2015-06-01

Bioinformatics for the Comparative Genomic Analysis of the Cotton (Gossypium) Polyploid Complex

Justin Thomas Page
Brigham Young University - Provo

Bioinformatics for the Comparative Genomic Analysis of the Cotton (*Gossypium*) Polyploid Complex

Justin Thomas Page

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Joshua A. Udall, Chair
Mark J. Clement
Steven M. Johnson
John S. K. Kauwe
Clinton J. Whipple

Department of Biology

Brigham Young University

June 2015

ABSTRACT

Bioinformatics for the Comparative Genomic Analysis of the
Cotton (*Gossypium*) Polyploid Complex

Justin Thomas Page
Department of Biology, BYU
Doctor of Philosophy

Understanding the composition, evolution, and function of the cotton (*Gossypium*) genome is complicated by the joint presence of two genomes in its nucleus (A_T and D_T genomes). Specifically, read-mapping (a fundamental part of next-generation sequence analysis) cannot adequately differentiate reads as belonging to one genome or the other. These two genomes were derived from progenitor A-genome and D-genome diploids involved in ancestral allopolyploidization. To better understand the allopolyploid genome, we developed PolyCat to categorize reads according to their genome of origin based on homoeo-SNPs that differentiate the two genomes. We re-sequenced the genomes of extant diploid relatives of tetraploid cotton that contain the A_1 (*G. herbaceum*), A_2 (*G. arboreum*), or D_5 (*G. raimondii*) genomes. We identified 24 million SNPs between the A-diploid and D-diploid genomes. These analyses facilitated the construction of a robust index of conserved SNPs between the A-genomes and D-genomes at all detected polymorphic loci. This index can be used by PolyCat to assign reads from an allotetraploid to its genome-of-origin. Continued characterization of the Gossypium genomes will further enhance our ability to manipulate fiber and agronomic production of cotton.

With new whole-genome re-sequencing data from 34 lines of cotton, representing all tetraploid cotton species, we explored the evolution of the cotton genome with greatly improved resolution and improved tools, including BamBam and PolyDog. Identifying SNPs and structural variants among these 34 lines and their extant diploid relatives, we clarified phylogenetic relationships among tetraploid species, including newly characterized species, and identify introgression between different species of cultivated cotton. We explored the evolution of homoeologs in the A_T- and D_T-genomes and especially the phenomenon of homoeologous conversion. Homoeologous conversion is rare in cotton, perhaps due to the vast difference in chromosome sizes in the two genomes. Several regions of the genome have been introgressed between *G. hirsutum* and *G. barbadense* resulting in superior cultivars, likely with beneficial alleles from both species and novel combinations of alleles. The genomic data provide a valuable resource for cotton researchers and breeders, who can freely access the data online at CottonGen.

Keywords: next-generation sequencing, allopolyploidy, bioinformatics, comparative genomics, cotton, *Gossypium*, HapMap

ACKNOWLEDGEMENTS

I would like to thank Don Jones and Cotton, Inc. for funding my research and education, Greg and Cynthia Page for teaching me to pursue that education, and Josh Udall for being my mentor, guide, boss, and friend throughout the process.

TABLE OF CONTENTS

TITLE PAGE i

ABSTRACT ii

ACKNOWLEDGEMENTS iii

TABLE OF CONTENTS iv

LIST OF TABLES vii

LIST OF FIGURES viii

PolyCat: A Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms 1

INTRODUCTION 1

MATERIALS AND METHODS 4

RNA-seq read categorization 4

Bisulfite sequencing 5

RESULTS 6

Homeologous SNP index 6

SNP-tolerant mapping efficiency 7

Read categorization of sequence reads 8

Allele-SNPs within individual allopolyploid genomes 10

DISCUSSION 11

The phylogenetic context of SNPs 11

Effectiveness of the PolyCat pipeline 14

REFERENCES 15

TABLES 22

FIGURES 24

Insights into the Evolution of Cotton Diploids and Polyploids from Whole-Genome Re-sequencing 31

INTRODUCTION 31

MATERIALS AND METHODS 34

Plant material 34

Acquisition of DNA sequence 35

Homoeo-SNP index 35

SNP identification and diversity analysis 36

Duplications and deletions 37

Polyploid conversion events 38

RESULTS 39

Intergenomic SNPs 39

Diversity and heterozygosity 43

Duplications and deletions ... 45
Polyploid conversion events ... 46
DISCUSSION ... 48
Genome resources for Gossypium ... 48
Insights into the genome biology of Gossypium ... 50
REFERENCES ... 57
TABLES ... 62
FIGURES ... 67

BamBam: Genome Sequence Analysis Tools for Biologists ... 70
FINDINGS ... 70
Single Nucleotide Polymorphisms ... 71
Copy Number Variants ... 72
Bisulfite-sequence Analysis ... 73
GeneVisitor ... 74
Allopolyploid Analysis ... 75
Scripts ... 75
Conclusions ... 76
REFERENCES ... 77
TABLES ... 80

Methods for Mapping and Categorization of DNA Sequence Reads from Allopolyploid Organisms ... 81
BACKGROUND ... 81
RESULTS ... 84
PolyDog Implementation ... 84
Comparative Analysis ... 85
Error Analysis ... 86
CONCLUSION ... 89
REFERENCES ... 93
FIGURES ... 95

DNA Sequence Evolution and Homoeologous Conversion in Allotetraploid Cotton ... 100
INTRODUCTION ... 100
MATERIALS AND METHODS ... 103
Sequence Data ... 103
Homoeo-SNP Identification ... 104
Mapping and Categorization ... 104
Single Nucleotide Polymorphisms ... 105
Organelle Genomes ... 106
Copy Number Variants ... 106
RESULTS ... 107

Mapping and Categorization 107
Single Nucleotide Polymorphisms 108
Phylogenies 110
Introgression 112
Copy Number Variants 113
Homoeologous Conversion 114
DISCUSSION 117
Evolution of Tetraploid Species 117
Domestication in Tetraploid Cotton 119
Homoeologous Conversion 121
REFERENCES 125
TABLES 131
FIGURES 138

LIST OF TABLES

Table 1 Contribution of different DNA and RNA sources to construction of a SNP index 22
Table 2 Composition of SNP index by SNP type 23
Table 3 Transitions and transversions in the homoeo-SNP index 62
Table 4 Summary of diploid WGS re-sequencing libraries 63
Table 5 Amount of molecular evolution between the A and D genomes of cotton 64
Table 6 Number of heterozygous loci in each accession 65
Table 7 SNPs attributable to specific areas of the phylogeny 66
Table 8 The core independent tools of BamBam 80
Table 9 Homoeo-SNPs identified between the A- and D-genomes 131
Table 10 SNPs and average diversity among sub-groups of diploids and tetraploids 132
Table 11 Evidence for introgression between AD_1 and AD_2 133
Table 12 Conserved copy number variants across sub-groups of tetraploids 134
Table 13 Possible large gene conversion events based on copy number variants 135
Table 14 Possible small gene conversion events based on genotype patterns 136
Table 15 Regions including 2 or more consecutive small gene conversion SNPs 137

LIST OF FIGURES

Figure 1 A diagram of the PolyCat read categorization process 24
Figure 2 Homoeo-SNPs in BS treatment 25
Figure 3 Histogram of SNP frequencies by gene 26
Figure 4 Mapping efficiency with and without SNP-tolerant mapping 27
Figure 5 Percentages of read categorization 28
Figure 6 SNPs in *G. hirsutum* and *G. tomentosum* compared with the SNP index 29
Figure 7 The phylogenetic contexts of SNPs within a polyploid genome 30
Figure 8 Plot of genes, homoeo-SNPs, duplications, deletions, and conversion events in the A-genomes 67
Figure 9 Premature stop codons found in each *Gossypium* genome 68
Figure 10 Neighbor-joining tree based on SNPs between genomes 69
Figure 11 Methods for read categorization 95
Figure 12 Categorization by each PolyDog step 96
Figure 13 Categorization results by method 97
Figure 14 PolyCat performance 98
Figure 15 Error rates in categorization 99
Figure 16 Mapping and categorization rates 138
Figure 17 Homoelog length and SNP density 139
Figure 18 Average diversity rate 140
Figure 19 Phylogenetic trees 142
Figure 20 Introgression 143
Figure 21 Duplicated and deleted genes 144

CHAPTER 1

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

INTRODUCTION

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 DNA2protein 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; Vaughn 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 their 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; Soltis 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), it is 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 coresident 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 coresident 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.

Sequence data from DNA treated with sodium bisulfite (for analysis of DNA methylation) 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 A_T- and D_T-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 A_2-genome of *G. aboreum* and the D_5-genome of *Gossypium raimondii*. The A_2-genome is more closely related to the A_T- genome than the D_5 genome to the D_T-genome (Senchina 2003; Flagel et al. 2012); however, the diploid D_5-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://cottonevolution.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.

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 (D_5), and 4,070,680,434 reads from G. arboreum (A_2) 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 A_2 and one for D_5. We compared the resulting pileups with each other and with the D_5 reference at each nucleotide position to identify homoeo-SNPs between the genomes, as well as allelic SNPs within the A_2 and D_5 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 A_T

and D_T 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 A_T or D_T genomes (Udall 2006a,b; Yang et al. 2006; Byers et al. 2012; Flagel et al. 2012). After mapping to the D_5-genome reference as described previously, PolyCat was used to compare the nucleotide at each SNP position to the SNP index and categorized it as A_T-genome or D_T-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 A_T-genome and D_T-genome matches. Reads with at least 75% (a user-specified parameter) of matches for one genome were categorized as A_T or D_T, 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 '2' 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 '2' 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 '2' 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 '2' strand (Figure 2). So PolyCat could use C-T SNPs to categorize '2' 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 (A_T, D_T, 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 (A_2 and D_5, 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 genic skew of a SNP collection and the tr/tv ratio.

SNPs were distributed evenly across the genome, reflecting the gene density of the G. raimondii genome. The average SNP density across all chromosomes was approximately 3.51 SNPs/kbp. 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 A_2, D_5, Maxxa, and a synthetic F_1 hybrid to the D_5 reference. The mappings were performed with and without SNP- tolerant mapping. For comparison, Bowtie also was used to map the WGS reads from A_2 and D_5 to the D_5 reference (Langmead et al. 2009).

The SNP-tolerant mapping substantially improved the mapping efficiency of reads from A_2 or allopolyploid cotton to the D_5-genome reference (Figure 4). The mapping efficiency of D_5 reads to the D- genome reference was unchanged. GSNAP mapped more A_2 reads than Bowtie, and a substantial increase of mapping efficiency was observed with SNP-tolerant mapping enabled. Of that increase, approximately 50% was observed whereas mapping A_2 BS-treated reads be- cause of the reduced sequence complexity typical of BS treatment (Lister and Ecker 2009; Laird 2010; Krueger et al. 2012). The overall mapping efficiency also improved for the allopolyploid reads since allopolyploid reads included both A_T-genome and D_T-genome reads. The improved efficiency of allopolyploid cotton reads was a result of accurate mapping of A-genome reads to the diploid D-genome reference.

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 D_5 reads categorized as A_T and 0.8% of A2 reads categorized as D_T). 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 A_2 than D_5. 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 A_2 and D_5, so it includes false homoeo-SNPs that are really allelic SNPs specific to A_2 or D_5. (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 homoeo-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 homeologous exchanges did not overlap a homoeo-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 number of historical exchanges between the two genomes.

Approximately one-half of the polyploid reads could not be categorized because they did not overlap a homoeo-SNP. The uncategorized fraction of reads varied by length and by quality of reads. In the reference genome, only 163 Mb of 749 Mbp were 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 '2' 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 sub-genomes of *G. hirsutum* and *G. tomentosum* (Figure 6). G. tomentosum had slightly more allele-SNPs, representing slightly more genes, than *G. hirsutum*. The D_T-genomes had more allele-SNPs, representing more genes, than their co-resident 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 (P_1) and Parent 2 (P_2) are combined into a common nucleus and form an F_1 (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 F_1 hybrid. Thus, SNPs between the diploid parents accurately predict homoeo-SNPs between the sub-genomes of the F_1, allowing for improvements in polyploid F_1 read mapping efficiency and read

categorization. For example, a sterile cotton diploid F_1 hybrid (a nascent allopolyploid) was created by a recent hybridization. Categorization of reads from F_1 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 homozygous or heterozygous individuals. SNP identification efforts in other species have used confusing, alternative notation (e.g., hemi-SNP, etc.) if the allele-SNPs were initially 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 autopomorphies 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 (A_T- and D_T-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.

REFERENCES

1. Adams, K. L., and J. F. Wendel, 2005 Polyploidy and genome evolution in plants. Curr. Opin. Plant Biol. 8: 135–141.
2. Bancroft, I., C. Morgan, F. Fraser, J. Higgins, R. Wells et al., 2011 Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. Nat. Biotechnol. 29: 762–766.
3. Birchler, J. A., N. C. Riddle, D. L. Auger, and R. A. Veitia, 2005 Dosage balance in gene regulation: biological implications. Trends Genet. 21: 219–226.
4. Byers, R. L., D. B. Harker, S. M. Yourstone, P. J. Maughan, and J. A. Udall, 2012 Development and mapping of SNP assays in allotetraploid cotton. Theor. Appl. Genet. 124: 1201–1214.
5. Chia, J.-M., C. Song, P. J. Bradbury, D. Costich, N. de Leon et al., 2012 Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44: 803–807.
6. Cokus, S., S. Feng, X. Zhang, Z. Chen, and B. Merriman, 2008 Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature 452: 215–219.
7. Coulondre, C., J. H. Miller, P. J. Farabaugh, and W. Gilbert, 1978 Molecular basis of base substitution hotspots in *Escherichia coli*. Nature 274: 775– 780.
8. Cronn, R. C., R. L. Small, T. Haselkorn, and J. F. Wendel, 2002 Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. Am. J. Bot. 89: 707–725.
9. Cui, L., 2006 Widespread genome duplications throughout the history of flowering plants. Genome Res. 16: 738–749.

10. de Peer, Y. V., S. Maere, and A. Meyer, 2009 The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. 10: 725.

11. Dubcovsky, J., and J. Dvorak, 2007 Genome plasticity a key factor in the success of polyploid wheat under domestication. Science 316: 1862–1866.

12. Durbin, R. M., D. L. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley et al., 2010 A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.

13. Flagel, L. E., and J. F. Wendel, 2009 Gene duplication and evolutionary novelty in plants. New Phytol. 183: 557–564.

14. Flagel, L. E., J. F. Wendel, and J. A. Udall, 2012 Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. BMC Genomics 13: 302.

15. Freeling, M., and B. C. Thomas, 2006 Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. 16: 805–814.

16. Gaeta, R. T., and J. C. Pires, 2010 Homoeologous recombination in allopolyploids: the polyploid ratchet. New Phytol. 186: 18–28.

17. Garber, M., M. G. Grabherr, M. Guttman, and C. Trapnell, 2011 Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods 8: 469–477.

18. Gore, M., J. Chia, R. Elshire, Q. Sun, and E. Ersoz, 2009 A first-generation haplotype map of maize. Science 326: 1115–1117.

19. Graveley, B. R., A. N. Brooks, J. W. Carlson, M. O. Duff, J. M. Landolin et al., 2010 The

developmental transcriptome of *Drosophila melanogaster*. Nature 471: 473–479.

20. Griffith, M., O. L. Griffith, J. Mwenifumbo, R. Goya, A. S. Morrissy et al., 2010 Alternative expression analysis by RNA sequencing. Nat. Methods 7: 843–847.
21. Harper, A. L., M. Trick, J. Higgins, F. Fraser, L. Clissold et al., 2012 Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. Nat. Biotechnol. 30: 798–802.
22. Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr et al., 2011 Ancestral polyploidy in seed plants and angiosperms. Na- ture 473: 97–100.
23. Kaufmann, K., J. M. Muiño, M. Østerås, L. Farinelli, P. Krajewski 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.
24. Kitzman, J. O., M. W. Snyder, M. Ventura, A. P. Lewis, R. Qiu et al. 2012 Noninvasive whole-genome sequencing of a human fetus. Sci. Transl. Med. 4: 137ra76.
25. Krueger, F., B. Kreck, A. Franke, and S. R. Andrews, 2012 DNA methylome analysis using short bisulfite sequencing data. Nat. Methods 9: 145–151.
26. Lai, K., C. Duran, P. J. Berkman, M. T. Lorenc, J. Stiller et al., 2012 Single nucleotide polymorphism discovery from wheat next-generation sequence data. Plant Biotechnol. J. 10: 743–749.
27. Laird, P. W., 2010 Principles and challenges of genome-wide DNA methylation analysis. Nat. Rev. Genet. 11: 191–203.
28. Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359.
29. Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009 Ultrafast and memory-

efficient alignment of short DNA sequences to the human genome. Genome Biol. 10: R25.

30. Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan et al.1000 Genome Project Data Processing Subgroup., 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.
31. Lister, R., and J. R. Ecker, 2009 Finding the fifth base: Genome-wide sequencing of cytosine methylation. Genome Res. 19: 959–966.
32. Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry et al., 2008 Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133: 523–536.
33. Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon et al., 2009 Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462: 315–322.
34. McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis et al., 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303.
35. McManus, C. J., J. D. Coolon, M. O. Duff, J. Eipper-Mains, B. R. Graveley et al., 2010 Regulatory divergence in *Drosophila* revealed by mRNA- seq. Genome Res. 20: 816–825.
36. Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, 2008 Mapping and quantifying mammalian transcriptomes by RNA- Seq. Nat. Methods 5: 621–628.
37. Nègre, N., C. D. Brown, L. Ma, C. A. Bristow, S. W. Miller et al., 2011 A cis-regulatory map of the *Drosophila* genome. Nature 471: 527–531.

38. Osborn, T. C., J. Chris Pires, J. A. Birchler, D. L. Auger, Z. Jeffery Chen et al., 2003 Understanding mechanisms of novel gene expression in polyploids. Trends Genet. 19: 141–147.

39. Park, P. J., 2009 ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. 10: 669–680.

40. Paterson, A. H., J. E. Bowers, Y. Van de Peer, and K. Vandepoele, 2005 Ancient duplication of cereal genomes. New Phytol. 165: 658–661.

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

42. Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter et al., 2007 Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.

43. Salmon, A., L. Flagel, B. Ying, J. A. Udall, and J. F. Wendel, 2009 Homoeologous nonreciprocal recombination in polyploid cotton. New Phytol. 186: 123–134.

44. Schranz, M. E., 2000 Novel flowering time variation in the resynthesized polyploid Brassica napus. J. Hered. 91: 242–246.

45. Senchina, D. S., 2003 Rate Variation Among Nuclear Genes and the Age of Polyploidy in *Gossypium*. Mol. Biol. Evol. 20: 633–643.

46. Soltis, D. E., P. S. Soltis, J. C. Pires, A. Kovarik, J. A. Tate et al., 2004 Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): cytogenetic, genomic and genetic comparisons. Biol. J. Linn. Soc. Lond. 82: 485–501.

47. Stajich, J. E., D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz et al., 2002 The Bioperl

toolkit: Perl modules for the life sciences. Genome Res. 12: 1611–1618.

48. Stebbins, G. L., 1950 Variation and Evolution in Plants. Columbia University Press, New York.

49. Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan et al., 2010 Transcript assembly and quantification by RNA-Seq reveals un-annotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28: 511–515.

50. Udall, J. A., 2006a A global assembly of cotton ESTs. Genome Res. 16: 441–450.

51. Udall, J. A., 2006b A novel approach for characterizing expression levels of genes duplicated by polyploidy. Genetics 173: 1823–1827.

52. Valouev, A., D. S. Johnson, A. Sundquist, C. Medina, E. Anton et al., 2008 Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat. Methods 5: 829–834.

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

54. Vaughn, M. W., M. Tanurdzic, Z. Lippman, H. Jiang, R. Carrasquillo et al., 2007 Epigenetic natural variation in *Arabidopsis thaliana*. PLoS Biol. 5: e174.

55. Wang, Z., M. Gerstein, and M. Snyder, 2009 RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10: 57–63.

56. Wendel, J. F., and R. C. Cronn, 2003 Polyploidy and the evolutionary history of cotton. Adv. Agronomy 78: 139–186.

57. Wilbanks, E., and M. Facciotti, 2010 Evaluation of algorithm performance in ChIP-seq peak detection. PLoS ONE 5: e11471.

58. Wood, T. E., N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon et al., 2009

The frequency of polyploid speciation in vascular plants. Proc. Natl. Acad. Sci. USA 106: 13875–13879.

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

60. Yang, S. S., F. Cheung, J. J. Lee, M. Ha, N. E. Wei et al., 2006 Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton. Plant J. 47: 761–775.

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

62. Zhang, X., J. Yazaki, A. Sundaresan, S. Cokus, S. Chan et al., 2006 Genome- wide high-resolution mapping and functional analysis of DNA methylation in *arabidopsis*. Cell 126: 1189–2001.

TABLES

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

Sequence Source	A_2	D_5	SRA IDs
ISU fiber, leaf, buds, floral parts, seed (RNA-seq)	1,032,531,096	931,721,308	SRA061240
BYU Petal RNA-seq	42,047,506	39,974,015	SRA061456
Whole G. Shotgun (Genomic DNA)	2,996,073,656	168,243,740	SRA062614
Total	4,070,652,258	1,139,939,063	SRA062615

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

Table 2 Composition of SNP index by SNP type

		Dt-genome			
		A	T	G	C
At-genome	A	0	190,935	132,443	409,059
	T	190,468	0	407,605	132,678
	G	117,349	363,240	0	86,903
	C	363,609	117,194	87,509	0

FIGURES

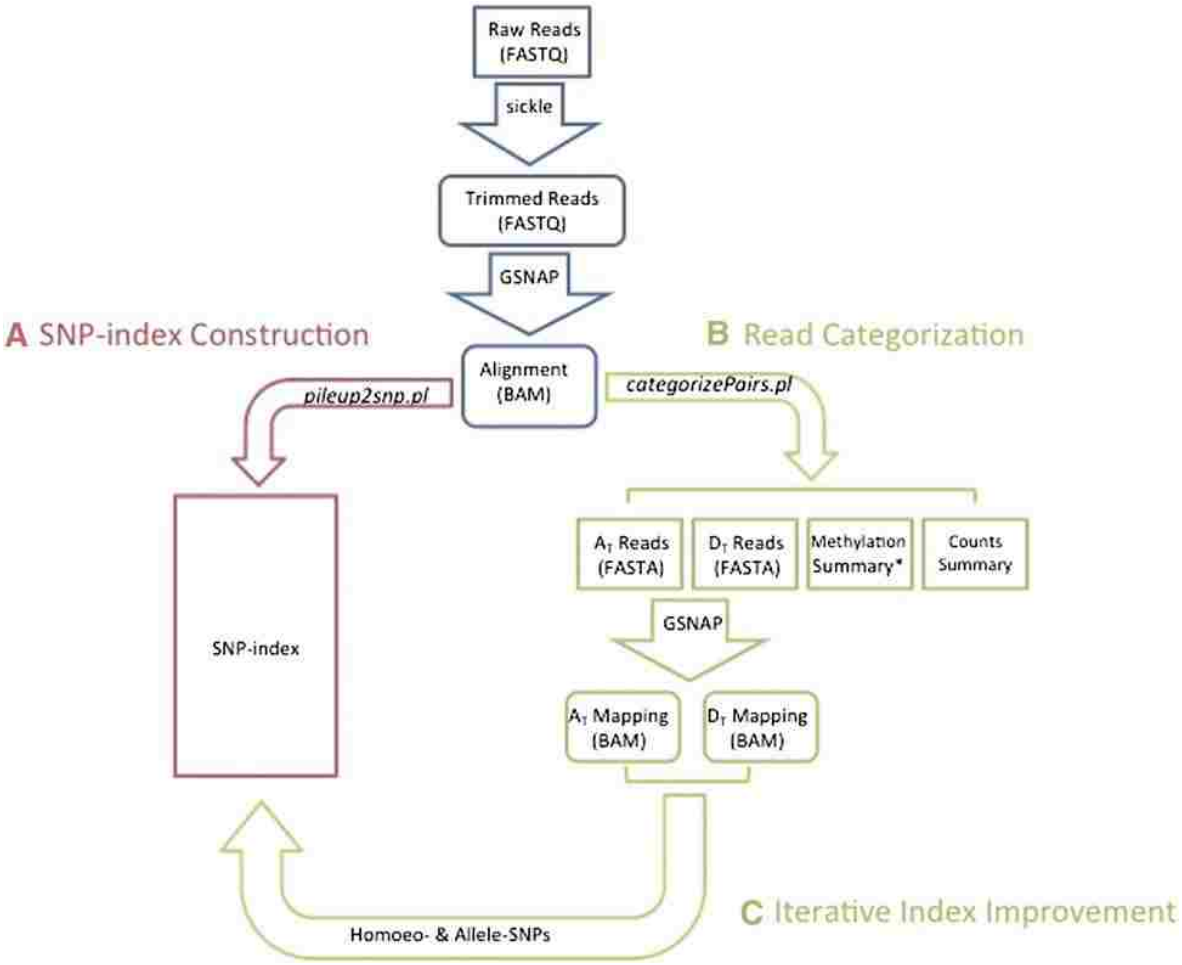


Figure 1 A diagram of the PolyCat read categorization process. (A) Reads from diploids are used to generate an index of homoeo-SNPs. (B) Reads from tetraploids are assigned to a genome based on the sequenced base at each overlapped SNP position. (C) Categorized reads from tetraploids can then be realigned into genome-specific assemblies and used to improve the SNP-index.

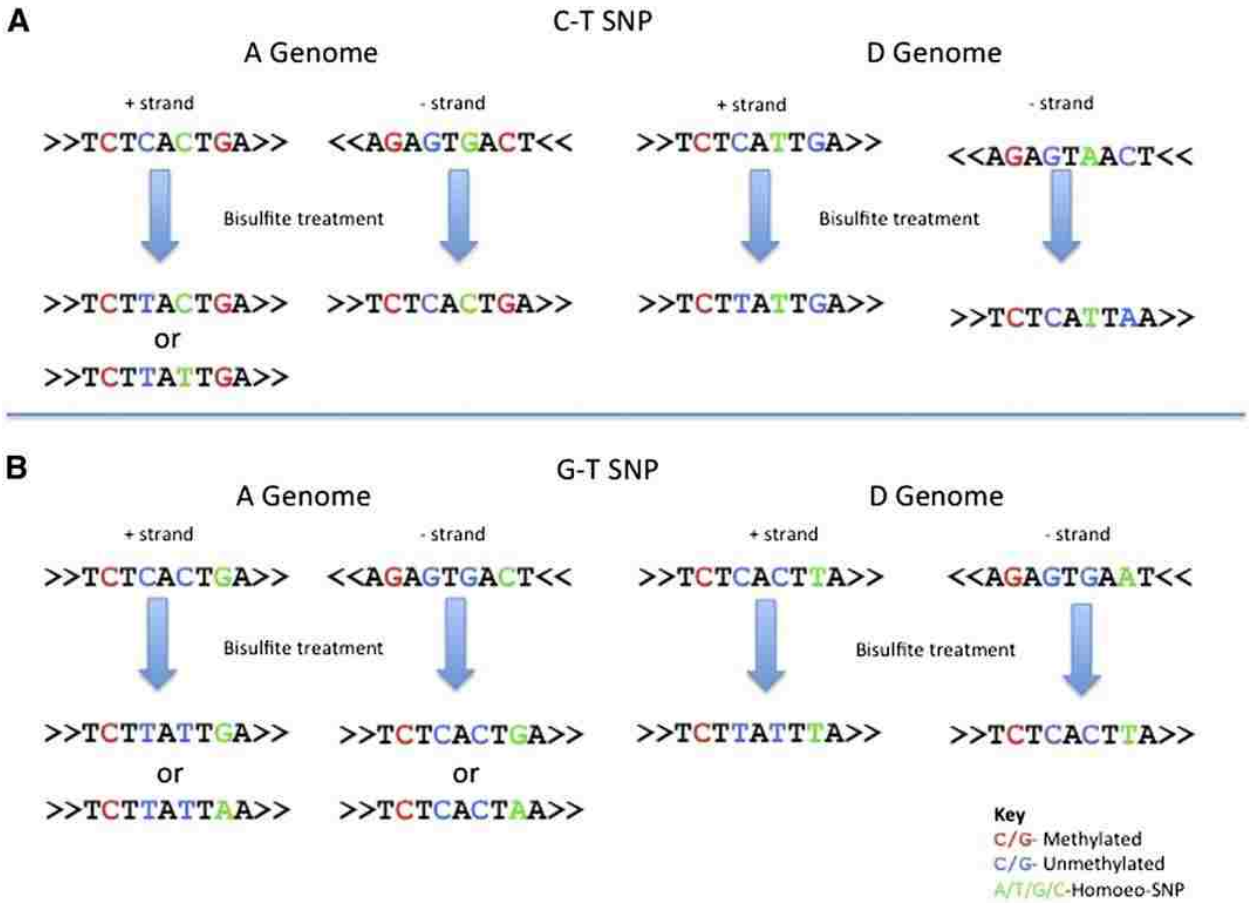


Figure 2 Homoeo-SNPs in BS treatment. (A) Suppose there is a C-T SNP on the '+' strand between the A and D genome (green characters). After BS treatment, reads 'descending' from the '+' strand may have a C or a T, depending on the methylation state. All reads from the '2' strand will have a 'C' at that 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 nucleotide was descended from the '2' 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 original read strandedness based of the frequency C/T and G/A transitions. (B) Suppose there is a G/T SNP; there is no ambiguity, then, about the genome origin of the original strand because A-genome reads will have a G or an A, whereas D-genome reads will have a T.

