PolyCat: A Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms

Justin T. Page,* Alan R. Gingle,† and Joshua A. Udall*†

*Plant and Wildlife Science Department, Brigham Young University, Provo, Utah 84062, and †Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia, 30602

ABSTRACT Read mapping is a fundamental part of next-generation genomic research but is complicated by genome duplication in many plants. Categorizing DNA sequence reads into their respective genomes enables current methods to analyze polyploid genomes as if they were diploid. We present PolyCat—a pipeline for mapping and categorizing all types of next-generation sequence data produced from allopolyploid organisms. PolyCat uses GSNAP’s single-nucleotide polymorphism (SNP)-tolerant mapping to minimize the mapping efficiency bias caused by SNPs between genomes. PolyCat then uses SNPs between genomes to categorize reads according to their respective genomes. Bisulfite-treated reads have a significant reduction in nucleotide complexity because nucleotide conversion events are confounded with transition substitutions. PolyCat includes special provisions to properly handle bisulfite-treated data. We demonstrate the functionality of PolyCat on allotetraploid cotton, Gossypium hirsutum, and create a functional SNP index for efficiently mapping sequence reads to the D-genome sequence of G. raimondii. PolyCat is appropriate for all allopolyploids and all types of next-generation genome analysis, including differential expression (RNA sequencing), differential methylation (bisulfite sequencing), differential DNA-protein binding (chromatin immunoprecipitation sequencing), and population diversity.

KEYWORDS read mapping polyploid cotton SNP homoeo-SNP

Read-mapping is a fundamental part of next-generation genomic research. Read-mapping was the essential first-step in pioneering studies of gene expression (Mortazavi et al. 2008; Wang et al. 2009), quantification of genome methylation (Lister et al. 2008; 2009), estimation of DNA—protein interactions (Park 2009; Wilbanks and Facciotti 2010), and assessment of population diversity (Sabeti et al. 2007; Durbin et al. 2010; Chia et al. 2012). Researchers have largely applied these methodologies to diploid genomes of model organisms, including Arabidopsis thaliana (Zhang et al. 2006; Vaughan et al. 2007; Cokus et al. 2008; Lister et al. 2008; Kaufmann et al. 2010), Drosophila melanogaster (Graveley et al. 2010; McManus et al. 2010; Nègre et al. 2011), and Homo sapiens (Mortazavi et al. 2008; Valouev et al. 2008; Lister et al. 2009; Trapnell et al. 2010).

Read-mapping will also be used to analyze the polyploid genomes of many important plants. It has been recently established that all seed plants are paleopolyploids, with all angiosperms sharing an additional event (Jiao et al. 2011). Thus, all flowering plants have undergone at least two paleopolyploid events in its history. Although all flowering plants have a history of whole-genome duplication (Stebbins 1950; Adams and Wendel 2005; Paterson et al. 2005; Cui 2006; Wood et al. 2009; 2011), ancient duplications do not significantly complicate read-mapping because duplicated loci diverge over time, permitting confident placement of a large majority of sequencing reads. On the other hand, more recent whole-genome duplications challenge read mapping by causing a twofold increase in chromosome number and DNA sequence while preserving gene order, coding and noncoding sequence, and chromosomal elements such as centromeres and telomeres. The increasing capacity of DNA sequencing will allow future studies to address the evolutionary and molecular hypothesis of recent polyploidization events (Osborn et al. 2003; Adams and Wendel 2005; de Peer et al. 2009; Flagel and Wendel 2009) and the effects of polyploidization on plant phenotypes (Gaeta and Pires 2010; Solits et al. 2004; Schranz 2000; Dubcovsky and Dvorak 2007). Accurate assignment of sequencing reads to their genomes-of-origin will be essential to elucidate the underlying principles and consequences of polyploid evolution.

