SNP Discovery in Complex Allotetraploid Genomes (Gossypium spp., Malvaceae) Using Genotyping by Sequencing

Authors: Logan-Young, Carla Jo, Yu, John Z., Verma, Surender K., Percy, Richard G., and Pepper, Alan E.

Source: Applications in Plant Sciences, 3(3)

Published By: Botanical Society of America

URL: https://doi.org/10.3732/apps.1400077

Applications in Plant Sciences

Applications in Plant Sciences 2015 3(3): 1400077

Genomic Resources Note

SNP discovery in complex allotetraploid genomes (*Gossypium* spp., Malvaceae) using genotyping by sequencing[1]

Carla Jo Logan-Young[2], John Z. Yu[3], Surender K. Verma[2], Richard G. Percy[3], and Alan E. Pepper[2,4]

[2]Department of Biology, Texas A&M University, College Station, Texas 77843 USA; and [3]USDA-ARS, Southern Plains Agricultural Research Center, 2881 F&B Road, College Station, Texas 77845 USA

- *Premise of the study:* Single-nucleotide polymorphism (SNP) marker discovery in plants with complex allotetraploid genomes is often confounded by the presence of homeologous loci (along with paralogous and orthologous loci). Here we present a strategy to filter for SNPs representing orthologous loci.
- *Methods and Results:* Using Illumina next-generation sequencing, 54 million reads were collected from restriction enzyme–digested DNA libraries of a diversity of *Gossypium* taxa. Loci with one to three SNPs were discovered using the Stacks software package, yielding 25,529 new cotton SNP combinations, including those that are polymorphic at both interspecific and intraspecific levels. Frequencies of predicted dual-homozygous (aa/bb) marker polymorphisms ranged from 6.7–11.6% of total shared fragments in intraspecific comparisons and from 15.0–16.4% in interspecific comparisons.
- *Conclusions:* This resource provides dual-homozygous (aa/bb) marker polymorphisms. Both in silico and experimental validation efforts demonstrated that these markers are enriched for single orthologous loci that are homozygous for alternative alleles.

Key words: genotyping by sequencing; *Gossypium*; interspecific; intraspecific; next-generation sequencing; polyploid; single-nucleotide polymorphisms.

Cottons (*Gossypium* L. spp.) provide the leading natural fiber for textiles, as well as an important seed product for feed, food, and oil (Campbell et al., 2010). The most widely grown species are the allotetraploids *G. hirsutum* L. (upland cotton) and *G. barbadense* L. (Sea Island cotton). These species are both descended from an allopolyploidization event involving an A-genome diploid species, related to modern *G. herbaceum* L. and *G. arboreum* L., and D-genome diploid species, related to modern *G. raimondii* Ulbr. (Percival and Kohel, 1990). Recent developments in next-generation sequencing (NGS) technology have lowered the cost of sequencing per base, and enabled the genotyping by sequencing (GBS) approach for developing informative single-nucleotide polymorphism (SNP) markers in species with large, complex genomes (Elshire et al., 2011), including species without a reference genome sequence (Glaubitz et al., 2014). In this study, we employed a simple and cost-effective GBS approach to identify intraspecific and interspecific SNPs within and between allotetraploid cottons *G. hirsutum* and *G. barbadense*.

[1]Manuscript received 7 August 2014; revision accepted 3 February 2015.

We would like to express our deepest gratitude to James Frelichowski and Jared Harris for their help with cotton germplasm resources. This work was funded by Cotton Incorporated Fellowship 08-380 (to C.J.L.Y.) and specific cooperative agreement 3091-21000-038-03 (A.E.P. and J.Z.Y.). S.K.V. was supported by a CREST Award from the Ministry of Science and Technology of India.

[4]Author for correspondence: apepper@bio.tamu.edu

doi:10.3732/apps.1400077

A major difficulty in the characterization and utilization of SNPs in polyploid species is determining whether a polymorphism detected by short-read NGS is the result of alternative alleles at a single locus or the presence of multiple homeologous loci. To identify markers that were likely to represent alternative alleles at a single orthologous locus, we used the Stacks bioinformatics pipeline (Catchen et al., 2011) as a filter to enrich for codominant markers composed of pairs of alleles that were homozygous in the respective taxa used for comparison (Fig. 1). Given high enough sequence coverage to accurately identify all relevant alleles and loci, the sstacks algorithm implemented by Stacks assigns an aa/bb marker type in this situation (Scenario 1). In contrast, detection of polymorphisms between homeologous loci will likely give rise to an ab/ab marker type prediction (Scenario 2). Detection of a polymorphism between paralogs within a subgenome will give rise to an aa/ab marker type (Scenario 3). Markers that are homozygous in one parent and heterozygous in the other parent will also give rise to an aa/ab marker type prediction (Scenario 4). A host of other combinations will give rise to either ab/ab or aa/bb patterns (not shown).

Given the array of possible scenarios that can give rise to candidate SNPs using GBS, we focused our efforts on markers with a simple aa/bb biallelic pattern. We considered these to be the best candidates for single loci with codominant polymorphisms between cotton accessions that would be useful for downstream applications such as genetic diversity studies, linkage and QTL mapping, genome-wide association studies (GWAS), and marker-assisted selection.

