undergraduate students Zach Liechty, Elisabeth Svedin, Prabin Bajgain, and Justin Page for their technical assistance.

References

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3(10):e3376

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. Plant J 51(5):910–918. doi:10.1111/j.1365-313X.2007.03193.x

Brubaker CL, Wendel JF (1994) Reevaluating the origin of domesticated cotton (*Gossypium hirsutum*; Malvaceae) using nuclear restriction fragment length polymorphisms (RFLPs). Am J Bot 81(10):1309–1326

Bundock PC, Eliott FG, Ablett G, Benson AD, Casu RE, Aitken KS, Henry RJ (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. Plant Biotechnol J 7(4):347–354. doi:10.1111/j.1467-7652.2009.00401.x

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6(5): e19379

Flagel LE, Wendel JF, Udall JA (2011) Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. Genome Biolo. (Submitted)

Fluidigm (2011) Fluidigm SNP genotyping analysis. Fluidigm Corp, San Francisco

Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES (2009) A first-generation haplotype map of maize. Science 326(5956):1115–1117. doi:10.1126/science.1177837

Gutiérrez O, Jenkins J, McCarty J, Wubben M, Hayes R, Callahan F (2010) SSR markers closely associated with genes for resistance to root-knot nematode on chromosomes 11 and 14 of Upland cotton. TAG Theor Appl Genet 121(7):1323–1337. doi:10.1007/s00122-010-1391-9

Hovav R, Chaudhary B, Udall JA, Flagel L, Wendel JF (2008) Parallel domestication, convergent evolution and duplicated gene recruitment in allopolyploid cotton. Genetics 179(3):1725–1733. doi:10.1534/genetics.108.089656

KBioscience (2009) PrimerPicker Lite for KASPar v.0.26. KBioscience Ltd, Hoddesdon

Kidwell KK, Osborn TC (1992) Simple plant DNA isolation procedures. In: Beckman JS, Osborn TC (eds) Plant genomes: methods for genetic and physical mapping. Kluwer Academic Publishers, Dordrecht, pp 1–13

Lacape J-M, Jacobs J, Arioli T, Derijcker R, Forestier-Chiron N, Llewellyn D, Jean J, Thomas E, Viot C (2009) A new interspecific, *Gossypium hirsutum* × *G barbadense* RIL population: towards a unified consensus linkage map of tetraploid cotton. TAG Theor Appl Genet 119(2):281–292. doi:10.1007/s00122-009-1037-y

Li W-H (1997) Molecular Evolution. Sinauer Associates, Sunderland

Lin Z, Zhang Y, Zhang X, Guo X (2009) A high-density integrative linkage map for *Gossypium hirsutum*. Euphytica 166(1):35–45. doi:10.1007/s10681-008-9822-2

Lu H, Romero S, Bernardo R (2002) Chromosomal regions associated with segregation distortion in maize. TAG Theor Appl Genet 105(4):622–628. doi:10.1007/s00122-002-0970-9

Maughan PJ, Yourstone SM, Jellen EN, Udall JA (2009) SNP discovery via genomic reduction, barcoding and 454-pyrosequencing in Amaranth. Plant Genome 2:260–270

Maughan PJ, Yourstone SM, Byers RL, Smith SM, Udall JA (2010) Single-nucleotide polymorphism genotyping in mapping populations via genomic reduction and next-generation sequencing: proof of concept. Plant Gen 3(3):166–178. doi:10.3835/plantgenome2010.07.0016

Morton BR, Bi IV, McMullen MD, Gaut BS (2006) Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. Genetics 172(1):569–577. doi:10.1534/genetics.105.049916

Percy RG, Wendel JF (1990) Allozyme evidence for the origin and diversification of *Gossypium barbadense* L. Theor Appl Genet 79:529–542

Rapp R, Udall J, Wendel J (2009) Genomic expression dominance in allopolyploids. BMC Biol 7(1):18

Reinisch AJ, Dong JM, Brubaker CL, Stelly DM, Wendel JF, Paterson AH (1994) A detailed RFLP map of cotton, *Gossypium hirsutum* × *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. Genetics 138(3):829–847

