Published January 9, 2015

Detection, Validation, and Application of Genotyping-by-Sequencing Based Single Nucleotide Polymorphisms in Upland Cotton

M. Sariful Islam, Gregory N. Thyssen, Johnie N. Jenkins, and David D. Fang*

Abstract

The presence of two closely related subgenomes in the allotetraploid Upland cotton, combined with a narrow genetic base of the cultivated varieties, has hindered the identification of polymorphic genetic markers and their use in improving this important crop. Genotyping-by-sequencing (GBS) is a rapid way to identify single nucleotide polymorphism (SNP) markers; however, these SNPs may be specific to the sequenced cotton lines. Our objective was to obtain a large set of polymorphic SNPs with broad applicability to the cultivated cotton germplasm. We selected 11 diverse cultivars and their random-mated recombinant inbred progeny for SNP marker development via GBS. Two different GBS methodologies were used by Data2Bio (D2B) and the Institute for Genome Diversity (IGD) to identify 4441 and 1176 polymorphic SNPs with minor allele frequency of $\geq$0.1, respectively. We further filtered the SNPs and aligned their sequences to the diploid *Gossypium raimondii* reference genome. We were able to use homeologous SNPs to assign 1071 SNP loci to the At subgenome and 1223 to the Dt subgenome. These filtered SNPs were located in genic regions about twice as frequently as expected by chance. We tested 111 of the SNPs in 154 diverse Upland cotton lines, which confirmed the utility of the SNP markers developed in such approach. Not only were the SNPs identified in the 11 cultivars present in the 154 cotton lines, no two cultivars had identical SNP genotypes. We conclude that GBS can be easily used to discover SNPs in Upland cotton, which can be converted to functional genotypic assays for use in breeding and genetic studies.

Published in The Plant Genome 8
doi: 10.3835/plantgenome2014.07.0034
© Crop Science Society of America
5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

COTTON (*Gossypium hirsutum* L.) is a major natural fiber crop, with estimated production of 116.7 million bales in the United States in 2013 (USDA, 2014). In the United States, the estimated return from cotton fibers and seed byproducts is more than five billion dollars annually (Wallace et al., 2008). Upland cotton (*Gossypium hirsutum* L.) represents over 95% of the total cotton fiber produced in the world (Fang et al., 2014). Both *G. hirsutum* and *G. barbadense* L. are allotetraploid ($2n = 4x = 52$) species and originated around 1 to 2 million years ago from interspecific crosses between an A-genome diploid (~1,700 Mb [megabase]) species and a D-genome (~900 Mb) diploid species (Wendel, 1989; Wendel and Cronn, 2003). Narrow genetic diversity within tetraploid cotton, as well as minor sequence divergence between At and Dt subgenomes (Doyle et al., 2008) have hindered identification of polymorphic molecular markers and their use in cotton improvement. Identification of large number of polymorphic molecular markers in Upland cotton will enable a more complete assessment of genetic structure of complex traits and will be valuable to cotton breeding programs.

M.S. Islam, G.N. Thyssen, and D.D. Fang, USDA-ARS-SRRC, Cotton Fiber Bioscience Research Unit, New Orleans, LA 70124; J.N. Jenkins, USDA-ARS, Genetics & Precision Agriculture Research Unit, Mississippi State, MS 39762; and D.D. Fang, USDA-ARS, Crop Genetics Research Unit, Stoneville, MS 38776. Received 30 July 2014. *Corresponding author (david.fang@ars.usda.gov).

Abbreviations: D2B, Data2Bio, LLC, Ames, IA; GBS, genotyping-by-sequencing; GRL, genome reduction level; GR-RSC, genome reduction on restriction site conservation; HR, heterozygosity rate; IGD, Institute for Genome Diversity, Cornell University, Ithaca, NY; JGI, Joint Genome Institute, USDOE, Walnut Creek, CA; KASP, competitive allele-specific PCR; MAF, minor allele frequency; Mb, megabase; NGS, next-generation DNA sequencing; PCR, polymerase chain reaction; RE, restriction enzymes; RIL, recombinant inbred line; SE, single end; SNP, single nucleotide polymorphism.

Next-generation DNA sequencing (NGS) technologies enable researchers to rapidly develop large numbers of SNP markers at a relatively low cost (Maughan et al., 2009). Inexpensive NGS technologies have been successfully used for whole-genome sequencing (Li et al., 2014; Wang et al., 2011; Xu et al., 2011), gene expression analysis (Harper et al., 2012; Naoumkina et al., 2014), and SNP discovery including small and large genome-size organisms, as well as complex polyploidy organisms with narrow genetic differences such as cotton (Byers et al., 2012; Gore et al., 2014) and wheat (*Triticum aestivum* L.; Poland et al., 2012). Several methods and techniques have been developed to discover SNPs and to genotype them in several organisms (Byers et al., 2012; Elshire et al., 2011; Gore et al., 2014; Poland et al., 2012; Wang et al., 2012). Among those methods, one promising, robust, and simple approach is GBS, which facilitates the detection of a wide range of SNPs using many individuals simultaneously. The GBS protocols usually use methylation sensitive restriction enzymes (RE) to produce reduced representation of the genome by targeting the genomic sequence flanking RE sites (Elshire et al., 2011; Poland et al., 2012). Compared with other similar methods, such as reduced representation libraries and restriction site-associated DNA, GBS library construction is more simplified and needs less DNA, avoids random shearing and size selection, and is completed in only two steps on plates followed by polymerase chain reaction (PCR) amplification of the pooled library (Elshire et al., 2011). Since it removes the prerequisite of detection and validation of polymorphism, GBS can be utilized in any polymorphic species and/or any segregating population with any number of individuals (Schnable et al., 2013). So far, a significant number of reports by different groups have detected SNPs in different crop species such as maize (Elshire et al., 2011), wheat, barley (*Hordeum vulgare* L.; Poland et al., 2012), cotton (Gore et al., 2014), rice (*Oryza sativa* L.; Spindel et al., 2013), soybean [*Glycine max* (L.) Merr.; Sonah et al., 2013], and sorghum [*Sorghum bicolor* (L.) Moench; Ma et al., 2012].

So far, in cotton, SNP development has progressed using different approaches (Byers et al., 2012; Gore et al., 2014; Van Deynze et al., 2009). The first extensive work on SNP development reported the characterization and mapping of >1000 SNPs from 270 loci based on EST sequencing using an interspecific population (Van Deynze et al., 2009). The first NGS-based SNP development in cotton employed the genome reduction on restriction site conservation (GR-RSC) method by using two accessions from *G. hirsutum* and two accessions from *G. barbadense* and was reported by Byers et al. (2012). They also used competitive allele-specific PCR (KASP) genotyping chemistry to convert hundreds of SNPs into functional genotyping assays which were mapped on the cotton genome. There is only one report of small-scale GBS in Upland cotton using a biparental population which developed and mapped 412 SNPs (Gore et al., 2014). We discovered thousands of SNP markers based on the GBS methods using a diverse set of Upland cotton cultivars, and validated a set of SNPs by converting them into functional SNP genotyping assays using KASP. So far, no other work has been reported in Upland cotton to develop large numbers of SNPs based on GBS technology.