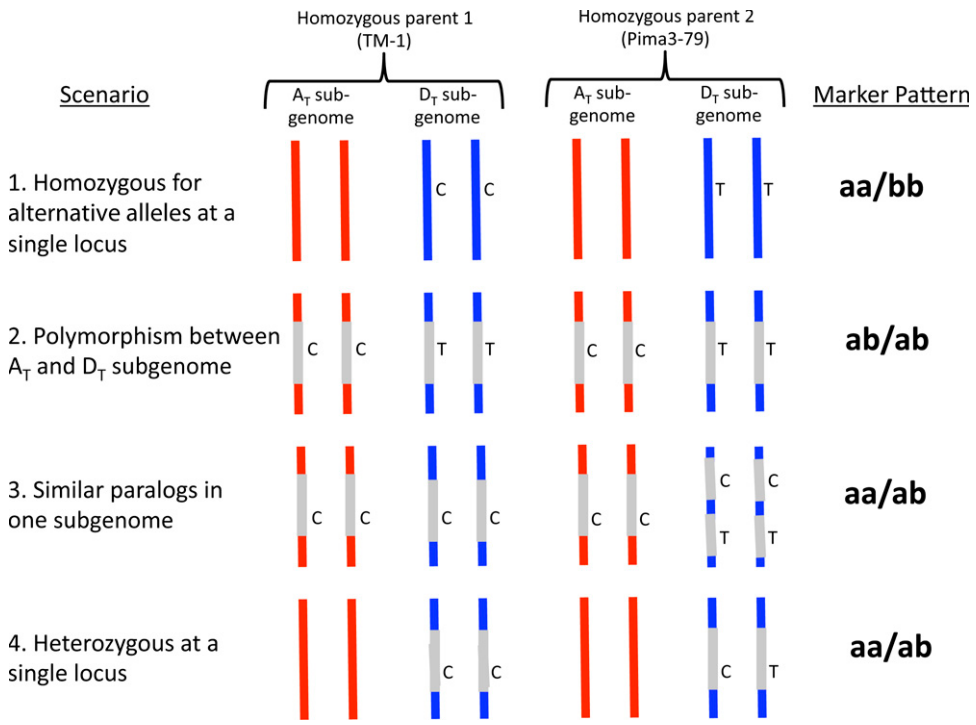

Fig. 1. Predicted marker type categories from the sstacks algorithm for four common genetic scenarios (out of many) that give rise to apparent GBS polymorphisms between two allotetraploids. Red lines indicate sequences that can be clearly assigned to the A_T subgenome, and blue lines indicate those that can be assigned to the D_T subgenome. Gray lines indicate regions of high sequence similarity between homeologs or paralogs (e.g., no differences outside of the SNP of interest). Marker type predictions are based on the assumption that there is adequate sequence coverage to accurately score all alleles at all relevant loci.

METHODS AND RESULTS

GBS was performed by a method similar to the general strategy outlined previously (Elshire et al., 2011), with major differences. Genomic DNAs from cotton taxa (Table 1) were extracted from liquid N_2 flash-frozen seedlings using a protocol described previously (Pepper and Norwood, 2001) with the addition of 1/10th volume of Plant RNA Isolation Aid (Ambion, Austin, Texas, USA) to the initial extraction buffer. Genomic DNA (250 ng) was digested with 10 units of restriction endonucleases *Hin*P1I or *Bsr*GI for 2 h at 37°C. *Hin*P1I is a CpG methylation-sensitive four-base cutter (G|CGC), while *Bsr*GI is a methylation-insensitive six-base cutter (T|GTACA). Adapter ligations were carried out in 50-µL reactions in the presence of 40 pmol of the restriction-enzyme appropriate combination of annealed, 6-bp bar-coded, Illumina-compatible P5 and P7 adapters (Appendix 1) and 1600 cohesive-end ligation units of T4 DNA ligase (New England Biolabs, Ipswich, Massachusetts, USA) for 1 h at 37°C. Because active restriction enzymes were still present in these reactions, a final incubation for 30 min at 37°C was performed to cleave any chimeric ligation products between genomic DNA fragments (the adapters were designed to abolish the *Hin*P1I or *Bsr*GI recognition site upon ligation). All samples were pooled and then purified using MinElute columns (QIAGEN, Valencia, California, USA). Single-copy regions of the plastid genome can be present in >1500 copies per cell in leaf tissue (Zoschke et al., 2007) and may thus represent a major contaminant of GBS libraries. In silico restriction digestion of the *G. hirsutum* plastid genome (Lee et al., 2006) showed that no fragments were present between 186 bp and 218 bp in *Hin*P1I digests and between 144 bp and 300 bp in *Bsr*GI digests. To exclude fragments originating from the plastid genome, and to achieve complexity reduction, size selection (190–210 bp for *Hin*P1I and 160–290 bp for *Bsr*GI) was performed using 2.5% agarose gels stained with Gel Green (Biotium, Hayward, California, USA). Size-selected fractions were treated with NEBNext Fill-in and ssDNA Isolation Module (New England Biolabs) to obtain single-stranded biotinylated fragments to use as template for PCR amplification with Illumina-compatible primers PCR 1.01 and PCR 2.01 (Appendix 1). PCR cycling conditions consisted of 98°C for 30 s followed by 20–30 cycles of 98°C for 12 s, 65°C for 30 s, 72°C for 30 s, with a final extension

TABLE 1. Seed sources, taxonomy, and preliminary GBS statistics for a set of diploid (A_1-27, D_5-1) and allotetraploid cottons.

Year	Scientific name	Name or designation	PI no.	Origin	*Bsr*GI sorted sequences	*Bsr*GI unique stacks	*Hin*P1I sorted sequences	*Hin*P1I unique stacks
2003	*G. herbaceum* L.	A_1-27	PI 408785	Peru	1,678,012	12,638	1,743,678	43,883
1989	*G. raimondii* Ulbr.	D_5-1	PI 530898	Ecuador	949,301	4215	15,323,076	8932
1984	*G. barbadense* L.	K-56	PI 274514	Sinchao Chico, Piura, Peru,	3,995,793	16,668	3,440,581	17,332
2005	*G. hirsutum* L.	TM-1	PI 607172	College Station, Texas, USA	2,752,301	22,862	2,609,330	10,792
2002	*G. barbadense* L.	Pima 3-79		Sacaton, Arizona, USA	2,756,143	23,077	1,816,093	9059
2005	*G. hirsutum* L.	TX-231	PI 163725	Zacapa, Zacapa, Guatemala	318,042	7895	592,566	1739

Note: PI no. = Plant Introduction number, National Plant Germplasm System.

time of 1 min at 72°C using Phusion polymerase (New England Biolabs). Because of the very narrow range of fragment sizes that were extracted in the size selection step, 20 cycles of PCR were required for amplification of *Bsr*GI libraries and 30 cycles were required for *Hin*P1I libraries. The samples were purified using Agencourt AMPure XP beads (Beckman Coulter, Brea, California, USA), then quantified using the AccuBlue High Sensitivity Quantitation Kit (Biotium) on the VICTOR X3 Multilabel Plate Reader (PerkinElmer, Akron, Ohio, USA). Samples were diluted to 10 nM, then provided to the Texas A&M AgriLife Genomics & Bioinformatics Services for sequencing on the Illumina GAII and/or HiSeq 2000 instrument (Illumina, San Diego, California, USA).

