BAC-End Sequence-Based SNP Mining in Allotetraploid Cotton (Gossypium) Utilizing Resequencing Data, Phylogenetic Inferences, and Perspectives for Genetic Mapping

Amanda M. Hulse-Kemp,*†1 Hamid Ashrafi,† Kevin Stoffel,† Xiuting Zheng,* Christopher A. Saski,§ Brian E. Scheffler,** David D. Fang,†† Z. Jeffrey Chen,‡‡ Allen Van Deynze,§ and David M. Stelly*†,††

*Department of Soil and Crop Sciences, Texas A&M University, College Station, Texas 77843, †Interdisciplinary Genetics Program, Texas A&M University, College Station, Texas 77843, ‡‡Seed Biotechnology Center, University of California, Davis, California 95616, §Clemson University Genomics Institute, Clemson University, Clemson, South Carolina 29634, **USDA-ARS, Genomics and Bioinformatics Research Unit, Stoneville, Mississippi 38776, ††USDA-ARS, Cotton Fiber Bioscience Research Unit, New Orleans, Louisiana 70124, and † mush Department of Molecular Biosciences, Center for Computational Biology and Bioinformatics, and Institute for Cellular and Molecular Biology, The University of Texas, Austin, Texas 78712

ORCID IDs: 0000-0001-9670-9433 (A.M.H.-K.); 0000-0002-2094-2367 (H.A.); 0000-0003-468-4119 (K.S.); 0000-0002-8574-4708 (X.Z.); 0000-0002-2780-4274 (C.A.S.); 0000-0003-1968-8952 (B.E.S.); 0000-0002-3468-4119 (K.S.); 0000-0002-8574-4708 (X.Z.); 0000-0002-2093-0577 (A.V.D.)

ABSTRACT A bacterial artificial chromosome library and BAC-end sequences for cultivated cotton (Gossypium hirsutum L.) have recently been developed. This report presents genome-wide single nucleotide polymorphism (SNP) mining utilizing resequencing data with BAC-end sequences as a reference by alignment of 12 G. hirsutum L. lines, one G. barbadense L. line, and one G. longicalyx Hutch and Lee line. A total of 132,262 intraspecific SNPs have been developed for G. hirsutum, whereas 223,138 and 470,631 interspecific SNPs have been developed for G. barbadense and G. longicalyx, respectively. Using a set of interspecific SNPs, 11 randomly selected and 77 SNPs that are putatively associated with the homeologous chromosome pair 12 and 26, we mapped 77 SNPs into two linkage groups representing these chromosomes, spanning a total of 236.2 cM in an interspecific F2 population (G. barbadense 3-79 × G. hirsutum TM-1). The mapping results validated the approach for reliably producing large numbers of both intraspecific and interspecific SNPs aligned to BAC-ends. This will allow for future construction of high-density integrated physical and genetic maps for cotton and other complex polyploid genomes. The methods developed will allow for future Gossypium resequencing data to be automatically genotyped for identified SNPs along the BAC-end sequence reference for anchoring sequence assemblies and comparative studies.

KEYWORDS
SNP genotyping cotton genomics resequencing BAC-derived SNPs intraspecific

Marker development in crop species has been an important aspect to facilitate genomics-based crop improvement. Single nucleotide polymorphisms (SNPs) are the most abundant form of markers in all organisms, because they have a possibility of occurring at every nucleotide position in the genome. This makes them ideal candidates for construction of high-density genetic maps, which can then be used for marker-based crop improvement and genetic analyses. In general, for marker-assisted selection, a large number of molecular markers are required in crop species as the majority of the traits of interest such as yield, drought and heat tolerance, nitrogen and water use efficiency, disease resistance, and fiber quality are complex and controlled by many genomic loci of small effect. Therefore, for marker-assisted breeding to be effective, markers for most of the regions controlling a trait need to be included in selection criteria (Mammadov et al. 2012). SNPs have been found to occur approximately every 60–120 bp in the maize genome (Ching et al. 2002), every 268 bp in the rice genome (Shen et al. 2002).
2004), and every 185–266 bp in the soybean genome (Lam et al. 2010). The markers developed to date in cotton have been limited mostly to amplified fragment length polymorphisms (AFLPs), restriction fragment length polymorphisms (RFLPs), and simple sequence repeats (SSRs). Some recent efforts to identify SNPs in cotton using transcriptome sequencing in intraspecific and interspecific cotton lines (Hulse-Kemp et al. 2014; Zhu et al. 2014; Ashrafi et al. 2015), single copy sequences between G. hirsutum and G. barbadense (Van Deynze et al. 2009), and a variety of reduced representation libraries (RRLs) techniques investigating combinations of G. hirsutum lines (Byers et al. 2012; Rai et al. 2013; Zhu et al. 2014; Islam et al. 2015; Gore et al. 2014) have been reported and have been increasing in efficiency using a diverse range of germplasm. These studies of cotton have generated anywhere from a few hundred SNPs to tens of thousands of SNPs. However, success with cotton has been limited compared to crop species such as maize (Elshire et al. 2011; Hansey et al. 2012), soybean (Hyten et al. 2010), barley, and wheat (Poland et al. 2012), which also used a broad range of germplasm and have been able to develop hundreds of thousands of SNPs.

Reasons for reduced success with SNP development within Gossypium compared to efforts in other crops are perhaps due to the evolutionary history of the cotton genome and recent polyploid generation. The most widely cultivated species of cotton, Gossypium hirsutum L., descends from a recently formed allotetraploid hybrid (1–2 MYA), 2n = 4x = 52, genomic equation 2[AD], (Wendel et al. 2009). Modern cultivated cotton, including Upland types, have undergone at least two independent bottleneck events, domestication, and elite cultivar selection, which further reduced the overall diversity that can be found in elite cultivars. The ancestral A- and D-like genomes are thought to have diverged only 5–10 MYA. Because of limited time for evolutionary divergence, homologous regions of the A- and D-subgenomes of cotton retain a very high similarity. Although reference sequences have recently become available for the diploid cottons, G. raimondii (extant relative to the allotetraploid D-subgenome) (Paterson et al. 2012), and G. arboreum (extant relative to the allotetraploid A-subgenome) (Li et al. 2014), a reference is not yet available for allotetraploid G. hirsutum.