Table 1. Eleven Upland cotton cultivars that were used for random-mated recombinant inbred population development and single nucleotide polymorphism discovery.[†]

	Cultivar	Place and source of origination
1	Acala Ultima	California Planting Cotton Seed Distributors (Shafter, CA)
2	Tamcot Pyramid	Texas A&M University (College Station, TX)
3	Coker 315	Coker Pedigreed Seed Co. (Hartsville, SC)
4	Stoneville 825	Stoneville Pedigreed Seed Co. (Stoneville, MS)
5	Fibermax 966	Bayer Crop Science (Lubbock, TX)
6	M240	USDA-ARS (Mississippi State, MS)
7	Paymaster HS26	Paymaster Technologies, Inc. (Plainview, TX)
8	Deltapine Acala 90	Delta and Pine Land Co. (Scott, MS)
9	Suregrow 747	Sure-Grow Co. (Centre, AL)
10	Phytogen PSC 355	Phytogen Seeds (Indianapolis, IN)
11	Stoneville 474	Stoneville Pedigreed Seed Co. (Stoneville, MS)

[†] This table was taken from Fang et al. (2014).

In this research, we first used GBS to identify polymorphic SNP markers using 11 diverse Upland cotton lines and random-mated recombinant inbred progeny derived from crosses using the 11 lines as parents. Then we tested a subset of the SNP markers in a panel of 154 Upland cotton lines to validate their polymorphisms. Our objectives were to (i) rapidly develop a large number of SNPs using GBS, (ii) convert those SNPs to functional genotyping assays, and (iii) validate the utility of the SNPs using diverse Upland cotton germplasm from around the world. Since the 11 Upland cotton cultivars represent the diverse pool of the U.S. cultivated cotton, SNPs identified in this research should be valuable to Upland cotton breeding.

Materials and Methods

Plant Materials

A set of 11 diverse Upland cotton lines (10 cultivars and one elite breeding line, Table 1) from major breeding programs across the United States were used as parents to develop a random-mated recombinant inbred population. The details of developing the random-mated recombinant inbred population were previously described by Jenkins et al. (2008) and Fang et al. (2014). The recombinant inbred lines used in the current research are C_5S_6 (five cycles of random-mating and six generations of self-pollination; Tables S1 and S2).

DNA Isolation, Library Preparation, and Sequencing

Recombinant inbred lines (RILs) and 11 parents were grown in a greenhouse in 2013 in New Orleans, LA. Young leaves were collected from each RIL, along with their parents, and stored at −80°C. The genomic DNA was extracted from frozen leaves following the protocol previously described (Islam et al., 2014) with an additional RNAase A digestion step before binding of DNA to the column. The quality and quantity of DNA was measured using a Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA) as well as on a 1.5% agarose gel. DNAs from 37 and 84 randomly selected RILs along with 11 parents (Tables S1 and S2) were sent to D2B, LLC (Ames, IA) and IGD (Cornell University, Ithaca, NY), respectively, for library preparation, sequencing, and subsequent bioinformatics. Data2Bio and IGD used different approaches to construct sequencing libraries. The protocols for library preparation and sequencing followed by D2B and IGD were described in Schnable et al. (2013) and Elshire et al. (2011), respectively. Briefly, D2B digested 48 DNA samples with two RE (*Nsp*I and *BfuC*I), followed by ligation with a single-stranded barcode oligonucleotide in one site. The restriction endonucleases *Nsp*I and *BfuC*I recognize a degenerate 5 bp sequence (RCATG, where R is A or G) and 4 bp sequence (GATC), respectively. The other site was ligated with an oligonucleotide which was complementary to amplification primer. On the other hand, IGD used *ApeK*I, a type II restriction endonuclease that recognizes a degenerate 5 bp sequence (GCWGC, where W is A or T) to digest 95 DNA samples. Ligation between *ApeK*I-cut genomic DNA and adaptor was completed after digestion and 96-plexing of samples was done for sequencing. Two primers were used for amplification. Data2Bio sequenced libraries using two different methods: (i) Genome Reduction Level (GRL) 3 (1 lane 100 bp single end [SE] hi-seq), and (ii) GRL2 (2 lanes 100 bp SE hi-seq). Meanwhile, IGD sequenced following only one method (1 lane 86 bp reads) using a Genome Analyzer 2000 (Illumina, Inc, San Diego, CA).

Processing of Illumina Raw Sequence Data and SNP Calling

In case of D2B, raw reads were trimmed initially using their own pipeline for low quality bases according to error tolerance rate $\leq$3%. The SNP sequences with >50 bases were aligned to consensus reference sequences. Uniquely mapped SNP sequences ($\leq$2 mismatches every 36 bp and <5 bases for every 75 bp as tails) were used for SNP discovery. Polymorphic bases of detected SNPs were supported by at least three reads of the respective SNP. A SNP was called as heterozygous in a given sample if at least two reads supported each of at least two different alleles and each of the two read types separately comprised >20% of the reads aligning to that site, and when the sum of the reads supporting those two alleles comprised at least 90% of all reads covering the site. The putative SNPs were then filtered on the basis of allele number = 2, minor allele frequency (MAF) $\geq$ 10%, heterozygosity rate (HR) $\leq$ 0.1, number of genotypes $\geq$ 2. Finally, good SNPs were detected by filtering again on the basis of missing rate (MR) $\leq$ 20%.

Detailed description of related bioinformatics from raw reads to SNP calling can be found in Glaubitz et al. (2014) and Elshire et al. (2011) for IGD. The SNP sequences provided by IGD were finally filtered in our laboratory. Filtering criteria were MR $\leq$ 20%, allele number = 2, MAF $\geq$ 10%, HR $\leq$ 0.1, number of genotypes $\geq$ 2. Polymorphism and MAF among the parents and RILs were checked separately for each of the filtered SNPs. The SNPs with > 20% MAF difference between average of parents and RILs were discarded. Monomorphic SNPs either among parents and/or RILs were also discarded. The SNP nomenclature was created in our laboratory starting with CFB (cotton fiber bioscience), followed by a serial number.

Alignment and Functional Annotation of SNPs

Alignment of filtered SNP markers from D2B and IGD to the USDOE Joint Genome Institute (JGI, Walnut Creek, CA) *G. raimondii* genome sequence (v.2.1; Paterson et al., 2012) was performed using BLASTN. The sequences of tags flanking SNPs were used as queries. The threshold was set as follows: percentage of length of alignment was $\geq$0.9; mismatches were $\leq$5; and at most, one $\leq$3 bp gap. When a tag gave significant alignments with a minimum bit score of 50, it was taken as having a real hit.