Because most read-mapping software has been written for the analysis of diploid genomes (Griffith et al. 2010; Wu and Nacu 2010;
Garber et al. 2011; Langmead and Salzberg 2012), they are unsuited for mapping sequencing reads from polyploid samples for two reasons. First, mapping reads from a polyploid to a related diploid genome results in differential mapping efficiencies because one co-identent genome matches the reference better than the other. Differential mapping efficiency biases subsequent comparisons of the two genomes and skews quantitative analyses. Second, existing tools cannot distinguish between the two genomes to assign quantitative results to one or the other. Other phenomena, such as copy number variation, cause different problems for interpreting read mapping results and are not the focus of this effort (Kitzman et al. 2012).

The problems related to analysis of polyploid data can be mitigated by a priori single-nucleotide polymorphism (SNP) identification within and between extant diploid relatives. Most of these SNPs are vertically inherited from diploid ancestors to allopolyploid derivatives, so they are present both between diploid relatives and between coidentent homeologous genomes of the allopolyploid. These “homoeo-SNPs” can be used to reduce mapping efficiency bias through the use of SNP-tolerant mapping, as with heterozygous genes in humans (Wu and Nacu 2010). After mapping, the genome of origin for individual reads can be identified based on a comparison between the bases at the homoeo-SNP locus and the respective bases of related diploid species—a process we call read categorization.

Bisulfite-treated data present additional challenges to read mapping and read categorization because transition SNPs cannot be distinguished from bisulfite (BS) conversion events. Because transition SNPs comprise a majority of all SNPs, including homoeo-SNPs, treatment with BS causes a majority of homoeo-SNPs to be potentially uninformative for categorizing BS sequencing (BS-seq) reads.

Here we present PolyCat: a pipeline for mapping and categorizing sequencing reads from allopolyploid genomes. PolyCat was developed and tested on data derived from various species of cotton (genus Gossypium). The most common form of domesticated cotton (Gossypium hirsutum) is an allopolyploid composed of homeologous AT and DT genomes, where the ‘T’ subscript indicates genomes within the tetraploid nucleus (Wendel and Cronn 2003). Two extant diploid cotton species have genomes closely related to those contained in the polyploid nucleus, namely the A2-genome of G. aboreum and the D5-genome of Gossypium raimondii. These A2-genome is more closely related to the A1-genome than the D5-genome to the D1-genome (Senchina 2003; Flagel et al. 2012); however, the diploid D5-genome recently was sequenced because of its smaller size (Paterson et al. 2012). This characterized trio of genomes was used to develop and evaluate the read mapping and read categorization of PolyCat.

The PolyCat source code and the current cotton SNP-index is publically available for other studies (http://cottonrevolution.info), along with a web portal in which evaluation sequence data sets may be submitted for mapping and categorizing. PolyCat produces genome-specific BAM files as output, which may be immediately used by most current bioinformatics tools for downstream analyses, such as differential expression (RNA-sequencing [RNA-seq]), differential methylation (BS-seq), differential DNA-protein binding (chromatin immunoprecipitation sequencing), and population diversity.

### Table 1: Contribution of different DNA and RNA sources to construction of a SNP index

<table>
<thead>
<tr>
<th>Sequence Source</th>
<th>A2</th>
<th>D5</th>
<th>SRA IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISU fiber, leaf, buds, floral parts, seed (RNA-seq)</td>
<td>1,032,531,096</td>
<td>931,721,308</td>
<td>SRA061240 SRA061456</td>
</tr>
<tr>
<td>BYU Petal RNA-seq</td>
<td>42,047,506</td>
<td>39,974,015</td>
<td>SRA062614</td>
</tr>
<tr>
<td>Whole G. Shotgen (Genomic DNA)</td>
<td>2,996,073,656</td>
<td>168,243,740</td>
<td>SRA062615</td>
</tr>
<tr>
<td>Total</td>
<td>4,070,652,258</td>
<td>1,139,939,063</td>
<td></td>
</tr>
</tbody>
</table>

SNP, single-nucleotide polymorphism; SRA, Sequence Read Archive (National Center for Biotechnology Information).

