Persistence of Subgenomes in Paleopolyploid Cotton after 60 My of Evolution

Simon Renny-Byfield,[‡,1] Lei Gong,[1] Joseph P. Gallagher,[1] and Jonathan F. Wendel[*,1]

[1]Department of Ecology, Evolution and Organismal Biology, Iowa State University

[‡]Present address: Department of Plant Sciences, University of California, Davis, Davis, CA

***Corresponding author:** E-mail: jfw@iastate.edu.

Associate editor: Susanne Renner

Abstract

The importance of whole-genome multiplication (WGM) in plant evolution has long been recognized. In flowering plants, WGM is both ubiquitous and in many lineages cyclical, each round followed by substantial gene loss (fractionation). This process may be biased with respect to duplicated chromosomes, often with overexpression of genes in less fractionated relative to more fractionated regions. This bias is hypothesized to arise through downregulation of gene expression through silencing of local transposable elements (TEs). We assess differences in gene expression between duplicated regions of the paleopolyploid cotton genome and demonstrate that the rate of fractionation is negatively correlated with gene expression. We examine recent hypotheses regarding the source of fractionation bias and show that TE-mediated, positional downregulation is absent in the modern cotton genome, seemingly excluding this phenomenon as the primary driver of biased gene loss. Nevertheless, the paleo subgenomes of diploid cotton are still distinguishable with respect to TE content, targeting of 24-nt-small interfering RNAs and GC content, despite approximately 60 My of evolution. We propose that repeat content per se and differential recombination rates may drive biased fractionation following WGM. These data highlight the likely importance of ancient genomic fractionation biases in shaping modern crop genomes.

Key words: **whole-genome duplication, gene fractionation, biased fractionation, transposable element, gene expression.**

Introduction

Whole-genome multiplication (WGM or polyploidy) is ubiquitous and cyclical in flowering plants and is thought to have played important roles in angiosperm diversification and the success of crop plants (Paterson et al. 2000, 2004; Bowers et al. 2003; Blanc and Wolfe 2004; Leitch et al. 2008; Soltis et al. 2009; Jiao et al. 2011, 2012; Renny-Byfield et al. 2014). The realization that all flowering plants are paleopolyploid indicates that, over time, a process of diploidization operates to return polyploids to a diploid-like condition (Wolfe 2001; Clarkson et al. 2005; Mandakova et al. 2010; Renny-Byfield et al. 2013), a transition that involves the large-scale loss of duplicate genes (Langham et al. 2004; Woodhouse et al. 2010, 2014; Schnable et al. 2011; Tang et al. 2012; Garsmeur et al. 2013). This process may lead to differential loss of duplicated genes from homoeologous genomic regions, a phenomenon termed biased fractionation (Langham et al. 2004; Thomas et al. 2006). This can have long-lasting ramifications and recent work has demonstrated that the biases in gene loss following the ancient WGM event shared by *Arabidopsis* and *Brassica* were propagated through a further WGM in the *Brassica* lineage (Woodhouse et al. 2014). Thus, Woodhouse et al. showed that biases in gene loss were heritable through multiple rounds of WGM.

Recent analyses in maize (Schnable et al. 2011) and *Brassica* (Cheng et al. 2012; Woodhouse et al. 2014) indicate that homoeologous regions experiencing greater gene loss tend to have lower levels of gene expression, such that for any given syntenic paralog pair, the gene with the highest expression is likely to reside on the genomic segment that has experienced less gene loss. Similarly, more recent polyploids also experience nonequivalence of gene expression between subgenomes, for example, cotton (Flagel et al. 2008; Hovav et al. 2008), coffee (Bardil et al. 2011), and *Tragopogon* (Buggs et al. 2010). These observations have led to the hypothesis that following WGM, differences in gene expression between duplicated regions drive differential gene loss (Freeling et al. 2012). This process is suggested to be accompanied by selection, where underexpressed genes contribute less to fitness than do their overexpressed homoeologous counterparts, and therefore are more likely to become dispensable (Freeling et al. 2012). The hypothesis is an attractive one in that nascent or evolutionarily young polyploids often exhibit biased homoeolog expression (Grover et al. 2012) suggesting that WGM, especially in allopolyploids (Garsmeur et al. 2013), may establish the initial conditions that set in motion biases in gene expression and potentially gene loss.

Importantly, earlier work (Hollister et al. 2011) demonstrated that in *Arabidopsis lyrata* and *A. thaliana*, there is a negative correlation between small interfering RNA (siRNA)-mediated methylation of TEs and expression of local genes. These observations led Woodhouse et al. propose a mechanistic underpinning for differences in gene expression between duplicate subgenomes; *Brassica* homoeologous

Article

regions experiencing greater gene loss were enriched for mapping of 24-nt siRNAs to transposable elements (TEs) surrounding resident genes, and these genes typically had lower expression compared with their homoeologous counterparts, providing a corollary for the earlier observations of Hollister et al. Given that at the time of WGM subgenomes may contain variable TE content (e.g., as in allopolyploidy) Woodhouse et al (2014) suggested that differential positional-effect, downregulation by local TEs may be variable between the two coresident subgenomes and that this may drive differences in gene expression between duplicate regions, potentially providing an explanation for biased fractionation.

The recent release of a cotton reference genome (Paterson et al. 2012) provided clear evidence of an ancient WGM in diploid cotton (*Gossypium raimondii*). Importantly, this WGM is unusual in comparison to other ancient WGM in that many duplicated regions appear to exist in five or six copies, all of similar ages (~60 Ma). This led Paterson et al. (2012) to suggest that the cotton WGM represented either a single 5- to 6-fold ploidy increase or multiple temporally closely adjacent events.

