GBE

Genome Biology and Evolution

SMBE

Functional Conservation of DNA Methylation in the Pea Aphid and the Honeybee

Brendan G. Hunt[1], Jennifer A. Brisson[2], Soojin V. Yi*,[1], and Michael A. D. Goodisman*,[1]

[1]School of Biology, Georgia Institute of Technology

[2]School of Biological Sciences, University of Nebraska

*Corresponding author: E-mail: michael.goodisman@biology.gatech.edu; soojin.yi@biology.gatech.edu.

Accepted: 13 September 2010

Abstract

DNA methylation is a fundamental epigenetic mark known to have wide-ranging effects on gene regulation in a variety of animal taxa. Comparative genomic analyses can help elucidate the function of DNA methylation by identifying conserved features of methylated genes and other genomic regions. In this study, we used computational approaches to distinguish genes marked by heavy methylation from those marked by little or no methylation in the pea aphid, *Acyrthosiphon pisum*. We investigated if these two classes had distinct evolutionary histories and functional roles by conducting comparative analysis with the honeybee, *Apis* (*Ap.*) *mellifera*. We found that highly methylated orthologs in *A. pisum* and *Ap. mellifera* exhibited greater conservation of methylation status, suggesting that highly methylated genes in ancestral species may remain highly methylated over time. We also found that methylated genes tended to show different rates of evolution than unmethylated genes. In addition, genes targeted by methylation were enriched for particular biological processes that differed from those in relatively unmethylated genes. Finally, methylated genes were preferentially ubiquitously expressed among alternate phenotypes in both species, whereas genes lacking signatures of methylation were preferentially associated with condition-specific gene expression. Overall, our analyses support a conserved role for DNA methylation in insects with comparable methylation systems.

Key words: comparative genomics, DNA methylation, epigenetics, insects, phenotypic plasticity, polyphenism.

Introduction

DNA methylation is an important epigenetic modification that plays a role in gene regulation in many organisms (Wolffe and Matzke 1999; Jaenisch and Bird 2003; Weber et al. 2007). Although DNA methylation occurs in all three domains of life, its genomic patterns show considerable variation among taxa (Hendrich and Tweedie 2003; Field et al. 2004; Suzuki and Bird 2008). For example, vertebrate genomes exhibit global patterns of methylation, but invertebrate genomes tend to display reduced or minimal levels of methylation (Suzuki and Bird 2008). Moreover, methylation of gene promoter regions in vertebrates leads to transcriptional repression (Wolffe and Matzke 1999; Jaenisch and Bird 2003; Weber et al. 2007; Zemach et al. 2010), but this relationship has not been observed in invertebrates. Instead, methylation primarily targets invertebrate gene bodies (Suzuki and Bird 2008; Xiang et al. 2010; Zemach et al. 2010). These contrasting patterns and effects have traditionally enforced the view that DNA methylation plays a fundamentally different role in vertebrate and invertebrate genomes.

The arrival of genome sequences from multiple insects now makes a greater understanding of the patterns and phenotypic consequences of DNA methylation more tangible (Honeybee Genome Sequencing Consortium 2006; Wang et al. 2006; The International Aphid Genomics Consortium 2010; The Nasonia Genome Working Group 2010; Walsh et al. 2010). Specifically, comparative genomic analysis can be used to determine whether targets of DNA methylation are conserved between taxa. Moreover, the inferred patterns of methylation can be used to test current hypotheses explaining the evolutionary persistence of DNA methylation (Yi and Goodisman 2009). For example, it has been hypothesized that gene body methylation may act to minimize spurious transcription patterns (Suzuki et al. 2007; Maunakea et al. 2010), which could explain observations of dense methylation in functionally conserved genes and genes with ubiquitous expression among tissues

in invertebrates (Suzuki et al. 2007; Foret et al. 2009; Xiang et al. 2010). It has also been suggested that DNA methylation persists in animals for genomic defense against transposable elements (Yoder et al. 1997, but see Regev et al. [1998]; Simmen et al. [1999]; Suzuki et al. [2007], and Xiang et al. [2010]). DNA methylation may also act as an important mechanism for genomic imprinting, which results in the differential expression of parental alleles (Reik and Walter 2001). Finally, de novo DNA methylation is hypothesized to play an important role in developmental responsiveness to environmental factors and the regulation of phenotypic plasticity, as is apparently the case in the honeybee (Jaenisch and Bird 2003; Kucharski et al. 2008; Maleszka 2008).