To do functional annotation, SNP loci were aligned to the JGI *G. raimondii* genome sequence (v.2.1) with the GSNAP software program (Paterson et al., 2012; Wu and Nacu, 2010). These loci were assigned to the At or Dt subgenome using the PolyCat software program (Page et al., 2013). The number of SNP loci per 5 Mb was calculated to generate a histogram for the reference genome, and the At and Dt subgenome assigned loci separately.

Validation of SNPs

A set of 111 SNPs evenly distributed along the whole *G. raimondii* genome were selected for validation using 154 diverse Upland cotton varieties from around the world (Table S3). Of the 111 SNPs, 75 and 36 were generated from D2B and IGD data, respectively. First, the selected 111 SNPs were converted to functional genotyping SNP assays using KASP technology (LGC Genomics, Beverly, MA). For each SNP, two allele-specific forward primers and one common reverse primer were designed. By using these primers, KASP assays were performed in a final reaction volume of 5.00 µL containing 1× KASP reaction mix (LGC Genomics), 0.07 µL of assay mix (12 µM each allele-specific forward primer and 30 µM reverse primer), and 20–25 ng of genomic DNA. The Bio-Rad CFX96 RT-PCR Thermal cycler (Bio-Rad Corporation, Hercules, CA) was used for the following cycling conditions: 15 min at 94°C; 10 touchdown cycles of 20 s at 94°C, and 60 s at 61 to 55°C (the annealing temperature for each cycle being reduced by 0.6°C per cycle); and 26 to 35 cycles of 20 s at 94°C and 60 s at 55°C. Fluorescence detection of the reactions was performed using Bio-Rad CFX96 RT-PCR, and the data were analyzed using the CFX96 manager software.

Table 2. Summary statistics of sequence reads and single nucleotide polymorphism (SNP) information produced by genotyping by sequencing.

Item	Data2Bio[†]	IGD[‡]
Total raw reads per lane (million)	379.80	235.32
Total trimmed reads per lane (million)	375.99	210.63
Average raw bp per sample per lane (million)	687.92	161.94
Total trimmed bp per sample per lane (million)	662.77	144.95
Average reads per SNP site/sample	51.20	7.90
Total tags after filter	123,942	73,632
Average length (base)	85	64
Total unfiltered SNPs	10,512	32,644
Total filtered SNPs	4,663	19,925
Total used SNPs	4,441	1,176
Total aligned SNPs	3,156	807

[†] Data2Bio results are combined from GRL2 and GRL3. Data2Bio, LLC, Ames, IA.

[‡] IGD, Institute for Genome Diversity, Cornell University, Ithaca, NY.

Results

Sequencing and Reads

Raw reads produced by D2B and IGD were quite different, since they used different methods during both library preparation and sequencing. A summary of sequence data generated from both D2B and IGD is included in Table 2. From a total of 379.8 million raw reads produced in D2B, approximately 3.8 million reads were discarded due to presence of low quality of sequences. An average of 662.77 million trimmed base pairs per sample were read in D2B. Results also revealed that D2B produced greater read depth per sequence site (on average, 51.2 reads). Among the 11 parents, the number of trimmed reads ranged from 2.43 ('Suregrow 747') to 8.83 ('Fibermax 966') million reads created in D2B, while average of RILs trimmed reads were 3.59 million (Table 3).

Approximately 24.6 million low quality reads were trimmed from 235.32 million total raw reads created in IGD. An average of 144.95 million trimmed base pairs per sample were read in IGD. On average, 7.9 reads per sequence site were found in IGD. With IGD data, the number of good reads ranged from 1.88 ('Acala Ultima') up to 2.85 (Fibermax 966) million reads among the 11 parents, and average of RILs was 2.25 million reads (Table 3).

SNP Discovery and Individual Genotype

A total of 123,942 filtered contig sequences were produced by D2B, while 73,632 contigs were produced by IGD (Table 2). Average length of contig sequences were 88 and 64 from data produced in D2B and IGD, respectively. A higher number of unfiltered SNPs were detected from data generated by IGD (32,644) compared with data generated by D2B (10,512). Similarly, the unfiltered to filtered SNPs ratio was also higher in IGD data (0.61) than D2B data (0.44). Surprisingly, a huge number of filtered SNPs were discarded from IGD data since they did not fulfill the criteria ($MR \leq 20\%$, allele number = 2, $MAF \leq 10\%$, $HR \leq 0.1$, number of genotypes ≥ 2) set up in our lab. All filtered SNPs were submitted to dbSNP (accessions numbers 1,387,933,573 to 1,387,939,389) and also listed in Tables S4 and S5. Finally, 4441 and 1176 good SNPs from D2B and IGD, respectively, were used for further analysis.

Percentages of genotypic categories of filtered SNPs created during GBS of 11 parents and RIL average were comparable in both D2B and IGD data except the homozygote alternate allele and missing categories for Acala Ultima (Table 3). The ranges of percentages of genotypic categories were 52.1 to 69.9 and 46.1 to 62% for homozygote major allele; 26.9 to 36.1 and 6.1 to 34.1% for homozygote minor allele; 0.2 to 13.2 and 1.0 to 19.8% for heterozygote; 1.0 to 7.3 and 3.5 to 28.1% for missing values in data generated from D2B and IGD, respectively. Comparing among cotton lines, Acala Ultima showed the highest rate of heterozygosity in both data from D2B (13.2%) and IGD (19.8%). Minor allele frequency is one of the effective indicators for prediction of the success rate of a SNP marker. In this study, the distribution of usable SNPs from data generated in D2B and IGD is illustrated in Fig. 1. The SNPs obtained from IGD data were evenly distributed across the MAF 0.10 to 0.50, while SNPs acquired from data D2B mostly clustered between MAF 0.12 to 0.22.

Alignment and Distribution of SNPs in the Cotton Genome

In total, 5617 (4441 and 1176 from D2B and IGD, respectively) high quality polymorphic SNPs contig sequences were aligned to the *G. raimondii* reference genome sequences. A total of 3963 (3156 and 807 from D2B and IGD, respectively) SNPs produced a significant hit, and could be aligned to the reference genome (Table S6). The remaining 1654 markers are most likely located in unique regions of the At subgenome that are not homeologous with the reference D5 genome. All aligned SNPs loci were assigned to the At or Dt subgenome and histograms were generated by counting SNPs in each 5 Mb interval (Fig. 2). Out of 3963 aligned SNPs, 1071 and 1223 were assigned to At and Dt subgenome, respectively, using software program PolyCat (Page et al., 2013). Histograms revealed that SNP loci were evenly distributed in the cotton reference genome, except for a few large clusters on chromosomes 1, 8, and 9. The highest number of SNPs loci were mapped on chromosome 9 between 40 and 50 Mb, followed by on chromosome 8 between 40 and 50 Mb. At subgenome assigned SNP loci were more or less evenly distributed to the whole genome, while Dt subgenome assigned SNPs produced two large clusters on chromosomes 1 and 9. To understand why SNPs on chromosome 9 between 40 and 50 Mb showed more diversity, we separated those SNPs and observed the genotype (Table S7). Cultivar FM966 accounts for most of the diversity in the interval, and 'Tamcot Pyramid' showed some heterozygosity.