Here we examine fractionation, utilizing the ancient WGM in the cotton lineage, extend the temporal window for studying biased fractionation, and demonstrate that the genomic footprints of the process persist despite 60 My of evolution. We assess differences in gene expression between duplicated regions and demonstrate that the rate of fractionation is negatively correlated with gene expression. We examine recent hypotheses regarding the source of fractionation bias and show that TE-mediated, positional downregulation is absent in the modern cotton genome, seemingly excluding this phenomenon as the primary driver of fractionation bias. We present evidence of other genomic features that distinguish the most fractionated (MF) and least fractionated (LF) components of the genome and suggest how these characteristics might drive differential gene loss. Our observations indicate that the impact of biased fractionation extends well beyond the time-scale over which it was originally identified, 10 Ma in maize (Schnable et al. 2009, 2011) and 20 Ma in *Brassica* (Wang et al. 2011; Cheng et al. 2012; Tang et al. 2012).

Results

Chromosome Reconstruction, Biased Fractionation, and Gene Expression

Using the SynMap tool of CoGe (https://genomevolution.org/CoGe, last accessed July 7, 2014), we identified blocks of genes in synteny between *G. raimondii* (diploid cotton) and its relative *Theobroma cacao* (cacao; SynMap output in supplementary file S1, Supplementary Material online). Importantly, comparative genome sequence data indicate that relative to cacao, the lineage that gave rise to modern diploid cotton experienced a 5- to 6-fold ploidy increase approximately 60 Ma (Paterson et al. 2012). Accordingly, we identified duplicate regions in the cotton genome resulting from the cotton-specific WGM (supplementary figs. S1 and S2, Supplementary Material online). The relative antiquity of WGM in cotton, however, makes identification of duplicate regions challenging compared with efforts in paleopolyploids where WGM, although still ancient, has occurred much more recently. Indeed, the genome of cotton is substantially rearranged relative to cacao (supplementary figs. S1 and S2, Supplementary Material online), and in some regions five or six duplicate segments per haploid cacao genome are evident in cotton. It was not possible to reconstruct all sets of chromosomes, as has been possible in maize (Schnable et al. 2011) and *Brassica* (Tang et al. 2012). We were, however, able to reconstruct a subset of the chromosomes, in the same manner as in Schnable et al. (2011); we restricted further analyses to those genes and genomic regions covered by the whole chromosome reconstructions (highlighted green in supplementary fig. S1, Supplementary Material online).

In total, we were able to reconstruct at least two ancestral cotton chromosomes for six of the ten preduplicated cacao chromosomes (Chromosomes 2, 6, 7, 8, 9, 10; table 1), allowing for comparisons of gene loss between homoeologous regions in the cotton genome (table 1 and supplementary fig. S1, Supplementary Material online). As the duplicated regions of the cotton genome originate from either a single event, or several temporally closely spaced events, the null expectation is that each duplicate chromosome should have the same (or almost the same) number of genes following duplication. We estimated gene loss by comparing the number of genes in each cotton reconstruction that share syntenic orthologs in the chocolate genome. Thus, our calculations are similar to those performed previously (Schnable et al. 2011), do not rely on aggregate gene content, and exclude gene insertions following the ancient WGM. The hypothesis of gene retention equivalence among homoeologs was evaluated using Pearson's chi-square test (table 1). Importantly, all tests indicate statistically significant deviation from random (equivalent) gene loss, providing strong evidence for biased gene fractionation between duplicate regions in the cotton genome. Thus, we are able to detect the signatures of chromosome-wide-biased fractionation following WGM on time-scales of approximately 60 Ma, far in excess of those previously reported in *Brassica* (Cheng et al. 2012; Tang et al. 2012; Woodhouse et al. 2014) and maize (Schnable et al. 2011).

Fractionation and Expression of Duplicate Genes

We binned chromosome reconstructions such that, for each preduplicated ancestral chromosome, the least fractionated of the postduplicated (cotton) reconstructions was designated LF and all others were grouped together and designated as MF. These classifications were used to group retained paralogous as belonging to LF or MF fractions of the genome. We then examined gene expression levels in three tissues and compared expression of syntenic paralogs residing in LF and MF fractions. In this case we took the most conservative approach and discarded any genes that lacked a suitable paralog for comparison, or genes that did not have a corresponding syntenic ortholog in the chocolate genome. Despite such a restrictive set of comparisons, this analysis revealed that genes on LF chromosomes are generally more highly expressed than

Table 1. Biased Fractionation in Cotton Following an Ancient (60 Ma) WGD.

Theobroma Cacao Chromosome	*Gossypium raimondii* Chromosome	Observed Number of Genes	Predicted Number of Genes (equivalent loss)	χ^2	*P* Value
2	5	929	785.5	52.43	4.5×10^{-13}
	8	642	785.5		
6	6	147	334	333.86	$<1 \times 10^{-15}$
	9	580	334		
	10	227	334		
7	2	420	325.5	58.95	1.6×10^{-14}
	13	225	325.5		
8	5	236	422	163.96	$<1 \times 10^{-15}$
	9	608	422		
9	4	343	575	433.79	$<1 \times 10^{-15}$
	9	981	575		
	13	400	575		
10	9	397	283.5	90.88	$<1 \times 10^{-15}$
	11	170	283.5		

NOTE.—A chi-square test was used to investigate nonequivalent gene loss in duplicated regions of the cotton genome. Reconstructed regions (chromosomes) are described by their position in *Theobroma cacao* and the corresponding region of the cotton genome (circled green in supplementary fig. S1, Supplementary Material online). We estimated the number of genes in each cotton region with syntenic orthologs in chocolate (observed number of genes), and compared these with the predicted number of genes based on equivalence of gene loss between duplicate regions within the cotton genome (as in Tang et al. 2012). All comparisons differ significantly from equivalent loss.