### MATERIALS AND METHODS

**Sequence preprocessing and SNP index generation from diploid-derived data**

Sickle (https://github.com/najoshi/sickle) was used to trim all sequence reads with a quality cutoff of 20. We used the Genomic Short-read Nucleotide Alignment Program (GSNAP) (Wu and Nacu 2010) to map 1,140,550,335 reads from G. raimondii (D5), and 4,070,680,434 reads from G. aboreum (A2) to the G. raimondii reference genome (Paterson et al. 2012), accepting only unique best hits and allowing for novel splice sites (Table 1). SAMtools (Li et al. 2009) was used to generate two pileups, one for A2 and one for D5. We compared the resulting pileups with each other and with the D5 reference at each nucleotide position to identify homoeo-SNPs between the genomes, as well as allelic SNPs within the A2 and D5 genomes with at least 4× coverage and a minor allele frequency of 40%. Sequences used in this effort are available through the National Center for Biotechnology Information Sequence Read Archive (Table 1).

**RNA-seq read categorization**

We illustrate read categorization with RNA-seq reads from cotton petals in two allopolyploid cotton species: G. hirsutum (cv. Maxxa Acala and referred to as Maxxa) and Gossypium tomentosum, an endemic polyploid cotton species of Hawaii. Because the cotton AT and DT genomes are more similar to their extant diploid relatives than they are to each other (Flagel et al. 2012), SNPs between diploids approximated SNPs between their respective allopolyploid genomes and were considered putative homoeo-SNPs. These SNPs were used to categorize reads from G. hirsutum and G. tomentosum as originating from either the A2 or D5 genomes (Udall 2006; Byers et al. 2012; Flagel et al. 2012). After mapping to the D5-genome reference as described previously, PolyCat was used to compare the nucleotide at each SNP position to the SNP index and categorized it as A1-genome or D5-genome (Figure 1), depending on its unique match in the SNP index. PolyCat then assigned each read to a category based on the number of A1-genome and D5-genome matches. Reads with at least 75% (a user-specified parameter) of matches for one genome were categorized as A1 or D5, accordingly. Reads with matches to both were categorized as chimeric (X). Reads without SNP positions or matches were categorized as unknown (N).

**Bisulfite sequencing**

Bisulfite treatment deaminates unmethylated cytosines to uracils. During subsequent polymerase chain reaction, the uracil is interpreted as a thymine for complementary strand synthesis. After sequencing, cytosine-to-thymine mismatches (C→T) between the read and the reference sequence indicate unmethylated cytosines on the sequenced ‘+’ strand. Guanine to arginine mismatches (G→A) indicate unmethylated cytosines on the sequenced ‘−’ strand. This conversion looks like a transition SNP and requires tracking by PolyCat to avoid data loss.

For BS-treated data, PolyCat first inferred the origin strand of each read by counting C→T and G→A conversions. More C→T
conversions indicated ‘+’ strand, whereas more G→A conversions indicated ‘−’ strand. Ambiguous strands were counted as half reads for both strands. For ‘+’ strand reads, PolyCat accepted a thymine as a match for a cytosine allele; for ‘−’ strand reads, PolyCat accepted an adenosine as a match for a guanine. Knowing the strand origin allowed PolyCat to maximize information from each SNP. Because transition SNPs comprised the majority of the SNP index (Table 2), most reads would be uncategorizable if transition SNPs were made uninformative. However, C-T SNPs were uninformative only on the ‘+’ strand, and G-A SNPs only on the ‘−’ strand (Figure 2). So PolyCat could use C-T SNPs to categorize ‘−’ strand reads and G-A SNPs to categorize ‘+’ strand reads to minimize data loss.

After categorizing each read, PolyCat reported the number of cytosines and thymines, or guanines and adenosines, at each cytosine or guanine reference position, along with the methylation context—CG, CHG, or CHH—according to the D5-genome reference (Lister and Ecker 2009). Separate columns reported the total number of cytosines and thymines, as well as the counts for each genome (A1, D1, X, or N).