Rong J, Abbey C, Bowers JE, Brubaker CL, Chang C, Chee PW, Delmonte TA, Ding X, Garza JJ, Marler BS, Park C-h, Pierce GJ, Rainey KM, Rastogi VK, Schulze SR, Trolinder NL, Wendel JF, Wilkins TA, Williams-Coplin TD, Wing RA, Wright RJ, Zhao X, Zhu L, Paterson AH (2004) A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). Genetics 166(1):389–417. doi:10.1534/genetics.166.1.389

Rong J, Feltus FA, Waghmare VN, Pierce GJ, Chee PW, Draye X, Saranga Y, Wright RJ, Wilkins TA, May OL, Smith CW, Gannaway JR, Wendel JF, Paterson AH (2007) Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. Genetics 176(4):2577–2588. doi:10.1534/genetics.107.074518

Shen X, Guo W, Zhu X, Yuan Y, Yu JZ, Kohel RJ, Zhang T (2005) Molecular mapping of QTLs for fiber qualities in three diverse lines in upland cotton using SSR markers. Mol Breed 15(2):169–181. doi:10.1007/s11032-004-4731-0

Smit AFA, Hubley R, Green P (1996–2010) RepeatMasker Open-3.0

Udall JA, Swanson JM, Haller K, Rapp RA, Sparks ME, Hatfield J, Yu Y, Wu Y, Dowd C, Arpat AB, Sickler BA, Wilkins TA, Guo JY, Chen XY, Scheffler J, Taliercio E, Turley R, McFadden H, Payton P, Klueva N, Allen R, Zhang D, Haigler C, Wilkerson C, Suo J, Schulze SR, Pierce ML, Essenberg M, Kim H, Llewellyn DJ, Dennis ES, Kudrna D, Wing R, Paterson AH, Soderlund C, Wendel JF (2006) A global assembly of cotton ESTs. Genome Res 16(3):441–450

Ulloa M, Meredith WR, Shappley ZW, Kahler AL (2002) RFLP genetic linkage maps from four $F_{2.3}$ populations and a joinmap of *Gossypium hirsutum* L. Theor Appl Genet 104:200–208. doi:10.1007/s001220100739

USDA (2011) Cotton and Wool Yearbook: Dataset. http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1282. http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1282

Van Deynze A, Stoffel K, Lee M, Wilkins T, Kozik A, Cantrell R, Yu J, Kohel R, Stelly D (2009) Sampling nucleotide diversity in cotton. BMC Plant Biol 9(1):125

Van Ooijen JW (2006) JoinMap 4.0, Software for the calculation of genetic linkage maps in experimental populations

Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Meth 5(3):247–252. http://www.nature.com/nmeth/journal/v5/n3/suppinfo/nmeth.1185_S1.html

Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett 7:537–546

Wallace TP, Bowman D, Campbell BT, Chee P, Gutierrez OA, Kohel RJ, McCarty J, Myers G, Percy R, Robinson F, Smith W, Stelly DM, Stewart JM, Thaxton P, Ulloa M, Weaver DB (2009) Status of the USA cotton germplasm collection and crop vulnerability. Genet Resour Crop Evol 56:507–532

Wendel JF, Cronn RC (2003) Polyploidy and the evolutionary history of cotton. Adv Agron 78:139–186

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

Wiedmann R, Smith T, Nonneman D (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genet 9(1):81

Zamir D, Tadmor Y (1986) Unequal segregation of nuclear genes in plants. Bot Gazette 147(3):355–358

Zhang F, Zhao Z (2004) The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. Genomics 84(5):785–795. doi:10.1016/j.ygeno.2004.06.015

Zhang Z-S, Hu M-C, Zhang J, Liu D-J, Zheng J, Zhang K, Wang W, Wan Q (2009) Construction of a comprehensive PCR-based marker linkage map and QTL mapping for fiber quality traits in upland cotton (*Gossypium hirsutum* L.). Mol Breed 24(1):49–61. doi:10.1007/s11032-009-9271-1

Zhang Z, Rong J, Waghmare V, Chee P, May O, Wright R, Gannaway J, Paterson A (2011) QTL alleles for improved fiber quality from a wild Hawaiian cotton, *Gossypium tomentosum*. TAG Theor Appl Genet 1–14. doi:10.1007/s00122-011-1649-x