The purpose of this study was to determine whether DNA methylation plays a conserved role in divergent insects with comparable DNA methylation systems. We provided insight into this question by comparing and contrasting the evolutionary signatures of DNA methylation in the genomes of the pea aphid, *Acyrthosiphon pisum*, and the honeybee, *Apis* (*Ap.*) *mellifera*.

Acyrthosiphon pisum diverged from *Ap. mellifera* more than 300 Ma (Gaunt and Miles 2002; Honeybee Genome Sequencing Consortium 2006), a time frame roughly equivalent to the divergence of modern birds and mammals (Kumar and Hedges 1998). Developmentally, *Ap. mellifera* undergoes full metamorphosis and possesses morphologically distinct larval, pupal, and adult stages. In contrast, *A. pisum* develops gradually and does not undergo metamorphosis. However, *A. pisum* and *Ap. mellifera* both serve as important models for understanding the evolution and development of phenotypic plasticity (Evans and Wheeler 2001; Brisson and Stern 2006; Honeybee Genome Sequencing Consortium 2006; Brisson 2010; The International Aphid Genomics Consortium 2010).

Specifically, aphids have a complex life cycle that alternates between asexual and sexual development. Asexual females exhibit a wing polyphenism in which they produce either winged or unwinged morphs depending on environmental cues (reviewed in Müller et al. 2001). During the sexual portion of the life cycle, males also produce winged or unwinged morphs. However, morph determination is genetic in males, and thus male wing dimorphism is referred to as a polymorphism (Smith and MacKay 1989). Honeybees, on the other hand, are highly social and dwell in large, predominantly female, colonies (Wilson 1971). Individuals partake in a remarkable division of labor, with a single queen typically dominating reproduction and workers engaged in tasks related to brood rearing, foraging, and colony defense (Wilson 1971). Queen and worker castes are developmentally determined by nutritional factors and exhibit dramatically different anatomy and behavior (Wheeler 1986; Evans and Wheeler 2001).

Importantly, both *Ap. mellifera* and *A. pisum* show evidence of widespread DNA methylation that is predominantly targeted to genes (Wang et al. 2006; Elango et al. 2009; Walsh et al. 2010). Consequently, patterns of genome methylation in *A. pisum* and *Ap. mellifera* can provide considerable insight into the function of gene methylation in insects, in particular, and invertebrates, in general.

In this study, we investigated the conservation of DNA methylation patterns in *A. pisum* and *Ap. mellifera* by first testing whether genes with similar functions are targeted by DNA methylation in both species. To achieve this aim, we examined patterns of functional enrichment among genes marked by relatively dense methylation and relatively sparse methylation. We further tested whether shared patterns of functional enrichment among DNA methylation targets are associated with conservation at the sequence level (Suzuki et al. 2007). Next, we examined whether *A. pisum* provided support for the hypothesis that genes with sparse methylation exhibit condition-specific gene expression (Elango et al. 2009; Foret et al. 2009). Finally, we synthesized our results with those from other recent investigations to advance a more comprehensive understanding of DNA methylation in insects. Overall, our results provide support for a remarkable level of conservation in gene methylation status and function over evolutionary time.

Materials and Methods

Gene Sequences

Analyses were conducted on mRNA transcript sequences because evidence suggests that DNA methylation preferentially targets exons in insects and other invertebrates (Wang et al. 2006; Suzuki et al. 2007; Elango et al. 2009; Xiang et al. 2010; Zemach et al. 2010). For *A. pisum*, the "ACYPmRNA" and the "ACYPproteins" official genes consensus sets were obtained from AphidBase (http://www.aphidbase.com). For *Ap. mellifera*, the "Amel_pre_release2" amino acid sequence official gene set (OGS) was obtained from BeeBase (http://www.beebase.org), and model RefSeq transcripts were downloaded from the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/Ftp). *Apis mellifera* OGS IDs were converted to RefSeq accessions using the "gene_info" and "gene2refseq" databases, also available from NCBI. For *Drosophila melanogaster*, "Release_5.21" transcript and protein sequence sets were obtained from flybase (http://flybase.org).

