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

CARLA JO LOGAN-YOUNG2, JOHN Z. YU3, SURENDER K. VERMA2, RICHARD G. PERCY3, AND ALAN E. PEPPER2,4

2Department of Biology, Texas A&M University, College Station, Texas 77843 USA; and 3USDA-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.

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 stacks 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.
between genomic DNA fragments (the adapters were designed to abolish the HinP1I or BsrGI 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 HinP1I digests and between 144 bp and 300 bp in BsrGI digests. To exclude fragments originating from the plastid genome, and to achieve complexity reduction, size selection (190–210 bp for HinP1I and 160–290 bp for BsrGI) 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 at 72°C for 5 min.

**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 N2 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 HinP1I or BsrGI for 2 h at 37°C. HinP1I is a CpG methylation-sensitive four-base cutter (G|CGC), while BsrGI 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 HinP1I or BsrGI 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 HinP1I digests and between 144 bp and 300 bp in BsrGI digests. To exclude fragments originating from the plastid genome, and to achieve complexity reduction, size selection (190–210 bp for HinP1I and 160–290 bp for BsrGI) 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 at 72°C for 5 min.

**TABLE 1.** Seed sources, taxonomy, and preliminary GBS statistics for a set of diploid (A<sub>1</sub>-27, D<sub>1</sub>-1) and allotetraploid cottons.

| Year | Scientific name | Name or designation | PI no. | Origin | BsrGI sorted sequences | BsrGI unique stacks | HinP1I sorted sequences | HinP1I unique stacks |
|------|-----------------|---------------------|-------|--------|------------------------|---------------------|------------------------|---------------------|
| 2003 | G. herbaceum L. | A<sub>1</sub>-27    | PI 408785 | Peru   | 1,678,012              | 12,638              | 1,743,678              | 43,883              |
| 1989 | G. raimondii Ulbr. | D<sub>1</sub>-1  | PI 530898 | Ecuador | 949,301              | 4215                | 15,323,076             | 8932               |
| 1984 | G. barbadense L. | K-56                | PI 274514 | Sinchao Chico, Piura, Peru, USA | 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             |
| 2004 | G. barbadense L. | Pima-3-79             | 2,756,143 | Sacaton, Arizona, USA | 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 BrSGI libraries and 30 cycles were required for HinP11 libraries. The samples were purified using Agencourt AMPure XP beads (Beckman Coulter, Brea, California, USA), then quantified using the AccuBlue High Sensitivity Quantitation Kit (Biouton) 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 GAIIL 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 pre-processed using the "process_radtags" script, in which the 6-bp barcodes were sorted and removed, yielding 60-bp fragments. The process радtag options included: -c (clean data) and -q (discard low quality reads). Barcode-sorted FASTQ files were processed in pairwise combinations using the "de novo_map.pl" script. The de novo map parameters included: -n 3 (match allowed between loci during catalog building), -m 2 (minimum number of identical, raw reads required to create a stack), 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 BrSGI and 10,792 unique HinP11 fragments from G. hirsutumTM-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 <e^-6. A very large proportion of HinP11 fragments (35–50%) showed some degree of sequence similarity to transcribed regions of the G. raimondii genome (Table 2). This proportion was far lower for BrSGI fragments (~3%). Because HinP11 is methylation sensitive, the HinP11 libraries may be enriched in transcripted regions, which are hypo-methylated in plant genomes (Feng et al., 2010). We also examined our GBS libraries for the presence of repetitive DNA using BLAST searches of all BrSGI and HinP11 against the TIGR Brassicaceae Repeat Database ver2_0_0 (Ouyang and Baell, 2004). We considered this database to be an appropriate proxy for cotton repetitive sequences because cotton (Malvaceae) and the Brassicaceae are sister taxa (Solís et al., 2000). The presence of fragments with similarity to known repetitive elements was extremely low (<2%) in both BrSGI and HinP11 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 BrSGI and HinP11 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 BrSGI fragments and 551 HinP11 fragments from G. hirsutum cultivar TM-1 was searched against all 183,035 CottonGen SNPs using the batch BLAST tool (http://www.cottongen.org/tools/batch_blast; searched 3 January 2015). For BrSGI, 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 HinP11, 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 BrSGI and HinP11 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 (>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 <e^-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 HinP11 and 1413 BrSGI 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 <e^-6) using total stacks from Gossypium hirsutum cv. TM-1 or selected aa/bb markers across all taxa as subject.* | BrSGI | HinP11 |
|---|---|---|
| Database | TM-1 | aa/bb | TM-1 | aa/bb |
| JGI CDS | 1059 | 109 | 3822 | 343 |
| (4.60%) | (3.10%) | (35.40%) | (38.20%) |
| JGI transcript | 2018 | 245 | 4943 | 448 |
| (8.80%) | (7.10%) | (45.80%) | (49.90%) |
| BGI CDS | 1941 | 341 | 3767 | 325 |
| (8.50%) | (9.80%) | (34.90%) | (36.20%) |
| Brassicaceae repeats | 102 | 0 | 170 | 5 |
| (0.45%) | (0.00%) | (1.50%) | (0.56%) |
| Plastid | 6 | 1 | 171 | 6 |
| (0.03%) | (0.03%) | (1.50%) | (0.67%) |
| Mitochondrial | 17 | 8 | 843 | 4 |
| (0.07%) | (0.23%) | (7.81%) | (0.45%) |
| Total | 22,862 | 3474 | 10,792 | 897 |

