Biological Sciences

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

Simon Renny-Byfield, Lei Gong, Joseph P. Gallagher, Jonathan F. Wendel\*

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

\*Corresponding author

email: jfw@iastate.edu

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

Abstract

The role of gene duplication in evolution has long been recognized. Indeed, ancient whole genome multiplication (WGM) is cyclical and ubiquitous in flowering plants, each round followed by substantial gene loss (fractionation). Fractionation may be biased with respect to homoeologs, often with over-expression of genes in less fractionated (LF) relative to more fractionated (MF) regions, hypothesized to arise through down-regulation of genes via silencing of local transposable elements (TEs). We report the genomic signatures of biased fractionation remain evident in the paleopolyploid genome of diploid cotton, despite 60 million years of evolution. Our data extend the evolutionary depth over which these signatures can be detected, suggesting that, hitherto, the long term impact of bias fractionation has not been fully realized. In cotton biased fractionation is associated with over-expression of genes in LF regions whereas MF regions exhibit elevated GC content, higher TE density and preferential mapping of siRNAs. Contrary to previous observations, we show that gene expression is not influenced by TE proximity, nor siRNA targeting, both of which are therefore unlikely to be primary drivers of biased fractionation. We propose repeat content *per se* and differential recombination rates may drive biased fractionation following WGM.

Significance Statement

Ancient whole genome multiplication (WGM) is universal in angiosperms and is followed by substantial gene loss (fractionation). Fractionation can be biased and correlated with over-expression of genes residing in the less fractionated regions, hypothesized to arise via silencing-effects of local transposable elements (TEs) and siRNAs. We report that genomic signatures of biased fractionation remain evident in cotton 60 million years after ancient WGM. The more fractionated regions have reduced GC content, higher TE density and preferentially map siRNAs. Importantly, we show that global gene expression is uncoupled from TE proximity and mapping of siRNAs, which are therefore unlikely to drive biased fractionation. Instead, we propose repeat content *per se* and differential recombination may be responsible for biased fractionation following WGM.

\body

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 (1-10). The realization that all flowering plants are multiply paleopolyploid indicates that, over time, a process of diploidization operates to return polyploids to a diploid-like condition (11-14). This transition involves reductions in genome size (15) and chromosome number (16, 17), as well as large-scale loss of duplicate genes (18-23). The latter process may lead to differential loss and retention of duplicated genes from homoeologous genomic regions and chromosomes, a phenomenon termed biased fractionation (19, 20, 23-25), although unbiased fractionation following WGM also is reported (23). The genomic signature of biased fractionation is that one set of homoeologous (duplicated) regions retains more genes than its counterparts.

Recent analyses in maize and *Brassica* (19, 26) indicate that regions experiencing greater rates of 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 (27-29), coffee (30) and *Tragopogon* (31). These observations have led to the hypothesis that following WGD, differences in gene expression between duplicated regions drive differential gene loss (32). This hypothesis is attractive in that nascent or evolutionarily young polyploids often exhibit biased homoeolog expression (27), suggesting that WGM, especially in allopolyploids (23), may establish the initial conditions that set in motion the mechanistic underpinnings of biased fractionation. This process is suggested to be accompanied by differential selection, where under-expressed genes contribute less to fitness than do their over-expressed homoeologous counterparts, and therefore are more likely to become dispensable (32).

Additional support for this hypothesis was generated by recent work involving *Brassica*; when compared to regions with lower levels of gene loss, regions experiencing greater gene loss were enriched for mapping of 24 nt siRNAs to transposable elements (TEs) surrounding resident genes (21). 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 (33). These observations of TE proximity, siRNA frequency, and relative gene expression of homoeologous genes in *Brassica* led to the suggestion (21) that positional-effect down-regulation by local TEs drives differences in gene expression between duplicate regions, thus providing a potential explanation for biased gene fractionation.

Here we demonstrate that the genomic footprints of biased fractionation following ancient (~ 60 mya) WGM persist 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.