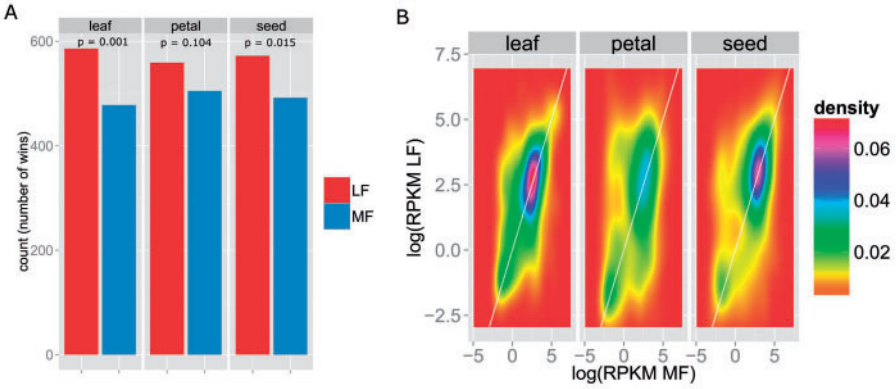

FIG. 1. Gene expression in LF and MF fractions of the genome. Syntenic paralogs were binned according to residency on LF or MF chromosome reconstructions, and compared for gene expression levels in three tissues, petal, leaf, and seed. Three biological replicates were used in each tissue comparison. (*A*) The number of genes more highly expressed in each category is given; significance deviation from an equal number of genes being more highly expressed in each category was examined using a cumulative binomial distribution with the odds of an LF gene being more highly expressed at 0.5. (*B*) Density plots comparing gene expression values for syntenic paralogs on LF and MF fragments of the genome. The white line in each plot indicates a 1:1 relationship and equal expression of syntenic paralogs.

those on MF (fig. 1); of the 1,064 paralogous gene pairs compared, LF genes were more highly expressed in 559–589 cases, depending on the tissue. To test for statistically significant deviation from a random number of upregulated genes in each group (LF or MF), we used a cumulative binomial distribution with the probability of LF being more highly expressed at 0.5. This result was statistically significant for expression data for both leaf and seed tissue, but not for petals (fig. 1). To examine the robustness of the bias, we restricted our analysis to those syntenic paralogs that varied in expression by greater than 2-fold (supplementary fig. S3*A*, Supplementary Material online). In this case, the bias was greater and statistically significant in all tissues examined. Furthermore, bias remained evident even when considering only syntenic paralogs exhibiting statistically significant differential gene expression and fold-change greater than 2, confirming previous results (supplementary fig. S3*B*, Supplementary Material online). Thus, the data indicate that genes on LF chromosomes are typically expressed at higher levels than their MF counterparts.

Local TE Density

Using TE annotations from the cotton reference genome (Paterson et al. 2012), we identified TEs near each gene and

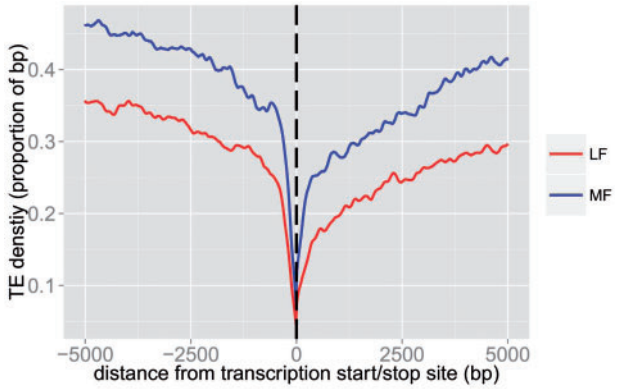

Fig. 2. TE density flanking genes in LF and MF fractions of the genome. Annotations from the cotton reference genome were used to identify TEs, and the proportion of base pairs annotated as TE-derived was calculated over sliding windows of 5,000 bp (10-bp increments), either side of transcription start/stop sites. Mean density for LF and MF genes is displayed separately. The dashed line indicates the start/stop site and the intervening genic region is excluded from the plot.

used sliding windows to calculate the average TE density in regions surrounding genes (fig. 2). Unsurprisingly, TE density is low close to transcription start/stop sites and increases further from genes. Importantly, a comparison of the local TE density for genes on LF and MF chromosomes revealed that TE density is higher in MF chromosomes (Wilcoxon's paired sum rank test, $V = 502{,}503$, $P \leq 0.0001$; fig. 2). In addition, genes residing on MF chromosomes generally have more internal TE insertions (supplementary fig. S4, Supplementary Material online) than do genes on LF fragments. Using the same sliding windows, we observed that GC content in both LF and MF fractions steadily decreases toward transcription start/stop sites and gradually increases following transcription start/stop sites (supplementary fig. S5, Supplementary Material online) and the GC content of LF chromosomes is significantly higher than that of MF 5,000 bp either side of genes models (Wilcoxon's paired sum rank test, $V = 0$, $P \leq 0.0001$).

