FUNCTIONAL EPIGENOMICS IN INSECTS USING NEXT-GENERATION SEQUENCING METHODS

A Dissertation
Presented to
The Academic Faculty

by

Xin Wu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biological Sciences

Georgia Institute of Technology
August 2021

FUNCTIONAL EPIGENOMICS IN INSECTS USING NEXT-GENERATION SEQUENCING METHODS

Approved by:

Dr. Soojin V. Yi, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Michael A.D. Goodisman
School of Biological Sciences
Georgia Institute of Technology

Dr. I. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. Amelia R.I. Lindsey
Department of Entomology
University of Minnesota

Dr. Christina M. Grozinger
Department of Entomology
Pennsylvania State University

Date Approved: [May 03, 2021]

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. Soojin Yi, whom I am deeply grateful to for all the continued guidance, advice, and support throughout my graduate school career. Your immense knowledge and experience have made the greatest contributions to my research and I will forever treasure your mentorship. I would like to offer my sincere thanks my committee members Dr. Amelia Lindsey and Dr. Christina Grozinger for their invaluable collaborations and instructions. I am also indebted to my other committee members, Dr. Michael Goodisman and Dr. King Jordan, for their feedback and insights on my research.

I am grateful to all of the current and past Yi lab colleagues whom have shared their time with me. Dr. Isabel Mendizabal, Dr. Iksoo Huh, Dr. Dan Sun, Dr. Thomas Keller, Paramita Chatterjee, Hyeonsoo Jeong, Devika Singh, Thomas Layman, Brandon Smith, and Ben Long – thank you for your help, advice, and friendship.

Last but not least, I would like to thank all of my family and friends for unwavering love and support, especially during the past year. You have made this journey tremendously fulfilling and memorable.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **iii**

LIST OF SYMBOLS AND ABBREVIATIONS **ix**

SUMMARY **x**

CHAPTER 1. Introduction **xi**

CHAPTER 2. Genomic distribution and characterization of Methylation islands in hymenopteran insects **1**
2.1 Introduction **1**
2.2 Results **3**
2.2.1 Identifying Methylation Islands in Seven Invertebrate Genomes 3
2.2.2 Characteristics of MIs 4
2.2.3 MIs Tend to Occur in Evolutionarily Conserved Genes and Amino Acids within MIs are More Conserved than those Outside MIs 7
2.2.4 The Presence of MIs Affects Gene Expression 10
2.2.5 Knockdown of DMNT3 Implicates MIs in Alternative Splicing 13
2.3 Discussion **15**
2.4 Methods **17**
2.4.1 Analysis of WGBS and RNA-seq Data 17
2.4.2 Identifying mCGs and MIs 17
2.4.3 Protein and Amino Acid Conservation Score 18
2.5 Acknowledgements **19**

CHAPTER 3. *Wolbachia*-mediated asexuality is linked to distinct epigenomic and transcriptomic changes **20**
3.1 Introduction **20**
3.2 Results **22**
3.2.1 Introgressing a Sexual Nuclear Genome into an Asexual Cytoplasm Infected with Wolbachia 22
3.2.2 Wolbachia Infection Results in DNA Methylation Changes in T. pretiosum 25
3.2.3 Gene Expression and Exon Usage is Associated with Wolbachia Infection 26
3.2.4 Differential Exon Usage but Not Differential Expression is Associated with Differential Methylation 28
3.3 Discussion **32**
3.4 Methods **34**
3.4.1 Rearing of Trichogramma lines 34
3.4.2 Introgression of Sexual Genome into Wolbachia Infected Cytoplasm 34
3.4.3 Nucleotide Extractions 35
3.4.4 RNA Sequencing 35
3.4.5 Genome Sequencing 35

3.4.6 Whole-genome Bisulfite Sequencing 35
3.4.7 Creation of Alternative Reference Genome 36
3.4.8 RNA-seq Analysis 36
3.4.9 Analysis of Transcriptional Noise 37
3.4.10 WGBS Data Processing 37
3.4.11 Using WGBS Data to Analyze Introgressed Regions 37
3.4.12 WGBS Data Analysis 38
3.5 Acknowledgements 38

CHAPTER 4. Lineage and parent-of-origin methylation patterns in A. mellifera using whole-genome bisulfite sequencing 39
4.1 Introduction 39
4.2 Results 40
4.2.1 Honey Bees Exhibit Both Lineage and Parent-specific DNA Methylation 41
4.2.2 Genes with Signatures of Parent-specific Methylation 44
4.2.3 Weak Association Between Allelic Methylation and Expression 49
4.3 Discussion 50
4.4 Methods 52
4.4.1 Biological Sample Collection 52
4.4.2 WGBS Library Construction and Sequencing 52
4.4.3 Creating N-masked Genomes 52
4.4.4 WGBS Data Processing 53
4.4.5 Differential Methylation Analysis 53
4.4.6 RNA-seq Processing 54
4.4.7 Gene Ontology 54
4.5 Acknowledgements 55

CHAPTER 5. Gene Body DNA Methylation is associated with reduced gene expression variability 56
5.1 Introduction 56
5.2 Results 57
5.2.1 Core promoter elements are significant contributors to gene expression variation 58
5.2.2 DNA methylation is anti-correlated with expression variation 61
5.3 Methods 61
5.3.1 Gene expression data 61
5.3.2 Data processing 62
5.3.1 Core promoter elements 62
5.3.2 Statistics 63

CHAPTER 6. Conclusions 64

APPENDIX A. Supplementary material for chapter 2 69
Figure A.1 Distance to nearest neighbor for control and mCGs. 69
Figure A.2 Sliding window algorithm. 70
Figure A.3 Distribution of MIs in key genic regions. 71
Figure A.4 Permutation of MIs at exon-intron boundaries. 72

APPENDIX B. Supplementary material for chapter 3 **73**
Figure B.1 DMPs in genes. **73**
Figure B.2 Methylated genes have higher and are more constitutively expressed. 74

APPENDIX C. Supplementary material for chapter 4 **75**
Table C.1 Differentially expressed genes for each reproductive state and genetic block. 75
Table C.2 Overlap between DMGs and DEGs. **76**

REFERENCES **77**

LIST OF TABLES

Table 2.1	– Genome composition summary of species used	3
Table 2.2	– Summary of MI related characteristics in each species	4
Table 2.3	– Comparison of odds ratios between genes with and without MIs	9
Table 2.4	– Correlation between “same state MI” and “different state MI” genes	12
Table 2.5	– Methylation statistics of control and DMNT3 knockdown bees	14
Table 4.1	– DMP distribution in each genetic block and reproductive state	41
Table 4.2	– DMG direction of bias for each genetic block and reproductive state	45

LIST OF FIGURES

Figure 2.1	– Variable methylation landscapes between humans and honey bees.	2
Figure 2.2	– MIs characterized by genomic region in seven *Hymenopterans*.	7
Figure 2.3	– MIs are overrepresented in evolutionarily conserved genes.	8
Figure 2.4	– Relationship between amino acid conservation and MIs and DNA methylation.	10
Figure 2.5	– Gene expression levels of MI- and non-MI genes based on sequence conservation.	11
Figure 2.6	– Average expression levels of exons inside (MI-exon) and outside of MIs (non-MI-exon).	13
Figure 3.1	– Introgression scheme used to create genetically homogeneous lines of *Wolbachia* infected and free *Trichogramma*.	25
Figure 3.2	– Comparing methylation and expression between *Wolbachia* infected and uninfected wasps.	27
Figure 3.3	– Transcriptional noise and *Wolbachia* infection.	30
Figure 3.4	– Two example genes that contain both differentially used exons and differentially methylated positions.	31
Figure 4.1	– Examples of mCGs showing parent-of-origin and lineage effects.	34
Figure 4.2	– DMP biases for A) genetic block A and B) genetic block B.	47
Figure 4.3	– Number of genes belonging to each bias category based on the worker reproductive state and their overlaps.	48
Figure 5.1	– Linear model covariate coefficients.	58

LIST OF SYMBOLS AND ABBREVIATIONS

DMNT	DNA methyltransferase
MI	Methylation island
WGBS	Whole-genome bisulfite sequencing
CO	Complete orthologous
IO	Incomplete orthologous
UG	Unique genes
CGI	CpG island
mCG	Methylated cytosine
GLM	Generalized linear model
PCR	Polymerase chain reaction
DMP	Differentially methylated position
DMG	Differentially methylated gene
DEG	Differentially expressed gene
DEU	Differential exon usage
CV	Coefficient of variation
FET	Fisher’s exact test
FDR	False discovery rate
GO	Gene ontology
O/E	Observed/Expected

SUMMARY

DNA methylation is a widespread epigenetic modification implicated in many important processes such as development, disease, and genomic imprinting. In well-studied mammalian systems, DNA methylation at gene promoters acts as a transcriptional repressor including playing a critical role in X chromosome inactivation. Despite the importance and prevalence of DNA methylation, essential model organisms such as *D. melanogaster* and *C. elegans* have experienced lineage-specific losses of genomic DNA methylation. This thesis focuses on a comprehensive epigenomics survey and investigation of the Hymenopteran insect order, a group of insects including wasps, bees and ants that have retained functional DNA methylation systems. This diverse group of insects allows us to gain new insights in to the function role of DNA methylation, especially in the context of gene expression regulation. I will first provide a general survey of the epigenetic landscape of insects, which have a completely different pattern compared to mammals, and offer a new approach to quantifying and analyzing DNA methylation in these organisms. Next, I investigate changes to DNA methylation and gene expression that accompany a bacterial infection and a drastic shift from sexual to asexual reproduction in a parasitoid wasp. I will then examine how the intricate honey bee society gives rise to allele-specific methylation and its potential relationship to allele-specific expression. Finally, I explore the importance of DNA methylation along with other promoter elements in regulating gene expression variation.

CHAPTER 1. INTRODUCTION

DNA methylation, typically referring to the methylation of the fifth carbon in cytosines in the CpG context, has ancient origins and is widespread in both eukaryotes and prokaryotes (Jones 2012; Greenberg and Bourc'his 2019). The enzymes responsible for this chemical modification, DNA methyltransferases (DMNTs), are a conserved set of proteins where DNMT3 is responsible for *de novo* methylation of cytosines while DNMT1 maintains faithful inheritance of methylation by the addition of methyl groups to hemimethylated DNA following replication (Bird 2002; Jones 2012; Greenberg and Bourc'his 2019). In mammals, CpG methylation has diverse roles in processes ranging from genomic imprinting, development, and cellular differentiation to cancer and neuropsychiatric diseases (Greenberg and Bourc'his 2019).

Traditionally, CpG methylation in animals has been viewed and studied in the context of a transcriptional repressor (Yoder, et al. 1997; Schubeler 2015; Greenberg and Bourc'his 2019). Specifically, methylation in promoter regions is associated with down-regulation of transcription (Bird 2002; Greenberg and Bourc'his 2019), as well as silencing of one copy of the X chromosome in therian female mammals (Sharp, et al. 2011). DNA methylation of repetitive genomic sequences is also associated with protecting the genome from transposable elements activity (Yoder, et al. 1997; Schubeler 2015). Yet, despite the prevalence and importance of DNA methylation, its function in other lineages remains poorly understood (Elango, et al. 2009; Sarda, et al. 2012). The recent burst of whole genome methylation profiling of diverse species (Feng, et al. 2010; Zemach, et al. 2010; Wang, et al. 2013; Galbraith, et al. 2015; Lindsey, Kelkar, et al. 2018) has greatly increased

our ability to survey both the presence of DNA methylation in previously unexplored species as well as study its function. Of particular interest to scientists are invertebrate lineages, where DNA methylation is widespread yet exhibit lineage-specific variation in terms of the extent, including a complete loss in some lineages (Glastad, et al. 2011; Yi 2012; Bewick, et al. 2017; Rosic, et al. 2018).

Hymenopteran insects, which include bees, wasps, and ants, have been focused on for their extreme diversity, importance to ecosystems, and presence of DNA methylation (Lyko, et al. 2010; Wang, et al. 2013). The advent of whole genome methylation studies in insects began with the publication of the honey bee (*Apis mellifera*) genome and discovery of a functional set of enzymes orthologous to vertebrate DNA methyltransferases (Honeybee Genome Sequencing 2006). In total, four CpG-specific DNMTs (two DNMT1 and two DMNT3s) were found to be expressed (Honeybee Genome Sequencing 2006), and the genome of the honey bee was found to only have a small fraction of the methylation of heavily methylated mammalian genomes (Lyko, et al. 2010; Zemach, et al. 2010). The subsequent sequencing of other Hymenopterans revealed similar methylome patterns – DNA methylation in insects was almost exclusively limited to the gene bodies of evolutionarily conserved genes and enriched in exons compared to introns (Lyko, et al. 2010; Wurm, et al. 2011; Wang, et al. 2013; Lindsey, Kelkar, et al. 2018). In the honey bee, queens and workers exhibit vastly different morphology and behaviors, yet share an identical genome (Honeybee Genome Sequencing 2006; Kucharski, et al. 2008). The specialized royal jelly diet fed to the queen-to-be was shown to modulate genome wide methylation patterns and was partly responsible for the phenotypic differences between queens and workers (Lyko, et al. 2010). Remarkably, the epigenetic states linked to

different phenotypes was found to be plastic and could be manipulated between behavioral subcastes (Herb, et al. 2012). However, direct causation, or even association, between changes in methylation and transcription mirroring mammalian systems have been difficult to establish in honey bee and other Hymenopterans (Elango, et al. 2009; Lyko, et al. 2010; Wang, et al. 2013; Galbraith, et al. 2015; Galbraith, et al. 2016).

In my dissertation research, I focused on the study of DNA methylation in Hymenopteran insects. My overarching goals were to further our understanding of the evolution of DNA methylation, as well as to investigate the specific roles of DNA methylation in the study species. In Chapter 2, we propose a method for detecting and quantifying units of methylated CpG clusters we refer to as "methylation islands" (MIs) in insects. This idea was inspired by clusters of hypomethylated CpGs are often found at transcriptionally active promoters in mammals called "CpG islands" (Bird 1992; Schubeler 2015). We employed high quality whole genome bisulfite sequencing datasets from seven Hymenopteran species to study the distribution and characteristics of these MIs. Additionally, we integrated RNA-seq are from three of the seven species to investigate potential functional associations between DNA methylation and transcription.

In Chapter 3, I studied epigenetic and transcriptomic changes that accompany a drastic shift from sexual to asexual reproduction associated with *Wolbachia* infection in the *Trichogramma pretiosum* wasp. *Wolbachia* is a highly successful endosymbiont that is widespread and has profound effects on host fitness (Werren, et al. 2008; Zug and Hammerstein 2012). In *Trichogramma* wasps, *Wolbachia* infection induces parthenogenesis in females, a mode of asexual reproduction where unfertilized eggs develop into diploid adult females that propagate this infection vertically (Stouthamer, et

al. 2010). Due to geographic isolation of infected and uninfected lines, we devised a clever introgression scheme to control for confounding genetic differences between uninfected sexually reproducing *Trichogramma* and *Wolbachia*-infected wasps. We then performed whole genome bisulfite sequencing in parallel with RNA-seq to investigate epigenetic and transcriptomic changes linked to such an extreme shift in reproductive physiology.

One of the many attractive qualities for studying honey bees is their extraordinary social structure. The typical queen produces offspring by mating with a multitude of males and the resulting differences in matrigene and patrigene relatedness among colony individuals has been hypothesized to contribute to parent-of-origin-specific expression (Haig 2000; Queller 2003). The kinship theory developed by David Queller predicts that the intragenomic conflict between the matrigenes and patrigenes due to differential fitness pressures should lead to parent-specific expression where the expression of an allele is dependent on the parent it was inherited from (Queller 2003). A previous study leveraging genotyping of European and Africanized reciprocal honey bee crosses found support for this theory using RNA-seq (Galbraith, et al. 2016), but the mechanisms behind these observations were not studied. In the fourth Chapter, I use the previously mentioned reciprocal crosses to investigate whether DNA methylation, the primary regulator of parent-specific expression in mammals and plants (Bird 2002; Queller 2003; Law and Jacobsen 2010), has a similar role in modulating parent-specific effects in the honey bee.

In the fifth Chapter, I investigated the role of DNA methylation in relation to variation of gene expression variation in insects. Gene expression levels may vary between individuals and within cell populations due to several mechanisms, including intrinsic factors such as the rate of transcription as well as extrinsic factors such as parasite infection

and cell cycle (Fraser, et al. 2004; Sanchez and Kondev 2008). It was previously proposed that DNA methylation may also affect gene expression variability (Sanchez and Kondev 2008; Huh, et al. 2013; Sevier, et al. 2016; Wu, et al. 2020b). It is hypothesized that natural selection has affected expression variability of highly expressed genes as a means to control for the inherent stochasticity involved in transcription and subsequent protein synthesis, which has been shown to be detrimental to organisms (Fraser, et al. 2004; Wang and Zhang 2011; Barroso, et al. 2018). Here, we gather high-quality RNA-seq datasets (8 honey bee and 12 *Drosophila*) to determine factors that contribute to gene expression variability. Importantly, DNA methylation is a known contributor to reducing gene expression variability (Huh, et al. 2013; Hunt, et al. 2013; Wang, et al. 2016) and the addition of *Drosophila* data allows us to ask whether the patterns of gene expression variability vary between the honey bee and a lineage that has lost ancestral gene body methylation.

The research in this thesis encompasses a detailed investigation into the relationship between DNA methylation and transcription in Hymenopteran insects and expands our current understanding of the function of the epigenome.

CHAPTER 2. GENOMIC DISTRIBUTION AND CHARACTERIZATION OF METHYLATION ISLANDS IN HYMENOPTERAN INSECTS

2.1 Introduction

The role of DNA methylation has been characterized extensively and plays important roles ranging from imprinting and disease to aging and development (Rainier and Feinberg 1994; Razin and Cedar 1994; Robertson and Wolffe 2000; Saze, et al. 2003). With the vast amount of sequencing in recent years, we have been able to dramatically expand the scope of DNA methylation profiling into previously unexplored lineages. This influx of genomic DNA methylation data has the potential to greatly increase our understanding of the phylogenetic distribution of DNA methylation and advance our knowledge of its function.

Traditionally viewed as repressor of transcription, we now have evidence that the function of DNA methylation is target dependent. When methylation occurs in gene regulatory regions such as promoters, downstream transcription is repressed (Jones 2012; Schubeler 2015). Similarly, DNA methylation at repetitive elements protects the genome from transposition of these elements (Yoder, et al. 1997; Schubeler 2015). In contrast, DNA methylation found in gene bodies is linked to active transcription, although whether it is the cause or effect remains unknown (Jones 2012). Though DNA methylation is widespread, some lineages including model organisms such as fruit flies and nematodes have experienced lineage-specific losses of methylation (Glastad, et al. 2011; Yi 2012; Rosic, et al. 2018). Of particular interest are insects from the order Hymenoptera due to being close relatives of fruit flies while also having functional DNA methylation systems (Lyko, et al. 2010; Hunt, et al. 2013; Wang, et al. 2013).

Interestingly, genomic methylation landscapes vary between species and are especially notable when comparing invertebrates to vertebrates. Vertebrate methylation, particularly mammals, is heavily methylated throughout the genome with the exceptions of clusters of hypomethylated CpGs known as "CpG islands" (Bird, et al. 1985; Bird 1992). These CpG islands are often used targets for methylation chips and as units to describe regions of methylation and their associations with transcription (Mendizabal, et al. 2014; Schubeler 2015). In contrast, invertebrate genomic methylation tends to be low. In hymenopteran insects, methylation is almost exclusively found within gene bodies and especially enriched in coding regions (Lyko, et al. 2010; Wang, et al. 2013; Bewick, et al. 2017; Lindsey, Kelkar, et al. 2018). Figure 2.1 shows a typical genomic region contrasting the methylation landscapes between honey bee and humans.

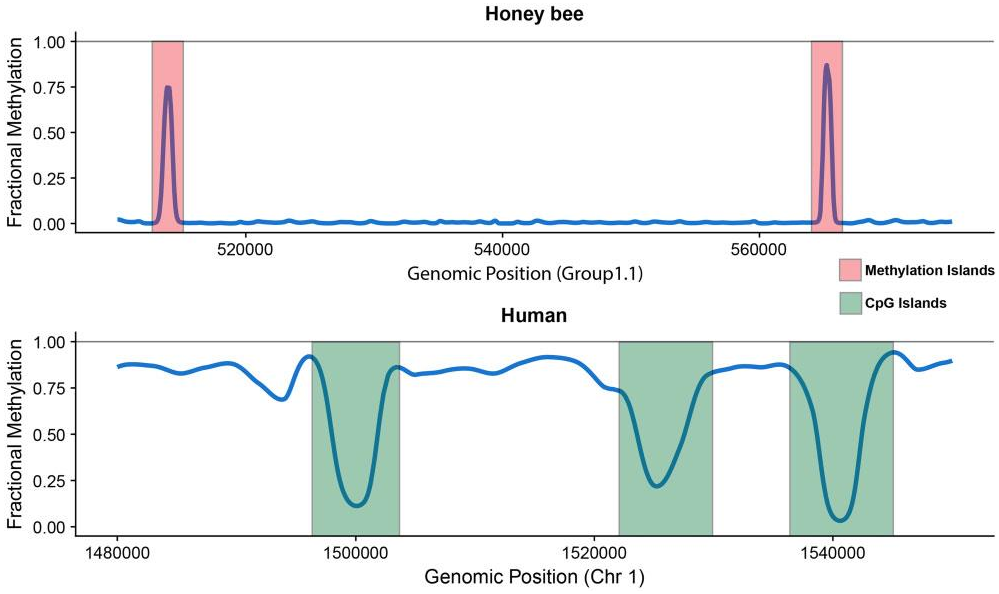

Figure 2.1 – Variable methylation landscapes between humans and honey bees. The honey bee genome is lowly methylated with only a few but clustered number of methylated CpGs. We termed these clusters "Methylation islands" which are usually around 250bp in length. In contrast, the human genome is heavily methylated throughout with regions of hypomethylated CpG islands that are around ~1kb in length.

