Persistence of sub-genomes in paleopolyploid cotton after 60 million years of evolution

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

[1]Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, IA 50014, USA.

[§]current address: Department of Plant Sciences, University of California, Davis. One Shields Avenue, Davis, CA, 95616, USA

[*]Corresponding author

email: jfw@iastate.edu

keywords: whole genome duplication, gene fractionation, biased fractionation, transposable element, gene expression

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 over-expression of genes in less fractionated (LF) relative to more fractionated (MF) regions. This bias is hypothesized to arise through down-regulation of gene expression via 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 down-regulation is absent in the modern cotton genome, seemingly excluding this phenomenon as the primary driver of biased gene loss. Nevertheless,the paleo sub-genomes of diploid cotton are still distinguishable with respect to TE content, targeting of 24nt-siRNAs and GC content, despite ~60 million years 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.

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; Bowers et al. 2003; Blanc and Wolfe 2004; Paterson et al. 2004; Leitch et al. 2008; Soltis et al. 2009; Jiao et al. 2011; Jiao et al. 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). This transition involves the large-scale loss of duplicate genes (Langham et al. 2004; Woodhouse et al. 2010; Schnable et al. 2011; Tang et al. 2012; Garsmeur et al. 2013; Woodhouse et al. 2014). 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).

Recent analyses in maize (Schnable et al. 2011) and *Brassica* (Cheng et al. 2012) 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 sub-genomes, e.g., 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 hypothesis is attractive 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 the mechanistic underpinnings of biased fractionation. This process is suggested to

be accompanied by selection, where under-expressed genes contribute less to fitness than do their over-expressed homoeologous counterparts, and therefore are more likely to become dispensable (Freeling et al. 2012).

Additional support for this hypothesis was generated by recent work involving *Brassica*; homoeologous regions experiencing greater gene loss were enriched for mapping of 24 nt siRNAs to transposable elements (TEs) surrounding resident genes (Woodhouse et al. 2014). As shown in *Arabidopsis*, gene expression may be down-regulated by positional effects associated with local TEs, which are silenced by the 24-nt siRNA machinery (Hollister et al. 2011). These observations of TE proximity, siRNA frequency, and relative gene expression of homoeologs in *Brassica* led to the suggestion (Woodhouse et al. 2014) that positional-effect, down-regulation by local TEs drives differences in gene expression between duplicate regions, thus providing a potential explanation for biased fractionation.

Here we extend the temporal window for studying biased fractionation and demonstrate that the genomic footprints of the process persist following ancient (~ 60 mya) WGM (Paterson et al. 2012) in the modern genome of diploid cotton (*Gossypium raimondii* L.). 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 down-regulation 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 and least fractionated 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 mya in maize (Schnable et al. 2009; Schnable et al.

2011)) and 20 mya 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 7/7/14), we identified blocks of genes in synteny between *Gossypium raimondii* (diploid cotton) and its relative *Theobroma cacao* (cacao; SynMap output in File S1). Importantly, comparative genome sequence data indicate that relative to cacao, the lineage that gave rise to modern diploid cotton experienced a 5-6 fold ploidy increase approximately 60 mya (Paterson et al. 2012). Accordingly, we identified duplicate regions in the cotton genome resulting from the cotton-specific WGM (Fig. S1 and S2). The relative antiquity of WGM in cotton, however, makes identification of duplicate regions challenging compared to efforts in paleopolyploids where WGM, although still ancient, has occurred much more recently. Indeed, the genome of cotton is substantially re-arranged relative to cacao (Fig. S1 and S2), and in some regions five or six duplicate segments per haploid cacao genome are evident in cotton. It was not possible to re-construct 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. (Schnable et al. 2011); we restricted further analyses to those genes and genomic regions covered by the whole chromosome reconstructions (highlighted green in Fig. S1).

In total, we were able to reconstruct at least two ancestral cotton chromosomes for six of the ten pre-duplicated 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, Fig. S1). Assuming that each duplicate chromosome has the same number of genes at the time of duplication, we

estimated gene loss by assessing the number of genes remaining in each reconstruction, in a manner similar to that used earlier (Tang et al. 2012). The hypothesis of gene retention equivalence among homoeologs was evaluated using Pearson's chi-squared 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 signatures of chromosome-wide biased fractionation following WGM on time-scales of ~60 mya, 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 pre-duplicated ancestral chromosome, the least fractionated of the post-duplicated (cotton) reconstructions was distinguished LF and all others were grouped together and designated as most fractionated (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 those on MF (Fig. 1); of the 1064 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 up-regulated 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. 1A). To examine the robustness of the bias we restricted our analysis to those syntenic paralogs that varied in expression by

greater than two-fold (Fig. S3 A). 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 two, confirming previous results (Fig. S3 B). Thus, the data indicate that genes on LF chromosomes are typically expressed at higher levels than their MF counterparts.