Enrichment of Mapped siRNAs on MF Chromosomes

The higher density of TEs surrounding MF genes prompted us to ask whether those same regions upstream and downstream of genes on MF chromosomes are enriched for TE-derived, 24-nt siRNAs. We mapped 24-nt siRNAs from leaves of *G. raimondii* to the cotton reference genome, averaging the number of uniquely mapped reads along sliding windows near genes of interest. Surrounding regions of MF genes have higher proportions of TE-derived siRNAs (Wilcoxon's paired sum rank test, $V = 327{,}477$, $P \leq 0.0001$; fig. 3). In addition to enrichment of siRNAs, the MF fraction of the genome also exhibits a sharp increase in siRNAs mapping approximately 200 bp upstream and approximately 800 bp downstream of the transcription start/stop site, a pattern absent in the LF fraction of the genome (enlarged in fig. 3*A*). To examine whether preferential targeting is a general characteristic of MF genes, rather than the result of a few outliers with excessive numbers of target siRNAs, we capped the number of mapped reads at a given locus to 10 (fig. 3*B*), thus eliminating the impact of such outliers. This new analysis rendered LF and MF fragments even more readily distinguishable than with no capping of read numbers.

TE Proximity and Gene Expression

All 37,000 cotton gene models were binned according the distance to the nearest TE, taking into account both up- and downstream insertions. Using RNA sequencing (RNA-seq) data from petal, leaf, and seed of diploid cotton (Renny-Byfield et al. 2014), we examined the relationship between TE proximity and gene expression. Importantly, expression was relatively uniform across all bins and across all tissues (fig. 4*A*). However, correlation analysis revealed a statistically significant negative correlation between TE proximity and expression (Pearson's correlation coefficient = −0.03, df = 111,667, $t = -10.166$, $P \leq 0.001$). Although statistically significant, the effect is minimal and in the opposite direction of expectation and indicates that TE proximity is likely to have only a small impact on levels of local gene expression. Furthermore, linear modeling revealed an R^2 value of less than 0.001, indicating that TE proximity explains only a very small fraction of variation in gene expression. We broadened our analysis to consider local TE density, rather than the nearest TE insertion (supplementary fig. S6, Supplementary Material online). Interestingly, in this case, local TE density seems to be more strongly negatively correlated with expression of nearby genes (Pearson's correlation coefficient = −0.125, df = 116,637, $t = -42.003$, $P \leq 0.001$, $R^2 < 0.001$); nevertheless, and similarly to TE proximity, only a very weak correlation was observed.

We separately binned genes into two pools based on the presence or absence of local siRNA-targeted TEs and compared the impact of TE proximity on gene expression in the two groups (fig. 4*B*). Analyses using linear modeling and analysis of variance (ANOVA) indicate that presence of siRNAs has little impact on the expression of nearby genes (supplementary table S1, Supplementary Material online).

Discussion

The extensive occurrence of episodically recurrent WGM in the history of flowering plants has only recently become evident (Paterson et al. 2000, 2004; Bowers et al. 2003; Blanc and Wolfe 2004; Leitch et al. 2008; Soltis et al. 2009; Jiao et al. 2011, 2012; Renny-Byfield et al. 2014). This legacy of genome multiplicity is exemplified by the genome of modern cotton, which in addition to experiencing more ancient WGM events, underwent a 5- to 6-fold ploidy increase near the start of the Tertiary, well after its divergence from cacao (Paterson et al. 2012). A universal attribute of WGM is the loss of duplicated genes (Langham et al. 2004), with many returning to single copy status through deletional processes, be they biased or unbiased with respect to homoeologous chromosomes. Examples of the former are phylogenetically widespread among angiosperms, as documented in recent analyses of maize, *Brassica*, *Arabidopsis*, and poplar

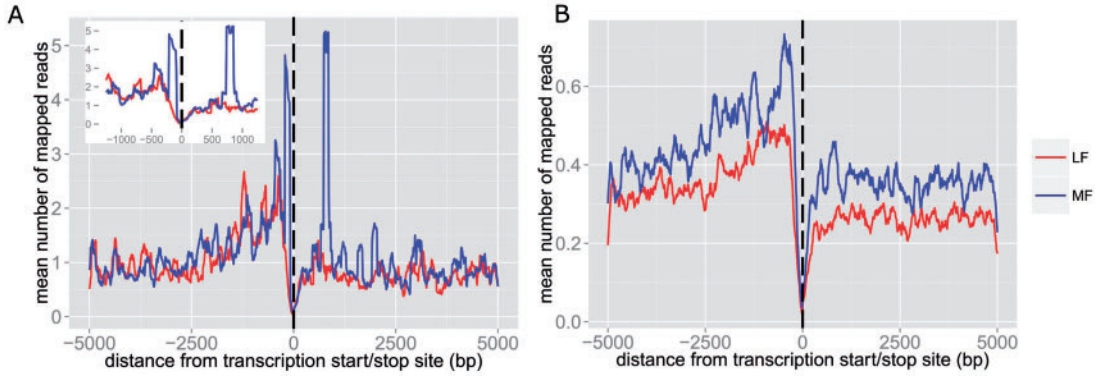

FIG. 3. Enrichment of 24-nt siRNAs mapped to TEs that flank MF genes. (*A*) All genes in LF and MF categories were assessed for uniquely mapped siRNAs, allowing no mismatches, 5,000 bp either side of transcription start/stop sites. In order to limit mapping of TE-derived siRNA, non-TE-derived genomic sequences were masked. The vertical dashed line indicates the start/stop site and the intervening genic region is excluded from the plot. The panel in the top left is a zoomed-in section of −1,250 to 1,250 bp showing a spike in 24-nt-siRNA abundance immediately up- and downstream of genes. (*B*) The same as in (*A*) except with the number of reads mapped to a given site capped at 10 in order to diminish the influence of a few highly targeted regions (i.e., outliers).