CpG islands have been a useful concept in many studies that have shed light on the functional role of epigenetic variation in vertebrate species, and we apply a similar concept here to investigate the function and distribution of DNA methylation clusters in insects. We refer to these clusters of methylated CpGs are "methylation islands" (MIs) and applied this concept to seven hymenopteran species with high quality genome assemblies and methylome data. We first identify these MIs throughout the genomes and characterize their distribution, followed by exploring the functional roles these MIs have on transcription using RNA-seq data.

2.2 Results

2.2.1 Identifying Methylation Islands in Seven Invertebrate Genomes

The seven species we selected (*Apis mellifera, Camponotus floridanus, Harpegnathos saltator, Nasonia vitripennis, Polistes canadensis, Solenopsis invicta, and Trichogramma pretiosum*) had well-annotated genomes along with whole-genome bisulfite sequencing (WGBS) data (Table 2.1). The fraction of methylated CpGs in the genome was low as expected, with all species examined having less than 1% (Table 2.1). The average fractional methylation of these methylated CpGs (mCGs) ranged from 0.44 to 0.74 while the global average of all CpGs ranged from 0.008 and 0.025 (Table 2.1). We tested to see if methylated CpGs were clustered based on previous findings, and found this to be the case (Wang, et al. 2013; Huh, et al. 2014). Specifically, the distance between neighbouring mCGs was significantly shorter than randomly selected CGs for all seven species.

Table 2.1 – Genome composition summary of the seven species used in this study and their basic methylation statistics.

Species	Genome Size (Mb)	# Protein-Coding Genes	# of mCGs (% of all CGs)	Avg. Fractional Methylation of mCGs
Apis mellifera	234.07	15,314	78,846 (0.78%)	0.584
Camponotus floridanus	232.68	11,042	85,746 (0.84%)	0.635
Harpegnathos saltator	294.46	11,838	112,212 (0.53%)	0.662
Nasonia vitripennis	295.78	13,354	114,261 (0.85%)	0.737
Polistes canadensis	211.21	9,876	15,744 (0.24%)	0.386
Solenopsis invicta	396.02	14,451	157,829 (0.98%)	0.526
Trichogramma pretiosum	196.22	13,200	60,298 (0.60%)	0.345

In order to capture these clustered mCGs, referred to as "methylation islands" (MIs), we developed a sliding window algorithm to search the genome for regions of dense mCGs and classified them as units of measurement for DNA methylation. In short, this algorithm labelled MIs as regions that are at least 200bp in length and contain >2% of mCGs (approximately a 3-fold enrichment compared to the genome average, Table 2.1).

2.2.2 Characteristics of MIs

Our sliding window approach captured thousands of MIs in each of the seven species. As we expected, the majority of mCGs in the genome were found within MIs even though the total length of MIs was only a small fraction of the genome size (Table 2.2). The average length of MIs in the genome was positively correlated with the number of mCGs (Pearson correlation coefficients = 0.97) rather than genome size (Tables 2.1 and 2.2). For instance, *P. canadensis* had the fewest MIs out of all the species with a total number of 1,342 even though its genome is 20 Mb larger than *T. pretiousum* which had 4,889 MIs.

Table 2.2 – Summary of MI related statistics in each of the seven species.

	Apis mellifera	*Camponotus floridanus*	*Harpegnathos saltator*	*Nasonia vitripennis*	*Polistes canadensis*	*Solenopsis invicta*	*Trichogramma pretiosum*
# of predicted MIs	5,126	6,327	8,375	9,644	1,342	10,574	4,889
# of mCGs in MIs (% of total mCGs)	29,254 (37.1%)	47,804 (55.8%)	78,490 (69.9%)	85,007 (74.4%)	8,293 (52.7%)	112,819 (71.5%)	30,141 (50%)
Total MI length (bp) (% of genome)	1,043,247 (0.45%)	1,803,969 (0.77%)	2,969,693 (1.01%)	3,355,006 (1.13%)	210,235 (0.099%)	4,291,930 (1.08%)	1,136,846 (0.58%)
Avg. MI length (bp)	213.15	286.12	355.59	348.88	157.66	406.89	233.53
Avg. mCG density per MI (# of mCGs/MI length)	0.03	0.02	0.03	0.02	0.07	0.03	0.03
# of MIs overlapping with genes[a] (% of all MIs)	4,958 (96.7%)	6,082 (96.1%)	7,845 (93.7%)	9,079 (94.1%)	1,020 (76%)	9,843 (93.1%)	4,603 (94.2%)
# of MIs overlapping exclusively with genes[a] (% of all MIs)	4,788 (93.4%)	5,961 (94.2%)	7,606 (90.8%)	8,873 (92.0%)	1,010 (75.3%)	9,477 (89.6%)	4,469 (91.4%)
# of MIs overlapping with exons/exclusively with exons (% of all MIs)	4,830/3,117 (94.2%/60.8%)	5,763/2,634 (91.1%/41.6%)	7,319/2,704 (87.4%/32.3%)	8,184/3,381 (84.9%/35.1%)	741/524 (55.2%/39.0%)	8,839/3,412 (83.6%/32.3%)	4,433/2,926 (90.7%/59.8%)
# of MIs overlapping with introns/exclusively with introns (% of all MIs)	1,794/178 (35.0%/3.5%)	3,404/382 (53.8%/6.0%)	5,011/592 (59.8%/7.1%)	5,739/1,206 (59.5%/12.5%)	478/273 (35.6%/20.3%)	6,300/1,160 (59.6%/11.0%)	1,881/242 (38.5%/4.9%)
# of MIs overlapping with exon–intron boundaries/ only one exon–intron boundary (% of all MIs)	1,637/705 (31.9%/13.8%)	3,051/1,312 (48.2%/20.7%)	4,461/1,690 (53.3%/20.2%)	4,672/1,635 (48.4%/17.0%)	205/92 (15.3%/6.9%)	5,252/2,123 (49.7%/20.1%)	1,649/611 (33.7%/12.5%)
# of MIs overlapping with promoters (% of all MIs)	172 (3.4%)	117 (1.8%)	146 (1.7%)	199 (2.1%)	30 (2.2%)	308 (2.9%)	213 (4.4%)

[a]Defined as the region spanning the transcript start site to the transcription termination site.

In *A. mellifera*, the majority of MIs overlapped with gene bodies (96.7%, with gene bodies defined as the region between the transcription start site and transcription termination site), especially exons (94.2%; Table 2.2). Furthermore, 60.8% of all MIs were exclusively within exons. MIs also overlapped with introns, but much less frequently. In *A. mellifera*, only 3.5% of MIs were exclusively overlapped with introns. Interestingly, 31.9% of *A. mellifera* MIs were found across exon-intron boundaries. Previous studies discussed the possibility of DNA methylation playing a role in alternative splicing by signalling splice junctions (Lyko, et al. 2010; Herb, et al. 2012; Li-Byarlay, et al. 2013; Galbraith, et al. 2015). Therefore, we asked if MIs were enriched at exon-intron boundaries. Our results show that this was in fact the case (empirical P value < 0.001) for all seven species.

It has been speculated that mCGs in insects were biased towards the 5' end of a gene (Lyko, et al. 2010; Hunt, et al. 2013; Wang, et al. 2013; Galbraith, et al. 2015). Using MIs as our unit of measurement for methylation, we found that they tended to be slightly biased towards the 3' end in *A. mellifera* and *T. pretiosum* (Figure 2.2B). In contrast, MIs in the four of the species (*C. floridanus*, *H. saltator*, *P. canadensis*, and *N. vitripennis*) displayed 5' bias (Figure 2.2B).

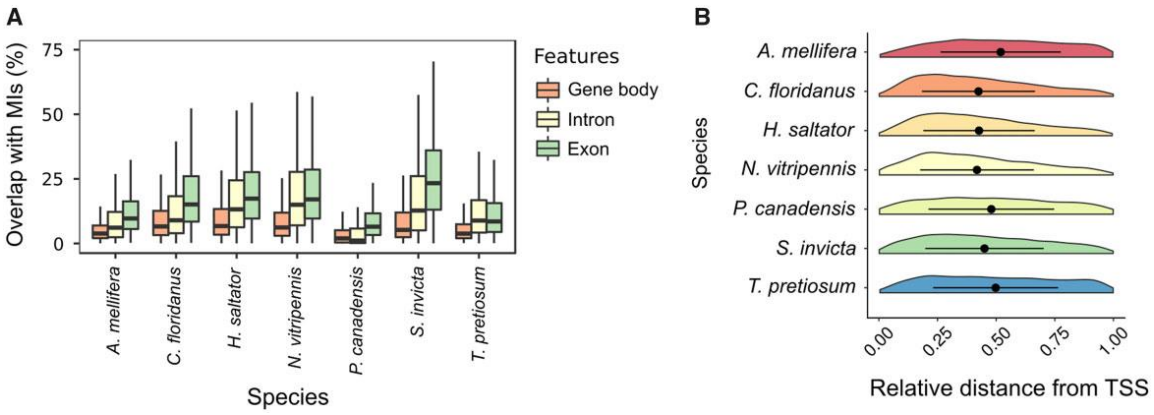

Figure 2.2 – MIs characterized by genomic region in seven Hymenopterans. A) Box plots showing whether MIs were found in gene bodies, exons, or introns. B) Violin plots displaying the position of MIs relative to the TSS of genes.

2.2.3 MIs Tend to Occur in Evolutionarily Conserved Genes and Amino Acids within MIs are More Conserved than those Outside MIs

Previous studies typically used a binary classification for genes, labelling them as either methylated or unmethylated based on the mean fractional methylation (Lyko, et al. 2010; Sarda, et al. 2012; Wang, et al. 2013). They showed that methylated genes were more evolutionarily conserved compared to unmethylated genes (Lyko, et al. 2010; Wang, et al. 2013; Galbraith, et al. 2015), and we used a similar approach to determine whether the presence of MIs in genes displayed a similar quality. We first determined a set of all orthologous genes shared in all seven species using protein sequences (Materials), yielding a total of 5,403 (44%) single copy orthologues out of 12,249 gene sets. We labelled these 5,403 genes as Complete Orthologues (CO). In the remaining gene sets, there were 6,429 (52%) that were found in two or more species which we classified as Incomplete

Orthologous gene (IO) sets. Finally, genes that were lineage-specific to each species were called the Unique Gene (UG) set (Figure 2.3A).

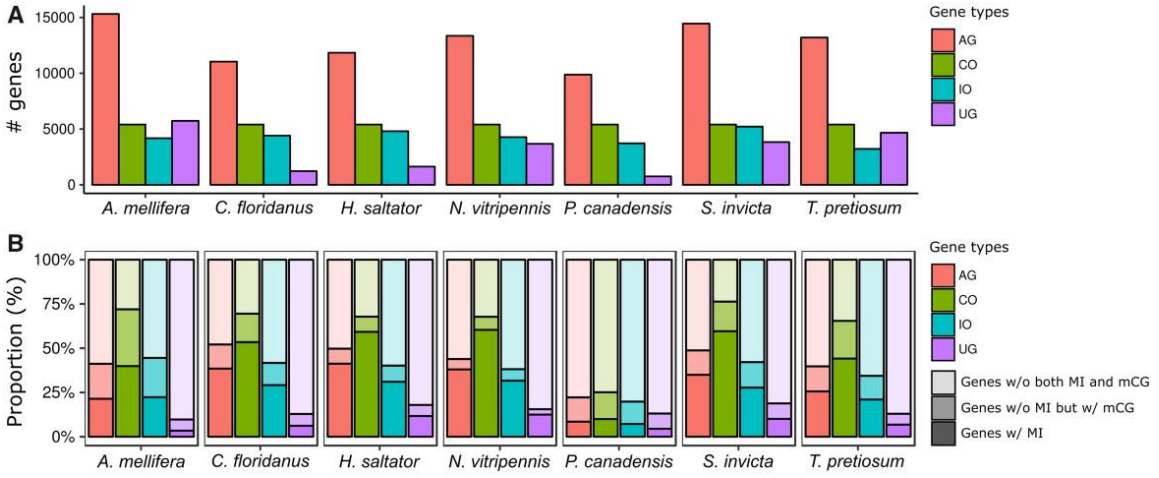

Figure 2.3 – MIs are overrepresented in evolutionarily conserved genes. A) Bar plots summarizing the number of genes classified as either all genes (AG), complete orthologous genes (CO), incomplete orthologous genes (IO), and unique genes (UG) in each species. B) The proportion of genes having different types of methylation features.

We followed by analyzing the frequency of genes with 1) MI, 2) without MI but at least one mCG, 3) without either MI or mCG in each gene set. We found that the proportion of genes with an MI is higher in the CO set compared to those in the IO and UG whereas the frequency of genes without MI but at least one mCG is comparable between CO and IO (Figure 2.3B). We next tested to see if genes with MIs were overrepresented in CO compared to IO with a Fisher's exact test, which yielded an average odds ratio of 3.1. In contrast, using the number of genes with an MI but at least one mCG resulted in an average odds ratio of 1.31. The odds ratios between the two tests were statistically significantly different, suggesting that clusters of mCGs, and therefore MIs, rather than individual mCGs, tend to be enriched in conserved genes (Table 2.3).

Table 2.3 – Statistical comparison of differences in Odds Ratios (OR) of genes with and without MIs using Z approximation.

Species	OR of Genes w/MI[a]	OR of Genes w/o MI but w/mCG[b]	Difference of Log. OR (δ)	SE(δ)	*P* Value
Apis mellifera	2.31	1.65	0.34	0.07	3.8E-07
Camponotus floridanus	2.79	1.33	0.74	0.07	2.2E-16
Harpegnathos saltator	3.23	0.93	1.25	0.08	2.2E-16
Nasonia vitripennis	3.29	1.14	1.06	0.09	2.2E-16
Polistes Canadensis	1.44	1.23	0.16	0.10	5.7E-02
Solenopsis invicta	3.84	1.19	1.17	0.07	2.2E-16
Trichogramma pretiosum	2.97	1.74	0.53	0.08	1.2E-11

NOTE.—Odds ratios were calculated and summarized in supplementary table S1, Supplementary Material online.
[a]Odds ratio of the number of genes with MIs and the number of the remaining genes between CO and IO types, respectively, were tested using Fisher's exact test.
[b]Odds ratio of the number of genes without MIs but with mCGs and the number of the remaining genes between CO and IO types, respectively, were tested using Fisher's exact test.

Additionally, we looked at whether the presence of DNA methylation and MIs was correlated with conservation status of individual amino acids. We first mapped the genomic coordinates of mCGs within coding regions to their corresponding positions in the protein sequence and quantified their conservation scores using the Jensen-Shannon (JS) divergence of protein sequence conservation (Capra and Singh 2007). We then applied a linear mixed model to predict the conservation scores of amino acids depending on the presence of mCG sites in the DNA sequence and the location of amino acids within or outside of MIs (Materials). We found that amino acids with mCGs had significantly higher conservation scores than those without mCGs (Figure 2.4). Moreover, amino acids within MIs had higher conservation scores when compared to amino acids outside MIs (*P* value $< 2.2 \times 10^{-16}$). Surprisingly, we also found that nucleotides that code for amino acids inside MIs that did not have any mCGs had comparable or higher conservation scores that amino acids than were inside MIs and had mCGs (Figure 2.4). While the relationship between the location of amino acids with respect to MIs and their conservation scores varied in different species, we consistently saw that sites within MIs had higher conservation scores that sites

outside of MIs. Our findings demonstrate that methylation islands had stronger association with protein sequence conservation than individual mCGs.

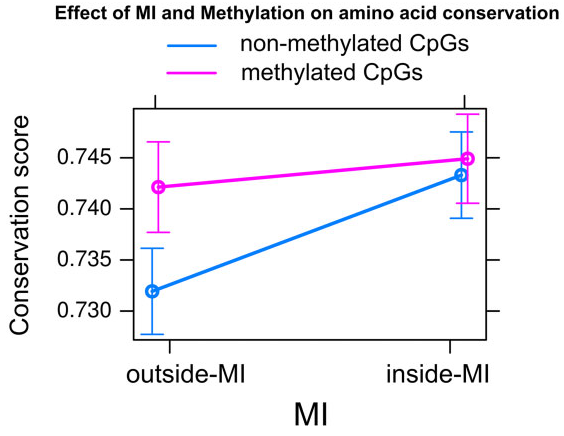

Figure 2.4 – Relationship between amino acid conservation and MIs and DNA methylation. We applied a linear mixed model to fit the conservation score of amino acids depending on if they located outside or inside MIs and whether they contained mCGs as the main factors and the interaction and random factors being gene and species, respectively. We used the Jensen-Shannon (JS) divergence to calculate the amino acid conservation score.

2.2.4 The Presence of MIs Affects Gene Expression

Previous studies provided evidence that gene body methylation tends to occur in evolutionarily conserved genes which also have constitutively and highly expressed (Elango, et al. 2009; Lyko, et al. 2010; Wang, et al. 2013; Galbraith, et al. 2015). We tested to see whether the presence of MIs had a similar pattern on gene expression. We normalized gene expression levels and compared them between MI- and non-MI- genes for three of the seven species that we had RNA-seq data for (Figure 2.5). In all three species, we found that MI-genes exhibited higher gene expression levels than non-MI genes. Furthermore, high conserved genes such as CO genes had higher expression levels than lowly conserve genes (IO and UG) in all species. These results agree with previous observations showing

a positive correlation between gene body methylation and gene expression and sequence conservation (Sarda, et al. 2012; Huh, et al. 2013; Hunt, et al. 2013). Notably, expression levels of MI genes remained consistently high regardless of conservation status while non-MI genes decrease in expression as conservation status decreased (Figure 2.5)

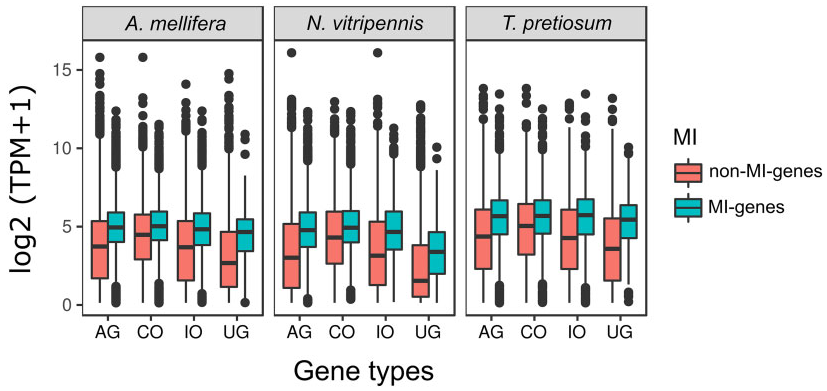

Figure 2.5 – Gene expression levels of MI- and non-MI genes based on sequence conservation. Gene expression levels are $\log_2$ transformed and normalized by gene length while the x-axis categorizes genes based on their conservation level (all genes [AG], complete orthologous genes [CO], incomplete orthologous genes [IO], and unique genes to each species [UG]).

We next described gene expression changes based on the gain or loss of MIs within conserved genes. Because it is difficult to directly compare expression levels between species, we tested how changes in MIs in CO genes affected gene expression between different species. First, each gene was classified as either being "same MI state" or "different MI state". "Same MI state" genes either lacked MI in both species or contained an MI in both while "different MI state" genes only had MIs in one species. Overall, there were a greater number of "same MI state" genes than "different MI state" genes in orthologous gene pairs which agrees with our previous observations (Table 2.4). We applied pairwise gene expression comparisons between the two groups for each species and found a significant difference in Spearman's rank correlation coefficients for all

pairwise comparisons between “same MI state” and “different MI state” genes. Moreover, “same MI state” genes showed stronger correlations which suggests that MIs in conserved genes are indeed associated with constitutively and highly expressed genes. (Table 2.4).

Table 2.4 – Pairwise correlation coefficients between “Same state MI” and “Different State MI” genes.

	Same State MIs		Different State MIs		
	Spearman's ρ	Number of Genes	Spearman's ρ	Number of Genes	*P* value
Apis mellifera–Nasonia vitripennis	0.607	3,590	0.557	1,768	9.30E-03
Apis mellifera–Trichogramma pretiosum	0.374	3,587	0.301	1,779	4.50E-03
Nasonia vitripennis–Trichogramma pretiosum	0.468	3,927	0.351	1,431	2.20E-16

NOTE.—The correlation coefficients were estimated between two species' gene expression level using Spearman's rho correlations.

We next tested whether MIs affected gene expression levels by comparing the relative expression of exons within MIs (MI-exon) and exons outside of MIs (non-MI-exon). The median expression level was higher for MI-exons than non-MI-exons and this was particularly highlighted for CO and IO genes (Figure 2.6). We saw this consistent pattern of higher expression of MI-exons regardless of species and gene conservation status, suggesting a robust relationship between the presence of MIs and levels of gene expression.