To investigate the structural, functional, and evolutionary impact of the filtered SNPs, we analyzed the annotations of the GBS SNP loci (Fig. 3). We found

Table 3. Summary of individual sequence and genotype information of filtered single nucleotide polymorphisms (SNPs) of 11 parents and their recombinant inbred line (RIL) average.

	Data2Bio[‡]						IGD[§]					
	Genotype						Genotype					
Sample[†]	Homozygote (major allele)	Homozygote (minor allele)	Heterozygote	Missing	Reads (million)	bp (million)	Homozygote (major allele)	Homozygote (minor allele)	Heterozygote	Missing	Reads (million)	bp (million)
	%						%					
AU	52.1	29.0	13.2	5.7	7.64	651.25	46.1	6.1	19.8	28.1	1.88	120.23
TP	57.7	36.1	0.9	5.2	4.50	386.00	59.5	34.1	1.3	5.0	2.30	147.23
Coker315	56.5	26.9	11.4	5.2	3.43	288.37	58.7	25.9	8.1	7.4	2.27	145.38
ST825	64.4	31.1	1.1	3.4	4.63	392.35	64.5	27.6	1.7	6.2	2.19	140.32
FM966	60.7	33.6	3.4	2.4	8.83	760.89	61.3	32.8	2.4	3.5	2.85	182.20
M240	62.4	35.2	0.2	2.2	3.78	324.11	63.7	28.2	1.5	6.6	2.41	154.47
HS26	64.4	32.8	0.3	2.5	6.01	505.51	54.8	38.7	1.3	5.2	2.49	159.34
DP90	57.6	30.9	5.4	6.2	3.36	279.10	60.2	27.7	7.7	4.4	2.50	160.18
SG747	63.6	31.0	0.5	4.9	2.43	205.23	66.6	27.6	1.0	4.9	2.36	151.06
PSC355	59.6	29.5	9.7	1.2	5.23	443.21	57.3	27.7	6.7	8.2	2.20	141.08
ST474	69.9	28.6	0.4	1.0	5.37	447.14	62.0	27.6	3.9	6.6	2.70	172.86
RIL avg.	58.8	32.4	1.6	7.3	3.59	303.33	58.5	29.0	5.9	6.6	2.25	143.93

[†] Cultivar name abbreviations: AU, Acala Ultima; TP, Tamcot Pyramid; ST, Stoneville; FM, Fibermax; DP, Deltapine; HS, Paymaster HS; PSC, Phytogen; SG, Suregrow.

[‡] Data2Bio results are combined from GRL2 and GRL3. Data2Bio, LLC, Ames, IA.

[§] IGD, Institute for Genome Diversity, Cornell University, Ithaca, NY.

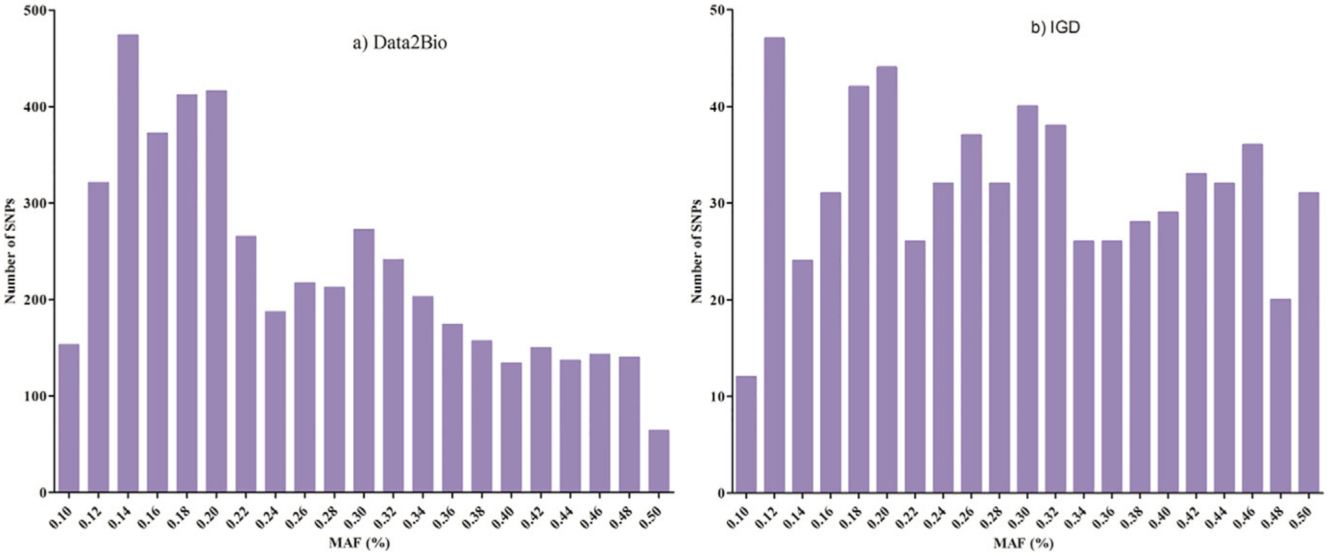

Figure 1. Distribution of filtered single nucleotide polymorphisms (SNPs) according to their minor allele frequency (MAF). (a) Data2Bio, Ames, IA. (b) Institute for Genome Diversity (IGD), Cornell University.

that 11.56% of these GBS SNPs resided in exons, while 9.94% were in introns. However, the published D genome is comprised of 5.95% exons and 6.88% introns (Paterson et al., 2012).

Validation and Utility of GBS SNPs

One-hundred-fifty-four diverse Upland cotton varieties, originating from 25 countries (Fang et al., 2013), were chosen for validation of a set of selected GBS SNPs (Table S3). Out of 111 selected GBS SNPs, 75 originated from D2B and 36 were from IGD data (Table 4). These selected SNPs reside on all 13 reference chromosomes and are evenly distributed along the whole *G. raimondii* genome (Table S8). Of the 75 SNPs from D2B data, 56 (74.7%) were amplified using KASP genotyping assay, while 18 (50.0%) out of 36 SNPs from IGD data were successfully converted into a KASP assay. Of the total 74 amplified SNPs, 58 were codominant, while remaining 16 were dominant. Sixty nine (93.2%) SNPs gave polymorphic results. Codominant polymorphic SNP assays produced three distinguishable clusters (homozygotes for 2 alleles and heterozygote), while dominant but polymorphic