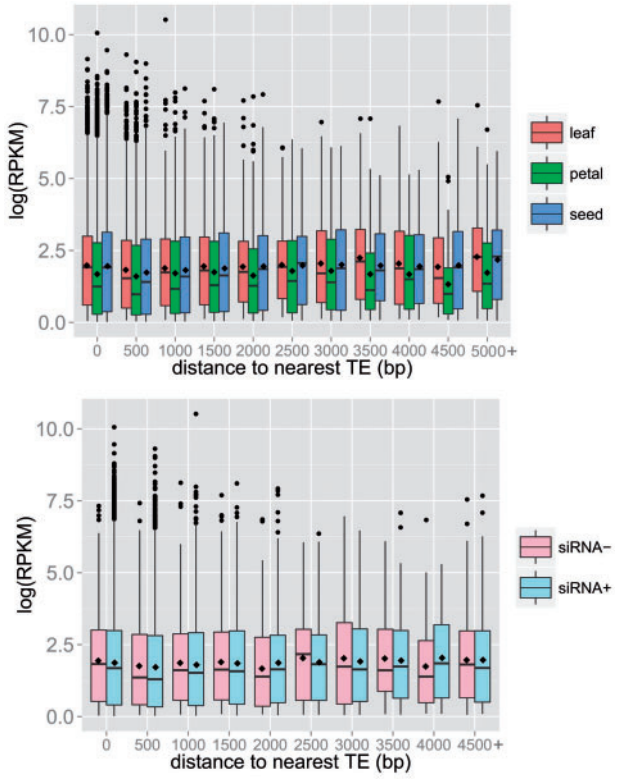

FIG. 4. TE proximity, siRNAs, and gene expression. (*A*) Boxplot of a transcriptomic analysis of 37,200 cotton genes examined for expression level and binned according the proximity to the nearest TE. Gene expression levels are given for each of three tissues. Boxes indicate 95% confidence intervals for median gene expression (solid black lines), mean expression is indicated by solid diamonds, and outliers are indicated with solid circles. (*B*) Expression of genes grouped into those with local siRNA producing TEs (siRNA+) and genes lacking local siRNA-inducing TEs (siRNA−). Linear modeling and ANOVA revealed that the presence or absence of siRNAs mapping to nearby TEs had no statistically significant effect on gene expression (supplementary table S1, Supplementary Material online).

(Salina et al. 2004; Woodhouse et al. 2010; Schnable et al. 2011; Sankoff and Zheng 2012; Tang et al. 2012; Garsmeur et al. 2013).

By reconstructing ancestral chromosomes of cotton, often in multiple copies, we demonstrate that this fractionation process has acted differentially on what initially were homoeologous chromosomes. Furthermore, this bias has occurred at the whole-chromosome level, and is evident in every chromosomal comparison we performed ($P < 0.0001$; table 1). Despite the antiquity of the genome multiplication in the cotton lineage, and the extensive genome rearrangement that has occurred between the genomes of modern cotton and cacao, we show here that fractionation in the cotton lineage remains evident after approximately 60 My of evolution (supplementary figs. S1 and S2, Supplementary Material online; table 1). This result extends our understanding of the scale, scope, and temporal depth of the diploidization processes and reveals the signatures of biased fractionation are evident far beyond the time-frame reported for other genomes; around 20 Ma in *Brassica* (Tang et al. 2012), 10 Ma post-WGM in maize (Schnable et al. 2011), and a few million years in *Arabidopsis* (Thomas et al. 2006). Considering that most paleopolyploid genomes appear chromosomally and functionally diploid it is noteworthy that such signatures, as well as differentiation of genomic content, remain evident after approximately 60 Ma. By extension, we deduce that the process of biased fractionation has a long-lasting legacy, raising a myriad questions regarding the functional and adaptive implications of biased fractionation over vast evolutionary timescales.

One of the primary correlates of biased fractionation in plants is differential expression of the duplicated genes that are retained on homoeologous segments. In maize and *Brassica*, biased fractionation is associated with overexpression of genes on the homoeologs experiencing less gene loss (Schnable et al. 2011; Tang et al. 2012; Woodhouse et al. 2014). We report similar results here, that is, that genes on LF chromosomes generally are more highly

expressed than their counterparts on MF chromosomes (fig. 1 and supplementary fig. S3, Supplementary Material online), demonstrating the same correlation between fractionation and gene expression in the cotton lineage as in other recent examples from plants. Remarkably, this quantitative difference in homoeolog expression levels remains detectable despite many tens of millions of years of evolution since WGM (fig. 1). It warrants mention here that although we can detect lower levels of expression in MF genes relative to LF, we have only a small sample of tissues from which to draw inference. Therefore, it remains possible that expression patterns (i.e., high levels of expression) in other tissue or time points may drive gene maintenance in both LF and MF fractions.

The correlation between biased fractionation and gene expression has led to a novel hypothesis regarding causation, that is, that selection has favored retention of the homoeolog with higher expression (Schnable et al. 2011). Under this hypothesis, conditions established at the time of genome merger (hybridization) and doubling (polyploidization), or those derived following autopolyploidization, generate widespread differences among homoeolog expression levels. These differences are commonly observed in cotton diploid hybrids and neoallopolyploids (Adams et al. 2003, 2004; Hovav et al. 2008; Yoo et al. 2013), showing that initial conditions may be sufficient to set in motion subsequent selective forces operating on gene expression. Moreover, in modern allopolyploid cotton, which traces to a hybridization event 1–2 Ma (Wendel and Cronn 2003), gene expression levels among linked genes on homoeologous segments may be correlated (Flagel et al. 2009). Collectively, these data indicate that homoeolog expression level differences may be established on temporal scales ranging from immediate, accompanying genome merger, to evolutionary timescales encompassing millions of years.

The question arises as to the proximate driver of expression level differences between MF and LF regions. An important insight in this respect stems from the work of Woodhouse et al. (2014), who noted that in *Brassica*, regions experiencing greater gene loss were enriched for mapping of 24-nt siRNAs to TEs near genes. This observation, in conjunction with the earlier demonstration (Hollister et al. 2011) that gene expression may be downregulated by positional effects associated with TEs, led to the suggestion that positional-effect downregulation drives differences in gene expression between duplicate regions, thus providing a potential explanation for biased fractionation.

Here we also show that 24-nt siRNAs are enriched in MF compared with LF regions (fig. 3), extending the scope of this phenomenon to other taxa. Thus, one might expect that genes residing near TEs are generally expressed at lower levels, particularly when close to TEs subject to siRNA silencing. Our data, however, show only a weak relationship between TE proximity and gene expression in cotton (fig. 4*A*). Furthermore, there is a slight negative correlation between gene expression and TE proximity, a trend that is opposite to expectations. Importantly, we observe a significant and better correlation between local TE density and gene expression (supplementary fig. S7, Supplementary Material online). This observation may reflect the contrasting genomic environments of heterochromatin (high TE prevalence and low gene expression) and euchromatin (low TE prevalence and high gene expression). Nevertheless, the presence of local TEs that are targeted by siRNA pathways has no apparent impact on local gene expression levels (fig. 4*B*). Yet, the correlation remains between homoeolog expression levels and fractionation bias (fig. 1 and table 1, respectively). Thus, siRNA-mediated, positional downregulation may not be the primary driver of biased gene loss, at least in cotton.

The foregoing observations raise the question as to whether homoeolog expression level differences are a consequence, rather than a driver, of biased fractionation. Others have previously predicted that differential TE load between subgenomes might be a determinant of LF and MF; that is, the subgenome with the lowest TE density would be less fractionated and overexpressed (Woodhouse et al. 2014). Our data generally support this notion, in that several genomic characteristics differentiate LF and MF genome fractions: Local TE density is higher in MF than LF chromosomes (fig. 2), whereas GC content is lower (supplementary fig. S5, Supplementary Material online).

Although our data do not support the notion that siRNA TE-mediated downregulation of genes is currently a driver of biased gene loss, it remains a possibility that such processes may be of crucial importance during the first flush following ancient WGM. Several processes, however, may obscure earlier patterns of siRNA-mediated silencing, including: 1) Inactivation of TEs such that siRNA-mediated silencing is no longer required (and thus is absent in the modern cotton genome); 2) the inability to identify divergent TEs, thus affecting our ability to associate TE presence with downregulation of genes; and 3) apparent TE content among homoeologous regions may be due to invasion of TE sequences postpaleopolyploidy, and are not reflective of conditions at the time of WGM. The latter process has almost certainly occurred, as previous work has implicated bursts of TE activity 0.5–2 Ma (Hu et al. 2010; Paterson et al. 2012), tens of millions of years after WGM. Such recent bursts of transposition may obscure signals of more ancient processes, such that it might be difficult to identify ancient patterns of TE-mediated downregulation of gene expression based on contemporary analysis of TEs and siRNAs.

With these caveats in mind, our observations, coupled with data showing that TE proximity appears to be decoupled with expression levels (fig. 4), we suggest the possibility that some other genomic feature associated with increased TE density content between subgenomes might drive differentiation in rates of gene loss following WGM.

In the present context, irrespective of the deletional mechanism, we forward the hypothesis that elevated TE density may have triggered a greater fixation of gene losses in MF chromosomes through indirect effects on recombination rate. In the case of allopolyploidy the subgenomes of a polyploid derive from independent species which at the time of unification may have quite distinct TE content (e.g., modern allopolyploids of cotton [Wendel et al. 2009] and *Nicotiana*

[Leitch et al. 2008]). Indeed, Garsmeur et al. (2013) suggested that biases in gene loss could be instigated by allopolyploidy, but that ancient WGMs where biased gene loss is absent are more likely to be derived from autopolyploidy. Furthermore, recombination rate has been shown to be negatively correlated with TE prevalence (Rizzon et al. 2002; Fontanillas et al. 2007) and GC content can be positively correlated with recombination rate (Birdsell 2002). Therefore, higher TE density and lower GC content in MF hint at a history of reduced recombination relative to LF, perhaps due to variable TE load and/or GC content at the time of unification. If true, this has important implications; selection is weaker in regions of low recombination (Hill and Robertson 1966), leading to the possibility that mildly deleterious gene deletions in LF fragments, where selection can operate strongly, are less likely to be fixed compared with similar deletions in MF fractions, where selection is weaker. Moreover, at the time of WGM, effective population size is likely to be very small, in which case, in the absence of effective selection on MF fragments, deletions could be fixed through drift. Although speculative, it seems to us possible that TE load (or perhaps some other determinant of recombination rate) may be a proximate evolutionary force responsible for the genesis of LF and MF genomic compartments.

Materials and Methods

Biased Fractionation

We examined gene fractionation following the cotton-specific WGM using the reference genome sequences for *G. raimondii* (Paterson et al. 2012) and its close relative, *T. cacao* (Argout et al. 2011). The SynMap function of the online tool CoGe (https://genomevolution.org/CoGe, last accessed July 7, 2014) was used to identify blocks of syntenic orthologs using BLASTN, relative gene order, and the following parameters: -D 50, -A 10, the Quota Align function to merge syntenic blocks, -Dm 80, and a ratio of syntenic depth of 6:1 (*G. raimondii*:*T. cacao*).