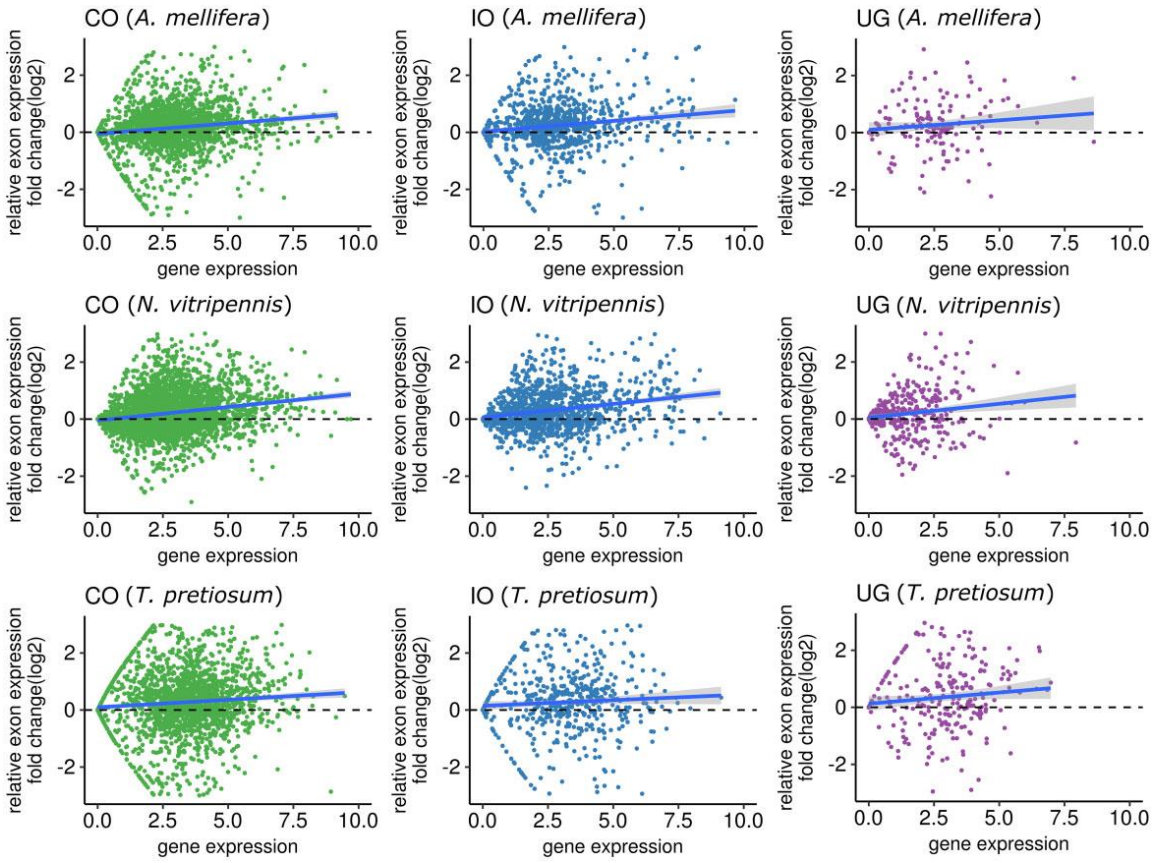

Figure 2.6 – Average expression levels of exons inside (MI-exon) and outside of MIs (non-MI-exon). We calculated the fold change between MI- and non-MI exons for each of the three gene conservation types. Each dot in the plot represents one gene. A locally weighted smoothing curve was applied to show the general trend of relative expression bias where values > 0 means higher expression of MI-exons compared to non-MI-exons. This analysis was done for A) *Apis mellifera*, B) *Nasonia vitripennis*, and C) *Trichogramma pretiosum*.

2.2.5 Knockdown of DMNT3 Implicates MIs in Alternative Splicing

We utilized *A. mellifera* gene expression data from a previous knockdown experiment of DMNT3 (Li-Byarlay, et al. 2013), the enzyme responsible for *de novo* methylation, to determine whether reduced genomic DNA affected transcription. Consistent with the function of DMNT3, we observed a modest reduction of both mCGs and MIs in the knockdown individual (Table 2.5). Overall, 89.8% of mCGs were shared

between control and knockdown samples which was also reflected in the 83.2% of shared MIs (Table 2.5). A total of 205 genes lost MIs in the knockdown sample, though we found no significant expression different in those genes. Gene ontology analysis of genes that showed similar methylation levels but lost MIs in the knockdown sample revealed functions related to nucleotide binding (*P* value = 0.017) and methyltransferase activity (*P* value = 0.032), though these were no longer significant following adjustment for false discovery rate.

Table 2.5 – Summary of methylation statistics in control and DNMT3 knockdown honey bees.

	Control	dnmt3 Gene Knockdown
# total mCG sites	78,846	75,897
# genes with mCG sites	6,308	6,277
Avg. # of mCGs per gene	12.3	11.9
# MIs (MI genes)	5,126 (3,280)	4,946 (3,207)
# MIs only present in group (MI genes)	501 (222)	372 (147)
# MIs at exon–intron boundary only present in group	116	38

Interestingly, 116 (23.1%) of the 501 MIs lost in the knockdown overlapped with exon-intron boundaries, suggesting that MIs lying at exon-intron boundaries tend to be excluded from the effects of DNMT3 knockdown ($P < 0.05$, Fisher's exact test). This observation is in line with the importance of DNA methylation, and subsequently MIs, at splicing sites (Li-Byarlay, et al. 2013). Additionally, the 327 MIs that were gained in the knockdown were significantly underrepresented at exon-intron boundaries (*P* value < 0.0001, Fisher's exact test), further indicating that splicing regulation may be affected in DNMT3 knockdown bees (Li-Byarlay, et al. 2013).

2.3 Discussion

A classical finding in mammalian epigenetics was the discovery of hypomethylated CpGs occurring in clusters, or "CpG islands" (CGIs) (Bird 1992; Bird 1995; Suzuki and Bird 2008), which have been useful markers for studying DNA methylation for decades (Suzuki and Bird 2008; Illingworth and Bird 2009; Yi 2017). The recent explosion in sequencing of methylome data of invertebrate species has provided an intriguing contrast between the different epigenetic landscapes of mammals and invertebrates (Figure 2.1). These differences bring about several interesting questions: in an otherwise unmethylated genome, do these rare mCGs occur in clusters? And if so, what functional roles do they play? To answer these questions, we used high quality methylome data from seven hymenopteran insects to characterize their methylation landscapes. Previously, methylation in insects was studied in the context of defining genes as either methylated or unmethylated, and measuring methylation based on the average fractional methylation level of a gene (Lyko, et al. 2010; Wang, et al. 2013; Lindsey, Kelkar, et al. 2018). While this approach provided meaningful insights into many aspects of invertebrate DNA methylation, taking averages of typically small numbers of mCGs may have diluted true signals of DNA methylation (Lyko, et al. 2010; Bonasio, et al. 2012; Wang, et al. 2013). However, these studies showed that DNA methylation occurred in clusters, a pattern we confirmed using the seven species here. We developed a sliding window algorithm to capture clusters of mCGs similar to the concepts for identifying CpG islands in mammals, reasoning that these clusters may represent functional units and therefore be conserved across closely related species similar to mammalian species (Illingworth and Bird 2009). This approach led to the identification of "methylation islands" (MIs) with a 3-fold

enrichment of methylation compared to the rest of the genome. Interestingly, mammalian CpG islands typically show a 3-fold enrichment of unmethylated CpGs (Gardiner-Garden and Frommer 1987; Jones and Takai 2001). Despite the similarity, criteria for defining CGIs are known to require adjustments depending the species, primarily due to differences in nucleotide composition (Matsuo, et al. 1993; Aerts, et al. 2004). Therefore, our definition and criteria for selecting MIs will likely require adjustments as well depending on the specific organism at hand.

One of the main consequences of CGIs was that genes containing them in their promoters had higher and more stable gene expression compared to genes without promoter CGIs (Aerts, et al. 2004; Elango and Yi 2008). This trend was consistent across diverse vertebrate species (Elango and Yi 2008). Here, we show that MIs in a group of insects have similar important implications for gene expression. First, they are overrepresented at exon-intron boundaries which is consistent with their proposed role of regulating alternative splicing (Flores, et al. 2012; Herb, et al. 2012; Li-Byarlay, et al. 2013; Galbraith, et al. 2015). This could potentially aid in discovering previously unannotated genes and their coding regions. In DNMT3 knockdown samples (Li-Byarlay, et al. 2013), MIs at exon-intron boundaries tended to be preserved at a rate higher than by random chance. Second, MI-genes exhibited higher and more stable gene expression compared to non-MI genes, a pattern that was mirrored at the exon level as well. This supports previous conclusions about the role of DNA methylation and inclusion of alternative transcripts. Further, we explored whether gain and loss of MIs influenced gene expression, which may reveal insights into cause-and-effect relationships between DNA methylation and gene expression. Though the available datasets are from fairly diverged species, we were

nevertheless able to show that expression levels were strongly correlated with MIs in coding regions across species. Our findings here offer insights into characteristics and functions of DNA methylation beyond single mCGs and implications of regions of methylation on transcription.

2.4 Methods

2.4.1 Analysis of WGBS and RNA-seq Data

Raw sequences for each species were downloaded from SRA and subjected to basically quality control such as adapter and low quality read trimming using Trim_galore! (Martin 2011). They were then aligned to their respective reference genomes and deduplicated using Bismark v0.14.4 (Krueger and Andrews 2011).

RNA-seq data from *A. mellifera*, *N. vitripennis*, and *T. pretiosum* were also downloaded from SRA. The reads were processed using FastQC to assess quality and adapters were removed with Trimmomatic (Bolger, et al. 2014a). We then aligned and quantified transcript count using Tophat2 and FeatureCount, respectively (Liao, et al. 2014; Ghosh and Chan 2016). Lowly expressed genes with fewer than 5 counts were removed from the analysis.

2.4.2 Identifying mCGs and MIs

Individual mCGs were identified using Bis-Class (Huh, et al. 2014), and we used a custom script for finding methylation islands based on individual mCGs. The process of identifying MIs is as follows:

1. Scaffolds are scanned in a 5’ to 3’ direction in 200bp windows. Each window is evaluated for its fraction of mCG which is calculated as the number of mCGs divided by the length of the window.
2. If window’s mCG fraction is < 0.02, the algorithm moves to the next downstream mCG which begins the new 200bp window. This process continues until a window has a mCG fraction of >= 0.02.
3. Once this occurs, the window is extended by 50bp and its mCG fraction is re-evaluated. This continues for as long as the mCG fraction remains < 0.02. As soon as the extended window’s mCG fraction falls below 0.02, extension is stopped and the previous mCG is chosen as the end position of the MI. As a result, the start and end of all MIs is always an mCG.
4. The algorithm then restarts at the next mCG, scanning a new 200bp window. Steps 2 and 3 are repeated until the end of the scaffold.

2.4.3 Protein and Amino Acid Conservation Score

ProteinOrtho with default settings was used to create orthologous gene sets (Lehner 2008). Each orthologous gene set including all protein sequences from each species was further analysed to calculate their conservation scores using Clustal-Omega (Sevier, et al. 2016). Individual amino acid conservation scores were calculated using the Jensen-Shannon (JS) divergence, a robust method for calculating protein sequence conservation (Capra and Singh 2007). We applied a linear mixed effects model with amino acid position (inside or outside MI) and the presence of mCGs as main factors along with the gene and species as the interaction and random factors. To avoid biased towards extremely short

proteins, we only included genes with at least five amino acids for each category in the analysis.

2.5 Acknowledgements

This study was supported by the National Science Foundation grant (MCB-1615664) and funds from the Georgia Institute of Technology to S.V.Y.

CHAPTER 3. *WOLBACHIA*-MEDIATED ASEXUALITY IS LINKED TO DISTINCT EPIGENOMIC AND TRANSCRIPTOMIC CHANGES

3.1 Introduction

Wolbachia is a highly successful and widespread endosymbiont that is estimated to infect 40-60 percent of all insect species (Hilgenboecker, et al. 2008; Zug and Hammerstein 2012). Its infection brings about wide ranging effects on its host fitness, including reproductive parasitism (Werren, et al. 2008). In the *Trichogramma* parasitoid wasps, *Wolbachia* induces parthenogenesis where female hosts convert to reproduce asexually (Stouthamer, et al. 1990; Stouthamer, et al. 1993; Stouthamer and Werren 1993). Typically, uninfected males develop from unfertilized haploid eggs while females result from fertilized, diploid eggs. Wasps that are infected with *Wolbachia* give rise to diploid female offspring through a fertilization-independent mechanism, spreading this infection along with its reproductive phenotype throughout the population (Stouthamer, et al. 2010). Some *Trichogramma* wasps become entirely dependent on *Wolbachia* to reproduce female offspring – these wasps are no longer able to fertilize their eggs, and cannot produce female offspring without *Wolbachia*-mediated diploidization (Stouthamer, et al. 2010). This scenario has been described as "symbiont addiction", where the infection leads to an evolutionary dependency on *Wolbachia* (Bennett and Moran 2015; Sullivan 2017).

Despite knowledge of *Wolbachia*'s ubiquity and ability to completely transform host reproductive physiology, the mechanisms surrounding the manipulation of its host and

induction of parthenogenesis are still poorly understood. Genes related to *Wolbachia*'s prophage are known to be responsible for cytoplasmic incompatibility (Beckmann, et al. 2017; LePage, et al. 2017; Lindsey, Rice, et al. 2018) and male-killing, but the strain infecting *Trichogramma* lack a prophage (Gavotte, et al. 2007; Lindsey, et al. 2016) and orthologs to genes known to manipulate reproductive behavior (Lindsey, et al. 2016). Despite this lack of knowledge, we do know that *Wolbachia* in *Trichogramma* arrests unfertilized eggs in the first mitotic division and prevents chromosome segregation (Stouthamer and Kazmer 1994).

One potential lead into the mechanism of parthenogenesis induction is *Wolbachia*'s manipulation of the host epigenome. It has been speculated that *Wolbachia* is capable of changing the host's heritable epigenetic modifications, especially DNA methylation and histone modifications (Bernstein, et al. 2007). For instance, in the fly *Drosophila simulans*, *Wolbachia* has been shown to modify chromatin reorganization during spermatogenesis (Harris and Braig 2003). Recently, there has been evidence of *Wolbachia* manipulation of the host epigenetic machinery in *Aedes aegypti* (Ye, et al. 2013; Zhang, et al. 2013), *Drosophila melanogaster* (Bhattacharya, et al. 2017), and *Cotesia plutellae* (Kumar and Kim 2017). While these studies indicate that *Wolbachia* may play a role in modifying host epigenetic systems, investigating this question on a genome level is difficult for several reasons. First, current insect model organisms such as flies have little to no genomic DNA methylation (Bewick, et al. 2017). Second, epigenetics are influenced by the underlying DNA sequence (Keller, et al. 2016; Yi 2017) and therefore it is necessary to separate the effects of the infection and the genetic background.

Trichogramma wasps, unlike flies, has a fully functioning DNA methylation system and genomic CpG methylation (Lindsey, Kelkar, et al. 2018). Despite this, there are several challenges when it comes to studying the effects of *Wolbachia* infection on the host epigenome. First, they are geographically widespread and therefore genetically diverse, thus differences in their methylomes are dependent on their diverse genetic backgrounds. Second, curing many *Wolbachia* infected lines is impossible due to their dependence on *Wolbachia* to reproduce, therefore we are unable to generate both infected and uninfected individuals from the same genomic background. The *Wolbachia* infected Trichogramma used in this study reproduce sexually at a reduced rate, where they are unable to maintain a self-sustaining population through fertilization. This does, however, enable us to introgress the genome of a sexually reproducing line into the cytoplasm of a *Wolbachia* infected cytoplasm via back-crossing multiple generations. With each generation, more and more of the sexual genome is introduced, eventually completely replacing the asexual genetic material and creating a line that is *Wolbachia* infected yet is able to be cured of the infection. These cured individuals are therefore genetically identical to the infected hybrids, allowing us to for the first time directly compare their epigenomes and transcriptomes in a genetically homogenous environment.

3.2 Results

3.2.1 Introgressing a Sexual Nuclear Genome into an Asexual Cytoplasm Infected with Wolbachia

For our introgression scheme, we used a total of four isofemales lines of *Trichogramma pretiosum* – two naturally sexually reproducing lines (“CA29” and “CA9”) and two *Wolbachia* infected, parthenogenesis lines (“Insectary” and “ES865”). We introgressed one uninfected genome into one *Wolbachia* infected cytoplasm – the CA29

genome into Insectary cytoplasm, and the CA9 genome into the ES865 cytoplasm (Figure 3.1A). The introgression pairs were determined based on the ability to track an introgression molecular marker (Methods). With each successive introgression generation, the fecundity of the hybrids decreased as expected given the increased cyto-nuclear incompatibilities (Figure 3.1B; GLM: Insectary: $\chi2 = 33.701$, $P < 0.0001$; ES865: $\chi2 = 44.372$, $P < 0.0001$). Over the entire introgression procedure, the sex ratios did not significantly change in the offspring produced by the *Wolbachia* infected females, an indicator of successful introgression (Figure 3.1C; GLM: Insectary: $\chi2 = 1.527$, $P = 0.2166$; ES865: $\chi2 = 2.943$, $P = 0.0862$). One of the pairs, the CA9 X ES865 cross, was less fecund than other which is common in some *Trichogramma* crosses due to disadvantageous cyto-nuclear incompatibilities (Stouthamer, et al. 1990; Stouthamer, et al. 1993; Stouthamer and Werren 1993; Stouthamer and Kazmer 1994). As a result, we used the CA29 X Insectary crosses as the source of our samples. We maintained a total of three independent isofemales lines, each of which were cured of *Wolbachia* following seven generations using antibiotics and subsequently restoring their ability to reproduce sexually. We found no other microbes in these wasps, meaning that the only difference between the cured and infected individuals was the presence of the *Wolbachia* infection. The infection was confirmed in each line using PCR (Methods). We then extract DNA and RNA from the infected and uninfected individuals in each of the three lines for RNA and whole-genome bisulfite sequencing (Methods).

We used the parental genomes (Insectary and CA29) along with the WGBS data of the introgressed hybrids to explore the genomics of the introgression. By using a tool to identify single nucleotide polymorphisms from WGBS data (Gao, et al. 2015) , we were

able to determine if the origin of each SNP was from the paternal (introgressed) or maternal (non-introgressed) parent. This approach allowed us to estimate the amount of non-introgressed genome in the generation seven hydrbids to assess the efficiency of introgression. Our results indicate that all three introgressions were extremely efficient, with two lines (B and C) showing greater than 99% introgression. Line A was less efficient, retaining about 5-8% of the original asexual genome. For unbiased comparisons in our analyses, we excluded the large amount of non-introgressed regions from Line A, although we obtained similar results regardless if these regions were included or not.

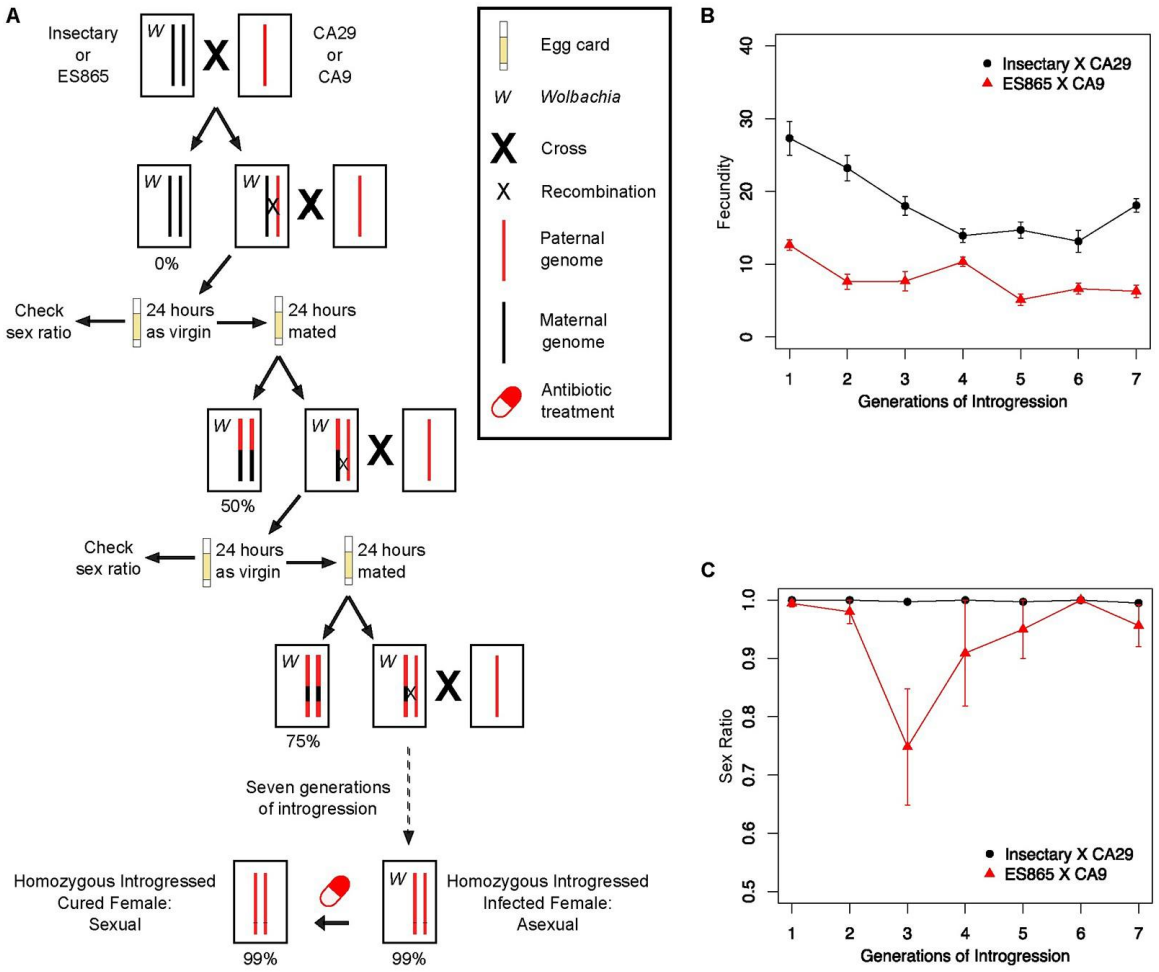

Figure 3.1 – **Introgression scheme used to create genetically homogeneous lines of *Wolbachia* infected and free *Trichogramma*.** A) We estimate that 95-99% of the asexual genome was replaced with the sexual genome after seven generations of introgression. We screened virgin wasps in each generation for sex ratio (proportion of female offspring) and fecundity prior to mating. This scheme was performed 3 times to create 3 isofemale lines. B) Wasp fitness and C) the efficiency of parthenogensis in each generation.