A total of ca. 54 million raw 76-bp paired-end reads were imported into the Geneious bioinformatics package (Drummond et al., 2012) to trim for quality (P = 0.05). The length of all fragments was trimmed to 66 bp for analysis using Stacks ver. 0.998 (Catchen et al., 2011). Sequences were preprocessed using the "process_radtags" script, in which the 6-bp barcodes were sorted and removed, yielding 60-bp fragments. The process radtag options included: -c (clean data) and -q (discard low quality reads). Barcode-sorted FASTQ files were processed in pairwise combinations using the "denovo_map.pl" script. The de novo map parameters included: -n 3 (mismatches allowed between loci during catalog building), -m 3 (minimum number of identical, raw reads required to create a stack), -M 2 (number of mismatches allowed between loci when processing a single individual), and -t (remove, or break up, highly repetitive RAD-Tags during ustacks). Output data files were loaded to MySQL tables (Oracle Corporation, Redwood Shores, California, USA), and SNPs between taxa were annotated in a pairwise manner. We obtained an average of ca. 2 million raw sequences and ca. 15,000 unique 'stacks' (loci) for each sample (Table 1).

To examine the partitioning of GBS markers into the gene-rich component of the genome, we used BLASTN (Altschul et al., 1990) to search the 22,862 unique *Bsr*GI stacks and 10,792 unique *Hin*P1I stacks from *G. hirsutum* TM-1 against predicted coding DNA sequences (CDSs) and transcripts from the diploid *G. raimondii* (Paterson et al., 2012; Wang et al., 2012) using a significance threshold of <1e-6. A very large proportion of *Hin*P1I fragments (~36–50%) showed some degree of sequence similarity to transcribed regions of the *G. raimondii* genome (Table 2). This proportion was far lower for *Bsr*GI fragments (~3–9%). Because *Hin*P1I is methylation sensitive, the *Hin*P1I libraries may be enriched in transcribed regions, which are hypomethylated in plant genomes (Feng et al., 2010). We also examined our GBS libraries for the presence of repetitive DNA using BLAST searches of all *Bsr*GI and *Hin*P1I against the TIGR Brassicaceae Repeat Database ver2_0_0 (Ouyang and Buell, 2004). We considered this database to be an appropriate proxy for cotton repetitive sequences because cotton (Malvaceae) and the Brassicaceae are sister taxa (Soltis et al., 2000). The presence of fragments with similarity to known repetitive elements was extremely low (<2%) in both *Bsr*GI and *Hin*P1I libraries. Organellar DNA contamination was examined using BLAST searches to the *G. hirsutum* plastid and mitochondrial genomes (Lee et al., 2006; Liu et al., 2013). The presence of fragments with similarity to the plastid and mitochondrial genomes was low in both libraries (Table 2).

For loci that were shared between any two taxa, polymorphisms were identified and categorized based on marker type predictions from Stacks. Across all taxa, totals of 18,073 and 5014 aa/bb pairwise SNP combinations were identified from the *Bsr*GI and *Hin*P1I libraries, respectively (Tables 3 and 4). Data for these marker sets have been submitted to the CottonGen SNP database (http://www.cottongen.org). To determine the extent of overlap between this SNP collection and those already available on the CottonGen database, a subset of 1475 *Bsr*GI fragments and 551 *Hin*P1I fragments from *G. hirsutum* cultivar TM-1 were searched against all 183,035 CottonGen SNPs using the batch BLAST tool (http://www.cottongen.org/tools/batch_blast; searched 3 January 2015). For *Bsr*GI, only two fragments (0.13%) had 100% matches in the CottonGen SNP database, and only 19 fragments (1.3%) had matches to similar sequences in the database (with up to three mismatches). For *Hin*P1I, only nine fragments (1.6%) had a 100% match in the CottonGen SNP database, and 27 fragments (4.9%) were similar up to three mismatches. Thus, the overlap between this and other cotton SNP collections was very low.

The proportion of aa/bb polymorphic loci to total (shared) loci was similar between *Bsr*GI and *Hin*P1I across all combinations of taxa (Fig. 2). In an intraspecific comparison within *G. barbadense* between cultivated variety Pima 3-79 and landrace K-56 (Peru), 6.7–8.1% of all markers showed aa/bb polymorphisms. In an intraspecific comparison within *G. hirsutum* between cultivated variety TM-1 and landrace TX-231 (Guatemala), 10.5–11.6% of markers showed aa/bb polymorphisms. An interspecific comparison between *G. barbadense* Pima 3-79 and *G. hirsutum* TM-1 showed the highest level of polymorphism (15–16.4%). These values correspond to approximate SNP frequencies of >0.0012–0.002 substitutions per base pair for intraspecific comparisons and >0.0028 substitutions per base pair for interspecific comparisons.

To test the efficacy of selecting aa/bb markers to enrich for orthologous loci SNPs, we employed both in silico and experimental validation. The in silico validation made use of the available A and D diploid genome sequences by determining whether the predicted aa/bb dual homozygous markers had the expected evolutionary pattern for sequences from Scenario 1 (Fig. 1). To perform this analysis, a subset of *G. hirsutum* TM-1 vs. *G. barbadense* Pima 3-79 aa/bb markers was searched against the complete genome sequences of *G. raimondii* (D-genome diploid) (Paterson et al., 2012; Wang et al., 2012) and *G. arboreum* (A-genome diploid) (Li et al., 2014) using BLASTN at a significance threshold of <1e-6. For each aa/bb marker, the tetraploid sequences and positive hits from diploid genomes were aligned using the map-to-reference tool implemented by the Geneious bioinformatics package (Drummond et al., 2012). Alignments were constructed for 549 *Hin*P1I and 1413 *Bsr*GI markers. Minimum spanning distance (smallest number of mutational changes to transition from one sequence to another) was used to classify the alignments into one of five categories