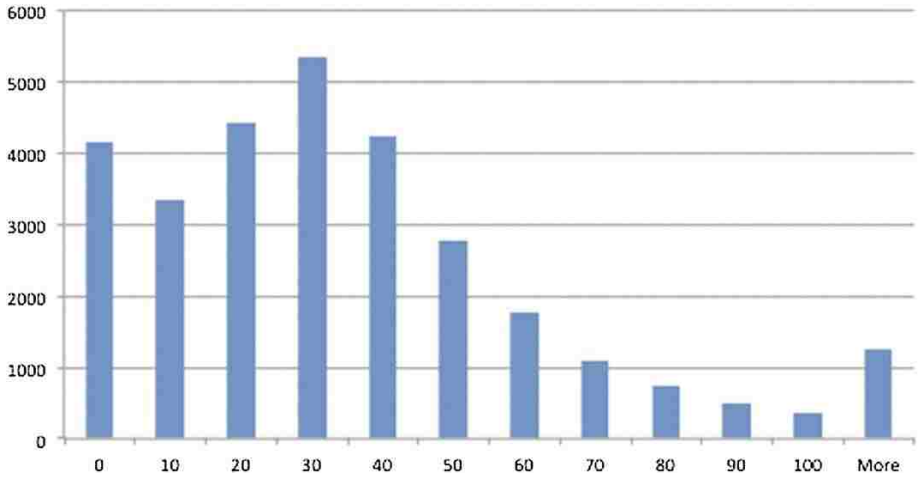


Figure 3 Histogram of SNP frequencies by gene as annotated in the initial draft of the D-genome reference sequence. Most genes (mode) had between 20 and 30 SNPs. A total of 7235 genes with low coverage (RNA-seq or WGS) from the diploid datasets were removed from the distribution.

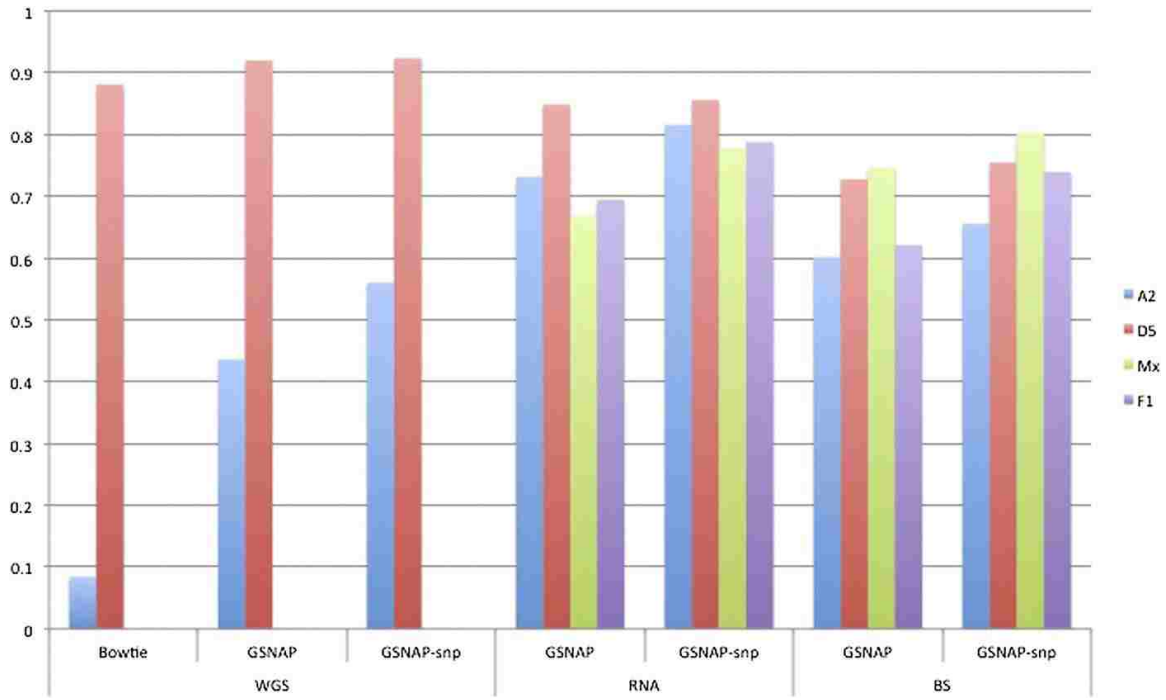


Figure 4 Mapping efficiency with and without SNP-tolerant mapping. Reads were mapped by Bowtie (WGS only), GSNAP, and GSNAP with SNP-tolerant mapping (GSNAP-snp). WGS reads from G. arboreum (A_2), G. raimondii (D_5), were mapped to the reference genome of G. raimondii. Subsequently, RNA-seq and BS-seq reads from A_2, D5, G. hirsutum (Mx) and the F_1 diploid hybrid (F_1) also mapped using SNP-tolerant mapping.

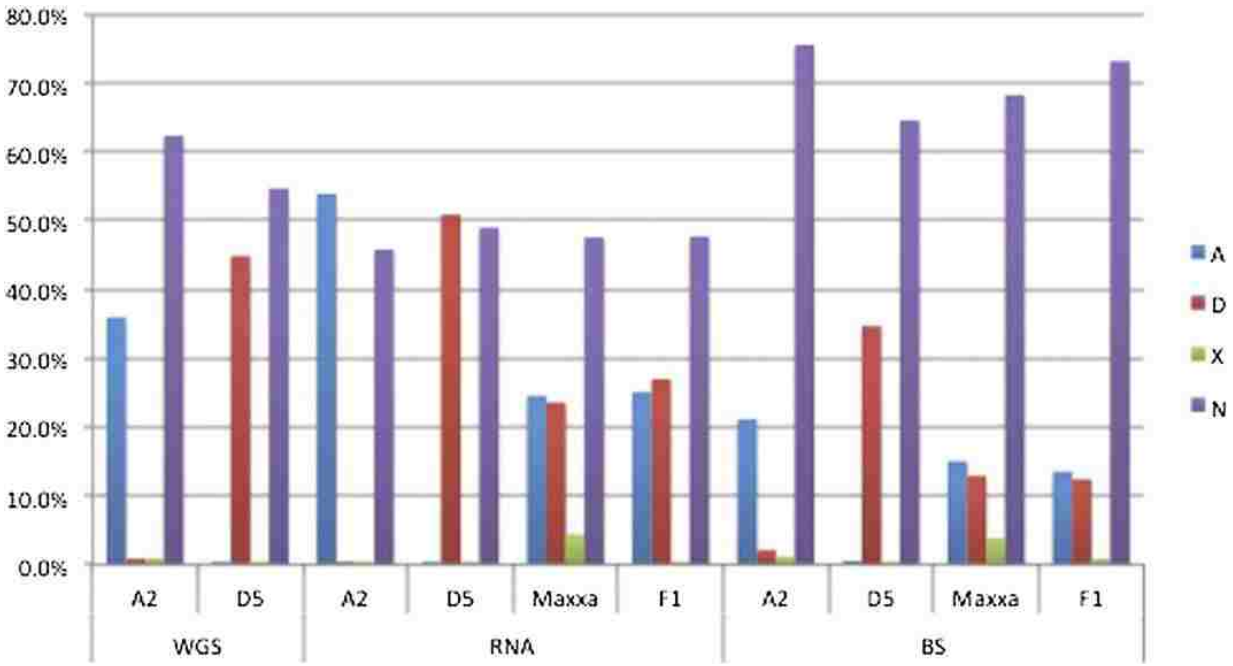


Figure 5 Percentages of read categorization. Reads were mapped to the *G. raimondii* reference with GSNAP and SNP-tolerant mapping, then categorized as A-genome (A), D-genome (D), chimeric (X), or unknown (N).

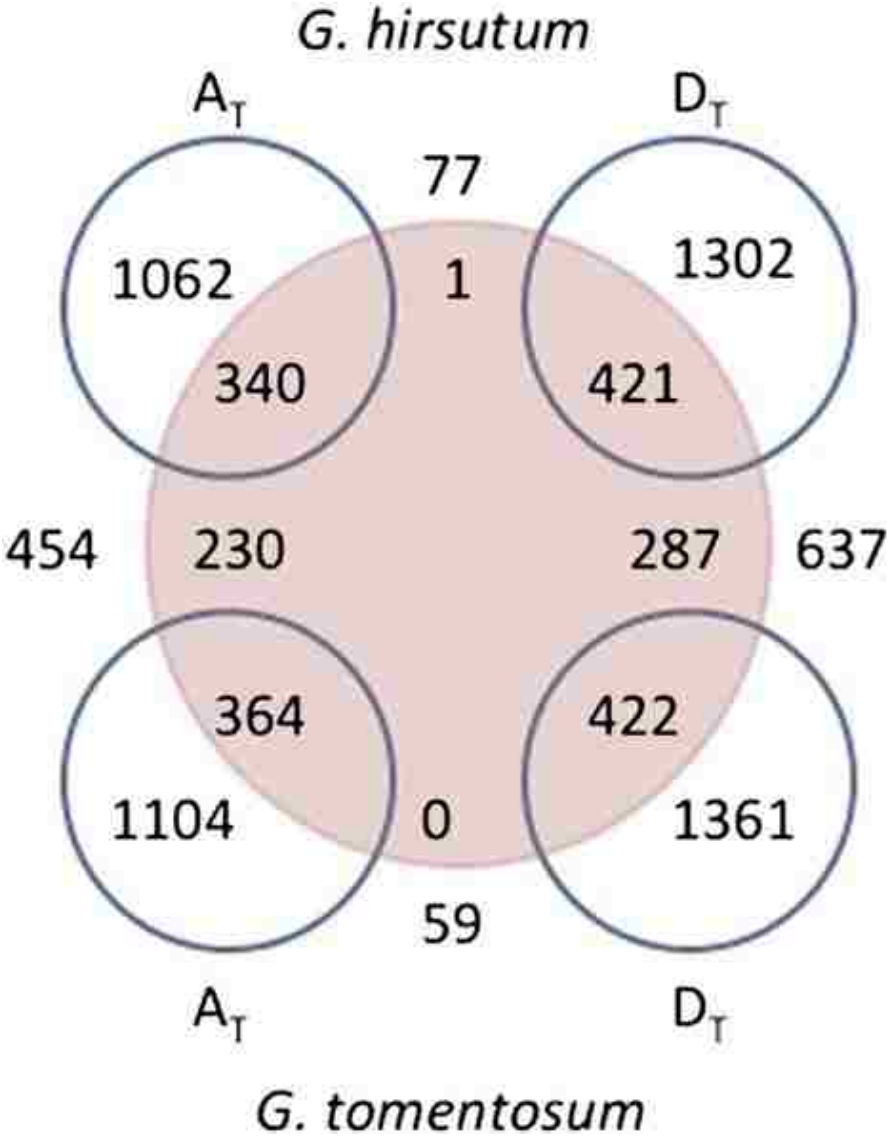


Figure 6 SNPs in *G. hirsutum* and *G. tomentosum* compared with the SNP index. Numbers inside blue circles represent the total number of SNPs for that genome, whereas underlined numbers between blue circles represent SNPs that are shared between two different allopolyploid genomes. This is not a formal Venn diagram because the numbers between blue circles are represented twice—once inside the circle(s) and once between the circles. They simply indicate the number of shared SNPs between the blue circles. Numbers inside the large red circle are indexed, while those outside were not contained within the SNP index.

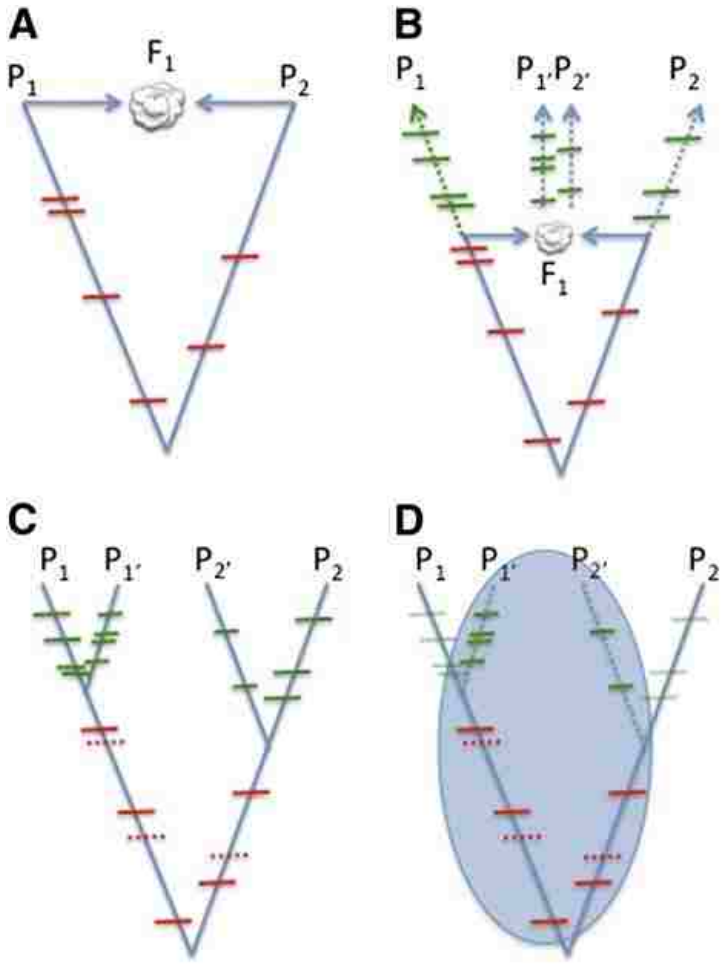


Figure 7 The phylogenetic contexts of SNPs within a polyploid genome. (A) Immediate formation of an F_1 is largely additive in terms of DNA content. SNPs between the contributing diploid genomes can be readily detected in the newly formed hybrid (red SNP marks) and F_1 reads can be readily categorized as originating from the P_1 or P_2 genome. (B) For most allopolyploids, a significant amount of time has passed since the initial genome duplication (represented by dashed arrows where time is on the y-axis). Nucleotide substitutions since polyploid formation (autapomorphies) resulted in allele-SNPs (green SNP marks). (C) SNPs can be placed within a classical phylogenetic context. Red and green SNP marks represent their respective SNP-types. Additional homoeo-SNPs (red, dashed SNP marks) were identified by comparing alignments of categorized reads (e.g., A_T- genome reference alignment to D_T-genome reference alignment). (D) The blue circle represents identified SNPs (allele- and homoeo-SNPs) that are useful for improving mapping efficiencies of allopolyploid samples. Potential false-positive homoeo-SNPs (i.e., diploid allele-SNPs) that are autapomorphic for each diploid do not negatively impact read mapping if one of the diploid alleles is common to one of the allopolyploid genomes.

CHAPTER 2

Insights into the Evolution of Cotton Diploids and Polyploids from Whole-Genome Re-sequencing

INTRODUCTION

We aligned re-sequencing reads to the existing D-genome sequence and discovered novel changes between the A-genomes and D-genomes in both diploid and polyploid plants. We identified single base differences throughout the genome between the diploid genomes and discovered that 978 genes of the D-genome reference sequence are consistently deleted in the A-genome. We discovered that approximately 900 Kbp of sequence in the polyploid genome have been converted from one genome to another in separate conversion events scattered across the genome. These discoveries help us better understand the dynamic nature of polyploid genomes and provide many avenues for further genomic research in cotton.

The genus *Gossypium* (cotton) includes approximately 45 diploid species that are divided into eight monophyletic groups, each designated by a single letter ("A" through "G" and "K," hereafter referred to as genome groups) (Wendel et al. 2012). Ancient hybridization between A and D diploids resulted in a new allopolyploid (AD) lineage in the New World approximately 1–2 million years ago (Wendel 1989). Two of the descendant allopolyploid species—*Gossypium hirsutum* (AD_1) *and Gossypium barbadense* (AD_2)—as well as two African-Asian A diploids—*Gossypium herbaceum* (A_1) and *Gossypium arboreum* (A_2)—were each independently domesticated for their long, spinnable, epidermal seed trichomes. These four species collectively provide the world's cotton fiber production, with more than 90% of this total being attributable to the cultivation of "upland cotton," *G. hirsutum* (Wendel and Cronn 2003). Understanding the

cotton genome is important for facilitating advances in crop variety development and utilization. In addition, insights into polyploid evolution in cotton may further our understanding of other polyploid crops.

Molecular studies and comparisons between diploid cotton species have revealed a genus with extraordinary genome dynamics. For example, there is a nearly three-fold variation in genome sizes among diploids (Wendel and Cronn 2003; Wendel et al. 2012), with the A-genome (1.7 Gbp) being nearly twice the size of the D-genome (0.9 Gbp), largely because of the proliferation of GORGE3 gypsy-like retrotransposons (Hawkins et al. 2006). Despite this size difference, comparative mapping studies have indicated that gene order and colinearity have been largely conserved between the diploid A-genomes and D-genomes (Brubaker et al. 1999), with the corollary that most genome size diversity reflects variation in the rates of proliferation and deletion of repetitive elements (Hawkins et al. 2009; Grover and Wendel 2010). Molecular phylogenetic and dating studies indicate that the A-genomes and D-genomes diverged approximately 5–10 million years ago. The F-genome of *Gossypium longicalyx* diverged from the A-genome after the A–D divergence, making it a suitable outgroup for a comparison of the A-genome diploids.

The respective A and D diploid genomes are closely related to the two homoeologous genomes in allopolyploid cotton, A_T and D_T ("T" denotes tetraploid), because allopolyploidization is thought to have occurred during the mid-Pleistocene era, or 1–2 million years ago (Wendel 1989). Consequently, genome differences between diploids A_2 and D_5 serve as a fair approximation of the differences between A_T and D_T tetraploid genomes (Udall 2006; Flagel et al. 2012). Thus, the existence of models of the diploid progenitors of allopolyploid cotton provide powerful reference points for inference of homoeology (e.g., of genes, transcripts,

RNA-seq reads) in allopolyploid cotton. The recent publication of the genome sequence of the D-genome diploid (*Gossypium raimondii*; Paterson et al. 2012) allows for the development of new analytical and comparative approaches for the genomics of both diploid and polyploid cotton. For example, a tool was recently created to assign the sequence reads of allopolyploid cotton to their respective genome after mapping reads from A-diploids and AD-polyploids to the *G. raimondii* (D5) reference sequence (Page et al. 2013).

Combined with the rapid increase in available sequence data, these new genomic approaches may facilitate molecular and traditional improvement efforts of cotton. For example, analysis of reads from diploids mapped to a single genome reference provides a straightforward method to identify single nucleotide polymorphisms (SNPs) between and within genomes, because each alignment of reads has identical relative positions (Page et al. 2013). In considering the relationships among sequences from A-genome, D-genome, and AD-genome cotton species, it is useful to distinguish between two classes of SNPs. Briefly, homoeo-SNPs are fixed differences that distinguish (and hence diagnose) the A-genomes and D-genomes. Allele-SNPs, however, are traditional segregating polymorphisms within a single genome, between the two alleles of an individual accession (i.e., heterozygosity) or between the corresponding homozygous alleles of different accessions. Allele-SNPs are those historically used by breeders to improve cotton cultivars in marker-assisted or genomic selection methods. Homoeo-SNPs add another layer to practical utility of allele-SNPs in that they provide a genomic feature to distinguish between duplicate gene copies. Homoeo-SNPs are also useful in an evolutionary context because their analysis offers insights into the molecular evolutionary properties of allopolyploid cotton and, more generally, allopolyploid genomes.

To better understand both diploid and allopolyploid cotton genomes, we performed deep

whole-genome re-sequencing of several diploid accessions of both A-genome and D-genome diploids. Our first objective was to determine all of the homoeo-SNPs between the A-genomes and D-genomes. Using reads from these diploids and publicly available reads from diploid and allopolyploid cottons, we compiled a database of SNPs between the various genomes studied. Our second objective was to describe genome evolution between the genomes that could be characterized by read coverage. We examined loci that are either duplicated or deleted in the A-genome species, based on coverage of A-genome reads mapped to the D_5-genome reference. Where those duplications or deletions overlap with genes, they may provide insight into the evolutionary basis for the phenotypic differences among diploids, including the production of spinnable fiber in A-genome diploid species. Our third objective was to document the extent of genome interaction based on sequence data in the polyploid (i.e., conversion events). A robust description of conversion events throughout the cotton genomes will serve as a bioinformatic aide to future genomic analyses of allopolyploid cotton.

MATERIALS AND METHODS

Plant material

Plant material was grown and harvested from greenhouses at Brigham Young University (D_5-2, D_5-31, A_1-155, A_2-34, A_2-1011), Iowa State University (D_5-4, D_5-53, A_2-4, A_1-73), and Texas A&M University (A_2-255). DNA was extracted from four accessions of *G. raimondii* (D_5-2, D_5-4, D_5-31, D_5-53), two accessions of *G. herbaceum* (A_1-73, A_1-155), and four accessions of *G. arboreum* (A_2-4, A_2-34, A_2-255, A_2-1011) using a Qiagen DNeasy plant kit.

Acquisition of DNA sequence

After shearing DNA with a Covaris instrument at the Huntsman Cancer Institute (Salt Lake City, UT), DNA libraries were prepared with the Illumina TruSeq V3 kit and sequenced by Beijing Genome Institute (BGI, Sacramento, CA), producing 100-bp paired-end reads. We assumed that the Illumina library construction process would perform equally well on high-quality DNA of the A-genomes and D-genomes. Approximately 40-times the genomic coverage was obtained for each library (Table 3). Reads from the diploids have been deposited in the NCBI Sequence Read Archive (SRA) under the fol- lowing entries: PRJNA202235, PRJNA202236, and PRJNA202239 for *G. arboreum*, *G. herbaceum*, and *G. raimondii*, respectively. Additional genomic sequence reads for *G. longicalyx* (F_1-1; SRR617255), *G. herbaceum* (A_1-97; SRR617256, SRR617284, SRR617704), and *G. hirsutum* cv. Maxxa (SRR617482) were obtained from the Sequence Read Archive. All reads were trimmed for quality with Sickle using a minimum phred quality threshold of 20 (https://github.com/ najoshi/sickle).

Homoeo-SNP index

An index of homoeo-SNPs between the A-genomes and D-genomes was produced by comparing sequence data from nine Gossypium diploids (A_1-97, A_1-155, A_2-34, A_2-255, A_2-1011 vs. D_5-2, D_5-4, D_5-31, D_5-53). First, all reads were mapped with GSNAP (Wu and Nacu 2010) using the options "-n1 -Q" (requiring unique best mapping for each read) to the 13 chromosomes of the D_5 reference sequence (Paterson et al. 2012). Second, alignment files were processed with SAMtools to produce sorted BAM files (Li et al. 2009). Third, we used InterSnp, a custom code built on the BAMtools API (https:// github.com/pezmaster31/bamtools) and

available as part of the BamBam package (http://udall-lab.byu.edu/Research/Software/BamBam.aspx) to call SNPs with at least 10-times coverage and a minimum minor allele frequency of 40%. Because the alignments from diploids used the same reference genome, homologous loci in the A-genomes and D-genomes were readily compared to identify SNPs between genomes (homoeo-SNPs). Homoeo-SNPs were called at a locus position of the D-genome reference when all diploid genomes with coverage at that particular locus were homozygous, all A-genome diploids had the same base, and all D-genome diploids had the same base, different from the A-genome diploids. Finally, loci with identified homoeo-SNPs were tabulated into a text file that was converted into a homoeo-SNP index for use by GSNAP and PolyCat.

SNP identification and diversity analysis

Using the homoeo-SNP index, we again mapped the sequence reads from the nine diploids, this time using the SNP-tolerant mapping ("-v" option) of GSNAP. We also mapped reads for three additional diploids (F_1-1, A_1-73, A_2-4) and one allopolyploid (*G. hirsutum* cv. Maxxa). GSNAP and SAMtools were otherwise used as noted. Reads from the tetraploid Maxxa were assigned to the A_T-genomes and D_T- genomes using PolyCat (Page et al. 2013). All SNPs (homoeo-SNPs and allele-SNPs) were called by InterSnp between the 13 resulting BAM files, one for each A or D diploid, one for the A_T-genome of Maxxa, and one for the D_T-genome of Maxxa. The number of heterozygous loci in each individual was summarized after filtering loci within conserved duplications. We constructed a neighbor-joining tree for the various diploid accessions, as well as the A_T-genomes and D_T-genomes, using the PHYLIP (Felsenstein 1989) program "neighbor" and default settings. The distance matrix consisted of the percentage of aligned sites that differed in pairwise comparisons.

We used homoeo-SNPs between the diploids to generate a "pseudo-A" genome, with the A alleles substituted into the D_5 reference. We did the same with the Maxxa homoeo-SNPs to make "pseudo-A_T" and "pseudo-D_T" genomes. Although these pseudo-genomes did not have indels or structural variations that are present in the actual A, A_T, and D_T genomes, the majority of gene sequences were conserved (Flagel et al. 2012; Paterson et al. 2012). Thus, these pseudo-genomes served to characterize the location of allele-SNPs within genes and other conserved noncoding sequences, and having each genome on the same "scale" greatly simplifies genome comparisons.

Duplications and deletions

We detected putative duplications (relative to the D_5 reference genome) in the other diploids using MACS (with default settings), a commonly used tool for ChIP-seq analysis (Zhang et al. 2008). It empirically models peaks in coverage of ChIP-seq reads using a dynamic Poisson distribution, thereby estimating the location of a DNA binding molecule. Here, we used MACS to call coverage peaks within WGS reads from the A-genome and F-genome diploids after alignment to the D-genome reference. Assuming the libraries from both A-genomes and D-genomes would be equally biased, reads from D5- 53 served as a control sample, estimating the expected coverage pattern. Peaks in A-genome coverage relative to D_5-53 represent putative duplicated sites that were sampled at a higher frequency during sequencing. To filter out false-positives, we also called coverage peaks in D5-2, D5-4, and D5-31, relative to D5-53. We used bedtools (Quinlan and Hall 2010) to compare peaks among and between the datasets and to identify their position relative to gene annotations in the D_5 version 2.1 (Paterson et al. 2012). Similarly, putative deletions in the test diploids were called if the reference sequence D_5-53 had

20-times or higher coverage at both ends, as well as at an additional point at least 200 bp from either end of a region of at least 1000 bp, and if the test diploid had near-zero coverage (3-times) at every point in that block. This detection was performed by Gapfall, part of the BamBam package (http://udall-lab. byu.edu). Blast2Go was used for an enrichment analysis (using the Fisher exact test) on genes duplicated or deleted in the A diploids (Conesa et al. 2005). Default B2G parameters were used.

Polyploid conversion events

We used two methods to identify possible nonreciprocal homoeologous or "gene conversion" events between the A_T-genomes and D_T- genomes of G. hirsutum cv. Maxxa. We first identified individual converted loci based on homoeo-SNPs, where reads from the A_T-genome carried the D-genome nucleotide, or vice versa. The second method used duplications and deletions to identify regions of conversion, but for deletion detection using 15-times the minimum coverage for the duplicated genome and less than four-times the coverage for the "deleted" genome. If a region spanning at least 1 Kbp was "duplicated" in the A_T-genome relative to the A diploids and "deleted" in the D_T-genome relative to the D diploids, then an A_T-biased conversion event was inferred. Similarly, an A_T-genome deletion and D_T- genome duplication suggested a D_T-biased conversion. These analyses of the polyploid genome were limited to regions that were present in both diploid genomes (A and D) because homoeo-SNPs could only be predicted in such regions.

RESULTS

Intergenomic SNPs

Similar to previous work on the detection and frequency of SNPs in genic regions (Page et al. 2013), we produced a robust index of 23,859,893 homoeo-SNPs between the genomes of diploid A-genome and D-genome cotton. These SNPs covered the genome of the D-genome reference sequence at a density of one SNP per 32.3 bases (Figure 8). This total number of SNPs is a dramatic increase from the number previously reported with the D-genome sequence (Paterson et al. 2012) and in genic sequences (Page et al. 2013). The index had a transition/transversion ratio of 1.92 (Table 3), similar to the Maize HapMap2 (Chia et al. 2012). This genome-wide SNP analysis confirmed our speculation that the previous ratio was downwardly biased in our gene-focused index. Across polymorphic nucleotide positions, there was not a significant difference between the GC biases of the A-genomes and D-genomes (45.4% and 45.1%, respectively). However, these values were higher than the genome-wide GC content, suggesting an increased likelihood for SNPs at G or C nucleotides, possibly because of the high frequency of C/T mutations caused by deamination of cytosines.

The genome-wide SNP index (SNP index 2.0) was based on comparisons of deep sequence coverage between multiple diploid A-genomes and D-genomes, so we anticipated it would be more robust and widely applicable for read mapping efforts of other diploids and allopolyploids than the previous index. The improved index increased mapping efficiency of A-genome reads to the D- genome reference sequence. With the SNP-tolerant mapping of GSNAP, more than 77% of A-genome reads mapped, reflecting a mapping improvement of approximately 15% compared to mapping without the SNP-index (Table 4). D-genome mapping was unaffected (95%). The error rate of categorization of WGS reads was less than 2%, as estimated by

categorizing the diploid reads and looking for incorrectly categorized reads. Although this error rate is slightly higher than that estimated for the original PolyCat index (Page et al. 2013), it is still an acceptable rate considering the increased fraction of reads that overlap a SNP between genomes ($70%, up from 50%), and considering WGS reads mapping to less conserved intergenic regions.