Large-scale SNP development in cotton has been hindered by the lack of a high-quality G. hirsutum reference which, in concert with high levels of similarity between subgenomes and low levels of diversity, makes it very difficult to unambiguously map short-read sequences obtained from next-generation sequencing technologies (Zhu et al. 2014). Recently, we developed three independent bacteria artificial chromosome (BAC) libraries, two generated from restriction-enzyme partial digestion (BstYI/HindIII) and one generated by random-shearing (C. A. Saksi, A. M. Hulse-Kemp, I. Schmutz, B. Liu, D. M. Stelly, J. A. Scheﬄer, D. C. Jones, D. G. Peterson, Z. J. Chen, and B. E. Scheﬄer). From these, 179,209 high-quality BAC-end sequences (BESs) were generated. Development of markers aligned to BAC-end sequences is desired for several reasons. First, the quality of Sanger-sequencing is currently the most widely cultivated species of cotton, G. hirsutum. Second, the markers can serve as a rigid interface between a BAC-based physical map with genetic maps. Third, taken together, long-range contiguity and marker distribution can begin to be contextualized on a genome-wide scale. Recent reports suggest the utility of BESs in simple sequence repeat (SSR) marker development in pigeon pea (Bohra et al. 2011), cotton (Freichowski et al. 2006), and peanut (Wang et al. 2012), and for SNP development via PCR-directed sequencing methods in apple (Han et al. 2009) and in Citrus (Ollitrault et al. 2012) to anchor genomic sequences.

Thus, the primary goals of the present study were: to develop large numbers of genomic-based SNPs for cultivated cotton, G. hirsutum; to compare the levels of diversity among elite cultivars, a wild accession, an additional tetraploid species (G. barbadense L.), and a diploid cotton species (G. longicalyx); to experimentally validate in silico–derived SNPs; and to demonstrate mapping ability and utility of developed SNPs. This article reports on utilizing BESs as a high-quality reference for SNP discovery through resequencing. The new SNPs will be a resource for SNP-based integration of the physical and genetic maps, and the methodology can also serve as a useful model in resolving other complex plant genomes.

**MATERIALS AND METHODS**

**Source of BAC clones and BAC-end sequencing**

Two complementary BAC libraries (BstYI and HindIII) and a random sheared BAC library from G. hirsutum genetic standard line Texas Marker-1 (TM-1) were used in this study. BAC DNA was isolated and sequenced by Sanger methods at USDA-ARS (Stoneville, MS). A total of 179,209 BAC-end sequences were retained after quality trimming and filtering (LIBGSS_039228) (Saksi et al. unpublished data).

**Plant material and DNA extraction**

The seed of 11 G. hirsutum lines (TM-1, Sealand 542, PD-1, Paymaster HS-26, M-240 RNR, Fibermax 832, Coker 312, SureGRow 747, Stoneville 474, Tamcot Sphinx, and TX0231) and the G. barbadense genetic standard line, 3-79, were planted at Texas A&M University, University of California Davis, or the USDA-ARS facility in New Orleans, Louisiana. Young leaf tissues were sampled from each plant and used to isolate genomic DNA using the Qiagen DNeasy (Qiagen, Valencia, USA) plant extraction kit following manufacturer instructions, including RNase digestion. DNA concentrations and qualitative absorbance values were determined using the Nanodrop Spectrophotometer (Thermo Fisher Scientific, Waltham, MA).

**Library preparation and sequencing**

Double-stranded DNA was quantiﬁed with a PicoGreen assay on the Synergy HT plate reader (Bio-Tek, Highland Park, IL). Libraries were prepared using an in-house protocol with individual barcoding. One and a half micrograms of each DNA sample was randomly sheared using a Bioruptor instrument (Diagenode, Denville, NJ), and then size selection was performed using AMPure XP beads to 300–500 bp. Fragments were end-repaired using NEBNext End Repair Module and then puriﬁed with AMPure beads. Next, an adenine was ligated at the end of the fragments, adapters were ligated, and the ﬁnal products were puriﬁed with AMPure beads. Enrichment PCR was performed for 14 cycles, and then the PCR product was run on a 1.5% gel to conﬁrm enrichment of product and size range. A ﬁnal round of puriﬁcation was performed using AMPure beads. The final libraries were assessed for quality with the Bioanalyzer (Agilent, Santa Clara, CA) to determine ﬁnal library size and concentration. Each sample was sequenced with paired-end sequencing (2 × 100 bp on two Illumina HiSeq2500 lanes. Raw, paired-read sequence ﬁles were uploaded to NCBI under BioProject PRJNA257223 and SRA numbers (SRX667500, G. hirsutum TM-1; SRX668168, G. hirsutum Sealand 542; SRX668322, G. hirsutum PD-1; SRX668354, G. hirsutum Paymaster HS-26; SRX669467, G. hirsutum M-240 RNR; SRX669468, G. hirsutum Fibermax 832; SRX669469, G. hirsutum Coker 312; SRX669470, G. hirsutum SureGRow 747; SRX669471, G. hirsutum Stoneville 474; SRX669472, G. hirsutum Tamcot Sphinx; SRX669473, G. hirsutum TX0231; SRX669474, G. barbadense 3-79). Raw reads for G. hirsutum Acala Maxxa and for G. longicalyx were
obtained from the NCBI Small Read Archive under numbers SRR617482 and SRR617704.

**SNP mining from G. hirsutum aligned to BAC-end sequences**

All raw sequence files were assessed for initial quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and the first 13 bases from each read were removed due to poor quality. The remaining reads were quality-trimmed, and adapters and any reads fewer than 40 bases were removed using the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). The files were assessed for final quality and concatenated into a single file to be utilized as single-end read data due to the size of the BESs being used as a reference. Once the sequences had been processed, they were imported into CLC Genomics Workbench v 6.0.2 (Valencia, CA). Reads from Sealand 542, PD-1, and 3-79 were aligned to the BAC-end sequence reference using different length and similarity fractions to determine optimal parameters. Iteration-1 utilized length fraction 0.70 and similarity fraction 0.99, whereas Iteration-2 utilized length fraction 0.99 and similarity fraction 0.98. SNPs were called using the Probabilistic variant caller in CLC Genomics Workbench using a minimum sequence depth of six, variant probability of 50.0, required presence in both forward and reverse reads, four maximum expected variants, and one standard genetic code. Theoretical homeo-SNP positions (variations between homeologous regions caused by ambiguous mapping of homeologous reads from different subgenomes in the allotetraploid) were determined by aligning the TM-1 Illumina-based sample, which is the same genotype as the reference, to call SNPs using the same parameters as all other samples. These identified homeo-SNP positions were removed from the SNPs identified in the other samples.

A random set of called 32 SNP positions was manually viewed to assess quality of alignments around the SNP positions, which included markers only found in Iteration-1 and markers only found in Iteration-2; markers found in both Iteration-1 and Iteration-2 from the SNPs discovered using *G. barbadense* 3-79 were selected for empirical testing. Half of the 32 markers were identified as overlapping transcriptome-derived SNPs as reported in Hulse-Kemp *et al.* (2014), whereas half of the SNPs did not overlap the Hulse-Kemp *et al.* dataset and thus were assumed to not be associated with the transcriptome. Primers were designed for KASP (LGC Genomics) SNP assays.