TABLE 2. BLASTN results (significance value <1e-6) using total stacks from *Gossypium hirsutum* cv. TM-1 or selected aa/bb markers across all taxa as subject.[a]

	*Bsr*GI		*Hin*P1I	
Database	TM-1	aa/bb	TM-1	aa/bb
JGI CDS	1059 (4.60%)	109 (3.10%)	3822 (35.40%)	343 (38.20%)
JGI transcript	2018 (8.80%)	245 (7.10%)	4943 (45.80%)	448 (49.90%)
BGI CDS	1941 (8.50%)	341 (9.80%)	3767 (34.90%)	325 (36.20%)
Brassicaceae repeats	102 (0.45%)	0 (0.00%)	170 (1.50%)	5 (0.56%)
Plastid	6 (0.03%)	1 (0.03%)	171 (1.50%)	6 (0.67%)
Mitochondrial	17 (0.07%)	8 (0.23%)	843 (7.81%)	4 (0.45%)
Total	22,862	3474	10,792	897

Note: BGI = Beijing Genomics Institute; CDS = coding DNA sequences; JGI = Joint Genome Institute.

[a] Results are searched against the databases indicated: JGI CDS and JGI transcript from *G. raimondii* (Paterson et al., 2012; Wang et al., 2012); Brassicaceae repeats from TIGR Brassicaceae Repeat Database ver2_0_0 (Ouyang and Buell, 2004); plastid from *G. hirsutum* plastid genome (Lee et al., 2006); and mitochondrial from *G. hirsutum* mitochondrial genome (Liu et al., 2013).

TABLE 3. Numbers of *Bsr*GI shared stacks (loci) and dual homozygous (aa/bb) marker loci across a set of intraspecific and interspecific combinations of *Gossypium* taxa.

Pairwise combination	Shared stacks	aa/bb Markers
A_1/D_5	413	216
Pima 3-79/A_1	3329	1538
Pima 3-79/D_5	2123	1057
Pima 3-79/K-56	10,623	859
Pima 3-79/TM-1	12,408	2040
Pima 3-79/TX-231	5322	1183
TM-1/A_1	3172	1550
TM-1/D_5	2003	1041
TM-1/K-56	8910	2171
TM-1/TX-231	5492	575
TX-231/A_1	2031	959
TX-231/D_5	1328	690
TX-231/K-56	4836	1512
K-56/A_1	3421	1616
K-56/D_5	2072	1066

TABLE 4. Numbers of *Hin*P1I shared stacks (loci) and dual homozygous (aa/bb) marker loci across a set of intraspecific and interspecific combinations of *Gossypium* taxa.

Pairwise combination	Shared stacks	aa/bb Markers
A_1/D_5	1201	502
Pima 3-79/A_1	2987	862
Pima 3-79/D_5	1387	351
Pima 3-79/K-56	931	62
Pima 3-79/TM-1	4921	740
Pima 3-79/TX-231	856	167
TM-1/A_1	3198	899
TM-1/D_5	1528	323
TM-1/K-56	906	182
TM-1/TX-231	961	111
TX-231/A_1	740	256
TX-231/D_5	444	119
TX-231/K-56	342	79
K-56/A_1	770	241
K-56/D_5	429	120

(Fig. 3). Dual-homozygous markers that were polymorphic between the cotton species at a single locus (Fig. 1, Scenario 1) were expected to give rise to the alignment pattern designated Category I. In contrast, polymorphisms between homeologs present in the A_T and D_T subgenomes (Fig. 1, Scenario 2) were expected to give an alignment similar to Category II (Fig. 3), in which the TM-1 and Pima 3-79 alleles were each most similar to fragments in one of the two diploid species. If a putative marker is actually a polymorphism between paralogs within a given subgenome, the expected alignment pattern would be exemplified by that shown for Category III (Fig. 3). For some markers, a likely subgenome identity could be discerned, but the marker showed unresolvable similarities to several paralogs within that subgenome (Category IV). This category may still manifest as aa/bb interspecific polymorphic markers, depending on the presence or absence of *Hin*P1I or *Bsr*GI flanking restriction sites. Finally, for some markers, the likely subgenome of the locus could not be determined (Category V) because of unresolvable similarities to fragments in both the A_T or D_T subgenomes. Again, these markers may still represent aa/bb markers, depending on flanking restriction sites. For the alignments of *Hin*P1I and *Bsr*GI markers (Table 5), positive BLAST hits were identified in one or both diploid genomes for 99.5% of *Bsr*GI fragments and 98.5% of *Hin*P1I fragments. Alignments for 83.5% of *Bsr*GI markers and 69.4% of *Hin*P1I markers (77.7% of markers overall) had the Category I pattern that was expected of aa/bb dual-homozygous single-locus markers, while only 1.6% of *Bsr*GI markers and 10.7% of *Hin*P1I markers (4.2% overall) had patterns indicating that they represented polymorphisms between homeologous loci from the A_T or D_T subgenomes (Category II) and only 1.5% of markers had alignment patterns suggesting that the polymorphism was between two paralogs in a given subgenome (Category III). Given that markers in both Categories IV and V can also, in principle, give rise to the aa/bb marker type, depending on flanking restriction sites, our in silico validation rate may actually be as high as 96.2% for *Bsr*GI and 89.3% for *Hin*P1I (94.2% overall).