We further reconstructed ancestral chromosomes (fig. 1), according to the logic of Schnable et al. (2011). Briefly, rearrangements on the same chromosome are presumed to be more frequent than exchanges between different chromosomes. Thus, segments of the cotton genome that reside on the same chromosome and are orthologous to the same chromosome of the cacao genome are assumed to originate from an ancestral chromosome in the common ancestor of the two species. Furthermore, under the assumption that gene loss and chromosomal rearrangements are more likely after than before a WGM event (Kasahara et al. 2007), we took gene content and gene order in the outgroup *T. cacao* to be representative of the ancestral model of the preduplicated cotton genome, as has been done previously (Schnable et al. 2011). We selected those reconstructions for which we had the greatest confidence of a full-length reconstruction, provided there was a homoeologous reconstruction of comparable quality with which to compare. We subsequently limited our analyses of fractionation to these selected regions (highlighted green in supplementary fig. S1, Supplementary Material online). Assuming that the number of genes on duplicated chromosomes is initially equal following duplication, we assessed differences in the numbers of remaining genes. Importantly, we only included genes with appropriate syntenic orthologs in the chocolate genome, and thus our analysis does not include genes that have inserted following the cotton WGM. We subsequently binned chromosome reconstructions such that, for each ancestral chromosome, the least fractionated homoeolog was designated as LF and all other reconstructions were designated as MF and used a Pearson's chi-square test to investigate biased fractionation between reconstructed cotton chromosomes as in Tang et al. (2012).

Gene Expression in LF and MF

We determined the overlap in gene content between reconstructions and compared the relative expression of syntenic paralogs on LF and MF fractions. In detail, we restricted our expression analysis to those genes with paralogs on both MF and LF fractions. We further restricted those genes under consideration by using only syntenic paralog pairs that had a corresponding syntenic ortholog in the unduplicated (relative to cotton) chocolate genome. We used gene expression data from petal, leaf, and seed, previously published in Renny-Byfield et al. (2014). Briefly, RNA-seq reads were filtered for quality using the program sickle (https://github.com/najoshi/sickle, last accessed July 7, 2014), and mapped to the cotton reference genome (Paterson et al. 2012) using GSNAP (http://research-pub.gene.com/gmap/, last accessed July 7, 2014). Mapped reads were sorted and indexed using SAMtools (http://samtools.sourceforge.net, last accessed July 7, 2014) and coverage of each gene annotation was calculated using custom perl scripts and normalized using reads per kilobase per million. We compared paralogs on LF and MF reconstructions using methods similar to Schnable et al. (2011) and Woodhouse et al. (2014), where for each comparable paralog pair the gene with the highest expression was declared the winner. In cases where there were three reconstructions (table 1) we compared expression of LF genes with their syntenic paralogs on either, or both of the MF homoeologs, depending on whether a syntenic paralog was present on each MF reconstruction. We tested for differential gene expression of syntenic paralogs using a student's *t*-test, correcting *P* values for a false discovery rate of 0.01 using the method of Benjamini and Hochberg (1995). Using a cumulative binomial distribution function we assessed the chances of observing the number of wins in LF, given a random chance of LF being more highly expressed for each gene (i.e., the probability of LF being more highly expressed for a given comparison is 0.5).

Mapping of siRNAs

Small RNA libraries, previously published in Gong et al. (2013) and produced from seedling leaves of *G. raimondii*, were analyzed (deposited in NCBI SRA database SRP017133). From this data set we extracted 24-nt siRNAs and mapped these to the cotton reference genome using bowtie 0.12.7 (Langmead et al. 2009). In order to determine the precise

TE from which a given siRNA derived, we restricted mapping to those reads to those with perfect and unique alignments. All 24-nt siRNAs were mapped to TEs with all non-TE-derived nucleotides masked. For each gene, sliding windows of 100 bp in size (moving by increments of 10) were used to characterize the siRNA distribution 5,000 bp up- and downstream of the transcription start/stop sites. Coverage of siRNAs was characterized by counting the number of mapped reads inside each window.

Local TE Density and GC Content

For each gene, we assessed the local GC content 5,000 bp either side of the transcription start/stop site in sliding windows of 100 bp, moving by increments of 10, using BEDtools (http://bedtools.readthedocs.org/en/latest/, last accessed July 7, 2014). We then grouped genes according to their status with respect to MF and LF and separately plotted average GC content over each window. We examined whether mean GC content in LF and MF groups was different using a Wilcoxon's sum rank test. Similarly, for each protein gene, the TE density (the proportion of TE-derived base pair) within each sliding window was estimated using BEDtools. We examined differences in TE density between LF and MF groups using Wilcoxon's sum rank test.

TE Density, Proximity, and Gene Expression

Repetitive DNA annotations from Paterson et al. (2012) were used to identify TEs. Using the "closest" functionality of BEDtools we ascertained the nearest TE to each gene, allowing for TEs inside gene models. Using the same expression data as above, each of the global set of approximately 37,200 gene annotations was assessed for expression level using three biological replicates. Genes were subsequently binned according to distance to nearest TE and expression was compared between bins using boxplots. The statistical relationship between TE proximity and the log of expression was examined using Pearson's correlation coefficient and linear modeling. All statistical analyses were performed in the statistical package R.

Supplementary Material

Supplementary file S1, table S1, and figures S1–S7 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank the National Science Foundation Plant Genome Program for funding. J.P.G. is funded by a Graduate Research Fellowship from the National Science Foundation.

References

Adams KL, Cronn R, Percifield R, Wendel JF. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc Natl Acad Sci U S A.* 100: 4649–4654.

Adams KL, Percifield R, Wendel JF. 2004. Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics* 168:2217–2226.

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, et al. 2011. The genome of *Theobroma cacao*. *Nat Genet.* 43:101–108.