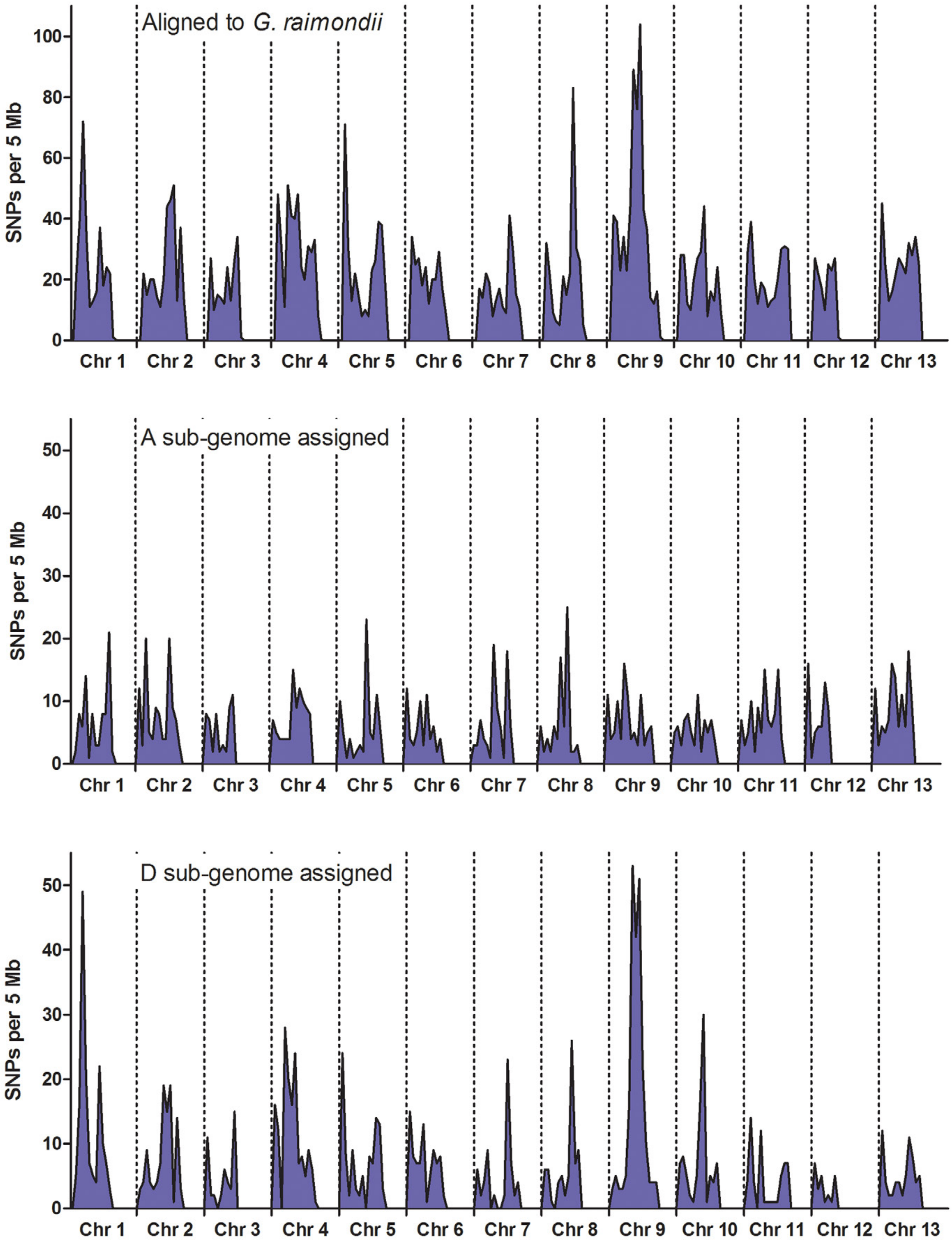

Figure 2. Histogram of genotyping by sequencing single nucleotide polymorphism (SNP) markers in 5 megabase (Mb) intervals of the *G. raimondii* reference genome.

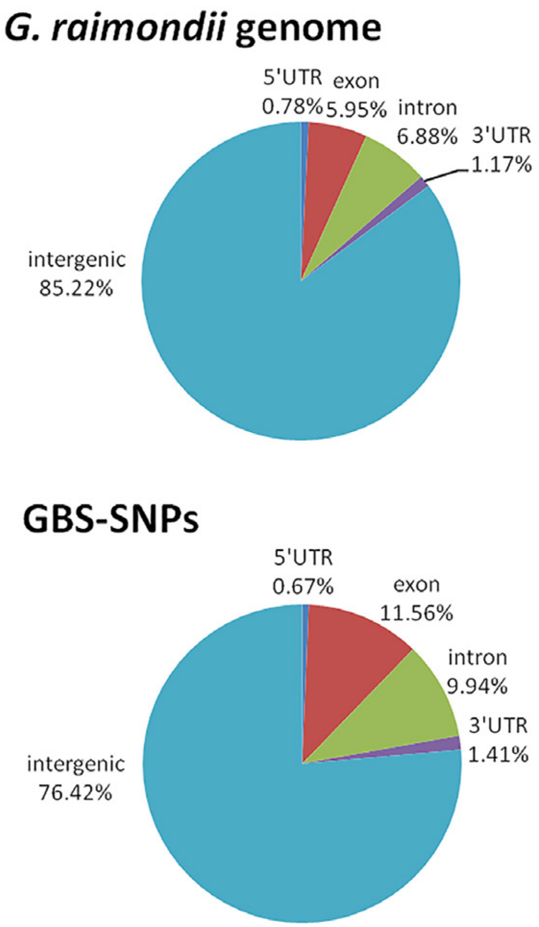

Figure 3. Comparison of the annotation of genotyping by sequencing (GBS) single nucleotide polymorphism (SNP) loci with the overall annotation of the *G. raimondii* reference genome. UTR, untranslated region.

markers gave two different clusters. Monomorphic SNP assays showed only one homozygote allele cluster (Fig. 4).

To investigate the potential utility of SNP assays in breeding, several observations could be made from the genotype pattern of 58 polymorphic and codominant KASP SNP assays within the 154 cotton lines. Of the 154 Upland cotton lines, no two individuals shared the same genotype across all SNP assays. Of the 11 parents initially used for GBS library construction, each had a unique genotype. Results also revealed that an average HR of 10.8% was observed across all SNP assays, with the highest HR of any assay being 34.8% (Table 4). Of the 58 assays tested, 52 (89.7%) had an MAF > 10% and 33 (57.0%) had an MAF > 20%, with an average MAF = 25.2%.

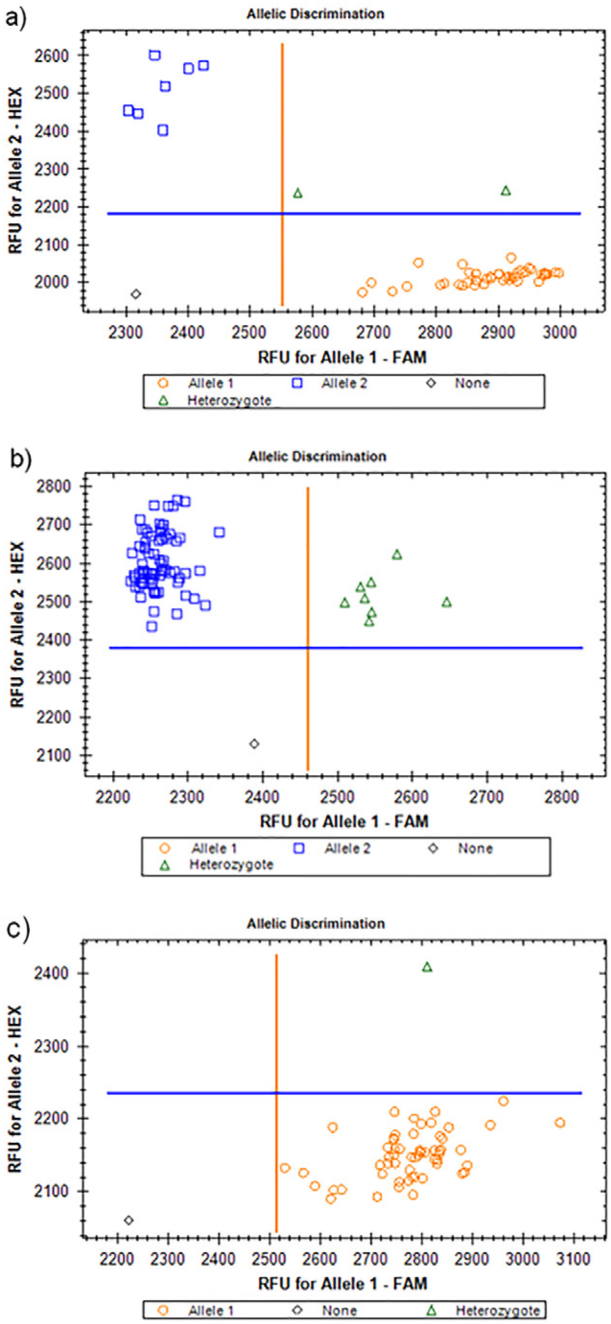

Figure 4. Sample genotyping plots of a diverse set of Upland cotton cultivars generated from Bio-Rad CFX96 manager software during competitive allele specific polymerase chain reaction (KASP) assay genotyping of selected genotyping by sequencing (GBS) single nucleotide polymorphisms (SNPs). (a) Codominant polymorphic SNP showed clear separation of two alleles; (b) SNP showed dominant reaction; (c) SNP was monomorphic.

Discussion

Our objective was to develop a large set of polymorphic markers with wide applicability to cultivated Upland cotton varieties. By sequencing 11 diverse cultivars and their random-mated progeny, we were able to identify 5617 SNPs that met our criteria for quality and minor allele frequency. We converted 111 of these SNPs into KASP assays, which we tested on 154 cultivars from 25 countries. Despite the narrow genetic base of cultivated cotton, no two cultivars showed identical genotypes even in this small subset of SNPs. The large-scale set of SNPs will no doubt further differentiate the haplotypes that exist in the elite cotton germplasm and be useful to genetic mapping in cotton.

Table 4. Competitive allele-specific polymerase chain reaction (KASP) assay genotyping information of selected single nucleotide polymorphisms (SNPs) using 154 diverse Upland cotton germplasm.

Item	Data2Bio[†]	IGD[‡]	Both
Total SNPs assayed	75	36	111
Functional SNPs	56	18	74
Functional SNPs percentage	74.7	50.0	66.7
Codominant	45.0	13.0	58.0
Dominant	11.0	5.0	16.0
Polymorphic	52.0	17.0	69.0
Percentage of polymorphic	92.9	94.4	93.2
Average heterozygosity, %	11.5	8.2	10.8
Average minor allele frequency, %	24.7	26.7	25.2
Average missing data, %	2.1	1.5	2.0

[†] Data2Bio, LLC, Ames, IA.

[‡] IGD, Institute for Genome Diversity, Cornell University, Ithaca, NY.

In this study, D2B generated a larger number of reads with higher individual site depth than IGD. This may be due to several reasons. One of the notable reasons was that D2B used altogether three lanes 100 bp SE Hi-seq during sequencing, while IGD used only one lane of 84 bp reads. The average individual base pairs generated from raw sequences of D2B and IGD were 831.1 and 151.5 Mb, respectively. Although the number of unfiltered SNPs was higher from IGD than D2B, the number of final usable SNPs discovered from D2B data was higher than that from IGD data. This may be due to the modified GBS protocol used by D2B which incorporates two RE. This enabled D2B to produce more uniform GBS libraries with different overhangs at each end of the digested fragments. They also used primers amplified on non barcode site to further reduce the number of targeted sites for analysis (Schnable et al., 2013).

Minor allele frequency is one of the key indicators to detect SNPs with real polymorphism and influence the success of SNPs in the subsequent genotypic assay, since SNPs with low MAF based on the NGS detection, were less likely to be polymorphic than SNPs with higher MAF. MAF also affects the application of SNPs as molecular markers by influencing the type of information provided by the markers in different populations. During mapping studies, it is desirable to maximimize the number of polymorphic markers by selecting SNPs with moderate MAF. It has been shown by simulation that map-independent imputation is significantly more accurate for markers with MAF > 0.1 (Rutkoski et al., 2013). Hence, we discarded SNPs having MAF < 10% among the 11 upland cotton cultivars and their random mated progenies. The SNPs having MAFs within clusters evenly spaced across the cotton genome will help to get available haplotypes with high-MAF SNPs (i.e., MAF $\geq$ 0.1), improving detection of heterozygosity for any assayed genotype sample, while lower-MAF SNPs (MAF $\leq$ 0.1) tend to come from a specific origin and improve detection of uncommon haplotypes.

Different groups have detected SNPs in crop species, including wheat, barley (Poland et al., 2012), cotton (Gore et al., 2014), rice (Spindel et al., 2013), soybean (Sonah et al., 2013), potato (*Solanum tuberosum* L.; Uitdewilligen et al., 2013), and sorghum (Ma et al., 2012). However, unlike the present study, most of these were conducted using only two parents and a mapping population to generate genetic maps (Gore et al., 2014; Ma et al., 2012; Poland et al., 2012; Spindel et al., 2013). A total of 10,120 and 129,156 SNPs were detected using a set of eight diverse varieties of soybean (Sonah et al., 2013) and 83 tetraploid potato varieties (Uitdewilligen et al., 2013), respectively. It is difficult to use these data to compare the underlying diversity of those crops with cotton, since our filtering criteria were not the same. Our 5617 high quality SNPs are significantly higher than the 412 GBS-based SNPs from cotton which were already used, along with 429 simple sequence repeat markers, to construct a linkage map and subsequent QTL analysis (Gore et al., 2014).

Some GBS approaches use methylation-sensitive RE, or RE which are predicted to cut less frequently in intergenic regions because of sequence biases. The IGD used *ApeK*I which is methylation sensitive, but D2B used both methylation sensitive *BfuC*I and methylation insensitive *Nsp*I (Elshire et al., 2011; Schnable et al., 2013). We observed that our SNPs were about twice as likely to be located in exons or introns as would be expected by chance. This bias increases the likelihood that the markers will divide gene-rich regions, and that the marker set may include actual causative mutations. Despite this bias, the markers we identified were distributed across all 26 *G. hirsutum* chromosomes. Using the reference *G. raimondii* genome and the PolyCat software to assign loci to subgenomes, we were able to assign physical map locations to 1071 SNP loci in the At subgenome and 1223 in the Dt subgenome. Therefore, on average, we expect one marker per 1.6 Mb in the 1.7 gigabase At subgenome, and one marker per 736 kilobase in the 900 Mb Dt subgenome (Wendel, 1989; Wendel and Cronn, 2003). Previous studies also demonstrated that SNPs identified based on various NGS technology, including GBS, were evenly distributed, such as in cotton, using the GR-RSC approach (Byers et al., 2012); in soybean, by using GBS method (Sonah et al., 2013). Well distributed markers in genes are especially useful for breeders and genetics researchers.

Validation of detected SNPs is essential to establish the utility of these predicted polymorphisms for practical plant breeding applications. A number of SNP genotyping platforms, such as GoldenGate, Infinium, Taqman, and KASP assays, are available to convert SNPs to functional genotypic assays. However, in this study, because of its cost-effective and flexible nature, the KASP assay was designed for 111 SNPs spaced across the cotton genome. Few reports are available on the development of KASP assays in crop plants. For example, Allen et al. (2011) developed 1114 KASP SNP assays for validation on 23 wheat cultivars and also incorporated

SNP markers into the genetic map of wheat. In the case of common bean (*Phaseolus vulgaris* L.), KASP assays have been developed for 94 SNPs and used for analyzing genetic diversity in 70 accessions (Cortes et al., 2011). Hiremath et al. (2012) developed KASP assays for 2005 SNPs in chickpea, and used these for genetic diversity analysis and genetic mapping in chickpea and comparative mapping in legumes. In cotton, Byers et al. (2012) developed KASP assays for 1052 SNPs that were used for genetic mapping in cotton and validated in 48 cotton varieties. Thus, these four studies highlight the significance of KASP assays for SNP genotyping on a large scale for genetics and breeding applications. In the present study, though 111 SNPs were attempted for conversion into KASP assays, only 74 (66.7%) markers could be successfully converted. The failure of the remaining SNP markers (33.3%) to be validated is likely due to the presence of duplicate or paralogous loci, incorrect primer design near the SNP, identification of fake SNPs initially, and/or the need to optimize PCR conditions. This conversion rate is higher than that of the other KASP study on cotton (35.8%; Byers et al., 2012). This rate of conversion from selected SNPs to functional KASP assays could probably be increased further with optimization of primer design and amplification conditions. Among the 74 successfully converted markers, five showed monomorphic reactions, although they were polymorphic among the 11 parents used in GBS genotyping. This may be associated with false SNP calling during GBS data analysis. To determine further usefulness of the GBS SNPs in practical breeding, we analyze the genotypic data of polymorphic KASP assays among 154 Upland cotton cultivars. The HR, MAF, and genotypic categories once again highlighted the diversity among the tested cotton cultivars.

Although GBS allows simultaneous SNP discovery and scoring, each GBS protocol will only identify and score SNPs that are adjacent to the RE sites used. We used multiple GBS protocols with different RE to collect our set of SNPs, and converted some of these to KASP assays. We anticipate that different sets of KASP markers will be used sequentially to map QTLs and genes in cotton varieties. Knowledge of the physical locations of the SNP loci makes this approach feasible and economical, even when compared with GBS. Since a small set of markers can assign a trait to a chromosome arm, only a small set of markers need be interrogated at first. As the genetic interval is narrowed, more markers may be needed than sites of a given RE exist on that interval. Since our set of diverse parents was proven to capture much of the diversity in Upland cotton cultivars, we may increase our set of markers by additional rounds of GBS with different RE on these 11 lines. Fine mapping of a trait would then proceed by KASP assays derived from SNPs that are located within the genetic interval. This approach would allow finer mapping than a single round of GBS, with the expense of only a few dozen KASP assays.

In conclusion, we have demonstrated the utility of GBS on carefully selected cultivars for the large-scale development of SNP markers in Upland cotton, an important crop with limited diversity. The markers reported here can be used in assays independently of the GBS technique and will be useful to breeders and cotton researchers.

Supplemental Information Available

Supplemental information is included with this article.

Table S1. Names of RILs and their crosses that were sent to Data2Bio for GBS library preparation and sequencing.

Table S2. Names of RILs and their crosses that were sent to IGD, Cornell University for library preparation and sequencing.

Table S3. The names, PI numbers, countries of original collection, and pedigree of 154 Upland cotton cultivars used in this study.

Table S4. List of filtered SNPs generated in Data2Bio.

Table S5. List of good SNPs used in this research originated from IGD, Cornell University

Table S6. List of SNPs aligned to *G. raimondaii* genome and their contig sequences.

Table S7. Allele produced by SNP loci mapped on chromosome 9 between 40 and 50 megabase (Mb).

Table S8. SNPs selected for KASP genotyping of 154 diverse Upland cotton lines.

Acknowledgments

This research was funded by United States Department of Agriculture-Agricultural Research Service CRIS projects 6435-21000-017-00D, 6406-21000-12-00D, and Cotton Incorporated Project 10-747. We thank Mrs. Ping Li for her help during validation. Our appreciation goes to Data2Bio, LLC, Ames, IA, and Institute for Genome Diversity, Cornell University, for their excellent support in library construction, sequencing, and bioinformatics. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U. S. Department of Agriculture, which is an equal opportunity provider and employer.