Normalized CpG Dinucleotide Content ($CpG_{O/E}$)

We used $CpG_{O/E}$ as a measure of the level of DNA methylation of genes (Saxonov et al. 2006; Suzuki et al. 2007; Weber et al. 2007; Yi and Goodisman 2009). $CpG_{O/E}$ acts

as a metric of levels of DNA methylation because methylation occurs predominantly on CpG dinucleotides in animals and methylated cytosines are hypermutable due to spontaneous deamination. This deamination causes a gradual depletion of CpG dinucleotides from methylated regions over time (Bird 1980). Consequently, genomic regions with relatively dense germline methylation have low $CpG_{O/E}$ and regions with little or no germline methylation maintain high levels of $CpG_{O/E}$. It is important to note that $CpG_{O/E}$ could be influenced by either the number of methylated CpG sites or the proportion of cells incurring methylation at a given locus. In addition, somatic mutations are not transmitted to progeny and therefore cannot influence $CpG_{O/E}$ in and of themselves. However, $CpG_{O/E}$ has been linked to empirically determined levels of DNA methylation in somatic tissues in insects, suggesting that many genes are universally methylated in germlines and soma (Foret et al. 2009; Xiang et al. 2010).

$CpG_{O/E}$ was calculated as described previously (Elango et al. 2009), from the gene sets above. Only RefSeq model sequences were used for analyses involving $CpG_{O/E}$ in *A. pisum* (except in the case of gene expression analysis, described below) because RefSeq models were used for *Ap. mellifera* in our analysis. Sequences with $CpG_{O/E}$ values of 0 were removed from further analysis.

Bimodal distributions of $CpG_{O/E}$ have previously been reported in both *Ap. mellifera* (Elango et al. 2009; Foret et al. 2009; Wang and Leung 2009) and *A. pisum* (Walsh et al. 2010). In this study, we used the NOCOM software package (Ott 1979) to estimate means, standard deviations, and proportions of two components of the mixture of normal distributions of $CpG_{O/E}$ for both *A. pisum* and *Ap. mellifera*. These distributions were plotted using R (R Development Core Team 2010), and their intersections were used as cutoffs to divide low $CpG_{O/E}$ and high $CpG_{O/E}$ gene classes.

Orthology

Three-way orthologs between *A. pisum*, *Ap. mellifera*, and *D. melanogaster* were identified by first performing pairwise BlastP comparisons of complete protein sequence sets with a cutoff of 1×10^{-5}, next identifying pairwise reciprocal best hits, and finally identifying orthologs with shared best hits among all pairwise comparisons (Altschul et al. 1997; Stajich et al. 2002). Orthologs determined in this manner were used for comparisons of $CpG_{O/E}$ and evolutionary distance between orthologs from *A. pisum* and *Ap. mellifera*.

Pairwise orthologs shared between *A. pisum* and *D. melanogaster* were identified by performing BlastP comparisons of complete protein sequence sets with a cutoff of 1×10^{-5} and identifying reciprocal best hits. Only orthologs with RefSeq model proteins in *A. pisum* were retained.

Sequence Divergence

In order to compare the evolutionary divergence of low $CpG_{O/E}$ and high $CpG_{O/E}$ orthologs between *A. pisum* and *Ap. mellifera*, a total of 2,222 orthologous protein sequences were first aligned using ClustalW (Thompson et al. 1994). Confidently, aligned gap-free columns were then extracted using Gblocks with default settings (Castresana 2000), and only long alignments ($\geq$100 amino acids) were kept for analysis. PAL2NAL was used to convert protein sequence alignments to corresponding codon alignments (Suyama et al. 2006). Finally, PAML was used to calculate rates of synonymous (d*S*) and nonsynonymous (d*N*) substitution with the "codeml" method (Yang 2007). Because synonymous substitution rates were predominantly saturated ($dS > 2$), measures of d*N* and DNA sequence percent identity were used to assess sequence divergence.