Local transposable element density

Using transposable element (TE) annotations from the cotton reference genome (Paterson et al. 2012), we identified TEs near each gene and 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=502503, p=<0.0001; Fig. 2). In addition, genes residing on MF chromosomes generally have more internal TE insertions (Fig. S4) 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 (Fig. S5) 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 = <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 =327477, p = <0.0001;Fig. 3). In addition to enrichment of siRNAs, the MF fraction of the genome also exhibits a sharp increase in siRNAs mapping ~200 bp upstream and ~800 bp downstream of the transcription start/stop site, a pattern absent in the LF fraction of the genome (Fig. 3).

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 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). Indeed, correlation analysis revealed a statistically significant negative correlation between expression and TE proximity (Pearson's correlation coefficient = -0.03, df = 111667, t = -10.166, p = <0.001). Although statistically significant, the effect is minimal and in the opposite direction of expectation, indicating that TE proximity is likely to have little impact on levels of gene expression. Furthermore, linear modeling revealed an R-squared value of <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 (Fig. S6). 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 = 116637, t = -42.003, p = < 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 ANOVA indicate that presence of siRNAs has little impact on the expression of nearby genes (Table S1).

Discussion

The extensive occurrence of episodically recurrent WGM in the history of flowering plants has only recently become evident (Paterson et al. 2000; Bowers et al. 2003; Blanc and Wolfe 2004; Paterson et al. 2004; Leitch et al. 2008; Soltis et al. 2009; Jiao et al. 2011; Jiao et al. 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 5-6 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 via 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 (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 ~ 60 million years of evolution (Fig. S1 and Fig. S2; 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 mya in *Brassica* (Tang et al. 2012), 10 mya 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 ~60 mya. 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 over-expression 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, i.e., that genes on LF chromosomes generally are more highly expressed than their counterparts on MF chromosomes (Fig. 1 and Fig. S3), 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).

The correlation between biased fractionation and gene expression has led to a novel hypothesis regarding causation, i.e., 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 neo-allopolyploids (Adams et al. 2003; Adams et al. 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 mya (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 Freeling and colleagues (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 down-regulated by positional effects associated with TEs, led to the suggestion that positional-effect down-regulation 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 to 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. Indeed, 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 down-regulation 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 sub-genomes might be a determinant of LF and MF; that is, the sub-genome with the lowest TE density would be less fractionated and over expressed (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 (Fig. S5). Given that ancient TEs are difficult to identify, we cannot exclude the possibility that differences in apparent TE content among homoeologous regions may be due to invasion of TE sequences post paleopolyploidy and are not reflective of conditions at the time of WGM. With this caveat in mind, our observations, coupled with data showing that TE proximity appears to be de-coupled with expression levels (Fig. 4), suggest that some other genomic feature associated with increased TE density content between sub-genomes 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 via indirect effects on recombination rate. Such a hypothesis may be testable given appropriate genetic maps, but currently suitable data are not available. Nevertheless, recombination rate has been shown to be negatively correlated with TE prevalence (Rizzon et al. 2002; Fontanillas et al. 2007). Similarly, GC content can be positively correlated with recombination rate (Birdsell 2002). Therefore, higher TE density and lower GC content in MF suggest a history of reduced recombination relative to LF. This is important as the impact of selection is weaker in regions of low recombination

(Hill and Robertso.A 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 to 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 via 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 *Gossypium raimondii* (Paterson et al. 2012) and its close relative, *Theobroma cacao* L. (Argout et al. 2011). The SynMap function of the online tool CoGe (https://genomevolution.org/CoGe) 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 (*Gossypium raimondii*: *Theobroma cacao*).

We further reconstructed ancestral chromosomes (Fig. 1), according to the logic of Schnable et al. (Schnable et al. 2011). Briefly, re-arrangements 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 re-arrangements are more likely after than before a WGM event (Kasahara et al. 2007), we took gene content and gene order in the

outgroup *Theobroma cacao* to be representative of the ancestral model of the pre-duplicated 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. 1). Assuming that the number of genes on duplicated chromosomes is initially equal following duplication, we assessed differences in the numbers of remaining genes and used a Pearson's chi-squared test to investigate biased fractionation between re-constructed cotton chromosomes as in (Tang et al. 2012). 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 most fractionated (MF).

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 7/7/14), and mapped to the cotton reference genome (Paterson et al. 2012) using GSNAP (http://research-pub.gene.com/gmap/; last accessed 7/7/14). Mapped reads were sorted and indexed using SAMtools (http://samtools.sourceforge.net; last accessed 7/7/14) and coverage of each gene annotation was calculated using custom perl scripts and normalized using reads per kilobase per million (RPKM). We compared paralogs on LF and MF reconstructions using methods similar to

Schnable et al. (Schnable et al. 2011) and Woodhouse et al. (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 *Gossypium raimondii* were analysed (deposited in NCBI SRA database SRP017133). From this dataset 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 ten) where used to characterize the siRNA distribution 5000 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 5000 bp either side of the transcription start/stop site in sliding windows of 100 bp, moving by increments of ten, using *BEDtools* (http://bedtools.readthedocs.org/en/latest/; last accessed 7/7/14). 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 if 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 bp) 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 ~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.