RESULTS

Homeologous SNP index

A SNP index largely composed of differences between homeologous loci was created by comparing the alignments of reads from A- and D-genome diploid species (A2 and D5, respectively) to the D5-genome reference. We refer to these single-nucleotide differences between homeologous loci as homoeo-SNPs. Our SNP index consisted of 2,633,689 SNPs (Table 2). Of these, 1,543,513 (~58.6%) were transitions (tr) and 1,055,479 were transversions (tv), a ratio of approximately 1.5 (34,697 SNPs had multiple allele possibilities in one of the two genomes and could not be classified). The gene-dense Maize HapMap 1 had a tr/tv ratio of approximately 1.0 (Gore et al. 2009), and the more uniform Maize HapMap 2 has a tr/tv ratio of approximately 2.0 (Chia et al. 2012), demonstrating a greater abundance of transition SNPs in intergenic regions in which natural selection does not prevent spontaneous cytosine to thymine mutations (Coulondre et al. 1978). These values, together with the cotton SNP-index tr/tv ratio of 1.5, suggest a correlation between the gene density of the G. raimondii genome. The average SNP density across all chromosomes was approximately 3.51 SNPs/kbp (Table S2). Chromosomes 6, 7, and 9 had slightly more than 4 SNPs/kbp, whereas Chromosomes 5, 10, and 13 had slightly less than 3 SNPs/kbp. The remaining chromosomes had between 3 and 4 SNPs/kbp.

A total of 1,123,129 SNPs were in annotated genes, including 579,259 in exonic sequence (9.4 SNPs/kbp). This increased SNP density in genes was likely due to increased sequence conservation between the A- and D-genomes (Cronn et al. 2002; Senchina 2003). The number of SNPs varied greatly between genes (Figure 3). A binomial distribution of genes with 9.4 SNPs/kbp and 1.6 kbp of average length predicted 0 genes with no coding homoeo-SNPs, but 4161 genes actually had no coding homoeo-SNPs. These data suggest strong purifying selection on these genes, possibly due to their connectedness (Birchler et al. 2005; Freeling and Thomas 2006).

SNP-tolerant mapping efficiency

SNPs between diploid relatives can approximate homoeo-SNPs between coresident genomes of an allopolyploid (Bancroft et al. 2011; Harper et al. 2012; Lai et al. 2012). SNP-tolerant mapping uses these SNPs to improve mapping efficiencies of sequence reads from allopolyploid genomes, but previous efforts (e.g., Brassica napus and Tuber aestivum) have not used SNP-tolerant mapping. To demonstrate the effectiveness of SNP-tolerant mapping, GSNAP (Wu and Nacu 2010) was used to map sequence reads from A2, D5, Maxxa, and a synthetic F1 hybrid to
than Bowtie, and a substantial increase of mapping efficiency also improved for the allopolyploid reads ince allo-

For comparison, Bowtie also was used to map the WGS reads from A2 and D5 to the D5 reference (Langmead et al. 2009). A greater rate of nonreciprocal homeologous exchange (6.8%) was recently detected in a global assembly of expressed sequence tags from *G. hirsutum* and *G. barbadense* (Flagel et al. 2012). If homeolo-
gous exchanges did not overlap a homo-SNP position or if they were larger than individual read (or expressed sequence tags), then they were not detected. Thus, these numbers likely underestimate the true num-

### Table 2 Composition of SNP index by SNP type

<table>
<thead>
<tr>
<th>At-genome</th>
<th>C-T SNP</th>
<th>D Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>190,935</td>
</tr>
<tr>
<td>T</td>
<td>190,468</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>117,349</td>
<td>363,240</td>
</tr>
<tr>
<td>G</td>
<td>363,609</td>
<td>117,194</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>At-genome</th>
<th>G-T SNP</th>
<th>D Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>190,935</td>
</tr>
<tr>
<td>T</td>
<td>190,468</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>117,349</td>
<td>363,240</td>
</tr>
<tr>
<td>G</td>
<td>363,609</td>
<td>117,194</td>
</tr>
</tbody>
</table>