Within the diploid index, genes had a median intergenomic SNP per base rate of 2.2% (range, 0–16.1%). Notably, there were 593 genes that had no unambiguous homoeo-SNPs between diploid A-genomes and D-genomes. Of these, 215 genes had one or more allele-SNPs (i.e., one diploid genome had two nucleotides, one matching the second diploid genome and the other nucleotide being novel). The remaining 378 genes were completely conserved across all accessions with no SNP differences. A Blast2Go enrichment analysis of these genes identified the following three enriched GO terms: NADH dehydrogenase (ubiquinone) activity (GO:0008137); NADH dehydrogenase (quinone) activity (GO:0050136); and NADH dehydrogenase activity (GO:0003954). Most of these genes were shorter than the average gene within the D-genome annotations (95 to 8113 bp with mean 810 6 786 SD for the 378 genes vs. 89 to 51,174 bp with mean 3249 6 2806 SD for all 37,223 genes) (Paterson et al. 2012).

In the polyploid, improved categorization of reads into its two separate genomes was enabled by the genomic SNP index. Using the SNPs from the diploids and the D-genome reference, PolyCat assigned more than 70% of mapped polyploid reads to the A_T- genomes or D_T-genomes. For the tetraploid Maxxa, a greater percentage of reads were assigned to the A_T-genome than to the D_T-genome, despite the fact that categorization only occurred in regions shared by the two genomes. This later criteria preempted the larger A-genome from an A_T categorization bias. The unexpectedly higher categorization rate of A_T reads may be partially

explained by the fact that A_1 and A_2 diploids are a two-fold better approximation of the A_T-genome than D_5 is of the D_T-genome. For example, nucleotide diversity appears to play a role in mapping efficiency among the diploid A-genome species. The most divergent line (A_1- 73) had the lowest mapping percentage of any of the of the A-genome diploid accessions. Because sequence divergence is less between the A-genome diploid and polyploid than the D-genome and polyploid, read categorization based on SNPs between the diploids would be more effective for the A_T-genome, resulting in the observed bias. To a much lesser degree, the A_T categorization bias may also be partially attributed to duplicated loci in the A_T-genome mapping to a single locus in the D_5 reference, although these artifacts were largely avoided by the detection of duplications.

Because of their recent common ancestry, many of the identified differences between the A-genome and D-genome diploids were retained between the A_T-genomes and D_T-genomes as homoeo-SNPs. A total of 20,828,020 homoeo-SNPs were identified between the A_T- genomes and D_T-genomes of the allopolyploid cultivar Maxxa. The difference between the number of approximately 20 million SNPs in the polyploid and the "retained ancestral" homoeo-SNPs ($16 million; 75.8%) were autapomorphic SNPs that were derived after the divergence of the A_T-genome and D_T-genome from the A-genome and D-genome common ancestor, respectively. This portion of the homoeo-SNPs (5,046,151; 24.2%) was only identified between the genomes in the allopolyploid and not in the comparison of the diploid genome sequences. These unique, homoeo-SNPs were found throughout the genome in 34,810 of the 37,223 annotated genes. We anticipate that additional polyploid autapomorphic SNPs will be identified as more polyploid genomes are re-sequenced.

For all of the annotated genes in the D_5-genome reference, an alignment of A, A_T, D_T,

and D-genomes was created, from which the amount of molecular evolution between the A-genomes and D-genomes of cotton was calculated (Table 5). The results of this effort concurred with our previously published work based on aligned EST contigs (Flagel et al. 2012), although SEs were much smaller because of the much larger dataset. We found slightly less divergence (dN and dS) between the polyploid genomes (n = 28,317) than between the diploid genomes (n = 30,874), although the difference is not significant. The different totals between the diploids and polyploids suggested that more than 2500 genes in the polyploid did not have sufficient polymorphism for an appropriate estimation of molecular evolution. We further investigated the alignments of these genes to ascertain whether their close sequence similarity was the result of gene conversion between homoeologous genomes. Of the genes without dN/dS estimates, 759 were found to only have sufficient polymorphisms between the tetraploid genomes (and not between the diploid genomes) and 3316 were found to have sufficient polymorphisms between the diploids (but not between the tetraploid genomes). This cumulative large difference between ploidy levels further suggested that gene conversions may play a role in reducing genetic diversity between genomes. However, only 106 and 42 genes were detected to overlap "conversion regions" in the diploids and tetraploid genome.

Of the 2,817,991 SNPs between diploids that fell within genes, 486,514 were inferred to be in exonic positions, including 248,599 that caused amino acid changes (i.e., nonsynonymous) compared to the reference sequence. Of these, there were 1651 genes with SNPs that resulted in premature stop codons in the pseudo-translation of A-genome transcripts, 1802 genes with premature stop codons in A_T, and 709 genes with premature stop codons in D_T (Figure 9). These genes were not excluded in estimates of molecular evolution. None of these gene sets had any enriched GO terms. The low level of D_T premature stops may simply reflect an ascertainment

bias of the annotated reference genome that was based on a diploid D genome. Many of the putative stop codons were found near the annotated end of the gene, suggesting that they might have only a minimal impact on protein function. Alternatively, their inference may reflect bioinformatic artifacts, such as imperfect gene annotation of the D-genome, or alternative stops that independently evolved in the A-genomes. Most of these alternative stops codons were within 10% of the 39 end of the gene. If one ignores the premature stop codons within the last 10% of the annotated genes, then 803 premature stops were shared between the diploid A-genomes and the A_T-genome (Figure 9). This result was marginally less than previously reported (Paterson et al. 2012), because we had the added power of multiple A-genome re-sequencing efforts.

Other SNPs disrupted a start or stop codon. We identified 806 genes with disrupted start codons (i.e., resulting in an amino acid distinct from that in the D_5 reference) in the A-genome, 703 in A_T, and 684 in D_T. These genes could have a longer or shorter coding sequence than as originally annotated. No GO terms were enriched in these gene sets. We also identified 831 genes with altered stop codons in the A-genome, 693 in A_T, and 437 in D_T, resulting in longer peptide sequences. Several GO terms (~20 in each genome) were enriched within these genes, and almost all were associated with photosynthesis. There were also 406 genes without a stop codon within the D_5 gene annotation, with the same photosynthesis GO terms being enriched.

Diversity and heterozygosity

In addition to creating an index of nucleotide differences between the diploid A-genomes and D-genomes, we detected unique nucleotide variation within and between individual accessions. Within a genome type (i.e., A or D), these types of SNPs are called allele-SNPs. The allelic genotype of each diploid and both genomes of the allopolyploid Maxxa were determined

at all polymorphic loci. A pairwise comparison between accessions found that the D_5 diploids had extremely low nucleotide diversity (1 million SNPs) between any two accessions, whereas a similar pairwise comparison between the A_1-genomes and A_2-genomes found that accessions were more diverse (4-5 million SNPs within A_1 or A_2; 6–8 million SNPs between A_1 and A_2). There were approximately twice as many SNPs between the A_1-genomes and A_2-genomes as within either of the two species. These results are not unexpected given the exceptionally low diversity found in a survey of allozyme diversity in *G. raimondii* (J. F. Wendel, unpublished data) and the appreciable levels of diversity in the chosen accessions of G. arboreum and *G. herbaceum* (Wendel et al. 1989).

In addition to having more fixed allele-SNPs between accessions, the A-genome diploids were more heterozygous than the D-genome diploids (13% and 1%, respectively; Table 6). In the A-genome diploids, heterozygous loci were approximately twice as frequent outside than inside of genes. This was not surprising, given the expectation of more intense purifying selection on coding sequences. Of course, these estimates of heterozygosity excluded loci that were duplicated in the A-genome. Interestingly, heterozygous loci in the D-genome diploids were equally common in genic and nongenic regions. This genomic difference likely reflects both the exceptionally low genetic diversity within the D-genome and a high level of generalized inbreeding. In this respect, we note that *G. raimondii* has a narrow natural range and presently exists as only scattered populations with very low effective population sizes.

A neighbor-joining tree was constructed and rooted based on the known relationship between the A/F-genome clade and the D-genome clade (Figure 10). Many fixed allele-SNPs (i.e., not heterozygous within a line) could be attributed to mutations occurring along specific branches of the phylogeny (Table 7). The tree correctly reconstructed the accepted relationships of the

diploids and their relationships to the two genomes of allopolyploid cotton (Senchina et al. 2003; Grover et al. 2012; Wendel et al. 2012). Specifically, and unsurprisingly, A_T and D_T were phylogenetically sister to the common ancestor of the [A_1 + A_2] clade and the D_5 clade, respectively. Our results agreed with previous reports that the A_1 or A_2 diploids were approximately twice as good of an approximation of the A_T-genome as the D_5 diploids were of the D_T-genome (Wendel and Cronn 2003; Senchina et al. 2003). The distance from D_T to D_5 was 0.14, and the distance from A_T to the common ancestor of A_1 and A_2 was 0.08. However, the distances from A_T to A_1 and A_2 were 0.10 and 0.11, respectively. Moreover, the distance from A_T to any individual A-genome diploid was 0.14, similar to the distance between D_T and any D-genome diploid. Although the group of A-genome diploids provided an approximation of A_T that was two-times better than a group of D_5 diploids did of D_T, any individual A or D diploid appeared to be equally similar to its A_T or D_T counterpart. The exceptionally low diversity among D_5 diploids explained the fact that a group of D_5- genome diploids was not significantly better for approximation of the D_T-genome than was a single D_5. However, multiple accessions of A diploids (A_1 and/or A_2) did provide a substantial (nearly two-times better) improvement in construction of the A_T-genome pseudo- sequence.

Duplications and deletions

Duplications were detected in A-genome diploids as coverage peaked across the D-genome reference sequence (Figure 8). Because these duplications were detected relative to the D-genome diploids, they represent events that occurred after the split of these two clades 5–10 million years ago (Wendel and Cronn 2003; Wendel et al. 2012). Thus, peaks shared by all A-

genome accessions represent pieces of the A-genome that are duplicated relative to the D-genome. These coverage peaks represent a mix of tandem and dispersed duplications, because the methodology used makes no distinction regarding genomic location of duplicated segments. There were 30,709 regions duplicated in all A-genome diploids but no duplication in D-genome diploids. These duplicated blocks overlapped 1007 genes, with a minimum overlap of 50% of the gene length. Only one GO term was enriched among these genes: structural constituent of ribosome (GO: 0003735).

In contrast to duplications, putatively deleted regions of the A-genome were detected with a higher degree of certainty because their diagnosis is based on lack of coverage rather than a quantitative difference in coverage (Figure 8). Some regions of the D-genome reference genome did not have any A-genome reads mapped to them, despite 40-times the WGS coverage based on the number of produced A-genome reads and correctly mapping D-genome reads to the same region. Each accession had a unique set of deleted regions, including genes. There were 25,408 regions deleted in all A-diploid genomes. The genomic regions included 978 annotated D-genome reference genes where the deleted region minimally overlapped 50% of the gene length. Among the genes within deleted regions, 118 GO terms were enriched compared to the population of GO terms within the annotated gene set (Paterson et al. 2012). Most of these deletion terms were associated with starch synthesis, tRNAs, or DNA repair mechanisms. Three hundred seventy-eight genes were completely deleted in the A-genome diploids relative to the D-genome diploids, meaning that the genic region was spanned by a single deletion block.

Polyploid conversion events

We used two different methods to detect conversion events in the allopolyploid genome

of Acala Maxxa (*G. hirsutum*). The first method identified historical nonreciprocal homoeologous recombination events (NRHR) at individual loci based on homoeo-SNPs within a polyploid genome. Earlier analyses in cotton used this method to detect conversion events (or NRHR events) based on comparative analysis of assembled EST sequences from diploid and allopolyploid cotton (Salmon et al. 2010; Flagel et al. 2012). The second method identified converted regions based on coverage patterns in the A_T- genomes and D_T-genomes relative to their respective diploid relatives. Here, we consider the NRHR events as "conversion" events regardless of whether they occur in coding or in noncoding sequences (*sensu amplo* of gene conversion).

Based on homoeo-SNPs (first method), 1,748,889 conserved SNPs in 29,576 genes in the diploid genomes suggested an A_T-biased allele conversion (a D_T nucleotide converted to the A_T nucleotide). In contrast, a total of 361,795 SNPs in 12,346 genes suggested a D_T-biased conversion. These data suggest a nearly five-fold bias based on homoeo-SNPs and a two-and-one-half–fold bias based on genes in favor of A_T-biased conversion, in stark contrast to the two-fold bias in favor of D_T-biased conversion reported previously (Paterson et al. 2012).

Based on coverage/deletion information (second method), conversion events were found in both directions across 882 Kbp of the D-genome reference sequence (Figure 8). These events ranged from 1 to 5470 bp, with a median of 337 bp. Two hundred fifty-nine regions suggested an A_T-biased conversion. These regions spanned 275 Kbp and overlapped the coding sequence of 19 genes. They also included 12,696 putative homoeo-SNPs (based on the diploids), none of which was detected within the tetraploid. However, 1213 regions showed a D_T-biased conversion. These regions spanned 607 Kbp and overlapped the coding sequence of 94 genes. They included 21,142 putative homoeo-SNPs, of which only three were also detected within the tetraploid. The

genes overlapped by these regions had no enriched GO terms and indicated a conversion bias in both directions, but with the D$_T$ direction most prominent, similar to that previously reported (Paterson et al. 2012). The events detected by the second method only included 1375 of the possibly conversion-related SNPs identified by the first method.

DISCUSSION

Genome resources for *Gossypium*

Using the diploid re-sequencing data, we created several useful resources for the *Gossypium* genome. First, a genome-wide map of the SNPs between the diploid A-genomes and D-genomes of cotton was created. Re-sequencing multiple accessions of each diploid enabled us to distinguish bases that were specific to a single accession from bases that are more representative of one diploid genome or another. It also allowed us to identify conserved genomic features shared by all A-genome or D-genome species and accessions. In that sense, the multiple accessions of each species acted as re-sequencing replications of the A-genome or D-genome "treatments." We have demonstrated that most SNPs identified between diploid genomes can be directly extrapolated to differences between the descendant allopolyploid genomes (i.e., homoeo-SNPs) because of their recent common ancestor. In addition, we also identified several million homoeo-SNPs that were unique to the Maxxa allopolyploid genome. These documented SNPs can be used for genome identification of individual sequence reads (Udall 2006; Flagel et al. 2012) or the development of genotyping assays (Van Deynze et al. 2009; Byers et al. 2012). With an index of homoeo-SNPs, read-mapping efficiency was significantly improved and future false-positive allele-SNPs can be filtered out of marker sets, resulting in more reliable allele-SNP assessment. In addition to the homoeo-SNPs, we also

identified allele- SNPs in the diploid cotton accessions. These SNP sets are available as Gbrowse tracks at CottonGen (http://www.cottongen.org) and as gff files at the Udall laboratory web site (http://udall-lab.byu.edu).

A second resource is the set of alignments of gene and protein sequences of the A-genomes, A_T-genomes, and D_T-genomes to accompany the previously published *G. raimondii* annotations (Paterson et al. 2012). These alignments were used to identify SNPs and to further refine our understanding of the molecular evolutionary differences between genomes. Because we have made this a public resource, any researcher investigating cotton now has homoeo-SNPs and allele-SNPs information for any target gene already identified. This simple, yet tedious, task has been a common obstacle of genetic research in polyploid cotton.

A third resource, and one that we suggest will be a fruitful topic for further investigation, is the description of putative duplications and deletions that distinguish the A-genomes and D-genomes and, hence, originated subsequent to their divergence from a common ancestor. These localized structural variations offer a rich source of sequences to mine for possible functional consequences, and to further our understanding of the mechanisms of copy number variation during genome evolution in plants.

Through our read-mapping efforts, we noticed that the limited number and the stochastic distribution of homoeo-SNPs could have implications for de novo genome assembly of polyploid cotton. Although 70% of the reads from allopolyploid cotton could be assigned to one of its two co-resident genomes, 30% of the reads that mapped to the D-genome reference did not overlap a homoeo-SNP and, hence, could not be categorized. Using an arbitrary length of 1000 bp, we found 47,399 unique loci where sequence reads of the A_T-genome and D_T-genome were indistinguishable when compared to each other and to the reference genome. Assuming sequence

read lengths less than 500 bp, these regions would likely co-assemble during a de novo whole-genome shotgun assembly with current read lengths. Consequently, co-assembled segments will create unique challenges of graph structure bifurcation (or higher branching) during the contig construction steps of de novo assembly. Part of this challenge could be addressed by generating reads with a greater likelihood of overlapping homoeo-SNPs, i.e., longer reads. Present data, however, suggest that de novo assembly of the allopolyploid cotton genome would not be successful if based on contemporary read lengths.

Insights into the genome biology of *Gossypium*

One of the intriguing results of this study is the insight it provides into the origin and frequency of indels during A-genome and D-genome divergence. Because of the lack of an outgroup sequence, none of the "duplications" or "deletions" described here is polarized, so their duplicate or deleted status is only relative to the single D-genome reference. Moreover, the methods used do not yield insights into the mechanistic underpinnings of the indels, which may conceivably entail a full spectrum of deletional mechanisms and processes of tandem and dispersed duplication.

Notwithstanding, the present study does reveal the scope and scale of the indel generating process during 5–10 million years of diploid evolution. Additionally intriguing are the genomic distributions of the duplicated and deleted regions. For example, chromosome 13 is notable for its high frequency of duplications, containing one-sixth (2850/17,102) of the total number of conserved duplications in the A-genome, yet only 2.9% (174/6072) of the deletions. Given that chromosome 13 comprises a mere 7.8% of the G. raimondii reference sequence (Paterson et al. 2012), the suggestion arises that there has been exceptional expansion and/or contraction of this

chromosome during the evolution of the two-fold size difference that distinguishes the A-genomes and D-genomes. Although some of this difference certainly reflects the expansion of transposable elements in the A-genome or possible contraction in the D-genome (Hawkins et al. 2009), it is unclear what genomic features have allowed more numerous rearrangements in chromosome 13 than within other chromosomes in the genome.

In addition to this broad-scale view of the contributions of duplications and deletions to *Gossypium* genome evolution, the data presented here offer a rich database that can be mined for potentially significant gene duplication and deletion. For example, gene loss has been associated with polyploidization (Shaked et al. 2001; Ozkan 2003; Han et al. 2005; Tate et al. 2009), but the deletions we have described in the A-genome occurred before polyploidization and include parts of approximately 1300 genes per accession. If these A-genome accessions were used as a "parental genome reference" for investigations of polyploidy, then the deletions common to the A_1-genomes, A_2- genomes, and ancestral A_T-genomes would be confounded with any putative deletions that occurred as a result of polyploidization. Thus, this initial database of duplications and deletions will be a useful research tool for investigations of the evolution of the *Gossypium* genome.

We observed that several genes involved in starch synthesis were deleted in the A-genome diploids, including seven genes with 1,4-alpha-glucan branching enzyme activity. It is tempting to speculate that these deletions increased the amount of glucose available in the A-genome diploid for cellulose synthesis and thereby played a role in the increased length of mature A-genome cotton fibers. Previous studies have documented altered carbon partitioning (Yong-Ling Ruan 1998) and altered starch accumulation (Chaohua et al. 2005) in fiberless *G. hirsutum* mutants. The deletion of starch genes in the A-genome may have been associated with

the opposite effect, resulting in more carbon being allocated to cellulose production and less to starch production. We caution that many of the deleted genes are members of gene families and the remaining paralogs may partially or fully compensate for their deletion in the A-genome diploids. Nevertheless, the deleted genes discovered in this study offer interesting avenues of future research of gene duplication and functional compensation.

Among genes with altered stop codons, we detected an enriched number of genes having photosynthesis-related functions. It appears unlikely that the altered stop codons are attributable to horizontal transfer of chloroplast genes to the nucleus because only four have high similarity to chloroplast genes. Although a biological explanation for this enrichment remained a mystery, it was likely that a portion of the enrichment of photosynthesis GO terms was an artifact of the gene annotation process. For example, the actual stop codon location may have been ambiguous because the original annotation of these genes actually had no stop codon. Perhaps, the initial gene annotation effort was simply unable to identify the full coding sequence and subsequent updates will include corrections to the original annotations. Regardless of the initial annotation, the enrichment of photosynthesis GO terms in genes with altered stop codons was interesting, but it was not attributable to differences between the A-genomes, D-genomes, A_T-genomes, and D_T-genomes.

We also identified homogenized genome regions from conversion events between the two homoeologous genomes of *G. hirsutum*. Gene conversion has been defined as the nonreciprocal transfer of genetic information between homologous sequences, leading to homogenization during meiotic or mitotic recombination (Szostak et al. 1983; Chen et al. 2007; Hsu et al. 2010; Jacquemin et al. 2011). Unlike analyses of pairs of genes (Drouin 2002; Mondragon-Palomino and Gaut 2005; Xu et al. 2008), whole-genome analysis of gene

conversion has been pioneered in rice (Xu et al. 2008; Wang et al. 2009; Jacquemin et al. 2011), a sequenced diploid genome with many closely related species also with sequenced genomes. By comparison of the diploid re-sequencing data to the publicly available WGS data of Maxxa Acala, we were able to identify conversion events between homoeologous sequences within the polyploid *Gossypium* genome using two different methods. The two methods of detecting homoeologous conversion events resulted in different directional biases.

Using the first method, we had previously used SNPs to estimate that up to 5% of the polyploid transcriptome had experienced "homoeologous gene conversion" (Salmon et al. 2010; Flagel et al. 2012). In both previous studies, identification of autapomorphic SNPs was not possible because of limited diploid sequencing data. Based on our current data, the presence of autapomorphic SNPs (and a liberal method of identification) appeared to have caused an overestimate in the amount of homoeologous conversion in genic regions. Thus, the genome sequence of a definitive outgroup is needed to unambiguously identify regions of conversion using SNP information alone. One dimension of the conversion events and the multi-alignment resource for all genes is the identification of loci where one of the two allopolyploid genomes has "overwritten" the other via a mechanism of reciprocal or nonreciprocal "gene conversion." At present, the functional consequences of these observations remain unexplored, but it is intriguing to ask whether these conversion events are functionally insignificant (which might, for example, be the case when only synonymous sites are involved) or if, instead, specific genes or regulatory sequences have been selectively "doubled" or "eliminated" by this unusual intergenomic aspect of allopolyploid speciation.

This first method contains an inherent bias in favor of A_T-biased conversion because of the greater genetic distance between D_T and D_5 compared to the distance between A_T and A_1 or

A_2. This bias, however, should only be approximately 50% based on our understanding of the genetic distances between A and A_T vs. D and D_T (the latter is 50% greater) (Flagel et al. 2012). In addition, the genotype pattern indicative of a conversion is indistinguishable from that caused by an autapomorphic SNP in the diploid species. For example, suppose an autapomorphic SNP in the A-genome ancestor of A_1 and A_2 (not shared with A_T) changes a C to a T at a given locus. Consequently, the D-genome diploids, D_T-genome, and A_T-genome would all have a C, whereas the A-genome diploids would have a T. So, the A_T nucleotide would appear the same as D and D_T, suggesting a D_T- biased allele conversion event, even though they simply shared the ancestral allele. These confounding autapomorphic SNPs would have occurred after the divergence of A_T but before the A_1–A_2 split, suggesting a D_T-biased conversion. However, an A_T-biased conversion would be suggested by an autapomorphic SNP occurring after the divergence of D_T, but before the most recent common ancestor of the D_5 diploids. Because the A-genome diploid is approximately a two-fold better approximation of the actual progenitor diploid than is the D-genome diploid (Wendel et al. 2012), these branch lengths are different (Figure 10). In fact, the phylogeny showed a distance of only 0.00447 for the branch corresponding to an autapomorphic SNP shared by all A diploids but not by A_T. However, the equivalent branch in the D-genome clade has an autapomorphic SNP distance of 0.05385. These numbers suggest that, in the absence of any actual conversion events, there should be more than 12-times (0.05385/ 0.00447) as many SNPs that look like A_T-biased conversion—because they are shared by A diploids but not by A_T—as SNPs that look like D_T- biased conversion. The difference between the 12-fold expected value for A_T-biased conversions as visualized by the branch lengths to the five-fold observed value could be explained by a bias toward D_T-biased

conversion events, as reported elsewhere and as detected by the second coverage-based method. Thus, we consider the conversion events detected by the SNP-based method (method 1) to be inaccurate based on autapomorphic SNPs, but we consider the conversion events of the coverage method (method 2) to be a conservative, yet relatively accurate, assessment of conversion between the polyploid *G. hirsutum* genomes.

The second method of conversion detection used deletion and coverage information to detect many separate events, and the direction of bias agreed with previous reports (Paterson et al. 2012). This method was very conservative and may represent a minimum amount of conversion events in the polyploid genome because of the uncertainty of the actual endpoints of conversion and the additional amounts of conversion suggested between homoeologous gene copies in the dN/dS analysis. The conversion events resulted in a loss of genomic diversity between the A$_T$- and D$_T$-genomes. Parts of at least 113 genes were included in conversion events between homoeologous chromosomes. Other investigations of genome evolution in rice have uncovered convergent evolution of ancient paralogs on Chr11 and Chr12 ($2.1 Mb) mediated by gene conversion including up to 180 genes (Jacquemin et al. 2011). The conversion events we have described were more recent (after polyploidization 1–2 Ma) and our inference space was limited to a single species of *G. hirsutum*. It will be interesting to see if other polyploid *Gossypium* genomes also have the same conversion events in their genomes, and to estimate the rate of gene conversion between homoeologous genomes.

Whole-genome re-sequencing of diploid *Gossypium* has identified insights into the genome evolution of cotton. These insights proved to be useful for characterization of the *G. hirsutum* genome via publicly available re-sequencing data. Additional de novo and re-sequencing efforts of polyploid *Gossypium* will continue to add to our understanding of the

cotton genome, thereby enhancing our ability to manipulate the fiber and agronomic characteristics of cotton.

REFERENCES

1. Brubaker, C. L., A. H. Paterson, and J. F. Wendel, 1999 Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome 42: 184–203.
2. Byers, R. L., D. B. Harker, S. M. Yourstone, P. J. Maughan, and J. A. Udall, 2012 Development and mapping of SNP assays in allotetraploid cotton. Theor. Appl. Genet. 124: 1201–1214.
3. Chaohua, C., W. Xiyuan, and N. I. Xiyuan, 2005 Observation of fiber ultrastructure of Ligon lintless mutant in upland cotton during fiber elongation. Chin. Sci. Bull. 50: 126–130.
4. Chen, J.-M., D. N. Cooper, N. Chuzhanova, C. Férec, and G. P. Patrinos, 2007 Gene conversion: mechanisms, evolution and human disease. Nat. Rev. Genet. 8: 762–775.
5. Chia, J.-M., C. Song, P. J. Bradbury, D. Costich, N. de Leon et al., 2012 Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44: 803–807.
6. Conesa, A., S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon et al., 2005 Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.
7. Drouin, G., 2002 Characterization of the gene conversions between the multigene family members of the yeast genome. J. Mol. Evol. 55: 14–23.
8. Felsenstein, J., 1989 PHYLIP–Phylogeny Inference Package (version 3.2). Cladistics 5: 163–166.
9. Flagel, L. E., J. F. Wendel, and J. A. Udall, 2012 Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. BMC Genomics 13: 302.

10. Grover, C. E., and J. F. Wendel, 2010 Recent insights into mechanisms of genome size change in plants. J. Bot. 2010: 1–8.
11. Grover, C. E., K. K. Grupp, R. J. Wanzek, and J. F. Wendel, 2012 Assessing the monophyly of polyploid *Gossypium* species. Plant Syst. Evol. 298: 1177–1183.
12. Han, F., G. Fedak, W. Guo, and B. Liu, 2005 Rapid and repeatable elimination of a parental genome-specific DNA repeat (pGc1R–1a) in newly synthesized wheat allopolyploids. Genetics 170: 1239–1245.
13. Han, M. V., and C. M. Zmasek, 2009 phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinformatics 10: 356.
14. Hawkins, J. S., H. Kim, J. D. Nason, R. A. Wing, and J. F. Wendel, 2006 Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res. 16: 1252–1261.
15. Hawkins, J. S., S. R. Proulx, R. A. Rapp, and J. F. Wendel, 2009 Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proc. Natl. Acad. Sci. USA 106: 17811–17816.
16. Hsu, C.-H., Y. Zhang, R. C. Hardison, E. D. Green, and W. Miller, 2010 An effective method for detecting gene conversion events in whole genomes. J. Comput. Biol. 17: 1281–1297.
17. Jacquemin, J., C. Chaparro, M. Laudie, A. Berger, F. Gavory et al., 2011 Long-range and targeted ectopic recombination between the two homeologous chromosomes 11 and 12 in *Oryza* species. Mol. Biol. Evol. 28: 3139–3150.
18. Krzywinski, M., J. Schein, İ. Birol, J. Connors, R. Gascoyne et al., 2009 Circos: an information aesthetic for comparative genomics. Genome Res. 19: 1639–1645.

19. Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan et al., 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

20. Mondragon-Palomino, M., and B. S. Gaut, 2005 Gene conversion and the evolution of three leucine-rich repeat gene families in *Arabidopsis thaliana*. Mol. Biol. Evol. 22: 2444–2456.

21. Ozkan, H., 2003 Nonadditive changes in genome size during allopolyploidization in the wheat (*Aegilops-triticum*) group. J. Hered. 94: 260–264.

22. Page, J. T., A. R. Gingle, and J. A. Udall, 2013 PolyCat: A resource for genome categorization of sequencing reads from allopolyploid organisms. G3 (Bethesda) 3: 517–525.

23. Paterson, A. H., J. F. Wendel, H. Gundlach, H. Guo, J. Jenkins et al., 2012 Repeated polyploidization of *Gossypium* genomes and the evo- lution of spinnable cotton fibres. Nature 492: 423–427.

24. Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

25. Salmon, A., L. Flagel, B. Ying, J. A. Udall, and J. F. Wendel, 2010 Homoeologous nonreciprocal recombination in polyploid cotton. New Phytol. 186: 123–134.

26. Senchina, D. S., I. Alverez, R. C. Cronn, B. Liu, J. Rong et al., 2003 Rate variation among nuclear genes and the age of polyploidy in *gossypium*. Mol. Biol. Evol. 20: 633–643.

27. Shaked, H., K. Kashkush, H. Ozkan, M. Feldman, and A. A. Levy, 2001 Sequence elimination and cytosine methylation are rapid and re- producible responses of the genome to wide hybridization and allopolyploidy in wheat. Plant Cell 13: 1749–1759.

28. Szostak, J. W., T. L. Orr-Weaver, R. J. Rothstein, and F. W. Stahl, 1983 The double-strand-break repair model for recombination. Cell 33: 25–35.

29. Tate, J. A., P. Joshi, K. A. Soltis, P. S. Soltis, and D. E. Soltis, 2009 On the road to diploidization? Homoeolog loss in independently formed populations of the allopolyploid *Tragopogon miscellus* (Asteraceae). BMC Plant Biol. 9: 80.

30. Udall, J. A., 2006 A global assembly of cotton ESTs. Genome Res. 16: 441– 450.

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

32. Wang, X., H. Tang, J. E. Bowers, and A. H. Paterson, 2009 Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. Genome Res. 19: 1026–1032.

33. Wendel, J. F., 1989 New World tetraploid cottons contain Old World cytoplasm. Proc. Natl. Acad. Sci. USA 86: 4132–4136.

34. Wendel, J. F., and Cronn, R. C., 2003 Polyploidy and the evolutionary history of cotton, pp. 139–186 in Advances in Agronomy, ed. D. Sparks, Elsevier, New York, NY.

35. Wendel, J. F., L. E. Flagel, and K. L. Adams, 2012 Jeans, genes, and genomes: cotton as a model for studying polyploidy, pp. 181–207 in Polyploidy and Genome Evolution, edited by P. S. Soltis, and D. E. Soltis. Springer, New York.

36. Wendel, J. F., and P. D. Olson, and J. McD. Stewart, 1989 Genetic diversity, introgression, and independent domestication of old world cultivated cottons. Am. J. Bot. 76: 1795–1806.

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

38. Xu, S., T. Clark, H. Zheng, S. Vang, R. Li et al., 2008 Gene conversion in the rice genome. BMC Genomics 9: 93.

39. Yong-Ling Ruan, P. S. C., 1998 A fiberless seed mutation in cotton is associated with lack of fiber cell initiation in ovule epidermis and alterations in sucrose synthase expression and carbon partitioning in developing seeds. Plant Physiol. 118: 399.

40. Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson et al., 2008 Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9: R137.

TABLES

Table 3 Transitions and transversions in the homoeo-SNP index

	A	G	C	T
A	-	2,495,527	626,075	1,003,583
G	2,547,739	-	353,034	647,148
C	644,840	352,239	-	2,544,261
T	1,003,739	628,050	2,492,619	-

Rows = A allele. Columns = D allele. There was an overall transition/transversion ratio of $1.92 and GC fractions of 45.4% (A genome) and 45.1% (D genome).

Table 4 Summary of nine diploid WGS re-sequencing libraries that were re-sequenced in this study and additional libraries (A1_97, F1_1, and Maxxa) obtained from the SRA

Accession	PI	Raw Pairs	Trimmed Reads	Raw Mapping %	Mapped % Using SNP Index 2.0
A1_155	630024	385,657,228	761,269,884	65.3	78.0
A1_73	485587	202,723,343	238,035,929	53.1	85.6
A1_97	529670	328,713,056	652,350,335	65.0	77.8
A2_1011	629339	412,420,252	816,274,495	58.3	73.8
A2_255	615756	300,406,057	595,289,591	61.1	75.5
A2_34	183160	367,844,399	729,370,248	62.3	76.5
A2_44	185788	78,180,657	153,728,823	63.3	76.9
A2_4	529707	343,470,023	686,940,046	48.9	82.4
D5_2	530899	152,913,856	304,706,886	95.6	95.3
D5_31	530928	217,334,954	428,323,703	95.8	95.5
D5_4	530901	310,387,080	616,432,521	95.1	94.8
D5_53	530950	188,469,224	375,193,268	96.2	96.0
F1_1	530986	534,258,839	1,055,751,863	71.1	79.1
Maxxa Acala	540885	463,761,132	919,898,042	72.5	79.8

Table 5 Amount of molecular evolution between the A and D genomes of cotton

		dN	dS	dN/dS
A vs. D	Mean	0.0094	0.0276	0.3726
N = 28,462	Median	0.0068	0.0256	0.2768
	SD	0.0106	0.0225	0.4236
A_T vs. D_T	Mean	0.0092	0.0266	0.3772
N = 26,156	Median	0.0066	0.0237	0.2843
	SD	0.0104	0.0228	0.4156

Table 6 Number of heterozygous loci in each accession, along with the percentage of total observable loci that were heterozygous

	Whole Genome		Genic Loci Only		Nongenic Loci Only	
Accession	n	%	N	%	n	%
F1_1	9,968,998	17.2	332,247	6.1	9,636,751	18.4
A1_73	2,963,374	7.1	126,260	2.6	2,837,114	7.7
A1_97	6,504,768	12.4	265,607	5.0	6,239,161	13.2
A1_155	7,549,531	13.9	322,095	6.0	7,227,436	14.8
A2_4	7,061,224	13.2	283,151	5.3	6,778,073	14.1
A2_34	6,826,660	12.9	270,384	5.1	6,556,276	13.7
A2_255	5,898,387	11.6	236,113	4.5	5,662,274	12.4
A2_1011	6,878,801	13.1	252,230	4.9	6,626,571	14.1
D5_2	193,418	0.3	20,536	0.4	172,882	0.3
D5_4	257,399	0.4	25,370	0.5	232,029	0.4
D5_31	178,290	0.3	20,723	0.4	157,567	0.3
D5_53	181,224	0.3	20,665	0.4	160,559	0.3
Maxxa.A	4,465,088	9.2	198,477	3.9	4,266,611	9.8
Maxxa.D	686,674	1.3	47,861	1.0	638,813	1.4

Table 7 SNPs attributable to specific areas of the phylogeny

Genome	SNPs	Deletions
All A	5,544,440	25,408
A_1	1,024,299	3809
A_2	1,152,825	2941
A_T	1,472,900	5247
All D	14,601,331	0
D_T	3,563,979	4518

As shown in Figure 10. Because of possible conversion events, it is not possible to determine how many SNPs were shared by all A or D diploids.

FIGURES

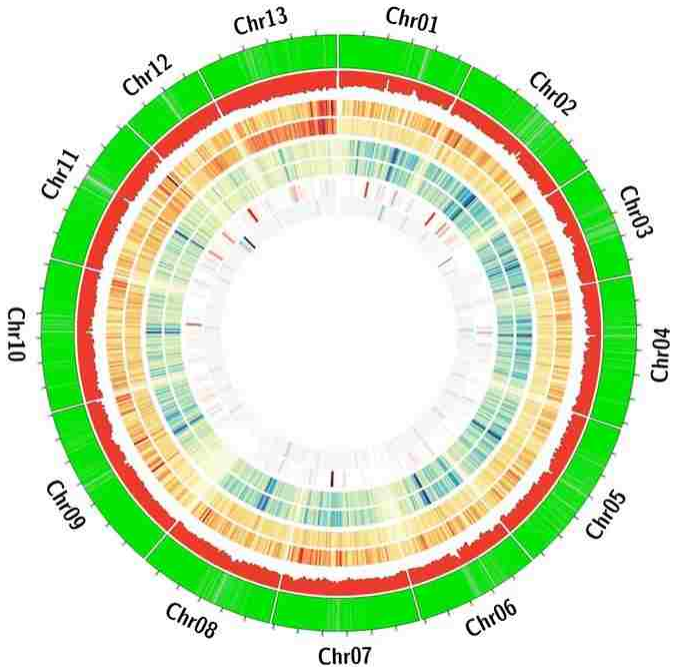


Figure 8 Plot of genes, homoeo-SNPs, duplications, deletions, and conversion events in the A-genomes, relative to the D_5 reference sequence, produced by Circos (Krzywinski et al. 2009). Considering the concentric circles from the outside inward, the outermost (and first) green circle indicates the location of annotated genes. The next circle (red) is a histogram of the number of homoeo-SNPs in a 1-Mbp window throughout the genome. The next two red (high-frequency) to yellow (low-frequency) circles are heat maps showing the location of duplications in the A_1 and A_2 genomes as compared to the D_5 genome (A_2 interior). The next two blue (high-frequency) to yellow (low-frequency) circles are heat maps showing the location of deletions in the A_1 and A_2 genomes as compared to the D_5 genome (A_2 interior). The final two circles show conversion events in the tetraploid G. hirsutum cv. Maxxa. The first circle shows conversion of loci to the A nucleotide on a red-to-yellow scale, whereas the innermost (and last) circle shows conversion of loci to the D nucleotide on a blue-to-yellow scale.

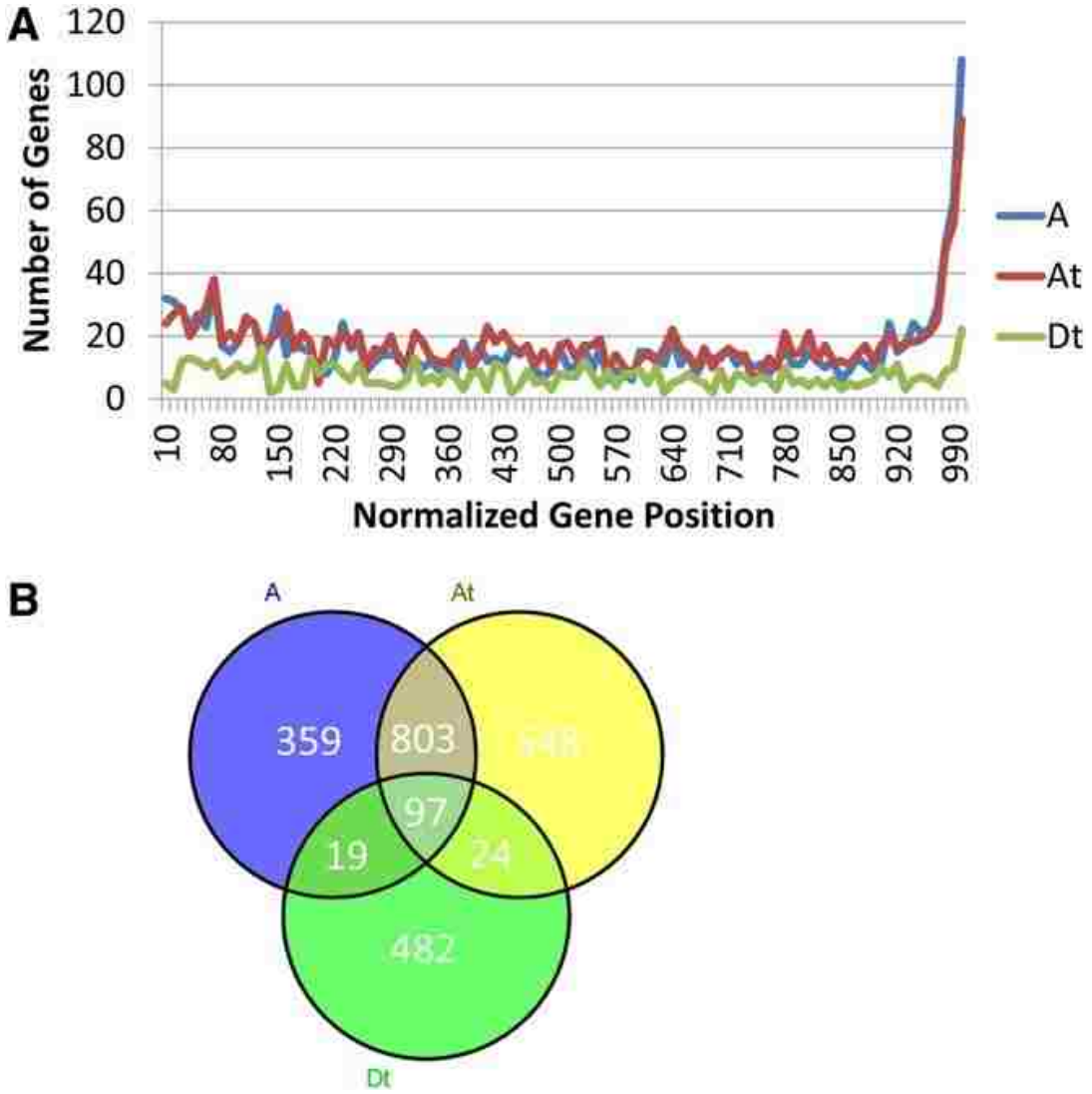


Figure 9 Premature stop codons found in each *Gossypium* genome. (A) Premature stop codons (compared to the annotations of the D-reference genome) were found in the A, A_T, and D_T genomes. (B) Common genes with premature stop codons in the first 90% of the gene.

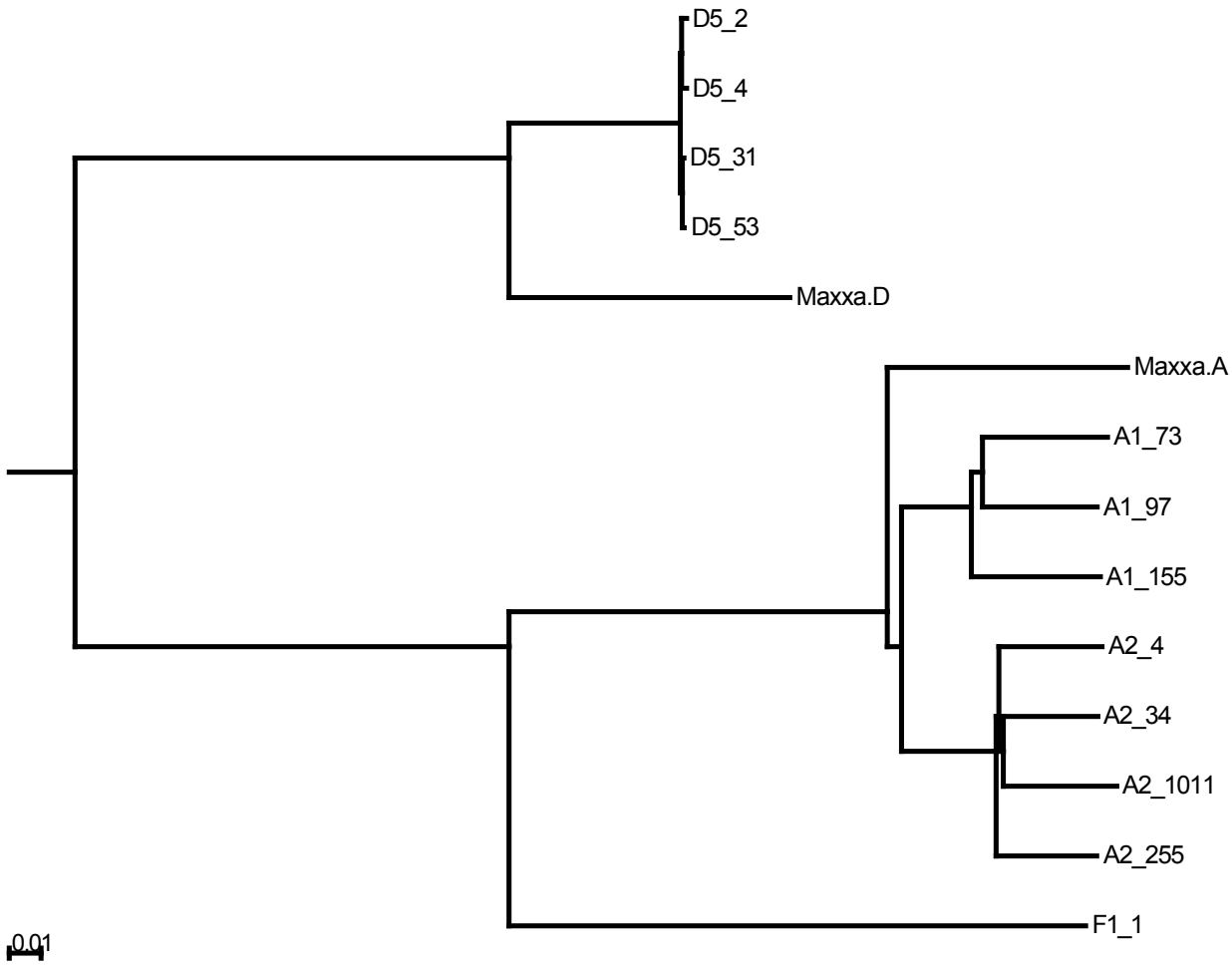


Figure 10 Neighbor-joining tree built by PHYLIP, based on SNPs between genomes. Units (as measured by the indicated scale) are percentage of represented polymorphic sites that differed between two individuals. Image rendered by Archaeopteryx (Han and Zmasek 2009).

CHAPTER 3

BamBam: Genome Sequence Analysis Tools for Biologists

FINDINGS

Massive amounts of data are involved in genome sequence research, requiring researchers to use supercomputing clusters and complex algorithms to analyze their sequence data. Genomic analyses frequently include next-generation sequencing to produce millions of short reads, followed by aligning of reads to a reference genome sequence with software like GSNAP and Bowtie 2 (Wu and Nacu 2010; Langmead and Salzberg 2012). These programs generate SAM files, the accepted standard for storing short read alignment data, which are subsequently compressed to BAM format via SAMtools (Li et al. 2009). The BAM files must then be analyzed and compared to produce meaningful results. Here we expand on the body of tools for analyzing and comparing BAM files.

We present BamBam, a package of bioinformatics tools to carry out a variety of genomic analyses on BAM files (Table 2). The included tools perform such tasks as counting the number of reads mapped to each gene in a genome (as for gene expression analyses), identifying SNPs (Single Nucleotide Polymorphisms) and CNVs (Copy Number Variants), and extracting consensus sequences. The purpose of BamBam is to provide a consistent framework to perform common tasks, without requiring extensive knowledge of computation or algorithms to select or interpret appropriate parameters.

The BamBam package includes several independent programs, briefly described below. The latest version of PolyCat is also included (Page, Gingle, and Udall 2013). The README in the download package provides example commands for various common analyses, including

phylogeny inference, molecular evolution estimation, methylation analysis, and differential expression analysis.

Single Nucleotide Polymorphisms

InterSnp calls SNPs between samples, represented by separate BAM files. InterSnp examines each position in the genome, assigning consensus alleles to each site for each sample. A SNP is called whenever two samples differ at the same position, producing a table with the genotypes of all samples at all polymorphic sites. The output is a table with the sequence name, position, and genotype for each sample at that site on each row, which can be readily processed by common command-line programs or scripts to calculate statistics or produce marker data for other programs.

Pebbles imputes genotypes using the K-nearest neighbor algorithm (Rutkoski et al. 2013; Troyanskaya et al. 2001). For each unknown genotype, Pebbles finds the samples that are most similar at nearby loci. Then it assigns a genotype to the unknown locus based on the weighted contributions of those neighbors. Pebbles operates on InterSnp output—a table of genotypes—and produces a file of the same form.

HapHunt uses K-means clustering to solve the haplotype-phasing problem, which consists of identifying all haplotypes in a sampled individual or population. Many programs have attempted to solve haplotype phasing and the closely related haplotype assembly problems using a variety of strategies, including Max-Cut, hidden Markov models, and dynamic programming (Bansal and Bafna 2008; B. L. Browning and Browning 2009; D. He and Eskin 2012). The K-means clustering algorithm (Figure 1) is an unsupervised machine learning algorithm, and is mathematically equivalent to Principle Component Analysis (Lloyd 1982; Ding and He 2004).

HapHunt first selects K reads distant from one another to serve as haplotype seeds. It assigns each other read to the haplotype with the closest consensus sequence. Then it recalculates the consensus sequences based on the reads in each haplotype and repeats the process of assigning each read to the haplotype with the closest consensus sequence. It repeats this process a given number of times, calculating a score at the end of each round based on the difference of the smallest interhaplotype distance and greatest intrahaplotype distance. This score favors clusterings in which haplotypes are individually compact and most distinct from one another. This score can optionally be scaled by the average size ratio for each pair of haplotypes, favoring clusterings that are more evenly divided. The consensus sequences of the final haplotypes are printed as an aligned FASTA file for each sequence in the original reference.

Copy Number Variants

Gapfall identifies large deletions between samples based on read coverage. It searches the genome for extended regions that have high coverage in one sample but no coverage in the other. A large region with no coverage could indicate a physical deletion (for genomic samples) or a deactivated gene (for RNA-seq). These putative deletions are reported as an annotation file that can be visualized with a genome browser such as IGV (Thorvaldsdóttir, Robinson, and Mesirov 2013).

Eflen identifies and extracts regions in a BAM file that are covered by at least a user-specified number of reads and outputs those regions as a GFF file. Provided with multiple BAMs, Eflen will identify regions that are covered in at least a user-specified fraction of those BAMs. This tool can be especially useful for analyzing GBS or RNA-seq data.

HMMph identifies CNVs between samples based on read coverage. BAM files must be provided for a control and for the sample of interest. The coverage ratio between those two BAM files is normalized by the total read coverage. Then the copy number of each locus in a sliding window is modeled based on a Poisson distribution in an untrained Hidden Markov Model (Zhao et al. 2013; Rabiner 1989).

Bisulfite-sequence Analysis

Bisulfite treatment converts unmethylated cytosines to thymines. **MetHead** summarizes methylation at all cytosine positions in the genome, based on BAM files of mapped bisulfite-treated reads. It totals the number of mapped cytosines and thymines at each position (indicating methylated and unmethylated states, respectively), then performs a one-tailed binomial test for the methylation of that site.

Different protocols are used for bisulfite treatment. If PCR is not performed after bisulfite treatment but before sequencing, then only 2 possibilities exist: conversions on the forward and reverse strand. But if PCR is performed, 4 possibilities exist (Figure 2). To properly count the number of cytosines and thymines in the 4-possibility protocol, the origin of the pre-PCR DNA fragment must be inferred. MetHead determines this—if necessary—by counting the number of C->T conversions and G->A conversions (indicative of a conversion on the reverse strand). It generates a BAM file with the orientation of each read matching its origin strand. That BAM can then be analyzed as if it were data produced by the 2-possibility protocol. Note that, in the produced BAM, the orientation of reads is not based on the direction in which the read was sequenced. Instead, the orientation of the read indicates the type of conversion caused by bisulfite treatment: C->G or A->T.

GeneVisitor

It is often useful to be able to compute on specific genomic intervals, such as genes. GeneVisitor provides a quick and easy way to do this, using an annotation file (GFF or BED format) to call a function on each indicated region of the genome. This class can be used by C++ programmers to run custom functions. In addition, pre-built tools utilize GeneVisitor without the need for programming.

Bam2Consensus converts one or more BAM files into a series of FASTA-formatted consensus sequences. If desired, multiple sequences—essentially unphased haplotypes—can be produced per BAM file, facilitating analyses of heterozygosity, nucleotide diversity, and molecular evolution. Suppose you have several BAM files representing different accessions of a species, all mapped to a common genome reference sequence. With a single command, Bam2Consensus can produce an aligned FASTA file for each gene, each containing the consensus sequences for each accession.

Bam2Fastq extracts mapped or unmapped reads from a BAM file, or from select regions of the BAM file.

Counter summarizes the number of reads mapped to each annotated region in one or more BAM files. RPKM (Reads Per Kilobase per Million mapped reads) normalization can be applied if desired. The output of Counter is a table of features and read counts, ready to be imported into EdgeR for differential expression analysis (M. D. Robinson et al. 2009).

SubBam extracts a subset of a BAM file. It can optionally modify the BAM file, changing the coordinates of mapped reads to match a new reference that is a subset of the original reference. Suppose you have WGS reads mapped to a reference sequence and are

interested in several loci. SubBam can produce BAMs that only contain the loci of interest, with a coordinate system corresponding to the position in the locus, rather than in the genome as a whole.

Allopolyploid Analysis

The latest version of **PolyCat** is included in BamBam. PolyCat uses an index of known homoeo-SNPs (polymorphisms that distinguish the genomes of an allopolyploid) to identify the source genome for each read in a library, which cannot be distinguished through typical next-generation sequencing protocols (Page, Gingle, and Udall 2013).

The MultiIndex class is used by PolyCat and MetHead, and can be used to make novel tools in C++. The MultiIndex is appropriate for random access to hundreds of millions of individual base positions in a genome sequence. It provides quick random access to base positions scattered across a genome sequence. Each sequence in the reference is indexed with a linked-list, with an index of landmark nodes spaced along the sequence at a resolution specified by the user (Figure 3).

Scripts

In addition to the core tools mentioned above, BamBam includes many Perl scripts, many of which use BioPerl modules (Stajich et al. 2002). Script functions include calculation of nucleotide diversity (π) and molecular evolution rates (Ka and Ks), paralog identification, differential expression with EdgeR (M. D. Robinson et al. 2009), summarization of results from MetHead, and summarization of genotype tables produced by InterSnp and Pebbles.

Conclusions

The BamBam tools form a simple interface between the researching biologist and the wealth of data contained in next-generation sequence alignments. They provide a means to efficiently identify interesting genomic features and summarize data, facilitating many next-generation sequence analysis experiments.

BamBam is freely available under the MIT license at http://sourceforge.net/projects/bambam/. It depends on both SAMtools and BAMtools (Li et al. 2009; Barnett, Garrison, and Quinlan 2011).

REFERENCES

1. Bansal, Vikas, and Vineet Bafna. 2008. “HapCUT: an Efficient and Accurate Algorithm for the Haplotype Assembly Problem.” http://bioinformatics.oxfordjournals.org/content/24/16/i153.short.
2. Barnett, D W, E K Garrison, and A R Quinlan. 2011. “BamTools: a C++ API and Toolkit for Analyzing and Managing BAM Files."
3. Browning, Brian L, and Sharon R Browning. 2009. “A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals.” *The American Journal of Human Genetics* 84 (2) (February): 210–223. doi:10.1016/j.ajhg.2009.01.005. http://linkinghub.elsevier.com/retrieve/pii/S0002929709000123.
4. Ding, Chris, and Xiaofeng He. 2004. “K-Means Clustering via Principal Component Analysis.” In, 29. New York, New York, USA: ACM Press. doi:10.1145/1015330.1015408. http://portal.acm.org/citation.cfm?doid=1015330.1015408.
5. Drummond, A J, B Ashton, S Buxton, and M Cheung. 2011. “Drummond: Geneious V5. 4 - Google Scholar.”
6. He, D, and E Eskin. 2012. “Hap-seqX: Expedite Algorithm for Haplotype Phasing with Imputation Using Sequence Data.” *Gene*. http://www.sciencedirect.com/science/article/pii/S0378111912015892.

7. Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4) (March 4): 357–359. doi:10.1038/nmeth.1923. http://www.nature.com/nmeth/journal/v9/n4/abs/nmeth.1923.html.
8. Li, H, B Handsaker, A Wysoker, T Fennell, and J Ruan. 2009. “The Sequence Alignment/Map Format and SAMtools." Bioinformatics 25: 2078–2079.
9. Lloyd, S. 1982. “Least Squares Quantization in PCM.” *IEEE Transactions on Information Theory* 28 (2) (March): 129–137. doi:10.1109/TIT.1982.1056489. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1056489.
10. Page, J T, A R Gingle, and J A Udall. 2013. “PolyCat: a Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms.” *G3 (Bethesda)* 3 (3) (March 8): 517–525. doi:10.1534/g3.112.005298.
11. Rabiner, Lawrence R. 1989. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.” *Proceedings of the IEEE* 77 (February 1): 1–30. http://axon.cs.byu.edu/~martinez/classes/678/Papers/rabinerHMM.pdf.
12. Robinson, M D, M D Robinson, D J McCarthy, D J McCarthy, G K Smyth, and G K Smyth. 2009. “edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1) (December 22): 139–140. doi:10.1093/bioinformatics/btp616. http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp616.
13. Rutkoski, Jessica E, Jesse Poland, Jean-Luc Jannink, and Mark E Sorrells. 2013. “Imputation of Unordered Markers and the Impact on Genomic Selection Accuracy..” *G3 (Bethesda, Md.)* 3 (3) (March): 427–439. doi:10.1534/g3.112.005363.

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23449944&retmode=ref&cmd=prlinks.

14. Stajich, J E, D Block, K Boulez, and S E Brenner. 2002. “The Bioperl Toolkit: Perl Modules for the Life Sciences.” *Genome Res*. 12: 1611–1618.
15. Thorvaldsdóttir, Helga, James T Robinson, and Jill P Mesirov. 2013. “Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration.” *Briefings in Bioinformatics* 14 (2) (March): 178–192. doi:10.1093/bib/bbs017. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22517427&retmode=ref&cmd=prlinks.
16. Troyanskaya, O, M Cantor, G Sherlock, P Brown, T Hastie, R Tibshirani, D Botstein, and R B Altman. 2001. “Missing Value Estimation Methods for DNA Microarrays.” *Bioinformatics* 17 (6) (June): 520–525. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=11395428&retmode=ref&cmd=prlinks.
17. Wu, T D, and S Nacu. 2010. “Fast and SNP-Tolerant Detection of Complex Variants and Splicing in Short Reads.” *Bioinformatics*.
18. Zhao, Min, Qingguo Wang, Quan Wang, Peilin Jia, and Zhongming Zhao. 2013. “Computational Tools for Copy Number Variation (CNV) Detection Using Next-Generation Sequencing Data: Features and Perspectives.” *BMC Bioinformatics* 14 (Suppl 11): S1. doi:10.1186/1471-2105-12-267. http://www.biomedcentral.com/1471-2105/14/S11/S1.

TABLES

Table 8 The core independent tools of BamBam

Section	Tool	Purpose
Single Nucleotide Polymorphisms	InterSnp	Call SNPs between two or more samples
	Pebbles	Impute genotypes in output from InterSnp
	HapHunt	Phase haplotypes with K-means
Copy Number Variants	GapFall	Identify deletions between two samples
	Elfen	Identify covered regions
	HMMph	Call copy number variants with HMM
Bisulfite-sequence Analysis	MetHead	Summarize base pair methylation in bisulfite-sequence data
GeneVisitor	Bam2Consensus	Generate consensus sequences from one or more samples
	Bam2Fastq	Extract mapped and unmapped reads from BAM files
	Counter	Summarize read coverage of sequences or regions
	SubBam	Extract subset of mapped reads
Allopolyploid Analysis	PolyCat	Categorize reads by genome based on similarity to parents
Scripts	Various	Various

CHAPTER 4

Methods for Mapping and Categorization of DNA Sequence Reads from Allopolyploid Organisms

BACKGROUND

Allpolyploid organisms are a type of polyploid in which two or more genomes from different ancestor species are brought together in a single nucleus. This genome doubling has radical effects on the genome. It causes immediate changes, termed "genomic shock", that affect the genetic and epigenetic state of the genome. In the long term, the genome doubling alters the course of evolution as two originally independent and self-sufficient genomes interact and develop together.

Allopolyploids are economically important to human society because there are many allopolyploid crops, including cotton, peanut, soybean, and Brassica. Analysis of these allopolyploids is complicated by the presence of multiple genomes. For example, single nucleotide polymorphisms (SNPs) that distinguish the co-resident genomes (homoeo-SNPs) can be confounded with SNPs that segregate in a Mendelian fashion (allele-SNPs).
Genome read categorization is the process of assigning DNA or RNA sequence reads from an allopolyploid organism to their singular genome of origin. Separating the genomes of an allopolyploid empowers researchers to identify true allele-SNPs and compare the parallel evolution of duplicated genes.

Common approaches to genome read categorization often involve the use of a single reference genome, belonging to a single diploid relative of one of the genomes from the allopolyploid, even if both diploid genome sequences are available. Sequence reads from diploid relatives of all constituent genomes are mapped to this reference, then SNPs distinguishing the

diploid relatives are inferred to represent homoeo-SNPs that would distinguish the genomes of the allopolyploid (Udall 2006). Reads from the allopolyploid can then be categorized to their genome of origin based on how closely it matches the haplotypes of the two parents. Note that, while a whole-genome reference sequence is desirable, the same strategy can be used with draft and/or transcriptome assemblies as a reference sequence. We previously developed PolyCat, which uses this approach (Page, Gingle, and Udall 2013). PolyCat considers homoeo-SNPs overlapped by each mapped read, and counts the bases at SNP locus to assign genome of origin for the read. If a threshold majority (default 75%) of counts match one of the genomes, the read is categorized to that genome. Multiple SNPs overlapped by a single read are evaluated for consistency of the genome assignment. Other tools have been developed using similar approaches to this problem, including HANDS and SNiPloid (Mithani et al. 2013; Peralta et al. 2013).

Along with read categorization, a researcher should consider a few issues when analyzing sequence data from an allopolyploid. First, if diploid A is used as the reference sequence, there will likely be an inherent mapping bias favoring reads from the A_T genome of the tetraploid over the B_T genome of the tetraploid (where the 'T' subscript distinguishes between the respective genomes in tetraploid nucleus). This can be alleviated through the use of GSNAP's SNP-tolerant mapping, which can take an index of known homoeo-SNPs identified between the diploid relatives and allow specified mismatches at those positions without penalizing the sequence alignment (Wu and Nacu 2010). Second, even when the mapping bias between diploids is accounted for, there may also be differences in the genetic distances between the tetraploid genomes and their respective diploid relatives. For example, the A genome could be better approximation of the A_T genome than the B genome is for the B_T genome. If a static SNP index

is being used, iterative development of SNP-indices may alleviate this problem by categorizing reads from a tetraploid then calling SNPs between the resulting genomes to generate a set of homoeo-SNPs that more closely represents the state of the tetraploid, rather than of the diploids. Finally, read categorization based on SNPs is limited by the ability of reads from one genome to map to the reference sequence of another genome. Wherever reads can map, homoeo-SNPs can potentially be identified. However, read categorization will only work if polymorphisms also exist at those loci.

Cotton species provide an excellent framework for the study of allopolyploidy and the development of specialized software. Allotetraploid cotton, which accounts for over 90% of cotton production worldwide, is the result of a hybridization and polyploidization event that occurred 1-2 million years ago (mya). At least 5 allotetraploid species arose from this single polyploidization. The parents of this event were A-genome and D-genome diploids. The A-genome diploids *Gossypium herbaceum* (A_1) and *G. arboreum* (A_2) are the closest extant diploid relatives of the allotetraploid A-genome (A_T), while *G. raimondii* (D_5) is the closest extant diploid relative of the allotetraploid D-genome (D_T). The A- and D-genomes diverged ~10 mya and both have 13 chromosomes. The A-genome is about twice the size of the D-genome (1.7 Gbp vs 0.9 Gbp), but the two genomes are largely collinear. The difference in length is largely made up of transposable elements. Allotetraploid cotton is a good model for research on polyploid genomes because the genome is relatively static and close diploid relatives are known for the genomes of the allotetraploid.

Here we present a new approach to read categorization that simultaneously uses data from two reference sequences, one for each genome of an allotetraploid. This dual-reference approach is implemented by our software called PolyDog. We compared the effectiveness of the

dual-reference method to the results of read categorization using either reference alone. We also compared the dual-reference method of mapping to a concatenation of two genome references, rather than to just one or the other (Figure 1).

PolyDog, along with PolyCat, is available for open source download as part of the BamBam package (https://sourceforge.net/projects/bambam/).

RESULTS

PolyDog Implementation

PolyDog processes two alignments (in BAM format) at once. These BAM files are made with the same set of reads, but are mapped to different references: in this case, the A_2 reference and the D_5 diploid reference, related to the A_T and D_T genomes of allotetraploid cotton. PolyDog examines each read on the basis of its mappings to both references and decides which reference the read matches more closely. Each read is analyzed based on 4 criteria:

- Whether the read mapped
- What mapping quality score (MAPQ) it had
- How long the alignment was
- How many bases matched the reference exactly (insertions and deletions are penalized as a mismatch)

These factors are factored serially, so the quality scores of the alignments are only considered if the read mapped to 1 or more locations; the alignment length is only considered if the read mapped in both references with equal MAPQ score, *etc*. If one mapping scores better than the other in a criterion, the read is categorized to the genome corresponding to the better mapping. In the tetraploid tests discussed below, nearly 75% of reads were categorized based on unique

mapping to one genome or the other (Figure 2). Differences in the length of reads aligned to each reference accounted for another 18%. Less than 1% of reads mapped to at least one reference but could not be categorized by any method. The relative contribution of each step will likely vary greatly based on the distance and nature of the relationship between the reference genome sequences.

When running PolyDog, reads are reported as belonging to the A-genome, D-genome, or unknown N-genome. These N reads are made up primarily of reads that map equally well to both reference sequences.

Comparative Analysis

Whole-genome shotgun reads were used to compare the different mapping and categorization methods. All reads were 100 bp paired-end Illumina reads.

Reads were mapped to 2 genome references. The 13 chromosomes of *G. arboreum* represented the A-genome, while the 13 chromosomes of *G. raimondii* represented the D-genome (Li et al. 2014; Paterson et al. 2012). Three allotetraploid species were used to test real application: *G. tomentosum* (AD_3), *G. darwinii* (AD_4) and *G. mustelinum* (AD_5). They each have 26 chromosomes (2n=4x=52), 13 from an A-genome ancestor and 13 from a D-genome ancestor. Mappings were performed using GSNAP (Wu and Nacu 2010). Only unique best mappings were accepted ("-n 1 -Q"). For PolyCat (but not for PolyDog or the full reference method), SNP-tolerant mapping was used ("-v" option) with the same set of homoeo-SNPs later used for read categorization by PolyCat.

PolyDog was run with paired-end support turned on, allowing fragments to be categorized as a single unit. Reads that mapped equally well to both references were rejected.

PolyCat was also run with paired-end support. A minimum vote majority of 75% per fragment was used. The SNP-index used for categorization was specific to each of the tetraploids. Initially, reads were mapped and categorized using a SNP-index based on homoeo-SNPs inferred from alignments of 6 A-genome and 4 D-genome diploids. Then SNPs were identified between allotetraploid reads categorized as A-genome and allotetraploid reads categorized as D-genome. Those SNPs were identified for each allotetraploid (AD_3, AD_4, and AD_5) and used for (re-) mapping and categorization in these tests.

For the full reference method, reads were "categorized" based on the reference chromosome they mapped to.

Error Analysis

Three diploids were used to test the accuracy of genome categorization by different methods: *G. herbaceum* (A_1-97), *G. arboreum* (A_2-34), and *G. raimondii* (D_5-2). These reads were treated in the same manner as the tetraploid reads: mapping with GSNAP followed by categorization by PolyCat and PolyDog. For the PolyCat tests, a SNP-index was used, based on homoeo-SNPs identified between 6 A-genome diploids and 4 D-genome diploids.

Categorizing diploid reads should be redundant because the genome of origin is already known for each read. But categorized diploid reads can be used as a useful measure for the accuracy of categorization methods, as every read from the D-genome diploid SHOULD be categorized as belonging to the D-genome. As such, the fraction of mapped reads that categorize to the A-genome instead of the D-genome approximates the error rate of that categorization method. Using an A-genome diploid, the fraction of mapped reads that categorize to the D-genome instead of the A-genome approximates the error rate. Note that this system for

measuring error rates only works because each read pair is mapped and categorized independently by all the methods analyzed in this study.

PolyDog was able to categorize slightly more reads than the full reference method, and both of these methods categorized far more reads than the PolyCat method, regardless of whether the A-genome or D-genome reference was used (Figure 3). The disadvantage of PolyCat is that it can only categorize reads in the homoeologous regions of the genome. The A-genome has hundreds of megabases of sequence that are not present in the D-genome, and even the smaller D-genome also has many regions that are absent in the A-genome. But PolyCat can only categorize reads where homoeo-SNPs are identified, and homoeo-SNPs can only be identified if the same region exists in both genomes.

PolyDog slightly outperformed the full reference method. This was largely because unique best mappings (GSNAP options -n 1 -Q) were required in the initial reference mapping. So a read that mapped equally well to the A-genome and D-genome versions of a locus would be unmapped in the full reference method. With PolyDog, however, the read would be mapped in both of the separate mappings. When PolyDog examines such mappings, it is able to investigate the difference between them with a finer resolution than GSNAP did when looking for the mapping. As a result, it may be able to assign the read to one genome. The ability of PolyDog to do this depends on the confidence thresholds used by the mapper and by PolyDog. But in general, PolyDog is and can be more aggressive than GSNAP in choosing a best mapping for a read because it is aware of the specific relationship between the two proposed mappings as pertaining to homoeologous loci.

With the PolyCat tests, more reads were mapped to the A-genome reference sequence than to the D-genome reference sequence (69.4% vs. 63.2%), but less reads were categorized

(Figure 4). The increased mapping rate is likely due to the large amount of non-homoeologous sequence in the A-genome. Allotetraploid reads from a non-homoeologous region can map to the A-genome but not the D-genome, thus increasing the mapping rate for the A-genome reference. However, the homoeologous portion of the genome is biologically the same size in both genomes, so it should be the same size in both references. But more homoeo-SNPs have been identified in the D-genome reference (28 M) than in the A-genome reference (15 M). As a result, PolyCat can analyze reads mapped to the D-genome reference with greater resolution. Thus, categorization rates were lower with an A-genome reference sequence.

PolyDog and the full reference method had higher error rates than the PolyCat methods (Figure 5). PolyCat is much more conservative, only using high confidence homoeo-SNPs and focusing on regions that can easily be distinguished by genome. Consequently, PolyCat categorizes far fewer reads but with a correspondingly low error rate. With the A-genome reference, PolyCat has less homoeo-SNPs to work with and thus categorizes even less reads with a correspondingly low error rate. Between PolyDog and the full reference method, PolyDog had a slightly lower error rate, likely for the same reason as it had a slightly higher categorization rate. The highest error rate of any method was less than 2.5% and most error rates were about 1%, suggesting that all these methods can be used to provide highly confident results (~99%).

In PolyDog and the full reference method, the highest error rate was observed in A_1-97 because the other two species (A_2-34 and D_5-2) were each represented in one of the reference genome sequences. In PolyCat, the homoeo-SNPs were based on diploids from all three species, so the A_1-97 did not have as much of a higher error rate. A_2-34 consistently exhibited a much lower error rate than either of the other species. The reason for this superior accuracy with A_2-34 is unclear.

The quality and completeness of the reference sequences can have a massive effect on error rate. This is readily observable in PolyDog output. PolyDog's error can be reported as the number of reads from a D-genome diploid that mapped to the D-genome reference but were ultimately (erroneously) categorized as A-genome reads. If you instead consider the number of D-genome reads that mapped to the A-genome reference and were (erroneously) categorized as A-genome reads, the number will likely be higher than with the previous measurement. This is because the reference being used is the wrong categorization type, so it's easier for reads mapped to that reference to look like the wrong categorization type. With D_5-2 reads, this increase of error as observed using a different reference sequence is about 2x (3.66 M -> 7.20 M reads). With A_1-97 and A_2-34 reads, this increase is 17x (4.04 M -> 141.25 M reads) and 35x (7.45 M -> 127.72 M reads), respectively. A likely cause for this asymmetry is the relative completeness and quality of the A- and D-genome reference sequences. This effect will vary greatly depending on the relative completeness of the reference sequences used, as well as the distance between the diploid relative and the allotetraploid being analyzed.

CONCLUSION

Using both reference sequences, either through PolyDog or the full reference method, is beneficial because it allows analysis of both the homoeologous and non-homoeologous portions of the genomes. However, there are still reasons to use a single reference sequence, such as with PolyCat. First, a reference sequence may only be available for one of the genomes in an allopolyploid, or the reference sequence of one genome may be largely incomplete. Second, if a SNP-index is used to properly alleviate mapping biases and categorize reads, a single reference sequence facilitates a comparison of homoeologs with each other. This can aid in the

identification of allele-SNPs and other Mendelian polymorphisms. It also facilitates direct quantitative comparison, as for gene expression analysis. In contrast, PolyDog can categorize reads in regions that are unique to one genome or the other. This may introduce a bias in the analysis. PolyDog would be better suited to qualitative analyses, such as genotyping loci and building phylogenetic trees, because it can use reads from the unique parts of the genome.

Perfectly conserved regions cannot be analyzed by read categorization because no difference in sequence identity can be exploited. Highly repetitive regions are also likely to be uncategorizable. It is possible that a region that is highly conserved between diploid species may have diverged in the polyploidy. The genome shock associated with polyploidization may cause almost immediate changes in the polyploidy. Or the presence of two copies of each gene in the same nucleus may result in divergent gene fates: neo-functionalization, sub-functionalization, or non-functionalization (Ohno 1970). Regardless, PolyCat may detect these polymorphisms because a read may stretch from a categorizable region into a non-categorizable region. Longer reads make this more likely. In addition, using paired-end data ("-p" option) allows a whole fragment to be categorized together, thereby reaching even further into an otherwise uncategorizable region. If SNPs are identified in this manner, a new index based on the allotetraploid itself may be constructed, facilitating further analysis. In addition, such an allotetraploid-specific SNP index has the benefit of not including homoeo-SNPs that resulted from autapomorphies in one of the diploid relatives.

A SNP-index is not used by PolyDog, and it is recommended that SNP-tolerant mapping not be used in preparing BAM files for analysis by PolyDog. This is because PolyCat and PolyDog act on fundamentally opposite principles in the mapping stage. PolyCat seeks to map reads from the "wrong" genome to a reference sequence (*e.g.*, map A-genome reads to a D-

genome reference). Then it categorizes the reads to sort out genomic identity. On the other hand, PolyDog seeks to NOT map reads from the "wrong" genome to a reference sequence. It is desirable (with PolyDog) that a read from the A-genome not map to the D-genome. Then PolyDog may easily recognize the genomic origin of the read.

PolyDog's advantage over the full reference method is that PolyDog can leverage the knowledge of the homoeologous relationship between loci in different genomes and distinguish it from the possible paralogous relationship between loci within a genome. In effect, PolyDog allows multiple hits when those multiple hits are on different genomes but disallows multiple hits when they're on the same genome. PolyDog does this by applying different standards to distinguish homoeologs from those used to distinguish paralogs. Stricter settings and larger margins are needed to confidently avoid paralogous mapping, but looser settings and minimal margins can be used to decide to which homoeolog a read belongs. PolyDog can require unique best mapping without having to discard reads that map comparably well to both genomes. In contrast, the full reference method must either 1) allow multiple hits for each read or 2) require unique best mapping. Option 1 allows a read to map to both genomes, but it also allows a read to map to multiple loci within one genome, which is often undesirable. Option 2 avoids this, but it also throws away some reads that map to homoeolgous loci. PolyDog takes the benefits of both. It maps each read to just a single locus in each genome separately. Then PolyDog analyzes and compares those best mappings from each genome. Thus, a read can map to two loci, but only if they're on different genomes.

While PolyDog performs better than the full mapping method, the difference is small. Because of this, another consideration becomes important in deciding which method to use for a specific experiment. PolyDog ultimately results in reference mappings made to each of the

reference genome sequences. For each reference, there is a single mapping for each genome. Thus for cotton, PolyDog produces a BAM file of A-genome reads mapped the A-genome reference, A-genome reads mapped to the D-genome reference, D-genome reads mapped to the A-genome reference, and D-genome reads mapped to the D-genome reference. These results can be very useful for comparisons where each genome should be considered separately (Rambani, Page, and Udall 2014). They can also be useful for identification of genetic markers such as SNPs (Thyssen et al. 2014). In contrast, the full reference method results in a single mapping of all reads against a concatenated reference. Such an output may be more appropriate for comparisons of species, where the character of the distinct genomes is not of interest.

When reference genome sequences are available for relatives of each genome of an allopolyploid, read categorization with PolyDog can leverage both sequences to maximize read categorizability with high (~99%) confidence. When dealing with unique best mappings for each read, PolyDog is a better option than simply mapping to a concatenation of the reference sequences, with higher categorization rates and lower error rates.

When a reference genome sequence is only available for the relative of one genome from the allopolyploid, read categorization based on homoeo-SNPs, whether through PolyCat, SNiPloid, HANDS, or some other tool, is an excellent and high confidence solution. However, analysis will be limited to regions that are present in both genomes, limiting the analysis of copy number variants.

Even if reference genome sequences are available for all genomes of the allopolyploid, it may still be preferable to use a single reference genome sequence followed by a tool like PolyCat. This will serve to minimize mapping and categorization biases between the genomes, facilitating quantitative analyses such as gene expression studies.

REFERENCES

1. Li, Fuguang, Guangyi Fan, Kunbo Wang, Fengming Sun, Youlu Yuan, Guoli Song, Qin Li, et al. 2014. "Genome Sequence of the Cultivated Cotton Gossypium Arboreum." *Nat Genet* 46 (6) (May 18): 567–572. doi:10.1038/ng.2987.
2. Mithani, Aziz, Eric J Belfield, Carly Brown, Caifu Jiang, Lindsey J Leach, and Nicholas P Harberd. 2013. "HANDS: a Tool for Genome-Wide Discovery of Subgenome-Specific Base-Identity in Polyploids." *BMC Genomics* 14 (1): 653. doi:10.1093/bib/bbs017.
3. Ohno, S. 1970. *Evolution by Gene Duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.
4. Page, J T, A R Gingle, and J A Udall. 2013. "PolyCat: a Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms." *G3 (Bethesda)* 3 (3) (March 8): 517–525. doi:10.1534/g3.112.005298.
5. Paterson, Andrew H, Jonathan F Wendel, Heidrun Gundlach, Hui Guo, Jerry Jenkins, Dianchuan Jin, Danny Llewellyn, et al. 2012. "Repeated Polyploidization of Gossypium Genomes and the Evolution of Spinnable Cotton Fibres." *Nature* 492 (7429) (December 19): 423–427. doi:10.1038/nature11798.
6. Peralta, Marine, Marie-Christine Combes, Alberto Cenci, Philippe Lashermes, and Alexis Dereeper. 2013. "SNiPloid: a Utility to Exploit High-Throughput SNP Data Derived From RNA-Seq in Allopolyploid Species." *International Journal of Plant Genomics* 2013 (8, article r86): 1–6. doi:10.1186/1471-2164-13-219.
7. Rambani, Aditi, Justin T Page, and Joshua A Udall. 2014. "Polyploidy and the Petal Transcriptome of Gossypium." *BMC Plant Biology* 14 (1): 3. doi:10.1016/j.molcel.2008.11.009.

8. Thyssen, Gregory N, David D Fang, Rickie B Turley, Christopher Florane, Ping Li, and Marina Naoumkina. 2014. “Next Generation Genetic Mapping of the Ligon-Lintless-2 (Li 2) Locus in Upland Cotton (Gossypium Hirsutum L.).” *Theoretical and Applied Genetics* (August 15). doi:10.1007/s00122-014-2372-1.
9. Udall, J A. 2006. “A Novel Approach for Characterizing Expression Levels of Genes Duplicated by Polyploidy.” *Genetics* 173 (3) (April 30): 1823–1827. doi:10.1534/genetics.106.058271.
10. Wu, T D, and S Nacu. 2010. “Fast and SNP-Tolerant Detection of Complex Variants and Splicing in Short Reads.” *Bioinformatics*.

FIGURES

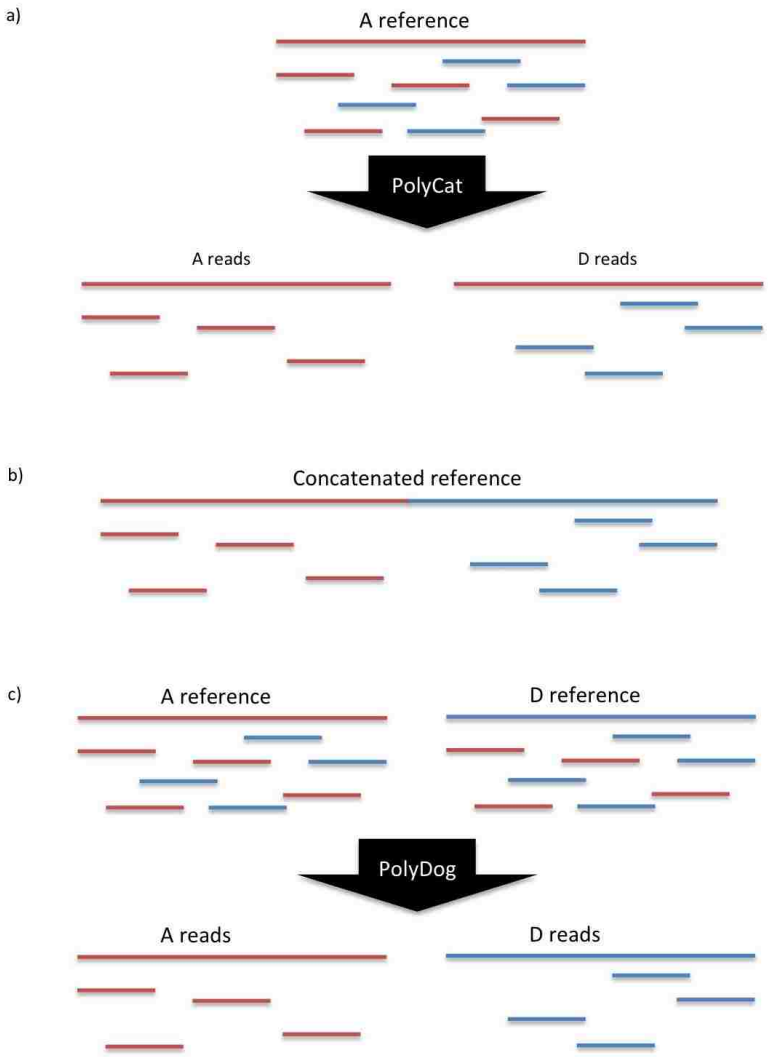


Figure 11 Methods for read categorization. In all images, black lines represent the A-genome while grey lines represent the D-genome. Idealized forms of these methods are shown, ignoring structural differences, perfect conservation, and other sources of complication and error. With PolyCat, (a) homoeo-SNPs are first identified between the consensus sequences for already known A- and D-genome reads. Then all reads are mapped to a single reference sequence (A-genome in the example) and PolyCat categorizes them by source genome. With the full reference method (b), reads are mapped to a concatenation of the A- and D-genome reference sequences, so reads will naturally map to the part of the reference that represents its appropriate genome. With PolyDog (c), the same set of reads from an allotetraploid is mapped to both the A- and D-genome references. Then PolyDog examines each pair of mappings and categorizes that read to its genome of origin.

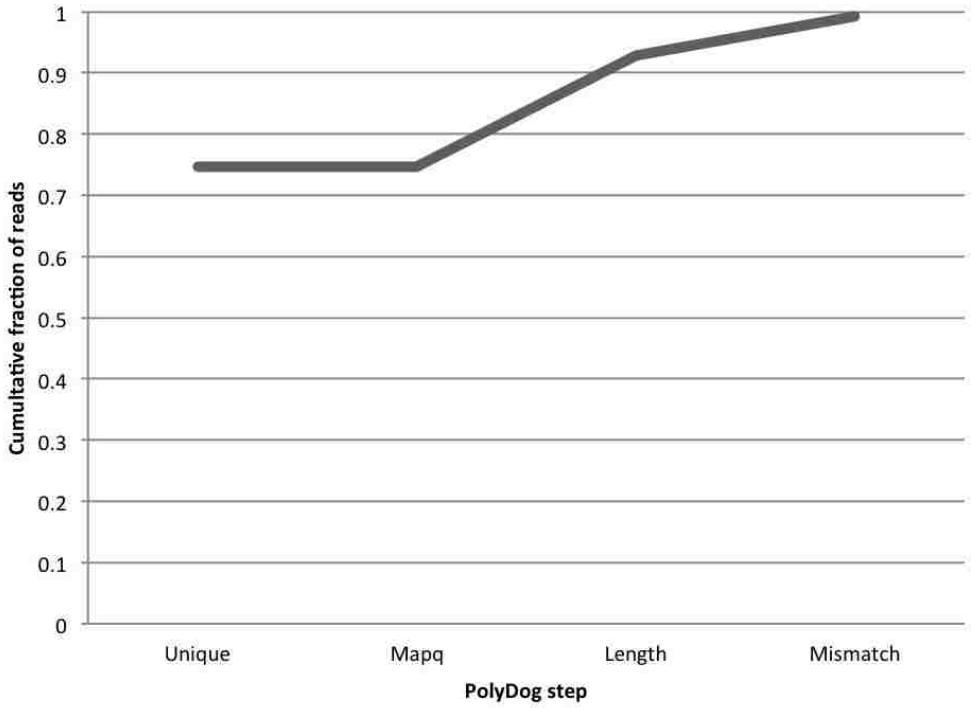


Figure 12 Categorization by each PolyDog step. Reads were first categorized by PolyDog based on unique mapping to one genome or another, then based on MAPQ, alignment length, and number of mismatches. In our tests, GSNAP did not calculate different MAPQ scores for each alignment, so MAPQ was not helpful in categorization. Fractions shown are relative to the total number of mapped reads, and are averaged over 3 allotetraploid datasets.

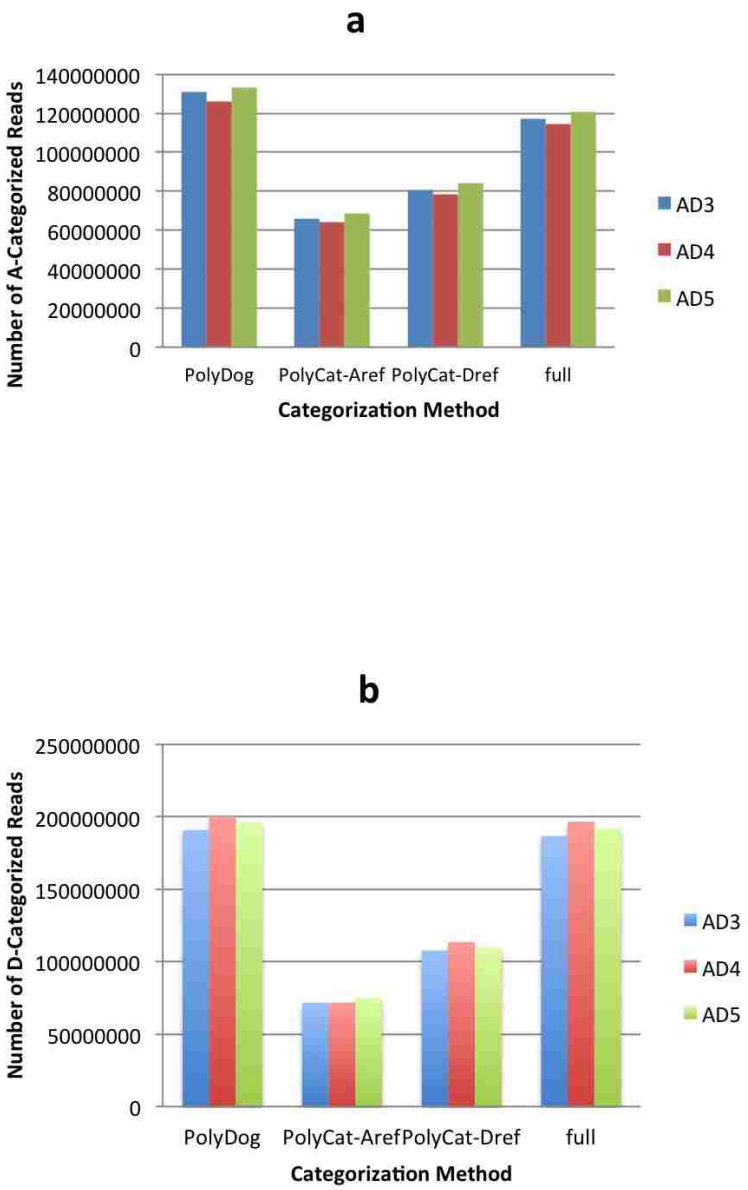


Figure 13 Categorization results by method. Number of reads categorized to the A-genome (a) and to the D-genome (b). Reads from three allotetraploid cotton species—AD_3 (blue), AD_4 (red), and AD_5 (green)—were categorized by PolyDog, PolyCat using the A-genome as reference, PolyCat using the D-genome as reference, and the full reference method.

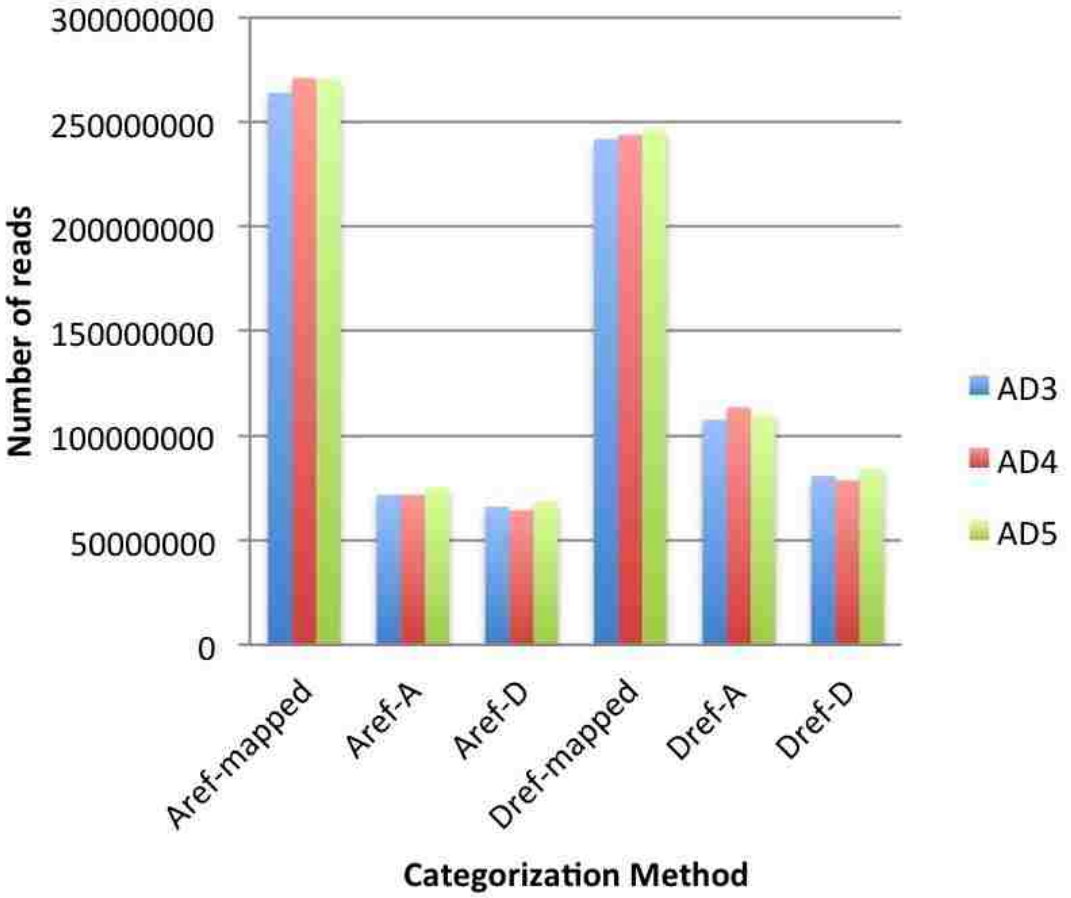


Figure 14 PolyCat performance. Reads from three allotetraploid cotton species—AD_3 (blue), AD_4 (red), and AD_5 (green)—were mapped by GSNAP with SNP-tolerant mapping and categorized by PolyCat using either the A-genome or D-genome reference sequence. The total number of mapped reads in each case is shown (Aref-mapped and Dref-mapped), as well as the number of reads categorized to the A (Aref-A and Dref-A) and D genomes (Aref-D and Dref-D).

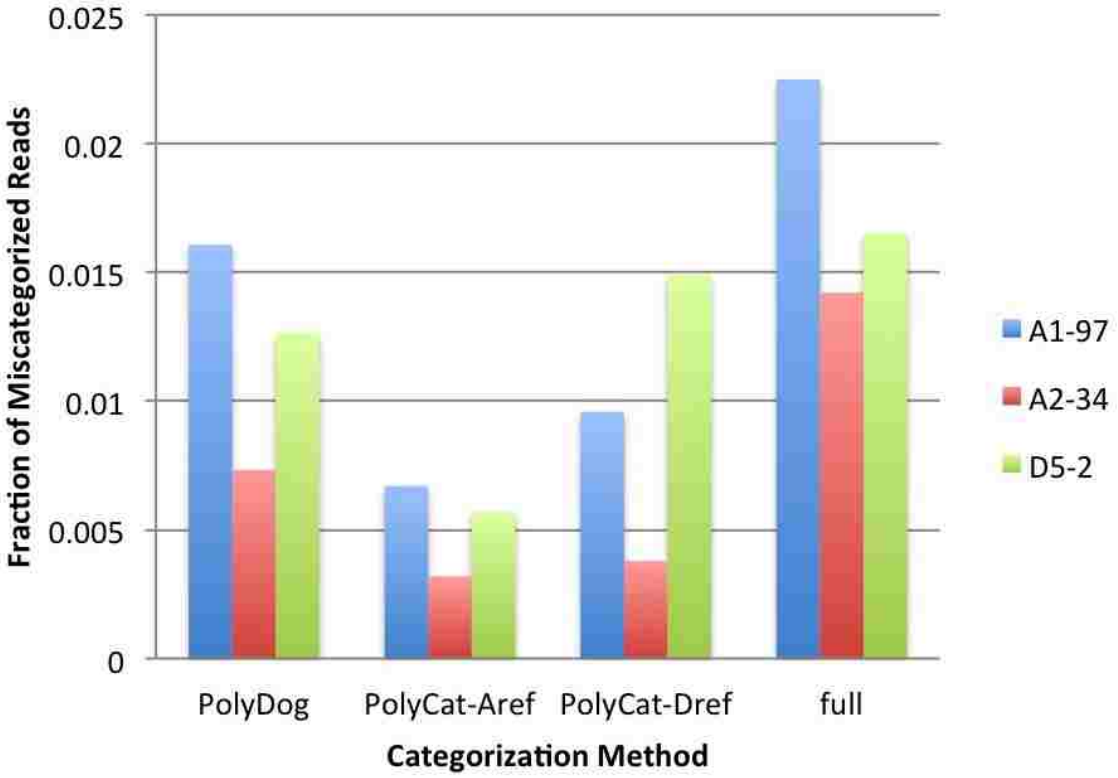


Figure 15 Error rates in categorization. Three diploid cotton species—A_1-97 (blue), A_2-34 (red), and D_5-2 (green)—were categorized by PolyDog, PolyCat using the A-genome as reference, PolyCat using the D-genome as reference, and the full reference method. The error rate shown is the number of reads categorized to the wrong genome divided by the number of mapped reads.

CHAPTER 5

DNA Sequence Evolution and Homoeologous Conversion in Allotetraploid Cotton

INTRODUCTION

Cotton (genus *Gossypium*) is an economically important crop, valued for its spinnable fiber which is produced on the seed of certain domesticated species, as well as for cottonseed oil used in food and cooking. The genus *Gossypium* consists of 8 genome groups: designated by letters A-G and K (Wendel et al. 2012). The A and D genomes diverged from one another about 5-10 million years ago (mya), with the A-genome in Africa and the D-genome in South America. The A- and D-genomes each have 13 chromosomes, but the A-genome length is ~1700 Mbp, nearly twice the ~900 Mbp length of the D-genome. The length difference is primarily due to retrotransposon activity in the A-genome (Hawkins et al. 2009; Grover and Wendel 2010). Based on genetic maps, the two genomes are largely collinear (Brubaker et al. 1999; Wang et al. 2013).

A polyploidization event approximately 1 million mya gave rise to six described AD allotetraploid species with genome length ~2400 Mbp, mostly native to Central and South America (Wendel 1989; Krapovickas et al. 2008; Grover et al. 2014). The six tetraploid cotton species arose from a single hybridization between unknown A-genome and D-genome diploid progenitors (Grover et al. 2012). Another unnamed island endemic of the Northern Line Islands is under consideration as a seventh tetraploid species. The A-genome donor was similar to extant *G. herbaceum* (A_1) and *G. arboreum* (A_2), while the closest extant diploid relative of the D-genome donor is likely *G. raimondii* (D_5) (Flagel et al. 2012). The two A-genome diploids are about twice as good of a predictor of the A_T-genome as D_5 is of the D_T-genome, indicating a greater genetic distance between D_5 and D_T than between A_1 or A_2 and A_T. There are two major

clades among the tetraploid species, one containing *G. hirsutum* (AD_1) and the other containing *G. barbadense* (AD_2) (Wendel and Albert 1992). *G. ekmanianum* (AD_6) and a possible seventh species belong to the AD_1 clade, and have only recently been described as distinct species separate from AD_1 (Grover et al. 2014). *G. darwinii* (AD_5) belongs to the AD_2 clade. *G. mustelinum* (AD_4) diverged from the other tetraploids prior to the divergence between the AD_1 and AD_2 clades, making it a useful outgroup for analyses of the cotton tetraploids. The position of *G. tomentosum* (AD_3) from Hawaii is either in the AD_1 clade or as an outgroup to the split between AD_1 and AD_2.

Two A-genome diploids— A_1 and A_2—and two tetraploids— AD_1 and AD_2—have been independently domesticated and produce long spinnable fiber. The remaining tetraploid species (AD_3 – AD_6) and other genome groups do not produce spinnable fiber and have not been domesticated. AD_1 is the source of the vast majority (~90%) of worldwide cotton production (Wendel and Cronn 2003). AD_2 accounts for another ~5%; its longer fibers are valued for high quality textiles. Attempts to produce stable AD_1 x AD_2 hybrids have resulted in fertile and productive F_1 hybrids, but development of hybrid seed is generally cost-prohibitive. In addition, hybrid breakdown, hybrid sterility, and selective elimination of genes make genomic resources difficult to develop (Zhang and Percy 2007). As such, introgression of genetic material from AD_2 into AD_1 (or vice versa) is of particular interest.

Homoeolog conversion—also called gene conversion, non-reciprocal homoeologous recombination, or homoeologous gene conversion—is a phenomenon in which an allele from one genome of a tetraploid overwrites its homoeolog in the other genome. For example, a D_T-genome allele overwrites its A_T-genome homoeolog, resulting in 4 copies of the D_T-genome allele and 0 of the A_T-genome allele instead of 2 of each. Homoeologous conversion has been

identified in various tetraploid groups, including *Brassica* and *Gossypium* (Salmon et al. 2009; Fujimoto et al. 2006; Teshima 2004). Homoeologous conversion is presumed to be caused by non-reciprocal homoeologous recombination, although the specific mechanism or cause for such events is still unknown. It has been hypothesized that homoeologous recombination is a major force in the evolution of desirable traits in allopolyploid crops, suggesting that it may be the reason that fiber traits in cotton have been selected on the D_T-genome (Gaeta and Chris Pires 2009). The majority of diversity among allopolyploid cotton species has been attributed to homoeologous conversion (Guo et al. 2014).

High-throughput sequencing technologies have provided the ability to analyze and compare many genomes from a single species or group. The results of these studies provide insight into genetic diversity, evolution, and specific traits of targeted species. Re-sequencing efforts in corn, tomato, and cotton diploids have investigated mutations, selection, and linkage disequilibrium (Chia et al. 2012; Lin et al. 2014; Page et al. 2013b). In this study, we apply Illumina technology to resequence and compare the genomes of 34 cotton tetraploids from 6 species at average coverage 23x per accession, whereas previous cotton tetraploid resequencing efforts have averaged only minimal coverage. We mapped reads to the diploid A- and D-genome reference sequences of *G. arboreum* and *G. raimondii*, as well as to the recently published drafts of the cotton tetraploid genomes, although subsequent analyses were based on the mappings to the diploid reference sequences (Li et al. 2014; Paterson et al. 2012; Li et al. 2015; Zhang et al. 2015). Mapping to the diploid sequences for this report is tenable because 1) the two tetraploid sequences have not arrived at a consensus for loci positions and 2) >25% of the draft sequences remain unanchored to either A_T- or D_T-genomes. Since our main focus compares A vs. D (or A_T vs. D_T), the comparison is only possible in regions present in *both* A and D genomes. We

account for the differences between the respective diploid and tetraploid genomes by adjusting the diploid reference sequences to the genotypes observed in the tetraploid species. Here, we examine the comparative evolution and genetic diversity of the polyploid cotton species and genomes.

MATERIALS AND METHODS

Various components of BamBam (version 1.4) and SAMtools (version 1.2), along with custom scripts built on BioPerl, were used to modify, summarize, and analyze aligned sequence data throughout the processes described below (Page et al. 2014; Li et al. 2009; Stajich et al. 2002).

Sequence Data

In total, over 18 billion 100 bp paired-end Illumina reads were generated by Huntsman Cancer Institute, BGI, University of California-Davis, and Mississippi State University across 33 accessions: 13 *G. hirsutum*, 15 *G. barbadense*, and 1 each of *G. tomentosum*, *G. mustelinum*, *G. darwinii*, *G. ekmanianum*, and 2 of the accessions endemic to the Wake Atoll National refuge. Illumina sequence data for the diploids—3 *G. herbaceum*, 4 *G. arboreum*, and 4 *G. raimondii*—and one additional *G. hirsutum* were obtained from SRA. For *Gossypiodes kirkii*—an outgroup of the *Gossypium* genus—40 million 36 bp single-end Illumina reads were obtained from NCGR. Reads were trimmed with Sickle (https://github.com/najoshi/sickle) using a PHRED quality threshold of 20.

Homoeo-SNP Identification

All reads were aligned to both the D_5 and A_2 reference genomes with GSNAP using the options "-n 1 –Q" to require unique best mappings (Paterson et al. 2012; Wu and Nacu 2010; Li et al. 2014). An index of homoeo-SNPs inferred from diploid whole genome resequencing was used for GSNAP SNP-tolerant mapping ("-v" option) (Page et al. 2013b). Reads were then categorized as originating in the A_T- or D_T-genome by PolyCat, using the same homoeo-SNP index used for the mapping (Page et al. 2013a). InterSnp (part of BamBam) was used to call SNPs between individuals with a minimum allele coverage of 5 reads per individual, and SNPs that consistently (75% of observed genotypes) manifested in one genome of a species—and were consistently (75%) absent in the other genome of that species—were called as homoeo-SNPs (Page et al. 2014). Five of these tetraploid-based homoeo-SNP indices were generated for each genome, one each for AD_1, AD_2, AD_3, AD_4, and AD_5, named D13.snp4.1 through D13.snp4.5 (or A13.snp2.1 through A13.snp2.5), respectively.

Mapping and Categorization

Having identified homoeo-SNPs that characterize the diploid A- and D-genomes (and the tetraploid A_T- and D_T-genomes), a modified A-genome reference sequence was constructed that used the D-genome reference as a template, but replaced the D-genome nucleotide with an A-genome nucleotide at positions with homoeo-SNPs (Page et al. 2013b). Having identified SNPs specific to each of the 5 traditional tetraploid cotton species, we also constructed modified reference sequences of A_T- and D_T-genomes where the ancestral nucleotide was replaced with the specific nucleotide for that species. We re-mapped the reads to their respective, modified genome sequences then categorized them to the A_T- and D_T-genomes via PolyDog (Page and

Udall 2015). This remapping enabled us to more accurately map and categorize reads because of the modified reference sequences, as well as to achieve more consistent coverage across the genome because of PolyDog. PolyDog allows for categorization of both reads of a paired set, even if only one of the reads overlaps a homoeo-SNP, increasing the portion of the genome that can be categorized.

Indel-induced mapping errors were corrected using GATK (DePristo et al. 2011). First, RealignerTargetCreator was run on a group of 20 A_T-genome BAM files and on 20 D_T-genome BAM files (representing all tetraploid species). Second, IndelRealigner was used on each individual BAM file to adjust read alignments around the indels identified in the first step: 3,692,540 loci in the A_2 reference and 2,195,978 loci in the D_5 reference.

Single Nucleotide Polymorphisms

SNPs and short indels were called--once for all A_T-genome BAM files and once for all D_T-genome BAM files--between the categorized genomes using InterSnp with a minimum coverage per allele of 5 reads and minimum frequency of 30% (Page et al. 2014). A neighbor-joining tree was constructed for each genome, bootstrapping 1000 sub-samples without replacement with 5% of SNPs in each sub-sample. Trees were generated by creating a distance matrix based on genotypes at all SNP loci, then running neighbor (from PHYLIP) with random sample ordering to build the actual tree (Felsenstein 2002). The 1000 trees from the bootstraps were combined with consense (from PHYLIP) to make a single consensus tree. Trees were visualized in Geneious (Drummond et al. 2011).

Small homoeologous conversions were analyzed by using PolyCat to categorize mapped reads from each tetraploid because PolyCat categorization allows for intergenomic analysis at a

nucleotide level (Page et al. 2013a). Then SNPs were called with InterSnp across all species and genomes (Page et al. 2014). Consensus genotypes were called for each species at sites that had coverage from at least 75% of individuals (10/13 for AD_1 and 11/14 for AD_2), and genotype patterns suggestive of homoeologous conversion in AD_1 or AD_2 were identified (e.g., A_2, A_T, and D_T have a C while D_5 has a T).

Organelle Genomes

The chloroplast and mitochondria genomes were analyzed by taking all unmapped reads from SNP-tolerant mappings to the D_5 reference sequence and remapping them to the *G. hirsutum* chloroplast and mitochondria sequences (Lee et al. 2006). SNPs were called by InterSnp with minimum allele coverage of 50 reads (Page et al. 2014). Then a phylogenetic tree was built and visualized as above, but with 100,000 bootstraps with 25% of the SNPs from the combined set of chloroplast and mitochondria SNPs in each bootstrap (Felsenstein 2002; Drummond et al. 2011).

Copy Number Variants

Copy number variants (CNVs) were called in the A_T- and D_T-genomes of each sample, relative to their respective diploid relatives, using CNVKit (Talevich et al. 2014). Reads from 3 diploid A_2 lines and 4 diploid D_5 lines were mapped and categorized in the same manner as the reads from the tetraploids, providing reference coverage profiles for the A- and D-genomes, which serve to normalize for biases in sequence coverage that are shared between diploid and tetraploid members of a common genome. The coverage of each tetraploid genome was compared to the reference coverage profile of its diploid relative. The gene annotations for each

reference sequence were provided as targets, and accessible regions of the genome were identified for filtering by a CNVKit utility script genome2access.py. Segments identified by CNVKit as having a log base 2 copy number of at least 1.0 were considered duplications in the tetraploid genome, and segments identified with a log base 2 copy number of -1.0 or less were considered deletions.

RESULTS

Mapping and Categorization

Approximately 60% of reads from tetraploids mapped to unique loci on the D_5 reference, while 70% mapped to unique loci on the A_2 reference (Fig. 1). The increased mapping percentage for the A_2 reference is likely because the A_T-genome is larger than the D_T-genome, so more reads drawn randomly from the tetraploid should be A-like than D-like. The difference is only 10% because much of the extra A-genome sequence is either repetitive (preventing unique mapping by short reads) or simply absent from the reference sequence. More reads were categorized by both PolyCat and PolyDog to the A_T-genome than to the D_T-genome. This is likely due to 1) the increased size of the A-genome and 2) the greater genetic distance between D_5 and D_T, which slightly decreases the effectiveness and accuracy of read categorization. When using the A_2 reference instead of the D_5 reference, the frequency of categorization was lower because less homoeo-SNPs have been defined in the A_2 reference SNP index. In addition, a greater fraction of the A_2 reference is non-homoeologous sequence, resulting in more reads that map to the reference but will ultimately be uncategorizable because they map to A-genome unique sequence. More reads overall were categorized by PolyDog than by PolyCat because PolyDog is able to categorize reads mapped to non-homoeologous regions (Page and Udall

2015). The end result of read mapping and categorization was read alignment (BAM) files for each genome (A_T and D_T) in each tetraploid accession. Downstream analyses focused on reads categorized by PolyDog for intragenomic analyses and by PolyCat for intergenomic analyses.

We also mapped and categorized reads from the diploids A_2 and D_5 to measure the error rates of each categorization method (number of A_2 reads categorized as D_T and number of D_5 reads categorized as A_T). PolyCat exhibited an error rate less than 1%. PolyDog more aggressively categorized reads, resulting in higher categorization rates (~97%) at the expense of higher error rates (~3%). The reads in error were either erroneously categorized or—unlike the error rate for PolyCat—could not be categorized because they were equally similar to the A_T- and D_T-genomes.

Single Nucleotide Polymorphisms

We analyzed evolutionary relationships by examining SNPs among cotton accessions. Within read alignments, we identified SNPs both between genomes (termed "homoeo-SNPs") and between accessions ("allele-SNP"). Homoeo-SNPs were first identified between the diploids A_2 and D_5 and then between the A_T- and D_T-genomes of AD_1, AD_2, AD_3, AD_4, and AD_5. Between 12.4 and 18.2 million homoeo-SNPs (23.9 million total unique loci) were found when using the A_2 reference and between 19.2 and 28.5 million homoeo-SNPs (35.6 million total unique loci) when using the D_5 reference (Table 1). There were 7.9 million and 11.2 million homoeo-SNPs on the A_2 and D_5 references, respectively, that were shared within all tetraploid species. There were also 6.6 million and 9.4 million homoeo-SNPs (using A_2 and D_5 references, respectively) that shared within all tetraploid species *and* they were also found between the diploid genomes. About 12-15% of homoeo-SNPs were in annotated genes. While this

percentage was higher than would be expected for homoeo-SNPs distributed randomly across the genome, it did not provide evidence for a bias towards genes. Homoeo-SNPs were simply easier to detect in genes because they were more conserved between genomes, whereas many non-genic regions were unique to one genome and/or difficult to map, preventing homoeo-SNP identification. There were 1,358 and 4,054 genes with no homoeo-SNPs identified in the tetraploid sequences aligned to the D_5 and A_2 reference sequences, respectively. The homoeo-SNP indices are available on CottonGen as A13.snp2.x and D13.snp4.x, where x=0 for homoeo-SNPs found in the diploids, x=1 for AD_1, x=2 for AD_2, etc.

We also identified allele-SNPs within genomes, between accessions of each species. For these comparisons, reads categorized by PolyDog to the A_T-genome were compared with each other and with the A-genome diploids using the A_2 reference sequence, while D_T-genome categorized reads and D-genome diploid reads were compared using the D_5 reference sequence. After filtering with a minor allele frequency of 10%, there were 15,864,224 and 10,437,663 allele-SNPs in the A_T- and D_T-genomes, respectively. In both AD_1 and AD_2, the number of A_T-genome allele-SNPs was about 1.5x the number of D_T-genome allele-SNPs (Table 2). However, after normalizing by genome size, average diversity (allele-SNPs per bp) in the D_T-genome was nearly 2x the average diversity in the A_T-genome.

We compared the number of SNPs in genes present in both the A_T- and D_T-genomes. There were 947,157 and 1,638,565 allele-SNPs in annotated genes in the A_T- and D_T-genomes, respectively. To identify homoeolog pairs in the annotations of the A_2 and D_5 reference sequences, we used BLASTP with a maximum e-value of $1e^{-20}$ to compare the peptide sequences of annotated A_2 and D_5 genes (Altschul et al. 1990). We found 26,782 gene pairs that were best reciprocal BLAST hits, suggesting homoeolog status. As expected, the lengths of these

homoeolog pairs were highly correlated (Pearson $r^2 = 0.744$, $_p$-value < 2.2e-16) (Figure 2A). The density of allele-SNPs was weakly correlated among allotetraploids (Pearson $r^2 = 0.321$, p-value < 2.2e-16) and among AD_1 cultivars (Pearson $r^2 = 0.261$, p-value < 2.2e-16; Figure 2B). There were 1,173 genes that had 0 allele-SNPs in the A_T-genome while the homoeolog had 5 or more allele-SNPs, and there were 1,835 genes that had 0 allele-SNPs in the D_T-genome while the homoeolog had 5 or more allele-SNPs.

We examined allele-SNP frequencies among different groups of accessions to identify regions of artificial selection associated with domestication. We analyzed the ratio of allele-SNP frequency in domesticated AD_1 lines (Coker 312, Deltapine 5690, Fibermax 832, Acala Maxxa, M-240, PD-1, SureGrow 747, Stoneville 474, Sealand-542, Tamcot sphinx, and Texas Marker 1) compared to the remainder of accessions in the AD_1 clade (AD_1, AD_6, and AD_7—excluding AD_3 because of its greater distance from the other species). We analyzed this ratio in a 1 Mbp window stepping by 500 Kbp (Fig. 3). The selection ratio (diversity of AD_1 cultivars / diversity of AD_1) ranged from 2.127 to 0.916 in the A_T-genome and from 1.794 to 0.831 in the D_T-genome. In the A_T-genome, there were 1,299 genes in 77 windows with a selection ratio less than 1.0 (less means stronger selection), while in the D_T-genome there were 5,152 genes in 228 windows with a selection ratio less than 1.0. These strongly selected regions include 1 gene on the A_T-genome (Cotton_A_25726) and one on the D_T-genome (Gorai.003G049300) that are CesA genes noted as being under strong selection in another study (Zhang et al. 2015).

Phylogenies

The A_T- and D_T-genome phylogenies positioned species consistent with previous observations (Wendel and Grover 2015): The tetraploids primarily split into two clades, one

containing AD_1 and the other containing AD_2. AD_4 is an outgroup to this split. AD_5 is closely related to AD_2, while AD_6 and AD_7 are close to AD_1. AD_3 is in the AD_1 clade, but diverged shortly after the AD_1 vs AD_2 split, making it a more distant relative of AD_1 than are AD_6 and AD_7 (Fig. 4). In the consensus bootstrap trees for the nuclear genomes, nearly all splits have 99-100% bootstrap support and only 2 splits (both within the AD_1 cultivars) have less than 90% support (80% and 82%). The cultivated varieties in AD_1 clustered together with wild AD_1 accessions nearby, and the same pattern was observed with AD_2 cultivars and wild accessions. The A_T- and D_T-genome trees largely agreed in regard to the topology of the AD_2 clade, with the exception of the positioning of a sub-clade containing the 3 cultivars: DP-340, GB-236, and Phy-76. The AD_1 clade was also similarly constructed in the A_T- and D_T-genome phylogenies, although the cultivars are so closely related to one another that their precise arrangements varied between trees. Outside of the AD_1 cultivars, the AD_1 wild accessions (TX-2094 and TX-231) were closest to the cultivar clade. Notably, GB-319 (although previously classified as an accession of AD_2) clustered with the wild AD_1 accessions. The two AD_7 accessions formed a clade external to the wild AD_1, and AD_6 was external to AD_7.