Experimental validation was performed using the PCR-based cleaved amplified polymorphic sequence (CAPS) marker method (Konieczny and Ausubel, 1993). Only a small proportion of markers were suitable for experimental validation based on the following criteria: (1) SNP had to be near the middle of the 60-bp sequence to allow for design of flanking primers, (2) flanking sequences had to have suitable G+C content for primer design (30–60%) and lack simple sequence repeats, (3) specific primers had to be designed (using the alignments) to avoid amplification of known paralogs and homeologs, and (4) the SNP had to occur within the recognition site of a commercially available restriction enzyme. Only 22 TM-1 vs. Pima 3-79 markers (three *Hin*P1I and 19 *Bsr*GI) met all of these criteria; all of these fell into Category I (above) when examined in evolutionary alignments. Primer pairs shown in Table 6 were used in PCR amplification with the KAPA3G Plant PCR kit (Kapa Biosystems, Wilmington, Massachusetts, USA), as per the manufacturer's recommended protocol. Amplification products were examined using 4% agarose gel electrophoresis (E-Gel EX, Life Technologies, Grand Island, New York, USA) before and after restriction digestion. Of the 22 CAPS markers, one marker (Bsr1616) yielded multiple PCR amplicons, none of which were of the predicted size. One marker (Bsr18072) showed unexpected partial digestion in both TM-1 and Pima 3-79 accessions. Of those markers that could be definitively scored, 20/21 (96%) showed the predicted pattern of restriction digestion for polymorphic markers that were homozygous within each of the two taxa examined.

CONCLUSIONS

Cultivated cottons have complex allotetraploid genomes with high levels of repetitive DNA and a small proportion of gene-encoding DNA (Li et al., 2014). These characteristics greatly complicate efforts to apply GBS approaches. Foremost

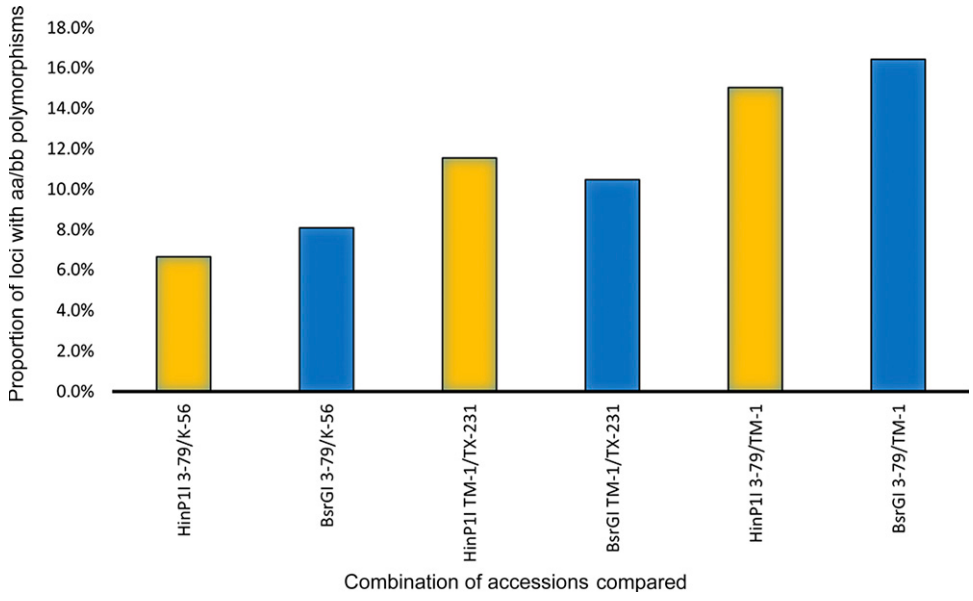

Fig. 2. *Bsr*G1 and *Hin*P1I GBS polymorphism in tetraploid *Gossypium* spp. The proportion of highly informative (aa/bb) markers relative to total shared loci (stacks) in intraspecific and interspecific pairwise comparisons is shown. 3-79 = *G. barbadense* cv. Pima 3-79; K-56 = *G. barbadense* accession K-56; TM-1 = *G. hirsutum* cv. TM-1; TX-231 = *G. hirsutum* accession TX-231.