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 (34) indicate that relative to cacao, the lineage that gave rise to modern diploid cotton experienced a 5-6 fold ploidy increase approximately 60 mya. Accordingly, we identified duplicate regions in the cotton genome resulting from the cotton-specific WGM (Fig. S1 and S2) previously reported in [??]. The relative antiquity of WGM in the cotton genome, however, makes identification of duplicate regions challenging compared to efforts in paleopolyploids where WGM, although still ancient, has occurred 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 (19) and *Brassica* (20)*.* We were, however, able to reconstruct a subset of the chromosomes, in the same manner as in Schnable et al. (19); 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 cotton regions (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 (20). The hypothesis of gene retention equivalence among homoeologs was evaluated using Pearson’s chi-squared test. 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 chrosmosome-wide bias fractionation following WGM on time-scales far in excess of those previously reported in *Brassica* and maize.

*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 identify retained paralogous genes in LF and MF chromosomes. 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 A); 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 of 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). 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 differential gene expression and fold-change greater than two, confirming previous results (Fig. S3). Thus, the data suggest that genes on LF chromosomes are typically expressed at higher levels than their MF counterparts.

*Local transposable element density* *and GC content*

Using transposable element (TE) annotations from the cotton reference genome (34), we identified TEs near each gene and used sliding windows to calculate the average TE density in regions surrounding each gene (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 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. 3). Furthermore, the GC content of LF chromosomes is significantly higher than that of MF in the 10,000 bp surrounding genes models (Wilcoxon’s paired sum rank test, V =0, p = <0.0001;Fig. 3).

*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 extracted siRNAs from leaves of *G. raimondii* and mapped the 24 nt siRNAs 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. 4), confirming observations also reported in *Brassica* (21). 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. 4).

*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 (35), we examined the relationship between TE proximity and gene expression. Importantly, expression was relatively uniform across all bins and across all tissues (Fig. 5 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. S5). 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-inducing TEs and compared the impact of TE proximity on gene expression in the two groups (Fig. 5 B). Analyses using linear modeling and ANOVA indicate that presence of siRNAs has little impact on the expression of nearby genes (Table S1).

Discussion

*Biased fractionation remains evident after 60 million years of evolution following genome multiplication*

The extensive occurrence of episodically recurrent WGM in the history of flowering plants has only recently become evident (2-5, 7, 8, 10, 36). 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). One interesting feature of neo- and meso-polyploids is genome downsizing (15), where the post-WGM genome partially returns to a diploid-like condition, by diploidization processes that are not well understood, and which occur at unknown rates. A universal attribute of this phenomenon is the loss of duplicated genes (18), 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 (19, 20, 22, 23, 25). Yet biased fractionation is not a universal feature of plant genomes; in *Musa* and *Glycine*, for example, genes appear to be randomly lost from duplicated regions (23).

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 process and reveals the signatures of bias fractionation are evident far beyond the time-frame reported for other genomes; around 12 mya post WGM in maize (Schnable and co) and a few million years in *Arabidopsis* and *Brassica* (Pires and others). Considering that most paleopolyploid genomes appear chromosomally and functionally diploid it is surprising that such signatures as well as differentiation of genomic content are still evident after eons. By extension we can deduce that the processes of bias fractionation has a long lasting legacy, hitherto not fully appreciated. Furthermore this leads to speculation as to what role such genomic differentiation might play over the course of angiosperm evolution and diversification.

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 (19-21). 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 genome multiplication (Fig. 1).

*What causes biased genome fractionation?*

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 (19). Under this hypothesis, conditions established at the time of genome merger (hybridization) and doubling (polyploidization), or those derived following polyploidization, generate widespread differences among homoeolog expression levels. These differences are commonly observed in cotton diploid hybrids and neo-allopolyploids (27, 37-40), 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 (41), gene expression levels among linked genes on homoeologous segments may be correlated (42). Finally, several recent studies in *Gossypium* have demonstrated that expression levels evolve (29, 42, 43). 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, which presumably creates the opportunity for differential selection. An important insight in this respect stems from the work of Freeling and colleagues (21), 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 demonstration (33) that gene expression may be down-regulated by positional effects associated with TEs, led to the suggestion (21, 32) 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. 4), with patterns similar to those recently described in *Brassica* (21). 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. 5 A). Furthermore, there is a slight negative correlation between gene expression and TE proximity, a trend that is opposite to expectations. Furthermore, the presence of local TEs that are silenced via siRNA pathways has no apparent impact on local gene expression levels (Fig. 5 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 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 (21). 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. 3). This observation, coupled with data showing that TE proximity appears to be de-coupled with expression levels (Fig. 5), suggests that some other genomic feature associated with increased TE density or GC 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. Recombination rate has been shown to be negatively correlated with TE prevalence (44, 45). Similarly, GC content can be positively correlated with recombination rate (46, 47). Therefore, higher TE density and lower GC content in MF suggest a history of reduced recombination relative to LF. This is important as selection is weaker in regions of low recombination (48, 49), 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. In other words, fixation of mildly deleterious deletions is likely to occur more often in the homoeolog where selection is less efficient, i.e. in MF. Moreover, at the time of polyploid formation, effective population size is likely to be very small, in which case, in the absence of effective selection on MF fragments, deletions could rapidly 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.

In this light we reiterate the observation of an additional genomic feature of note that differentiates LF and MF fractions, i.e., GC content (Fig. 3). The clear differentiation of LF and MF reconstructions suggests that at the time of duplication, these chromosomes may have had variable TE and nucleotide contents. This observation could support the proposition that biased fractionation stems from the initial conditions established by allopolyploidy, as suggested (23). It would be enlightening to examine the genomic characteristics studied here in other paleopolyploids where biased fractionation is evident, to assess the plausibility of the scenario forwarded.

Methods

*Biased fractionation*

We examined gene fractionation following the cotton-specific WGM using the reference genome sequences for *Gossypium raimondii* Ulbr. (34) and its close relative, *Theobroma cacao* L. (50). 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. (19). 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 (51), we took gene content and gene order in the outgroup *Theobroma cacao* to be representative of the ancestral model of thepre-duplicatedcotton genome, as has been done previously (19). We selected those reconstructions for which we had the greatest confidence of a full-length reconstruction, provided there was a homoelogous 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 (20). 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. (35). 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 (34) 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. (19) and Woodhouse et al. (21), 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 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 of seedling leaves from *Gossypium raimondii* were constructed and sequenced on the Illumina GA II sequencing platform (deposited in NCBI SRA database SRP017133 (52)). From the dataset we extracted 24-nt siRNAs and mapped these to the cotton reference genome using bowtie 0.12.7 (53). 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 GC content and local TE density*

For each genewe assessedthe 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 the 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 the TE annotations of (34) and *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 (34) 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

1. Blanc G & Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16(7):1667-1678.

2. Bowers JE, Chapman BA, Rong JK, & Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422(6930):433-438.

3. Jiao Y*, et al.* (2012) A genome triplication associated with early diversification of the core eudicots. *Genome Biology* 13(1).

4. Jiao Y*, et al.* (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature* 473(7345):97-100.

5. Leitch AR & Leitch IJ (2008) Perspective - genomic plasticity and the diversity of polyploid plants. *Science* 320(5875):481-483.

6. Paterson AH (2005) Polyploidy, evolutionary opportunity, and crop adaptation. *Genetica* 123(1-2):191-196.

7. Paterson AH*, et al.* (2000) Comparative genomics of plant chromosomes. *Plant Cell* 12(9):1523-1539.

8. Paterson AH, Bowers JE, & Chapman BA (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(26):9903-9908.

9. Renny-Byfield S & Wendel JF (2014) Doubling down on genomes: Polyploidy and crop plants. *American Journal of Botany* 101(10):1-15.

10. Soltis DE*, et al.* (2009) Polyploidy and angiosperm diversification. *American Journal of Botany* 96(1):336-348.

11. Clarkson JJ*, et al.* (2005) Long-term genome diploidization in allopolyploid *Nicotiana* section *Repandae* (Solanaceae). *New Phytologist* 168(1):241-252.

12. Mandakova T, Joly S, Krzywinski M, Mummenhoff K, & Lysak MA (2010) Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. *Plant Cell* 22(7):2277-2290.

13. Renny-Byfield S*, et al.* (2013) Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. *The Plant Journal* 74(5):829-839.

14. Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics* 2(5):333-341.

15. Leitch IJ & Bennett MD (2004) Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society* 82(4):651-663.

16. Abrouk M*, et al.* (2010) Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends in Plant Science* 15(9):479-487.

17. Lim KY*, et al.* (2007) Sequence of events leading to near-complete genome turnover in allopolyploid *Nicotiana* within five million years. *New Phytologist* 175(4):756-763.

18. Langham RJ*, et al.* (2004) Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* 166(2):935-945.

19. Schnable JC, Springer NM, & Freeling M (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(10):4069-4074.

20. Tang H*, et al.* (2012) Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. *Genetics* 190(4):1563-1574.

21. Woodhouse MR*, et al.* (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(17):6527-6527.

22. Woodhouse MR*, et al.* (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *Plos Biology* 8(6):e1000409.

23. Garsmeur O*, et al.* (2013) Two evolutionarily distinct classes of paleopolyploidy. *Molecular Biology and Evolution* 31(2):448-454.

24. Schnable PS*, et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326(5956):1112-1115.

25. 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 Research* 16(7):934-946.

26. Cheng F*, et al.* (2012) Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *Plos One* 7(5).

27. Grover CE*, et al.* (2012) Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist* 196(4):966-971.

28. Rapp RA, Udall JA, & Wendel JF (2009) Genomic expression dominance in allopolyploids. *BMC Biology* 7:18.

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

30. 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 Phytologist* 192(3):760-774.

31. Buggs RJ*, et al.* (2010) Characterization of duplicate gene evolution in the recent natural allopolyploid *Tragopogon miscellus* by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular Ecology* 19:132-146.

32. Freeling M*, et al.* (2012) Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Current Opinion in Plant Biology* 15(2):131-139.

33. Hollister JD*, et al.* (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(6):2322-2327.

34. Paterson AH*, et al.* (2012) Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* 492(7429):423–427.

35. Renny-Byfield S*, et al.* (2014) Ancient gene duplicates in *Gossypium* (Cotton) exhibit near-complete expression divergence. *Genome Biology and Evolution* 6(3):559-571.

36. Van de Peer Y, Maere S, & Meyer A (2009) The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* 10(10):725-732.

37. Adams KL, Percifield R, & Wendel JF (2004) Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics* 168(4):2217-2226.

38. Adams KL & Wendel JF (2005) Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* 8(2):135-141.

39. Hovav R*, et al.* (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(16):6191-6195.

40. Yoo MJ, Szadkowski E, & Wendel JF (2013) Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* 110(2):171-180.

41. Wendel JF & Cronn RC (2003) Polyploidy and the evolutionary history of cotton. *Advances in Agronomy, Vol 78*, ed Sparks DL), Vol 78, pp 139-186.

42. Flagel LE, Chen L, Chaudhary B, & Wendel JF (2009) Coordinated and fine-scale control of homoeologous gene expression in allotetraploid cotton. *Journal of Heredity* 100(4):487-490.

43. Yoo M-J & Wendel JF (2014) Comparative evolutionary and developmental dynamics of the Cotton (*Gossypium hirsutum*) fiber transcriptome. *PLoS Genet* 10(1):e1004073.

44. Fontanillas P, Hartl DL, & Reuter M (2007) Genome organization and gene expression shape the transposable element distribution in the *Drosophila melanogaster* euchromatin. *Plos Genetics* 3(11):2256-2267.

45. Rizzon C, Marais G, Gouy M, & Biemont C (2002) Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. *Genome Research* 12(3):400-407.

46. Birdsell JA (2002) Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Molecular Biology and Evolution* 19(7):1181-1197.

47. Meunier J & Duret L (2004) Recombination drives the evolution of GC-content in the human genome. *Molecular Biology and Evolution* 21(6):984-990.

48. Hill WG & Robertso.A (1966) Effect of linkage on limits of artificial selection. *Genetical Research* 8(3):269-&.

49. Kliman RM & Hey J (1993) Reduced natural-selection associated with low recombination in *Drosophila melanogaster Molecular Biology and Evolution* 10(6):1239-1258.

50. Argout X*, et al.* (2011) The genome of *Theobroma cacao*. *Nature Genetics* 43(2):101-108.

51. Kasahara M*, et al.* (2007) The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447(7145):714-719.

52. Gong L, Kakrana A, Arikit S, Meyers BC, & Wendel JF (2013) Composition and expression of conserved microRNA genes in diploid cotton (*Gossypium*) species. *Genome Biology and Evolution* 5(12):2449-2459.

53. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10(3).

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. GC content in LF and MF fractions of the genome. Shown are the proportions of GC nucleotides in regions surrounding LF and MF genes 5000 bp either side of gene transcription start/stop sites. The vertical dashed line indicates the start/stop site and the intervening genic region is excluded from the plot.

Figure 4. 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 5. 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).