Acknowledgements

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

References

Adams, K. L., R. Cronn, R. Percifield, and J. F. Wendel. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific

reciprocal silencing. Proceedings of the National Academy of Sciences of the United States of America **100**:4649-4654.
Adams, K. L., R. Percifield, and J. F. Wendel. 2004. Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. Genetics **168**:2217-2226.
Argout, X., J. Salse, J.-M. Aury, M. J. Guiltinan, G. Droc, J. Gouzy, M. Allegre, C. Chaparro, T. Legavre, S. N. Maximova, M. Abrouk, F. Murat, O. Fouet, J. Poulain, M. Ruiz, Y. Roguet, M. Rodier-Goud, J. F. Barbosa-Neto, F. Sabot, D. Kudrna, J. S. S. Ammiraju, S. C. Schuster, J. E. Carlson, E. Sallet, T. Schiex, A. Dievart, M. Kramer, L. Gelley, Z. Shi, A. Berard, C. Viot, M. Boccara, A. M. Risterucci, V. Guignon, X. Sabau, M. J. Axtell, Z. Ma, Y. Zhang, S. Brown, M. Bourge, W. Golser, X. Song, D. Clement, R. Rivallan, M. Tahi, J. M. Akaza, B. Pitollat, K. Gramacho, A. D'Hont, D. Brunel, D. Infante, I. Kebe, P. Costet, R. Wing, W. R. McCombie, E. Guiderdoni, F. Quetier, O. Panaud, P. Wincker, S. Bocs, and C. Lanaud. 2011. The genome of *Theobroma cacao*. Nature Genetics **43**:101-108.
Bardil, A., J. D. de Almeida, M. C. Combes, P. Lashermes, and B. Bertrand. 2011. Genomic expression dominance in the natural allopolyploid *Coffea arabica* is massively affected by growth temperature. New Phytologist **192**:760-774.
Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Methodological **57**:289-300.
Birdsell, J. A. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Molecular Biology and Evolution **19**:1181-1197.
Blanc, G., and K. H. Wolfe. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell **16**:1667-1678.
Bowers, J. E., B. A. Chapman, J. K. Rong, and A. H. Paterson. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422**:433-438.
Buggs, R. J., N. M. Elliott, L. J. Zhang, J. Koh, L. F. Viccini, D. E. Soltis, and P. S. Soltis. 2010. Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid *Tragopogon mirus*. New Phytologist **186**:175-183.
Cheng, F., J. Wu, L. Fang, S. Sun, B. Liu, K. Lin, G. Bonnema, and X. Wang. 2012. Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. Plos One **7**.
Clarkson, J. J., K. Y. Lim, A. Kovarik, M. W. Chase, S. Knapp, and A. R. Leitch. 2005. Long-term genome diploidization in allopolyploid *Nicotiana* section *Repandae* (Solanaceae). New Phytologist **168**:241-252.
Flagel, L., J. Udall, D. Nettleton, and J. Wendel. 2008. Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. Bmc Biology **6**:16.
Flagel, L. E., L. Chen, B. Chaudhary, and J. F. Wendel. 2009. Coordinated and fine-scale control of homoeologous gene expression in allotetraploid cotton. Journal of Heredity **100**:487-490.