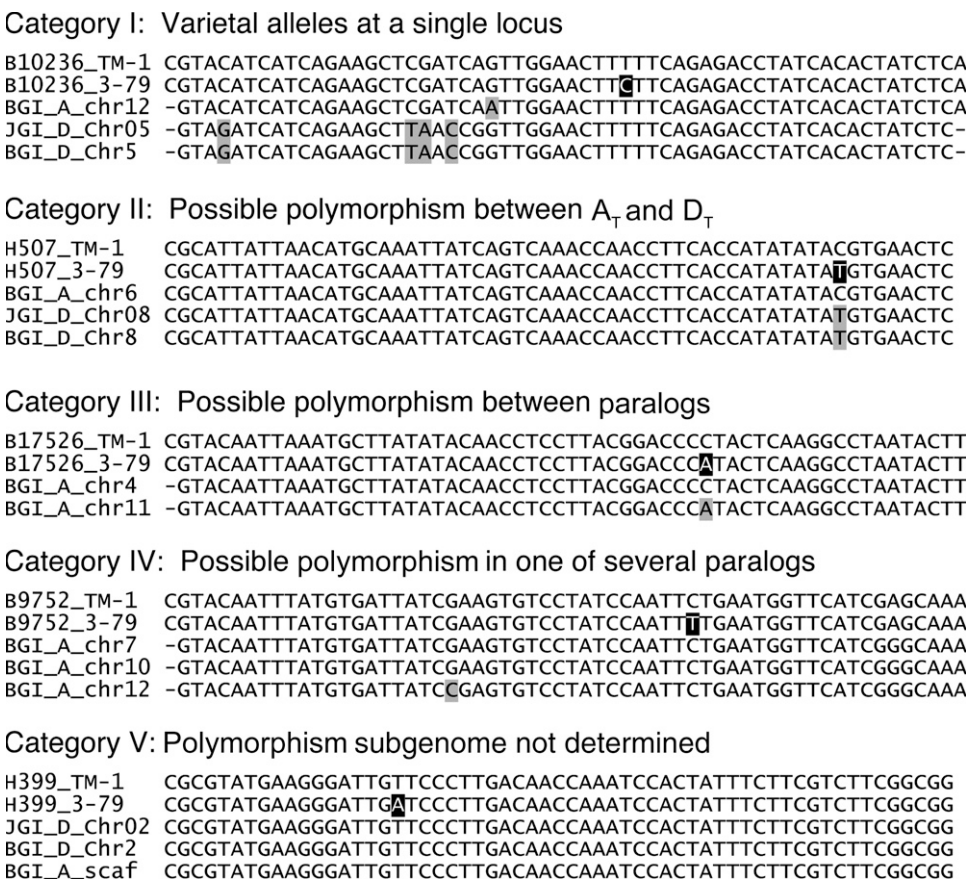

Fig. 3. Representative examples of the five categories of sequence alignments observed in TM-1 vs. Pima 3-79 polymorphic markers with aa/bb marker type assignment from Stacks. Nucleotides on a black background indicate the site of the key Pima 3-79 polymorphism relative to the TM-1 reference sequence. Nucleotides on a gray background indicate additional mismatches relative to the TM-1 reference sequence. The top two lines in each category indicate the TM-1 and Pima 3-79 fragment sequences, respectively. The prefix B indicates *Bsr*G1 markers, and H indicates *Hin*P1I. Additional lines in the alignment represent fragments from diploid genomes along with chromosomal assignments. BGI_A = *Gossypium arboreum* (Li et al., 2014); JGI_D = *G. raimondii* (Paterson et al., 2012); BGI_D = *G. raimondii* (Wang et al., 2012); scaf = scaffold.

among these difficulties is the presence of homeologous gene copies (homeologs) inherited from the diploid ancestors. Furthermore, all plant genomes have paralogous regions arising from gene duplication processes other than allotetraploidization (such as tandem duplication). To filter out the confounding polymorphisms between homeologs and between paralogs, we selected for markers with a dual-homozygous aa/bb marker prediction from the Stacks algorithm. The likelihood of such a pattern arising from more than one orthologous locus by a mutational process or by gene conversion was considered to be small compared to the straightforward interpretation of alternative alleles at a single locus. The resulting filtered marker set was highly enriched for markers with evolutionary patterns that were consistent with alternative, codominant alleles at a single locus within a particular subgenome. The application of this filter also reduced the number of markers with sequence similarity to repetitive elements and organellar genomes (Table 5). BLASTN searches against *G. raimondii* transcript and CDS databases indicated that markers derived from the methylation-sensitive restriction enzyme *Hin*P1I were highly enriched in gene-related sequences (Table 4). Thus we consider this marker set to be highly informative for mapping traits in gene-encoding regions of the genome.

Table 5. Categorization of marker alignments of aa/bb markers polymorphic between *Gossypium hirsutum* TM-1 and *G. barbadense* Pima 3-79. Alignments included TM-1 and 3-97 alleles, along with any BLAST hits (1e-6) to sequenced A- and D-genome diploid species (Paterson et al., 2012; Wang et al., 2012; Liu et al., 2013). The five categories are described in the text and illustrated in Fig. 3.

Category	*Bsr*GI	*Hin*P1I
Total fragments aligned	1413	549
Category I	1183	381
Category II	24	59
Category III	30	0
Category IV	99	16
Category V	77	93
Category V without BLAST hits to diploids	7	8
Fragments assigned to a subgenome	1312	397
Fragments assigned to A_T	841	234
Fragments assigned to D_T	471	163

The total set of 18,073 *Bsr*GI-derived and 5014 *Hin*P1I-derived polymorphic markers selected by this strategy can be used in a variety of applications across a range of taxa. For example, 921 SNPs between the photoperiodic *G. barbadense* landrace K-56 (Peru) and the photoperiod-independent cultivar Pima 3-79 could be used for mapping the photoperiodism trait. They

TABLE 6. Cleaved amplified polymorphic sequence validation of 22 aa/bb markers that are polymorphic between *Gossypium hirsutum* TM-1 and *G. barbadense* Pima 3-79.

Locus	Primer sequences (5′–3′)	Enzyme	Predicted cut[a]	Cut TM-1	Cut 3-79
Bsr1195	F: CGTACACAAAGTATTTAGAGAATATAA	*Mlu*CI	Pima 3-79		X
	R: CAAAAAGGTACGTTCCATGAAAAG				
Bsr1616[b]	F: CGTACACATGGTGAACACTTAGTAC	*Bfa*I	TM-1	(Multiple amplicons)	
	R: GTAGACAAGAGAGCTACGAGATAAAC				
Bsr3721	F: CACGTCCTAGGACACGGGCTAT	*Nla*III	Pima 3-79		X
	R: GTGTGACCGTGTGTGGCACACTA				
Bsr5368	F: CGTACAATTAGGTGTTTCGCTCTTAG	*Nla*III	TM-1	X	
	R: AGCTCTAGTATCATAACTACAGTTAGC				
Bsr7080	F: CGTACATGGAACTTTTTAAGGAGGC	*Alu*I	TM-1	X	
	R: ACATTTAATGCAAGTGCATGTAT				
Bsr7402	F: CGTACAAGACTCACCCACAAGT	*Taq*I	TM-1	X	
	R: GGCTTGATGCTGGGATTATATACAC				
Bsr9628	F: CGTACAATAGAGTTACAATAAACTCG	*Taq*I	Pima 3-79		X
	R: GTTTTTGCCGAACTTTATTCATAACA				
Bsr12910	F: CGTACAGTCAACCGCCTTAAAAATTTA	*Mse*I	TM-1	X	
	R: CTTTTACGGTGTTTTTGTTTTGACATC				
Bsr13288	F: CATCAGCATAAGGAACACGTGGCAC	*Hpy*CH2IV	Pima 3-79		X
	R: TTGACGGAATAACCAGACAAGAACA				
Bsr14160	F: CGTACATGAGTACTAAAGAGATTGG	*Nla*III	TM-1	X	
	R: GATATCTTTAATAGGGGGTGCAAC				
Bsr17257	F: CAAAGACCTCCCCCACCTACTTC	*Hpa*II	TM-1	X	
	R: TCAGCACCCTGTGGTACCTCAAG				
Bsr17701	F: CAACAACCTGCCTCACCTGCTTC	*Mlu*CI	TM-1	X	
	R: TTAGCACCTTATGGCATCTCAGGA				
Bsr18072	F: CGTACAAGAACCTCCCCCACC	*Hpa*II	TM-1	X	X
	R: CAGCACCCTGTGGCATCTCTG				
Bsr18083	F: CGTACAAACCTGAGATTTCAGGTC	*Hpy*CH2IV	Pima 3-79		X
	R: CCCTGATATGTATTGGTCGGGC				
Bsr18484	F: CGTACATTAACCCGGTTCAGGTG	*Nla*III	Pima 3-79		X
	R: ACTGGATCCATTAGTTAGAATCGGG				
Bsr18818	F: CGTACAGTTATAAGAGAAATTCCAC	*Bfa*I	TM-1	X	
	R: CTCTTCAACCCCTTGTTTTGTGATC				
Bsr20063	F: CGTACATGATAAGGACAAGAGTATT	*Mse*I	Pima 3-79		X
	R: CAGTTTTGTCCGGTACGGTCTGGCA				
Bsr20113	F: CGTACAACAATCATACAAGGAAT	*Nla*III	TM-1	X	
	R: GTCTCTAGACCCGTTCCTTCATG				
Bsr20829	F: CGTACAACTCAAGTGTACCACT	*Taq*I	Pima 3-79		X
	R: TTCCTGTTGAATTTATCTGAAATATC				
Hin2726	F: CGCATGCATGTTAGCAAGCAGTG	*Hpy*CH4V	Pima 3-79		X
	R: CGTGATTCGACGAAAACCAATC				
Hin3799	F: CCAGTTCTATCATGGCAAGATTCC	*Hpa*II	TM-1	X	
	R: GGAAGTTTCAACGAGAGAGTTGAAAG				
Hin9147	F: CAGCCCACCACTTTTCCTTACC	*Bfa*I	TM-1	X	
	R: TGTGCAGAATTGAGGGTTGCCT				