Gene Ontology

Gene ontology (GO) annotations for *D. melanogaster* orthologs of *A. pisum* proteins were used to analyze enrichment of biological process terms (Ashburner et al. 2000). GO biological process term enrichment was determined by comparing orthologs of low $CpG_{O/E}$ and high $CpG_{O/E}$ genes separately with a background composed of both low $CpG_{O/E}$ and high $CpG_{O/E}$ orthologs using the DAVID bioinformatics database functional annotation tool (Dennis et al. 2003). A Benjamini multiple-testing correction of the EASE score (a modified Fisher exact *P* value; Hosack et al. 2003) was used to determine statistical significance of GO term enrichment.

EST Mapping

Acyrthosiphon pisum expressed sequence tags (ESTs), previously used to characterize differential gene expression underlying developmental differences, sex differences, female wing polyphenism, and wing morph differences (Brisson et al. 2007), were mapped to the *A. pisum* official genes consensus set (OGS) to aid in assessing the relationship between the degree of differential gene expression among phenotypic classes and $CpG_{O/E}$. EST sequences were compared with all OGS mRNA sequences by BlastN (Altschul et al. 1997). To be considered a match, EST query sequences were required to have >50% sequence alignment to an OGS hit, >95% identity of the aligned sequence, and reciprocal best hits resulting from BlastN analysis of the OGS query against an EST database. GLEAN as well as RefSeq gene models were accepted in this case to map a greater proportion of microarray data.

Gene Expression

Brisson et al. (2007) previously examined the gene expression differences underlying distinct phenotypes in *A. pisum*

using cDNA microarrays (Wilson et al. 2006). Specifically, microarrays were utilized to determine the degree of differential gene expression in comparisons of 1) fourth instar juveniles versus adults (compared within unwinged males, within winged males, within unwinged asexual females, and within winged asexual females), 2) males versus asexual females (compared within winged fourth instars, within unwinged fourth instars, within winged adults, and within unwinged adults), 3) polyphenic winged versus unwinged females (compared within fourth instars and within adults), and finally, polymorphic winged versus unwinged males (compared within fourth instars and within adults).

For the present study, we calculated the mean of the absolute value of $\log_2$-transformed ratios across multiple comparisons to measure the degree of differential gene expression. In this manner, we combined data from all pairwise comparisons of 1) development, 2) sex, 3) female wing polyphenism, and 4) male wing polymorphism. The mean of $\log_2$-transformed gene expression ratios across all 12 pairwise comparisons was also calculated. We further divided each of these measures into two bins at a mean $|\log_2$ expression ratio$|$ value of 0.5, with genes below this threshold roughly corresponding to genes with similar expression between groups and genes above this value roughly corresponding to genes with differential expression between groups.

We also revisited analysis previously described and published by Elango et al. (2009), which demonstrated that high $CpG_{O/E}$ genes were overrepresented among genes that were differentially expressed between queen and worker castes (Grozinger et al. 2007). For the present manuscript, we analyzed NCBI transcript sequences rather than introns and exons combined, to remain consistent with our analyses of aphid gene expression.

Finally, Foret et al. (2009) previously used an oligonucleotide microarray representing the honeybee OGS (Honeybee Genome Sequencing Consortium 2006) to assess the expression breadth of genes among the following tissues in *Ap. mellifera*: antenna, brain, whole-body larva, hypopharyngeal gland, ovary, and thorax. They further demonstrated that low $CpG_{O/E}$ genes were vastly overrepresented among genes with ubiquitous expression (Foret et al. 2009). We expanded upon their analysis by splitting genes into six classes based upon the number of tissues with observed expression. To do so, we utilized lists of genes expressed in each tissue, along with a fasta file of sequences used to design the array. To map sequences with generic microarray identifiers to honeybee model RefSeq transcripts, we compared the sequences using BlastN (Altschul et al. 1997). To be considered a match, array query sequences were required to have >50% sequence alignment to a model RefSeq transcript hit and >98% identity for the aligned sequence. We then generated a numeric count of the number of tissues in which each gene was expressed (integers from 1 to 6) and recorded the $CpG_{O/E}$ for each associated model RefSeq transcript. Data for expression breadth and $CpG_{O/E}$ were obtained in this manner for a total of 7,576 *Ap. mellifera* genes.

Additional Analysis

Statistical tests (rank sum tests and correlations) were performed using either R (R Development Core Team 2010) or the JMP statistical software package (SAS Institute Inc.). Proportional Venn diagrams were generated using the Venn Diagram Plotter available from Pacific Northwest National Laboratory (http://omics.pnl.gov).

Results

We divided genes into low and high $CpG_{O/E}$ classes based on the bimodal distributions of $CpG_{O/E}$ observed in *A. pisum* ($CpG_{O/E}$ cutoff = 0.82; fig. 1*A*) and *Ap. mellifera* ($CpG_{O/E}$ cutoff = 0.72; fig. 1*B*). These two classes of genes roughly correspond to genes incurring relatively dense versus relatively sparse methylation (Saxonov et al. 2006; Suzuki et al. 2007; Weber et al. 2007; Elango et al. 2009; Foret et al. 2009; Wang and Leung 2009; Yi and Goodisman 2009; Xiang et al. 2010).

To gain insight into the evolutionary maintenance of genes with different levels of methylation, we first investigated whether genes belonging to distinct $CpG_{O/E}$ classes showed differences in their conservation of $CpG_{O/E}$ status over evolutionary time. A total of 2,339 three-way orthologs were identified with nonzero $CpG_{O/E}$ values in *A. pisum*, *Ap. mellifera*, and *D. melanogaster*. By comparing the $CpG_{O/E}$ classification of orthologs in *A. pisum* and *Ap. mellifera* from this data, we found that genes with high $CpG_{O/E}$ exhibited considerably less conservation of $CpG_{O/E}$ status than genes with low $CpG_{O/E}$ (fig. 2, table 1; Pearson's Chi-squared test with Yates' continuity correction $P = 0.0075$). Thus, patterns of dense DNA methylation have been more conserved over evolutionary time than patterns of sparse DNA methylation in *A. pisum* and *Ap. mellifera*.

We next determined whether the differential conservation of low $CpG_{O/E}$ and high $CpG_{O/E}$ status was associated with differential conservation of nucleotide and amino acid sequence. We found that genes from the low $CpG_{O/E}$ class in *A. pisum* and *Ap. mellifera* both harbored significantly greater proportions of genes with detectable three-way orthologs than genes from the high $CpG_{O/E}$ class (table 2; Pearson's Chi-squared test with Yates' continuity correction $P < 1 \times 10^{-15}$). We also found that DNA sequence conservation was significantly higher between *A. pisum* and *Ap. mellifera* orthologs from the low $CpG_{O/E}$ class than orthologs from the high $CpG_{O/E}$ class (Kruskal–Wallis rank sum test $P = 0.0003$; fig. 3*A*, supplementary table S1, Supplementary Material online). Both of these results suggested

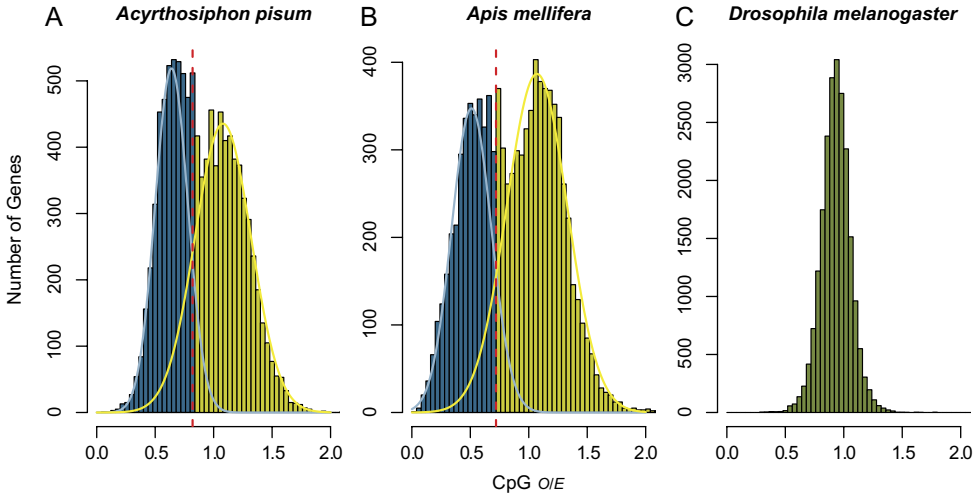

Fig. 1.—Distributions of normalized CpG dinucleotide content ($CpG_{O/E}$). (*A*) *Acyrthosiphon pisum* and (*B*) *Apis mellifera* exhibit bimodal distributions of $CpG_{O/E}$ among genes, signifying variation in germline DNA methylation levels. Dashed red lines indicate cutoffs used to divide low $CpG_{O/E}$ genes (blue) from high $CpG_{O/E}$ genes (yellow). In contrast to *A. pisum* and *Ap. mellifera*, (*C*) *Drosophila melanogaster* has a unimodal distribution of $CpG_{O/E}$ and does not exhibit substantial levels of CpG methylation.

that densely methylated genes, as a whole, were considerably more conserved at the sequence level than sparsely methylated genes. However, in contrast to the results obtained from analysis of ortholog loss and DNA sequence identity, amino acid substitution rates among genes with detectable three-way orthologs were slightly higher among low $CpG_{O/E}$ genes than high $CpG_{O/E}$ genes (Kruskal–Wallis rank sum test $P = 0.0012$; fig. 3*B* and supplementary fig. S1 and tables S1 and S2, Supplementary Material online). Furthermore, an alternate analysis, presented in our supplementary material, also found that densely methylated genes with detectable orthologs exhibited slightly higher rates of amino acid substitution than sparsely methylated genes.

To investigate whether genes with different levels of methylation were associated with specific functions, we next tested for enrichment of GO biological process terms in 4,404 *A. pisum* genes with *D. melanogaster* orthologs. We found that functions related to cellular metabolic processes were overrepresented among low $CpG_{O/E}$ genes (table 3). In contrast, functions associated with cellular signaling, behavior, and environmental stimulus were overrepresented among high $CpG_{O/E}$ genes (table 3).

We also found that six of the top ten enriched functional terms for *A. pisum* low $CpG_{O/E}$ genes were among the top ten enriched functional terms in *Ap. mellifera* low $CpG_{O/E}$ genes (table 3; Elango et al. 2009). In contrast, only two of the top ten high $CpG_{O/E}$ functional enrichment terms

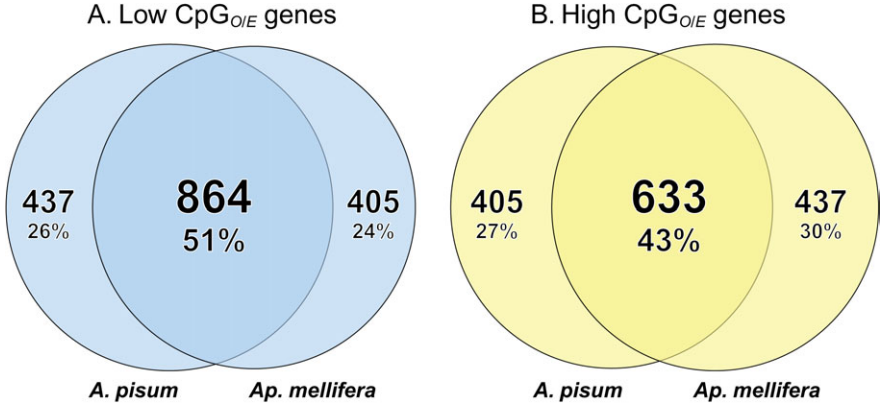

Fig. 2.—Pan-genomic high $CpG_{O/E}$ status is less conserved than low $CpG_{O/E}$ status. Analysis of orthologs in *Acyrthosiphon pisum* and *Apis mellifera* show that a higher proportion of (*A*) low $CpG_{O/E}$ genes are conserved with respect to normalized CpG content than (*B*) high $CpG_{O/E}$ genes. Each circle represents the number of genes from one species belonging to the designated $CpG_{O/E}$ class; overlap designates the number of orthologs with agreement in $CpG_{O/E}$ classification in both species.

Table 1
Contingency Table of $CpG_{O/E}$ Conservation between *Acyrthosiphon pisum* and *Apis mellifera*

	Conserved $CpG_{O/E}$ Status with *Ap. mellifera*	Nonconserved $CpG_{O/E}$ Status with *Ap. mellifera*	Proportion Conserved (%)
A. pisum low $CpG_{O/E}$ genes	864	437	66.4
A. pisum high $CpG_{O/E}$ genes	633	405	61.0

NOTE.—Conservation differs significantly between low $CpG_{O/E}$ genes and high $CpG_{O/E}$ genes (Pearson's Chi-squared test with Yates' continuity correction $P = 0.0075$).

were in agreement between *A. pisum* and *Ap. mellifera* (table 3; Elango et al. 2009). Thus, the function of low $CpG_{O/E}$ genes appears to be relatively conserved over evolutionary history.

Finally, we investigated whether $CpG_{O/E}$ measures were associated with patterns of gene expression among distinct phenotypic groups in *A. pisum* using microarray data for 1,347 genes (Brisson et al. 2007). We analyzed the degree of differential gene expression between developmental stages (development; 4th instar vs. adult), between sexes (sex; male vs. asexual female), between environmentally sensitive asexual female wing phenotypes (female wing polyphenism; winged vs. unwinged), and between genetically determined male wing phenotypes (male wing polymorphism; winged vs. unwinged).

Our results suggested that genes with low levels of DNA methylation exhibited complex, condition-specific regulation of gene expression: differential gene expression, when combined for all pairwise comparisons of alternate phenotypes, displayed a significant positive correlation with $CpG_{O/E}$ in *A. pisum* (Pearson product-moment correlation $P < 0.001$; table 4, fig. 4*A*). This signal was primarily driven by development, sex, and female wing polyphenism, which each demonstrated that differential gene expression was significantly associated with high $CpG_{O/E}$ (table 4; fig. 4*A*). Differential gene expression between male wing morphs was not significantly associated with $CpG_{O/E}$ in *A. pisum*, although the trend was in the same direction as the other tests (table 4, fig. 4*A*).

We also reanalyzed data linking gene expression to methylation levels in *Ap. mellifera* to illustrate that differential gene expression between caste phenotypes (Elango et al. 2009) and gene expression breadth (Foret et al. 2009) were also each associated with $CpG_{O/E}$ (fig. 4*B* and *C*). Specifically, genes with differential expression between *Ap. mellifera* queens and workers, and those expressed in few *Ap. mellifera* tissues, preferentially exhibited high $CpG_{O/E}$. Overall, our results reveal that genes with condition-specific regulation are associated with higher $CpG_{O/E}$ and lower levels of DNA methylation than ubiquitously expressed genes in both *A. pisum* and *Ap. mellifera*.

Discussion

Gene Evolution and DNA Methylation

We have reported distinct levels of conservation of DNA methylation status for orthologs with heavy methylation (low $CpG_{O/E}$) and sparse methylation (high $CpG_{O/E}$) in the pea aphid, *A. pisum*, and the honeybee, *Ap. mellifera* (fig. 2, table 1). In particular, a greater proportion of orthologs maintain low $CpG_{O/E}$ status than high $CpG_{O/E}$ status over evolutionary time. Thus, genes that were presumably densely methylated in the ancestor of *A. pisum* and *Ap. mellifera* were more likely to remain methylated through evolutionary time, whereas genes with sparse methylation were less likely to maintain their low methylation status.

Furthermore, we found that heavily methylated genes had a greater number of detectable orthologs and exhibited greater DNA sequence conservation than genes with sparse methylation (table 2; fig. 3*A*). In line with these results, a prior study also found that genes with signatures of methylation were enriched among orthologs that could be identified between distantly related taxa (Suzuki et al. 2007). Thus, heavily methylated genes, overall, appear to be more conserved at the sequence level than sparsely methylated genes. This observation is particularly striking because DNA methylation increases the occurrence of mutations at CpG sites and might be expected to lead to rapid DNA sequence divergence (Elango et al. 2008). One possible explanation for the observed trend, however, is that orthologs with consistently low $CpG_{O/E}$ over evolutionary history have