Bardil A, de Almeida JD, Combes MC, Lashermes P, Bertrand B. 2011. Genomic expression dominance in the natural allopolyploid *Coffea arabica* is massively affected by growth temperature. *New Phytol.* 192:760–774.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Stat Soc Series B Methodol.* 57:289–300.

Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol.* 19:1181–1197.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16:1667–1678.

Bowers JE, Chapman BA, Rong JK, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422:433–438.

Buggs RJ, Elliott NM, Zhang LJ, Koh J, Viccini LF, Soltis DE, Soltis PS. 2010. Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid *Tragopogon mirus*. *New Phytol.* 186:175–183.

Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, Bonnema G, Wang X. 2012. Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PLoS One* 7:e36442.

Clarkson JJ, Lim KY, Kovarik A, Chase MW, Knapp S, Leitch AR. 2005. Long-term genome diploidization in allopolyploid *Nicotiana* section *Repandae* (Solanaceae). *New Phytol.* 168:241–252.

Flagel L, Udall J, Nettleton D, Wendel J. 2008. Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. *BMC Biol.* 6:16.

Flagel LE, Chen L, Chaudhary B, Wendel JF. 2009. Coordinated and fine-scale control of homoeologous gene expression in allotetraploid cotton. *J Hered.* 100:487–490.

Fontanillas P, Hartl DL, Reuter M. 2007. Genome organization and gene expression shape the transposable element distribution in the *Drosophila melanogaster* euchromatin. *PLoS Genet.* 3: 2256–2267.

Freeling M, Woodhouse MR, Subramaniam S, Turco G, Lisch D, Schnable JC. 2012. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Curr Opin Plant Biol.* 15:131–139.

Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. 2013. Two evolutionarily distinct classes of paleopolyploidy. *Mol Biol Evol.* 31:448–454.

Gong L, Kakrana A, Arikit S, Meyers BC, Wendel JF. 2013. Composition and expression of conserved microRNA genes in diploid cotton (*Gossypium*) species. *Genome Biol Evol.* 5:2449–2459.

Grover CE, Gallagher JP, Szadkowski EP, Yoo MJ, Flagel LE, Wendel JF. 2012. Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytol.* 196:966–971.

Hill WG, Robertson A. 1966. Effect of linkage on limits of artificial selection. *Genet Res.* 8:269–294.

Hollister JD, Smith LM, Guo Y-L, Ott F, Weigel D, Gaut BS. 2011. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proc Natl Acad Sci U S A.* 108:2322–2327.

Hovav R, Udall JA, Chaudhary B, Rapp R, Flagel L, Wendell JF. 2008. Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. *Proc Natl Acad Sci U S A.* 105:6191–6195.

Hu GJ, Hawkins JS, Grover CE, Wendel JF. 2010. The history and disposition of transposable elements in polyploid *Gossypium*. *Genome* 53: 599–607.

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* 13:R3.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473:97–100.

Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447:714–719.

Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M. 2004. Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* 166:935–945.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25.

Leitch IJ, Hanson L, Lim KY, Kovarik A, Chase MW, Clarkson JJ, Leitch AR. 2008. The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). *Ann Bot.* 101:805–814.

Mandakova T, Joly S, Krzywinski M, Mummenhoff K, Lysak MA. 2010. Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. *Plant Cell* 22:2277–2290.

Paterson AH, Bowers JE, Burow MD, Draye X, Elsik CG, Jiang CX, Katsar CS, Lan TH, Lin YR, Ming RG, et al. 2000. Comparative genomics of plant chromosomes. *Plant Cell* 12:1523–1539.

Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A.* 101:9903–9908.

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.

Renny-Byfield S, Gallagher JP, Grover CE, Szadkowski E, Page JT, Udall JA, Wang X, Paterson AH, Wendel JF. 2014. Ancient gene duplicates in *Gossypium* (Cotton) exhibit near-complete expression divergence. *Genome Biol Evol.* 6:559–571.

Renny-Byfield S, Kovarik A, Kelly LJ, Macas J, Novak P, Chase MW, Nichols RA, Pancholi MR, Grandbastien M-A, Leitch AR. 2013. Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. *Plant J.* 74:829–839.

Rizzon C, Marais G, Gouy M, Biemont C. 2002. Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. *Genome Res.* 12:400–407.

Salina EA, Numerova OM, Ozkan H, Feldman M. 2004. Alterations in subtelomeric tandem repeats during early stages of allopolyploidy in wheat. *Genome* 47:860–867.

Sankoff D, Zheng C. 2012. Fractionation, rearrangement and subgenome dominance. *Bioinformatics* 28:i402–i408.

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci U S A.* 108:4069–4074.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* 326: 1112–1115.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng CF, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *Am J Bot.* 96:336–348.

Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, Wang X, Freeling M, Pires JC. 2012. Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. *Genetics* 190:1563–1574.

Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* 16:934–946.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet.* 43:1035–1039.

Wendel JF, Brubaker C, Alvarez I, Cronn R, Stewart JM. 2009. Evolution and natural history of the cotton genus. In: Paterson AH, editor. Genetics and genomics of cotton plant genetics and genomics: crops and models. New York: Springer Science. p. 3–22.

Wendel JF, Cronn RC. 2003. Polyploidy and the evolutionary history of cotton. In: Sparks DL, editor. Advances in agronomy. Vol. 78. Waltham, Massachusetts: Academic Press. p. 139–186.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet.* 2:333–341.

Woodhouse MR, Cheng F, Pires JC, Lisch D, Freeling M, Wang X. 2014. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids (vol 111, pg 5283, 2014). *Proc Natl Acad Sci U S A.* 111:5283–5288.

Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M. 2010. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biol.* 8:e1000409.

Yoo MJ, Szadkowski E, Wendel JF. 2013. Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* 110: 171–180.