[a] Predicted cut site is based on an alignment of GBS fragment sequences.
[b] Bsr1616 yielded multiple PCR amplicons, none of which matched the expected size.

could also be employed as a resource for marker-assisted conversion of photoperiodic germplasm to photoperiod independence (Percy, 2009). Similarly, our collection included 686 SNPs between the photoperiodic *G. hirsutum* landrace TX-231 (Guatemala) and the photoperiod-independent cultivar TM-1. Finally, our collection included ca. 2000 SNP markers that can be applied to the TM-1 × Pima 3-79 interspecific recombinant inbred line (RIL) population (Yu et al., 2012) by providing a linkage-based framework for ongoing genome sequencing and chromosome assembly efforts in the allotetraploid cottons *G. hirsutum* and *G. barbadense*.

These new markers add to existing collections of cotton SNPs developed from: (1) comparative transcriptome sequencing, (2) shallow depth genome sequencing, (3) genome reduction based on restriction site conservation (GR-RSC), detected by (4) Roche 454 pyrosequencing, and (5) genotyping by sequencing (Van Deynze et al., 2009; Byers et al., 2012; Lacape et al., 2012; Rai et al., 2013; Zhu et al., 2014). Because of the small fragment size (50 bp) and intrinsic similarities between orthologs and paralogs, the SNP loci described here may be of limited use for non-sequence-based genotyping approaches (e.g., KASPR, Illumina GoldenGate) (Hyten et al., 2008; Byers et al., 2012). However, this unique GBS approach can facilitate the discovery of large sets of informative markers that can be employed to genotype extensive collections of biological samples and experimental populations (RILs, F_2, backcross) using barcoding and multiplex sequencing strategies (Elshire et al., 2011). Currently, a single lane on the Illumina HiSeq 2500 instrument can be used to genotype 96 samples at an average coverage of ca. 2 million reads per sample (the depth of coverage used in this work).

For some experimental purposes, the most flexible and cost-effective approach may be to use a "white list" of marker sequence and polymorphism data, such as that provided here, to

design a targeted set of oligonucleotides that capture and enrich selected genomic fragments for resequencing using available NGS technologies. It is important to note that the overall strategy for marker discovery and annotation that we have provided in this study can be extended to any species, including those that are allopolyploid.

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215: 403–410.

Byers, R., D. Harker, S. Yourstone, P. Maughan, and J. Udall. 2012. Development and mapping of SNP assays in allotetraploid cotton. *Theoretical and Applied Genetics* 124: 1201–1214.

Campbell, B. T., S. Saha, R. Percy, J. Frelichowski, J. N. Jenkins, W. Park, C. D. Mayee, et al. 2010. Status of the global cotton germplasm resources. *Crop Science* 50: 1161–1179.

Catchen, J. M., A. Amores, P. Hohenlohe, W. Cresko, and J. H. Postlethwait. 2011. Stacks: Building and genotyping loci de novo from short-read sequences. *G3: Genes|Genomes|Genetics* 1: 171–182.

Drummond, A. J., B. Ashton, S. Buxton, M. Cheung, A. Cooper, J. Heled, M. Kearse, et al. 2012. Geneious, version R6 for Windows. Website: http://www.geneious.com/ [accessed 6 February 2015].

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.

Feng, S., S. E. Jacobsen, and W. Reik. 2010. Epigenetic reprogramming in plant and animal development. *Science* 330: 622–627.

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.

Hyten, D., Q. Song, I.-Y. Choi, M.-S. Yoon, J. Specht, L. Matukumalli, R. Nelson, et al. 2008. High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. *Theoretical and Applied Genetics* 116: 945–952.

Konieczny, A., and F. M. Ausubel. 1993. A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. *Plant Journal* 4: 403–410.

Lacape, J.-M., M. Claverie, R. O. Vidal, M. F. Carazzolle, G. A. Guimarães Pereira, M. Ruiz, M. Pré, et al. 2012. Deep sequencing reveals differences in the transcriptional landscapes of fibers from two cultivated species of cotton. *PLoS ONE* 7: e48855.

Lee, S. B., C. Kaittanis, R. K. Jansen, J. B. Hostetler, L. J. Tallon, C. D. Town, and H. Daniell. 2006. The complete chloroplast genome sequence of *Gossypium hirsutum*: Organization and phylogenetic relationships to other angiosperms. *BMC Genomics* 7: 61–73.

Li, F., G. Fan, K. Wang, F. Sun, Y. Yuan, G. Song, Q. Li, et al. 2014. Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nature Genetics* 46: 567–572.