Fontanillas, P., D. L. Hartl, and M. Reuter. 2007. Genome organization and gene expression shape the transposable element distribution in the *Drosophila melanogaster* euchromatin. Plos Genetics **3**:2256-2267.
Freeling, M., M. R. Woodhouse, S. Subramaniam, G. Turco, D. Lisch, and J. C. Schnable. 2012. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Current Opinion in Plant Biology **15**:131-139.
Garsmeur, O., J. C. Schnable, A. Almeida, C. Jourda, A. D'Hont, and M. Freeling. 2013. Two evolutionarily distinct classes of paleopolyploidy. Molecular Biology and Evolution **31**:448-454.
Gong, L., A. Kakrana, S. Arikit, B. C. Meyers, and J. F. Wendel. 2013. Composition and expression of conserved microRNA genes in diploid cotton (*Gossypium*) species. Genome Biology and Evolution **5**:2449-2459.
Grover, C. E., J. P. Gallagher, E. P. Szadkowski, M. J. Yoo, L. E. Flagel, and J. F. Wendel. 2012. Homoeolog expression bias and expression level dominance in allopolyploids. New Phytologist **196**:966-971.
Hill, W. G., and Robertso.A. 1966. Effect of linkage on limits of artificial selection. Genetical Research **8**:269-&.
Hollister, J. D., L. M. Smith, Y.-L. Guo, F. Ott, D. Weigel, and B. S. Gaut. 2011. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proceedings of the National Academy of Sciences of the United States of America **108**:2322-2327.
Hovav, R., J. A. Udall, B. Chaudhary, R. Rapp, L. Flagel, and J. F. Wendell. 2008. Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. Proceedings of the National Academy of Sciences of the United States of America **105**:6191-6195.
Jiao, Y., J. Leebens-Mack, S. Ayyampalayam, J. E. Bowers, M. R. McKain, J. McNeal, M. Rolf, D. R. Ruzicka, E. Wafula, N. J. Wickett, X. Wu, Y. Zhang, J. Wang, Y. Zhang, E. J. Carpenter, M. K. Deyholos, T. M. Kutchan, A. S. Chanderbali, P. S. Soltis, D. W. Stevenson, R. McCombie, J. C. Pires, G. K.-S. Wong, D. E. Soltis, and C. W. dePamphilis. 2012. A genome triplication associated with early diversification of the core eudicots. Genome Biology **13**.
Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, Y. Hu, H. Liang, P. S. Soltis, D. E. Soltis, S. W. Clifton, S. E. Schlarbaum, S. C. Schuster, H. Ma, J. Leebens-Mack, and C. W. dePamphilis. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature **473**:97-100.
Kasahara, M., K. Naruse, S. Sasaki, Y. Nakatani, W. Qu, B. Ahsan, T. Yamada, Y. Nagayasu, K. Doi, Y. Kasai, T. Jindo, D. Kobayashi, A. Shimada, A. Toyoda, Y. Kuroki, A. Fujiyama, T. Sasaki, A. Shimizu, S. Asakawa, N. Shimizu, S.-i. Hashimoto, J. Yang, Y. Lee, K. Matsushima, S. Sugano, M. Sakaizumi, T. Narita, K. Ohishi, S. Haga, F. Ohta, H. Nomoto, K. Nogata, T. Morishita, T. Endo, T. I. Shin, H. Takeda, S. Morishita, and Y. Kohara. 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature **447**:714-719.

Langham, R. J., J. Walsh, M. Dunn, C. Ko, S. A. Goff, and M. Freeling. 2004. Genomic duplication, fractionation and the origin of regulatory novelty. Genetics **166**:935-945.
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 Biology **10**.
Leitch, I. J., L. Hanson, K. Y. Lim, A. Kovarik, M. W. Chase, J. J. Clarkson, and A. R. Leitch. 2008. The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). Annals of Botany **101**:805-814.
Mandakova, T., S. Joly, M. Krzywinski, K. Mummenhoff, and M. A. Lysak. 2010. Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. Plant Cell **22**:2277-2290.
Paterson, A. H., J. E. Bowers, M. D. Burow, X. Draye, C. G. Elsik, C. X. Jiang, C. S. Katsar, T. H. Lan, Y. R. Lin, R. G. Ming, and R. J. Wright. 2000. Comparative genomics of plant chromosomes. Plant Cell **12**:1523-1539.
Paterson, A. H., J. E. Bowers, and B. A. Chapman. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proceedings of the National Academy of Sciences of the United States of America **101**:9903-9908.
Paterson, A. H., J. F. Wendel, H. Gundlach, H. Guo, J. Jenkins, D. Jin, D. Llewellyn, K. C. Showmaker, S. Shu, J. Udall, M.-j. Yoo, R. Byers, W. Chen, A. Doron-Faigenboim, M. V. Duke, L. Gong, J. Grimwood, C. Grover, K. Grupp, G. Hu, T.-h. Lee, J. Li, L. Lin, T. Liu, B. S. Marler, J. T. Page, A. W. Roberts, E. Romanel, W. S. Sanders, E. Szadkowski, X. Tan, H. Tang, C. Xu, J. Wang, Z. Wang, D. Zhang, L. Zhang, H. Ashrafi, F. Bedon, J. E. Bowers, C. L. Brubaker, P. W. Chee, S. Das, A. R. Gingle, C. H. Haigler, D. Harker, L. V. Hoffmann, R. Hovav, D. C. Jones, C. Lemke, S. Mansoor, M. U. Rahman, L. N. Rainville, A. Rambani, U. K. Reddy, J.-k. Rong, Y. Saranga, B. E. Scheffler, J. A. Scheffler, D. M. Stelly, B. A. Triplett, A. Van Deynze, M. F. S. Vaslin, V. N. Waghmare, S. A. Walford, R. J. Wright, E. A. Zaki, T. Zhang, E. S. Dennis, K. F. X. Mayer, D. G. Peterson, D. S. Rokhsar, X. Wang, and J. Schmutz. 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature **492**:423–427.
Renny-Byfield, S., J. P. Gallagher, C. E. Grover, E. Szadkowski, J. T. Page, J. A. Udall, X. Wang, A. H. Paterson, and J. F. Wendel. 2014. Ancient gene duplicates in *Gossypium* (Cotton) exhibit near-complete expression divergence. Genome Biology and Evolution **6**:559-571.
Renny-Byfield, S., A. Kovarik, L. J. Kelly, J. Macas, P. Novak, M. W. Chase, R. A. Nichols, M. R. Pancholi, M.-A. Grandbastien, and A. R. Leitch. 2013. Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. The Plant Journal **74**:829-839.
Rizzon, C., G. Marais, M. Gouy, and C. Biemont. 2002. Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. Genome Research **12**:400-407.

Salina, E. A., O. M. Numerova, H. Ozkan, and M. Feldman. 2004. Alterations in subtelomeric tandem repeats during early stages of allopolyploidy in wheat. Genome **47**:860-867.
Sankoff, D., and C. Zheng. 2012. Fractionation, rearrangement and subgenome dominance. Bioinformatics **28**:i402-i408.
Schnable, J. C., N. M. Springer, and M. Freeling. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proceedings of the National Academy of Sciences of the United States of America **108**:4069-4074.
Schnable, P. S.D. WareR. S. FultonJ. C. SteinF. S. WeiS. PasternakC. Z. LiangJ. W. ZhangL. FultonT. A. GravesP. MinxA. D. ReilyL. CourtneyS. S. KruchowskiC. TomlinsonC. StrongK. DelehauntyC. FronickB. CourtneyS. M. RockE. BelterF. Y. DuK. KimR. M. AbbottM. CottonA. LevyP. MarchettoK. OchoaS. M. JacksonB. GillamW. Z. ChenL. YanJ. HigginbothamM. CardenasJ. WaligorskiE. ApplebaumL. PhelpsJ. FalconeK. KanchiT. ThaneA. ScimoneN. ThaneJ. HenkeT. WangJ. RuppertN. ShahK. RotterJ. HodgesE. IngenthronM. CordesS. KohlbergJ. SgroB. DelgadoK. MeadA. ChinwallaS. LeonardK. CrouseK. ColluraD. KudrnaJ. CurrieR. F. HeA. AngelovaS. RajasekarT. MuellerR. LomeliG. ScaraA. KoK. DelaneyM. WissotskiG. LopezD. CamposM. BraidottiE. AshleyW. GolserH. KimS. LeeJ. K. LinZ. DujmicW. KimJ. TalagA. ZuccoloC. FanA. SebastianM. KramerL. SpiegelL. NascimentoT. ZutavernB. MillerC. AmbroiseS. MullerW. SpoonerA. NarechaniaL. Y. RenS. WeiS. KumariB. FagaM. J. LevyL. McMahanP. Van BurenM. W. VaughnK. YingC. T. YehS. J. EmrichY. JiaA. KalyanaramanA. P. HsiaW. B. BarbazukR. S. BaucomT. P. BrutnellN. C. CarpitaC. ChaparroJ. M. ChiaJ. M. DeragonJ. C. EstillY. FuJ. A. JeddelohY. J. HanH. LeeP. H. LiD. R. LischS. Z. LiuZ. J. LiuD. H. NagelM. C. McCannP. SanMiguelA. M. MyersD. NettletonJ. NguyenB. W. PenningL. PonnalaK. L. SchneiderD. C. SchwartzA. SharmaC. SoderlundN. M. SpringerQ. SunH. WangM. WatermanR. WestermanT. K. WolfgruberL. X. YangY. YuL. F. ZhangS. G. ZhouQ. ZhuJ. L. BennetzenR. K. DaweJ. M. JiangN. JiangG. G. PrestingS. R. WesslerS. AluruR. A. MartienssenS. W. CliftonW. R. McCombieR. A. Wing, and R. K. Wilson. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science **326**:1112-1115.
Soltis, D. E., V. A. Albert, J. Leebens-Mack, C. D. Bell, A. H. Paterson, C. F. Zheng, D. Sankoff, C. W. dePamphilis, P. K. Wall, and P. S. Soltis. 2009. Polyploidy and angiosperm diversification. American Journal of Botany **96**:336-348.
Tang, H., M. R. Woodhouse, F. Cheng, J. C. Schnable, B. S. Pedersen, G. Conant, X. Wang, M. Freeling, and J. C. Pires. 2012. Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. Genetics **190**:1563-1574.
Thomas, B. C., B. Pedersen, and M. Freeling. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Research **16**:934-946.

Wang, X.H. WangJ. WangR. SunJ. WuS. LiuY. BaiJ.-H. MunI. BancroftF. ChengS. HuangX. LiW. HuaJ. WangX. WangM. FreelingJ. C. PiresA. H. PatersonB. ChalhoubB. WangA. HaywardA. G. SharpeB.-S. ParkB. WeisshaarB. LiuB. LiB. LiuC. TongC. SongC. DuranC. PengC. GengC. KohC. LinD. EdwardsD. MuD. ShenE. SoumpourouF. LiF. FraserG. ConantG. LassalleG. J. KingG. BonnemaH. TangH. WangH. BelcramH. ZhouH. HirakawaH. AbeH. GuoH. WangH. JinI. A. P. ParkinJ. BatleyJ.-S. KimJ. JustJ. LiJ. XuJ. DengJ. A. KimJ. LiJ. YuJ. MengJ. WangJ. MinJ. PoulainJ. WangK. HatakeyamaK. WuL. WangL. FangM. TrickM. G. LinksM. ZhaoM. JinN. RamchiaryN. DrouP. J. BerkmanQ. CaiQ. HuangR. LiS. TabataS. ChengS. ZhangS. ZhangS. HuangS. SatoS. SunS.-J. KwonS.-R. ChoiT.-H. LeeW. FanX. ZhaoX. TanX. XuY. WangY. QiuY. YinY. LiY. DuY. LiaoY. LimY. NarusakaY. WangZ. WangZ. LiZ. WangZ. Xiong, and Z. Zhang. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. Nature Genetics **43**:1035-U1157.

Wendel, J. F., and R. C. Cronn. 2003. Polyploidy and the evolutionary history of cotton. Pp. 139-186 *in* D. L. Sparks, ed. Advances in Agronomy, Vol 78.

Wolfe, K. H. 2001. Yesterday's polyploids and the mystery of diploidization. Nature Reviews Genetics **2**:333-341.

Woodhouse, M. R., F. Cheng, J. C. Pires, D. Lisch, M. Freeling, and X. Wang. 2014. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids (vol 111, pg 5283, 2014). Proceedings of the National Academy of Sciences of the United States of America **111**:6527-6527.

Woodhouse, M. R., J. C. Schnable, B. S. Pedersen, E. Lyons, D. Lisch, S. Subramaniam, and M. Freeling. 2010. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. Plos Biology **8**:e1000409.

Yoo, M. J., E. Szadkowski, and J. F. Wendel. 2013. Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity **110**:171-180.

Figures and Figure Legends

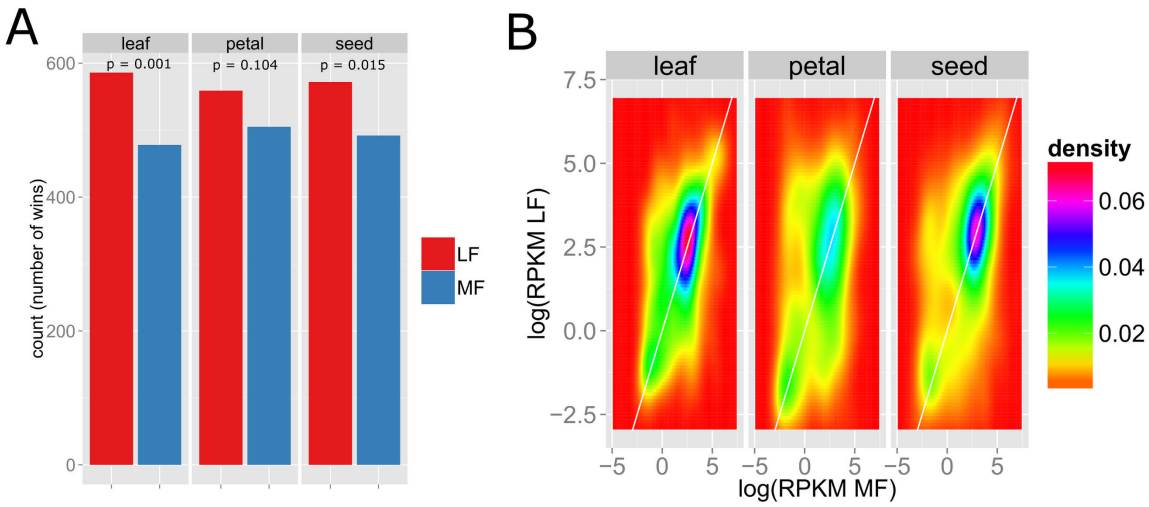

Figure 1. Gene expression in LF and MF fractions of the genome. Syntenic paralogs were binned according to residency on least fractionated (LF) or most fractionated (MF) chromosome reconstructions, and compared for gene expression levels in three tissues, petal, leaf and seed. Three biological replicates 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.

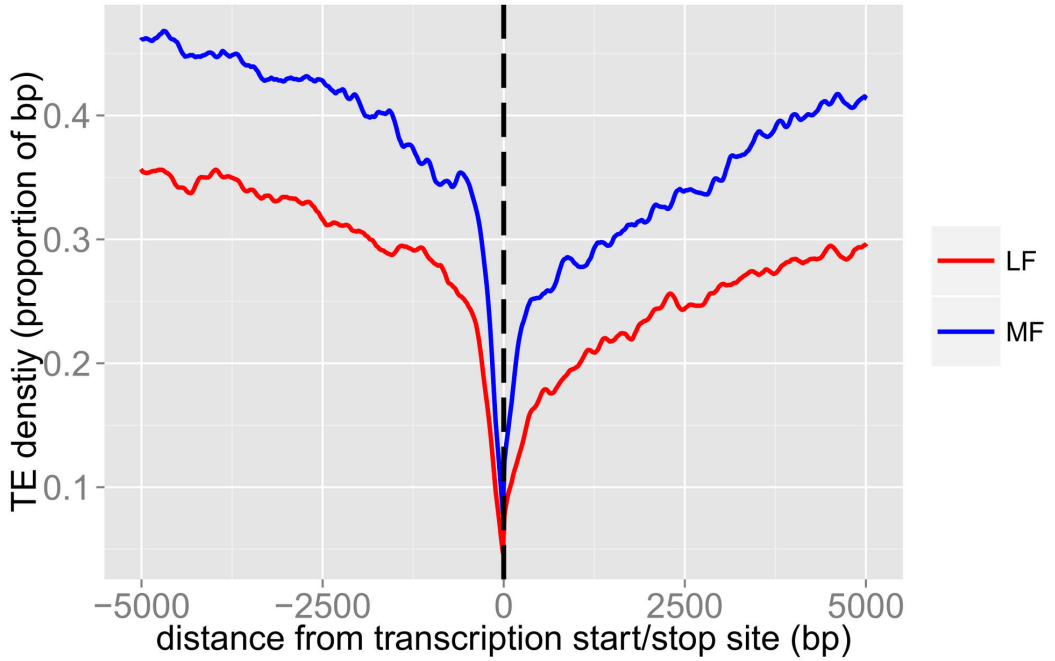

Figure 2. Transposable element (TE) density flanking genes in LF and MF fractions of the genome. Annotations from the cotton reference genome were used to identify TEs, and the proportion of base pairs annotated as TE-derived was calculated over sliding windows of 5000 bp (10 bp increments), either side of transcription start/stop sites. Mean density for LF and MF genes are displayed separately. The dashed line indicates the start/stop site and the intervening genic region is excluded from the plot.

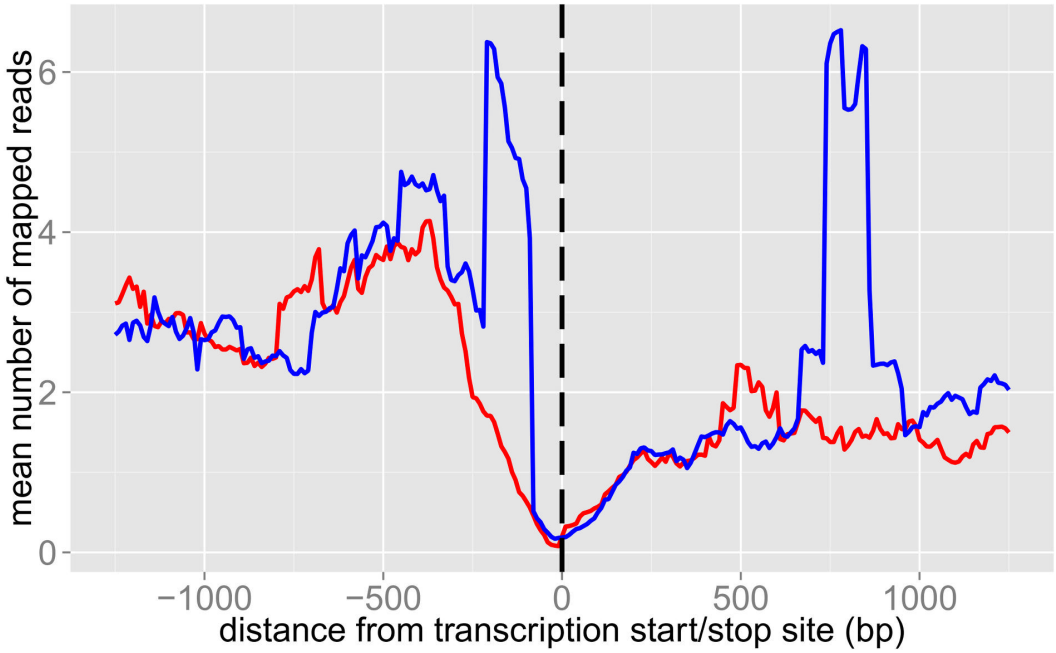

Figure 3. Enrichment of 24-nt siRNAs mapped to TEs that flank MF genes. All genes in LF and MF categories were assessed for uniquely mapped siRNAs, allowing no mismatches, 1,250 bp either side of transcription start/stop sites. I n order to limit mapping of TE-derived siRNA non-TE derived genomic sequences were masked. The vertical dashed line indicates the start/stop site and the intervening genic region is excluded from the plot.