In the chloroplast and mitochondria, 1,417 and 331 SNPs were identified, respectively. After bootstrap tree construction, a consensus neighbor-joining tree was created based on these organelle genome data. The organelle phylogeny disagreed with both nuclear phylogenies. There were numerous minor differences in the fine placements within species, which are unremarkable, given the relatively small number of SNPs and common differences in maternal inheritance. The one major disagreement was in the placement of AD_3, AD_4, and AD_5, which formed a small clade near the AD_1 species. This position would be understandable for AD_3 and possibly even AD_4, but it is a drastic shift for AD_5, which is (in previous observations and the nuclear-based

trees) more closely related to AD_2. If this result could be confirmed by other data, it could suggest the possibility of multiple crosses giving rise to the extant tetraploid species, rather than a single cross as generally supported (Grover et al. 2012). From this tree alone, it is not clear that the chloroplast and mitochondria are more closely related to the A-genome parent than to the D-genome parent, as previously observed, suggestive of a maternal A-genome donor to the original tetraploid (Wendel 1989). Considering that the genetic distance between the A-genome diploids and A_T-genome is significantly shorter than the genetic distance between D_5 and the D_T-genome, it may be that the A-genome donor was not the maternal parent of the original hybrid, as previously thought.

Introgression

We found introgression of AD_2 alleles into AD_1 cultivars by identifying SNPs between the wild AD_1 lines (TX-231, TX-2094, and GB-319) and the AD_2 lines (excluding GB-319). There were 3,558,401 and 1,913,744 diagnostic SNPs on the A_T- and D_T-genomes, respectively. Using a novel application of PolyCat, reads from each AD_1 cultivar were categorized as AD_1-like or AD_2-like. Regions with at least 10x coverage of AD_2-like reads were identified with Eflen (part of BamBam) (Page et al. 2014). Genes in these introgressed regions were identified with BEDTools (Quinlan and Hall 2010). On average, each AD_1 accession had 6.8 Mbp (containing 1,605 genes) of introgression on the A_T-genome and 3.8 Mbp (containing 1,934 genes) of introgression on the D_T-genome (Table 3; Figure 5). The larger number of genes with introgression on the D_T-genome could indicate artificial selection acting preferentially on the D_T-genome during the domestication process, consistent with the large number of fiber QTLs found on the D_T-genome in previous studies (Rong et al. 2007).

We performed a similar analysis to look for introgression of AD_1 alleles into AD_2 cultivars. Between the AD_1 cultivars and the wild AD_2 lines (everything except DP-340, Phy-76, and GB-236), we identified 5,217,270 and 2,803,879 diagnostic SNPs on the A_T- and D_T-genomes, respectively. We then used PolyCat to categorize reads from DP-340, Phy-76, and GB-236 as AD_1-like or AD_2-like. On average, each AD2 cultivar had 18.4 Mbp (containing 1,731 genes) of introgression on the A_T-genome and 5.0 Mbp (containing 1,679 genes) of introgression on the D_T-genome. Interestingly, GB-236—which is an obsolete cultivar—had far fewer genes with evidence of introgression than the other cultivars. There was an especially large difference in the number of introgressed genes on the A_T-genome, suggesting that more recent breeding has targeted more A_T-genome genes. In addition, the A_T-genome has far more noise in the introgression signal than the D_T-genome, suggesting that relaxed selective pressures have allowed the accumulation of more mutations that give false positive introgression at isolated loci.

These introgressed regions contain hundreds of different genes that may have been individual targets of introgression, or they may have been remnants linkage 'drag' after a limited number of backcrosses recovered a suitable genotype. Further research is needed to associate the genes targeted by introgression and the historical breeding objectives of individual cultivar pedigrees.

Copy Number Variants

Copy number variants (CNVs) indicate regions of historic duplication and/or deletion, and there are various strategies used to identify them (Zhao et al. 2013). CNVs were detected in the BAM alignment files by comparing sequence coverage in sliding windows across the pseudomolecules, using CNVKit (Talevich et al. 2014). Both deletions and duplication CNVs

were detected by CNVkit, using a log base 2 threshold of 1.0 for duplications and -1.0 for deletions. Deletions in the A_T-genome were the longest and most common type of copy number variant, with ~69 blocks and 19,000 Kbp per accession (Fig. 6). Deletions in the D_T-genome were much less frequent, with ~31 blocks and less than 5,000 Kbp per accession. Duplications were considerably less frequent than deletions, with less than 10 blocks and 1,000 Kbp per accession. In the D_T genome, a similar number of duplications were found in AD_1 and AD_2, but A_T-genome duplications were more common in AD_1 than in AD_2. No pattern in frequency of duplications or deletions appeared to distinguish wild and domesticated lines. In comparisons between species, AD_4 had few duplications and deletions, and had a particularly low number of D_T-genome duplications.

Deletions were much more conserved than duplications, although this is likely related to the larger number of deletions detected, making shared deletions more likely to occur at random (Table 4). Duplications in the A_T-genome were more conserved than duplications in the D_T-genome, but duplications differed greatly from accession to accession, even among the closely related AD_1 cultivars. Generally, conservation rates of CNVs were higher in cultivars than in wild accessions, again this could have been the result of an evolutionarily recent shared ancestry.

Homoeologous Conversion

We compared copy number variants in tetraploids to identify duplications of one loci and a corresponding deletion at its respective homoeolog (*i.e.* a 'conversion' event) using the homoeolog gene pairs identified above. Very few putative homoeologous conversions of this type were detected (Table 5). Almost all genes that appeared to be converted were contained in one large possible conversion event, which was detected in several accessions; however, various

facts suggest that it was not a true conversion event: 1) the accessions exhibiting this possible conversion are not monophyletic. They include accessions of AD_1 and AD_2, but not the other members of those species. 2) The duplication associated with this possible conversion event is ubiquitous among tetraploid lines, suggesting that it may be an artifact, while the deletion associated with the possible conversion occurred in only a subset of the individuals with the duplication. 3) If the conversion event were real and occurred in a common ancestor of AD_1 and AD_2 but was only detectable in certain descendants, then we would expect it to be a much smaller event, as recombination would fractionate the affected region since AD_1 and AD_2 divergence 1-2 mya. For these reasons, we conclude that this possible conversion event—the only major event suggested by the data—was not a true conversion event and is instead an artifact of the analyses employed. Based on this conclusion, large blocks of homoeologous conversion within the cotton polyploid nucleus must be extremely rare, if they exist at all.

Other conversion events could be small remnants of a larger homoeologous conversion region that was subsequently fractionated by homologous recombination. These small blocks of interspersed, converted loci were detected by a parsimony-based method, similar to that employed by other studies (Guo et al. 2014). Reads were categorized to the A_T- and D_T-genomes with PolyCat, in order to allow intergenomic comparison at a nucleotide level. Genotypes were called using InterSnp with a minimum allele coverage of 5 reads. Polymorphic loci were selected where 75% of individuals had a genotype. These were tested for a genotype pattern indicative of homoeologous conversion in *G. hirsutum* and *G. barbadense* by comparing the tetraploid genotypes to the diploids (as a proxy ancestor genotype). Precisely, we looked for loci where the A_2 and D_5 alleles disagreed, but the A_T- and D_T-genome alleles matched each other *and* one of the two diploid alleles. Except for a significant caveat, such a simple pattern would suggest that

the allele of the matching diploid relative had overwritten the allele of the differing relative resulting in homoeologous conversion.

The caveat to simple homoeologous conversion detection lies in the diploid genomes that were used as reference proxies. The diploids A_2 and D_5 do not precisely represent the true progenitors of the A_T- and D_T-genomes (Wendel and Cronn 2003). Mutations that have occurred in the extant diploid after their divergence from the progenitors of the polyploid will result in false positive events of simple conversion detection because both tetraploid genomes will match the diploid that didn't have the mutation. To correct for this, we use the outgroup AD_4, which diverged from the other polyploid lineages shortly after polyploidization. If a putative homoeologous conversion was detected in AD_4 as well as in AD_1 and/or AD_2, then it was likely due either to a conversion event immediately after (or coincidental) with polyploidization or to an autapamorphic mutation in one of the diploid lines (Salmon et al. 2009; Flagel et al. 2012). For example, the equivalence of $A_2 = A_T = D_T \neq D_5$ could be due to a mutation in the D_5 lineage, rather than to a homoeologous conversion. Using the D_5 reference sequence, of 1,322,948 putative A-dominant events found in AD_1 that could be analyzed using AD_4, only 52,680 (4.0%) were likely homoeologous conversion events under these criteria (similar numbers were observed for AD_1 and AD_2). The remaining 1,270,268 were false positives (autapamorphies in the D_5 diploid) or occurred immediately after polyploidization (Table 6). More—but still relatively few—likely D-dominant conversion events were found: 65,276 (6.7%) out of 979,045. When using the A_2 reference sequence, however, more A-dominant than D-dominant conversions were detected, suggesting that the choice of reference sequence may be a source of further false positive events. Similar ratios of true and false conversion events were observed in AD_1 and AD_2, and a little less than half of the likely homoeologous conversion loci were shared

by AD_1 and AD_2, suggesting events prior to the division between the AD_1 and AD_2 clades. Many more regions of consecutive homoeologous conversion SNPs were detected as dominant for the same genome as the reference (Table 7). Regardless of the reference, slightly more D_T-dominant ancient homoeologous conversion regions (with 2 or more consecutive homoeologous conversion SNPs) were detected. On the A_2-reference sequence, fewer consecutive SNPs representing fewer regions were found, but they overlapped more genes. These results suggest that the contribution of homoeologous conversion to genetic diversity in cotton may be less important than previously asserted (Guo et al. 2014; Li et al. 2015).

DISCUSSION

Evolution of Tetraploid Species

Many fates are possible for duplicate gene pairs arising from a polyploidization event (Ohno 1970). In cotton, disparate outcomes have been observed for numerous pairs of genes in terms of expression level, but no general pattern has been observed as especially dominant (Rambani et al. 2014; Samuel Yang et al. 2006; Flagel et al. 2012; Yoo et al. 2012). The current study examines genotypic differences, instead of gene expression levels, and we find no obvious, generalized patterns that are consistently different between the genomes of the cotton polyploid nucleus. SNP density was not strongly correlated—positively or negatively—between homoeologs. That is, a gene with high SNP density in the A_T-genome was not much more or much less likely to have a homoeolog with high SNP density in the D_T-genome. If a positive correlation had been found, it would suggest that genes under strict selective pressure in the A-genome are also under strict selective pressure in the D-genome, resulting in strong positive correlation between the SNP densities of A_T- and D_T-genome homoeologs. On the other hand, a

negative correlation would have suggested that homoeologs tend to act as duplicate gene pairs in which one homoeolog is able to carry out the original function of the gene, while the other homoeolog is released from selective pressure, able to accumulate mutations, and develops new or more specialized functions or loses its original function (Ohno 1970). The lack of either a strong positive or a negative correlation pattern suggests that selection pressures on duplicate gene pairs likely vary pair-by-pair depending on their function and that selection pressures on these functions are small to modest across most genes in the genome. These findings also suggest that more than 1-2 million years is needed for generalized sub-functionalization throughout the genome.

Our results cast light on the phylogenetic relationships among tetraploid species, including a newly characterized species, *G. ekmanianum* (AD_6), and a possible new species of the Wake Island Atolls (Krapovickas et al. 2008; Grover et al. 2014). Previous work had constructed cotton phylogenies based on select genes or loci (Senchina 2003). However, we use an unprecedented breadth and depth of data in cotton with SNPs from across the nuclear genome, resulting in over 48 million allele-SNPs. Other studies have disputed the species status of AD_6 and suggested that it is merely wild AD_1 (Coppens d'Eeckenbrugge and Lacape 2014). However, our results show that AD_6 and AD_7 are both external to the wild AD_1 accessions TX—231 and TX-2094 and they are also distant from the AD_1 cultivars. AD_6 and AD_7 form distinct clades that cannot be considered as one monophyletic species. We conclude that the species status for AD_6 (*G. ekmanianum*) and proposed AD_7 are supported by whole genome sequence data. GB-319, although labeled an AD_2 line, is clearly not AD_2, as it consistently clusters within the AD_1 clade. Genotypic (SSR) and phenotypic data also suggest that GB-319 is a wild AD_1 rather than AD_2, corroborating this result (Richard Percy et al., personal communication).

These allele-SNPs form the beginning of a Cotton HapMap, similar to the database of SNPs constructed for the maize HapMap (Chia et al. 2012). Our homoeo-SNP indices augment this database, resulting in a database of over 70 million SNPs among cotton species, organized according to their status as homoeo-SNPs between genome groups and allele-SNPs within genome groups. These SNPs are available for visualization and download on CottonGen (http://www.cottongen.org/).

Domestication in Tetraploid Cotton

Artificial selection associated with domestication causes a genetic bottleneck in all domesticated plant species. This bottleneck results in cultivars having less genetic diversity compared to wild lines, as seen in WGS data of recent studies of soybean and tomato (Zhou et al. 2015; Lin et al. 2014). This phenomenon was observed in the AD_1—and to a lesser degree AD_2—cultivars, as manifested in the tight clustering of cultivars within the SNP-based phylogenetic trees. Small amounts of genetic diversity impose limits on the genetic potential of cotton breeding, since limited genetic diversity remained after domestication. Based on these genomic sequences, significant genetic diversity exists in wild accessions of both *G. hirsutum* and *G. barbadense*. Some of the wild accessions sequenced here could be used for sources of additional genetic diversity in breeding programs. An effort to sequence all of the genetic diversity within cultivated and wild cotton accessions could provide a comprehensive perspective to inform genetic improvement of cotton.

Domestication increased the conservation of copy number variants (duplications and deletions) among cultivars as opposed to wild cotton lines. This reflects the bottleneck effect associated with domestication. It has been observed that selection acted more heavily on the D_T-

genome than on the A_T-genome in AD_1 (Rong et al. 2007). However, our study shows that A_T-genome duplications were more (~2x) conserved than D_T-genome duplications in AD_1 cultivars, although not in AD_2. This suggests that, while selection may have acted more strongly on D_T-genome nucleotide sequence changes, selection also acted on A_T-genome duplications. Also, A_T-genome deletions were more conserved than D_T-genome deletions in AD_2 but not in AD_1. As these AD_2 lines were mostly wild, it's unlikely that this conservation was caused by artificial selection for those deletions. Rather, these deletions likely occurred in the ancestral AD_2 line that gave rise to the modern species, possibly contributing to the speciation and fiber quality that distinguish AD_2 from other tetraploid cotton species. Both of these findings (greater duplication in A_T of AD_1 and deletions in D_T of AD_2) support the independent domestication events for these two species.

Evidence of past attempts to introduce desirable traits from AD_1 into AD_2, or *vice versa*, was detected in the introgression detected in AD_1 and AD_2 cultivars. Some regions—including large regions of A_T-Chr8—exhibited evidence of introgression in all AD1 cultivars, suggesting an earlier event, while other, larger regions—e.g., D_T-Chr09—evidenced introgression in a smaller number of cultivars, suggesting more recent introgression in the pedigrees of these lines. Breeders have long attempted to introduce genes for disease resistance, fiber quality, and other traits between AD_1 and AD_2, and we are now able to see the footprint left by those efforts (Zhang and Percy 2007; Paterson 2009). We can observe an obsolete cultivar (GB-236) both in its lack of genes commonly introgressed in other cultivars and its increased noise as selective pressure is decreased. In addition to introducing specific, targeted traits, new combinations of introgression may provide an additional source of diversity for the extremely narrow germplasm of AD_1 cultivars.

Homoeologous Conversion

In diploid organisms, gene conversion is considered a by-product of recombination where one allele is reconstructed using the second allele as a template (Pessia et al. 2012). In polyploids, a conversion event that uses homoeologous loci as a template is also possible to result in conversion between genomes (Salmon et al. 2009; Flagel et al. 2012). To distinguish between the traditional definitions of homoeologous conversion, we refer the sequence-based events found between genomes sharing a nucleus as homoeologous conversions. They were likely caused by historical non-reciprocal homoeologous recombination and it results in a region of a chromosome that is converted to the genotype of its homoeolog. Assuming this region was larger than the size of an Illumina sequence read (100 bp in this study), reads originating in the converted area would be incorrectly categorized as belonging to the homoeologous genome. For example, if the D_T-genome overwrites a section of the A_T-genome, then reads from that region were categorized as D_T-genome, even though they were from an A_T-genome chromosome.

Two different methods can be used to search for two different types of homoeologous conversion: large blocks of homoeologous conversion and small, interspersed regions of SNP patterns that suggest homoeologous conversions. Large blocks of homoeologous conversion manifests as duplication in one genome and deletion in the homoeologous region of the other genome Overlapping duplication/deletion events have been detected in *Brassica* in whole genome sequencing data and these coverage and SNP patterns were attributed to non-reciprocal homoeologous recombination events (Sharpe et al. 1995; Chalhoub et al. 2014; Udall 2005), reminiscent of chromosome rearrangements observed by RFLP patterns in *B. napus* (Udall 2005; Sharpe et al. 1995). Because these large blocks of conversion have not been fully dissected by

homologous recombination, they are likely recent events between the genomes. Non-reciprocal homoeologous recombination has not been detected in cotton using RFLPs, SSRs, or sequence data, as it has in *Brassica*. Because homoeologous chromosomes in cotton have radically different sizes, it is not surprising that homoeologous conversion may occur relatively frequently on an evolutionary scale in other polyploid species, but only rarely in tetraploid cotton, where detection of non-reciprocal homoeologous recombination has not been detected by any molecular technology. We did not detect any likely homoeologous conversion events in polyploid cotton using CNV approaches to analyze WGS sequence data.

SNPs can also be used to detect putative homoeologous conversion as a consecutive pattern of shared nucleotides between diploid and tetraploid genotypes along the chromosome. The vast majority of such pattern occurrences—both in our analysis and in that done by Guo et al.—were positioned before the divergence of AD_4 from the other polyploid species (Guo et al. 2014). As little time is estimated to have passed between the polyploidization event itself and the divergence of AD_4, it is far more likely that these apparent homoeologous conversions are actually due to mutations arising in one of the diploid genomes during the 1-2 million years that separated the extant diploids from the true progenitors of the tetraploid genomes. Many mutations undoubtedly arose during that time, and nearly every such mutation would result in an apparent homoeologous conversion in the tetraploids (Flagel et al. 2012). This hypothesis is supported by the fact that apparent autapomorphies were much more common in D_5 than in A_2, reflecting the greater genetic distance between D_5 and D_T than between A_2 and A_T.

Examining putative homoeologous conversion events via SNP patterns, we observed 5% (or less) likely homoeologous conversions (as opposed to likely autapamorphic mutations). This value is consistent with EST work predating the use of the reference sequences, which also

suggested the possibility of autapamorphic SNPs yielding false positives for homoeologous conversion detection (Flagel et al. 2012). Homoeologous conversion detection was biased to favor dominant conversion events for the genome corresponding to the reference sequence used in the analysis, supporting the idea that many detected homoeologous conversions are artifacts of analysis. The larger size of the A-genome may result in false mappings of A_T-genome reads to the D_5 reference. Indeed, we observe a higher rate of tetraploid reads categorized as A_T by PolyCat. Unlike PolyDog, Polycat only examines homoeologous regions, which should limit the A_T- and D_T-genomes to nearly identical sequence lengths, despite the difference in total genome length for the A_T- and D_T-genomes (Page and Udall 2015). Thus, the increased number of reads categorized by PolyCat to the A_T-genome rather than the D_T-genome may represent false positive mappings. This increase could be due to duplications in the A-genome, resulting in two A-genome loci corresponding to a single locus in the D-genome reference. It could also be due to the closer relationship between A_T-genome and its diploid relative than between the D_T-genome and its diploid relative. Regardless of the specific cause, these false positive mappings may result in an overestimate of D-dominant homoeologous conversion events, as detected by both Guo et al. and the current study.

Resequencing the tetraploid genome of cotton provided insights into domestication, introgression, and homoeologous conversion in both *G. hirsutum* and *G. barbadense*. Our whole genome sequencing data supports the previously described independent domestication of these two polyploid species. The large genome-wide collection of SNPs between and within genomes provided an unprecedented examination of the single-nucleotide genetic diversity within the cotton genome, but a comprehensive assessment is not entirely complete. Additional re-sequencing of wild and domesticated cotton accessions will identify rare

alleles (important for diversity starved crops), provide sufficient power for robust estimates of linkage disequilibrium, and further identify regions of unique sequence evolution. The resequencing data also provided insights into the polyploid nature of the tetraploid cotton genome. Allopolyploidy has been a key aspect of cotton evolution and necessitates special computational considerations of short read sequence data because of the close sequence similarity of homoeologs. In general, the sequences of homoeologs of cotton genome have not significantly changed after polyploidization, though some exceptions can be found in individual gene pairs. Further research is needed to identify any association between these exceptions and the phenotype of modern cotton.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW. 1990. Basic local alignment search tool. *Journal of molecular ….*

Brubaker CL, Paterson AH, Wendel JF. 1999. Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. *Genome*.

Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, et al. 2014. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* **345**: 950–953.

Chia J-M, Chia J-M, Song C, Song C, Bradbury PJ, Bradbury PJ, Costich D, Costich D, de Leon N, de Leon N, et al. 2012. Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet* **44**: 803–807. http://dx.doi.org/10.1038/ng.2313.

Coppens d'Eeckenbrugge G, Lacape J-M. 2014. Distribution and Differentiation of Wild, Feral, and Cultivated Populations of Perennial Upland Cotton (Gossypium hirsutum L.) in Mesoamerica and the Caribbean ed. X. Zhang. *PLoS ONE* **9**: e107458.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.

Drummond AJ, Ashton B, Buxton S, Cheung M. 2011. Drummond: Geneious v5. 4 - Google Scholar.

Felsenstein J. 2002. {PHYLIP} (Phylogeny Inference Package) version 3.6a3. *(2002)*.

Flagel LE, Wendel JF, Udall JA. 2012. Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. http://www.biomedcentral.com/1471-2164/13/302.

Fujimoto R, Sugimura T, Nishio T. 2006. Gene conversion from SLG to SRK resulting in self-compatibility in Brassica rapa. *FEBS Letters* **580**: 425–430.

Gaeta RT, Chris Pires J. 2009. Homoeologous recombination in allopolyploids: the polyploid ratchet. *New Phytologist* **186**: 18–28.

Grover CE, Grupp KK, Wanzek RJ, Wendel JF. 2012. Assessing the monophyly of polyploid Gossypium species. *Plant Syst Evol* **298**: 1177–1183.

Grover CE, Wendel JF. 2010. Recent Insights into Mechanisms of Genome Size Change in Plants. *Journal of Botany* **2010**: 1–8.

Grover CE, Zhu X, Grupp KK, Jareczek JJ, Gallagher JP, Szadkowski E, Seijo JG, Wendel JF. 2014. Molecular confirmation of species status for the allopolyploid cotton species, Gossypium ekmanianum Wittmack. *Genet Resour Crop Evol* 1–12.

Guo H, Wang X, Gundlach H, Mayer KFX, Peterson DG, Scheffler BE, Chee PW, Paterson AH. 2014. Extensive and Biased Intergenomic Nonreciprocal DNA Exchanges Shaped a Nascent Polyploid Genome, Gossypium (Cotton). *Genetics* **197**: 1153–1163.

Hawkins JS, Proulx SR, Rapp RA, Wendel JF. 2009. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proceedings of the National Academy of Sciences* **106**: 17811–17816.

Krapovickas A, Seijo G, Seijo JG. 2008. Gossypium ekmanianum (Malvaceae), algodon silvestre de la Republica Dominicana. *Bonplandia*.

Lee S-B, Kaittanis C, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H. 2006. The complete chloroplast genome sequence of Gossypium hirsutum: organization and phylogenetic relationships to other angiosperms. *BMC Genomics* **7**: 61.

Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, et al. 2015. Genome

sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. *Nat Biotechnol*.

Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C, et al. 2014. Genome sequence of the cultivated cotton Gossypium arboreum. *Nat Genet* **46**: 567–572.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J. 2009. The sequence alignment/map format and SAMtools. ....

Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, et al. 2014. Genomic analyses provide insights into the history of tomato breeding. *Nat Genet* **46**: 1220–1226.

Ohno S. 1970. *Evolution by gene duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.

Page JT, Gingle AR, Udall JA. 2013a. PolyCat: A Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms. *G3: Genes|Genomes|Genetics* **3**: 517–525.

Page JT, Huynh MD, Liechty ZS, Grupp K, Stelly D, Hulse AM, Ashrafi H, Van Deynze A, Wendel JF, Udall JA. 2013b. Insights into the Evolution of Cotton Diploids and Polyploids from Whole-Genome Re-sequencing. *G3: Genes|Genomes|Genetics* **3**: 1809–1818.

Page JT, Liechty ZS, Huynh MD, Udall JA. 2014. BamBam: genome sequence analysis tools for biologists. *BMC Research Notes* **7**: 1–5.

Page JT, Udall JA. 2015. Methods for mapping and categorization of DNA sequence reads from allopolyploid organisms. *BMC Genetics* **16**: S4.

Paterson A. 2009. *Genetics and Genomics of Cotton*. Springer Science & Business Media.

Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC,

Shu S, Udall J, et al. 2012. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature* **492**: 423–427.

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. 2012. Evidence for Widespread GC-biased Gene Conversion in Eukaryotes. *Genome Biology and Evolution* **4**: 675–682.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. **26**: 841–842. http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq033.

Rambani A, Page JT, Udall JA. 2014. Polyploidy and the petal transcriptome of Gossypium. *BMC Plant Biology* **14**: 3.

Rong J, Feltus FA, Waghmare VN, Pierce GJ, Chee PW, Draye X, Saranga Y, Wright RJ, Wilkins TA, May OL, et al. 2007. Meta-analysis of Polyploid Cotton QTL Shows Unequal Contributions of Subgenomes to a Complex Network of Genes and Gene Clusters Implicated in Lint Fiber Development. *Genetics* **176**: 2577–2588.

Salmon A, Flagel L, Ying B, Udall JA, Wendel JF. 2009. Homoeologous nonreciprocal recombination in polyploid cotton. *New Phytologist* **186**: 123–134.

Samuel Yang S, Cheung F, Lee JJ, Ha M, Wei NE, Sze S-H, Stelly DM, Thaxton P, Triplett B, Town CD, et al. 2006. Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton. *The Plant Journal* **47**: 761–775.

Senchina DS. 2003. Rate Variation Among Nuclear Genes and the Age of Polyploidy in Gossypium. **20**: 633–643. http://mbe.oupjournals.org/cgi/doi/10.1093/molbev/msg065.

Sharpe AG, Parkin IAP, Keith DJ, Lydiate DJ. 1995. Frequent nonreciprocal translocations in the amphidiploid genome of oilseed rape (Brassica napus). *Genome*.

Stajich JE, Block D, Boulez K, Brenner SE. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome*.

Talevich E, Shain AH, Bastian BC. 2014. CNVkit: Copy number detection and visualization for targeted sequencing using off-target reads. *bioRxiv* 010876.

Teshima KM. 2004. The Effect of Gene Conversion on the Divergence Between Duplicated Genes. *Genetics* **166**: 1553–1560.

Udall JA. 2005. Detection of Chromosomal Rearrangements Derived From Homeologous Recombination in Four Mapping Populations of Brassica napus L. *Genetics* **169**: 967–979.

Wang Z, Zhang D, Wang X, Tan X, Guo H, Paterson AH. 2013. A Whole-Genome DNA Marker Map for Cotton Based on the D-Genome Sequence of Gossypium raimondii L. *G3: Genes|Genomes|Genetics* **3**: 1759–1767.

Wendel JF. 1989. New World tetraploid cottons contain Old World cytoplasm. *Proceedings of the National Academy of Sciences of the United States of America* **86**: 4132–4136. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=16594050&retmode=ref&cmd=prlinks.

Wendel JF, Albert VA. 1992. Phylogenetics of the cotton genus (Gossypium): character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and …. *Systematic Botany*.

Wendel JF, Cronn RC. 2003. Polyploidy and the evolutionary history of cotton. In *Advances in Agronomy*, Vol. 78 of, pp. 139–186, Elsevier http://linkinghub.elsevier.com/retrieve/pii/S0065211302780048.

Wendel JF, Flagel LE, Adams KL. 2012. Jeans, Genes, and Genomes: Cotton as a Model for Studying Polyploidy. In *Polyploidy and Genome Evolution*, pp. 181–207, Springer Berlin

Heidelberg, Berlin, Heidelberg.

Wendel JF, Grover CE. 2015. Taxonomy and Evolution of the Cotton Genus, Gossypium. In *Cotton*, *Agronomy Monograph*, American Society of Agronomy, Inc., Crop Science Society of America, Inc., and Soil Science Society of America, Inc.

Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*.

Yoo MJ, Szadkowski E, Wendel JF. 2012. Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* **110**: 171–180.

Zhang JF, Percy R. 2007. Improving upland cotton by introducing desirable genes from pima cotton. *World Cotton Research Conference-4, Lubbock, Texas, USA, 10-14 September 2007*.

Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, Zhang J, Saski CA, Scheffler BE, Stelly DM, et al. 2015. Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. *Nat Biotechnol*.

Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14**: S1. http://www.biomedcentral.com/1471-2105/14/S11/S1.

Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, et al. 2015. resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol* **33**: 384–389.

TABLES

Table 9 Homoeo-SNPs identified between the A- and D-genomes of the diploids and A_T- and D_T-genomes of the tetraploids

	A2-reference			D5-reference		
	Genomic	**Genic**		**Genomic**	**Genic**	
Diploids	15,618,185	2,090,126	13.4%	28,540,537	3,009,100	10.5%
AD_1	18,253,297	2,303,433	12.6%	24,908,821	3,069,346	12.3%
AD_2	17,286,282	2,224,161	12.9%	24,776,502	3,003,401	12.1%
AD_3	12,574,385	2,044,681	16.3%	19,235,460	2,742,627	14.3%
AD_4	12,442,214	1,973,277	15.9%	19,274,313	2,656,550	13.8%
AD_5	12,914,212	2,017,762	15.6%	19,809,248	2,719,911	13.7%

Table 10 SNPs and average diversity (# pairwise differences / # polymorphic sites covered by both individuals) among sub-groups of diploids and tetraploids

	A_T		D_T	
Group	**SNPs**	**Diversity**	**SNPs**	**Diversity**
All	27,447,974	0.165%	21,476,013	0.285%
AD	15,864,224	0.132%	10,437,663	0.179%
AD_1-clade	9,555,028	0.060%	6,574,982	0.099%
AD_1-dom	7,875,126	0.048%	5,610,018	0.092%
AD_2	9,489,947	0.048%	6,376,241	0.085%

Table 11 Evidence for introgression between AD_1 and AD_2

		A_T Introgression		D_T Introgression	
	Accession	**Length**	**Genes**	**Length**	**Genes**
AD_2 into AD_1	Coker-312	6,352,569	1,608	3,495,471	1,887
	DP-5690	2,508,579	640	1,137,344	651
	FM-832	6,650,745	1,809	5,314,686	2,433
	Maxxa	8,558,032	1,998	4,596,128	2,313
	MS-240	8,427,778	1,770	3,757,081	1,878
	PD-1	6,556,943	1,564	3,325,321	1,920
	Sealand	7,123,054	1,670	3,858,004	1,985
	SG-747	7,656,250	1,604	3,630,665	1,890
	ST-474	6,970,667	1,661	3,704,376	1,936
	Tamcot-sphinx	6,606,198	1,750	8,510,980	2,440
	TM-1	6,817,357	1,605	3,803,304	1,945
	Average	**6,748,016**	**1,607**	**4,103,033**	**1,934**
AD_1 into AD_2	DP-340	21,707,123	2,146	5,878,386	1,938
	Phy-76	18,255,558	1,819	4,879,839	1,555
	GB-236	15,326,627	1,228	4,265,336	1,543
	Average	**18,429,769**	**1,731**	**5,007,854**	**1,679**

Table 12 Conserved copy number variants across sub-groups of tetraploids

Type	Group	Total Genes	Found in at least 50% of accessions	
A-duplication	AD1-domesticates	193	34	18%
	AD1-clade	258	30	12%
	All	307	21	7%
D-duplication	AD1-domesticates	188	24	13%
	AD1-clade	286	20	7%
	All	387	21	5%
A-deletion	AD1-domesticates	785	405	52%
	AD1-clade	937	385	41%
	All	1539	369	24%
D-deletion	AD1-domesticates	604	273	45%
	AD1-clade	764	216	28%
	All	1114	196	18%

Table 13 Possible large gene conversion events based on copy number variants by accession (A) and by gene (B)

A

Accessions	# Genes
AD_3	15
AD_5	16
AD_7	17
DP-340	15
FM-832	17
GB-287	15
GB-362	15
GB-398	15
GB-618	15
GB-67	16
MS-240	15
Phy-76	15
SG-747	15
TX-231	15
Maxxa	2

B

D_5 gene	A_2 gene	# Accessions
Gorai.012G000600	Cotton_A_20329	14
Gorai.012G000700	Cotton_A_20330	14
Gorai.012G000900	Cotton_A_20331	14
Gorai.012G001000	Cotton_A_20332	14
Gorai.012G001100	Cotton_A_20334	14
Gorai.012G001200	Cotton_A_20335	14
Gorai.012G001500	Cotton_A_20338	14
Gorai.012G001600	Cotton_A_20339	14
Gorai.012G001700	Cotton_A_20340	14
Gorai.012G001800	Cotton_A_20341	14
Gorai.012G001900	Cotton_A_20342	14
Gorai.012G002000	Cotton_A_20343	14
Gorai.012G002100	Cotton_A_20344	14
Gorai.012G002200	Cotton_A_20345	14
Gorai.012G002300	Cotton_A_20346	14
Gorai.012G000400	Cotton_A_33577	1
Gorai.007G329600	Cotton_A_35502	3
Gorai.007G329800	Cotton_A_35503	3
Gorai.001G162500	Cotton_A_10073	1
Gorai.002G268400	Cotton_A_00274	1
Gorai.002G268600	Cotton_A_00273	1

Table 14 Possible small gene conversion events based on genotype patterns across diploids and tetraploids

Type	**A2-reference**					
Species	**AD_1**		**AD_2**		**Both AD_1 and AD_2**	
Autapamorphy in A_2	605,147	36.1%	605,514	36.5%	599,400	37.4%
Autapamorphy in D_5	1,005,869	60.1%	996,741	60.1%	977,987	61.1%
A-dominant Conversion	41,759	2.5%	37,960	2.3%	17,130	1.1%
D-dominant Conversion	21,593	1.3%	19,493	1.2%	6,406	0.4%
	D_5-reference					
	AD_1		**AD_2**		**Both AD_1 and AD_2**	
Autapamorphy in A_2	913,769	39.7%	912,441	40.0%	900,332	41.2%
Autapamorphy in D_5	1,270,268	55.2%	1,259,994	55.2%	1,235,851	56.5%
A-dominant Conversion	52,680	2.3%	50,528	2.2%	21,857	1.0%
D-dominant Conversion	65,276	2.8%	58,898	2.6%	27,757	1.3%

Table 15 Regions including 2 or more consecutive small gene conversion SNPs

		A_2-reference		D_5-reference	
Type	**Number**	**AD_1**	**AD_2**	**AD_1**	**AD_2**
A_T-dominant	SNPs	3145	2662	2491	2636
	Regions	818	699	640	697
	Total Length (Kbp)	1636	1327	413	499
	Genes	144	143	8	6
D_T-dominant	SNPs	401	183	10769	8383
	Regions	100	45	2661	2097
	Total Length (Kbp)	747	209	3766	2793
	Genes	29	14	60	50

FIGURES

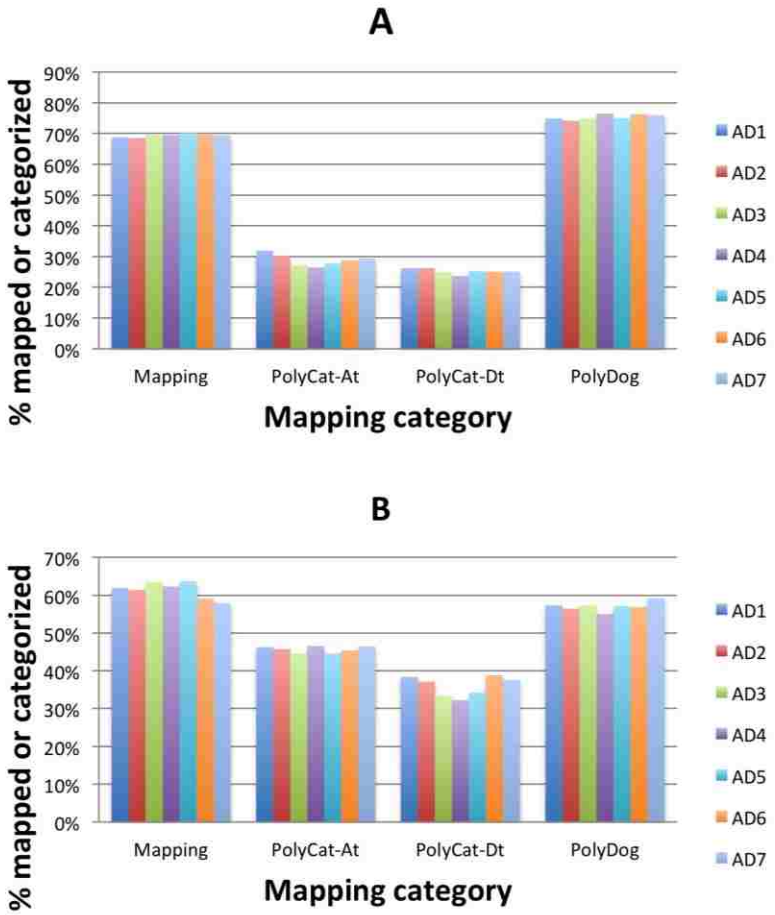


Figure 16 Mapping and categorization rates. The percentage of trimmed reads successfully mapped (by GSNAP) from each species (AD_1-AD_7) to the A_2 reference (A) and the D_5 reference (B) is shown. For each reference, the percentage of mapped reads from each species (AD_1-AD_7) categorized to the A_T-genome by PolyCat, D_T-genome by PolyCat, or to the genome of the reference sequence by PolyDog is also shown.

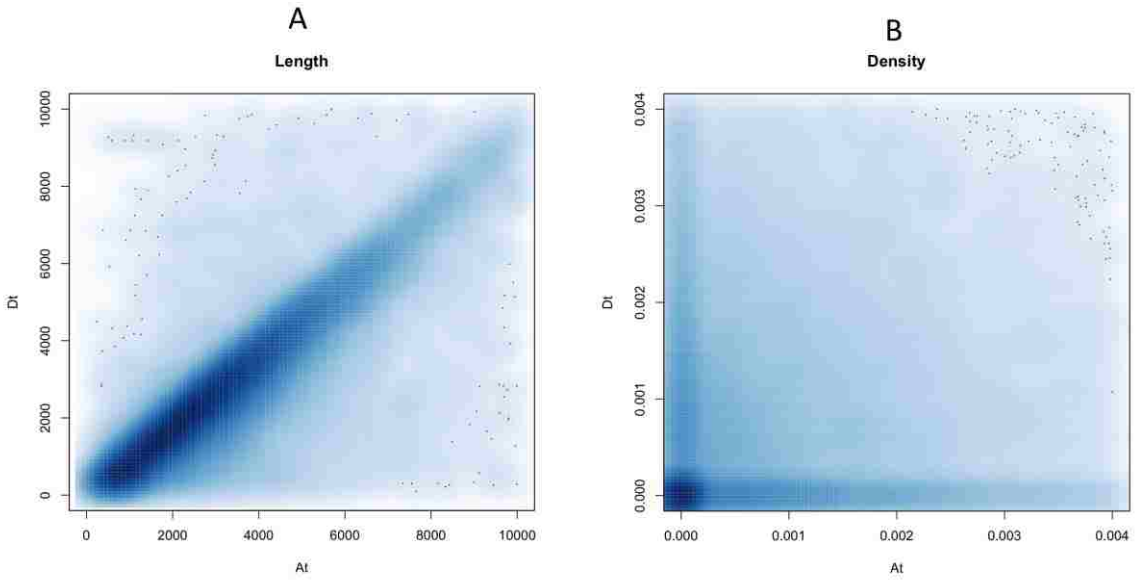


Figure 17 Homoelog length and SNP density. Homoeolog pairs in the A2 and D5 reference annotations were compared on the basis of length (A) and SNP density (B), with the A-genome copy on the X-axis and the D-genome copy on the Y-axis. Darker blue color indicates a larger number of homoeolog pairs with those values.

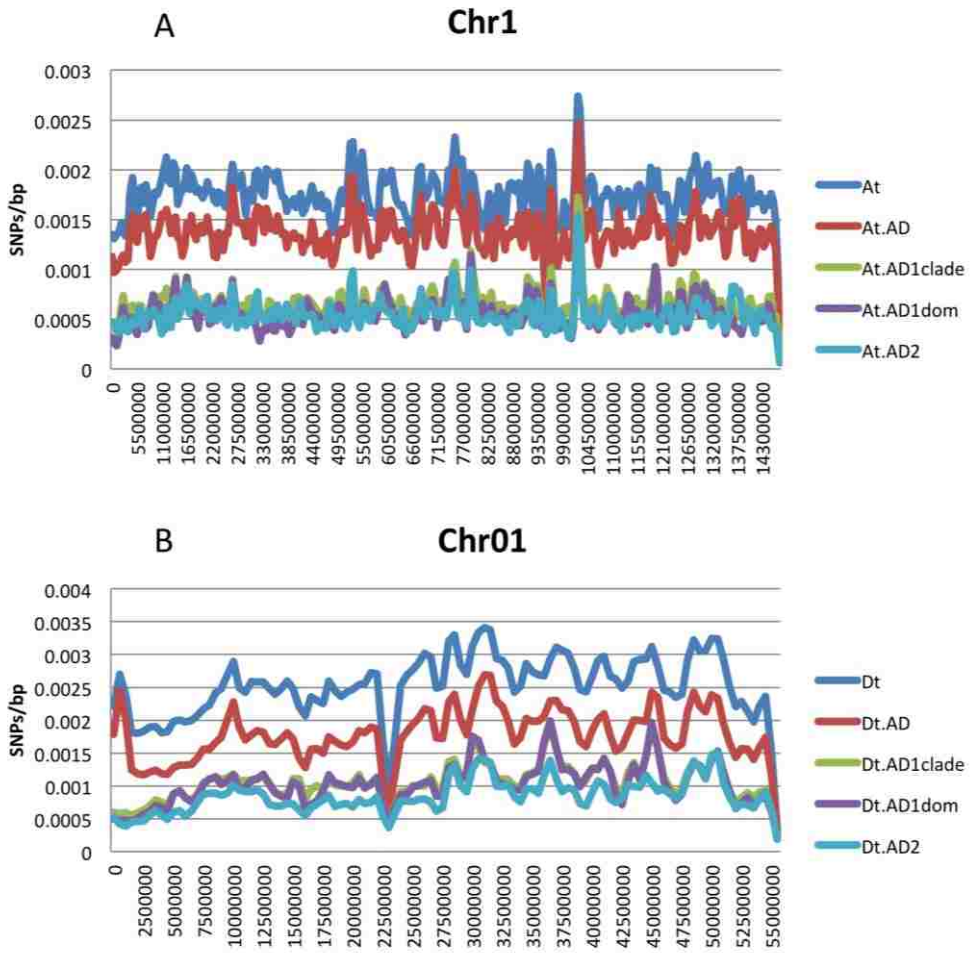
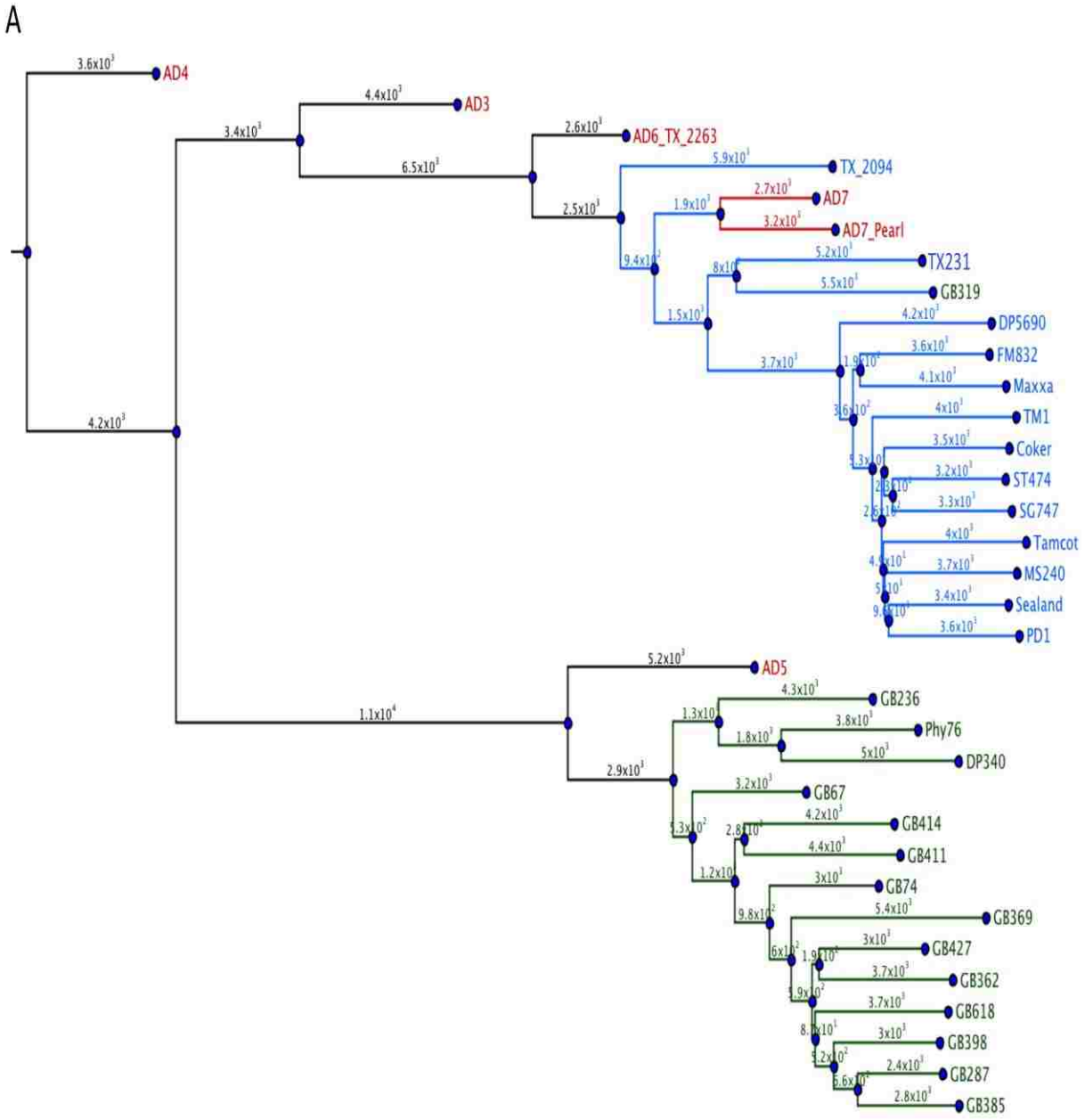


Figure 18 Average diversity rates. Diversity was measured in a sliding window across Chr1 of the A_2 reference (A) and Chr01 of the D_5 reference (B). Diversity was measured among different groups: all diploids and tetraploids (dark blue), all tetraploids (red), all lines from AD_1, AD_6, and AD_7 (green), all AD_1 cultivars (purple), and all lines from AD_2 (light blue).


A
AD4
AD3
AD6_TX_2263
TX_2094
AD7
AD7_Pearl
TX231
GB319
DP5690
FM832
Maxxa
TM1
Coker
ST474
SG747
Tamcot
MS240
Sealand
PD1
AD5
GB236
Phy76
DP340
GB67
GB414
GB411
GB74
GB369
GB427
GB362
GB618
GB398
GB287
GB385

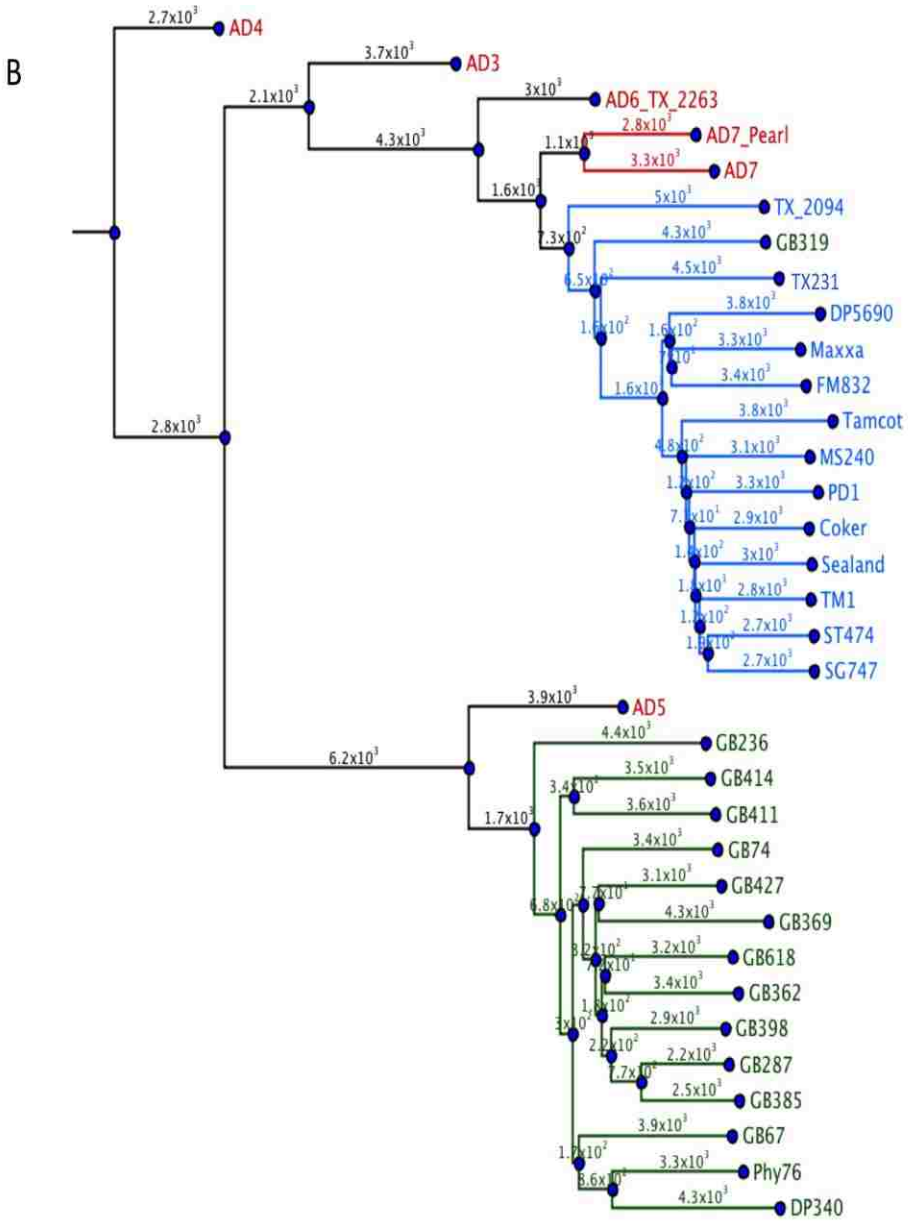


Figure 19 Phylogenetic trees. Neighbor-joining trees were constructed (by PHYLIP) based on a distance matrix representing SNPs between each pair of accessions in the A-genome (A) and D-genome (B). Only tetraploid lines are shown, with the root representing the point of connection to the diploid relatives. Individuals from AD_1 are colored blue, AD_2 colored green, and other species colored red.

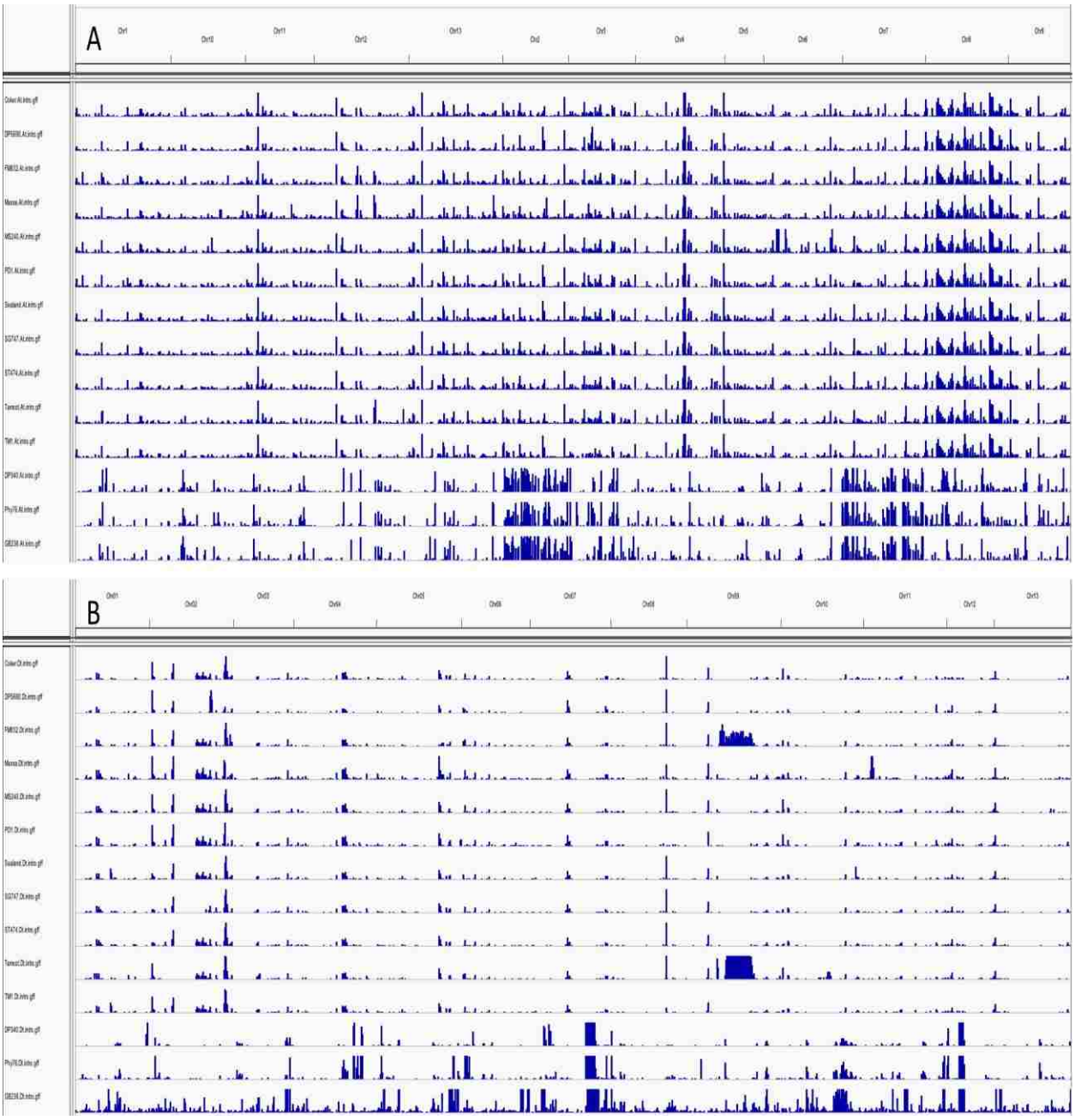


Figure 20 Introgression. Regions of introgression were identified in the A_T-genome (A) and in the D_T-genome (B). For each cultivar, blue indicates regions of introgression from the other species.

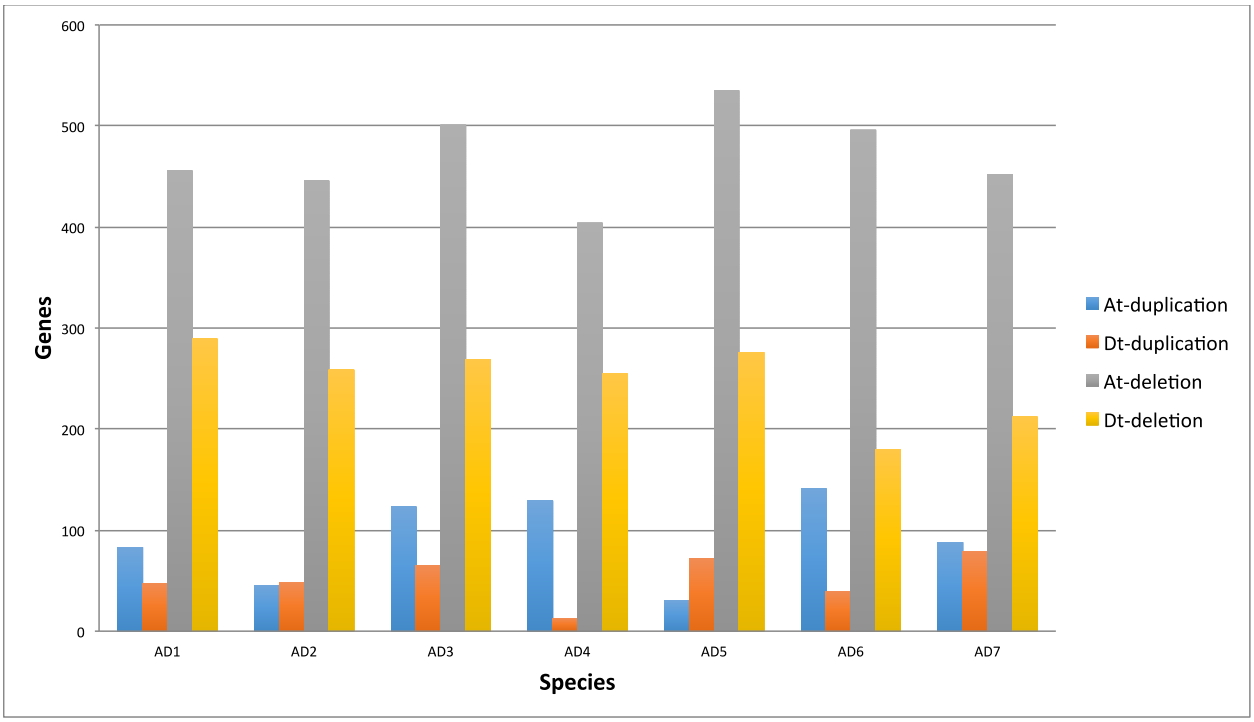


Figure 21 Duplicated and deleted genes. Duplications and deletions were identified in each genome of each species, relative to the extant diploid relative. So blue indicates duplications in A_T-genomes compared to A_2 individuals. CNVkit identified duplications and deletions, with a minimum threshold of 2-fold difference.