References

Allen, A.M., G.L. Barker, S.T. Berry, J.A. Coghill, R. Gwilliam, S. Kirby, et al. 2011. Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). Plant Biotechnol. J. 9:1086–1099. doi:10.1111/j.1467-7652.2011.00628.x

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. doi:10.1007/s00122-011-1780-8

Cortes, A.J., M.C. Chavarro, and M.W. Blair. 2011. SNP marker diversity in common bean (*Phaseolus vulgaris* L.). Theor. Appl. Genet. 123:827–845. doi:10.1007/s00122-011-1630-8

Doyle, J.J., L.E. Flagel, A.H. Paterson, R.A. Rapp, D.E. Soltis, P.S. Soltis, and J.F. Wendel. 2008. Evolutionary genetics of genome merger and doubling in plants. Annu. Rev. Genet. 42:443–461. doi:10.1146/annurev.genet.42.110807.091524

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379. doi:10.1371/journal.pone.0019379

Fang, D.D., J.N. Jenkins, D.D. Deng, J.C. McCarty, P. Li, and J. Wu. 2014. Quantitative trait loci analysis of fiber quality traits using a random-mated recombinant inbred population in Upland cotton (*Gossypium hirsutum* L.). BMC Genomics 15:397. doi:10.1186/1471-2164-15-397

Fang, D.D., L.L. Hinze, R.G. Percy, P. Li, D.D. Deng, and G. Thyssen. 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. doi:10.1007/s10681-013-0886-2

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, and E.S. Buckler. 2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS ONE 9:e90346. doi:10.1371/journal.pone.0090346

Gore, M.A., D.D. Fang, J.A. Poland, J. Zhang, R.G. Percy, R.G. Cantrell, G. Thyssen, and A.E. Lipka. 2014. Linkage map construction and quantitative trait locus analysis of agronomic and fiber quality traits in cotton. The Plant Genome 7 (1). doi:10.3835/plantgenome2013.07.0023

Harper, A.L., M. Trick, J. Higgins, F. Fraser, L. Clissold, R. Wells, et al. 2012. Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. Nat. Biotechnol. 30:798–802. doi:10.1038/nbt.2302

Hiremath, P.J., A. Kumar, R.V. Penmetsa, A. Farmer, J.A. Schlueter, S.K. Chamarthi, et al. 2012. Large-scale development of cost-effective SNP marker assays for diversity assessment and genetic mapping in chickpea and comparative mapping in legumes. Plant Biotechnol. J. 10:716–732. doi:10.1111/j.1467-7652.2012.00710.x

Islam, M.S., L. Zeng, C.D. Delhom, X. Song, H.J. Kim, P. Lim, and D.D. Fang. 2014. Identification of cotton fiber quality quantitative trait loci using intraspecific crosses derived from two near-isogenic lines differing in fiber bundle strength. Mol. Breed. 34:373–384. doi:10.1007/s11032-014-0040-4

Jenkins, J.N., J.C. McCarty, O.A. Gutierrez, R.W. Hayes, D.T. Bowman, C.E. Watson, and D.C. Jones. 2008. Registration of RMUP-C5, a random mated population of Upland cotton germplasm. J. Plant Reg. 2:239. doi:10.3198/jpr2008.02.0080crg

Li, F., G. Fan, K. Wang, F. Sun, Y. Yuan, G. Song, et al. 2014. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat. Genet. 46:567–572. doi:10.1038/ng.2987

Ma, X.F., E. Jensen, N. Alexandrov, M. Troukhan, L. Zhang, S. Thomas-Jones, et al. 2012. High resolution genetic mapping by genome sequencing reveals genome duplication and tetraploid genetic structure of the diploid *Miscanthus sinensis*. PLoS ONE 7:e33821. doi:10.1371/journal.pone.0033821

Maughan, P.J., S.M. Yourstone, E.N. Jellen, and J.A. Udall. 2009. SNP discovery via genomic reduction, barcoding, and 454-pyrosequencing in amaranth. Plant Genome 2:260. 10.3835/plantgenome2009.08.0022. doi:10.3835/plantgenome2009.08.0022

Naoumkina, M., G. Thyssen, D.D. Fang, D.J. Hinchliffe, C. Florane, K.M. Yeater, J.T. Page, and J.A. Udall. 2014. The Li2 mutation results in reduced subgenome expression bias in elongating fibers of allotetraploid cotton (*Gossypium hirsutum* L.). PLoS ONE 9:e90830. doi:10.1371/journal.pone.0090830

Page, J.T., A.R. Gingle, and J.A. Udall. 2013. PolyCat: A resource for genome categorization of sequencing reads from allopolyploid organisms. G3: Genes Genomes Genet. 3:517–525. doi:10.1534/g3.112.005298.

Paterson, A.H., J.F. Wendel, H. Gundlach, H. Guo, J. Jenkins, D. Jin, et al. 2012. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492:423–427. doi:10.1038/nature11798

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. doi:10.1371/journal.pone.0032253

Rutkoski, J.E., J. Poland, J.-L. Jannink, and M.E. Sorrells. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3: Genes Genomes Genet. 3:427–439.

Schnable, P.S., S. Liu, and W. Wu. 2013. Genotyping by next-generation sequencing. U.S. Patent Appl. no. 13/739,874.

Sonah, H., M. Bastien, E. Iquira, A. Tardivel, G. Legare, B. Boyle, et al. 2013. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS ONE 8:e54603. doi:10.1371/journal.pone.0054603

Spindel, J., M. Wright, C. Chen, J. Cobb, J. Gage, S. Harrington, et al. 2013. Bridging the genotyping gap: Using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theor. Appl. Genet. 126:2699–2716. doi:10.1007/s00122-013-2166-x

Uitdewilligen, J.G., A.M. Wolters, B.B. D'Hoop, T.J. Borm, R.G. Visser, and H.J. van Eck. 2013. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS ONE 8:e62355. doi:10.1371/journal.pone.0062355

USDA. 2014. Cotton and wool yearbook: Dataset. http://usda.mannlib.cornell.edu/usda/ers/CWS//2010s/2014/CWS-03-12-2014.pdf (accessed 25 July 2014).

Van Deynze, A., K. Stoffel, M. Lee, T.A. Wilkins, A. Kozik, R.G. Cantrell, et al. 2009. Sampling nucleotide diversity in cotton. BMC Plant Biol. 9:125. doi:10.1186/1471-2229-9-125

Wallace, T.P., D. Bowman, B.T. Campbell, P. Chee, O.A. Gutierrez, R.J. Kohel, et al. 2008. Status of the USA cotton germplasm collection and crop vulnerability. Genet. Resour. Crop Evol. 56:507–532. doi:10.1007/s10722-008-9382-2

Wang, S., E. Meyer, J.K. McKay, and M.V. Matz. 2012. 2b-RAD: A simple and flexible method for genome-wide genotyping. Nat. Methods 9:808–810. doi:10.1038/nmeth.2023

Wang, X., H. Wang, J. Wang, R. Sun, J. Wu, S. Liu, et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. Nat. Genet. 43:1035–1039. doi:10.1038/ng.919

Wendel, J.F. 1989. New world tetraploid cottons contain old world cytoplasm. Proc. Natl. Acad. Sci. USA 86:4132–4136. doi:10.1073/pnas.86.11.4132

Wendel, J.F., and R.C. Cronn. 2003. Polyploidy and the evolutionary history of cotton. Adv. Agron. 78:139–186. doi:10.1016/S0065-2113(02)78004-8

Wu, T.D., and S. Nacu. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26:873–881. doi:10.1093/bioinformatics/btq057

Xu, X., S. Pan, S. Cheng, B. Zhang, D. Mu, P. Ni, et al. 2011. Genome sequence and analysis of the tuber crop potato. Nature 475:189–195. doi:10.1038/nature10158