3.2.2 Wolbachia Infection Results in DNA Methylation Changes in T. pretiosum

Our first analysis compared genome-wide methylation changes between infected and uninfected wasps at CpG sites. In total, 106,475 cytosines were methylated (mCGs) in at least one sample (Huh, et al. 2014). Of all the mCGs, we found a total of 340

differentially methylated positions (DMPs) (FDR-adjusted Q < 0.05). 317 were found within gene bodies with the other 23 DMPs being intergenic. The majority of DMPs (238, or 70%) were hypermethylated in the infected wasps, meaning that their levels of fractional methylation were higher compared to the uninfected individuals (Figure 3.2A). The 317 genic DMPs were distributed across 84 genes, which we defined as "differentially methylated genes" (DMGs). These DMGs were enriched for functions relating to embryonic axis specification, pattern specification, and oocyte development, which is concordant with speculation that *Wolbachia* is at least in part manipulating egg development and cell division mechanics by targeting the host epigenome (Medzhitov, et al. 1997; Sun, et al. 2004).

3.2.3 Gene Expression and Exon Usage is Associated with Wolbachia Infection

We next performed differential expression analysis using a negative binomial generalized linear model (Love, et al. 2014b) and identified 59 differentially expressed genes (DMGs; FDR Q < 0.05; Figure 3.2B). 45 (76%) of DMGs were up-regulated in the infected group ($\chi 2$ test, $P < 10^{-15}$) with an average of 4.72-fold change compared to the cured group. These DMGs were not enriched for any gene ontology terms, mostly because the majority of these genes were functionally unannotated. In fact, 35 of the 59 DEGs were specific to the *Trichogramma* lineage (Lindsey, Kelkar, et al. 2018), suggesting that *Wolbachia* infection may be inducing a host-specific response, or potentially a host-specific method of manipulation by *Wolbachia*.

We also looked to determine whether exon usage differed between the infection groups using a generalized linear model (Anders, et al. 2012). In total, 685 genes containing

1,012 exons were classified as differentially used exons, although once again these genes were not enriched for any gene ontology terms.

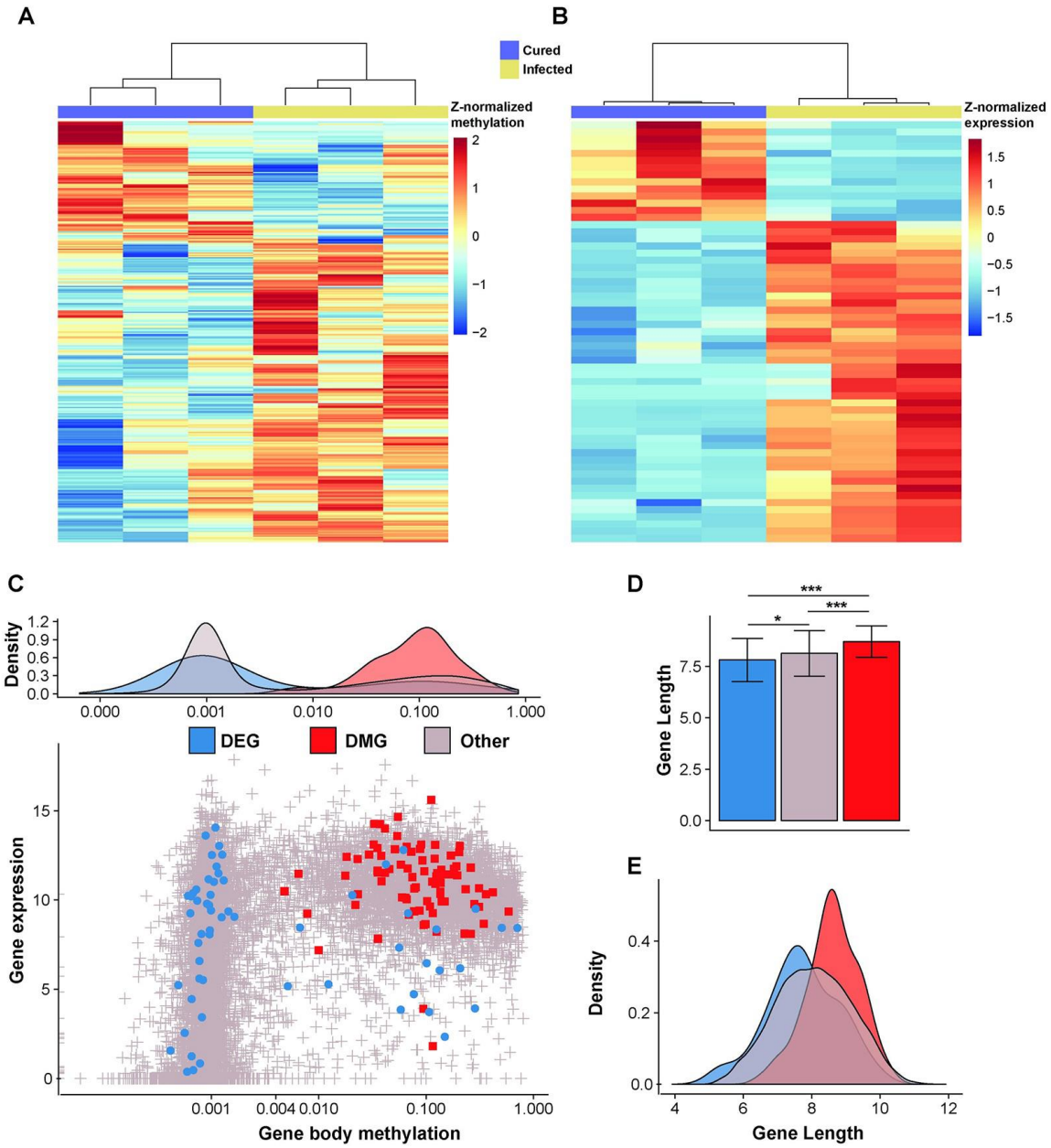

Figure 3.2 – **Comparing methylation and expression between *Wolbachia* infected and uninfected wasps.** A) Heatmap of 340 differentially methylated positions, most of which (239/340) were hypermethylated in the infected wasps compared to the uninfected wasps. B) 59 differentially expressed genes, 39 of which were up-regulated in the infected wasps. C) Gene body methylation (log_{10} transformed) and gene expression (log_2 transformed) for DMGs, DEGs, and the rest of the genes in the genome. The expecte bimodal gene body

methylation distribution is shown above. D) Gene length and E) gene length densities for each gene classification.

3.2.4 Differential Exon Usage but Not Differential Expression is Associated with Differential Methylation

Despite changes to both methylation and expression as a result of *Wolbachia* infection, we found no overlap between DEGs and DMGs. However, there was some concordance in the direction of change in both of these processes. 32 of the 39 genes that were up-regulated in the infected wasps also had higher, but not statistically significant methylation. Gene body methylation has also been shown to regulate expression variability, typically by reducing transcriptional noise (Bird 1995; Huh, et al. 2013). Based on our previous analyses, we expected infected wasps to have lower transcriptional noise due to an overall increase in methylation. We tested this hypothesis by constructing a linear model using transcriptional noise (coefficient of variation of gene expression (Huh, et al. 2013)) as the response variable and gene body methylation, gene expression, gene length, and infection status as explanatory variables (Figure 3.3A). Our results indicate that *Wolbachia* infected wasps do indeed have lower transcriptional noise compared to uninfected wasps (Figure 3.3B and 3C), even when the increased DNA methylation is taken into account.

We also tested to see whether differential methylation was associated with differential exon usage since one potential role of DNA methylation is regulating alternative splicing (Ding, et al. 2016; Arsenault, et al. 2018; Li, et al. 2018). In our list of differentially used exons, only 5 overlapped with DMPs. However, this overlap was

statistically significant due to the low number of DMPs genome-wide (Odds ratio = 4.40, Fisher's exact test, P = 0.0071). Furthermore, of the 685 genes containing differentially used exons, 14 overlapped with DMGs which was also statistically significant (Fisher's exact test, Odds ratio = 3.29, $P = 3.14 \times 10^{-4}$). While the number of overlaps between differential exon usage and differential methylation is low, the fact that the overlaps are statistically significant supports the role of methylation in alternative splicing (Flores, et al. 2012; Foret, et al. 2012; Lev Maor, et al. 2015). Figure 3.4 depicts two examples of such overlap.

A

Predictor	β estimate	t value	Significance
Intercept	2.10	103.01	$< 10^{-15}$
Gene body methylation	-0.43	-23.24	$< 10^{-15}$
Gene expression	-0.16	-205.91	$< 10^{-15}$
Gene length	-0.017	-6.69	2.3×10^{-11}
Infection status	-0.27	-56.01	$< 10^{-15}$

Adjusted $R^2 = 0.74$

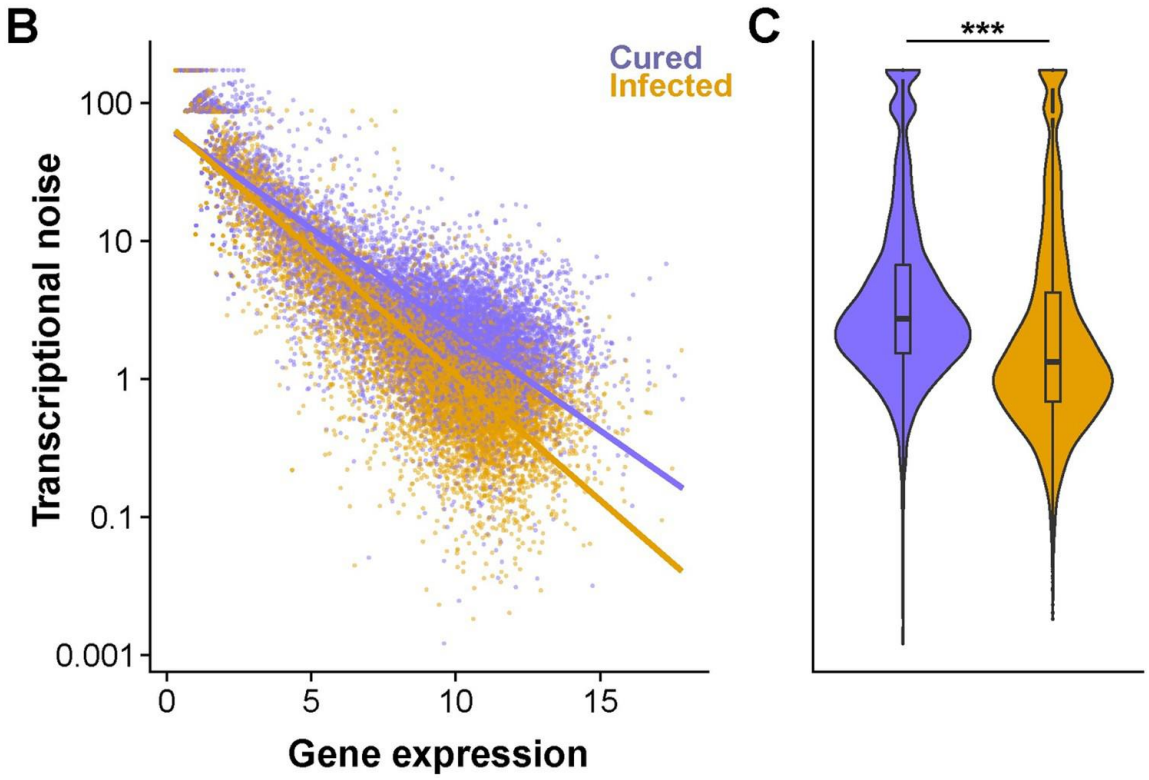

Figure 3.3 – **Transcriptional noise and *Wolbachia* infection.** A) Linear model results using transcriptional noise (coefficient of variation of gene expression) as the response vector and gene body methylation, expression, length, and *Wolbachia* infection status as explanatory variables. B) Infected wasps have lower transcriptional noise than uninfected wasps. C) Violin plot comparing significant differences in transcriptional noise between the two infection groups (Student's t-test, $P < 10^{-15}$).

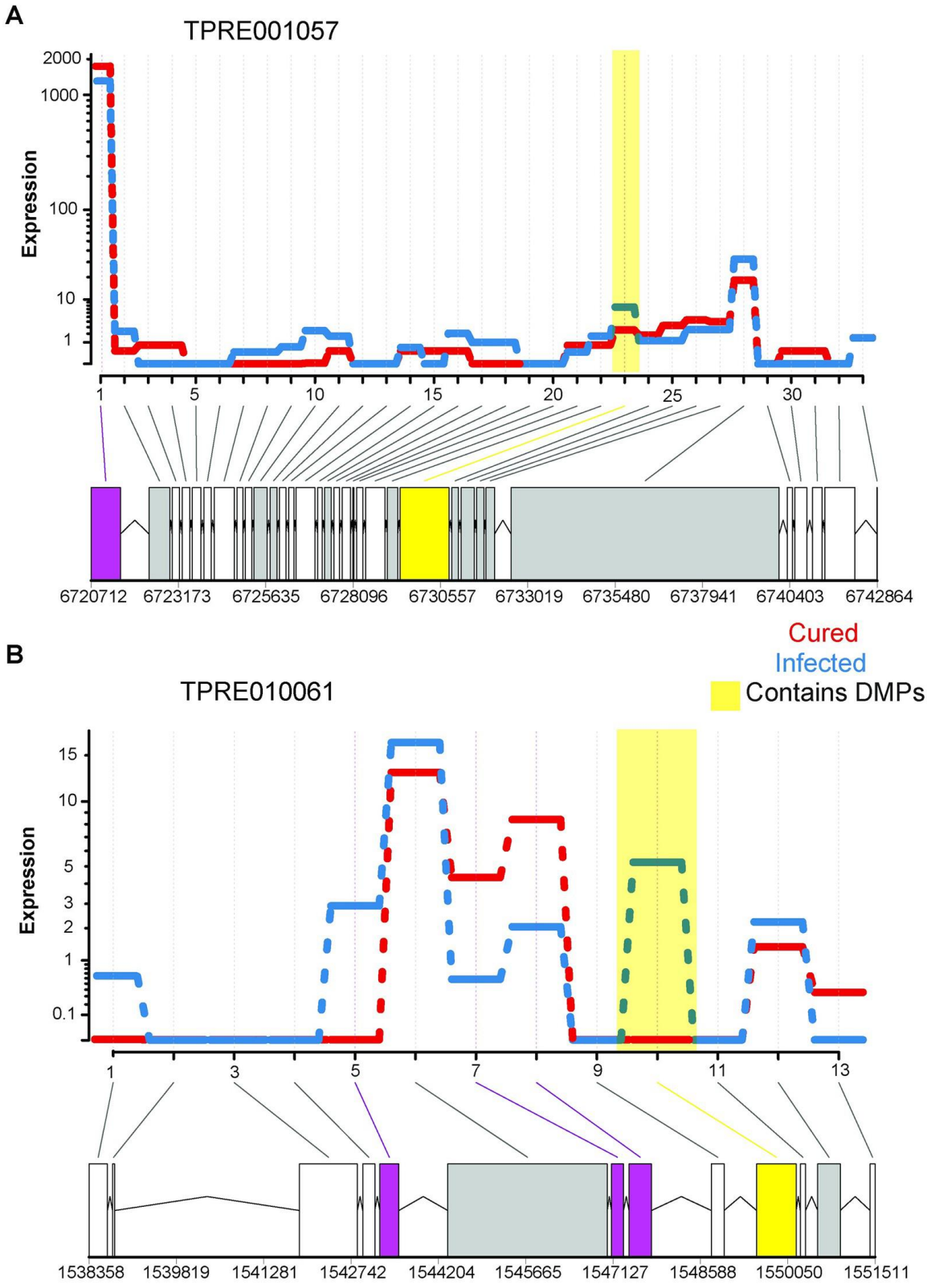
A
TPRE001057
Expression
2000
1000
100
10
1
1
5
10
15
20
25
30
6720712
6723173
6725635
6728096
6730557
6733019
6735480
6737941
6740403
6742864
B
Cured
Infected
Contains DMPs
TPRE010061
Expression
15
10
5
3
2
1
0.1
1
3
5
7
9
11
13
1538358
1539819
1541281
1542742
1544204
1545665
1547127
1548588
1550050
1551511

Figure 3.4. **Two example genes that contain both differentially used exons and differentially methylated positions.** Purple boxes represent differentially used exons and those that also contain DMPs are highlighted in yellow. A) An ortholog of *D. melanogaster CG14299* with 2 differentially used exons (exon 1 and 23). Exon 23 also contains 6 DMPs. B) An ortholog of *D. melanogaster Mzt1*, with 4 differentially used exons (exons 5,7,8,and 10). Exon 10 contains 8 DMPs.

3.3 Discussion

Wolbachia's successful and widespread infection of wasps presents an interesting and useful model for studying molecular mechanisms behind infection and reproductive manipulation. Here, we use a *Wolbachia*-mediated parthenogenesis system, controlled for genetic background by a clever introgression scheme, to describe major methylome and transcriptome changes that accompany a drastic change in reproductive physiology. Our system comes with the major advantage of controlling for differences in genetic background (Keller, et al. 2016; Yi 2017) by creating two genetically homogeneous groups as well as using an organism that has global DNA methylation (Lindsey, Kelkar, et al. 2018).

From our system, we saw global changes in both DNA methylation and gene expression as a result of *Wolbachia* infection. On the methylation side, we found 340 DMPs spread across 84 genes. This number compares favorably with genes associated with *Wolbachia* infection in *A. aegypti* (Ye, et al. 2013) and a viral infection in *A. mellifera* (Galbraith, et al. 2015). This overall pattern from several insect species suggest that perhaps only a small subset of the genome is subject to changes in DNA methylation in

response to an outside infection. Humans, in comparison have an even smaller number of genes change in methylation as a result of disease, estimated to be around 0.5% (Liu, et al. 2013; Dayeh, et al. 2014; Mendizabal, et al. 2019). Differentially expressed genes tend to be evolutionarily conserved and are enriched in functions related to egg maturation and cell division. These functions support the role of *Wolbachia* acting as a disruptor of chromosome segregation and arresting the egg in mitosis (Lindsey, et al. 2016).

In contrast to DMGs, differentially expressed genes had completely different characteristics. DEGs tended to be unmethylated following *Wolbachia* infection, and have unknown functions due to being specific to the *Trichogramma* lineage (Lindsey, Kelkar, et al. 2018). This suggests that *Wolbachia* may induce host-specific responses to infection and may explain the lack of horizontal transfer out of the *Trichogramma* lineage (Raychoudhury, et al. 2009). Even though there was no direct link between differential methylation and expression at the gene level, our study did discover potential relationships between these two processes. At the genome level, we saw an overall increase in both global DNA methylation and transcription which mirrors the pattern of viral infection in honey bees (Galbraith, et al. 2015). Additionally, infection reduced gene expression variability, or transcriptional noise, although it is unclear what the mechanism behind this observation is. We also showed that expression at the exon level was significantly altered as a result of *Wolbachia* infection, and that these differentially used exons tended to contain DMPs. This observation supports previous studies that link DNA methylation to roles in regulating alternative transcripts and splicing (Li-Byarlay, et al. 2013; Galbraith, et al. 2015; Arsenault, et al. 2018). One potential pitfall of our experimental design is the pooling of individuals used for our data, though it was necessary due to the extremely small size of

the wasps. As a result, our samples were heterogeneous and therefore may have diluted methylation and expression signals.

3.4 Methods

3.4.1 Rearing of Trichogramma lines

Trichogramma pretiosum colonies were kept in 12 x 75 mm glass tubes and incubated in 24 ˚C with a 16:8 hour light:dark cycle. Four isofemales lines were used here. The "Insectary" line originates from Peru and has been kept since 1966 (Lindsey, et al. 2016) while the "ES865" line started in Hawaii in 2011. Both lines are infected with *Wolbachia* that induces parthenogenesis and have been resistant to curing by antibiotics (using rifampicin) to restore sexual reproduction (Russell and Stouthamer 2011). The other two lines, "CA29" and "CA9" are highly inbred and come from females collected in California in 2008. Neither of these two lines are infected with *Wolbachia*.

3.4.2 Introgression of Sexual Genome into Wolbachia Infected Cytoplasm

We Introgressed the CA9 genome into the ES865 cytoplasm and the CA29 genome into the Insectary cytoplasm (Figure 3.1A). Females from the *Wolbachia* infected cytoplasm were crossed with uninfected males which produced female hybrids that were heterozygous. There female hybrids were then backcrossed with the original uninfected male strain, a process that was repeated for a total for seven generations. A total of 3 independent isofemale lines were created using this introgression scheme. After three generations, individuals in each line were split, with one being cured of the *Wolbachia* infection using rifampicin (Stouthamer, et al. 1990). Cured wasps were allowed to

"recover" from the effects of antibiotics for three generations prior to being used for sequencing.

3.4.3 Nucleotide Extractions

Newly emerged wasps of less than 48 hours were collected and sex sorted based on antennal morphology. Approximately 500 females were used for each biological replicate for a total of six samples – three infected and three uninfected replicates. The pools were then homogenized and split evenly for DNA and RNA extraction using Qiagen DNeasy and RNeasy kits, respectively.

3.4.4 RNA Sequencing

RNA-seq libraries were created using NovoGene based on the standard eukaryotic workflow. Final library quality and quantity was assessed using the Agilent 2100 Bioanalyzer and Qubit 2.0, respectively (Panaro, et al. 2000; Mardis and McCombie 2017). Libraries were then multiplexed and sequenced on the Illumina HiSeq 4500 platform with 150 paired-end reads.

3.4.5 Genome Sequencing

Genomic libraries were prepared using a modified version of an illumina compatible protocol (Urich, et al. 2015). DNA was extracted and fragmented using the Covaris machine using a 200bp target peak size protocol. The size selection was performed according to a previous protocol (Urich, et al. 2015).

3.4.6 Whole-genome Bisulfite Sequencing

We used a previously published protocol to create our WGBS libraries (Urich, et al. 2015). Bisulfite treatment was performed using the MethylCode Bisulfite conversion kit (Life technologies). DNA was treated with CT conversion reagent for 10 minutes and 10ng of unmethylated lambda phage DNA was added as control. Libraries were diluted and sequenced on the Illumina HiSeq X machine for 150bp paired-end reads, yielding between 100-200 million reads per sample.