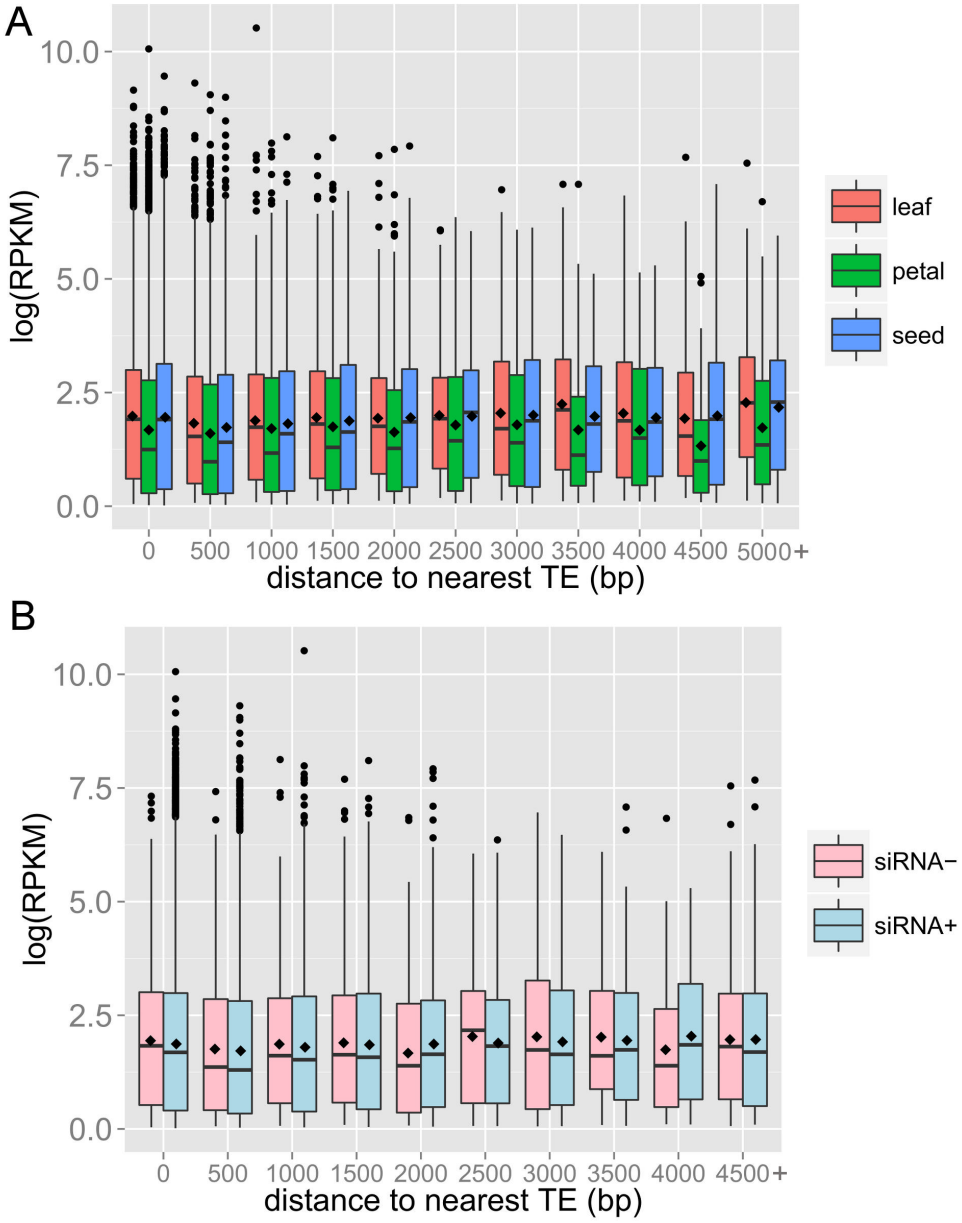

Figure 4. TE proximity, siRNAs and gene expression. (A) Boxplot of a transcriptomic analysis of 37,200 cotton genes examined for expression level and binned according the proximity to the nearest TE. Gene expression levels are given for each of three tissues. Boxes indicate 95% confidence intervals for 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 (Table S1).

Tables

Table 1 Biased fractionation in cotton following an ancient (60 mya) WGD.

T. cacao chromosome	*G. raimondii* chromosome (block numbers)	observed	predicted	χ^2	*p* value
2	**5** (137,138,139)	929	785.5	52.43	4.5×10^{-13}
	8 (179,184,185)	642	785.5		
6	**6** (149,150)	147	477	333.86	$<1 \times 10^{-15}$
	9 (190)	580	477		
	10 (33,34,36)	227	477		
7	**2** (88,86,89)	420	325.5	58.95	1.6×10^{-14}
	13 (76,75)	225	325.5		
8	**5** (133,132)	236	422	163.96	$<1 \times 10^{-15}$
	9 (191)	608	422		
9	**4** (113,114,130)	343	862	433.79	$<1 \times 10^{-15}$
	9 (188,189)	981	862		
	13 (79,80,81,82)	400	862		
10	**9** (195)	397	283.5	90.88	$<1 \times 10^{-15}$
	11 (44)	170	283.5		