*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 Baell, 2004); plastid from G. hirsutum plastid genome (Lee et al., 2006); and mitochondrial from G. hirsutum mitochondrial genome (Liu et al., 2013).

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

Table 3. Numbers of BrSGI 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 |
|---|---|---|
| A1/D3 | 413 | 216 |
| Pima 3-79/A1 | 3329 | 1538 |
| Pima 3-79/D3 | 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/A1 | 3172 | 1550 |
| TM-1/D | 2003 | 1041 |
| TM-1/K-56 | 8910 | 2171 |
| TM-1/TX-231 | 5492 | 575 |
| TX-231/A1 | 2031 | 959 |
| TX-231/D | 1328 | 690 |
| TX-231/K-56 | 4856 | 1512 |
| K-56/A1 | 3421 | 1616 |
| K-56/D | 2072 | 1066 |

http://www.bioone.org/loi/apps
(Fig. 3). Dual-homozygous markers that were polymorphic between the cotton species at a single locus (Fig. 1, Scenario I) were expected to give rise to the alignment pattern designated Category I. In contrast, polymorphisms between homologs present in the A₄ and D₁ 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 HinP11 or BsrGI flanking restriction sites. Finally, for some markers, the likely subgenome of the locus could not be determined (Category V) because of unresolved similarities to fragments in both the A₄ or D₁ subgenomes. Again, these markers may still represent aa/bb markers, depending on flanking restriction sites. For the alignments of HinP11 and BsrGI markers (Table 5), positive BLAST hits were identified in one or both diploid genomes for 99.5% of BsrGI fragments and 98.5% of HinP11 fragments.

**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
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 \textit{G. raimondii} transcript and CDS databases indicated that markers derived from the methylation-sensitive restriction enzyme \textit{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.

The total set of 18,073 \textit{Bsr}GI-derived and 5014 \textit{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 \textit{G. barbadense} landrace K-56 (Peru) and the photoperiod-independent cultivar Pima 3-79 could be used for mapping the photoperiodism trait. They among these difficulties is the presence of homoeologous gene copies (homeologs) inherited from the diploid ancestors. Furthermore, all plant genomes have paralogous regions arising from gene duplication processes other than allopolyploidization (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 \textit{G. raimondii} transcript and CDS databases indicated that markers derived from the methylation-sensitive restriction enzyme \textit{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.

The total set of 18,073 \textit{Bsr}GI-derived and 5014 \textit{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 \textit{G. barbadense} landrace K-56 (Peru) and the photoperiod-independent cultivar Pima 3-79 could be used for mapping the photoperiodism trait. They
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, F2, 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 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, F2, 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

---

**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: CTTACCAAGTATTTAGAAGAAATGAA | *Mlu*I | Pima 3-79 | X |
| Bsr1616\(^b\) | F: CTTACCAAGTATTTAGAAGAAATGAA | *Bfa*I | TM-1 | (Multiple amplicons) |
| Bsr3721 | F: TCTTTTACCCTGCTTGTCCT | *Nla*III | Pima 3-79 | X |
| Bsr5368 | F: TCTTTTACCCTGCTTGTCCT | *Nla*III | TM-1 | X |
| Bsr7080 | F: TCTTTTACCCTGCTTGTCCT | *Ala* | TM-1 | X |
| Bsr7402 | F: TCTTTTACCCTGCTTGTCCT | *Taq*I | TM-1 | X |
| Bsr9628 | F: TCTTTTACCCTGCTTGTCCT | *Taq*I | Pima 3-79 | X |
| Bsr12910 | F: TCTTTTACCCTGCTTGTCCT | *Mse*I | TM-1 | X |
| Bsr13288 | F: TCTTTTACCCTGCTTGTCCT | *Hpy*CH2IV | TM-1 | X |
| Bsr14160 | F: TCTTTTACCCTGCTTGTCCT | *Hpa*I | TM-1 | X |
| Bsr1757 | F: TCTTTTACCCTGCTTGTCCT | *Mio*Cl | TM-1 | X |
| Bsr18072 | F: TCTTTTACCCTGCTTGTCCT | *Hpi*I | TM-1 | X |
| Bsr18083 | F: TCTTTTACCCTGCTTGTCCT | *Hpy*CH2IV | Pima 3-79 | X |
| Bsr18484 | F: TCTTTTACCCTGCTTGTCCT | *Nla*III | Pima 3-79 | X |
| Bsr18818 | F: TCTTTTACCCTGCTTGTCCT | *Bfa*I | TM-1 | X |
| Bsr20063 | F: TCTTTTACCCTGCTTGTCCT | *Mse*I | Pima 3-79 | X |
| Bsr20113 | F: TCTTTTACCCTGCTTGTCCT | *Nla*III | TM-1 | X |
| Bsr20829 | F: TCTTTTACCCTGCTTGTCCT | *Bfa*I | Pima 3-79 | X |
| Hin2726 | F: TCTTTTACCCTGCTTGTCCT | *Hpy*CH4V | Pima 3-79 | X |
| Hin3799 | F: TCTTTTACCCTGCTTGTCCT | *Hpa*I | TM-1 | X |
| Hin9147 | F: TCTTTTACCCTGCTTGTCCT | *Bfa*I | TM-1 | X |

\(^a\) 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.

---

http://www.bioone.org/loi/apps
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. FREILICHSOWSKI, J. N. JENKINS, W. PARK, C. D. MAYEEL, ET AL. 2010. Status of the global cotton germplasm resources. Crop Science 50: 1161–1179.

CATCHEN, J. M., A. AMORES, P. HOBENSHOIRE, W. CRESKO, AND J. H. POSTLETHWAITE. 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. HELIÉ, M. KEARSE, ET AL. 2012. Geneious, version R6 for Windows. Website: http://www.geneious.com/ [accessed 6 February 2015].

ELSHERE, 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. ELSHERE, 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, L.-Y. CHOI, M.-S. YOON, J. SPEICH, L. MATUKAMILLI, 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.

LACAPÉ, 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. KATTANIS, 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 amplissicaulis var. barbarae (Brassicaceae), a rare serpentine endemic plant: A molecular phylogenetic perspective. American Journal of Botany 88: 1479–1489.

PERCY, 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.

RAI, K. M., S. K. SINGH, A. BHARDWAI, V. KUMAR, D. LAKHANI, 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.

ZHOU, 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.

ZOSCHEK, R., K. LIEBE, 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.

http://www.bioone.org/loi/apps
APPENDIX 1. Oligonucleotides used for adapters and primers in this study. Note: [Btn] = 5′-biotinylated.

1. **HinP1I Illumina-compatible adapters**

   **P5 (forward) end: Top strand**
   - A1T-1: ACACCTTTCTCCCTACACGAGGCTCTTCCGATCTATACGC
   - A1T-2: ACACCTTTCTCCCTACACGAGGCTCTTCCGATCTC
   - A1T-3: ACACCTTTCTCCCTACACGAGGCTCTTCCGATCTCTTGAC
   - A1T-4: ACACCTTTCTCCCTACACGAGGCTCTTCCGATCTGC
   - A1T-5: ACACCTTTCTCCCTACACGAGGCTCTTCCGATCTGC
   - A1T-6: ACACCTTTCTCCCTACACGAGGCTCTTCCGATCTGC
   - A1T-7: ACACCTTTCTCCCTACACGAGGCTCTTCCGATCTGC
   - A1T-8: ACACCTTTCTCCCTACACGAGGCTCTTCCGATCTGC
   - A1T-9: ACACCTTTCTCCCTACACGAGGCTCTTCCGATCTGC
   - A1T-10: ACACCTTTCTCCCTACACGAGGCTCTTCCGATCTGC

   **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] CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

   **P7 (reverse) end: Bottom strand**
   - A2B-0: CGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG

2. **BsrGI 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**
   - A2B-0: CGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG

3. **Preparation of Illumina-compatible adapters for HinP1I and BsrGI 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 HinP1I, and B1–B12 (forward) and B2-0 (reverse) for BsrGI.

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

   - **PE_PCR_Primer_1.01:**
     - AATGATACGGGACCACCGAGATCTACACTCTTTCTCCCTACACGAGC
   - **PE_PCR_Primer_2.01:**
     - CAACGCAGAAGAGCGGTTCAGCAGGAATGCCGAGACCG

http://www.bioone.org/loi/apps