3.4.7 Creation of Alternative Reference Genome

The GATK best practices pipeline (Urich, et al. 2015) was used to detect high quality SNPs with confidence in the CA29 line and added to the published *Trichogramma* reference genome (from the Insectary line) (Lindsey, Kelkar, et al. 2018). This alternative reference genome was used for subsequence alignment of WGBS and RNA-seq data.

3.4.8 RNA-seq Analysis

Reads were trimmed for low quality and adapters using Trimmomatic v.0.35 (Bolger, et al. 2014b). They were then mapped to the alternative reference genome using the CA29 SNPs (see above) (Lindsey, Kelkar, et al. 2018) with tophat2 v. 2.2.1 (Kim, et al. 2013). Gene counts were generated using HTSeq (Anders, et al. 2015b) and differential expression analysis carried out using DESeq2 (Love, et al. 2014b). Gene expression was measured by the normalized count generated using the "estimateSizeFactors" function from DESeq2.

Differential exon usage was performed using the DEXseq (Anders, et al. 2012) package. Expression at the exon-level was quantified with their raw counts and normalized

using the “estimateSizeFactors” function. Differential exon usage was modeled based on the following linear model: Exon count ~ sample + exon + infection status:exon. Exons significance was assessed at the FDR < 0.05 (Benjamini and Hochberg 1995) level.

3.4.9 Analysis of Transcriptional Noise

We used the percent coefficient of variation of gene expression to measure transcriptional noise (Huh, et al. 2013), which was used as the response variable in the following linear regression model: $\log_{10}$(transcriptional noise) ~ gene body methylation + $\log_2$(gene expression) + $\log_{10}$(gene length) + *Wolbachia* infection status. The linear model was performed in R version 3.3.2 (R Core Team 2014) using the “lm” function.

3.4.10 WGBS Data Processing

Reads were trimmed to filter out low quality reads and remove adapter sequences using Trim Galore! (Martin 2011). They were then aligned to the alternate reference genome with Bismark using the parameters --score_min L,0,-0.4 (Krueger and Andrews 2011). Additionally, the reads were aligned to the lambda genome (GenBank Accession: J02459.1) as a way of measuring the bisulfite conversion efficiency. Aligned reads were deduplicated and CpG counts from both minus and plus strands were combined. Each CpG was classified as either “methylated” or “unmethylated” using Bis-Class (Huh, et al. 2014).

3.4.11 Using WGBS Data to Analyze Introgressed Regions

To assess the efficiency of introgression, we mapped our WGBS reads to both the paternal and maternal genomes separately. We then used BS-SNPer (Gao, et al. 2015) to call SNPs using WGBS data with stringent parameters to retain high quality SNPs with

confidence. The origin of each SNP was determined by comparing it to the original maternal and paternal genomes, with maternal SNPs considered as non-introgressed. We then labelled putative non-introgressed regions as clusters of maternal SNPs – they started with a maternal SNP and were followed in close succession by additional maternal SNPs within 10kb. Genes and CpGs belonging to non-introgressed regions were removed from subsequent analyses.

3.4.12 WGBS Data Analysis

We retained mCGs that were methylated in at least one of the six samples, leaving 106,475 CpGs for differential methylation analysis (Huh, et al. 2017). We then used RADMeth (logit link) package (Dolzhenko and Smith 2014) to model individual CpGs in a beta-binomial regression to identify CpGs that were differentially methylated between the two infection groups (DMPs). The initial list of DMPs were corrected for multiple testing at a FDR threshold of 0.05 (Benjamini and Hochberg 1995).

3.5 Acknowledgements

The authors would like to thank Robert J. Schmitz and Nicholas A. Rhor for their WGBS protocol and training on library preparation. This work was supported by the National Science Foundation (DEB 1501227 to ARIL, and MCB 1615664 to SVY), the United States Department of Agriculture (NIFA 194617 to RS and NIFA 2016-67011-24778 to ARIL); and Robert and Peggy van den Bosch Memorial Scholarships to ARIL.

CHAPTER 4. LINEAGE AND PARENT-OF-ORIGIN METHYLATION PATTERNS IN A. MELLIFERA USING WHOLE-GENOME BISULFITE SEQUENCING

4.1 Introduction

Several theories have been proposed to explain the origins of parent-specific expression (e.g.,(Patten, et al. 2014)), including Haig's kinship theory of intragenomic conflict (Haig 2000; Pegoraro, et al. 2017). The kinship theory predicts that parent-specific expression arises due to maternal and paternal genes having different selection pressures, such as in a scenario where one female reproduces with multiple males for offspring. In this scenario, matrigenes may favor traits that promote equal survival among siblings whereas patrigenes support traits that focus on individual "selfish" fitness (Haig 2000; Pegoraro, et al. 2017). Evidence for this theory has been reported in mammals and plants, though social insects such as honey bees where it is especially applicable have not yet been studied in this context (Haig 2000; Wilkins and Haig 2003). In a honey bee colony, the vast differences in matrigene and patrigene relatedness among individuals lends itself as an ideal example for studying both kinship theory and its role in regulating social behaviors (Queller 2003; Kocher, et al. 2015; Galbraith, et al. 2016; Pegoraro, et al. 2017).

Previous studies in insects have shown support for the kinship theory (Bonasio, et al. 2012; Lonsdale, et al. 2017). For example, Kocher et al. (2015) showed parent-specific expression patterns across different developmental stages, behavioral states, and tissues. Galbraith (Galbraith, et al. 2016) showed that worker ovary size and activation timing were

dependent on the parental phenotype, an observation that is consistent with predictions of kinship theory (patrigenes should favor worker reproduction). Furthermore, Galbraith et al. (Galbraith, et al. 2016) showed that patrigenes were upregulated compared to matrigenes in reproductive tissues of both reproductive and sterile workers in reciprocal crosses of Africanized and European bees.

Studies supporting the kinship theory, however, failed to address the mechanisms behind parent-specific expression. In other lineages such as mammals and plants, parent-specific expression is primarily regulated via the epigenome and DNA methylation (Reik and Walter 2001; Bird 2002; Queller 2003; Law and Jacobsen 2010). The honey bee does possess a functional DNA methylation system and has genomic CpG methylation, albeit at a much lower frequency than the aforementioned organisms (Wang, et al. 2006; Lyko, et al. 2010). Rather than being ubiquitous through the genome, DNA methylation in honey bees is sparse and almost exclusive to gene bodies and coding regions (Elango, et al. 2009; Lyko, et al. 2010; Galbraith, et al. 2015).

In this study, we take samples from the previous study of reciprocal crosses between Africanized and European honey bees (Galbraith, et al. 2016) to look for signatures of parent-specific methylation using whole-genome bisulfite sequencing (WGBS). Samples consisted of sterile as well as reproductive workers, allowing us to study allelic methylation patterns based on parent, lineage, and reproductive state differences. We can then investigate whether parent-specific methylation exists in honey bees and if it is associated with parent-specific expression.

4.2 Results

4.2.1 Honey Bees Exhibit Both Lineage and Parent-specific DNA Methylation

To study allelic patterns of DNA methylation, we used a list of informative SNPs that allowed us to assign reads based on their allelic origin (Methods). We performed our DNA methylation analysis for each block separately, allowing us to increase the scope of our analysis by using the large amount of SNPs that were unique to each genetic block. In genetic block A, we had 213,056 informative SNPs allowing us to examine 48,745 methylated CpGs (mCGs) and 5,613 methylated genes. In block B, there were 214,504 informative SNPs, overlapping with 41,764 methylated CpGs and 5,359 methylated genes.

We used a linear model to assess each individual mCG and its methylation levels based on variation in parent-of-origin and lineage effects (Methods). The significant mCGs from this model were referred to as differentially methylated positions (DMPs) and summarized in Table 4.1 based on their bias. Figure 4.1 shows examples of DMPs showing both types of allelic methylation biases.

Table 4.1 – Summary of DMPs in each block and reproductive state based on their direction of allelic bias.

	Block A		Block B	
	Sterile	**Reproductive**	**Sterile**	**Reproductive**
Parent-of-origin				
Maternal bias	132	190	208	189
Paternal bias	148	218	216	188
Lineage				
Africanized bias	333	921	696	727
European bias	410	948	829	964

The strongest factor affecting DNA methylation was the lineage effect, which was the effect due to either Africanized or European alleles. In genetic block A, 743 mCGs showed lineage-specific methylation in sterile workers and 1,868 showed lineage-specific methylation in reproductive workers. In genetic block B, 1,525 mCGs showed lineage-specific methylation in sterile workers and 1,691 showed lineage-specific methylation in reproductive workers (Table 4.1). We also saw a greater number of European biased mCGs compared to Africanized biased mCGs in both genetic blocks and reproductive statuses. In all cases other than reproductive workers in block A, these differences were statistically significant (Table 4.1; X^2 test, $P < 0.05$).

There were also hundreds of mCGs that displayed parent-specific methylation effects (Table 4.1 and Figure 4.2; Figure 4.1). In block A, there were 280 DMPs showing parent-of-origin effects in sterile workers (132 maternal and 148 paternal; Table 4.1 and Figure 4.2). In the reproductive workers, we saw a total of 408 parent-of-origin DMPs (190 maternal and 218 paternal; Table 4.1 and Figure 4.2). The increase in paternal biased DMPs was a significant increase over maternal biased DMPs (X^2 test, $P < 0.01$; Table 4.1 and Figure 4.2). In block B, we saw 208 maternal biased DMPs and 216 paternal biased DMPs in sterile workers as well as 189 maternal and 188 paternal biased DMPs in the reproductive workers (Table 4.1 and Figure 4.2). In all allelic bias categories, we observed a greater number of DMPs in the reproductive workers compared to the sterile workers (X^2 test, $P < 0.05$ for all directions of bias) in genetic block A but for none of the categories in block B.

We found significant overlaps of DMPs between workers of different reproductive states. 69 parent-of-origin DMPs overlapped between sterile and reproductive workers in

block A while 119 parent-of-origin DMPs were shared in block B. Both overlaps were highly significant enrichments compared to a null expectation of no association (Fisher's exact test, $P < 0.01$ for both comparisons). However, a large number of DMPs were still specific to each reproductive state. In block A, 211 and 339 parent-of-origin DMPs were specific to sterile and reproductive workers, respectively. In block B, 305 parent-of-origin DMPs were specific to sterile workers and 258 parent-of-origin DMPs specific to reproductive workers. Furthermore, 189 sterile-specific and 191 reproductive-specific DMPs are shared across blocks which is also a highly significant overlap in both cases (Fisher's exact test, $P < 0.01$ for both comparisons). These overlaps suggest common, robust factors affecting genome-wide DNA methylation that are independent of reproductive status and genetic block.

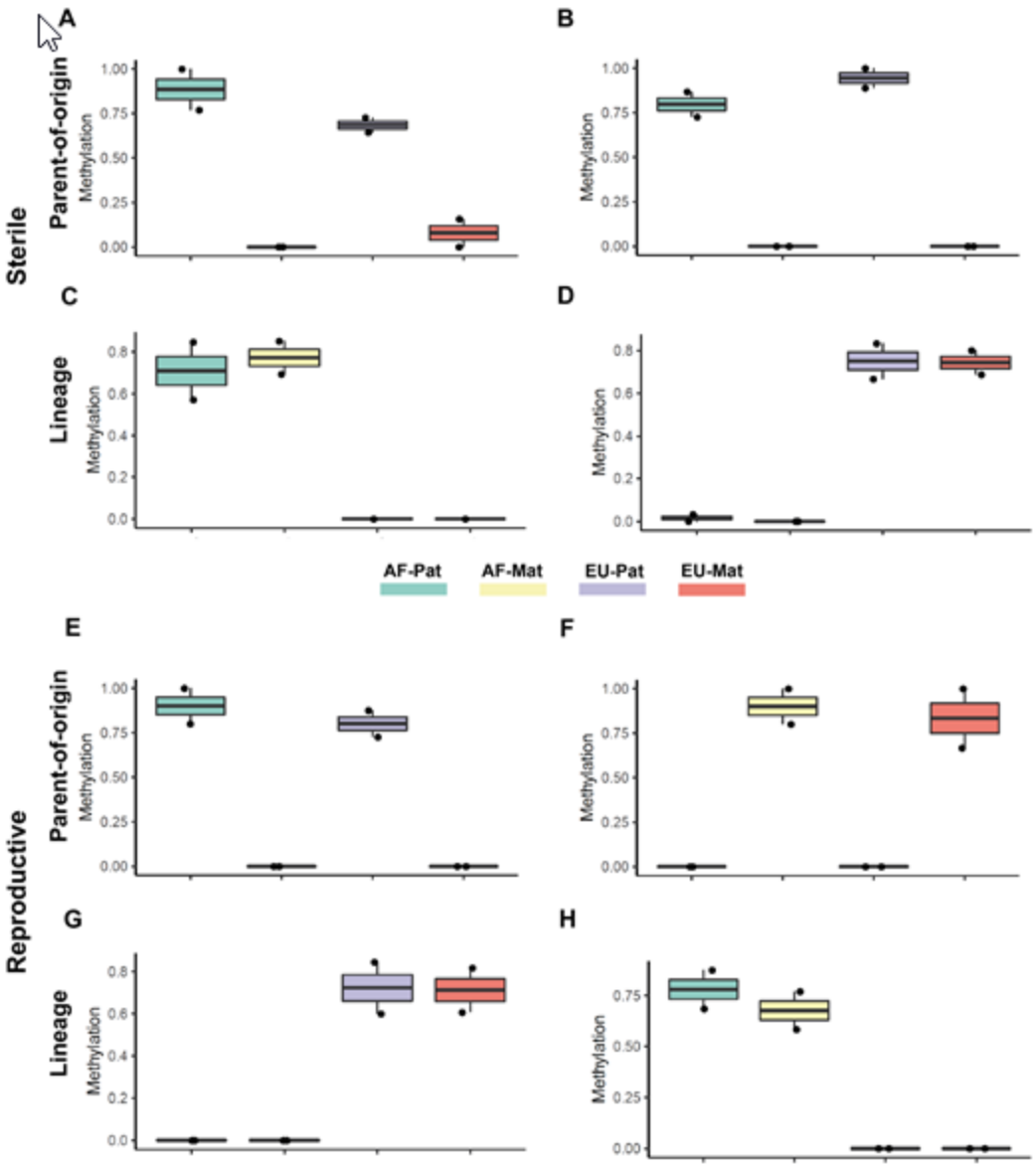

Figure 4.1 – **Examples of mCGs showing parent-of-origin and lineage effects.** A) and B) DMPs showing parent-of-origin bias in sterile workers. C) An example of Africanized biased DMP and D) European biased DMP in sterile workers. E) and F) show paternal and maternal biased DMPs, respectively. G) Lineage biased DMP in reproductive workers and H) DMPs biased towards Africanized and European workers.

4.2.2 Genes with Signatures of Parent-specific Methylation

Genes containing DMPs showing the same direction of allelic methylation bias were defined as differentially methylated genes (DMGs)(Methods). For example, parent-of-origin DMPs in block A were found across 179 and 230 genes in the sterile and reproductive workers, respectively, and these genes are subsequently referred to as parent-of-origin differentially methylated genes (Table 4.2). Interestingly, the majority of parent-of-origin DMGs contained just a singular DMP (sterile average: 1.21 DMPs; reproductive average: 1.24 DMPs).

Table 4.2 – DMGs for all directions of allelic bias based on genetic block and worker reproductive status.

	Block A		Block B	
	Sterile	**Reproductive**	**Sterile**	**Reproductive**
Parent-of-origin				
Maternal bias	82	113	140	127
Paternal bias	97	117	126	106
Lineage				
Africanized bias	165	313	258	259
European bias	201	314	293	321

NOTE.—DMGs contain DMPs that show the same direction of allele-specific bias.

To take advantage of the information provided by the two different genetic blocks, we combined DMGs from both blocks for gene ontology (GO), pathway and comparative analyses. GO terms for sterile parent-of-origin DMGs included protein glycosylation, ATP binding functions, and involved in fatty acid degradation. Reproductive parent-of-origin DMGs were enriched for functions involving intracellular protein transport and mRNA surveillance pathways.

We observed moderate but significant overlaps between parent-of-origin DMGs of the two reproductive states in both blocks. Thus, these were the genes which showed parent-of-origin effects in both sterile and reproductive workers. Specifically, there were 16 DMGs showing maternal bias (Fisher's exact test, $P < 0.01$) and 30 DMGs showing paternal bias (Fisher's exact test, $P < 0.01$) overlapping between sterile and reproductive workers in block A. In block B, there were 45 maternal DMGs (Fisher's exact test, $P < 0.01$) and 35 paternal DMGs (Fisher's exact test, $P < 0.01$) overlapping between sterile and reproductive workers. Though none of the overlapping gene sets were enriched for specific GO terms, they nevertheless mirrored the DMP results and reinforce the idea of a common set of genes that are differentially methylated due to parent-of-origin effects.

Interestingly, there was significant overlap between genes showing lineage differential methylation and parent-of-origin differential methylation (Figure 4.3). We found 46 DMGs exhibiting both lineage and parent-of-origin biases in block A sterile workers (Fisher's exact test, $P < 0.01$), and 83 DMGs showing both biases in reproductive workers (Fisher's exact test, $P < 0.01$; Figure 4.3). In block B, sterile workers and reproductive workers had 96 and 83 genes belonging to lineage and parent-of-origin DMGs. Functions of genes that show both types of allele-specific methylation did not deviate from the enriched GO terms of their respective reproductive states, which were generally focused on cell energy metabolism and signal transduction. Since these genes exhibit both lineage and parent-of-origin differential methylation, they may be particularly labile in terms of allele-specific methylation.

We next examined parent-of-origin DMGs that were unique to sterile and reproductive workers to investigate the relationship between parent-specific methylation

and reproductive phenotype. There were a total of 133 sterile-specific DMGs and 184 reproductive-specific DMGs in block A and 266 sterile-specific DMGs and 233 reproductive-specific DMGs in block B. 12 such DMGs were commonly found in sterile workers of both blocks whereas 22 DMGs were common between the reproductive workers (Fisher's exact test, $P < 0.05$ for both comparisons). While these overlaps were statistically significant, they did not exhibit any significant functional enrichment in our analysis, likely due to the small number. In comparison, DMGs specific to sterile workers in block A were enriched for GO terms associated with protein deubiquitination while reproductive-worker specific DMGs were enriched for functions such as mRNA surveillance pathway and hydrolase activity. For block B, sterile-specific DMGs were enriched for GO terms related to protein glycosylation and signal transduction whereas reproductive-specific DMGs showed enriched GO terms such as intracellular transport.

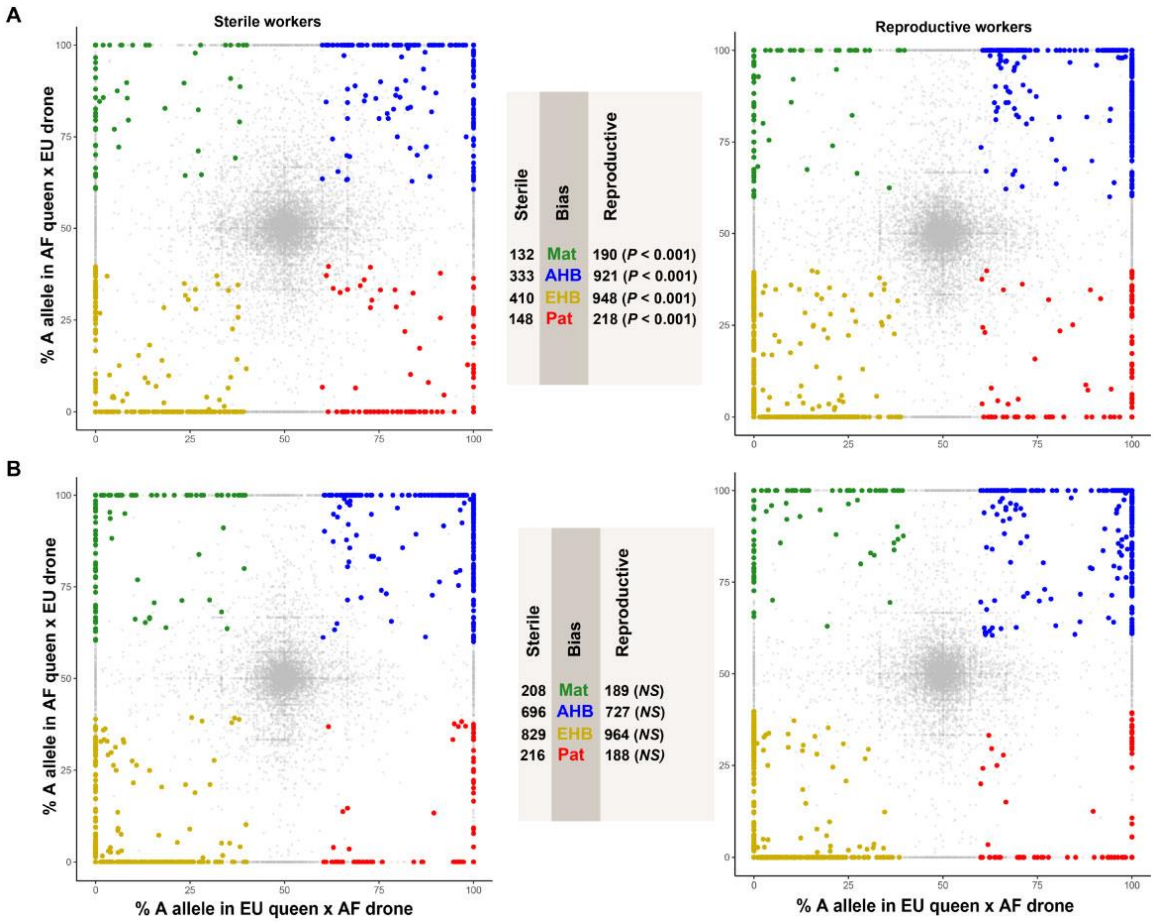

Figure 4.2 – **DMP biases for A) genetic block A and B) genetic block B.** Each dot represent a DMP and is shown as the relative percentage of Africanizedmethylation in each cross (Methods).

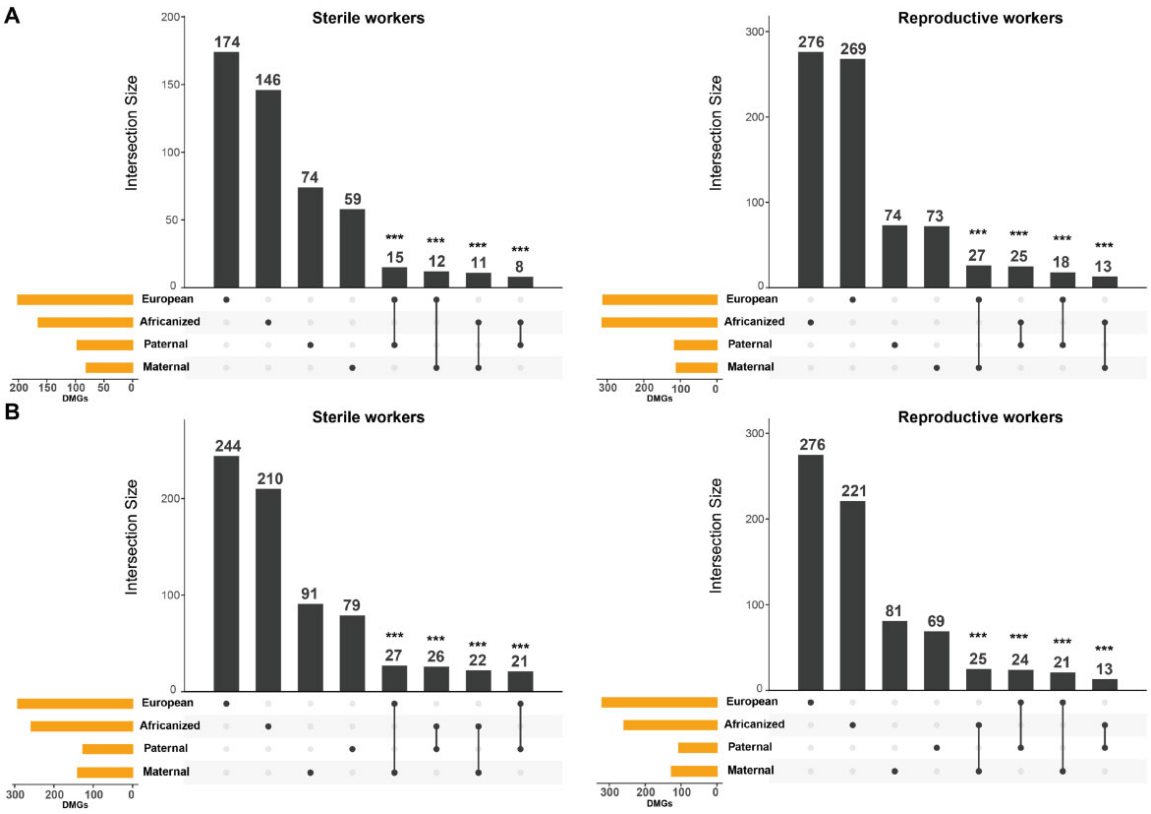

Figure 4.3 – **Number of genes belonging to each bias category based on the worker reproductive state and their overlaps.** A) DMG and overlap summary for genetic block A. B) DMG and overlap summary for genetic block B.

4.2.3 Weak Association Between Allelic Methylation and Expression

We investigated if parent-of-origin expression and methylation were correlated by comparing the previously obtained RNA-seq dataset (Galbraith, et al. 2016) with our current results. The individuals from the RNA-seq study are sisters of the individuals in the current study. To make the results comparable to the methylation results, we re-analyzed the RNA-seq data using the same analysis pipeline as the current study (Methods). Our results recapitulated trends from the previous study, and while the number of genes in each category was different from the original study, they were all subsets of the

genes from Galbraith et al. 2016. For both genetic blocks, there was a significantly more patrigene bias compared to matrigene bias as well as bias towards reproductive workers compared to sterile workers (Fisher's exact test, $P < 0.01$ for all comparisons). Interestingly, we found that differentially expressed genes (DEGs) varying due to parent-of-origin and lineage effects were almost exclusive to reproductive workers in both genetic blocks. Additionally, there was essentially no overlap between allelic DMGs and allelic DEGs in either genetic blocks. In fact, the only overlap we observed was in reproductive workers for the lineage effect in block A and there was a complete lack of overlap for any parent-of-origin genes in both genetic blocks.

4.3 Discussion

Our study uses the power of reciprocal crosses to understand lineage and parent-of-origin effects on genome-wide DNA methylation and how these effects differ between reproductive and sterile workers. We found very strong lineage effects which agrees with many previous studies showing that DNA methylation is highly influenced by the genetic background (Jones 2012; Smith and Meissner 2013; Mendizabal, et al. 2014; Yi 2017). Our analysis also indicates that some of the CpGs in the honey bee genome show variation consistent with parent-of-origin effects. The numbers of DMPs and DMGs showing a parent-of-origin effect were 2-3 fold smaller than those exhibiting lineage effects, indicating that parent-of-origin effect is not as strong as genetic background effects. Nevertheless, the numbers of genes exhibiting parent-of-origin effects range between 3.2 ~ 9.9 % of genes analyzed, similar ranges as observed in mammals (Luedi, et al. 2007;

Ferguson-Smith and Bourc'his 2018). We also observed that many genes harbored both parent-of-origin DMPs and lineage-specific DMPs in both blocks (Figure 4.3). This observation could potentially indicate that some positions or some genes in the honey bee genome tend to be labile in terms of epigenetic modification, and potentially targets of regulation for a many different factors.

Interestingly, we found that, with the exception of the paternal category, there was an increase in both DMP and DMG numbers in the reproductive workers compared to the sterile workers (X^2 test, $P < 0.05$ for all comparisons) in block A. This observation mirrored the increase of parent-of-origin effect in reproductive workers at the level of gene expression (Galbraith et al. 2016). However, in block B, this pattern was not observed (except a modest increase in European biased DMPs, Table 4.1, X^2 test, $P < 0.05$). One possibility is that this difference could have arisen due to the different ages of the workers between the two genetic blocks – though all the reproductive workers were confirmed to have activated ovaries, since workers in block A were 4 days older, they were likely more reproductively mature, which could manifest in clearer DNA methylation difference between worker castes.

Previous work on parent-of-origin gene expression supported the prediction that worker ovary activation was associated with biased expression of patrigenes, with a stronger paternal bias in reproductive workers compared to sterile workers (Galbraith, et al. 2015). Our re-analysis of the RNA-seq data recapitulated this finding, though we did not see the same patterns in our DNA methylation analysis. In terms of the link between DNA methylation and gene expression, we observed almost no overlap between parent-specific gene expression and methylation. This could indicate that either DNA methylation

does not affect parent-of-origin gene expression, or that the effect of DNA methylation is indirect. It is worth noting that studies in insects thus far suggest that differential DNA methylation does not directly correlate with differential gene expression (Galbraith, et al. 2015; Arsenault, et al. 2018; Wu, et al. 2020a). Rather, DNA methylation may affect other aspects of gene expression such as gene expression variability or alternative splicing (Huh, et al. 2013; Hunt, et al. 2013; Wang, et al. 2013; Galbraith, et al. 2015; Arsenault, et al. 2018).

4.4 Methods

4.4.1 Biological Sample Collection

Samples were collected based on the previous study (Galbraith, et al. 2016). We obtained 8 sterile and 8 reproductive workers equally from both genetic blocks and from both types of reciprocal crosses. These samples came from the same crosses as those used for the Galbraith et al. 2016 transcriptomic study. DNA was extracted from the ovaries and abdominal fat bodies for bisulfite sequencing library construction.

4.4.2 WGBS Library Construction and Sequencing

WGBS libraries were made according to a Illumina compatible protocol (Urich, et al. 2015). Bisulfite treatment of genomic DNA was performed using the MethylCode Bisulfite Conversion Kit (Life Technologies, Cat. No. MECOV-50). Finished libraries were diluted and sequenced on the Illumina HiSeq X machine using 150bp paired-end reads.

4.4.3 Creating N-masked Genomes

SNPs for the parents of each cross were from the previous study (Galbraith, et al. 2016). For each cross, we removed ambiguous SNPs and SNPs with a Phred quality score of < 30, as well as C -> T and T -> C SNPs. We also removed any SNPs that had fewer than 5 coverage in either their European or Africanized alleles. Using this stringent filtering criteria, we ended with 213,056 and 214,504 informative SNPs for genetic blocks A and B, respectively. A custom python script was used to generate one N-masked genome for each genetic block based on the final list of informative SNPs.

4.4.4 WGBS Data Processing

Raw reads were trimmed for low quality and adaptors using Trim_galore! (Martin 2011) and aligned to the respective N-masked genome using default Bismark parameters (Krueger and Andrews 2011). We then use SNPSplit (Krueger and Andrews 2016) to assign each read as either European or Africanized origin based on the list of informative SNPs for the genetic block. We then applied the binomial test for each CpG site using the deamination rate as the probability of success and an FDR threshold of < 0.05 (Benjamini and Hochberg 1995) to label each CpG as "methylated" or "unmethylated" (Lyko, et al. 2010; Wang, et al. 2013; Galbraith, et al. 2015). Only CpGs that were methylated in at least one sample were retained for downstream analyses (Huh, et al. 2019).

4.4.5 Differential Methylation Analysis

The DSS package (Park and Wu 2016) was used to find CpGs that were differentially methylated (DMPs). For the model, we included parent-of-origin (either maternal or paternal) and lineage (European or Africanized) as explanatory variables. We applied this model separately for each gene block. Additionally, each significant CpG was

required to exhibit at least 60% relative allele-specific methylation bias in both reciprocal crosses ($European_{mother}$ x $Africanized_{father}$ and $Africanized_{mother}$ x $European_{father}$), similar to previous calculation of allele-specific expression bias (Kocher, et al. 2015; Galbraith, et al. 2016). The relative allele-specific methylation is the percent of fractional methylation (Galbraith, et al. 2015; Lindsey, Kelkar, et al. 2018) of one allele relative to the sum of the fractional methylation of both alleles. Differentially methylated genes (DMGs) for each explanatory variable in the model were defined as genes that contain DMPs that all showed the same direction of bias (Galbraith, et al. 2015; Kocher, et al. 2015).

4.4.6 RNA-seq Processing

We re-analyzed the data from (Galbraith, et al. 2016) using the same pipeline and criteria as the methylation analysis to provide a consistent comparison between the two datasets. Briefly, RNA-seq reads were aligned to their respective N-masked genome HISAT2 and then assigned to an allele using SNPSplit (Krueger and Andrews 2016). HTSeq (Anders, et al. 2015a) with default parameters was used to count the allele-separated reads. We used DESeq2, which applies a similar linear model as DSS, and the same model variables as the methylation analysis to find differentially expressed genes. Significant genes were further corrected for FDR at a threshold of 0.1 (Benjamini and Hochberg 1995).

4.4.7 Gene Ontology

Gene ontology was performed using the DAVID bioinformatics Functional Annotation tool (Huang da, et al. 2009). Enriched GO terms were considered significant at $P < 0.05$ with the background gene list set to all protein coding genes in the honey bee genome.

4.5 Acknowledgements

We thank Tom Glenn (retired, Glenn Apiaries) for generating the honey bee crosses and samples used in this study. This work was supported by the National Science Foundation [grant number MCB-0950896 to C.M.G. and S.V.Y.]

CHAPTER 5. GENE BODY DNA METHYLATION IS ASSOCIATED WITH REDUCED GENE EXPRESSION VARIABILITY

5.1 Introduction

Population-level data on gene expression brings new opportunities to understand genomic factors that associate with variability of gene expression. Gene expression levels may vary between individuals and within cell populations due to several mechanisms, including intrinsic factors such as the rate of transcription and epigenetic regulation (Sanchez and Kondev 2008; Huh, et al. 2013; Sevier, et al. 2016; Wu, et al. 2020b) as well as extrinsic factors such as parasite infection and cell cycle (Fraser, et al. 2004; Sanchez and Kondev 2008; Wu, et al. 2020b).

Previous studies of gene expression variability from wide ranging taxa have discovered that highly expressed genes tend to have reduced variability between individuals (Bird 1995; Choi and Kim 2008; Huh, et al. 2013; Wu, et al. 2020b). It is hypothesized that natural selection has shaped expression variability of highly expressed genes as a means to control for the inherent stochasticity involved in transcription and subsequent protein synthesis, which has been shown to be detrimental to organisms (Fraser, et al. 2004; Wang and Zhang 2011; Barroso, et al. 2018). Genes that are constitutively highly expressed are typically essential housekeeping genes whose noise are therefore minimized by natural selection (Fraser, et al. 2004; Wang and Zhang 2011; Barroso, et al. 2018).

Other traits that were shown to significantly associate with gene expression variability include gene length, presence of a TATA box, initiator motifs, and disease and infection (Huh, et al. 2013; Ravarani, et al. 2016; Faure, et al. 2017; Wu, et al. 2020b). The presence of a TATA box has been shown to have a strong impact on increasing gene expression noise, with other core promoter elements such as initiator motifs and GC motifs being associated with higher gene expression noise to a much lesser degree (Faure, et al. 2017). These observations indicate that genomic features can play significant roles in shaping gene expression variability.

Gene body DNA methylation, which is an ancestral form of epigenetic regulation in animal genomes, is negatively associated with gene expression variability in humans (Huh et al. 2013), indicating that they may reduce transcriptional noise. Studies in insects also supported this observation (Hunt et al. 2013, Wu et al. 2020, Wang et al. 2016). However, the relative contributions of these different genomic features have not been examined systematically in insects. In this study, we aim to elucidate relative contributions and roles of different genomic features on gene expression variability.

In addition, some lineages, notably the order Diptera that includes the model insect *Drosophila melanogaster*, has lost DNA methylation (Sarda, et al. 2012). Given that DNA methylation is implicated in the regulation of gene expression variability, it is of interest to examine whether the patterns of gene expression variability vary between honey bee, from the hymenopteran lineage possessing ancestral gene body methylation, and Drosophila.

5.2 Results

5.2.1 Core promoter elements are significant contributors to gene expression variation

For each dataset, we first modeled gene expression variation, quantified as the coefficient of variation (Huh, et al. 2013; Islam, et al. 2014; Fan, et al. 2016), using a linear model based using the following co-variates: mean gene expression, gene length, presence of a TATA box, and presence of an initiator motif (Methods). Our main motivation was to evaluate the impact of DNA methylation on gene expression variability. However, for data sets in honey bees, matching data on DNA methylation are lacking. Therefore, for honey bee data sets, we included CpG O/E as an additional covariate which is an approximate measure of DNA methylation (Elango, et al. 2009).

Here, we discuss the impacts of gene expression, TATA box, initiator motifs, and gene lengths. The effects of DNA methylation are discussed in a separate section later. As expected, mean gene expression was strongly anti-correlated with gene expression variation and was by far the most significant term with the largest coefficient in the linear model in all datasets (Huh, et al. 2013; Islam, et al. 2014; Fan, et al. 2016; Wu, et al. 2020b) (Figure 5.1A and Table A.1). Following mean expression, the presence of a TATA box in the gene promoter region was a significant term in all but 3 fly datasets (Lindsey et al. 2020, Miozzo et al. 2020, and Thackray et al. 2018) and in all but 2 honey bee datasets (Doublet et al. 2016 and Galbraith et al. 2016; Table A.1). With the exception of one fly study (Lehmann et al. 2020), the TATA box factor was positively correlated with gene expression variation in all datasets in which the term was significant and is consistent with previous findings that have reported that genes with TATA boxes are associated with high noise (Blake, et al. 2003; Lehner 2008; Ravarani, et al. 2016; Faure, et al. 2017). The other

core promoter element, presence of initiator motif, was only significant in approximately half of the studies (6 out of 12 fly studies; 3 out of 8 honey bee studies; Table A.1). The direction of correlation for the initiator motif was also less consistent than the previous two discussed factors, as the coefficient was positive in 4 of the 6 fly datasets it was significant in and in 2 of the 3 honey bee datasets it was significant in (Table A.1). Lastly, gene length, while a significant term in the majority of datasets, also failed to display a consistent direction of correlation in either fly or honey bee datasets. In conclusion, in the linear models, we observed a strong and significant anti-correlation between mean expression and expression variability along with consistent, though not always significant, correlation between the presence of a TATA box and expression variability (Figure 5.1A). The other promoter element, initiator motifs, failed to display a consistent relationship with expression variability.

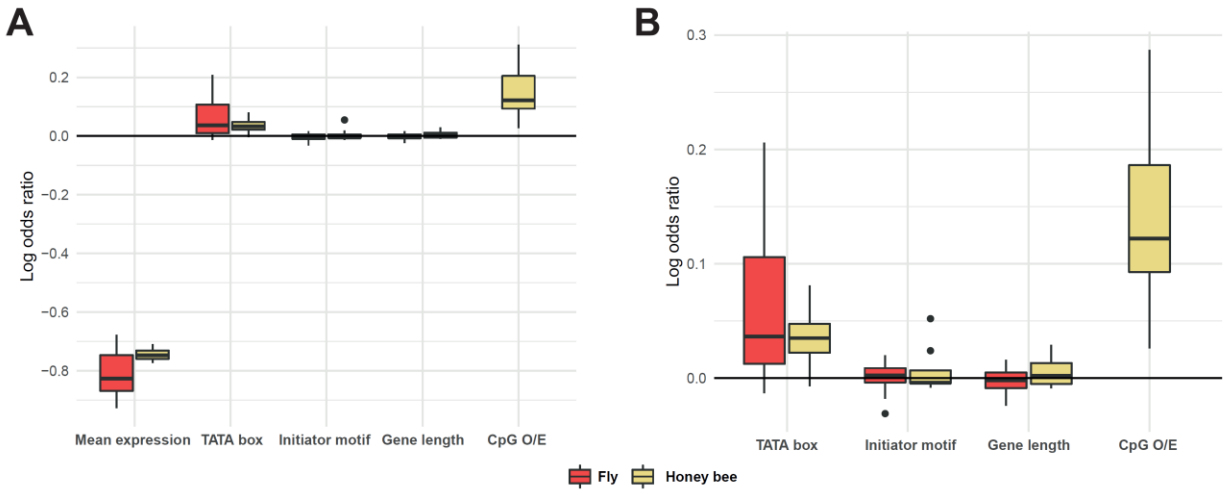

Figure 5.1 – Linear model covariate coefficients. A) Box plot of log ratio of covariate coefficients including mean expression, core promoter elements, gene length, and CpG O/E (for honey bee datasets) from the full linear model. B) Covariate coefficients with mean regressed out using a quadratic model (Methods).

Because of the strong effects of mean gene expression on the linear model, we applied another strategy to control for this effect. We first regressed out mean expression using a quadratic model (Methods). We used the quadratic model as it was shown to have fairly unbiased residual distributions for our data (Figure A.1) and previously applied to model the relationship between gene expression and expression variability (Alemu, et al. 2014). The residual from this regression would reflect the remaining variation independent of gene expression, which then can be interrogated for other genomic factors. This analysis yielded almost identical results as our initial linear models, though at the cost of heavily reduced R^2 values across the board (Table A.2). For the TATA box term, the significance at the $P < 0.05$ threshold and the direction of correlation remained the same for all honey bee studies. Similarly, the P-value for the TATA box term was nearly the same for the fly datasets, with only one study, Thackray et al. 2018, having a small change going from $P = 0.055$ in the full model to $P = 0.048$ (Table A.2). For the initiator motif term, the direction and significance remained the same for all fly studies and only changed for one honey bee study (Rutter et al. 2019) (Table A.2). Gene length, as with the other covariates, was the same across all studies with the exception of Brown et al. 2020, which was no longer statistically significant after regressing out the effects of gene expression (Table A.2). Due to the expected strong effects of mean expression on expression variability, there was a sharp drop off in R^2 values across the board. By regressing out gene expression, only 3 fly and 2 honey bee studies had models explaining more than 10% of the variance in expression variability. Nevertheless, the results of both linear model approaches indicate that the presence of a TATA box in the gene promoter region is consistently correlated with higher expression variability (Figure 5.1B).

We also used a partial correlations approach to examine effects of covariates free from the effects of gene expression. Specifically, we separately applied partial correlations for each numerical variable (gene length for both organisms in addition to CpG O/E for honey bee) while controlling for mean expression. Using this method, gene length was a significant term in 10 fly and 6 honey bee datasets (Table A.2).

5.2.2 DNA methylation is anti-correlated with expression variation

We utilized CpG O/E as a proxy measurement for Gene body DNA methylation in the honey bee (Elango, et al. 2009) datasets, as *Drosophila* lacks genomic DNA methylation and displays a unimodal CpG O/E distribution unlike the honey bee (Figure A.2). In all of our statistical methods (full linear model, linear model with mean expression regressed out, and partial correlations), the CpG O/E term was highly significantly and positively correlated with gene expression variation (Figure 5**.**1 and Table A.1-3). The value of the coefficient was highly consistent across all methods, including the full linear model, linear model with mean expression regressed out, and partial correlations, respectively (Figure 5.1 and Table A.1-3). Outside of mean expression, which was by far the most significant and impactful covariate, CpG O/E displayed strong and stable correlation with gene expression variation across all honey bee datasets. Since CpG O/E itself is negatively correlated with DNA methylation, these results align with previous findings in both mammals and insects that DNA methylation is associated with reduced gene expression variation (Huh, et al. 2013; Wu, et al. 2020b).

5.3 Methods

5.3.1 Gene expression data

We analyzed a total of 20 RNA-seq datasets for this study, 12 of which are from fly (*Drosophila melanogaster*) and 8 from honey bee (*Apis mellifera*) (Table A.1). Our fly datasets were chosen from a diverse set of laboratories as well as recently published with at least 10 samples (no more than 2 years old). The honey bee studies were all of the RNA-seq datasets we could access, as well as being fairly recent and a minimum of 10 samples (one from 2012, the rest were from 2016-2020).

5.3.2 Data processing

Reads for each study were trimmed to remove low quality reads and adaptors using default Trim_galore! (Martin 2011) settings. Trimmed reads were then aligned to their respective genomes, amel 4.5 and dmel r6.33 for honey bee and fly, respectively, using HISAT2 with soft clipping disabled (parameter setting: --sp 1000,1000). Following alignment, gene counts were generated with HTSeq (Anders, et al. 2015a) default parameters and imported into R (Team 2014) for further downstream analyses. Gene expression for each study was quantified and normalized using the "estimateSizeFactors" function in the DESeq2 package (Love, et al. 2014a). To remove lowly expressed genes, we removed genes with counts less than 5 and also required a gene to be expressed in at least 10% of all samples in the study. Gene expression variation was measured as the percent coefficient of variation (CV) of gene expression (Huh, et al. 2013) and CpG O/E values for the honey bee genome was calculated as previously described (Lindsey, Kelkar, et al. 2018).

5.3.1 Core promoter elements

Core promoter element designations for TATA boxes and initiator motifs were obtained from the Eukaryotic Promoter Database (Cavin Perier, et al. 1998; Dreos, et al. 2017). Briefly, promoter classifications for each organism were downloaded from the database using the "EPDnew selection tool" as done in a previous study (Faure, et al. 2017).

5.3.2 Statistics

For our full linear model, gene expression variation was used as the response variable for the following quadratic model: $\log_{10}(CV) \sim \log_2(expression) + \log_2(expression)^2 + \log_{10}(gene\ length)$ + TATA box + Initiator motif + X, where X are additional covariates from each experiment based on its metadata file. In our second set of linear models, we first regressed out the effect of gene expression with $\log_{10}(CV) \sim \log_2(expression) + \log_2(expression)^2$ and then using the residuals as the response variable mirroring the full linear model: $residuals \sim \log_{10}(gene\ length)$ + TATA box + Initiator motif + X. Partial correlation was performed using the "pcorr" function in R with gene expression as the variable that was controlled for and gene length and CpG O/E (honey bee studies only) as the response variables.

CHAPTER 6. CONCLUSIONS

The incredible pace of technical advances of multi-omics methods has allowed researchers to greatly expand profiling of DNA methylation throughout previously unexplored lineages (Zemach, et al. 2010; Bewick, et al. 2017). This thesis is centered on characterizing DNA methylation in the hymenopteran insect lineage, an emerging system for epigenetic research (Lyko, et al. 2010; Glastad, et al. 2011; Herb, et al. 2012), and its functional relationship with transcription. The hymenopterans include bees, wasps, and ants, providing an astonishing amount of diversity to study behavioral, molecular, and evolutionary hypotheses.

Chapter 2 provides a general survey of DNA methylation in the hymenopteran order by characterizing its distribution in seven organisms and presenting a method for identifying units of methylation. The idea was inspired by the concept of "CpG islands" that are characterized in mammals, which are dense regions of hypomethylated CpGs often found in the promoters of actively transcribed genes (Bird 1992; Schubeler 2015). We developed an analogous, but entirely different, concept and applied it to the overall hypomethylated insect genome that has clusters of methylated CpGs. By using a sliding window approach to capture these clusters of hypermethylated of CpGs, we developed units of methylation term "methylation islands" (MIs) that could be compared across species to find potentially underlying functional consequences. Indeed, we discovered that MIs were functional units that were enriched in evolutionarily conserved genes and overrepresented at exon-intron boundaries, supporting previous findings that gene body methylation is associated with increased transcription and has roles in splicing (Flores, et

al. 2012; Herb, et al. 2012; Li-Byarlay, et al. 2013; Galbraith, et al. 2015). We also found that MI gain and loss in coding regions was significantly correlated with up- and down-regulation in expression, respectively. While studies with paired epigenomic and transcriptomic data are currently limited, these preliminary findings suggest that methylation islands in insects and other lineages has the potential to offer new insights into epigenetic regulation.

How changes in methylation, whether due to intrinsic or extrinsic causes, affect gene transcription is another question at the forefront of epigenetics. In Chapter 3, we demonstrated that both epigenomic and transcriptomic changes accompanied a drastic alteration in reproductive physiology due to *Wolbachia* infection in *Trichogramma pretiosum*. The transition from sexual reproduction to parthenogenesis is a phenomenon in arthropods (Werren, et al. 2008), but the mechanism by which *Wolbachia* induces this phenotype remain unclear. By devising an innovative introgression scheme, we created genetically identical infected and uninfected wasp strains in order to make comparisons free from the effects of the divergent genetic background. We discovered that *Wolbachia* infection and the resulting parthenogenesis phenotype was indeed accompanied by both genome-wide DNA methylation and transcriptomic changes. Differentially methylated genes were associated with functions related to oocyte development and cell division, seemingly fitting in with *Wolbachia*'s potential manipulation of meiosis (Werren, et al. 2008; Lindsey, Kelkar, et al. 2018). However, differentially expressed genes tended to be lineage-specific genes with unknown functions, potentially pointing to host-specific responses to infection. Despite *Wolbachia* infection affecting both epigenomic and transcriptomic processes, as well as increasing levels of methylation and transcription, we

found little overlap between differentially methylated and expressed genes. These results indicate and support previous findings that changes in DNA methylation do not directly cause changes in transcription (Lyko, et al. 2010; Wang, et al. 2013; Galbraith, et al. 2015).

Parent-of-origin expression, where the allele from one parent is preferentially expressed over the other, has been long observed in mammals and plants and found to be regulated in part regulated by DNA methylation (Reik and Walter 2001; Bird 2002; Law and Jacobsen 2010). The kinship theory of intragenomic conflict predicts that the differential relatedness between matrigenes and patrigenes in social insects such as the honey bee should lead to parent-specific expression (Queller 2003). Evidence for this theory was found in a previous study utilizing reciprocal crosses of European and Africanized bees (Galbraith, et al. 2016), yet whether this phenomenon was associated with epigenetic regulation was unknown. Chapter 4 sampled bees from the same crosses as the aforementioned study to investigate whether predictions from the kinship theory applied to DNA methylation and whether it was regulating parent-specific expression. Our results indicated that the lineage effect was the strongest, which was in line with previous studies in other species showing that DNA methylation was highly influenced by the background genetics (Jones 2012; Smith and Meissner 2013; Mendizabal, et al. 2014; Yi 2017). More importantly, we showed, for the first time, evidence of parent-specific methylation in insects. Interestingly, genes displaying parent-specific methylation significantly overlapped with those exhibiting lineage-specific methylation, but not with those displaying parent-specific expression. These finding suggest that certain CpGs in the honey bee genome may be particularly modifiable to methylation changes, and that allele-specific DNA methylation is not directly responsible for allele-specific gene expression.

Given the lack of direct association between DNA methylation and transcription, Chapter 5 deals with an alternate hypothesis proposing that methylation may affect gene expression variability, which may largely reflect transcriptional noise (Bird 1995; Blake, et al. 2003; Arias and Hayward 2006; Huh, et al. 2013), rather than the total amount of transcripts. We gathered a wealth of RNA-seq datasets to test the impact of DNA methylation on gene expression variability in honey bees. We tested this in the context of other variables previously shown to affect gene expression variability (Huh, et al. 2013; Faure, et al. 2017). In addition, we included *Drosophila* data to see whether patterns in expression variability vary for lineages that have lost DNA methylation. We found that levels of gene expression had by far the most profound effect on the expression variability, with genes having high expression having decreased expression variability. The presence of a TATA box in the gene promoter was consistently positively correlated with gene expression noise which has been a well-established pattern in other organisms (Hornung, et al. 2012; Zoller, et al. 2015; Faure, et al. 2017). Controlling for the effect of gene expression using two different methods provided support for these results. Finally, we show that DNA methylation as approximated using CpG O/E is significantly and consistently anti-correlated with gene expression variability across all datasets.

In summary, the chapters outlined in this thesis provide an extensive examination of the functional role of DNA methylation in the hymenopteran order. We provided a comprehensive survey of distribution of DNA methylation in the order along with a novel method of finding and characterizing clusters of methylated CpGs. The subsequent studies demonstrated that genome-wide methylation was highly labile, subject to change as a result of genetic and infectious forces. And while we consistently found a lack of

direction association between levels of DNA methylation and gene transcription, we did observe strong effects of methylation on gene expression variation. With the continued proliferation of sequencing technologies and studies, incorporation of additional methods such as chromatin accessibility assays and single-cell genomics can hopefully further elucidate the role of DNA methylation in the insect lineage.

APPENDIX A. SUPPLEMENTARY MATERIAL FOR CHAPTER 2

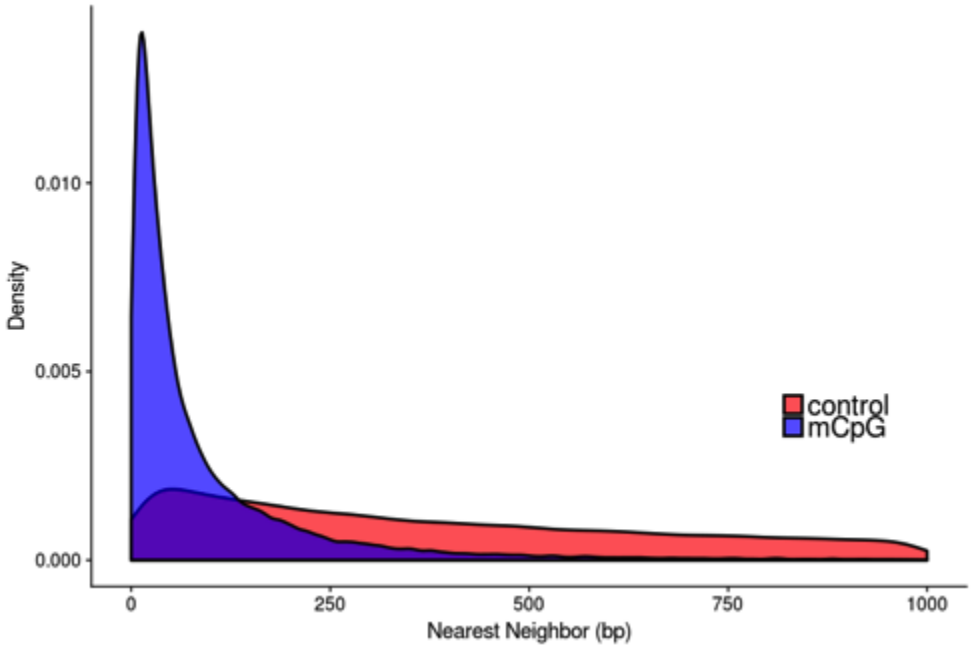

Figure A.1 Distance to nearest neighbor for control and mCGs.

Density plot for each type of CG in *A. mellifera* (n = 78,846 for both mCpG and control CpGs)

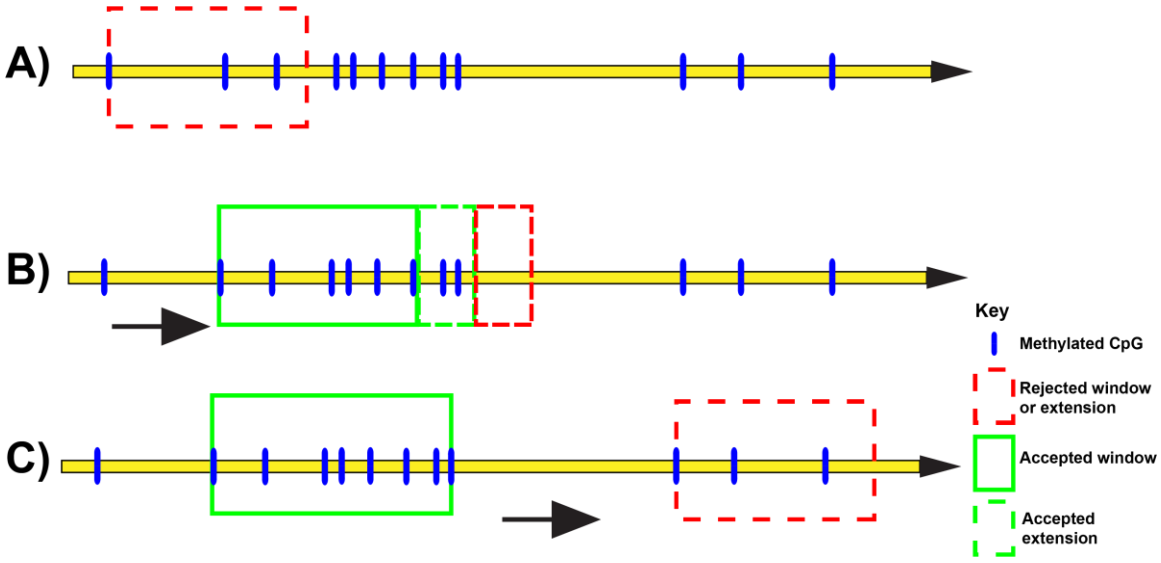

Figure A.2 Sliding window algorithm.

Sliding window approach used to identify MIs. A) The window moves in a 5' -> 3' fashion and calculates the mCG fraction of windows until a window meets the mCG fraction threshold (0.02 for this study). B) A window that satisfies the threshold is extended by 50bp a time until the entire region (original window + extension) falls below the threshold. C) The MI is terminated at the last mCG of the previously evaluated region and the evaluation mCG fraction of windows at the next downstream mCG starts.

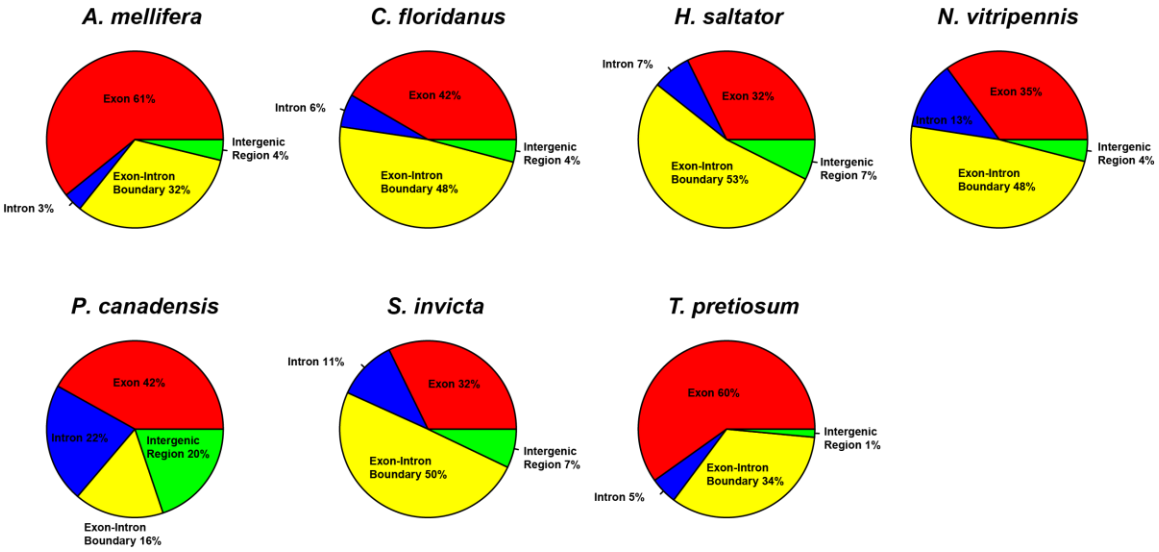

Figure A.3 Distribution of MIs in key genic regions.

Pie charts showing percentage of MIs found within exons, introns, exon-intron boundaries, and intergenic regions for all seven species.

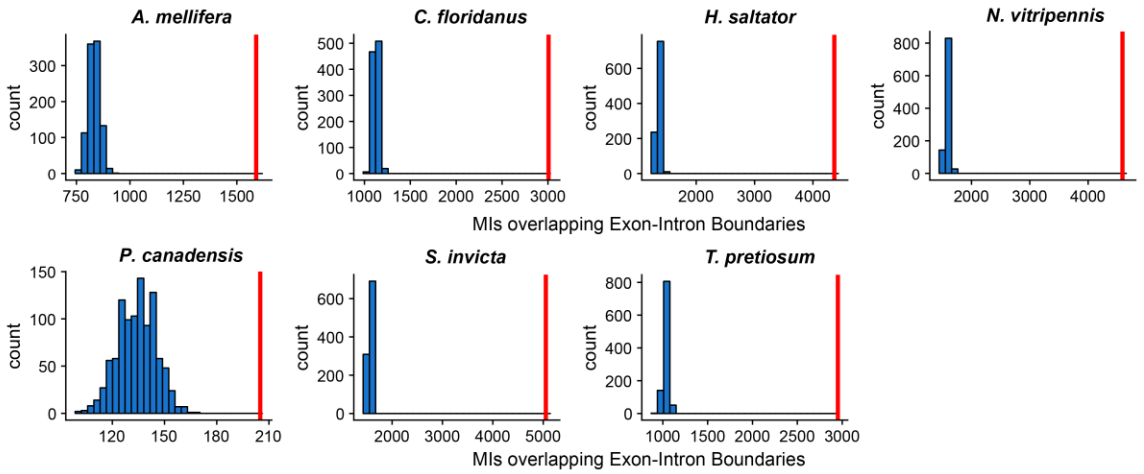

Figure A.4 Permutation of MIs at exon-intron boundaries.

Empirical evidence showing that the expected number of MIs (blue bars) is much lower than the observed (Red line) over 1000 permutations.

APPENDIX B. SUPPLEMENTARY MATERIAL FOR CHAPTER 3

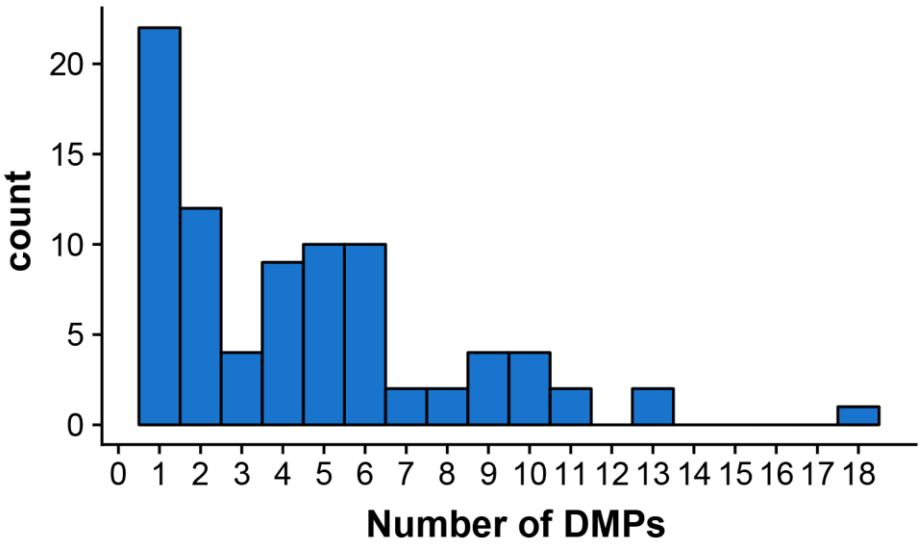

Figure B.1 DMPs in genes.

Number of DMPs found within differentially methylated genes.

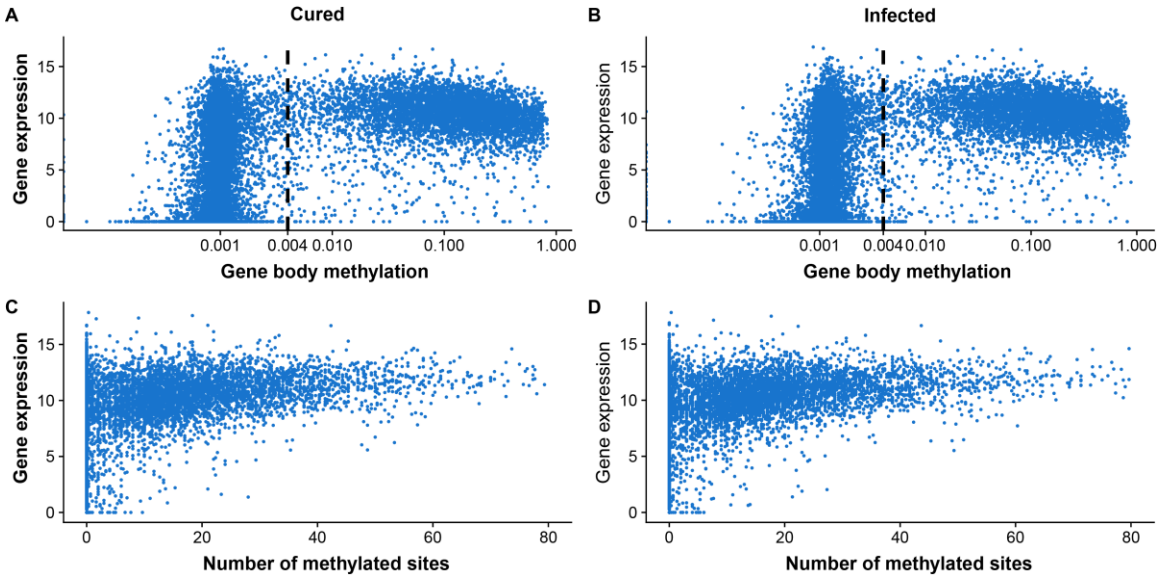

Figure B.2 Methylated genes have higher and are more constitutively expressed.

Methylated genes, defined as having >0.004 gene body methylation, show higher gene expression than unmethylated genes (<0.004 gene body methylation) in both A) cured and B) infected wasps.

APPENDIX C. SUPPLEMENTARY MATERIAL FOR CHAPTER 4

Table C.1 Differentially expressed genes for each reproductive state and genetic block.

RNA-seq data from the previous study was re-analyzed using the same pipeline was the WGBS data.

		Block A		Block B	
		Sterile	Reproductive	Sterile	Reproductive
Parent-of-origin	Maternal bias	1	3	0	1
	Paternal bias	3	53	1	30
Lineage	Africanized bias	0	20	1	3
	European bias	2	31	3	5

Table C.2 Overlap between DMGs and DEGs.

We found almost no overlap between DEGs and DMGs showing the same direction of allelic bias.

		Block A		Block B	
		Sterile	Reproductive	Sterile	Reproductive
Parent-of-origin	Maternal bias	0	0	0	0
	Paternal bias	0	0	0	0
Lineage	Africanized bias	0	1	0	0
	European bias	0	3	0	0

REFERENCES

REFERENCES

Aerts S, Thijs G, Dabrowski M, Moreau Y, De Moor B. 2004. Comprehensive analysis of the base composition around the transcription start site in Metazoa. BMC Genomics 5:34.

Alemu EY, Carl JW, Jr., Corrada Bravo H, Hannenhalli S. 2014. Determinants of expression variability. Nucleic Acids Res 42:3503-3514.

Anders S, Pyl PT, Huber W. 2015a. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31:166-169.

Anders S, Pyl PT, Huber W. 2015b. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31:166-169.

Anders S, Reyes A, Huber W. 2012. Detecting differential usage of exons from RNA-seq data. Genome Res 22:2008-2017.

Arias AM, Hayward P. 2006. Filtering transcriptional noise during development: concepts and mechanisms. Nat Rev Genet 7:34-44.

Arsenault SV, Hunt BG, Rehan SM. 2018. The effect of maternal care on gene expression and DNA methylation in a subsocial bee. Nat Commun 9:3468.

Barroso GV, Puzovic N, Dutheil JY. 2018. The Evolution of Gene-Specific Transcriptional Noise Is Driven by Selection at the Pathway Level. Genetics 208:173-189.

Beckmann JF, Ronau JA, Hochstrasser M. 2017. A Wolbachia deubiquitylating enzyme induces cytoplasmic incompatibility. Nat Microbiol 2:17007.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 57:289-300.

Bennett GM, Moran NA. 2015. Heritable symbiosis: The advantages and perils of an evolutionary rabbit hole. Proceedings of the National Academy of Sciences 112:10169-10176.

Bernstein BE, Meissner A, Lander ES. 2007. The mammalian epigenome. Cell 128:669-681.

Bewick AJ, Vogel KJ, Moore AJ, Schmitz RJ. 2017. Evolution of DNA Methylation across Insects. Mol Biol Evol 34:654-665.

Bhattacharya T, Newton ILG, Hardy RW. 2017. Wolbachia elevates host methyltransferase expression to block an RNA virus early during infection. PLoS Pathog 13:e1006427.

Bird A. 2002. DNA methylation patterns and epigenetic memory. Genes Dev 16:6-21.

Bird A. 1992. The essentials of DNA methylation. Cell 70:5-8.

Bird A, Taggart M, Frommer M, Miller OJ, Macleod D. 1985. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell 40:91-99.

Bird AP. 1995. Gene number, noise reduction and biological complexity. Trends Genet 11:94-100.

Blake WJ, M KA, Cantor CR, Collins JJ. 2003. Noise in eukaryotic gene expression. Nature 422:633-637.

Bolger AM, Lohse M, Usadel B. 2014a. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Bolger AM, Lohse M, Usadel B. 2014b. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Bonasio R, Li Q, Lian J, Mutti NS, Jin L, Zhao H, Zhang P, Wen P, Xiang H, Ding Y, et al. 2012. Genome-wide and caste-specific DNA methylomes of the ants Camponotus floridanus and Harpegnathos saltator. Curr Biol 22:1755-1764.

Capra JA, Singh M. 2007. Predicting functionally important residues from sequence conservation. Bioinformatics 23:1875-1882.

Cavin Perier R, Junier T, Bucher P. 1998. The Eukaryotic Promoter Database EPD. Nucleic Acids Res 26:353-357.

Choi JK, Kim YJ. 2008. Epigenetic regulation and the variability of gene expression. Nat Genet 40:141-147.

Dayeh T, Volkov P, Salo S, Hall E, Nilsson E, Olsson AH, Kirkpatrick CL, Wollheim CB, Eliasson L, Ronn T, et al. 2014. Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. PLoS Genet 10:e1004160.

Ding XL, Yang X, Liang G, Wang K. 2016. Isoform switching and exon skipping induced by the DNA methylation inhibitor 5-Aza-2'-deoxycytidine. Sci Rep 6:24545.

Dolzhenko E, Smith AD. 2014. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinformatics 15:215.

Dreos R, Ambrosini G, Groux R, Cavin Perier R, Bucher P. 2017. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. Nucleic Acids Res 45:D51-D55.

Elango N, Hunt BG, Goodisman MA, Yi SV. 2009. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera. Proc Natl Acad Sci U S A 106:11206-11211.

Elango N, Yi SV. 2008. DNA methylation and structural and functional bimodality of vertebrate promoters. Mol Biol Evol 25:1602-1608.

Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, Kaper F, Fan JB, Zhang K, Chun J, et al. 2016. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Methods 13:241-244.

Faure AJ, Schmiedel JM, Lehner B. 2017. Systematic Analysis of the Determinants of Gene Expression Noise in Embryonic Stem Cells. Cell Syst 5:471-484 e474.

Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al. 2010. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A 107:8689-8694.

Ferguson-Smith AC, Bourc'his D. 2018. The discovery and importance of genomic imprinting. Elife 7.

Flores K, Wolschin F, Corneveaux JJ, Allen AN, Huentelman MJ, Amdam GV. 2012. Genome-wide association between DNA methylation and alternative splicing in an invertebrate. BMC Genomics 13:480.

Foret S, Kucharski R, Pellegrini M, Feng S, Jacobsen SE, Robinson GE, Maleszka R. 2012. DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. Proc Natl Acad Sci U S A 109:4968-4973.

Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB. 2004. Noise minimization in eukaryotic gene expression. PLoS Biol 2:e137.

Galbraith DA, Kocher SD, Glenn T, Albert I, Hunt GJ, Strassmann JE, Queller DC, Grozinger CM. 2016. Testing the kinship theory of intragenomic conflict in honey bees (Apis mellifera). Proc Natl Acad Sci U S A 113:1020-1025.

Galbraith DA, Yang X, Nino EL, Yi S, Grozinger C. 2015. Parallel epigenomic and transcriptomic responses to viral infection in honey bees (Apis mellifera). PLoS Pathog 11:e1004713.

Gao S, Zou D, Mao L, Liu H, Song P, Chen Y, Zhao S, Gao C, Li X, Gao Z, et al. 2015. BS-SNPer: SNP calling in bisulfite-seq data. Bioinformatics 31:4006-4008.

Gardiner-Garden M, Frommer M. 1987. CpG islands in vertebrate genomes. J Mol Biol 196:261-282.

Gavotte L, Henri H, Stouthamer R, Charif D, Charlat S, Bouletreau M, Vavre F. 2007. A Survey of the bacteriophage WO in the endosymbiotic bacteria Wolbachia. Mol Biol Evol 24:427-435.

Ghosh S, Chan CK. 2016. Analysis of RNA-Seq Data Using TopHat and Cufflinks. Methods Mol Biol 1374:339-361.

Glastad KM, Hunt BG, Yi SV, Goodisman MA. 2011. DNA methylation in insects: on the brink of the epigenomic era. Insect Mol Biol 20:553-565.

Greenberg MVC, Bourc'his D. 2019. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol 20:590-607.

Haig D. 2000. The kinship theory of genomic imprinting. Annual Review of Ecology and Systematics 31:9-32.

Harris HL, Braig HR. 2003. Sperm chromatin remodelling and Wolbachia-induced cytoplasmic incompatibility in Drosophila. Biochem Cell Biol 81:229-240.

Herb BR, Wolschin F, Hansen KD, Aryee MJ, Langmead B, Irizarry R, Amdam GV, Feinberg AP. 2012. Reversible switching between epigenetic states in honeybee behavioral subcastes. Nat Neurosci 15:1371-1373.

Hilgenboecker K, Hammerstein P, Schlattmann P, Telschow A, Werren JH. 2008. How many species are infected with *Wolbachia*? - A statistical analysis of current data. Fems Microbiology Letters 281:215-220.

Honeybee Genome Sequencing C. 2006. Insights into social insects from the genome of the honeybee Apis mellifera. Nature 443:931-949.

Hornung G, Bar-Ziv R, Rosin D, Tokuriki N, Tawfik DS, Oren M, Barkai N. 2012. Noise-mean relationship in mutated promoters. Genome Res 22:2409-2417.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44-57.

Huh I, Wu X, Park T, Yi SV. 2019. Detecting differential DNA methylation from sequencing of bisulfite converted DNA of diverse species. Brief Bioinform 20:33-46.

Huh I, Wu X, Park T, Yi SV. 2017. Detecting differential DNA methylation from sequencing of bisulfite converted DNA of diverse species. Brief Bioinform.

Huh I, Yang X, Park T, Yi SV. 2014. Bis-class: a new classification tool of methylation status using bayes classifier and local methylation information. BMC Genomics 15:608.

Huh I, Zeng J, Park T, Yi SV. 2013. DNA methylation and transcriptional noise. Epigenetics Chromatin 6:9.

Hunt BG, Glastad KM, Yi SV, Goodisman MA. 2013. Patterning and regulatory associations of DNA methylation are mirrored by histone modifications in insects. Genome Biol Evol 5:591-598.

Illingworth RS, Bird AP. 2009. CpG islands--'a rough guide'. FEBS Lett 583:1713-1720.

Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lonnerberg P, Linnarsson S. 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods 11:163-166.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13:484-492.

Jones PA, Takai D. 2001. The role of DNA methylation in mammalian epigenetics. Science 293:1068-1070.

Keller TE, Lasky JR, Yi SV. 2016. The multivariate association between genomewide DNA methylation and climate across the range of Arabidopsis thaliana. Mol Ecol 25:1823-1837.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg S. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology 14:R36.

Kocher SD, Tsuruda JM, Gibson JD, Emore CM, Arechavaleta-Velasco ME, Queller DC, Strassmann JE, Grozinger CM, Gribskov MR, San Miguel P, et al. 2015. A Search for Parent-of-Origin Effects on Honey Bee Gene Expression. G3 (Bethesda) 5:1657-1662.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27:1571-1572.

Krueger F, Andrews SR. 2016. SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes. F1000Res 5:1479.

Kucharski R, Maleszka J, Foret S, Maleszka R. 2008. Nutritional control of reproductive status in honeybees via DNA methylation. Science 319:1827-1830.

Kumar S, Kim Y. 2017. An endoparasitoid wasp influences host DNA methylation. Sci Rep 7:43287.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11:204-220.

Lehner B. 2008. Selection to minimise noise in living systems and its implications for the evolution of gene expression. Mol Syst Biol 4:170.

LePage DP, Metcalf JA, Bordenstein SR, On J, Perlmutter JI, Shropshire JD, Layton EM, Funkhouser-Jones LJ, Beckmann JF, Bordenstein SR. 2017. Prophage WO genes recapitulate and enhance Wolbachia-induced cytoplasmic incompatibility. Nature 543:243-247.

Lev Maor G, Yearim A, Ast G. 2015. The alternative role of DNA methylation in splicing regulation. Trends Genet 31:274-280.

Li-Byarlay H, Li Y, Stroud H, Feng S, Newman TC, Kaneda M, Hou KK, Worley KC, Elsik CG, Wickline SA, et al. 2013. RNA interference knockdown of DNA methyl-transferase 3 affects gene alternative splicing in the honey bee. Proc Natl Acad Sci U S A 110:12750-12755.

Li S, Zhang J, Huang S, He X. 2018. Genome-wide analysis reveals that exon methylation facilitates its selective usage in the human transcriptome. Brief Bioinform 19:754-764.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923-930.

Lindsey ARI, Kelkar YD, Wu X, Sun D, Martinson EO, Yan Z, Rugman-Jones PF, Hughes DST, Murali SC, Qu J, et al. (Lindsey2018 co-authors). 2018. Comparative genomics of the miniature wasp and pest control agent *Trichogramma pretiosum*. BMC Biology 16:54.

Lindsey ARI, Rice DW, Bordenstein SR, Brooks AW, Bordenstein SR, Newton ILG. 2018. Evolutionary Genetics of Cytoplasmic Incompatibility Genes cifA and cifB in Prophage WO of Wolbachia. Genome Biol Evol 10:434-451.

Lindsey ARI, Werren JH, Richards S, Stouthamer R. 2016. Comparative genomics of a parthenogenesis-inducing *Wolbachia* symbiont. G3: Genes|Genomes|Genetics 6:2113-2123.

Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, et al. 2013. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol 31:142-147.

Lonsdale Z, Lee K, Kiriakidu M, Amarasinghe H, Nathanael D, O'Connor CJ, Mallon EB. 2017. Allele specific expression and methylation in the bumblebee, Bombus terrestris. PeerJ 5:e3798.

Love MI, Huber W, Anders S. 2014a. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550.

Love MI, Huber W, Anders S. 2014b. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15:550.

Luedi PP, Dietrich FS, Weidman JR, Bosko JM, Jirtle RL, Hartemink AJ. 2007. Computational and experimental identification of novel human imprinted genes. Genome Res 17:1723-1730.

Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, Maleszka R. 2010. The honey bee epigenomes: differential methylation of brain DNA in queens and workers. PLoS Biol 8:e1000506.

Mardis E, McCombie WR. 2017. Library Quantification: Fluorometric Quantitation of Double-Stranded or Single-Stranded DNA Samples Using the Qubit System. Cold Spring Harb Protoc 2017:pdb prot094730.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10-12.

Matsuo K, Clay O, Takahashi T, Silke J, Schaffner W. 1993. Evidence for erosion of mouse CpG islands during mammalian evolution. Somat Cell Mol Genet 19:543-555.

Medzhitov R, Preston-Hurlburt P, Janeway CA, Jr. 1997. A human homologue of the Drosophila Toll protein signals activation of adaptive immunity. Nature 388:394-397.

Mendizabal I, Berto S, Usui N, Toriumi K, Chatterjee P, Douglas C, Huh I, Jeong H, Layman T, Tamminga CA, et al. 2019. Cell type-specific epigenetic links to schizophrenia risk in the brain. Genome Biol 20:135.

Mendizabal I, Keller TE, Zeng J, Yi SV. 2014. Epigenetics and Evolution. Integrative and Comparative Biology 54:31-42.

Panaro NJ, Yuen PK, Sakazume T, Fortina P, Kricka LJ, Wilding P. 2000. Evaluation of DNA fragment sizing and quantification by the Agilent 2100 Bioanalyzer. Clinical Chemistry 46:1851-1853.

Park Y, Wu H. 2016. Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics 32:1446-1453.

Patten MM, Ross L, Curley JP, Queller DC, Bonduriansky R, Wolf JB. 2014. The evolution of genomic imprinting: theories, predictions and empirical tests. Heredity (Edinb) 113:119-128.

Pegoraro M, Marshall H, Lonsdale ZN, Mallon EB. 2017. Do social insects support Haig's kin theory for the evolution of genomic imprinting? Epigenetics 12:725-742.

Queller DC. 2003. Theory of genomic imprinting conflict in social insects. BMC Evol Biol 3:15.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria: URL http://www.R-project.org/.

Rainier S, Feinberg AP. 1994. Genomic imprinting, DNA methylation, and cancer. J Natl Cancer Inst 86:753-759.

Ravarani CN, Chalancon G, Breker M, de Groot NS, Babu MM. 2016. Affinity and competition for TBP are molecular determinants of gene expression noise. Nat Commun 7:10417.

Raychoudhury R, Baldo L, Oliveira DC, Werren JH. 2009. Modes of acquisition of Wolbachia: horizontal transfer, hybrid introgression, and codivergence in the Nasonia species complex. Evolution 63:165-183.

Razin A, Cedar H. 1994. DNA methylation and genomic imprinting. Cell 77:473-476.

Reik W, Walter J. 2001. Genomic imprinting: parental influence on the genome. Nat Rev Genet 2:21-32.

Robertson KD, Wolffe AP. 2000. DNA methylation in health and disease. Nat Rev Genet 1:11-19.

Rosic S, Amouroux R, Requena CE, Gomes A, Emperle M, Beltran T, Rane JK, Linnett S, Selkirk ME, Schiffer PH, et al. 2018. Evolutionary analysis indicates that DNA alkylation damage is a byproduct of cytosine DNA methyltransferase activity. Nat Genet 50:452-459.

Russell JE, Stouthamer R. 2011. The genetics and evolution of obligate reproductive parasitism in *Trichogramma pretiosum* infected with parthenogenesis-inducing *Wolbachia*. Heredity 106:58-67.

Sanchez A, Kondev J. 2008. Transcriptional control of noise in gene expression. Proc Natl Acad Sci U S A 105:5081-5086.

Sarda S, Zeng J, Hunt BG, Yi SV. 2012. The evolution of invertebrate gene body methylation. Mol Biol Evol 29:1907-1916.

Saze H, Mittelsten Scheid O, Paszkowski J. 2003. Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis. Nat Genet 34:65-69.

Schubeler D. 2015. Function and information content of DNA methylation. Nature 517:321-326.

Sevier SA, Kessler DA, Levine H. 2016. Mechanical bounds to transcriptional noise. Proc Natl Acad Sci U S A 113:13983-13988.

Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, Dupre Y, Antonarakis SE. 2011. DNA methylation profiles of human active and inactive X chromosomes. Genome Res 21:1592-1600.

Smith ZD, Meissner A. 2013. DNA methylation: roles in mammalian development. Nat Rev Genet 14:204-220.

Stouthamer R, Breeuwer JAJ, Luck RF, Werren JH. 1993. Molecular-identification of microorganisms associated with parthenogenesis. Nature 361:66-68.

Stouthamer R, Kazmer DJ. 1994. Cytogenetics of microbe-associated parthenogenesis and its consequences for gene flow in *Trichogramma* wasps. Heredity 73:317-327.

Stouthamer R, Luck RF, Hamilton WD. 1990. Antibiotics cause parthenogenetic *Trichogramma* (Hymenoptera, Trichogrammatidae) to revert to sex. Proceedings of the National Academy of Sciences 87:2424-2427.

Stouthamer R, Russell JE, Vavre F, Nunney L. 2010. Intragenomic conflict in populations infected by Parthenogenesis Inducing Wolbachia ends with irreversible loss of sexual reproduction. BMC Evol Biol 10:229.

Stouthamer R, Werren JH. 1993. Microbes associated with parthenogenesis in wasps of the genus *Trichogramma*. Journal of Invertebrate Pathology 61:6-9.

Sullivan W. 2017. *Wolbachia*, bottled water, and the dark side of symbiosis. Molecular Biology of the Cell 28:2343-2346.

Sun H, Towb P, Chiem DN, Foster BA, Wasserman SA. 2004. Regulated assembly of the Toll signaling complex drives Drosophila dorsoventral patterning. EMBO J 23:100-110.

Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9:465-476.

Team RC. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria: URL http://www.R-project.org/;.

Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. 2015. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nat Protoc 10:475-483.

Wang X, Werren JH, Clark AG. 2016. Allele-Specific Transcriptome and Methylome Analysis Reveals Stable Inheritance and Cis-Regulation of DNA Methylation in Nasonia. PLoS Biol 14:e1002500.

Wang X, Wheeler D, Avery A, Rago A, Choi JH, Colbourne JK, Clark AG, Werren JH. 2013. Function and evolution of DNA methylation in Nasonia vitripennis. PLoS Genet 9:e1003872.

Wang Y, Jorda M, Jones PL, Maleszka R, Ling X, Robertson HM, Mizzen CA, Peinado MA, Robinson GE. 2006. Functional CpG methylation system in a social insect. Science 314:645-647.

Wang Z, Zhang J. 2011. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. Proc Natl Acad Sci U S A 108:E67-76.

Werren JH, Baldo L, Clark ME. 2008. Wolbachia: master manipulators of invertebrate biology. Nat Rev Microbiol 6:741-751.

Wilkins JF, Haig D. 2003. What good is genomic imprinting: The function of parent-specific gene expression. Nature Reviews Genetics 4:359-368.

Wu X, Lindsey ARI, Chatterjee P, Werren JH, Stouthamer R, Yi SV. 2020a. Distinct epigenomic and transcriptomic modifications associated with Wolbachia-mediated asexuality. PLoS Pathog.

Wu X, Lindsey ARI, Chatterjee P, Werren JH, Stouthamer R, Yi SV. 2020b. Distinct epigenomic and transcriptomic modifications associated with Wolbachia-mediated asexuality. PLoS Pathog 16:e1008397.

Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, Nipitwattanaphon M, Gotzek D, et al. 2011. The genome of the fire ant Solenopsis invicta. Proc Natl Acad Sci U S A 108:5679-5684.

Ye YH, Woolfit M, Huttley GA, Rances E, Caragata EP, Popovici J, O'Neill SL, McGraw EA. 2013. Infection with a Virulent Strain of Wolbachia Disrupts Genome Wide-Patterns of Cytosine Methylation in the Mosquito Aedes aegypti. PLoS One 8:e66482.

Yi S. 2012. Birds do it, bees do it, worms and ciliates do it too: DNA methylation from unexpected corners of the tree of life. Genome Biol 13:174.

Yi SV. 2017. Insights into Epigenome Evolution from Animal and Plant Methylomes. Genome Biol Evol 9:3189-3201.

Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13:335-340.

Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328:916-919.

Zhang G, Hussain M, O'Neill SL, Asgari S. 2013. Wolbachia uses a host microRNA to regulate transcripts of a methyltransferase, contributing to dengue virus inhibition in Aedes aegypti. Proc Natl Acad Sci U S A 110:10276-10281.

Zoller B, Nicolas D, Molina N, Naef F. 2015. Structure of silent transcription intervals and noise characteristics of mammalian genes. Mol Syst Biol 11:823.

Zug R, Hammerstein P. 2012. Still a host of hosts for Wolbachia: analysis of recent data suggests that 40% of terrestrial arthropod species are infected. PLoS One 7:e38544.