Approximately one-half of the polyploid reads could not be categorized because they did not overlap a homo-SNP. The uncategorized fraction of reads varied by length and by quality of reads. In the reference genome, only 163 Mb of 749 Mb were observed with SNP-tolerant mapping enabled. Of that increase, approx-

### Read categorization of sequence reads

After mapping, PolyCat categorized each read based on matches to the SNP index (Figure 5). To test accuracy, reads from diploids were also categorized. Most reads were assigned to their correct genome (0.3% of D5 reads categorized as A and 0.8% of A2 reads categorized as D). Erroneous categorization occurred most frequently in BS-treated reads (2.1%). A small number of reads from diploids (<1%) were categorized as chimeric, indicating nucleotide matches at separate loci (within a read) to both the A- and D-genomes. Chimeric reads were slightly more common in A2 than D5. The low level of erroneous or chimeric categorization shows that PolyCat successfully categorized the vast majority of sequence reads.

For allopolyploid reads, erroneous categorization was impossible to definitively identify, but the rate of chimeric categorization was low, albeit greater than in reads from diploids (4.4% in RNA-seq and 3.8% in BS-treated reads). Two factors may explain the increase in chimeric categorization in reads from allopolyploids: (1) The SNP index was based on A2 and D5, so it includes false homo-SNPs that are really allelic SNPs specific to A2 or D5. (2) After polyploidization, gene (or intergene) conversion events between the allopolyploid genomes could have replaced the nucleotides of one genome with the nucleotides of the other. At homo-SNP positions, conversion events can be detected in reads from an allopolyploid (Salmon et al. 2009; Flagel et al. 2012), and the rate of nonreciprocal homeologous exchange had been extrapolated to be approximately 2% between the two genomes (Salmon et al. 2009). A greater rate of nonreciprocal homeologous exchange (6.8%) was recently detected in a global assembly of expressed sequence tags from *G. hirsutum* and *G. barbadense* (Flagel et al. 2012). If homeolo-
gous exchanges did not overlap a homo-SNP position or if they were larger than individual read (or expressed sequence tags), then they were not detected. Thus, these numbers likely underestimate the true num-

#### Figure 2 Homoeo-SNPs in BS treatment.

(A) Suppose there is a C-T SNP on the ‘+’ strand between the A and D genome (green characters). After BS treatment, reads ‘descending’ from the ‘+’ strand may have a C or a T, depending on the methylation state. All reads from the ‘−’ strand will have a ‘C’ at that SNP position, regardless of methylation state. And in this case, all reads from the D genome will have a T, regardless of the strand. Thus, a T base at the SNP is uninformative because it could be from the D genome or an unmethylated A genome. However, if it were known that the T nucleo-
tide was descended from the ‘−’ strand, then the T would be fully informative (i.e., it would indicate the read was unambiguously from the D-
genome in this example). As mentioned in Materials and Methods, we impute the ori-

- **Table 2 Composition of SNP index by SNP type**

<table>
<thead>
<tr>
<th>SNP type</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-T SNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D Genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-T SNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D Genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Figure 2 Homoeo-SNPs in BS treatment.

(A) Suppose there is a C-T SNP on the ‘+’ strand between the A and D genome (green characters). After BS treatment, reads ‘descending’ from the ‘+’ strand may have a C or a T, depending on the methylation state. All reads from the ‘−’ strand will have a ‘C’ at that SNP position, regardless of methylation state. And in this case, all reads from the D genome will have a T, regardless of the strand. Thus, a T base at the SNP is uninformative because it could be from the D genome or an unmethylated A genome. However, if it were known that the T nucleo-
tide was descended from the ‘−’ strand, then the T would be fully informative (i.e., it would indicate the read was unambiguously from the D-
genome in this example). As mentioned in Materials and Methods, we impute the origi-

- **Table 2 Composition of SNP index by SNP type**

<table>
<thead>
<tr>
<th>SNP type</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-T SNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D Genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-T SNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D Genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Figure 2 Homoeo-SNPs in BS treatment.

(A) Suppose there is a C-T SNP on the ‘+’ strand between the A and D genome (green characters). After BS treatment, reads ‘descending’ from the ‘+’ strand may have a C or a T, depending on the methylation state. All reads from the ‘−’ strand will have a ‘C’ at that SNP position, regardless of methylation state. And in this case, all reads from the D genome will have a T, regardless of the strand. Thus, a T base at the SNP is uninformative because it could be from the D genome or an unmethylated A genome. However, if it were known that the T nucleo-
tide was descended from the ‘−’ strand, then the T would be fully informative (i.e., it would indicate the read was unambiguously from the D-
genome in this example). As mentioned in Materials and Methods, we impute the origi-

- **Table 2 Composition of SNP index by SNP type**

<table>
<thead>
<tr>
<th>SNP type</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-T SNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D Genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-T SNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D Genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Figure 2 Homoeo-SNPs in BS treatment.

(A) Suppose there is a C-T SNP on the ‘+’ strand between the A and D genome (green characters). After BS treatment, reads ‘descending’ from the ‘+’ strand may have a C or a T, depending on the methylation state. All reads from the ‘−’ strand will have a ‘C’ at that SNP position, regardless of methylation state. And in this case, all reads from the D genome will have a T, regardless of the strand. Thus, a T base at the SNP is uninformative because it could be from the D genome or an unmethylated A genome. However, if it were known that the T nucleo-
tide was descended from the ‘−’ strand, then the T would be fully informative (i.e., it would indicate the read was unambiguously from the D-
genome in this example). As mentioned in Materials and Methods, we impute the origi-

- **Table 2 Composition of SNP index by SNP type**

<table>
<thead>
<tr>
<th>SNP type</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-T SNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D Genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-T SNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D Genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Figure 2 Homoeo-SNPs in BS treatment.

(A) Suppose there is a C-T SNP on the ‘+’ strand between the A and D genome (green characters). After BS treatment, reads ‘descending’ from the ‘+’ strand may have a C or a T, depending on the methylation state. All reads from the ‘−’ strand will have a ‘C’ at that SNP position, regardless of methylation state. And in this case, all reads from the D genome will have a T, regardless of the strand. Thus, a T base at the SNP is uninformative because it could be from the D genome or an unmethylated A genome. However, if it were known that the T nucleo-
tide was descended from the ‘−’ strand, then the T would be fully informative (i.e., it would indicate the read was unambiguously from the D-
genome in this example). As mentioned in Materials and Methods, we impute the origi-

- **Table 2 Composition of SNP index by SNP type**

<table>
<thead>
<tr>
<th>SNP type</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-T SNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D Genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G-T SNP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D Genome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
within 100 bp (the length of Illumina HiSeq reads in our dataset) of a homoeo-SNP, resulting in a 21.78% theoretical probability of any whole genome shotgun read being categorized. Genic regions (120 Mbp) had a greater density of putative homoeo-SNPs than intergenic regions because of our large collection of diploid RNA-seq data. In these regions, the theoretical probability of categorization was higher (60.7%) than the remainder of the genome (Figure 5). These data illustrate the dependency of polyploid reads categorization on SNP density.

The BS-treated reads had a decreased level of uncategorized reads because of the information loss caused by BS conversion. Each transition homoeo-SNP was only informative for half of the reads (C-T SNPs for ‘+’ strand reads and A-G SNPs for ‘−’ strand reads). Although the same portion of the genome (120 Mbp) could have been theoretically be categorized after BS treatment, the combination of transitions confounded with BS treatment and of uneven distribution of homoeo-SNP density (e.g., < single homoeo-SNP/read) caused fewer reads to be categorized in some regions than would have been
otherwise categorized had only one of the individual causes been a factor.

**Allele-SNPs within individual allopolyploid genomes**

After read categorization, SAMtools (Li et al. 2009) was used to call allele-SNPs within each genome-specific assembly (A_T and D_T). These allele-SNPs represented loci that were heterozygous within the subgenomes of *G. hirsutum* and *G. tomentosum* (Figure 6). *G. tomentosum* had slightly more allele-SNPs, representing slightly more genes, than *G. hirsutum* (Table S3). The D_T-genomes had more allele-SNPs, representing more genes, than their coresident A_T-genomes. Approximately 75% of allele-SNPs were novel (i.e., not indexed). A small number of indexed homoeo-SNPs also appeared as allele-SNPs within the genome-specific assemblies. These SNPs may reflect homeologous gene conversion events, or they may be false homoeo-SNPs.

By comparing the A_T and D_T alignments, we found that only a small number of novel homoeo-SNPs were identified in genic regions (77 in *G. hirsutum* and 59 in *G. tomentosum*), which suggests that most existing homoeo-SNPs between the A_T- and D_T- genomes were identified using the diploid genomes as surrogates. Therefore, increased sequencing of tetraploid transcriptomes will only minimally augment the number of “new” homoeo-SNPs; however, it would likely decrease the number of false-positive homoeo-SNPs resulting from diploid specific nucleotides.

**DISCUSSION**

**The phylogenetic context of SNPs**

Read mapping in polyploid genomes is a natural application of DNA sequencing, although the practical challenges of mapping to the duplicated loci of polyploid genomes have not received much attention. These challenges include (1) mapping duplicated reads to a single reference genome, (2) the difference in similarity between the subgenomes of an allopolyploid and the diploid reference sequence, (3) gene conversion, (4) allopolyploid autapomorphies, and (5) diploid autapomorphies. Carefully classified SNPs can be used to address some of these challenges, despite the lack of a read-mapping program capable of mapping to a duplicated reference genome. For evolutionary and plant improvement studies, reads are best classified within a phylogenetic context using SNP positions and their corresponding nucleotides.

In the simplest case involving allopolyploid formation, the genomes of Parent 1 (P1) and Parent 2 (P2) are combined into a common nucleus and form an F1 (Figure 7A). Assuming that such a sexually reproducing hybrid could be created, little nucleotide substitution will have occurred between the parental genomes and their counterparts within the polyploid F1 hybrid. Thus, SNPs between the diploid parents accurately predict homoeo-SNPs between the subgenomes of the F1, allowing for improvements in polyploid F1 read mapping efficiency and read categorization. For example, a sterile cotton diploid F1 hybrid (a nascent allopolyploid) was created by a recent hybridization.
Categorization of reads from F₁ had fewer chimeric (X) reads than reads from the natural allopolyploids (Figure 5).

This simple model of polyploidization lacks the passage of time since polyploid formation, during which additional nucleotide substitutions will have accumulated (autapomorphies in the diploid and polyploid genomes; Figure 7B). The nucleotide substitutions within each genome after polyploid formation are called allele-SNPs because (1) they occurred independently in various allopolyploid individuals (e.g., accessions) and (2) they originated in only one genome and in only one of two germline chromosomes. After a single base substitution, drift, selection, or both will move the allele frequency of the derived base toward fixation or elimination. Thus, allele-SNPs can be found within individual genomes where a particular accession is heterozygous or by the comparison of two different homozygous accessions. These allele-SNPs would independently assort during meiosis after nucleotide substitution, regardless if they were identified in a heterozygote as opposed to a homozygote (Bancroft et al. 2011; Harper et al. 2012). We do not use that context-dependent terminology in cotton.

Allele-SNPs can be identified by remapping categorized reads to the reference sequence and searching the alignments using common SNP-finding tools developed for diploid genomes (Li et al. 2009; McKenna et al. 2010). As an example, by using SAMtools we identified more than 1000 new allele-SNPs within both allopolyploid genomes of G. hirsutum and G. tomentosum (Figure 6). These allele-SNPs would be the most useful type of SNPs for cotton improvement because they have been bioinformatically discriminated from homoeo-SNPs and because they could be expected to segregate in Mendelian fashion (Van Deynze et al. 2009; Byers et al. 2012; Yu et al. 2012).

Comparison of independent alignments of categorized reads identified a limited number of new homoeo-SNPs because the extant diploid relatives used for initial homoeo-SNP identification were not perfect surrogates for the actual ancestral genomes that formed the ancestral allopolyploid, AND because of autapomorphic substitutions since polyploid formation (Figure 7). Resequencing multiple diploid accessions from each genome could identify the true diploid autapomorphies and reduce the number of SNPs erroneously classified as homoeo-SNPs. With our current dataset, these two SNP types were indistinguishable in our SNP index. Fortunately, the rate of false-positive homoeo-SNP (or false-positive allele-SNPs) had a negligible impact on read mapping because neither allele was penalized as a mismatch during SNP-tolerant read mapping. Thus, PolyCat used a conservative approach where if any SNP were included in the index (regardless of its source) its respective bases would be essentially masked during mapping.

Finally, SNPs can be placed on a traditional phylogenetic tree, but only a portion of those SNPs (homoeo-SNPs and allele-SNPs within the allopolyploid) impact mapping of sequence reads from allopolyploids (Figure 7). Allele-SNPs identified in subsequent re-sequencing of additional allopolyploid accessions can be easily added to the SNP index. Thus, improvement and extension of the PolyCat’s SNP index will be an iterative process (although SNP discovery will likely reach a saturation point and plateau). The combination of both types of SNPs (homoeo- and allelic) was included in the cotton SNP index for read mapping, and a similar collection of SNPs could be compiled for other allopolyploid genomes such as Brassica napus (Bancroft et al. 2011; Harper et al. 2012) and Triticum aestivum (Lai et al. 2012).

**Effectiveness of the PolyCat pipeline**

The SNP index and read categorization process facilitated the analysis of allopolyploid cotton by reducing the bias in mapping efficiency between the two genomes and by providing a means to separate data generated for each allopolyploid genome (A₁- and D₁-genomes in cotton). Mapping all sequence reads to a single genome reference allowed for an aligned, comparative analysis between the two genomes within a given accession, as well as for more accurate analyses between accessions. Although these tools have been developed for cotton, they can be readily applied to any allopolyploid by providing an appropriate genome reference FASTA file, SNP index, and sequencing reads.

PolyCat is ultimately limited by the density of homoeo-SNPs across the genome. Reads belonging to a particular region of the genome can only be categorized if it has one or more homoeo-SNPs because every categorized read must overlap at least one SNP. The use of longer reads could improve the rate of categorization.

PolyCat is written in C++ and Perl, using BamTools (https://github.com/pezmaster31/bamtools) and Bioperl (Stajich et al. 2002). The custom scripts, the cotton SNP index, and a demo web application for demonstration of allopolyploid cotton read categorization are...
available online (http://bioinfo3.pgml.uga.edu/polyCat/upload.html). In the online version, 1 GB of sequence reads (non BS-seq) in FASTQ format can be categorized by PolyCat in approximately 15 min. Additional sequencing and development of software algorithms and tools will provide continued insights into polyploid genomes, their interactions, and their resultant phenotypes.

ACKNOWLEDGMENTS

We thank Michael Gore and Jonathan F. Wendel for valuable comments and discussion. We also thank all primary contributors of public cotton sequence data, particularly Jeremy Schmutz, Andrew H. Paterson (D-genome WGS and reference sequences), and Allen Van Deynze (A-genome WGS sequence). We thank the Brigham Young University Mentored Environment Grants, the National Science Foundation Plant Genome Program, and Cotton Inc. for generous support.

LITERATURE CITED


Cui, L., 2006 Widespread genome duplications throughout the history of flowering plants. Genome Res. 16: 738–749.


Kauffman, K., J. M. Muňho, M. Østergaard F. Farinelli, P. Krajelewski et al., 2010 Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). Nat. Protoc. 5: 457–472.