Liu, G., D. Cao, S. Li, A. Su, J. Geng, C. E. Grover, S. Hu, and J. Hua. 2013. The complete mitochondrial genome of *Gossypium hirsutum* and evolutionary analysis of higher plant mitochondrial genomes. *PLoS ONE* 8: e69476.

Ouyang, S., and C. R. Buell. 2004. The TIGR Plant Repeat Databases: A collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Research* 32: D360–D363.

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

Pepper, A. E., and L. E. Norwood. 2001. Evolution of *Caulanthus amplexicaulis* var. *barbarae* (Brassicaceae), a rare serpentine endemic plant: A molecular phylogenetic perspective. *American Journal of Botany* 88: 1479–1489.

Percival, A. E., and R. J. Kohel. 1990. Distribution, collection, and evaluation of *Gossypium*. *In* N. C. Brady [ed.], Advances in agronomy, vol. 44, 225–256. U.S. Department of Agriculture, Agricultural Research Service, Southern Crops Research Laboratory, College Station, Texas, USA.

Percy, R. G. 2009. The worldwide gene pool of *Gossypium barbadense* L. and its improvement. *In* A. H. Paterson [ed.], Genetics and genomics of cotton, 53–68. Plant genetics and genomics: Crops and models, Vol. 3. Springer, New York, New York, USA.

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

Soltis, D. E., P. S. Soltis, M. W. Chase, M. E. Mort, D. C. Albach, M. Zanis, V. Savolainen, et al. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Botanical Journal of the Linnean Society* 133: 381–461.

Van Deynze, A., K. Stoffel, M. Lee, T. Wilkins, A. Kozik, R. Cantrell, J. Yu, et al. 2009. Sampling nucleotide diversity in cotton. *BMC Plant Biology* 9: 125.

Wang, K., Z. Wang, F. Li, W. Ye, J. Wang, G. Song, Z. Yue, et al. 2012. The draft genome of a diploid cotton *Gossypium raimondii*. *Nature Genetics* 44: 1098–1103.

Yu, J. Z., R. J. Kohel, D. D. Fang, J. Cho, A. Van Deynze, M. Ulloa, S. M. Hoffman, et al. 2012. A high-density simple sequence repeat and single nucleotide polymorphism genetic map of the tetraploid cotton genome. *G3: Genes|Genomes|Genetics* 2: 43–58.

Zhu, Q.-H., A. Spriggs, J. M. Taylor, D. Llewellyn, and I. Wilson. 2014. Transcriptome and complexity-reduced, DNA-based identification of intraspecies single-nucleotide polymorphisms in the polyploid *Gossypium hirsutum* L. *G3: Genes|Genomes|Genetics* 4: 1893–1905.

Zoschke, R., K. Liere, and T. Börner. 2007. From seedling to mature plant: *Arabidopsis* plastidial genome copy number, RNA accumulation and transcription are differentially regulated during leaf development. *Plant Journal* 50: 710–722.

APPENDIX 1. Oligonucleotides used for adapters and primers in this study. *Note*: [Btn] = 5′-biotinylated.

1. *Hin*P1I Illumina-compatible adapters

P5 (forward) end: Top strand

A1T-1 ACACTCTTTCCCTACACGACGCTCTTCCGATCTATCACGC
A1T-2 ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGATGTC
A1T-3 ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTAGGCC
A1T-4 ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGACCAC
A1T-5 ACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGTGC
A1T-6 ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCAATC
A1T-7 ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGATCC
A1T-8 ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTTGAC
A1T-9 ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATCAGC
A1T-10 ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGCTTC
A1T-11 ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCTACC
A1T-12 ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTGTAC

P5 (forward) end: Bottom strand

A1B-1 CGGCGTGATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
A1B-2 CGGACATCGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
A1B-3 CGGGCCTAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
A1B-4 CGGTGGTCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
A1B-5 CGGCACTGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
A1B-6 CGGATTGGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
A1B-7 CGGGATCTGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
A1B-8 CGGTCAAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
A1B-9 CGGCTGATCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
A1B-10 CGGAAGCTAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
A1B-11 CGGGTAGCCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
A1B-12 CGGTACAAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

P7 (reverse) end: Top strand

A2T-0 [Btn]CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTC

P7 (reverse) end: Bottom strand

A2B-0 CGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG

2. *Bsr*GI Illumina-compatible adapters

P5 (forward) end: Top strand

Used oligonucleotides A1T-1 through A1T-12 (provided above).

P5 (forward) end: Bottom strand

B1B-1 GTACGCGTGATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
B1B-2 GTACGACATCGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
B1B-3 GTACGGCCTAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
B1B-4 GTACGTGGTCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
B1B-5 GTACGCACTGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
B1B-6 GTACGATTGGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
B1B-7 GTACGGATCTGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
B1B-8 GTACGTCAAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
B1B-9 GTACGCTGATCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
B1B-10 GTACGAAGCTAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
B1B-11 GTACGGTAGCCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
B1B-12 GTACGTACAAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

P7 (reverse) end: Top strand

A2T-0 [Btn]CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTC

P7 (reverse) end: Bottom strand

B2B-0 GTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG

3. Preparation of Illumina-compatible adapters for *Hin*P1I and *Bsr*GI GBS

Dried oligonucleotides from the supplier were dissolved in TEN (10 mM Tris pH 8, 1 mM EDTA, 50 mM NaCl) to a final concentration of 100 mM. Corresponding "top" and "bottom" oligonucleotides (10 µL) were combined, and additional TEN was added to a total volume of 40 µL. The sample was heated to 95°C, then cooled to room temperature over a period of 2 h. The addition of 60 µL of TEN resulted in working stocks at a concentration of 10 pmol of double-stranded DNA adapter per milliliter. These adapters were designated A1–A12 (forward) and A2-0 (reverse) for *Hin*P1I, and B1–B12 (forward) and B2-0 (reverse) for *Bsr*GI.

4. Modified Illumina-compatible PCR primers used for GBS library amplification

PE_PCR_Primer_1.01:

AATGATACGGGACCACCGAGATCTACACTCTTTCCCTACACGACGACGCTCTTCCGATCT

PE_PCR_Primer_2.01:

CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT