2012-12-13

Transcriptome and Methylation Analysis of Gossypium Petal Tissue

Aditi Rambani
Brigham Young University - Provo

Transcriptome and Methylation Analysis of *Gossypium* Petal Tissue

Aditi Rambani

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

Master of Science

Joshua A Udall, Chair
Craig Coleman
William Evan Johnson

Department of Plant and Wildlife Sciences

Brigham Young University

December 2012

ABSTRACT

Transcriptome and Methylation Analysis of *Gossypium* Petal Tissue

Aditi Rambani
Department of Plant and Wildlife Sciences, BYU
Master of Science

Polyploidization instantly doubles all genome content by combining two genomes that have markedly different methylation and gene expression levels. This process may be accompanied by genetic and epigenetic changes in each genome. Sequencing of the transcriptome (RNA-seq) and the methylome (bisulfite treated libraries whole genome libraries) were used to measure gene expression and methylation levels of genic regions of allopolyploid cotton petals and petals of their diploid relatives. Many differentially expressed genes detected by RNA-seq were consistent with expression levels previously detected by microarrays. RNA-seq results also reconfirmed the presence of general polyploid gene expression trends like expression level dominance and homoeologous expression biases in *Gossypium* polyploid species. Expression biases between A- and D-genome homoeologs and expression level dominance was characterized for thousands of genes in tetraploids and a diploid F_1-hybrid. Unlike the results of microarray study previously done we found a slightly greater number of genes showing A-genome bias vs genes showing D-genome bias. More commonly the overall expression level from homoeologs of polyploid is heterotic i.e the expression level is greater than the average of the expression levels from the two parent genomes. In addition, genome methylation (CG, CHG, and CHH contexts) of each genome was assessed in the diploid and tetraploid samples. The A- and D-genomes had distinct levels of DNA methylation for each context. DNA methylation may be independently regulating homoeologous expression levels of a small number of genes.

Keywords: allotetraploid, cotton , transcriptome, RNA sequencing, duplicate gene expression, homoelogous gene expression bias, expression level dominance, bisulfite sequencing, methylome, DNA methylation, correlation of DNA methylation and gene expression

DEDICATION

I dedicate this work to my dear friend Krishna and my family. Their unconditional love and blessings made it all possible.

ACKNOWLEDGEMENTS

I am grateful to all those who helped me with this project. This project would not have been possible without the help and support of the kind people around me, to only some of whom it is possible to give particular mention here. Foremost is my advisor Dr. Joshua Udall, I am very thankful for his guidance, encouragement and patience over the past two years. I sincerely appreciate the direction and input my committee members Dr. Coleman and Dr. Evan Johnson provided for this project. I would also like to thank Justin Page for all the help with bioinformatics work and Kara Grupp for sending all the plant material used in this project. I am grateful to Brigham Young University for maintaining such a beautiful environment on campus. I would also like to thank my fellow students and staff at Department of Plant and Wildlife Sciences for their cooperation and help throughout my project. Finally, I would like to thank my family for constant support that helped me maintain right perspective and attitude towards my work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Gossypium Petal transcriptome analysis

INTRODUCTION

The genus of *Gossypium* originated about 10 million years ago and consists of approximately 45 diploid species (Wendel & Cronn, 2003). The five polyploid species (and possibly more) of this genus were formed 1-2 million years ago (Grover *et al.,* 2012b). The cells of modern alloteraploid contains two distinct diploid genomes denoted by A_T and D_T. The genome content and percent identity of the two genomes in the tetraploid nucleus are most closely related to the A_2 genome of *G. arboreum* and D_5 genome of *G. raimondii* (Senchina *et al.,* 2003). Since formation, these polyploids species have independently evolved and their monophyletic origin makes this genus an ideal system to study effects of polyploidization and independent domestication. Only two polyploids species produce spinnable fiber used by the textile industry. Superior cotton fiber qualities and yields have made accessions of tetraploid *G. hirsutum* more widely grown in cultivation than the other species, *G. barbadense* (Brubaker *et al.,* 1999). Global transcriptome analysis of *Gossypium* over the last decade has revealed many interesting transcriptomic consequences of polyploidization and domestication (Adams, 2007; Chaudhary *et al.,* 2009; Flagel & Wendel, 2009; Flagel *et al.,* 2008; Grover *et al*., 2012a; Grover *et al.,* 2004; Hovav *et al.,* 2008; Rapp *et al.,* 2009; Salmon *et al.,* 2005).

Polyploidization causes an immediate, simultaneous duplication of all DNA (including genes) and some the genomic consequences of polyploidization can be dramatic (Salmon *et al.,* 2005; Chelaifa *et al.,* 2010; Huang *et al.,* 2012). In cotton, duplicate

genes do not always contribute equally to the transcriptome during different stages of growth or stress. Using microarrays, it was observed that some duplicated gene pairs showed extreme expression bias of single genome (mimicked monoallelic expression) and while other duplicated gene pairs showed intermediate expression of both genomes (Flagel *et al.,* 2009; Hovav *et al.,* 2008). In petal tissue, it was found that about 76% of homeolog expression biases observed were determined immediately after genomic merger and 24 % were determined under evolutionary forces over time (Flagel *et al.,* 2008). In addition to cotton, homeologous expression bias has been reported in other natural and synthetic allopolyploid species (Bottley *et al.,* 2006; Buggs *et al.,* 2010; Koh *et al.,* 2010; Wang *et al.,* 2004; Chang *et al.,* 2010; Gaeta *et al.,* 2007; Auger *et al.,* 2009; Chelaifa *et al.,* 2010).

Another consequence of genomic merger is expression level dominance, which was first observed in leaf tissue of a synthetic *Gossypium* tetraploid (Rapp *et al.,* 2009; Grover *et al.,* 2012). Expression level dominance was determined if a gene was differentially expressed between the diploid parents, frequently its combined expression from homeologs is statistically equal to expression of only one of the parent donors. This dominance trend is also present in petal tissue and other natural *Gossypium* polyploids (Flagel & Wendel, 2009). Expression level dominance has been observed in other polyploid species such as *Coffea* (Bardil *et al.,* 2011) and *Spartina* (Chelaifa *et al.,* 2010) and wheat (Qi *et al.,* 2012). Factors that give rise to expression level dominance are still unclear, but interaction of regulatory machinery from two distinct genomes is one explanation (Osborn, 2003). External factors could also play a role since temperature was shown to influence the

magnitude and direction of expression level dominance in *Coffea* species (Bardil *et al.,* 2011).

A microarray platform for studying transcript contributions of the two co-resident cotton genomes made it possible to quantify individual expression of homeologous genes (Udall, 2009). However, a more accurate assessment of transcriptome composition is possible through RNA-seq technology because gene expression measurement by RNA-seq is not influenced by probe specificity, a prior template sequence, and cross-hybridization (Costa *et al.,* 2010). Here, we used RNA-seq to measure gene expression in several polyploid accessions of cotton within a phylogenetic framework.

METHODS

Plant Material

Six accessions were used in our study: *G. arboreum* (2x=2n=26, A_2), *G. raimondii* (2x=2n=26, D_5), *G. tomentosum* (4x=2n=52, AD_3), *G. hirsutum* cv. Acala Maxxa (4x=2n=52, AD_1; referred to as Maxxa), *G. hirsutum* cv. TX2094 (referred to as Tx) and a sterile diploid synthetic F_1-hybrid between A_2 and D_5 (1x = 1n = 26; F_1) (Table 1). The diploid synthetic F_1-hybrid was created by a hand pollination between reduced gametes of diploids *G. arboreum* (A_2) and *G. raimondii* (D_5), and its somatic cells only contains 13 chromosomes from each extant diploid genome (Table 1).

Petal tissue was collected from plants growing under controlled greenhouse conditions at the Pohl Conservatory, Iowa State University, USA. Tissue was harvested at the time of full petal expansion after dawn but before pollination. Taking one flower from three different plants made three biological replicates for experiments. Harvested tissue was flash frozen in liquid nitrogen and stored at -80^{o} C until RNA and DNA extraction.

RNA extractions, RNA-Seq Libraries and Sequencing

RNA samples were extracted from the three replicates using a modified hot borate method (Wan & Wilkins, 1994). RNA samples were quantified using Ribogreen (Invitrogen Inc., Grand Island, NY) and their quality was evaluated on an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA). As described by Illumina, cDNA was sheared by sonication to a 200-400 bp fragment size (Covaris Inc., Woburn, MA). RNA seq libraries were prepared according to the Illumina TruSeq RNA library prep kit protocol and sequenced on an Illumina HiSeq using v.2 chemistry at the Huntsman Cancer Center, SLC, UT.

Data Analysis

Quality filtering and Quantitative assessment of RNA Seq reads

Reads were filtered and trimmed using sickle with a phred quality threshold of 20 (https://github.com/najoshi/sickle,). Diploid and tetraploid sequencing reads were individually mapped using GSNAP (Wu & Nacu, 2010) to the diploid genome reference of *G. raimondii* (Paterson *et al., in press*). Tetraploid reads were categorized in two groups, A_T and D_T, using PolyCat (Page *et al. in press*) (Table 2). We assessed the transcript abundance for each gene and converted raw read counts to RPKM (reads per kilobase per million mapped reads).

Petal Transcriptome Analysis

Universal Probability of expression Codes (UPC) uses a mixed-model approach to quantify the probability of gene expression in a sample (Piccolo *et al.* 2011 *unpublished thesis*). Which genes were actively expressed in the petal tissue were determined using UPC for each accession (Figure 1). Active genes in all the accessions were called 'common genes'

and they were used to generate GO annotations for the petal transcriptome through Blast2GO. BLASTX was performed on the Fulton Super Computer at BYU. Blast2GO visual tools were employed to build pie charts depicting gene ontology (Figure 2). Utilizing GO annotations and Enzyme Codes (EC) the KEGG ids were assigned to each gene and the transcript abundance was calculated for KEGG pathways (Figure 3).

Differential expression analysis

The R-package EdgeR was used to normalize expression data and perform differential expression analysis (McCarthy *et al.,* 2012). Genes with less than 30 reads were filtered from the analysis. Two factors were used as explanatory variables in model design matrix: 'accession' with four levels (diploid F_1-hybrid, *G. tomentosum*, *G. hirsutum* TX2094, *G. hirsutum* Maxxa) and 'genome' with two levels (A-genome or D-genome). A simple single factor experiment with 8 levels was used to detect genes differentially expressed between two genomes for each accession. A single, nested interaction design was used to determine genes significantly differentially expressed between accessions. EdgeR performs exact test for the NB distribution coefficients to provide p-values and false discovery rates (q-values) for all the genes. Genes with <0.05 FDR were considered differentially expressed (Table 3).

Expression Phylogeny

Simple phylogeny based on expression levels of the genes from all the accessions was built using the neighbor-joining algorithm, with sum of squared differences across all the genes used as the distance between accessions (Figure 4). We built another phylogeny using the neighbor-joining algorithm based on differential homoeologous gene expression levels between all the accessions (Figure 5).

Expression Level Dominance Analysis

To analyze expression level dominance, every gene was analyzed for each polyploid accession and characterized according to the relationships between the RPKM values of the different genomes. Genes without expression in petals as determined by UPC were excluded from analysis. Each gene was categorized after comparison of A_2 and D_5 expression to the total expression of the polyploidy. A matrix was constructed with the number of genes that fit into each combination of classes from the two comparisons (Table 4).

RESULTS

Measurement of Gene Expression

A large total of RNA-seq reads were generated from three replicates of each accession (Table 1). Maxxa and *G. tomentosum* had the most RNA-seq reads (> 40M each) and diploid D5 had the least amount of RNA-seq reads (~37M). Each of these reads was mapped (or aligned) to the D-genome reference sequence (v. 2.2.1) containing an initial set of gene annotations. Not all the reads mapped to the reference genome sequence. Perhaps, this is because the initial draft of the D-genome reference did not have all of the genes annotated (Paterson *et al.,* 2012 *in press*) or transcripts mapped to genomic regions outside of annotated genes. Of the annotated genes, 80% had at least one mapped read.

The genome origin was identified for approximately 50% of mapped reads. If the mapped reads overlapped a homoeo-SNP position (SNPs between the A- and D-genomes), they were categorized as belonging to one of the two genomes or as a chimeric read because it had A- *and* D-genome bases at different loci (A-Reads, D-reads, and X-reads, respectively; Page *et al.* 2012 *in press*). If they did not overlap a homoeo-SNP position, the

read was unable to be categorized as originating from either the A- or D-genome (N reads) (Table 2). The remainder of the mapped reads did not overlap a homoeo-SNP in each of the polyploid samples and they were not categorized. This result is not unexpected given the limited divergence between the A_T- and D_T-genome in coding sequences (Flagel *et al.,* 2012).

Based on the UPC analysis, only 45-50% of the genome is expressed in petals. This total is lower than the number of cotton genes found to be expressed in fiber tissue (75-90%) at each developmental stage (Hovav *et al.,* 2008). Out of 37,224 genes annotated in the reference D-genome, 15,497 genes were commonly expressed in the petal tissue of all the polyploid accessions (Table 3). This amount of commonly expressed genes represented approximately 85% of expressed genes in each accession. Many of the commonly expressed genes may have been involved with essential functions of petal tissue (Figure 1). Using Blast2GO, we assigned GOIDs to the 15,497 common genes based on their RefSeq Blast hits and categorized them into three separate gene ontologies according to their putative function (Figure 2).

The cellular component (CC) ontology had the highest number of assigned GOIDs (88%) followed by the biological process (BP) ontology (17%) and molecular process (MP) ontology (9%). The most abundant GO terms of CC were cytoplasm related (cytoplasm (28%) and cytoplasmic part (27%)) (Figure 2). Cellular protein metabolic processes (31%) and kinase activity (41%) were the most plentiful GO terms for the BP and MP ontologies respectively (Figure 2). Similar distributions among categories have also been reported from the petal tissue of other species like *Dianthus* (Tanase *et al.,* 2012) and Safflower (Li *et al.,* 2012). Enzyme-coding genes were identified and their role in KEGG enzymatic

pathways was determined. Total 4,565 genes were assigned an enzyme code id, but these genes only corresponded to 654 different enzymes (*i.e.* many genes were members of large gene families). These 654 enzymes were found to be part of 93 different enzymatic pathways in petal tissue. The enzymatic pathways can be divided into four general categories: Metabolic pathways, Biosynthetic pathways, Degradation pathways and Signaling pathways (Figure 3). Transcript abundance of genes involved in metabolic pathways like starch and amino sugar metabolism was highest in petal tissue compared to other enzymatic pathways. Amongst biosynthetic pathways, biosynthesis of amino acids and flavonoids were most abundant whereas other processes like wax and pigment synthesis had smaller representation.

Differential Gene Expression

The phylogenetic relationships of *Gossypium* species have been well characterized. It is also possible to use gene expression levels to visualize these evolutionary relationships (Flagel *et al.,* 2009). Our RNA-seq results support these previous findings where the *Gossypium* expression phylogram had branching patterns similar to the genetic phylogram. A single phylogenetic tree that had two main branches containing the A_T- and D_T-genomes, respectively, illustrated the expression level differences of each genome (Figure 5). As expected, the respective genomes of the two *G. hirsutum* accessions, Maxxa and TX2094, were most closely related and clustered together. Differential expression analysis showed that there were only 692 genes differentially expressed between these two accessions of *G. hirsutum.* There were 1,394 genes differentially expressed between the two accessions of *G. hirsutum* and *G. tomentosum*. The diploid F_1-hybrid was found have expression patterns more closely related to the diploids species than the natural polyploids (Figure 5). The

diploid F_1-hybrid had 2,671 genes differentially expressed between it and the natural polyploids.

The two co-resident genomes of polyploid nuclei do not always contribute equally to the transcript pool (as measured by RNA-seq). In cotton, unequal contribution by the A_T- and D_T-genomes to the transcript pool of any single gene is referred to as 'genome bias'. Approximately 20% of genes expressed in petals had a significant bias towards the A_T- or D_T-genome (Table 3). A slight overall bias towards A-genome was observed in all the accessions. To compare the homoelogous expression biases between accessions transcript contributions of 15,497 commonly expressed homoeologous gene pairs were evaluated. The homoeologous expression phylogeny had the same topology as the expression tree and summarized homoeologous expression biases (Figure 6). The diploid F_1-hybrid was relatively distant from natural tetraploids in the phylogeny and it had the least number of biased genes. More than half of the biased genes (1,195) in synthetic diploid F_1-hybrid were biased in it and not in all the other accessions (Figure 4). Among the natural tetraploids, TX2094 had the highest number of biased genes followed by G. *tomentosum* and then Maxxa (Table 3).

Expression Level Dominance and homoeologous gene expression trends

Expression level dominance refers to a comparison of total expression of a duplicate gene pair in a polyploid nucleus to the expression level measured in diploid ancestors (or surrogates thereof; Grover *et al.,* 2012). In this case, comparison of interest is between the sizes of the total transcript pool, instead of their respective constitutions. Thus, gene expression in a tetraploid is called additive if it is equal to the average expression of the two diploid parents (the mid-parent value) and non-additive when it is unequal. All

plausible expression combinations between a tetraploid and its two parental diploids have been described in 12 separate possible expression categories (Rapp *et al.,* 2009). Homoeolog expression levels of polyploids were compared to expression levels of the diploid accessions (Table 4). Very few genes displayed an additive expression pattern where the polyploid genome had an expression level equal to the average of the two diploids (categories I and XII). The majority of expressed genes had equal expression levels in the polyploid and the diploids (A=P=D; where 'A' represents the diploid gene expression of A_2, 'D' represented the diploid gene expression of D_5, and 'P' represented the polyploid gene expression [A_T + D_T] of the individual tetraploid accessions). This result indicated that most genes had a finite limit or a functional limit to the amount of gene expression in the polyploid genome because 2x gene copies did not result in 2x expression (i.e. dosage compensation). Other categories of expression level dominance were also interpreted as evidence for dosage compensation because the polyploid expression level was equal to one of the two diploids (II, IV, IX, and XI). Of these four categories, the polyploid genome consistently had many more genes with expression levels equal to the higher of the two diploids (II and IV; 994 genes) than genes with expression levels equal to the lower of the two diploids (IX and XI; 110 genes). If categories IV and IX were more frequent than categories II and XI, then the A genome would be considered to be the expression level dominant genome and vice versa. None of these accessions appeared to exhibit expression level dominance; and considered jointly, the degree of expression level dominance was not significantly different than 1 ($\chi 2$ test; $p > 0.05$).

In the remainder of the categories, the polyploid genome had a more extreme expression levels than either of the two diploid genomes (categories III, V, VI, VII, VIII and

X) and constituted approximately 40% of expressed genes in each accession. In these categories, there were 20x-89x more cases where the polyploid genome had a greater expression level (compared to both diploid genomes) than a lower expression level (compared to both diploid genomes). If this were a diploid study of F1 hybrids, these patterns of gene expression would be considered as a heterotic pattern because most the majority of these genes display expression over-dominance. In addition, the over-dominant categories V, VI, and VIII outnumbered the categories of expression level dominance (II and IV) nearly 4:1. Thus, the most frequent exception to equal expression of polyploid and diploid genomes is a gene expression level that is non-additive and 'heterotic'.

DISCUSSION

Gossypium Petal transcriptome

RNA sequencing technology has emerged as an excellent tool for transcriptomic studies (Costa *et al.,* 2010). It is being extensively employed for gene discovery and detection of differential gene expression between different developmental. Fiber tissue has been main focus of many transcriptome studies of Gossypium species since it is the most economically important tissue of the plant (Udall, 2009). Gossypium petal transcriptome analysis through microarray technology has revealed several interesting polyploid duplicate gene expression trends (Flagel *et al.,* 2008; Rapp *et al.,* 2009; Grover *et al.,* 2012). None of the previous studies documented functional annotation of genes expressed in the petal tissue of Gossypium. For this project we performed deep sequencing of petal transcriptome of six Gossypium species using Illumina high throughput sequencing platform. The expression levels and functional annotation of transcripts were determined by reference mapping to *G. raimondii* reference genome (Paterson *et al.,* 2012 *in press*).

It was found only 45-50% of the Gossypium genome is expressed in the petal tissue at full expansion stage. Such low genic expression diversity could be due to simple cell histology and function of the petal tissue. About 70-85% of the genes expressed in the petals of four polyploid accessions were common. Petals suffice to a very basic function and has a very short life of one day, for this reason it may have undergone canalized evolution and we see lack of genic expression diversity. Though majority of genes expressed in the petal tissue are common for all four accessions they differ in levels of expression. The expression level variation can arise due to environmental pressures and reflects different natural histories of the accessions. The expression variation most likely preceded the genetic variation and we see that the phylogenetic relationships can be clearly seen through simple neighbor joining tree based on gene expression levels (Figure 5).

Distributing transcripts in GO categories developed a molecular snap shot of the petal tissue. The cellular component ontology that includes multi-subunit enzymes and other protein complexes was most abundant GO category (88%). Petal cells undergo rapid elongation to reach full petal expansion stage. Actin cytoskeleton helps with cell elongation by transporting vesicles and organelles to the site of growth from cytoplasm. The cytoplasm (28%) and cytoplasmic parts (27%) were most represented under cellular component GO category. About 17 % of transcripts fell under biological processes GO category and under this category cellular protein metabolic processes (31%) were most prominent. Petal tissue is an energy sink tissue for plant reproduction where starch and sucrose are mobilized from photosynthetic organs and broken down to sugars that function as precursors to essential primary and secondary metabolites (Muhlemann *et al.*, 2012). This was confirmed by looking at transcript abundance of different KEGG pathways.

It was found that most enzymes expressed in petal tissue were involved in starch and sucrose metabolism pathways (Figure 3).

Amongst biosynthetic pathways, biosynthesis of amino acids and flavonoids had most number of transcripts. Synthesis of wax and pigments also occur in petal tissue particularly at the petal base but only a low level of the transcripts coding for these enzymes were detected. Different genes are upregulated or downregulated at different developmental stages of the cell and at the time of harvest submerging tissue in liquid nitrogen freezes the molecular activity of the cell. The tissue for our study was harvested at full petal expansion stage. Flavonoids are mainly involved in production of fragrant volatile chemicals that attract pollinators, so these genes are activated when the petals are fully expanded. Pigment biosynthesis genes were expressed less than flavonoid biosynthetic pathways because pigment biosynthesis likely precedes petal expansion and anthesis.

Heterotic pattern of homoelogous gene expression and expression level dominance

The cumulative expression from the homoeologs of the allopolyploids was compared to the expression levels of the diploid parents and categorized into 12 expression state categories (Table 4). The expression states can be broadly described as additive gene expression (allopolyploid expression equal to the average expression of the diploid parents), non additive expression (allopolyploid expression NOT equal to the average expression of the diploid parents) and expression level dominance (when the diploid parent expression levels are unequal and the allopolyploid expression equals to only one of the diploid parents and not the average). If the gene expression regulatory factors change proportionally with the change in ploidy level an additive type of gene

expression state is observed. In our study we found that non-additive expression state with transgressive upregulation of expression (categories V, VI, VIII) was more common than an additive state of expression. Heterotic gene expression patterns can arise more easily in allotetraploids formed by fusion of two divergent diploid parents. The diploid parents *G. raimondii* (D_5) and *G. arboreum* (A_2) diverged from a common ancestor around 7MYA and evolved on different continents. Over the generations heterozygosity remains fixed in allotetraploids because the chromosomes from divergent genomes are unable to pair and intergenomic recombination is prevented. This fixed heterozygosity at many loci can result in 'over dominance' and heterotic expression patterns.

The microarray-based studies in the past on synthetic and natural Gossypium allopolyploids had found 'expression level dominance' categories to be more common over other expression categories (Rapp *et al.,* 2009; Flagel *et al.,* 2009) but we did not see this in our data. The non-additive categories with heterotic gene expression patterns accounted for 39% of the total genes and it is was higher than the expression level dominance categories (29%) (categories II, IV, IX, XI). Even in the 'expression level dominance' categories the upregulation is seen more frequently as the diploid parent with higher expression shows the 'expression level dominance' more often (eg. P=A if A>D or P=D if D>A). This dominance does not seem to favor one sub genome more over the other and occurs in similar numbers for both A and D genome. The gene redundancy in allopolyploids can help compensate for the low expressing recessive alleles of one subgenome with dominant alleles of other subgenome giving rise to such 'expression level dominance'. There could be other factors involved behind this phenomenon for example in *Coffea* species the 'expression level dominance' was found to be affected by the temperature

(Bardil *et al.,* 2011). In this project we have made an attempt only to elucidate duplicate gene expression trends that can arise because of many different factors that still need to be explored.

Homoeologous Gene Expression Bias

RNA sequencing technology was used to sequence petal transcriptome and all the RNA reads were categorized as A_T and D_T based on homoeSNPs to determine transcript contribution of two subgenomes towards transcriptome pool. Adam *et al.,* 2003 first reported expression bias in synthetic and natural gossypium polyploids for eight homoeologous genes. Using microarray technology homoeologus expression biases have since been observed in different tissues from all five known Gossypium polyploids and synthetic diploid F1 hybrids (Flagel *et al.,* 2008). Flagel *et al.,* 2008 reported that 'D' sub-genome is favored more over 'A' in petal tissue as more number of genes showed bias towards D sub-genome, this is in contrast to our findings. The number of genes showing bias favored both genomes almost equally (Table 3) with slightly more inclination towards A genome. Homoeologous genome bias towards 'A' sub-genome has been reported in ovular tissue of gossypium allopolyploids (Yang *et al.,* 2006). The previous studies that reported a D genome expression bias in petal tissue were done using microarray technology (Udall 2009; Flagel *et al.,* 2009). High sequence identity between homoeologous loci makes a probe based microarray technology more error prone and reduces accuracy of the estimates.

Evolution of duplicate gene expression

We compared expression patterns of synthetic diploid hybrid F1 with natural polyploids and diploids to see what shifts in expression patterns occur due to hybridization event and what are due to long-term evolutionary forces. Post genomic merger, about 15 % of the genes in F1 synthetic hybrid acquired significant bias towards one of the sub-genomes. These are the number of genes that are set up for sub- or neo- functionalization immediately post-genomic merger. Eventually under evolutionary pressures other genes replace many of these biased genes, about 40-44% of genes biased in F1 were found biased in natural tetraploids. A microarray study in past has also observed this trend where about 44% of genes were only found biased in F1 and not in AD_1 accession (Flagel *et al.,* 2008). They concluded that besides mechanism of sub-neo- and non- functionalization these differences between F1 and natural polyploids could arise because the synthetic diploid hybrid F1 parents are not exactly same as the diploid progenitors of natural polyploids. The homoeologous biases do not strongly favor one sub genome over other, a slight 'A' genome bias found is probably introduced at the time of genomic merger and is subsequently maintained in the natural polyploids (Table 3).

The 'expression level dominance' phenomenon appears in synthetic diploid hybrid F1 much more frequently than in natural polyploids (Table 4). We also see slightly more number of genes showing 'D' genome 'expression level dominance' like reported before but this favorability was less drastic in our case compared to these earlier reports (Rapp *et al.,* 2009; Flagel *et al.,* 2009). Genes for whom diploid parents have unequal expression more commonly show either 'expression level dominance' or transgressive upregulation. In synthetic diploid hybrid F1 'expression level dominance' categories have higher number of

genes whereas in natural polyploids transgressive upregulation categories have more number of genes. The synthetic diploid hybrid F1also has lesser number of genes showing additive expression compared to natural polyploids. We can conclude that over time under evolutionary pressures the genes tend to adopt a dosage balanced additive expression level or move to a more favorable transgressive upregulated level that may be conferring some selection advantage.

REFERENCES

Adams, K.L, Cronn, R., Percifield, R., Wendel, J.F. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc Natl Acad Sci USA 100: 4649-4654.

Adams, K. L. (2007). Evolution of Duplicate Gene Expression in Polyploid and Hybrid Plants. *Journal of Heredity*, *98*(2), 136–141. doi:10.1093/jhered/esl061

Bardil, A., de Almeida, J. D., Combes, M. C., 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. doi:10.1111/j.1469-8137.2011.03833.x

Brubaker, C. L., Paterson, A. H., & Wendel, J. F. (1999). Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. *Genome*, *42*(2), 184–203.

Chaudhary, B., Flagel, L., Stupar, R. M., Udall, J. A., Verma, N., Springer, N. M., & Wendel, J. F. (2009). Reciprocal Silencing, Transcriptional Bias and Functional Divergence of Homeologs in Polyploid Cotton (Gossypium). *Genetics*, *182*(2), 503–517. doi:10.1534/genetics.109.102608

Chelaifa, H., Monnier, A., & Ainouche, M. (2010). Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species Spartina× townsendii and Spartina anglica (Poaceae). *New Phytologist*, *186*(1), 161–174.

Costa, V., Angelini, C., De Feis, I., & Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-Seq. *Journal of Biomedicine and Biotechnology*, *2010*.

Flagel, L. E., & Wendel, J. F. (2009). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New*

Phytologist, *186*(1), 184–193. doi:10.1111/j.1469-8137.2009.03107.x

Flagel, L. E., Wendel, J. F., & Udall, J. A. (2012). Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton. *BMC Genomics*, *13*(1), 302. doi:10.1186/1471-2164-13-302

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*(1), 16. doi:10.1186/1741-7007-6-16

Grover, C. E., Gallagher, J. P., Szadkowski, E. P., Yoo, M. J., Flagel, L. E., & Wendel, J. F. (2012a). Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist*.

Grover, C. E., Grupp, K. K., Wanzek, R. J., & Wendel, J. F. (2012b). Assessing the monophyly of polyploid Gossypium species. *Plant Systematics and Evolution*, *298*(6), 1177–1183. doi:10.1007/s00606-012-0615-7

Grover, C. E., Kim, H. R., Wing, R. A., Paterson, A. H., & Wendel, J. F. (2004). Incongruent patterns of local and global genome size evolution in cotton. *Genome Research*, *14*(8), 1474–1482.

Hovav, R., Udall, J. A., Hovav, E., Rapp, R., Flagel, L., & Wendel, J. F. (2008). A majority of cotton genes are expressed in single-celled fiber. *Planta*, *227*(2), 319–329.

Li, H., Dong, Y., Yang, J., Liu, X., Wang, Y., Yao, N., Guan, L., et al. (2012). De Novo Transcriptome of Safflower and the Identification of Putative Genes for Oleosin and the Biosynthesis of Flavonoids. (A. Tramontano, Ed.)*PLoS ONE*, *7*(2), e30987. doi:10.1371/journal.pone.0030987.t001

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of

multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research*, *40*(10), 4288–4297.

Muhlemann, J. K., Maeda, H., Chang, C. Y., San Miguel, P., Baxter, I., Cooper, B., Perera, M. A., et al. (2012). Developmental Changes in the Metabolic Network of Snapdragon Flowers. *PLoS ONE*, *7*(7), e40381.

Qi, B., Huang, W., Zhu, B., Zhong, X., Guo, J., Zhao, N., Xu, C., et al. (2012). Global transgenerational gene expression dynamics in two newly synthesized allohexaploid wheat (Triticum aestivum) lines. *BMC Biology*, *10*(1), 3. doi:10.1186/1741-7007-10-3

Rapp, R. A., Udall, J. A., & Wendel, J. F. (2009). Genomic expression dominance in allopolyploids. *BMC Biology*, *7*(1), 18. doi:10.1186/1741-7007-7-18

Salmon, A., Ainouche, M. L., & Wendel, J. F. (2005). Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). *Molecular Ecology*, *14*(4), 1163–1175.

Tanase, K., Nishitani, C., Hirakawa, H., Isobe, S., Tabata, S., Ohmiya, A., & Onozaki, T. (2012). Transcriptome analysis of carnation (Dianthus caryophyllus L.) based on next-generation sequencing technology. *BMC Genomics*, *13*(1), 292.

Udall, J. A. (2009). The Gossypium Transcriptome. *Genetics and Genomics of Cotton*, 1–29.

Wan, C. Y., & Wilkins, T. A. (1994). A Modified Hot Borate Method Significantly Enhances the Yield of High-Quality RNA from Cotton (Gossypium hirsutum L.). *Analytical biochemistry*, *223*(1), 7–12.

Wendel, J. F., & Cronn, R. C. (2003). Polyploidy and the evolutionary history of cotton. *Advances in Agronomy*, *78*, 139–186.

Wu, T. D., Wu, T. D., Nacu, S., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex

variants and splicing in short reads. *Bioinformatics*, *26*(7), 873–881.

doi:10.1093/bioinformatics/btq057

TABLES

Table 1. List of plant materials used in this study

Species Name	Genome Designation	Accession	Ploidy level	Location	Read Number
G. arboreum	A_2	AKA8401	diploid	Africa	39,229,888
G. raimondii	D_5	GN33	diploid	South America	36,756,492
G. hirsutum	AD_1	Maxxa	tetraploid	Mexico	43,247,980
G. hirsutum	AD_1	TX2094	tetraploid	Yucatan Peninsula	38,350,345
G. tomentosum	AD_3	WT936	tetraploid	Hawaii	42,047,506
G. arboreum X G. raimondii	A_2 x D_5	Unnamed	F_1-haploid	NA	39,974,015

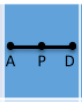
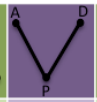
Table 2. Number of reads (Millions) that were categorized from reference mapping RNA-seq reads

Accessions	A-Reads	D-Reads	X Reads	N Reads	Mapped Total	Mapped %
G. arboreum	16.5	0.1	0	14	30.8	73.30%
G. raimondii	0	17.1	0	16.5	33.9	84.70%
Diploid F_1-Hybrid	8	8.4	0.1	15.1	32	78.80%
G. hirsutum Maxxa	7	6.7	1.2	13.5	28.6	77.70%
G. hirsutum Tx2094	8	7.7	1.4	15.7	33.1	76.60%
G. tomentosum	7.3	6.9	1.3	14.2	29.8	77.60%
Total	46.8	46.9	4.1	88.9	188.2	78.10%

Table 3. The number of genes expressed in each *Gossypium* accession, the total number shared by every accession, and the number of genes found to have unequal transcript contribution of both genomes (A_T and D_T) to the transcript pool (genome bias).

Accession	Total Expressed	Total Bias	A_T Bias	D_T Bias	% A Bias	% D Bias
Diploid F_1-hybrid	18,871	3,014	1,560	1,454	51%	48%
G. tomentosum	18,295	3,250	1,691	1,558	52%	47%
G. hirsutum Maxxa	18,832	3,022	1,570	1,452	51%	48%
G. hirsutum Tx2094	18,180	3,353	1,724	1,628	51%	48%
Common	15,497	757	396	305		

Table 4. Number of genes in 12 categories listed in first column where 'A' = expression from A genome, 'D'= expression level from D genome, and the 'P' = expression level from polyploid. The position of letters A, D and P indicate the level of expression relative to the other. Columns shaded blue show additive expression, in green show upregulation, in purple show down regulation and in orange show expression level dominance.

Categories	0	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
													
F1 Synthetic Diploid Hybrid	2,769	148	2,805	1	2,553	757	660	18	3,492	52	3	82	151
G. tomentosum	3,358	261	1,939	10	2,039	1,073	1,040	43	3,690	166	15	117	167
G. hirsutum Maxxa	3401	227	1,863	19	1,856	1,109	1,027	80	3,535	205	21	162	195
G. hirsutum TX2094	3,510	219	1,929	8	1,872	998	941	52	3,488	171	17	169	219

FIGURES

Figure 1. Venn Diagram for genes expressed in all the accessions above the background expression level.

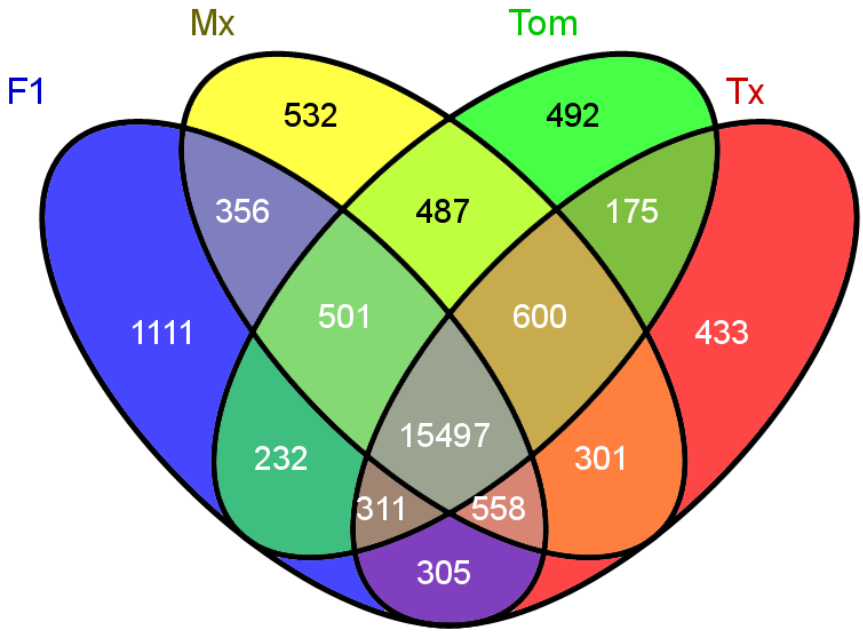
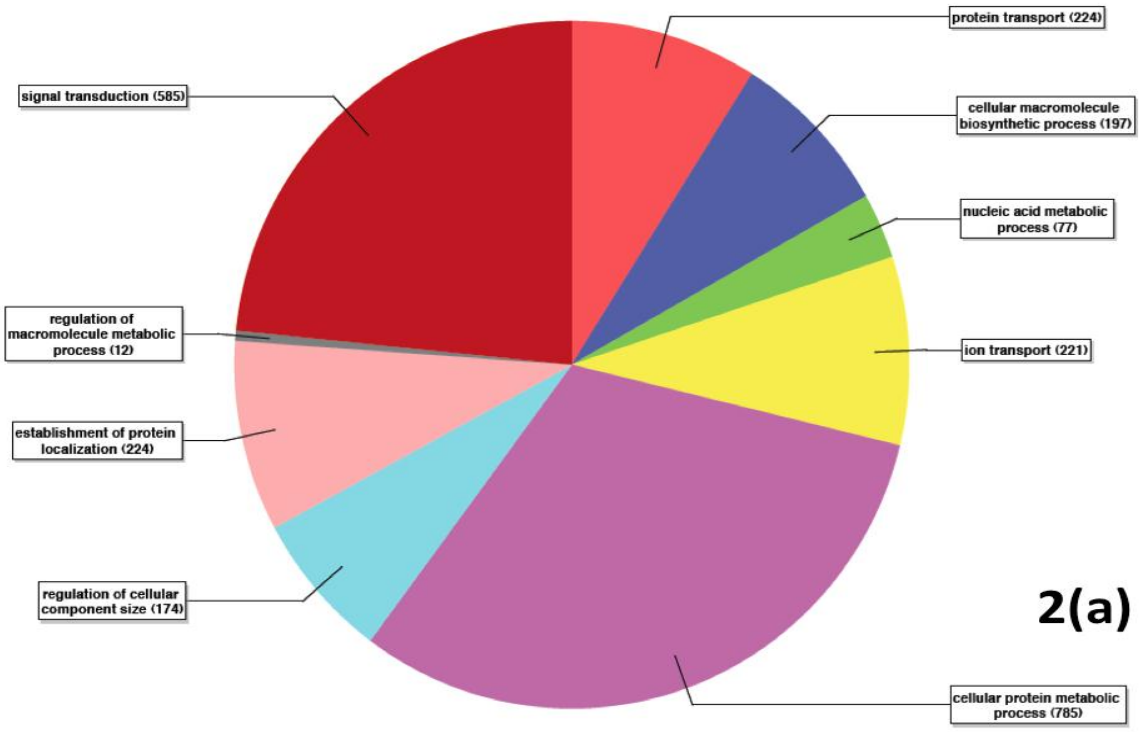
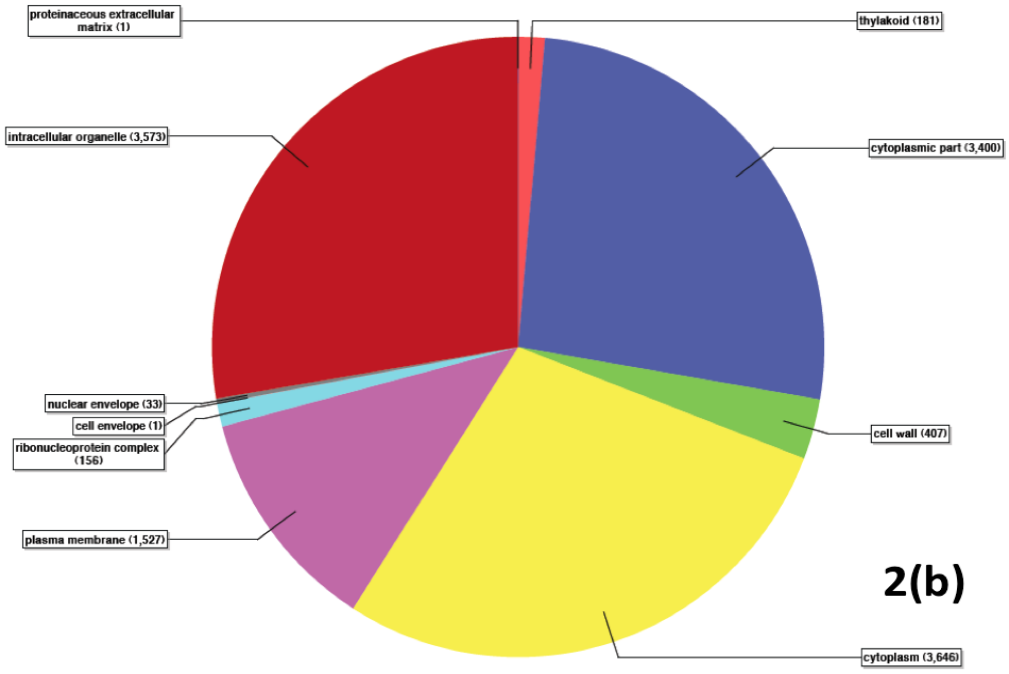


Figure 2. GO Terms Distribution for petal transcriptome (a) Biological processes (b) Cellular Component (c) Molecular Processes




proteinaceous extracellular matrix (1)
thylakoid (181)
intracellular organelle (3,573)
cytoplasmic part (3,400)
nuclear envelope (33)
cell envelope (1)
ribonucleoprotein complex (156)
cell wall (407)
plasma membrane (1,527)
cytoplasm (3,646)
2(b)

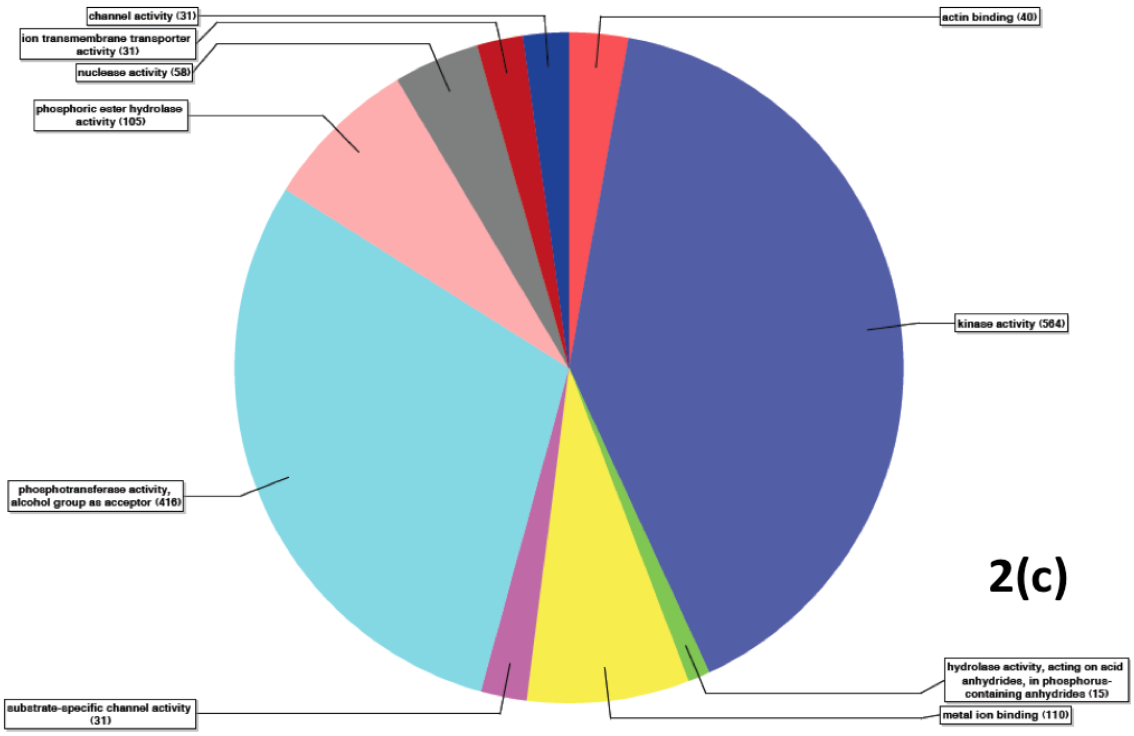

channel activity (31)
ion transmembrane transporter activity (31)
nuclease activity (58)
phosphoric ester hydrolase activity (105)
actin binding (40)
kinase activity (564)
phosphotransferase activity, alcohol group as acceptor (416)
substrate-specific channel activity (31)
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides (15)
metal ion binding (110)
2(c)

Figure 3. Quantification of genes according to KEGG processes

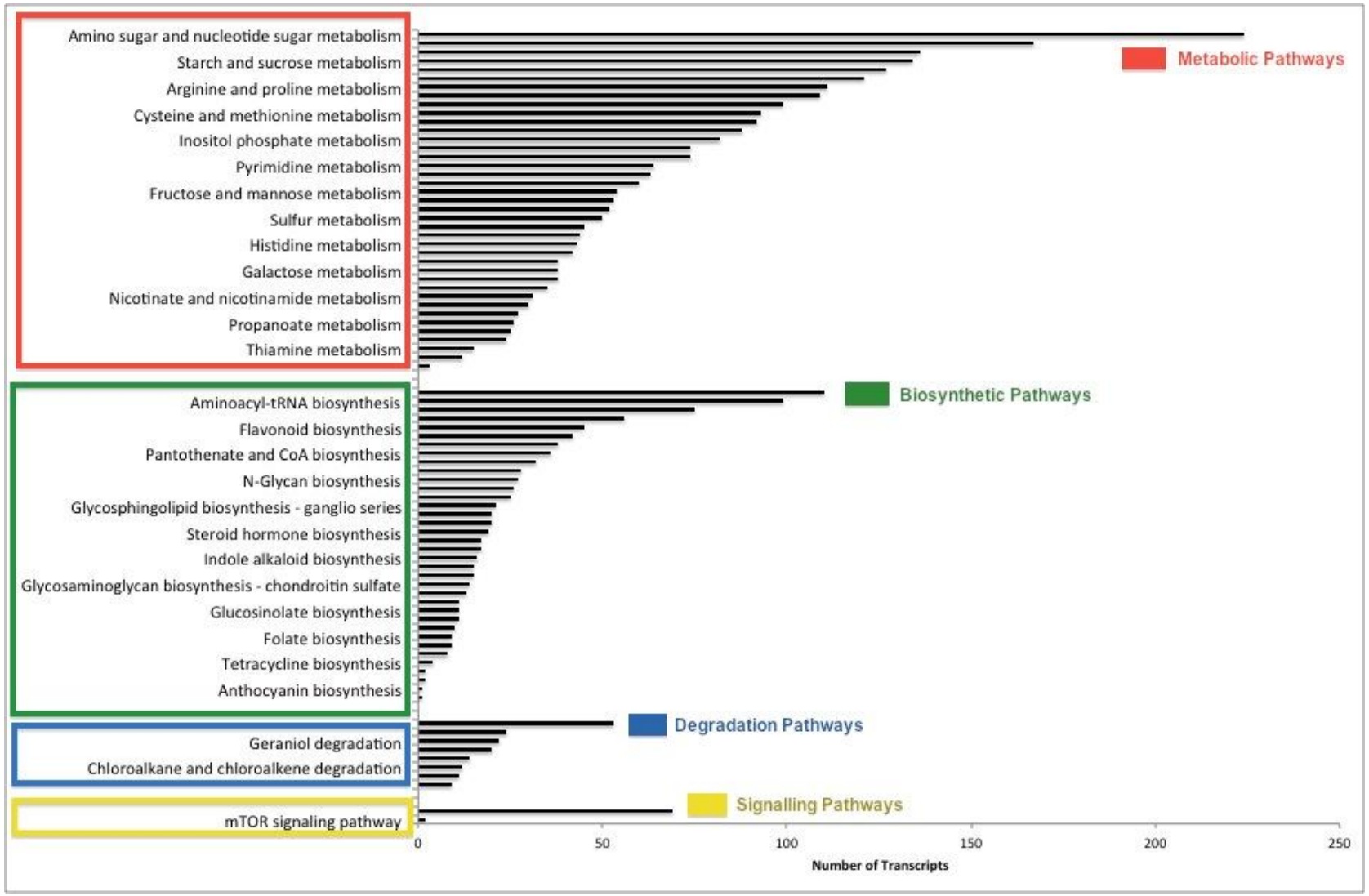


Figure 4. Venn Diagram for number of genes showing homoelogous expression bias in each accession.

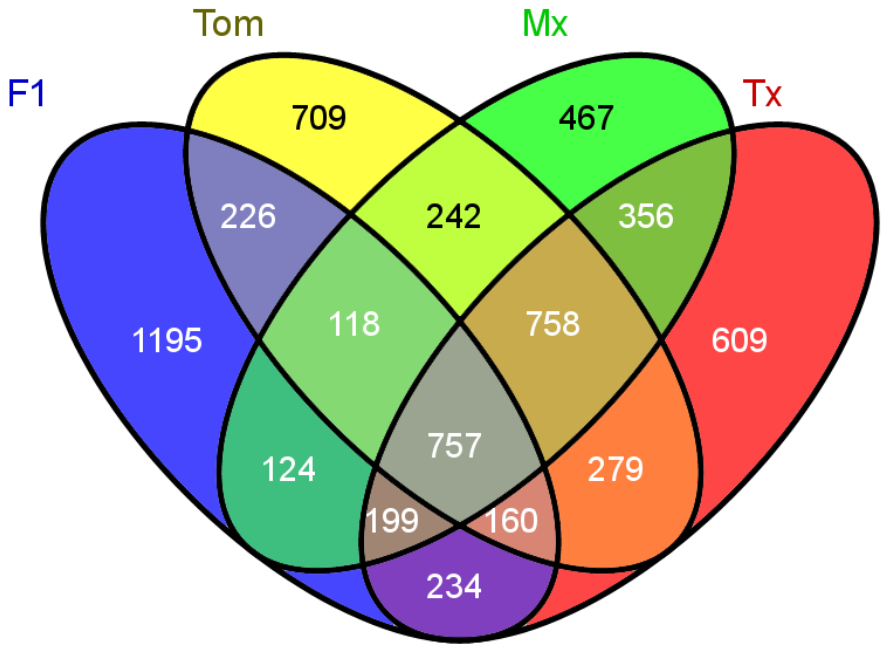


Figure 5. Phylogenetic tree based on gene expression from all the accessions where F1= diploid F_1-hybrid; Mx= *G. hirsutum* var Maxxa; Tx = *G. hirsutum* var TX2094; Tom = *G. tomentosum*, _A = A_T and _D = D_T

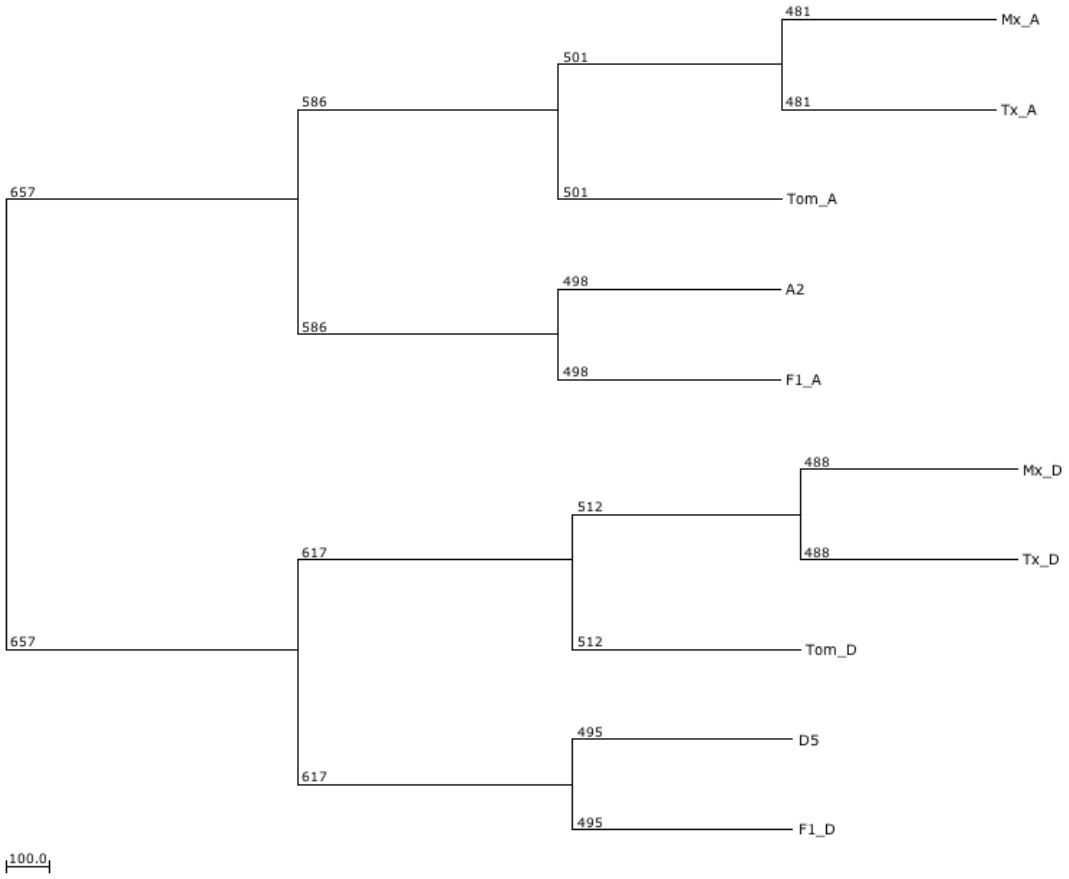
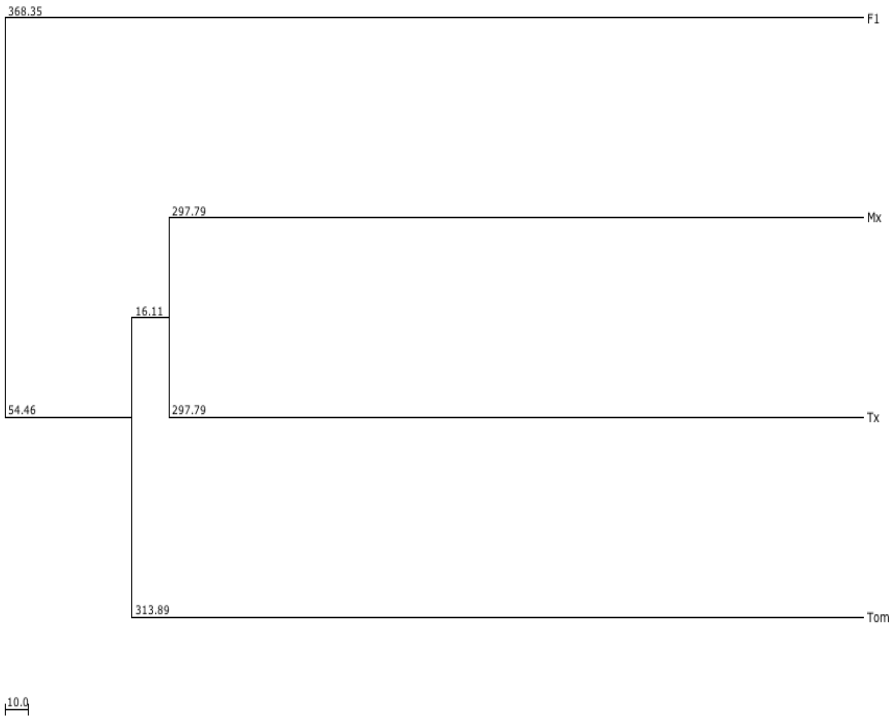


Figure 6. A phylogenetic tree based on the amount of expression divergence between homoeologous gene pairs (F1= diploid F_1-hybrid; Mx= *G. hirsutum* var Maxxa; Tx = *G. hirsutum* var TX2094; Tom = *G. tomentosum*)



CHAPTER 2

DNA Methylation in homeologous genes and its correlation with expression for Gossypium petal tissue

INTRODUCTION

Genome doubling has been used as a means for speciation by angiosperms, and most known species have undergone polyploidization (Fawcett et al. 2009; Soltis et al. 2009; DEBODT et al. 2005). Evolutionarily, several fates are possible for duplicated genes, and these possibilities have been extensively reviewed (Flagel and Wendel 2009; Osborn et al. 2003; Soltis et al. 2010; Udall and Wendel 2006). One possible fate for duplicated genes is functional divergence by accumulation of favorable (or unfavorable) point mutations. Change in phenotype can be achieved fairly quickly post-allopolyploidization without any genetic changes. When two genomes are united into a single nucleus, every gene is instantly duplicated and interaction between the regulatory machinery of the two genomes results in altered gene expression and gives rise to phenotypic diversity (Chen 2007; Rieseberg and Willis 2007; Doyle et al. 2008; Jackson and Chen 2010; Parisod et al. 2010). Hybrid vigor is another example of altered gene expression due to interaction of different parental alleles. Interactions like dominance, overdominance, and epistatis are observed between homologous loci in diploid hybrids (Schnable and Springer 2013). Similarly, interactions like homeologous expression bias and expression level dominance have been reported between homeologous loci for several allopolyploid species (Grover et al. 2012a). The molecular mechanisms behind the non-additive expression observed in diploid hybrids, diploid homoploids and allopolyploids are still under speculation.

Epigenetic regulators like DNA methylation, histone acetylation, or small RNA-mediated silencing may work in conjunction or independently to give rise to altered patterns or levels of gene expression (Henderson and Jacobsen 2007; Wang et al. 2009; Ghildiyal and Zamore 2009). Methylation at the 5_{th} carbon residue of cytosine (mC) is probably the most studied epigenetic factor and because of its prominence in eukaryotic genomes has been described as the "5^{th} base" (Lister et al. 2009). Methylated cytosine occurs in three different contexts—CG, CHG, and CHH. The regulatory role of cytosine methylation is context- and region-dependent, presumably because each context is controlled by different cellular mechanisms. CHG and CHH methylation have been reported to repress expression upstream and downstream of a gene respectively (Suzuki and Bird 2008; Li et al. 2012a). Gene body methylation is commonly found in CG context and is mostly correlated with upregulated gene expression (Cokus et al. 2008; Takuno and Gaut 2011). In recent times whole genome epigenetic profiles have been built using new technologies, including methods based on bisulfite sequencing and chromatin immunoprecipitation (ChIP) (Laird 2010; Su et al. 2011). Deviations from parental methylation patterns have been observed in several allopolyploids using the methylation-sensitive amplified polymorphism (MSAP) technique (Shaked et al. 2001; Madlung et al. 2004; Salmon et al. 2005; Lukens et al. 2006).

It is impossible to accurately detect the context and region of cytosine methylation with traditional MSAP techniques or tiling array-based methods. Though these traditional techniques have been useful in characterizing general variation in whole genome methylation patterns due to polyploidization (Paun et al. 2010). It has been deduced that most epigenetic changes due to ploidy level are generated by hybridization and not genome

doubling (VERHOEVEN et al. 2009). Chen 2010 described allopolyploids as "doubled interspecific hybrid" where heterozygosity is permanently fixed, drastic changes in epigenetic marks may be necessary for stabilization of an allopolyploid. There have been no studies to correlate methylation changes and expression deviations caused by allopolyplodization. In Arabidopsis, diploid hybrid methylation levels were found non-additive more often at loci that were differentially methylated between parents; some of these changes corresponded with expression differences (Shen et al. 2012; Greaves et al. 2012). Such correlation was absent in rice diploid hybrids, and it is possible that overall expression from alleles may be under other regulatory factors in addition to DNA methylation (Chodavarapu et al. 2012).

Heterosis is permanently fixed in allopolyploid species making them more profitable to cultivate over their diploid parents. One example is seen in *Gossypium* species. Approximately 95% of cotton fiber comes from the polyploid cotton species *Gossypium hirsutum* (AD_1, Upland cotton), with another 4% from the polyploid *Gossypium barbadense* (AD_2, Pima or Egyptian cotton) (USDA, 2012). *Gossypium* polyploid cotton species have two genomes—A_T and D_T—in their nuclei, where the 'T' subscript refers to the tetraploid nucleus. The genome content and DNA sequence of the two tetraploid genomes are closely related to the A_2 genome of *G. arboreum* and D_5 genome of *G. raimondii*, respectively (Wendel and Cronn 2003). Consequently, homologous duplicates of nearly all genes of both diploid genomes can be found on homoeologous chromosomes. The *Gossypium* genus has a well-established phylogenetic framework and is a model system to study polyploidy (Wendel et al. 2012). We utilized the phylogenetic framework of Gossypium to observe the inheritance of methylation and the effect of polyploidization on epigenetic marks.

Gene expression from duplicate genes was correlated with cytosine methylation, throwing light on the molecular mechanisms behind observed genome expression biases.

MATERIALS AND METHODS

Plant Material

Five accessions were used in our study: *G. arboreum* (2x=2n=26, A_2), *G. raimondii* (2x=2n=26, D_5), *G. tomentosum* (4x=2n=52, AD_3), *G. hirsutum* cv. Maxxa (4x=2n=52, AD_1) and a sterile diploid hybrid between A_2 and D_5 (1x = 1n = 26; F_1) (Table 1). The diploid hybrid F_1 was created by a hand pollination between reduced gametes of diploids *G. arboreum* (A_2) and *G. raimondii* (D_5), and its somatic cells only contained 13 chromosomes from each extant diploid genome.

Petal tissue was collected from plants growing under controlled greenhouse conditions at the Pohl Conservatory, Iowa State University, USA. Tissue was harvested at time of full petal expansion after dawn but before pollination from three flowers of different plants of each accession (3 biological replicates). Harvested tissue was flash frozen in liquid nitrogen and stored at -80^{o} C until RNA and DNA extraction.

RNA extractions and RNA-Seq Libraries

RNA samples were extracted from the three replicates using a modified hot borate method (ref11). RNA samples were quantified using Ribogreen (Invitrogen Inc., Grand Island, NY) and their quality was evaluated on an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA). As described by Illumina, cDNA was sheared by sonication to a 200-400 bp fragment size (Covaris Inc., Woburn, MA). RNA-Seq libraries were

prepared according to the Illumina TruSeq RNA library prep kit protocol and sequenced on an Illumina HiSeq using v.2 chemistry.

Bisulfite treatment and BS-Seq Libraries

Whole genome bisulfite-sequencing (BS-Seq) libraries were sheared and prepared using the Illumina TruSeq Library Prep kit (Illumina, San Diego, CA) with minor modifications necessitated by the bisulfite treatment. Five micrograms of genomic DNA from three reps of allotetraploids and unmethylated lambda DNA were sheared using the Covaris as per manufacturer instructions. Sheared gDNA was spiked with 1-2% of fragmented, unmethylated lambda DNA (Promega, Madison, WI). Fragmented DNA was quantified using Picogreen. Cytosine methylated adapters provided by Illumina were ligated to the blunt ends of fragmented DNA. The ligated DNA was treated with sodium bisulfite using MethylCode™ Bisulfite Conversion Kit (Invitrogen, Grand Island, NY). A size range (250-350bp) of fragments were selected on a LabChipXT (Caliper Life Sciences, Hopkinton, MA). Post-size selection samples were enriched by 16 cycles of PCR using 2.5U uracil-insensitive PfuTurboC_xHotstart polymerase (Agilent, Santa Clara, CA), 5ul 10X PfuTurbo Buffer, 0.4ul of 100nM dNTPs, 5ul of Illumina Truseq oligo mix (PCR temperatures: 95°C for 2min, then 10 cycles of 98°C for 15 sec, 60°C for 30 sec, 72°C for 4 min, end with 72°C for 10 min). The reaction products were cleaned using Agencourt AMPure XP beads (Beckman Coulter, Inc.,Brea,CA). Cleaned PCR products were enriched further using Illumina's protocol for enrichment of libraries. Libraries were validated on an Agilent Bioanalyzer and quantified using Illumina sequencing primers in a qPCR reaction to sequencing.

Sequencing

We sequenced the transcriptomes and methylomes of *G. arboreum* (A_2), *G. raimondii* (D_5), *G. hirsutum* cv. Maxxa, *G. tomentosum*, and the diploid hybrid on an Illumina HiSeq using v.2 chemistry (Table 1). We trimmed all reads with SICKLE (https://github.com/najoshi/sickle), using a phred quality threshold of 20.

Unmethylated lambda phage DNA was mixed with each library (1-2%). The trimmed lambda reads were mapped to the lambda phage reference genome sequence using GSNAP (Wu & Nacu, 2010). The number of sequenced cytosines and thymines was tallied at each cytosine position, and the percentage of unconverted cytosines was calculated as an estimate of percent methylation.

RPKM calculation from RNA reads

Diploid RNA sequencing reads were individually mapped using GSNAP to the diploid genome reference of *G. raimondii* (Wu et al. 2010; Wang et al. 2012). Reads from tetraploids were categorized into two group—A_T and D_T—using PolyCat (Page et al. 2012). Once categorized, the reads were mapped to the diploid genome reference of *G. raimondii* to assess the transcript abundance for every gene. Raw read counts were normalized to RPKM values (Table 3).

Detection of mC in Diploid Whole Genomes

The genomes of D_5 and A_2 were analyzed for DNA methylation by mapping bisulfte (BS)-treated sequence reads to their genome reference sequences. GSNAP was used to map D_5 reads to the D_5 reference sequence. A_2 reads were mapped to a reconstructed A_2 reference sequence. The reconstructed A-genome sequence was based on the A_2 consensus

sequence of 4,070,680,434 non BS-treated reads mapped to the D_5 reference (Page et al., 2012). Some base-pair positions within the D_5 references did not have any mapped A_2-genome reads. These positions could not be reconstructed and were represented as N's within the A_2 sequence. This reconstruction was necessary because an A-genome reference sequence was not available. This strategy also provided a comparative genome framework since each orthologous base of the diploid genomes had the same genome position. Since the A_2-genome is nearly double the size of the D_5-genome, the use of a reconstructed genome sequence assumed that the two genomes had co-linear gene order, though undetected (or unknown) macro chromosome rearrangements would not bias or hinder local read mapping and subsequent assessment of DNA methylation. Copy number variants between the two diploid genomes could have contributed to quantitative anomalies of read mapping results, but these were assumed to be negligible.

The number of cytosines and thymines was tallied at each cytosine position in each corresponding reference sequence. At every position with at least 5x coverage, a cytosine was called methylated (mC) if at least 75% of the mapped BS-treated reads had cytosines at that position (*i.e.*, unconverted). The cytosine was called partially methylated (pmC) if between 25% and 75% of the BS-treated reads contained cytosines at that position. The cytosine was called unmethylated if less than 25% of the BS-treated reads had a cytosine at that position. Each mC and pmC was assigned to a context, according to the corresponding reference sequence. Loci were discarded if one of the diploids did not have at least 5x coverage, or if the methylation context (CG, CHG, or CHH) differed between the 2 references.

Detection of mC in Genic Regions of Polyploid Genomes

Methylation of genic regions in A_2, D_5, *G. hirsutum*, *G. tomentosum*, and the diploid hybrid was analyzed. "Genic regions" included the annotated regions of all genes in the *G. raimondii* v2.1 draft annotation, including UTRs, exons, and introns, and an additional 1 Kbp upstream and downstream of each gene. Analysis of mC in polyploids was limited to genic regions because the genic regions were largely conserved between the A- and D-genomes (Grover et al. 2012b), and because PolyCat was better able to categorize reads where diploid RNA-seq and WGS reads could both be used to identify homoeo-SNPs between genomes (Page et al., 2012).

Initially the reads from the polyploid genomes were mapped to the D_5 reference sequence with GSNAP, using SNP-tolerant mapping to reduce the mapping efficiency bias between the A- and D-genomes (Page et al. 2012). The assemblies were then processed with PolyCat, which categorized each read mapped to a genic region according to its genome of origin (A_T or D_T), or as having an unknown genomic origin. In order to provide an unbiased comparison between diploid and tetraploid accessions, this same process was performed on reads from the A_2 and D_5 samples. mC's and pmC's were called as above. Positions of cytosines were not included in the analysis if one of the tetraploid genomes (A_T of D_T) did not have at least 3x coverage, or if the context differed between the 2 references.

Methylation percentage at all mapped cytosine positions was plotted in a sliding window across the "average" gene, separately for each context. For the upstream and downstream regions, a step size of 50 and a window width of 100 were used. For coding regions, a step size of 1/20 of the gene length and a window width of 1/10 of the gene length were used.

Phylogenetic trees were generated based on the pattern of methylated cytosines in genic regions. Essentially, all bases except cytosine were removed from sequence alignments and the cytosines were coded to represent C or mC at a particular base. The patterns of methylation were used to estimate relationships between genomes using standard phylogenetic techniques. Two trees were created for each methylation context (mC's and pmC's, respectively) using the nearest neighbor algorithm where the Euclidian distance between bit vectors represented the pattern of mC's or pmC's across all genes. Loci were not included in the analysis if the context differed between the A- and D-genomes (e.g. a homoeo-SNP adjacent to an mC could change its context from CG to CHG).

Correlation

Pearson correlation was calculated between the log base 2 RPKM value of each gene and a sliding window of percent methylation across all genes.

To analyze the relationship between expression and methylation of genome-biased genes, a 4x4 contingency table was constructed for each accession, region (upstream, gene body, and downstream), and context (CG, CHG, CHH). Each gene was categorized according to the differences in percent methylation and expression (RPKM) between the two genomes, with the more expressed genome being defined as 100% expression. 4 categories were used for each dimension: A >> D, A > D, D > A, and D >> A, where >> represents a difference greater than 50% and > represents a difference greater than 25% but less than 50%. Genes that differed by less than 25% were excluded from this analysis. Chi-squared analysis was performed on each contingency table to test for significant patterns.

RESULTS

Diploids

Whole genome methylation analysis was performed separately on the diploids A_2 and D_5 through bisulfite sequencing (BS-seq). An important consideration of BS-seq experiments is the conversion efficiency of non-methylated C -> T. A spike-in control of non-methylated lambda DNA indicated that both diploid BS-seq libraries had a conversion rate of 99.4%.

Approximately 350 million raw reads were produced for each diploid (Table 1). A greater percentage of D_5 reads mapped to the respective reference genome than A_2 reads. The lower percentage of A_2 mapped reads was likely due to A_2 regions that were not represented in the reconstructed A_2 reference. Indeed, the reconstructed A_2-genome sequence was approximately 63% the length of the assembled 749Mb D-reference genome. Consequently, it represented a much smaller percentage of the 1.7 Gb *G. arboreum* genome than the assembled reference sequence of the 0.9 Gb *G. raimondii* genome.

Since the A-genome reference sequence was reconstructed by using the D-genome reference as a template, both reference sequences shared the same coordinates across the cotton genome, so we considered the aligned nucleotides in both reference sequences to be homologous loci. At homologous loci, 81,861,614 cytosine loci were shared between the A_2 and D_5 genomes with at least 5x coverage in each genome. While these may not be all of the cytosines in the cotton genome, these were the loci that could be evaluated for methylation.

Prior to an analysis of mC context, the accuracy of genome assignment for each read was evaluated for the diploid BS-seq reads (Figure 1). This assessment provided an estimate of the pipeline accuracy rate and provided context for the assignment of polyploid

reads. The diploid read categorization in the BS-seq reads had a very low percentage of A_2 reads that were categorized as 'D' or *vice versa*. In addition, the fraction of reads categorized as 'X' (the SNP database indicated a chimeric read with different bases matching both A and D genome bases) was also low in each diploid genome. In other NGS datasets, ~50% of reads mapping to genic regions can be categorized as originating from the A- or D-genome because of the natural and uneven distribution of homoeo-SNPs. However in BS-seq data, most nucleotide transitions were fully confounded with the BS-treatment because C->T and G->A transitions were the most frequent types of homoeo-SNP between the A- and D-genomes (Page et al. 2012). Thus, only a portion of the total reads that overlapped a homoeo-SNP were possible to categorize.

A summary of methylation at the ~81M cytosine bases identified three striking differences in methylation between the A_2 and D_5 genomes (Table 2). 1) The D_5-genome had approximately three times the number of fully methylated cytosines (mC; >75%) as the A_2-genome in the CG and CHG contexts. 2) The A_2-genome had many more partially methylated cytosines (pmC; $25\% < x < 75\%$) than the D_5-genome in all contexts. However, the much higher number of CHH positions compared to CG and CHG skewed that average. If only CG and CHG contexts were considered, the A_2-genome had approximately 5 times as many pmC's as the D_5-genome. 3) The two diploid genomes had very different context distributions of mC's. In A_2, the number of mC's were nearly equivalently distributed between contexts, with mCG being slightly more frequent than either mCHG or mCHH (Table 2). In D_5, mCG context accounted for almost half of all mC's, and mCHH accounted for only a very small portion of all mC's (Figure 2A). In genic regions, the distribution of mC was more evenly distributed between contexts than it was genome-wide, where the mCG

context contributed to more than 50% of the total mC. A_2 had fewer sites of fully methylated CG and CHG than D_5 (Table 2), but many more pmC than the D_5 genome particularly in the CHH context (Figure 2B).

Polyploids

DNA methylation was also quantified for the diploid hybrid, *G. hirsutum*, and *G. tomentosum* with the same analysis used for the natural diploids A_2 and D_5. Approximately 700M reads were generated for each accession (Table 1). A spike-in control of non-methylated lambda was also included in the polyploid libraries and all libraries had about 95% or higher bisulfite conversion rate.

DNA methylation of polyploid genomes can be assessed as the sum of methylation at homoeologous loci or it can be assessed individually by genome.

A genome-wide assessment of mC within a single genome of a polyploid nucleus requires a method for attributing reads to their genome of origin. Homoeo-SNPs are single nucleotide differences between the A- and D-genomes at homoeologous positions and they can be used to categorize overlapping sequencing reads to either genome in a density dependent manner (Page et al. 2012). One indication of accurate genome assignment within the polyploid sample was the number of X reads. The number of X reads in the diploid hybrid was lower than in the natural polyploids suggesting that categorization worked properly in the polyploid samples, but that nucleotide substitution differences between the extant diploid genomes and extant polyploid genomes preclude a higher rate of categorization without additional homoeo-SNP data from the natural polyploid genomes.

Based on mapped and categorized reads, the genic regions of the A_T and D_T genomes had distinct levels of methylation in the three canonical contexts: CG, CHG, and CHH. In the

CG context, the level of full methylation was greater in the D_T-genome, but the level of partial methylation was greater in the A_T-genome. Interestingly, the full-CG methylation difference between genomes of the diploid F_1-hybrid was greater than the difference found between the diploid genomes (5.4% difference in the diploid F_1-hybrid *vs.* 4.6% difference in diploids) while the difference of full-CG methylation between the A_T and D_T-genomes was 0.6% and 0.8% in Maxxa and *G. tomentosum*, respectively. Because the level full-CG methylation in the diploid F_1-hybrid was similar to that of the diploids, polyploidization alone does not appear to reset DNA methylation between genomes. Perhaps, methylation could be the symptom or the cause of unsuccessful efforts to double its chromosome number and restore fertility.

In the CHG context, the level of full-CHG and partial-CHG methylation was consistently greater within the A-genome (A_2 or A_T) than the D-genome. These trends matched CHG levels found in the genic regions of diploid genomes. The level of full-CHG and partial-CHH methylation was greater in the D-genome (D_5 or D_T) than the A-genome (0.5% and 1.4%, respectively). Unlike the CHG context, the tendency of CHH methylation in the polyploid genomes did not match those found in diploid genomes where the A_2-genome had a greater amount of full-CHH and partial-CHH methylation than that of the D_5-genome (0.1% and 0.8% greater, respectively). These results indicated that after polyploidization the CHH methylation increased in the D-genome relative A-genome. A comparison of *G. hirsutum* and *G. tomentosum* methylation levels indicated that DNA methylation levels are consistent in all contexts between polyploid species.

A sliding window of methylation was generated for all genes (+/- 1000 bp) annotated in the D_5 reference sequence. A meta-analysis of these annotations showed that

methylation decreased dramatically in all contexts immediately upstream and downstream of coding regions, relative to intergenic regions (Figure 3). Coding regions themselves were highly methylated in CG context immediately after the translation start site, but maintained relatively low in CHG and CHH contexts. mCG was much higher in the coding sequence than mCHG or mCHH, especially in coding regions. In each context, the intergenic regions appeared to be more highly methylated than the regions immediately adjacent to the coding start site.

Phylogenetic trees were generated based on the mC's (Figure 4) and pmC's (data not shown) in the genic regions. The trees of the mC contexts were largely identical, containing a basal node that separated A- and D-genomes from each other. Within the A- or D-genome branch, the respective genomes of diploid and diploid F_1-hybrid were more closely related to each other than they were to any of the other genomes. Likewise within the A- or D-genome branch, the respective genomes of the natural polyploids (*G. hirsutum and G. tomentosum*) were also more closely related to each other than they were to any of the other genomes. Thus, the mC patterns produced a relationship that reflected the known phylogenetic relationships among the genomes and accessions, though the level of A_2 pmC was distinct from the other genomes. The trees of the pmC contexts similarly had a basal node dividing the A- and D-genomes, but the other relationships could not be reconstructed correctly.

In the CHH methylation context, the level of polyploidy had a larger impact on determining the genome relationships than the genome identity. For example, the diploid and diploid F_1-hybrid A- and D-genomes were more closely related to each other than they were to their respective genomes in the natural polyploids. In all trees, the greatest

divergences of methylation were contained in terminal branches, indicating that, while there was an observable pattern of mC's and pmC's across accessions and genomes, most of the variation was between accessions.

Comparison of genic mC between diploid and polyploid genomes

In general, the natural diploids had fewer mC's than the polyploids. The A_2 genome had the lowest methylation in every context (Figure 2). The D_5 genome was moderately more methylated than A_2 in the CG context, but the A_2 genome had a bit more methylation in the CHG and CHH contexts. Both the A_2 and D_5-genomes were less methylated than their respective orthologous genomes in the polyploids in every context. In the CHG and CHH context, the polyploid genomes were more methylated than the diploid genomes by a factor of 2 or 4 respectively (Table 2). Thus, polyploidization appears to be associated with an increased level of methylation in every context.

DNA Methylation and gene expression

In other plants, DNA methylation has been associated with transcriptional regulation of genes. Cytosine methylation plays a significant role in gene expression regulation for *Gossypium*. Except for CG methylation, mCs were negatively correlated to expression in all other contexts and regions (Table 2; Figure 5). CG methylation had significant negative correlation with expression only in upstream regions. Because CG methylation in the body was positively correlated to expression in all accessions, it may be involved with upregulation of a gene (Table 2).

A 4X4 contingency table was built to analyze relationship between genes showing significant expression and methylation biases. A chi-squared test was used to obtain

significance values. Only CHH methylation downstream of a gene had significant correlation with expression (Table 4). Upstream and gene body methylation significantly regulate expression but do not cause expression biases between genomes. It can be concluded that the repressive action of downstream CHH methylation is only important epigenetic factor behind expression biases.

DISCUSSION

G. ramondii epigenome

Reference mapping of bisulfite converted reads made it possible to look at cytosine methylation with single base pair resolution. 16% of all cytosines in the *G. raimondii* genome are methylated in petal tissue. This is 3 times higher Arabidopsis flower buds and 1.5 times lower than rice panicles (Lister et al. 2008; Li et al. 2012b). Methylation was much higher in the genome as a whole than in genic regions alone. Around 57 % of *G. raimondii* genome is transposable elements and such repetitive regions are usually highly methylated in all sequence contexts (Wang et al. 2012; Cokus et al. 2008). Also, Gossypium has very small chromosomes; therefore, a relatively higher portion of the genome lies in pericentromeric regions that are enriched for methylation (Lister et al. 2008; Li et al. 2012b) . CG sites were methylated more often than CHG sites, and the CHH sites were only sparingly methylated. Hypermethylation of CG sites and hypomethylation of CHH sites have also been reported in Arabidopsis and rice. CG methylation is most prevalent in the genome followed by CHG and then CHH. Though the proportion of CG methylation is comparable to Arabidopsis, the methylation in CHG context is much higher and CHH methylation is much lower. CHG methylation usually accumulates in pericentromeric regions and TE elements these regions are found more abundantly in Gossypium genome than Arabidopsis (Wang et

al. 2012; Cokus et al. 2008; Lister et al. 2008; Kato et al. 2003). The coding regions are only found associated with methylation in CG context (Zhang et al. 2006; Zilberman et al. 2006; Cokus et al. 2008; Lister et al. 2008). Therefore, CG methylation comprised a larger percentage of the methylation in genic regions (92%) than in the genome as a whole (58%).

Relative Methylation of the A- and D-genomes

The letters A through G plus K are used to denote the genomes of 45 diploid species of genus *Gossypium*. Chromosome number for all the diploid species is the same (n=13), but they have wide range of genome sizes from 2500 Mb in the K genome to 900 Mb in the D genome (Wendel and Cronn 2003) All the diploid species in *Gossypium* have retained collinear gene order and by mapping A genome BS treated reads to D genome reference we could compare methylation at the common loci in both genomes (Brubaker et al. 1999) In whole genome and genic analysis, and in diploids and polyploids, the D-genome had more fully methylated sites than the A-genome, while the A-genome had more partially methylated sites than the D-genome (Figure 2; Table 2). Partial methylation is less common in the CG context of the D genome, as it is in Arabidopsis (Lister et al. 2008). The A genome is more similar to rice which is not as heavily methylated. On average about 44% of cytosines were found methylated at a CG site in rice.

The difference between pmCs and mCs between the A and D genomes is more pronounced in non-genic regions. The two diploid species diverged 5-10 million years ago from a common ancestor but still share some common transposable elements (Grover et al. 2004). The D genome is half the physical size of the A genome in diploid as well as polyploid genomes. Much of the inflated size of the A genome is due to repetitive elements

(Desai et al. 2006). Activation of dormant TE through demethylation might have led to inflation and A genome reorganization. Multiple copies of these TE in the A genome could be differentially methylated giving rise to a higher percentage of pmCs in non-genic regions compared to D genome (Figure 2).

Effect of Polyploidization on Methylation

Polyploidization took place 1-2 million years ago in *Gossypium* sps (Wendel and Cronn 2003).In our study design we included a synthetic diploid hybrid (F1) made by artificial hybridization of diploid species most closely related to diploid progenitors of natural polyploids. A comparison of methylation marks in diploid parents and the diploid hybrid enabled us to study the changes introduced immediately after genomic merger. The changes in methylation landscape are necessary for stabilization after the 'genomic shock' of a hybridization event (Chen 2007). We noted increased methylation level in genic regions for both sub-genomes (A_T and D_T) of the hybrid, indicating that additional methylation marks are put on the parent genomes when they merge in a single nucleus. Increase in methylation levels have also been reported from synthetic polyploids of Arabidopsis (Madlung et al. 2004; Beaulieu et al. 2009), brassica (Xu et al. 2008), wheat (Shaked et al. 2001) and dandelion (VERHOEVEN et al. 2009).Present day natural polyploid species had 1-2 millions years to reset and add on to the methylation marks that allopolyplodization introduced in their common ancestor. Both natural polyploids have higher methylation levels than the diploid hybrid, and *G. hirsutum* has lower methylation than *G. tomentosum*. Such difference could arise because these two natural polyploids evolved in very different ecotypes. *G. tomentosum* is endemic to the Hawaiian islands whereas *G. hirsutum* is a domesticated species (DeJoode and Wendel 1992). Environmental

differences caused notable divergence in the methylation profiles of three sister allopolyploid taxa of orchids and selection pressure plays important role in designing methylation landscape of a species (Paun et al. 2010).

Inheritance of Methylated Sites

Polyploidization facilitates speciation by setting up a new genomic and methylation landscape for selection to act upon (Doyle et al. 2008). Epigenetic factors are stably inherited and are more susceptible to changes in environment than genetic factors (Robertson and Wolf 2012). Epigenetic variation is introduced in every generation and is present even in genetically identical lines (Johannes et al. 2008). This variation may occur as a result of errors in maintenance of methylation in genomes like random mutations or it can be introduced by environment and stabilized under selective pressures. Epigenetic variation is not completely independent of genetic variation. There are many genetic factors that can significantly influence epigenetic marks, including transposable elements (Furner and Matzke 2011), small RNA production (Zhai et al. 2008) and the genes responsible for methylation maintenance enzymes or histone modifications (Cokus et al. 2008). In cotton, methylation patterns are highly conserved between related individuals. Even with a relatively small portion of the genome to analyze, the phylogenetic relationship of individuals and genomes could be predicted by epigenetic variation between them. The signal was strong enough to persist through the increase in overall methylation incidental to polyploidization. The sub-genomes of tetraploid cotton retain the methylation characteristics of their ancestral diploids. The phylogenetic tree based on cytosine methylation of genic regions has the same topology as the gene tree for these accessions , with the exception of the mCHH-based tree. Discrepancies in the CHH context methylation

tree have also been reported in rice (Li et al. 2012b). Evolutionary forces molded cytosine methylation marks in similar way in *Gossypium* and rice species.

Interactions between methylation and gene expression

Numerous studies have established 5-methyl-cytosine as a cue for repression of expression (Cedar and Bergman 2012). It is thought that methylation serves as a host-defense system preventing rampant transposition of TE or retroelements (Hirochika et al. 2000). Differential methylation between endosperm and embryo suggest that this defensive system has subsequently been adapted by plants for imprinting genes critically important during seed development (Gehring et al. 2009). For a long time it was assumed that methylation was mostly confined to endosperm due to early investigations of imprinting. However, it is now recognized that DNA methylation may play a larger role in development and routine regulation of gene expression. Whole genome sequencing of sodium bisulfite treated samples (BS-seq) in Arabidopsis showed that hypermethylation in promoter region is negatively correlated to expression (Lister et al. 2008; Cokus et al. 2008).We observed the same relation in Gossypium accessions: methylation upstream of a gene was found to repress expression regardless of context (Figure 5). A significant negative correlation with expression was observed for nonCG methylation downstream of a gene. Li et al. observed in a recent study on rice that CHH methylation downstream of a gene was significantly associated with lower expression. They also revisited Arabidopsis methylation data and found the same correlation. The role of methylation in a CG context is more varied. It was found negatively correlated with expression upstream of a gene but positively correlated in gene body. Its role downstream is ambiguous. Gene body methylation is found almost exclusively in CG context in Arabidopsis and has been found to

be associated with functionally important genes (Takuno and Gaut 2011). It has been proposed that CG body methylation may be necessary for accurate splicing or may restrict leaky expression (Zilberman and Henikoff 2007; Maunakea et al. 2010).

Correlation analyses between expression and cytosine methylation for all the genes confirmed the regulatory role of this epigenetic mark (Table 3, Figure 5). Whether expression regulation by 5-methyl-cytosine results in expression level dominance and homoelogous expression biases observed in polyploids still needs to be determined. The first instance of correlation of DNA methylation and histone modification with expression dominance was noted in an allopolyploid of Arabidopsis (Chen and Pikaard 1997). We built a 4x4 contingency table to compare methylation and expression levels between homoeologous genes (Table 3). Only methylation in CHH context downstream seems to have significant impact on polyploid expression. The CHH methylation tree does not conform to known relationships of these species, indicating that these marks change significantly post merger (Figure 10 C). It was previously believed that repression through methylation was caused by simple prevention of binding of transcription factors to the promoters, but it is now clear that methylation regulates expression through more complex mechanisms that involve interactions with histones (Okitsu and Hsieh 2007; Cedar and Bergman 2012). Chromatin remodeling makes genic regions unavailable for transcription, which is a more efficient way of achieving repression at a genome-wide scale. An epigenetic cue like CHH methylation downstream of a gene could interact with histones to alter the chromatin folding and result in genome wide repression of duplicate genes.

REFERENCES

Beaulieu J, Jean M, Belzile F. 2009. The allotetraploid Arabidopsis thaliana–Arabidopsis lyrata subsp. petraea as an alternative model system for the study of polyploidy in plants. *Molecular Genetics and Genomics* **281**: 421–435.

Brubaker CL, Paterson AH, Wendel JF. 1999. Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. *Genome* **42**: 184–203.

Cedar H, Bergman Y. 2012. Programming of DNA methylation patterns. *Annual Review of Biochemistry* **81**: 97–117.

Chen ZJ. 2007. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu Rev Plant Biol* **58**: 377.

Chen ZJ, Pikaard CS. 1997. Epigenetic silencing of RNA polymerase I transcription: a role for DNA methylation and histone modification in nucleolar dominance. *Genes & development* **11**: 2124–2136.

Chodavarapu RK, Feng S, Ding B, Simon SA, Lopez D, Jia Y, Wang GL, Meyers BC, Jacobsen SE, Pellegrini M. 2012. Transcriptome and methylome interactions in rice hybrids. *Proceedings of the National Academy of Sciences* **109**: 12040–12045.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**: 215–219.

DEBODT S, MAERE S, VANDEPEER Y. 2005. Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution* **20**: 591–597.

DeJoode DR, Wendel JF. 1992. Genetic diversity and origin of the Hawaiian Islands cotton, Gossypium tomentosum. *American Journal of Botany* 1311–1319.

Desai A, Chee PW, Rong J, May OL, Paterson AH. 2006. Chromosome structural changes in diploid and tetraploid A genomes of Gossypium. *Genome* **49**: 336–345.

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF. 2008. Evolutionary Genetics of Genome Merger and Doubling in Plants. *Annu Rev Genet* **42**: 443–461.

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proceedings of the National Academy of Sciences* **106**: 5737–5742.

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. *New Phytologist* **183**: 557–564.

Furner IJ, Matzke M. 2011. Methylation and demethylation of the Arabidopsis genome. *Current Opinion in Plant Biology* **14**: 137–141.

Gehring M, Bubb KL, Henikoff S. 2009. Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science* **324**: 1447–1451.

Ghildiyal M, Zamore PD. 2009. Small silencing RNAs: an expanding universe. *Nature Reviews Genetics* **10**: 94–108.

Greaves IK, Groszmann M, Ying H, Taylor JM, Peacock WJ, Dennis ES. 2012. Trans chromosomal methylation in Arabidopsis hybrids. *Proceedings of the National Academy of Sciences* **109**: 3570–3575.

Grover CE, Gallagher JP, Szadkowski EP, Yoo MJ, Flagel LE, Wendel JF. 2012a. Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist*.

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

Grover CE, Kim HR, Wing RA, Paterson AH, Wendel JF. 2004. Incongruent patterns of local and global genome size evolution in cotton. *Genome Research* **14**: 1474–1482.

Henderson IR, Jacobsen SE. 2007. Epigenetic inheritance in plants. *Nature* **447**: 418–424.

Hirochika H, Okamoto H, Kakutani T. 2000. Silencing of retrotransposons in Arabidopsis and reactivation by the ddm1 mutation. *THE PLANT CELL ONLINE* **12**: 357–368.

Jackson S, Chen ZJ. 2010. Genomic and expression plasticity of polyploidy. *Current Opinion in Plant Biology* **13**: 153–159.

Johannes F, Colot V, Jansen RC. 2008. Epigenome dynamics: a quantitative genetics perspective. *Nature Reviews Genetics* **9**: 883–890.

Kato M, Miura A, Bender J, Jacobsen SE, Kakutani T. 2003. Role of CG and Non-CG Methylation in Immobilization of Transposons in Arabidopsis. *Current Biology* **13**: 421–426.

Laird PW. 2010. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics* **11**: 191–203.

Li X, Zhu J, Hu F, Ge S, Ye M, Xiang H, Zhang G, Zheng X, Zhang H, Zhang S, et al. 2012a. Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics* **13**: 300.

Li X, Zhu J, Hu F, Ge S, Ye M, Xiang H, Zhang G, Zheng X, Zhang H, Zhang S, et al. 2012b. Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics* **13**: 300.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133**: 523–536.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**: 315–322.

Lukens LN, Pires JC, Leon E, Vogelzang R, Oslach L, Osborn T. 2006. Patterns of sequence loss and cytosine methylation within a population of newly resynthesized Brassica napus allopolyploids. *PLANT PHYSIOLOGY* **140**: 336–348.

Madlung A, Tyagi AP, Watson B, Jiang H, Kagochi T, Doerge RW, Martienssen R, Comai L. 2004. Genomic changes in synthetic Arabidopsis polyploids. *The Plant Journal* **41**: 221–230.

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**: 253–257.

Okitsu CY, Hsieh CL. 2007. DNA methylation dictates histone H3K4 methylation. *Molecular and cellular biology* **27**: 2746–2757.

Osborn TC, Pires C, Birchler JA, Auger DL, Jeffery CZ, Lee HS, Comai L, Madlung A, Doerge RW, Colot V. 2003. Understanding mechanisms of novel gene expression in polyploids. *Trends in Genetics* **19**: 141–147.

Parisod C, Holderegger R, Brochmann C. 2010. Evolutionary consequences of autopolyploidy. *New Phytologist* **186**: 5–17.

Paterson AH. 2010. Cotton. In *Biotechnology in Agriculture and Forestry*, Vol. 65 of, pp. 45–63, Springer Berlin Heidelberg.

Paun O, Bateman RM, Fay MF, Hedren M, Civeyrel L, Chase MW. 2010. Stable Epigenetic Effects Impact Adaptation in Allopolyploid Orchids (Dactylorhiza: Orchidaceae). *Molecular Biology and Evolution* **27**: 2465–2473.

Rieseberg LH, Willis JH. 2007. Plant speciation. *Science* **317**: 910–914.

Robertson AL, Wolf DE. 2012. The role of epigenetics in plant adaptation. *Trends in Evolutionary Biology* **4**: e4.

Salmon A, Ainouche ML, Wendel JF. 2005. Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). *Molecular Ecology* **14**: 1163–

1175.

Schnable PS, Springer NM. 2013. Progress Toward Understanding Heterosis in Crop Plants. *Annu Rev Plant Biol* **64**: null.

Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA. 2001. Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *THE PLANT CELL ONLINE* **13**: 1749–1759.

Shen H, He H, Li J, Chen W, Wang X, Guo L, Peng Z, He G, Zhong S, Qi Y, et al. 2012. Genome-Wide Analysis of DNA Methylation and Gene Expression Changes in Two Arabidopsis Ecotypes and Their Reciprocal Hybrids. *The Plant Cell* **24**: 875–892.

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

Soltis DE, Buggs RJA, Doyle JJ, Soltis PS. 2010. What we still don't know about polyploidy. *Taxon* **59**: 1387–1403.

Su Z, Han L, Zhao Z. 2011. Conservation and divergence of DNA methylation in eukaryotes: New insights from single base-resolution DNA methylomes. *epigenetics* **6**: 134–140.

Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics* **9**: 465–476.

Takuno S, Gaut BS. 2011. Body-Methylated Genes in Arabidopsis thaliana Are Functionally Important and Evolve Slowly. *Molecular Biology and Evolution* **29**: 219–227.

Udall JA, Wendel JF. 2006. Polyploidy and Crop Improvement. *Crop Science* **46**: S–3.

VERHOEVEN KJ, van DIJK PJ, BIERE A. 2009. Changes in genomic methylation patterns during the formation of triploid asexual dandelion lineages. *Molecular Ecology* **19**: 315–324.

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S. 2012. The draft genome of a diploid cotton Gossypium raimondii. *Nature Genetics*.

Wang X, Elling AA, Li X, Li N, Peng Z, He G, Sun H, Qi Y, Liu XS, Deng XW. 2009. Genome-Wide and Organ-Specific Landscapes of Epigenetic Modifications and Their Relationships to mRNA and Small RNA Transcriptomes in Maize. *The Plant Cell* **21**: 1053–1069.

Wendel JF, Cronn RC. 2003. Polyploidy and the evolutionary history of cotton. *Advances in Agronomy* **78**: 139–186.

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

Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**: 873–881. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20147302&retmode=ref&cmd=prlinks.

Xu Y, Zhong L, Wu X, Fang X, Wang J. 2008. Rapid alterations of gene expression and cytosine methylation in newly synthesized Brassica napus allopolyploids. *Planta* **229**:

471–483.

Zhai J, Liu J, Liu B, Li P, Meyers BC, Chen X, Cao X. 2008. Small RNA-directed epigenetic natural variation in Arabidopsis thaliana. *PLoS Genet* **4**: e1000056.

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SWL, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE. 2006. Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis. *Cell* **126**: 1189–1201.

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2006. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics* **39**: 61–69.

Zilberman D, Henikoff S. 2007. Genome-wide analysis of DNA methylation patterns. *Development* **134**: 3959–3965.

TABLES

Table 5. Sequencing results for each accession, with the total number of reads after trimming, the number of reads mapped to the D_5 reference and to the lambda phage sequence, and the bisulfite conversion rate.

Accession	Reads	Cotton	Lambda	Conversion
G. arboreum	332,107,534	257,094,061	1,724,455	99.4%
G. raimondii	365,021,156	323,291,802	3,062,929	99.4%
F_1-hybrid	760,163,066	589,574,013	6,770,503	99.4%
G. hirsutum	701,274,772	512,076,942	4,621,456	99.4%
G. tomentosum	714,789,765	529,919,156	10,894,289	94.7%

Table 6. Methylation in each context, with the total number of sites analyzed in each context and the percentage of those sites with at least 75% methylation (mC's) and between 25% and 75% methylation (pmC's), for the whole genome analysis of A2 and D5 (WGS) and the genic analysis of polyCat-categorized reads for all genomes of A_2, D_5, F_1 hybrid, *G. hirsutum* (Mx), and *G. tomentosum* (Tom).

	Accession	mCG %	pmCG %	CG	mCHG %	pmCHG %	CHG	mCHH %	pmCHH %	CHH
WGS	A_2	34.9	30.4	8,778,778	20.9	28.3	10,743,449	2.4	18.4	75,007,136
	D_5	78.3	3.2	24,674,734	55.5	11.0	26,857,081	2.1	14.6	176,872,528
Genic	A_2	19.1	3.3	1,440,464	2.2	2.6	2,361,309	0.5	2.7	10,939,280
	D_5	23.7	2.2	2,208,486	1.7	1.2	3,607,410	0.4	1.9	17,539,342
	F_1-A_t	21.4	2.0	758,360	4.1	1.8	1,205,415	2.0	1.8	5,764,140
	F_1-D_t	26.8	1.7	818,401	3.9	1.2	1,272,218	2.9	2.2	6,429,987
	Mx-A_t	26.2	2.7	614,904	4.7	1.9	940,802	2.3	2.0	4,141,206
	Mx-D_t	26.9	2.4	582,779	4.5	1.6	878,835	3.0	2.3	4,019,990
	Tom-A_t	27.6	2.5	579,732	5.5	1.6	883,752	3.1	1.8	3,991,663
	Tom-D_t	28.4	2.1	552,013	5.3	1.3	826,582	4.1	2.0	3,878,097

Table 7. Significance values for correlation between expression and methylation in different contexts/regions

Accession	Region	Context	Pearson Correlation Coefficient	P value for Pearson Correlation Coefficient
Dipploid Hybrid F_1	UP	CG	-0.149	0
		CHG	-0.218	0
		CHH	-0.018	0.004
	BODY	CG	0.058	0
		CHG	-0.281	0
		CHH	-0.137	0
	DOWN	CG	-0.055	0
		CHG	-0.276	0
		CHH	-0.085	0
G. hirsutum Maxxa	UP	CG	-0.162	0
		CHG	-0.2	0
		CHH	-0.025	0.001
	BODY	CG	0.051	0
		CHG	-0.269	0
		CHH	-0.124	0
	DOWN	CG	0.069	0
		CHG	-0.218	0
		CHH	-0.074	0
G. tomentosum	UP	CG	-0.141	0
		CHG	-0.206	0
		CHH	-0.022	0.003
	BODY	CG	0.06	0
		CHG	-0.242	0
		CHH	-0.117	0
	DOWN	CG	-0.002	0.814
		CHG	-0.207	0
		CHH	-0.073	0

Table 8. Chi Square test significance values from the Contingency table comparing methylation differences in significantly biased homeologs.

Accessions	CG			CHG			CHH		
	up	body	down	up	body	down	up	body	down
Diploid Hybrid F_1	0.28	0.67	0.38	0.43	0.71	0.82	0.06	0.49	0
G. hirsutum Maxxa	0.26	0.24	0.36	0.65	0.39	0.07	0.43	0.37	0
G. tomentosum	0.42	0.52	0.6	0.85	0.89	0.59	0.41	0.77	0

FIGURES

Figure 7. PolyCat results for A_2, D_5, F_1 hybrid (F_1), *G. hirsutum* (Mx), and *G. tomentosum* (Tom). Reads are categorized as A-genome (A), D-genome (D), chimeric (X), or uncategorizable (N).

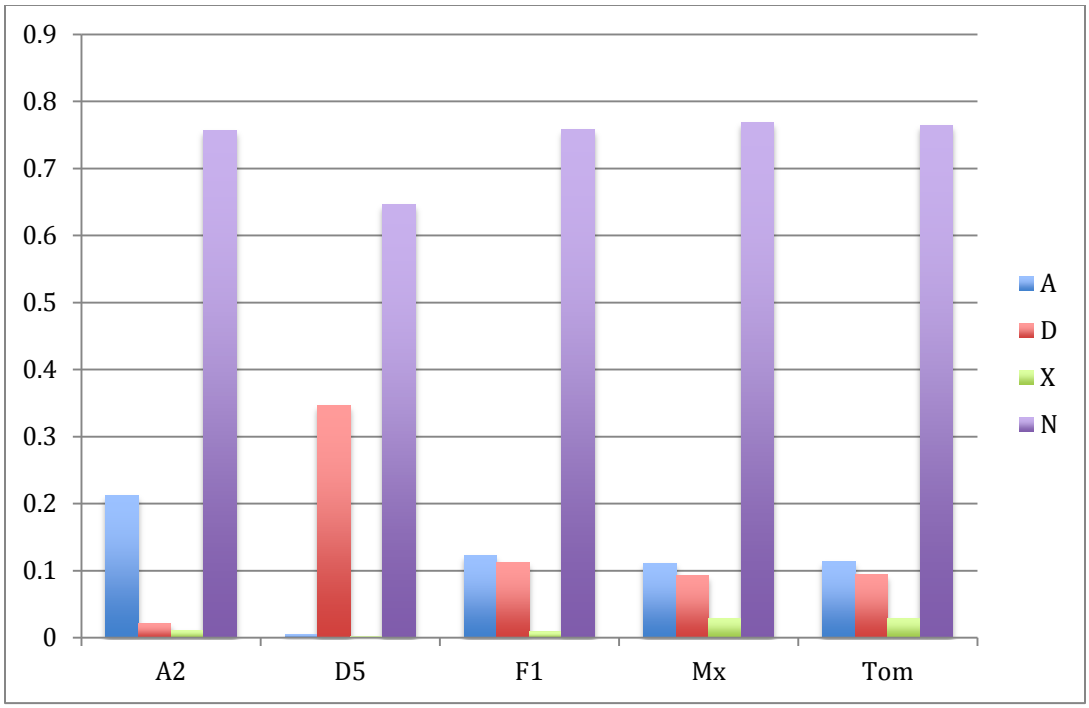
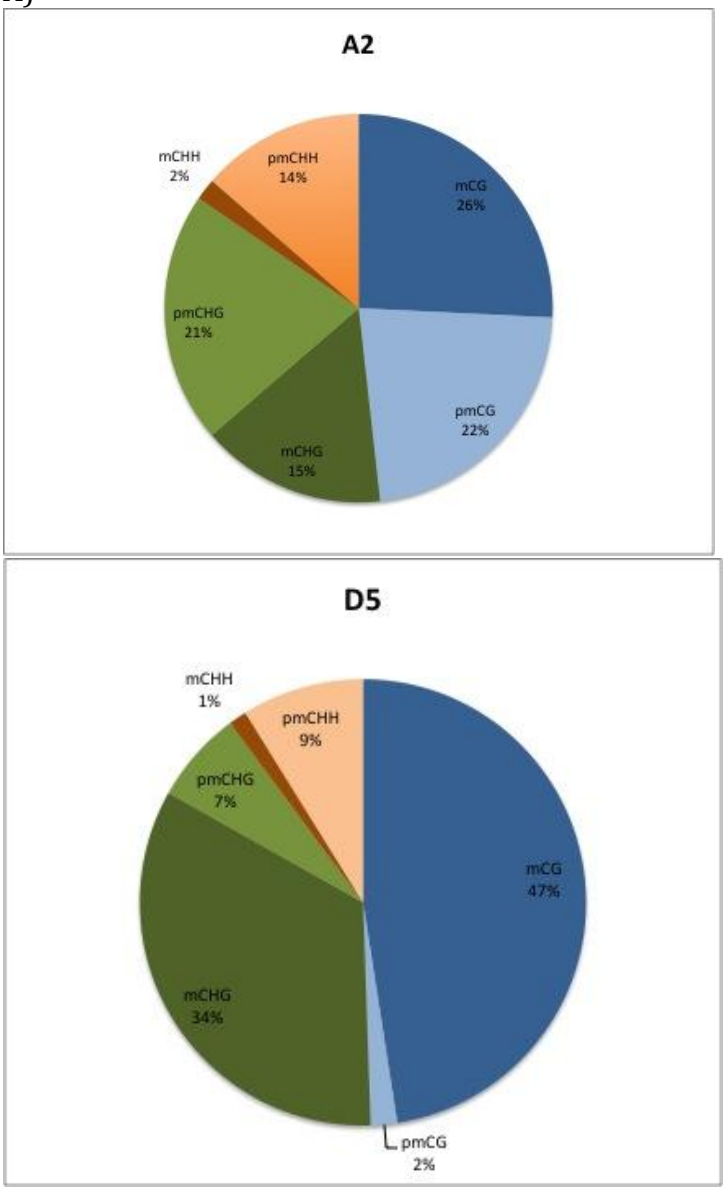


Figure 8. Distribution of methylated and partially methylated cytosines in the three contexts - CG, CHG, and CHH. A)Relative proportions of methylated (mC) and partially methylated (pmC) cytosines in the whole diploid genomes of *G. arboreum* (A_2)and *G. raimondii* (D_5). B) Context percentage of methylated (mC) and partially methylated cytosines (pmC) in genic regions of two diploid accessions - *G. raimondii* (D5); *G. arboreum* (A2); two polyploid accessions - *G. hirsutum* (Mx), G. tomentosum (Tom) and one diploid synthetic hybrid (F1).

A)



B)

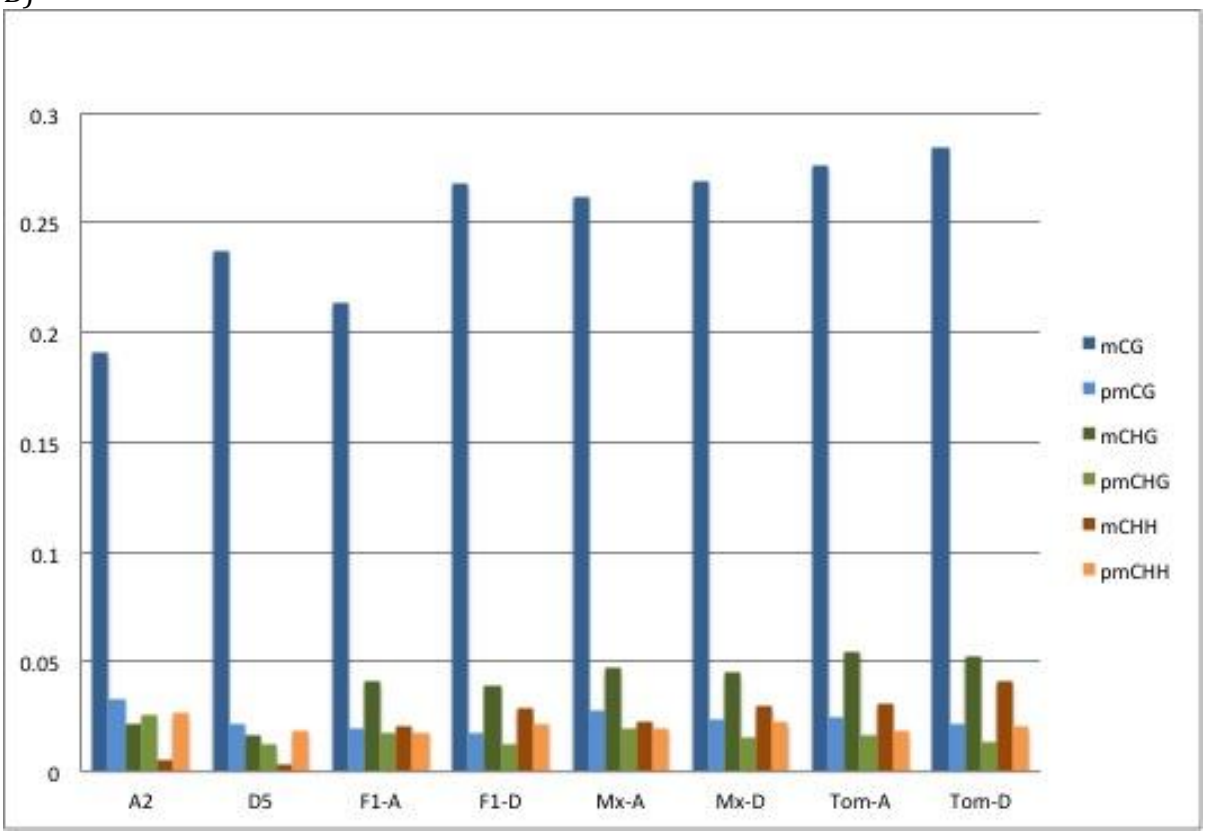
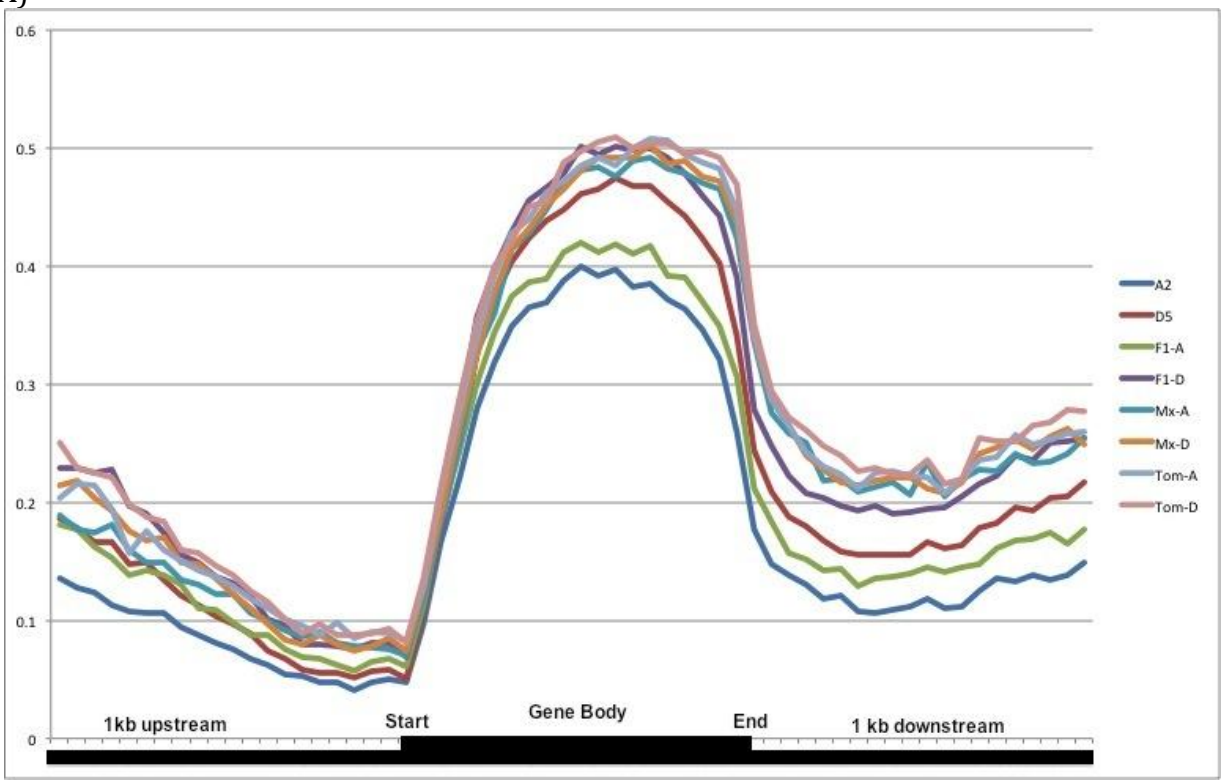
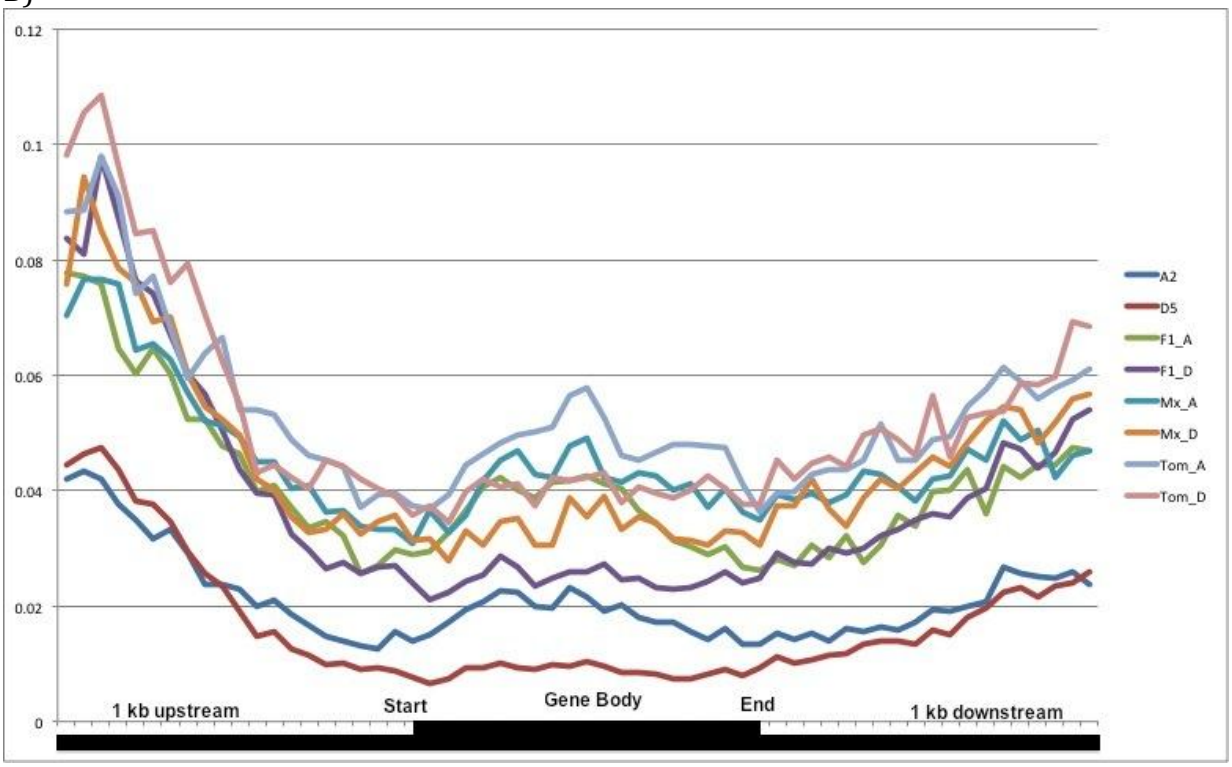

0.3
0.25
0.2
0.15
0.1
0.05
0
A2
D5
F1-A
F1-D
Mx-A
Mx-D
Tom-A
Tom-D
mCG
pmCG
mCHG
pmCHG
mCHH
pmCHH

Figure 9. Average methylation for each context in a sliding window across all genes. The length of each gene was adjusted to allow levels of methylation to be comparable across genes. A) Methylation in the CG context B) methylation in the CHG context and C) Methylation in the CHH context.

A)



B)


0.12
0.1
0.08
0.06
0.04
0.02
0
1 kb upstream
Start
Gene Body
End
1 kb downstream
A2
D5
F1_A
F1_D
Mx_A
Mx_D
Tom_A
Tom_D

C)

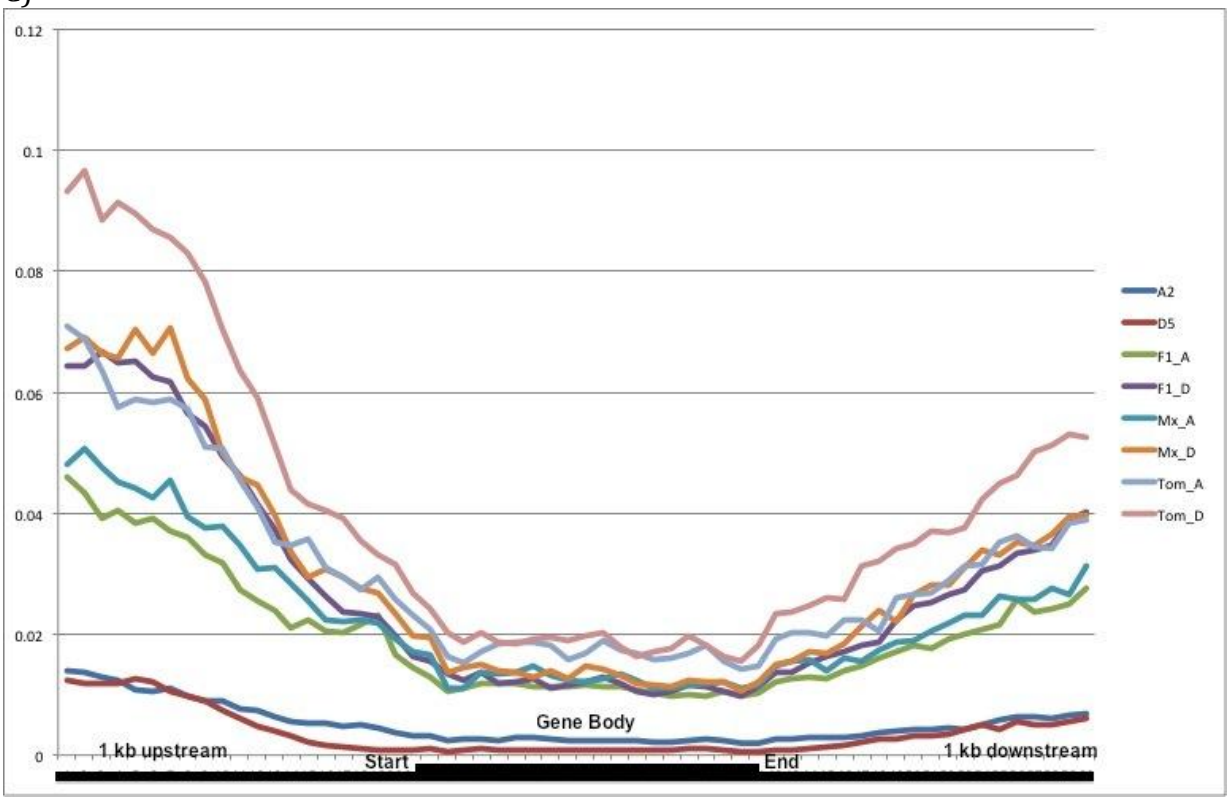
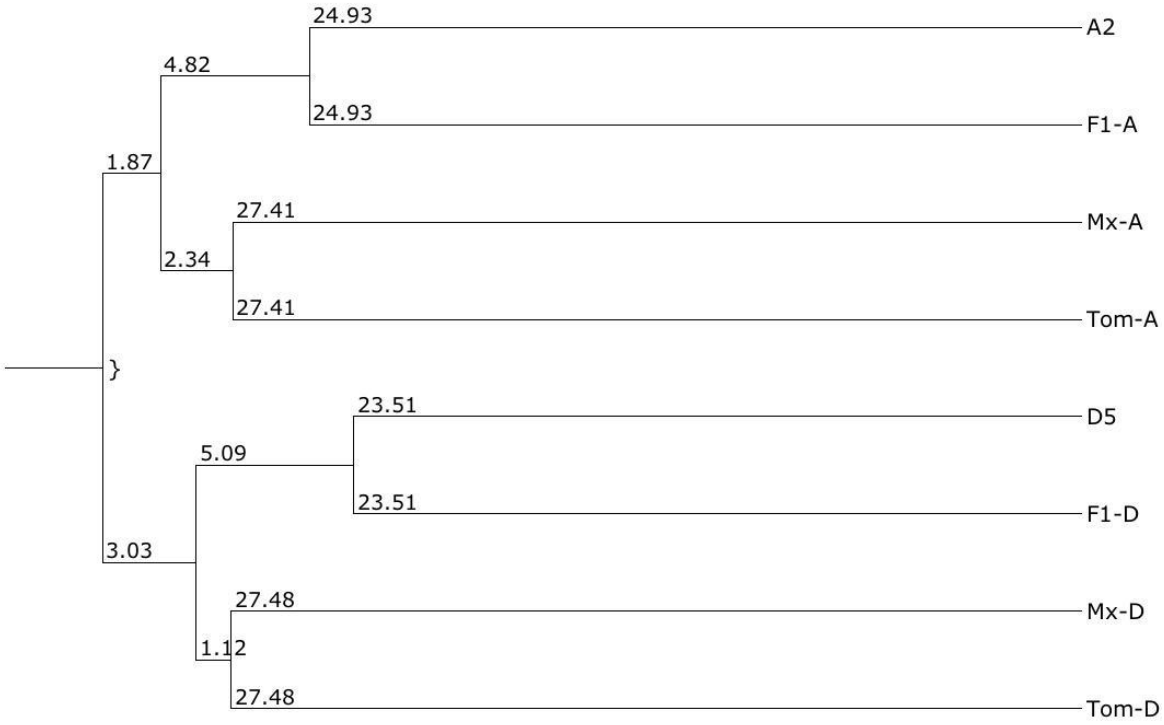
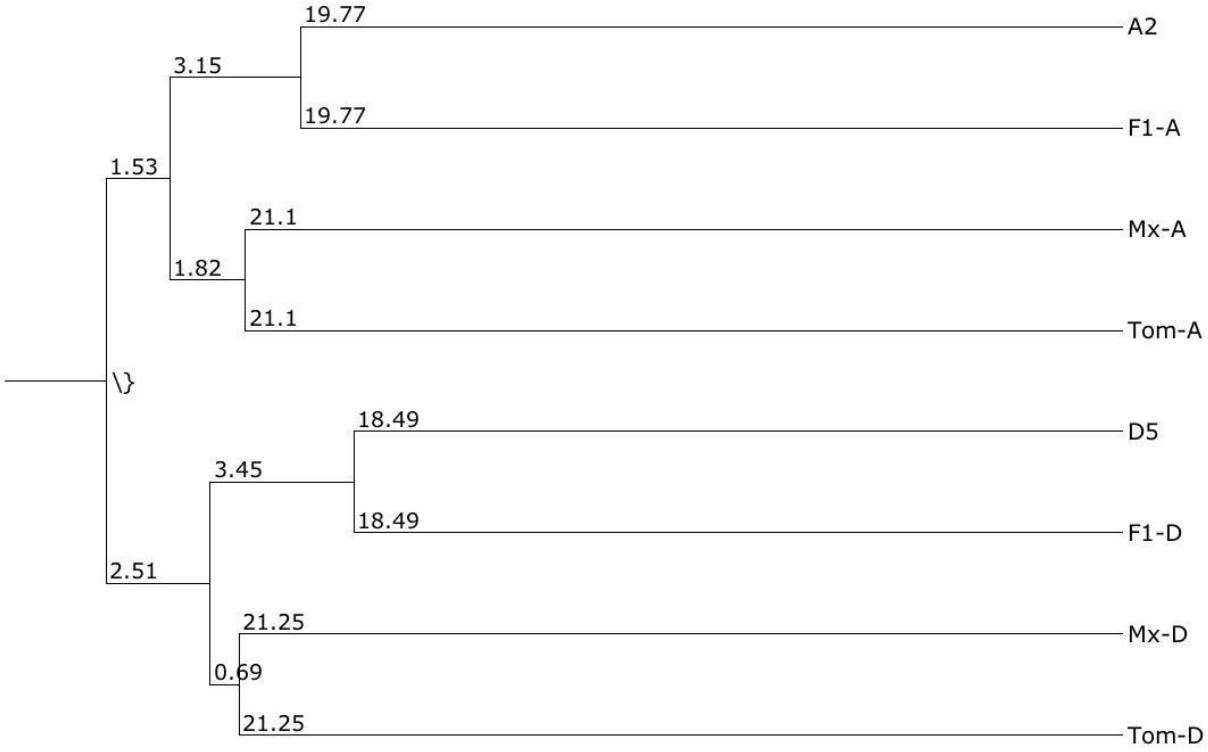

0.12
0.1
0.08
0.06
0.04
0.02
0
A2
D5
F1_A
F1_D
Mx_A
Mx_D
Tom_A
Tom_D
Gene Body
1 kb upstream
Start
End
1 kb downstream

Figure 10. Dendograms based on patterns of mC's in each methylation context for the genic regions of A_2, D_5, F_1 diploid hybrid (F_1-A and F_1-D), *G. hirsutum* (Mx-A and Mx-D), and *G. tomentosum* (Tom-A and Tom-D). The numerical numbers at the nodes are the branch lengths of the Euclidean distance between bit vectors representing the patterns of each genome of each accession for A) methylation in the CG context B) methylation in the CHG conext and C) Methylation in the CHH context.

A)



B)


19.77
A2
3.15
19.77
F1-A
1.53
21.1
Mx-A
1.82
21.1
Tom-A
\}
18.49
D5
3.45
18.49
F1-D
2.51
21.25
Mx-D
0.69
21.25
Tom-D

C)

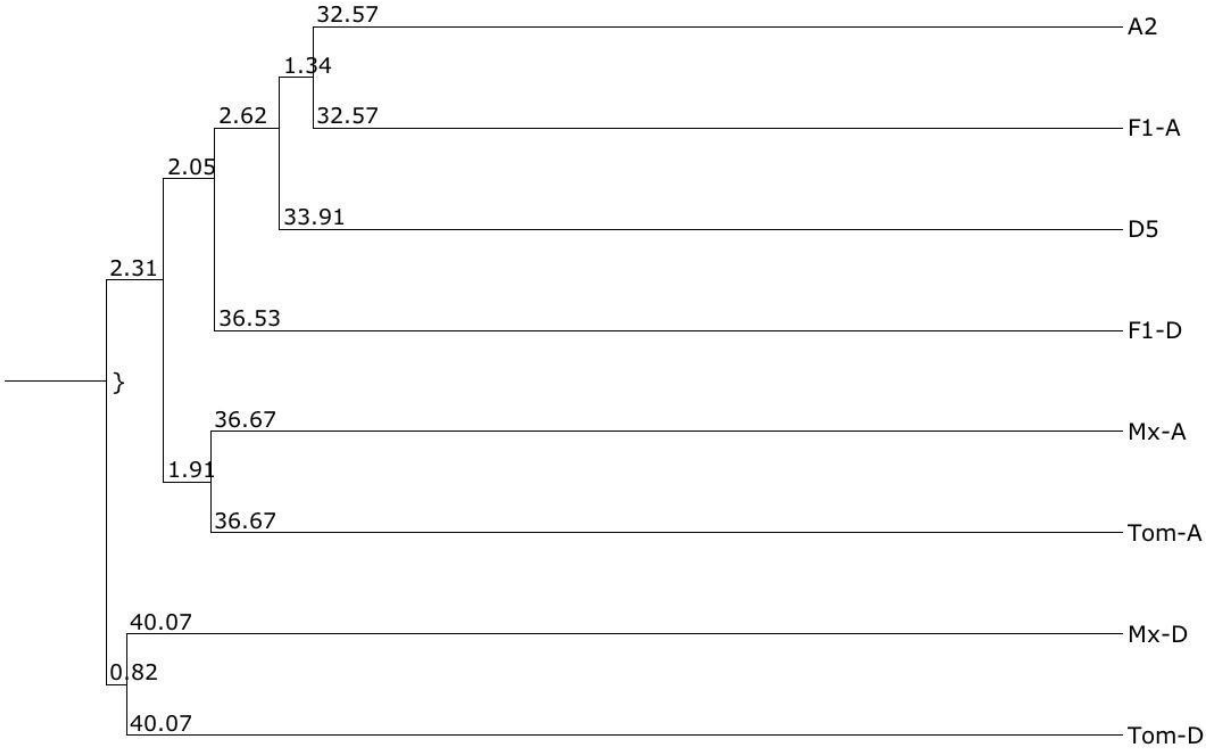
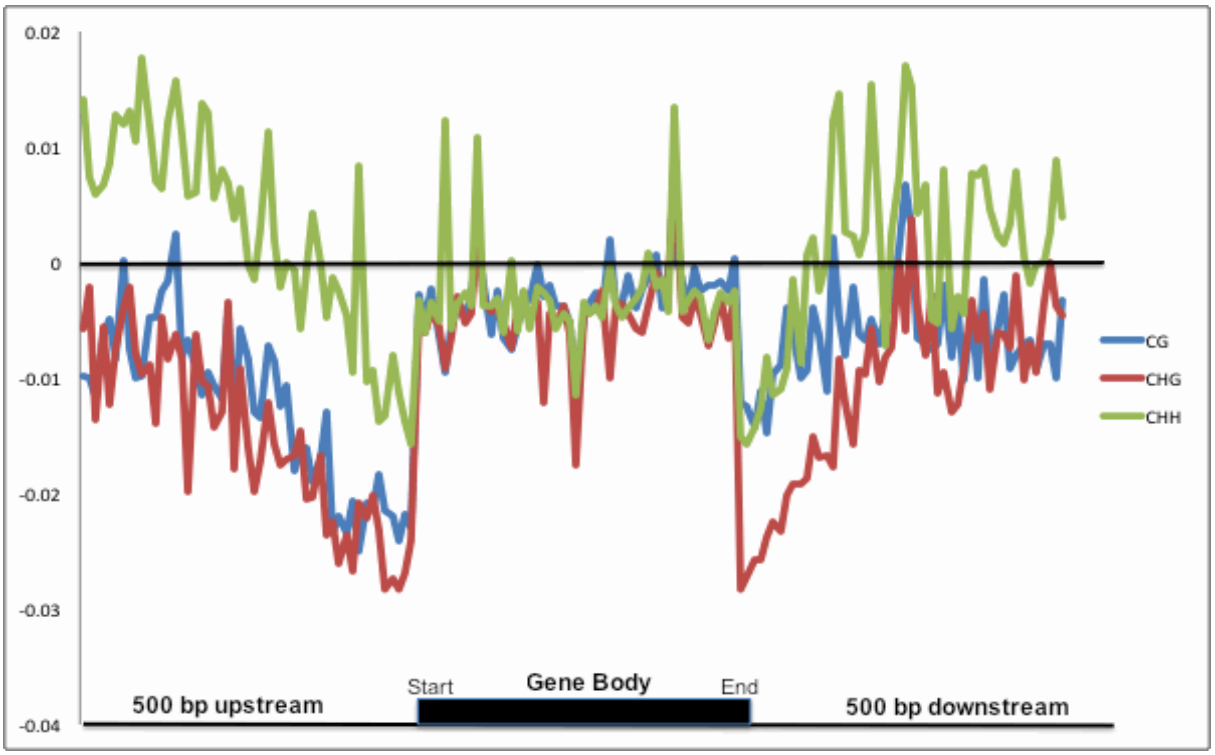
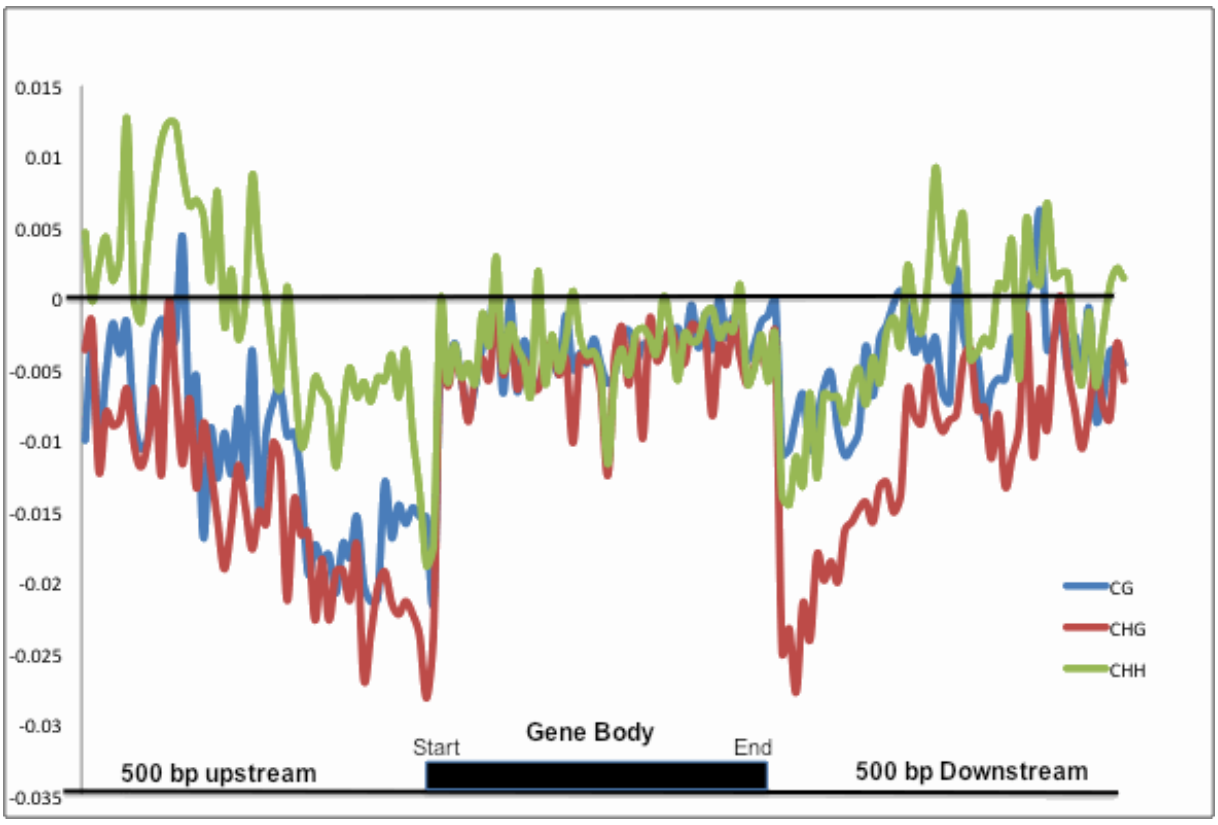

32.57
A2
1.34
2.62
32.57
F1-A
2.05
33.91
D5
2.31
36.53
F1-D
}
36.67
Mx-A
1.91
36.67
Tom-A
40.07
Mx-D
0.82
40.07
Tom-D

Figure 11. Correlation between methylation and expression across the average gene. for A) diploid F_1-hybrid B) *G. hirsutum* and C) *G. tomentosum*

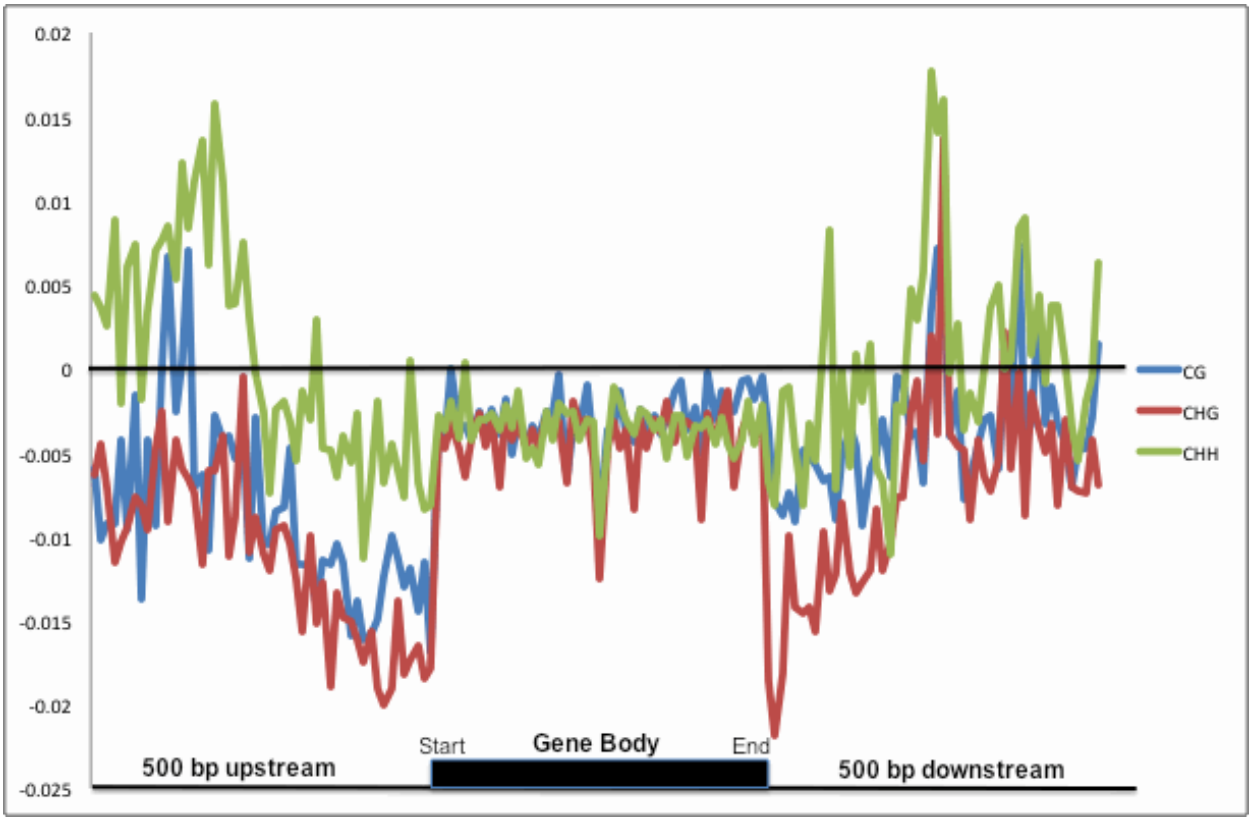
A)



B)


0.015
0.01
0.005
0
-0.005
-0.01
-0.015
-0.02
-0.025
-0.03
-0.035
CG
CHG
CHH
Gene Body
Start
End
500 bp upstream
500 bp Downstream

C)


0.02
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
-0.02
-0.025
CG
CHG
CHH
Start
Gene Body
End
500 bp upstream
500 bp downstream