### Table 1. KASP assay screening results for *G. barbadense* derived markers

| Type                          | Marker Name | KASP Result | Identified in *G. barbadense* Mapping | Homeo-SNP Removal | Final Result |
|-------------------------------|-------------|-------------|--------------------------------------|-------------------|-------------|
|                               |             |             | iteration 1 | iteration 2 | iteration 1 | iteration 2 | iteration 1 | iteration 2 |
| Not Transcriptome-Associated  | GH_Tbb001A07f_381 | Good        | Yes        | Yes        | Retain     | Retain     | Good        | Good        |
| GH_Tbb001A07f_486             | Bad         | No          | Yes        | —          | —          | —          | —           | —           |
| GH_Tbb001A23r_531             | Bad         | No          | Yes        | —          | Retain     | —          | —           | —           |
| GH_Tbb001A23r_614             | Good        | Yes         | Yes        | —          | Retain     | Good       | Good        | —           |
| GH_Tbb001D22r_204             | Good        | Yes         | Yes        | —          | Retain     | Good       | Good        | —           |
| GH_Tbb001D22r_445             | Bad         | Yes         | No         | —          | Retain     | —          | —           | —           |
| GH_Tbb001D22r_511             | Bad         | Yes         | Yes        | —          | Retain     | Bad        | —           | —           |
| GH_Tbb001B05f_180             | Good        | Yes         | Yes        | Remove     | Remove     | —          | —           | —           |
| GH_Tbb001B05f_564             | Bad         | Yes         | Yes        | Remove     | Remove     | —          | —           | —           |
| GH_Tbb001C03f_116             | Good        | No          | Yes        | —          | Retain     | —          | —           | —           |
| GH_Tbb001C03f_401             | Good        | Yes         | Yes        | Retain     | Retain     | Good       | Good        | —           |
| GH_Tbb001F01r_117             | Bad         | Yes         | Yes        | Remove     | Remove     | —          | —           | —           |
| GH_Tbb001F01r_310             | Bad         | Yes         | No         | Remove     | —          | —          | —           | —           |
| GH_Tbb001A17r_218             | Bad         | Yes         | Yes        | Remove     | Remove     | —          | —           | —           |
| GH_Tbb001F06f_303             | Bad         | No          | Yes        | —          | Remove     | —          | —           | —           |

| Transcriptome-Associated      | GH_Tbb004J20r_76 | Good        | Yes        | Yes        | Remove     | Remove     | —           | —           |
| GH_Tbb004J20r_348             | Good        | Yes         | Yes        | Retain     | Retain     | Good       | Good        | —           |
| GH_Tbb053N14f_270             | Good        | Yes         | Yes        | Retain     | Retain     | Good       | Good        | —           |
| GH_Tbb036B20r_583             | Good        | No          | Yes        | —          | Retain     | —          | —           | —           |
| GH_Tbb162H20f_547             | Bad         | Yes         | Yes        | Remove     | Remove     | —          | —           | —           |
| GH_Tbb046002r_64              | Bad         | Yes         | Yes        | Remove     | Remove     | —          | —           | —           |
| GH_Tbb046002r_138             | Bad         | Yes         | Yes        | Remove     | Remove     | —          | —           | —           |
| GH_Tbb046002r_418             | Bad         | Yes         | Yes        | Remove     | Remove     | —          | —           | —           |
| GH_Tbb069A06f_248             | Bad         | Yes         | No         | Retain     | —          | —           | Bad         | —           |
| GH_Tbb069A06f_304             | Bad         | Yes         | Yes        | Remove     | Remove     | —          | —           | —           |
| GH_Tbb034D07r_276             | Good        | Yes         | Yes        | Retain     | Retain     | Good       | Good        | —           |
| GH_Tbb030P12r_332             | Bad         | Yes         | Yes        | Remove     | Remove     | —          | —           | —           |
| GH_Tbb119O19f_465             | Maybe       | Yes         | Yes        | Retain     | Retain     | N/A        | N/A         | —           |

| GH_Tbb055K17f_57              | Bad         | Yes         | Yes        | Remove     | Remove     | —          | —           | —           |
| GH_Tbb055K17f_506             | Bad         | Yes         | Yes        | Remove     | Remove     | —          | —           | —           |
| GH_Tbb023C021f_162            | Good        | Yes         | No         | Remove     | Retain     | —          | —           | —           |
| GH_Tbb023C021f_537            | Good        | Yes         | Yes        | Retain     | Retain     | Good       | Good        | —           |

**Note:** KASP assay screening results for *G. barbadense*-derived markers for mapping parameter optimization in CLC Genomics Workbench. Iteration-1 was performed using the following mapping parameters: 0.70 length fraction and 0.99 similarity fraction. Iteration-2 was performed using the following mapping parameters: 0.99 length fraction and 0.98 similarity fraction.
using BatchPrimer3 (allele-specific primers and allele-flanking primers; Tm: optimum, 57°C; minimum, 55°C; maximum, 60°C; max difference, 2°C; product size: minimum, 50 bp; optimum, 50 bp; maximum, 100 bp). Primers were synthesized and diluted according to KASP developer (LGC Genomics, Hoddesdon, UK) instructions. KASP assays were run on a “G. barbadense screening panel” (Supporting Information, Table S1) containing 12 samples including TM-1 (Stelly Lab), TM-1 (USDA), 3-79 (×2), F1 – 3-79×TM-1 (×2), RIL01-04 (3-79×TM-1), and water nontemplate control (×2) according to the manufacturer’s suggested PCR conditions. Plates were read using the Pherastar at 38, 44, and 50 cycles, and then they were analyzed using the KlusterCaller program. SNP assays were labeled as “good” if samples produced clean clusters that allowed for differentiation and scoring of the two parents and the F1 genotypes, or as “bad” if no definable clusters were produced or no amplification occurred. (These definitions of “good” and “bad” are used subsequently throughout the rest of the article.) Markers included in the “good” and “bad” categories for Iteration-1 and Iteration-2 were calculated (Table 1) to determine optimal parameter settings.

All 12 G. hirsutum samples and the G. barbadense sample were analyzed using Iteration-2 parameters. The G. longicalyx sample was analyzed using the same length fraction but similarity fraction of 0.96. Samples were processed individually and homo-SNPs identified using TM-1 were removed. Unique SNP positions were collated into a master file and required 100% homozygous identification in at least one G. hirsutum line. Positions where coverage was 1 to 2 SDs above the average coverage in the homozygous sample(s) were removed, because regions with high coverage are likely to be associated with

**Figure 1** Distance between polymorphisms developed in (A) G. hirsutum, (B) G. barbadense, and (C) G. longicalyx relative to the G. hirsutum-derived BAC-end sequences.
A random set of 32 markers developed from *G. longicalyx* were selected and primers were designed. KASP assays were run per manufacturer’s instructions on a “*G. longicalyx* screening panel” that contained *G. longicalyx* (x2), *G. hirsutum* cv. TM-1 (x2), synthetic allotetraploid “FADD” (x2), 2 BC1F1 samples (FADD × TM-1), *G. barbadense* line 3-79, *G. hirsutum* accession TX0231, and two non-template controls (Table S3). Plates were cycled and analyzed as previously mentioned. SNP sequences and primers for all screened markers are listed in Table S4.

### Interspecific linkage mapping

Good markers obtained in screening of the *G. barbadense* SNPs were used to genotype 118 F2 (3-79×TM-1) individuals, two parents (*G. hirsutum* line TM-1 and *G. barbadense* line 3-79), and F1 (3-79×TM-1) as controls. KASP assays were run using the Fluidigm system in 96.96 dynamic array format, utilizing multiple arrays, and read using the Fluidigm BioMark HD (Fluidigm, San Francisco, CA). Clustering for genotyping was performed using Fluidigm SNP Genotyping Analysis software. Genotypes of the 118 F2s for successfully genotyped markers were imported into JoinMap 4.1 (Van Ooijen 2011), and identical markers were removed (Table S5) and linkage mapped using the maximum likelihood algorithm and Haldane’s mapping function with default parameters. Linkage groups were determined using LOD scores of 5.0 or more. Cytogenetic stocks including F1 hypo-aneuploids, each deficient for a known *G. hirsutum* chromosome, were also genotyped for the same markers.

### Phylogenetic analysis

The VCF file produced from GATK that included genotypes from all 14 samples was imported into R (R Development Core Team 2010). The SNPRelate package (Zheng et al. 2012) was used to perform a principle component analysis, and eigenvalue1 and eigenvalue2 were used to visualize samples. A distance matrix was determined between samples and then used for hierarchical clustering over 10,000 permutations (z.threshold = 15, outlier.n = 2). The clustering results were used to create a dendrogram tree of the 14 samples.

### RESULTS

**Candidate SNPs derived from BAC-end sequences**

We identified a total of 132,262 intraspecific SNPs for *G. hirsutum* that occurred, on average, every 888 base pairs (bp). The distribution of base pairs between adjacent intraspecific SNPs found on the same
BES is, on average, 76 bp (Figure 1A). Interspecific SNPs for *G. barbadense* and *G. longicalyx* occurred at a much higher rate than intraspecific SNPs, which was expected due to longer divergence time between the species. A total of 223,138 interspecific SNPs between *G. barbadense* and *G. hirsutum* were determined. Although SNPs were similarly spaced (78 bp) as the intraspecific SNPs, a larger number of SNPs was identified compared to intraspecific SNPs because they occurred on more BESs (Figure 1B). A total of 470,631 interspecific SNPs were identified between *G. longicalyx* and *G. hirsutum*. The distance distribution between adjacent SNPs found on the same BES from *G. longicalyx* is quite different from the *G. hirsutum* and *G. barbadense* distributions, and it shows that SNPs are more likely to be in close proximity if found on the same BES due to an overall elevated number of SNPs (Figure 1C). This difference is expected and reflects the greater divergence of *G. longicalyx*. Overlap of SNPs identified in multiple species is shown in Figure 2. Considering all identified polymorphisms across all three species, a SNP was identified between *G. longicalyx*, *G. hirsutum* and *G. barbadense*. The percentage of missing loci was related to evolutionary divergence between the samples analyzed and the reference, with *G. longicalyx*, *G. barbadense*, and wild *G. hirsutum* line TX0231 exhibiting the largest percentages of missing data (in decreasing order). High levels of missing data in *G. longicalyx* were expected because as a diploid, it would not contain loci from both A- and D-subgenomes of the tetraploid species, but rather would contain mostly only A-subgenome-like loci because it is closely related to the A-genome diploids (Phillips and Strickland 1966). Homozygous differences, which are positions for which two samples are both homozygous for a different base at a single locus, between samples were counted for each pair of samples (Table 4). Overall the number of differences between pairs of cultivated *G. hirsutum* samples was quite variable, ranging from 5562 (Paymaster HS-26 vs. Fibermax 832) to 29,052 (TM-1 vs. Tamcot Sphinx). Differences between cultivated samples and the uncultivated *G. hirsutum* TX0231 were, in general, quite similar, as was the case between all *G. hirsutum* samples with *G. barbadense* and *G. longicalyx*.

To investigate the functional characteristics of the identified SNPs, all 179,209 BESs used as a reference were first repeat masked, which resulted in a total of 13 million masked bases (11% of total BES length). The masked bases were largely composed of interspersed repeats and retroelements. The remaining bases were aligned to the Genbank nonredundant protein database, which resulted in 9.48% of all BESs having a hit to coding sequences. When all determined SNPs were considered, a similar proportion of genic SNPs (~7%) was identified in *G. hirsutum* and *G. barbadense*, whereas a larger proportion was identified in *G. longicalyx* (Table 5).

**Conversion of in silico SNPs to assays**

*G. barbadense* SNPs developed in silico were randomly and nonrandomly (based on theoretical location on chromosomes 12 and 26 based on alignment to the D3 sequence) selected for primer design and experimental screening. Initially, a random set of 32 *G. barbadense* markers was

---

**Table 2 Distribution of SNP types identified in *G. hirsutum*, *G. barbadense*, and *G. longicalyx***

| Total SNP | G. hirsutum | G. barbadense | G. longicalyx | Overall |
|-----------|-------------|---------------|---------------|---------|
| 132,262 | 100.0% | 187,355 | 100.0% | 450,577 | 100.0% |
| M (A/C) | 11,146 | 8.4% | 16,558 | 8.8% | 42,802 | 9.5% | 70,506 | 9.2% |
| R (A/G) | 45,444 | 34.4% | 64,420 | 34.4% | 139,410 | 30.9% | 249,274 | 32.4% |
| W (A/T) | 11,992 | 9.1% | 16,021 | 8.6% | 60,070 | 13.3% | 88,083 | 11.4% |
| S (C/G) | 45,408 | 34.3% | 64,574 | 34.5% | 139,620 | 31.0% | 249,602 | 32.4% |
| K (G/T) | 11,410 | 8.6% | 16,594 | 8.9% | 43,255 | 9.6% | 71,259 | 9.3% |
| Total transition | 90,852 | 68.7% | 128,994 | 68.9% | 279,030 | 61.9% | 498,876 | 64.8% |
| Total transversion | 41,410 | 31.3% | 58,361 | 31.1% | 171,547 | 38.1% | 271,318 | 35.2% |

---

**Table 3 Description of missing data and heterozygous loci in the final VCF file for *G. hirsutum*, *G. barbadense*, and *G. longicalyx***

| Species | Sample | Missing (No.) | Missing (%) | Heterozygous (No.) | Heterozygous (%) |
|---------|--------|---------------|-------------|-------------------|-----------------|
| *Gossypium hirsutum* | TM-1 | 2144 | 0.28% | 16,880 | 2.19% |
| | Sealand 542 | 4180 | 0.54% | 22,460 | 2.93% |
| | PD-1 | 4435 | 0.58% | 27,039 | 3.53% |
| | Paymaster HS-26 | 4450 | 0.58% | 38,201 | 4.99% |
| | M-240 RNR | 4416 | 0.54% | 30,450 | 3.97% |
| | Fibermax 832 | 4994 | 0.65% | 31,862 | 4.16% |
| | Coker 312 | 4604 | 0.60% | 21,831 | 2.85% |
| | SureGrow 747 | 4187 | 0.54% | 25,839 | 2.82% |
| | Stoneville 474 | 4058 | 0.53% | 21,554 | 2.81% |
| | Tamcot Sphinx | 5934 | 0.77% | 22,342 | 2.92% |
| | Acala Maxxa | 5159 | 0.67% | 20,443 | 2.81% |
| | TX0231 | 12,057 | 1.57% | 24,974 | 3.29% |
| *Gossypium barbadense* | 3-99 | 43,530 | 5.65% | 20,443 | 2.81% |
| | F1-1 | 221,786 | 28.80% | 30,204 | 5.51% |
Table 4 Pairwise comparison of homozygous different genotypes between 14 resequenced samples using BCFtools command gtcheck

| Genotype       | Sealand 542 | PD-1 | M-240 | Fibermax | Coker | SureGrow | Stoneville | Tamcot | Sphinx | Maxxa TX0231 | 3-79 |
|---------------|-------------|------|-------|----------|-------|----------|------------|--------|--------|---------------|------|
| Sealand 542   | 15,755      | 11,672 | 12,249 | 13,484   | 13,462 | 13,549   | 12,854     | 14,552 | 13,668 | 11,657        | 20,141 |
| PD-1          | 15,755      | 11,672 | 12,249 | 13,484   | 13,462 | 13,549   | 12,854     | 14,552 | 13,668 | 11,657        | 20,141 |
| M-240         | 15,755      | 11,672 | 12,249 | 13,484   | 13,462 | 13,549   | 12,854     | 14,552 | 13,668 | 11,657        | 20,141 |
| Fibermax      | 15,755      | 11,672 | 12,249 | 13,484   | 13,462 | 13,549   | 12,854     | 14,552 | 13,668 | 11,657        | 20,141 |
| Coker         | 15,755      | 11,672 | 12,249 | 13,484   | 13,462 | 13,549   | 12,854     | 14,552 | 13,668 | 11,657        | 20,141 |
| SureGrow      | 15,755      | 11,672 | 12,249 | 13,484   | 13,462 | 13,549   | 12,854     | 14,552 | 13,668 | 11,657        | 20,141 |
| Stoneville    | 15,755      | 11,672 | 12,249 | 13,484   | 13,462 | 13,549   | 12,854     | 14,552 | 13,668 | 11,657        | 20,141 |
| Tamcot        | 15,755      | 11,672 | 12,249 | 13,484   | 13,462 | 13,549   | 12,854     | 14,552 | 13,668 | 11,657        | 20,141 |
| Sphinx        | 15,755      | 11,672 | 12,249 | 13,484   | 13,462 | 13,549   | 12,854     | 14,552 | 13,668 | 11,657        | 20,141 |
| Maxxa TX0231  | 15,755      | 11,672 | 12,249 | 13,484   | 13,462 | 13,549   | 12,854     | 14,552 | 13,668 | 11,657        | 20,141 |
| 3-79          | 15,755      | 11,672 | 12,249 | 13,484   | 13,462 | 13,549   | 12,854     | 14,552 | 13,668 | 11,657        | 20,141 |

Selected to determine optimal parameters for homeo-SNP removal. Screening produced a total of 13 SNPs (40.6%) that were found to have acceptable clustering patterns, of which seven (53.8%) were associated and six (46.2%) were not associated with the transcriptome, respectively (Table 1). Thus, no difference was seen for selecting markers based on association with the transcriptome. When attempting to determine the best parameters for in silico SNP calling, results from Iteration-1 and Iteration-2 were compared to determine which iteration resulted in the set of markers with the highest overall validation rate. Taking into account the 13 good markers, Iteration-1 produced a success rate of 61.1% (11/18), whereas Iteration-2 produced a success rate of 84.6% (11/13); therefore, Iteration-2 mapping parameters (0.99 length fraction and 0.98 similarity fraction) were utilized for final mapping of all samples, except for G. longicalyx, as noted in the Materials and Methods (Table S6).

Utilizing the SNPs identified in G. barbadense, following homeo-SNP removal, an additional 96 interspecific SNPs with successfully designed primers were selected for screening based on inferred positioning on chromosomes 12 and 26. A total of 77 (or 80.2%) markers were categorized as good markers (Table S7).

A total of 48 intraspecific SNP markers were screened on 32 G. hirsutum samples and G. barbadense line 3-79, which produced 40 good markers or an 83.3% success rate (Table S8). To obtain primers for the 48 markers, 87 SNPs were randomly selected for primer design because primer design was successful in 56% of cases. Sample call rates ranged from 0.175 to 1.000. Excluding the two lowest samples, all samples had greater than 72.5% call rate. Marker call rates ranged from 0.545 to 0.970.

A total of 32 G. longicalyx interspecific SNP markers amenable to successful primer design were selected for screening. Primer design for this diploid species was significantly more difficult compared to the other species due to the elevated number of SNPs and their close proximity within a BES (Figure 1C). On screening, a total of 31 (96.9%) SNPs were determined to produce good assays (Table S9). Two of the markers listed as good but exhibit very close clusters due to the higher sequence conservation in genic regions. Markers were also aligned to the A2 genome (D5) reference genome. The starkly different percentage of mapped markers is likely to lead to difficulties for cluster separation with additional samples. So, conservatively, 29 would be regarded as good markers, i.e., a 90.6% success rate.

Alignment of markers to diploid cotton genomes

When all markers from all three species identified were aligned to the high-quality G. raimondii (D5) reference (Paterson et al. 2012) genome, 38.8% of markers could be aligned to the genome. This was distinctly different from transcriptome-derived markers reported by Hulse-Kemp et al. (2014), of which 75.9% could be aligned to the D5 genome. The starkly different percentage of mapped markers is likely due to the higher sequence conservation in genic regions. Markers were also aligned to the G. arboreum (A2) draft genome (Li et al. 2014). The percentages of markers aligning to the A2 and D5 reference genomes were very similar for G. hirsutum (70.6%/40.9%) and G. barbadense (69.6%/41.1%), whereas the percentage was very different for G. longicalyx (55%/18.4%). A larger fraction of markers that can be aligned to the A2 genome is expected for G. longicalyx because it is an F-genome diploid that is closely related to A-genome diploids, including G. arboreum, and it is relatively distant from the D5 species, G. raimondii (Wendel et al. 2009).

Utilizing the alignment information to both A- and D-diploid reference genomes, markers can theoretically be localized to either the A-subgenome or the D-subgenome in the tetraploid by considering whether the markers align uniquely to either the A2 or D5. In both G. hirsutum and G. barbadense, ~52–53% of markers aligned uniquely to...
A2, so they can be putatively localized to the A-subgenome, whereas 22–24% of markers aligned uniquely to D5 can be putatively localized to the D-subgenome. Because the A2 is approximately twice the size of the D5, it is expected that a larger percentage of markers should be uniquely localized to A2 (Hendrix and Stewart 2005). G. longicalyx had approximately the same proportion of A-specific markers, but the proportion of D-specific markers is considerably lower (12.6%), which is expected because it is closely related to A-genome diploids.

Interspecific linkage mapping of candidate SNPs and anchoring of contigs

From the random and nonrandom (selected for chromosomes 12 and 26 based on alignment to D5 reference sequence) screened SNPs identified from G. barbadense, genotypes were obtained for 88 markers that represent 67 unique BAC-end sequences. These 88 markers were screened against a population of 118 F2 (3-79 × TM-1) individuals and F1 hypo-aneuploid stocks. On linkage mapping of the F2s with JoinMap 4.1 software, two linkage groups representing a total of 236.2 cM were obtained (Figure 3). As expected, due to being randomly selected from the entire data set, all of the markers tested from the randomly selected set (11) were not linked and were listed as singletons (Table S10). The resulting two linkage groups were identified as allotetraploid homeologous chromosomes 12 and 26 by loss of heterozygosity in F1 hypo-aneuploid samples and by alignment information to the D5 reference genome.

PCA and dendrogram analysis

Principal component analysis with the SNPRelate program was able to successfully separate G. hirsutum samples from the other species, G. barbadense and G. longicalyx (Figure S2). The analysis also showed a slight difference with cultivated G. hirsutum lines and the one wild G. hirsutum line TX0231. Similarly a dendrogram compiled with the SNPRelate program (Figure 4) was able to separate the wild species and indicated G. longicalyx as the out-group, as expected. It also showed the kinship coefficient between G. hirsutum samples to be extremely high (near 1) and the individual dissimilarity, or difference between individuals, to be very low (close to 0).

DISCUSSION

Reliability of the SNP-based integration of physical and genetic maps

The resulting linkage maps demonstrate the feasibility of integrating physical maps by utilizing markers associated with BAC-end sequences for genetic mapping. In this way, a large number of markers can be identified using resequencing data and can be genetically mapped with a moderately sized mapping population to obtain genetic maps associated with BAC resources. We demonstrate that even with an F2 mapping population of 118 individuals, we are able to obtain recombination events within the length of the BAC (average 120 kilobases) and are able to separate markers associated with the forward and reverse reads of a BAC in some instances. Additionally, multiple

| Table 5 Annotation of BAC-derived SNPs |
|--------------------------------------|
| **Species** | **BES-derived SNPs Identified** | **Genic** | **Nongenic** | **Overall** |
|            | No. | %     | No. | %     |
| G. hirsutum | 10,329 | 7.81% | 121,933 | 92.19% | 132,262 |
| G. barbadense | 13,274 | 7.08% | 174,081 | 92.92% | 187,355 |
| G. longicalyx | 97,366 | 21.61% | 353,211 | 78.39% | 450,577 |
| Total       | 120,969 | 15.71% | 649,225 | 84.29% | 770,194 |

Figure 3 Linkage groups determined utilizing 118 interspecific (G. barbadense line 3-79 × G. hirsutum line TM-1) F2 samples for BAC-end-derived SNPs in JoinMap.
Dendrogram of 14 Gossypium Samples using BES-derived SNP

![Dendrogram](image)

Figure 4 Dendrogram produced by hierarchical clustering utilizing BAC-end sequence-associated SNPs for 12 G. hirsutum samples (TM-1, Sealant 542, PD-1, Paymaster HS-26, M-240 RNR, Fibermax 832, Coker 312, SureGrow 747, Stoneville 474, Tamcot Sphinx, Acala Maxxa, TX0231), G. barbadense (3-79), and G. longicalyx using the SNPRelate package in R.

markers from the same BES were assayed on the population and linkage mapping places of all those markers except for two from GH_TBh104B19r in the same mapping position. The accuracy of mapping BES-associated SNPs makes it possible to utilize SNP GH_TBh104B19r in the same mapping position. The accuracy of linkage mapping places of all those markers except for two from markers from the same BES were assayed on the population and linkage mapping places of all those markers except for two from GH_TBh104B19r in the same mapping position. The accuracy of mapping BES-associated SNPs makes it possible to utilize SNP mapping for ordering and orienting BACs and BAC contigs when SNPs are identifiable on both forward and reverse end sequences of single BACs, or when multiple SNPs are identified within a given BES.

Factors that affect development of BES-associated SNPs in cotton

Development of genomic-based SNPs in cotton has largely been hindered by availability of a high-quality reference, which has led to development efforts largely targeting transcriptome-based SNPs or SNPs identified using reduced representation libraries. In general, genomic enablement for modern cultivated cotton through marker-assisted breeding is constrained by extremely low diversity and large identical regions between the subgenomes. Because of the high similarity between homeologous chromosomes and particularly genic regions, it is extremely difficult to use short next-generation sequencing technology-derived reads to uniquely localize them to a reference sequence. The availability of high-quality BAC-end sequences from a newly developed BAC library resource has provided a high-quality reference of Sanger reads. With a relatively even distribution across the subgenomes, the BESs are a superior reference with very high quality, allowing mapping of short NGS reads obtained from resequencing that have much higher error rates. In this study, this is achieved by using very high stringencies over the entire length of quality-trimmed NGS reads. However, this mapping approach is not feasible with references that are not specific to allotetraploid cotton. When the reference is known to be of much higher quality compared to typical next-generation assemblies, assay development from in silico-derived SNPs using this reference is also greatly enhanced. The flanking sequence based on the BES reference should theoretically be correct for TM-1, which can represent cultivated cotton with expectedly low levels of overall diversity and high levels of synteny; this hypothesis holds up to experimental testing as large number of SNPs are able to be successfully genotyped. This is likely due to the automatic inclusion of correct haplotype information for homeo-SNP alleles, which has been shown to increase percentage of co-dominant markers and overall success rates within polyploid cotton (Islam et al. 2015).

Although success rates of the BES-associated SNPs are higher than those of previous NGS and RRL approaches (Byers et al. 2012; Rai et al. 2013; Gore et al. 2014; Van Deynze et al. 2009; Hulse-Kemp et al. 2014; Zhu et al. 2014; Islam et al. 2015), all in silico SNPs were not verified. This could be due to many different reasons, primarily because the BES reference only represents a small portion of the cotton genome, mostly noncoding intergenic regions. Thus, although reads are required to have unique mapping to the reference, it is still possible that additional instances of the sequence exist in the overall cotton genome. This is particularly likely due to ancient paleo-polyploid events that have been discovered in the cotton evolutionary lineage (Paterson et al. 2012). Even in largely unique genic regions, gene families and duplicates are common, which will create falsely identified SNPs that, when experimentally assayed, will produce unidentifiable clusters and result in unreliable or bad markers.

Utilization of BES-associated markers in G. hirsutum germplasm

The availability of large numbers of intraspecific SNP markers is essential for marker-assisted breeding in G. hirsutum lines, particularly with elite cultivars, which exhibit very low levels of polymorphism. Within a typical elite-by-elite single cross, only a very small percentage of markers will subsequently segregate (typically <5%), so it is difficult to map large numbers of markers from intraspecific crosses. To date, this has led to use of interspecific crosses for the majority of mapping in allotetraploid cotton, such as the one utilized in this effort of G. hirsutum (line TM-1) by G. barbadense (line 3-79) (Yu et al. 2012). However, in this study we see that even within the small set of 48 experimentally tested intraspecific markers (Table S8), most of the 32 tested G. hirsutum lines can be uniquely identified. Even the three TM-1 samples obtained from different labs had a small number of identified differences. The KASP assays were largely successful for most samples, except for Fibermax 966 and Deltapine 90, which likely had a much lower DNA concentration when measured via nanodrop, so it was not reaching a genotyping end-point consistently with the other samples and thus led to many uncalled genotypes. Call frequencies for individual markers showed greater variability than
sample call frequencies, which is likely due to the need for optimization of PCR conditions for different marker sets. However, when using the Fluidigm system, this is difficult because all marker sets are run under the same conditions. It was found that while SNPs were identified in a relatively small number of samples from U.S. lines, polymorphisms were also transferrable to other germplasm sources, such as the three Australian samples (Delta Opal, Scot 70, and Siokra 1-4) and one Indian line (MCU-5) included on the panel. This indicates that ascertainment bias may not be very large, and that sample lineage is likely more important than the country of germplasm origin.

Although only small differences exist between lines, the PCA and hierarchical clustering analyses were able to distinguish relationships among the samples. In the PCA, the wild *G. hirsutum* line TX0231 was markedly different from the cultivated samples, and the Tamcot Sphinx (a MAR program–derived sample) was the next most unique sample, which was expected due to the assumed diversity of those samples. Within the cultivated lines, when zoomed-in to a very small section of the plot, the samples appear to occupy three different clusters. Cluster 1 contains TM-1 and Fibermax 832, cluster 2 contains Acala Maxxa, Paymaster HS-26, M-240 RNR, and PD-1, and cluster 3 contains Sealand 542, SureGrow 747, Stoneville 474, and Coker 312. In Fang et al. (2013), three of the included samples, SureGrow 747, Stoneville 474, and Fibermax 832, were all identified as being in the same group (group 6). Although we found two of the samples in the same cluster, Fibermax 832 was found in a separate cluster. Fang et al. (2013) also identified Acala Maxxa and Paymaster HS-26 to occur in the same group (group 7), which correlates with our analysis where both are in cluster 2. The PCA results correlate with the hierarchical clustering analysis with the shared samples showing the greatest relationship: Acala Maxxa and Paymaster HS-26 from Fang group 7, and SureGrow 747 and Stoneville 474 from Fang group 6. The TM-1 sample, which was the sample utilized as a reference, was also found to have the largest individual dissimilarity after the MAR sample (Tamcot Sphinx). It may be possible that this information is correct because TM-1 was a line established in the mid 20th century (Kohel et al. 1970); however, it is also likely that due to ascertainment bias from using this sample as the reference, it looks more different than the other cultivated *G. hirsutum* lines.

Development of BES-associated markers is helpful for integration of physical and genetic maps
The abundance of markers and accuracy of localization using SNPs associated with BAC-end sequences will be extremely helpful for integrating an allotetraploid cotton physical map with genetic maps. On finalization of an allotetraploid cotton physical map, BES-associated SNPs can also be utilized to integrate unplaced contigs and singleton to enhance the completion of a quality draft reference sequence. Fine-mapping utilizing BES-associated SNPs with a large population can also be used to correct ordering and localization of contigs. Integrating genetic maps with quantitative trait loci (QTL) mapping will allow for utilization of BAC resources for QTL cloning and fine-scale investigation of important regions in the cotton genome.

CONCLUSIONS
The largest set of intraspecific and interspecific SNPs for cultivated cotton to date have been developed. These SNPs are associated with BACs and will serve as an interface between future physical and genetic maps. They were developed using genomic sequences from multiple lines and species aligned to BAC-end sequences generated by Sanger sequencing, which provided a high-quality reference. Experimental validation was highly successful and indicated that the SNPs will allow for future high-density mapping. Furthermore, additional lines can be resequenced and quickly genotyped for the identified SNP positions as shown here using the GATK software. The developed markers complement currently available genic-based SNPs and simple sequence repeat markers, and provide a largely evenly distributed set of markers for mapping the entire cotton genome with high-density. This will promote more extensive genomic-based studies and breeding of cotton.

ACKNOWLEDGMENTS
The authors acknowledge the Whole Systems Genomics Initiative (WSGI) for providing computational resources and systems administration support for the WSGI HPC cluster, and Kranti Konganti for assistance with utilizing the WSGI resources. The authors thank Fei Wang for maintaining the AgriGenomics Laboratory where KASP assays were performed, Alan Pepper (TAMU) for providing DNA for TX0231, and Iain Wilson’s group (CSIRO) for providing DNA for Australian and Indian lines included in *G. hirsutum* screening. This work was supported by funding from Cotton Incorporated grant no. 03–430 and the National Science Foundation, Plant Genome Research Program (NSF-PGRP) grant no. IOS1025947.

LITERATURE CITED
Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215(3): 403–410.
Ashrafi, H., A. M. Hulse-Kemp, F. Wang, S. S. Yang, X. Guan et al., 2015 A Long-Read Transcriptome Assembly of Cotton (*Gossypium hirsutum* L.) and Intraspecific Single Nucleotide Polymorphism Discovery. The Plant Genome 8: 1–14.
Blenka, A., D. D. Fang, J. F. Rami, O. Garsmeur, F. Luo et al., 2012 A high density consensus genic map of tetraploid cotton that integrates multiple component maps through molecular marker redundancy check. PLoS ONE 7: e45739.
Bohra, A., A. Dubey, R. K. Saxena, R. V. Pennmesta, K. N. Poornima et al., 2011 Analysis of BAC-end sequences (BESs) and development of BES-SSR markers for genetic mapping and hybrid purity assessment in pigeonpea (*Cajanus* spp.). BMC Plant Biol. 11: 56.
Byers, R. L., D. B. Harker, S. M. Yourstone, P. J. Maughan, and J. A. Udall, 2012 Development and mapping of SNP assays in allotetraploid cotton. Theor. Appl. Genet. 124: 1201–1214.
Ching, A., K. S. Caldwell, M. Jung, M. Dolan, O. S. Smith et al., 2002 SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. 3: 19.
Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto et al., 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6: e19379.
Fang, D. D., L. L. Hinze, R. G. Percy, P. Li, D. Deng et al., 2013 A microsatellite-based genome-wide analysis of genetic diversity and linkage disequilibrium in Upland cotton (*Gossypium hirsutum* L.) cultivars from major cotton-growing countries. Euphytica 191: 391–401.
Frelichowski, J. E., M. B. Palmer, D. Main, J. P. Tomkins, R. G. Cantrell et al., 2012 Development and mapping of SNP assays in allotetraploid cotton. The Plant Genome 7(1): 128–430.
Goecks, J., A. Nekrutenko, and J. Taylor, 2010 Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11: R86.
Gordon, D. C. Abajian, and P. Green, 1998 Consed: a graphical tool for computational research in the life sciences. Genome Biol. 10: 288.
Gore, M. A., D. D. Fang, J. A. Poland, J. Zhang, R. G. Percy et al., 2014 Linkage map construction and quantitative trait locus analysis of agronomic and fiber quality traits in cotton, The Plant Genome 7(1) 128–430.
Han, Y., D. Chagne, K. Gasic, E. H. Rikkerink, J. E. Beever et al., 2009 BAC-end sequence-based SNPs and bin mapping for rapid integration of physical and genetic maps in apple. Genetics 93: 282–288.
Hansey, C. N., B. Vaillancourt, R. S. Sekhon, N. De Leon, S. M. Kaepller et al., 2012  Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing. PLoS ONE 7: e33071.

Hendrix, B., and J. M. Stewart, 2005  Estimation of the nuclear DNA content of Gossypium species. Ann. Bot. (Lond.) 95: 789–797.

Hulse-Kemp, A. M., H. Ashrafi, X. Zheng, F. Wang, K. A. Hoegenauer et al., 2014  Development and bin mapping of gene-associated interspecific SNPs for cotton (Gossypium hirsutum L.) introgression breeding efforts. BMC Genomics 15: 945.

Hyten, D. L., S. B. Cannon, Q. J. Song, N. Weeks, E. W. Fickus et al., 2010  High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genomics 11: 38.

Islam, M. S., G. N. Thyssen, J. N. Jenkins, and D. D. Fang, 2015  Detection, validation and application of genotyping-by-sequencing based single nucleotide polymorphisms in upland cotton (Gossypium hirsutum L.), The Plant Genome 8(1).

Van Ooijen, J. W., 2011  Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. Genet. Res. 93: 343–349.

Kohel, R., T. Richmond, and C. Lewis, 1970  Texas marker-1. Description of a genetic standard for Gossypium hirsutum L. Crop Sci. 10: 670–671.

Lam, H. M., X. Xu, X. Liu, W. B. Chen, G. H. Yang et al., 2010  Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. 42: 1053–1059.

Li, F. G., G. Y. Fan, K. B. Wang, F. M. Sun, Y. L. Yuan et al., 2014  Genome sequence of the cultivated cotton Gossypium arboreum. Nat. Genet. 46: 567–572.

Mammadov, J., R. Aggarwal, R. Buyyarakupu, and S. Kumpatla, 2012  SNP markers and their impact on plant breeding. Int. J. Plant Genomics 2012: 728398.

Olitrault, P., J. Terol, A. Garcia-Lor, A. Berard, A. Chauveau et al., 2012  SNP mining in C. clementina BAC end sequences;转让ability in the Citrus genus (Rutaceae), phylogenetic inferences and perspectives for genetic mapping. BMC Genomics 13: 13.

Paton, A. H., J. F. Wendel, H. Gundlach, H. Guo, J. Jenkins et al., 2012  Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492: 423–427.

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

Poland, J. A., P. J. Brown, M. E. Sorrells, and J. L. Jannink, 2012  Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7: e32253.

R Development Core Team, 2010  R: A language and environment for statistical computing, edited by R Development Core Team. R Foundation for Statistical Computing, Vienna.

Rai, K. M., S. K. Singh, A. Bhardwaj, V. Kumar, D. Lakhwani et al., 2013  Large-scale resource development in Gossypium hirsutum L. by 454 sequencing of genic-enriched libraries from six diverse genotypes. Plant Biotechnol. J. 11: 953–963.

Shen, Y. J., H. Jiang, J. P. Jin, Z. B. Zhang, B. Xi et al., 2004  Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiol. 135: 1198–1205.

Taraio-Graovac, M., and N. Chen, 2009  Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics Chapter 4: Unit 4: 10.

Van deynze, A., K. Stoffel, M. Lee, T. A. Wilkins, A. Kozik et al., 2009  Sampling nucleotide diversity in cotton. BMC Plant Biol 9: 125.

Wang, H., R. V. Pennmetsa, M. Yuan, L. Gong, Y. Zhao et al., 2012  Development and characterization of BAC-end sequence derived SSRs, and their incorporation into a new higher density genetic map for cultivated peanut (Arachis hypogaea L.). BMC Plant Biol. 12: 10.

Wendel, J. F., C. Brubaker, I. Alvarez, R. Cronn, and J. M. Stewart, 2009  Evolution and natural history of the cotton genus, pp. 3–22 in Genetics and Genomics of Cotton, edited by A. H. Paterson. Springer, New York.

Yu, J.Z., R.J. Kohel, D.D. Fang, J. Cho, A. Van Deynze et al., 2012  A high-density simple sequence repeat and single nucleotide polymorphism genetic map of the tetraploid cotton genome. G3 (Bethesda) 2: 43–58.

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie et al., 2012  A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28: 3326–3328.

Zhu, Q.H., A. Spriggs, J.M. Taylor, D. Llewellyn, and I. Wilson, 2014  Transcriptome and complexity-reduced, DNA-based identification of intraspecific single-nucleotide polymorphisms in the polyploid Gossypium hirsutum L. G3 (Bethesda) 4:1893–1905.

Communicating editor: E. Akhunov