Table 2
Ortholog Detection among Low $CpG_{O/E}$ and High $CpG_{O/E}$ Genes

	Acyrthosiphon pisum			*Apis mellifera*		
	Three-Way Orthology	No Three-Way Orthology	Proportion with Three-Way Orthology (%)	Three-Way Orthology	No Three-Way Orthology	Proportion with Three-Way Orthology (%)
Low $CpG_{O/E}$	1,301	3,309	28.2	1,269	2,331	35.3
High $CpG_{O/E}$	1,038	4,818	17.7	1,070	4,790	18.3

NOTE.—Ortholog detection differs significantly between low $CpG_{O/E}$ genes and high $CpG_{O/E}$ genes (Pearson's Chi-squared test with Yates' continuity correction $P < 1 \times 10^{-15}$ for both *A. pisum* and *Ap. mellifera*, each analyzed separately).

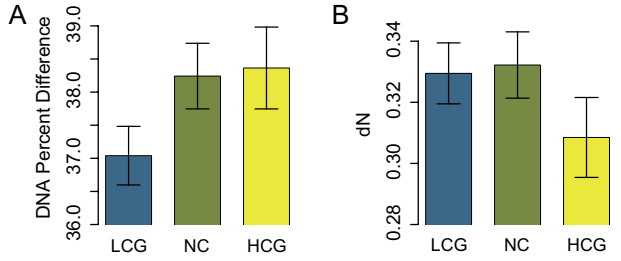

Fig. 3.—High $CpG_{O/E}$ genes exhibit significantly greater nucleotide divergence but lower amino acid divergence when compared with low $CpG_{O/E}$ genes with three-way orthology. (*A*) DNA percent difference is significantly higher between *Acyrthosiphon pisum* and *Apis mellifera* for conserved high $CpG_{O/E}$ orthologs (HCG) and orthologs with nonconserved $CpG_{O/E}$ status (NC) than those with conserved low $CpG_{O/E}$ status (LCG; Kruskal-Wallis rank sum test $P = 0.0003$). (*B*) In contrast, the nonsynonymous substitution rate (d*N*) is lower for conserved high $CpG_{O/E}$ orthologs compared with orthologs with nonconserved $CpG_{O/E}$ status or low $CpG_{O/E}$ status (Kruskal-Wallis rank sum test $P = 0.0012$). Means with 95% confidence intervals are plotted.

fewer total CpG dinucleotides than methylated genes with intermediate $CpG_{O/E}$, and thus do not incur new mutations at a comparable rate (Suzuki et al. 2009). Another possibility is that genes targeted by DNA methylation may be under greater functional constraint, as a class, than unmethylated genes.

Surprisingly, in contrast to our results from analysis of DNA sequence identity, we found that densely methylated genes with detectable orthologs may be under less constraint at the amino acid level than their sparsely methylated counterparts (fig. 3*B* and supplementary fig. S1, Supplementary Material online). Apparently, *A. pisum* and *Ap. mellifera* high and low $CpG_{O/E}$ genes that do not retain detectable orthologs in *D. melanogaster* differ more from each other, in terms of evolutionary constraint at the protein level, than do high and low $CpG_{O/E}$ genes with detectable orthologs (table 2 and supplementary tables S1 and S2, Supplementary Material online; fig. 3 and supplementary fig. S1, Supplementary Material online). It remains unclear why this may be the case, but our results suggest that different classes of genes may behave differently with respect to the interaction between selective constraints or mutability and methylation status.

Gene Expression and DNA Methylation

In the present study, we add to the emerging view that genes with ubiquitous expression in insects are preferentially targeted by DNA methylation (Elango et al. 2009; Foret et al. 2009; Xiang et al. 2010). Specifically, genes with similar expression levels among phenotypic groups exhibit evolutionary signatures of significantly higher levels of DNA methylation than genes with differential expression between

Table 3
Top 10 Enriched GO Biological Process Terms by $CpG_{O/E}$ Class for *Acyrthosiphon pisum*

$CpG_{O/E}$ Class	Accession	GO Biological Process Term	Fold Enrichment in Class	Top Ten in *Apis mellifera*[a]	Significance[b]
Low	GO:0044260	Cellular macromolecule metabolic process	1.15	No	1.72×10^{-10}
	GO:0044237	Cellular metabolic process	1.11	Yes	1.53×10^{-09}
	GO:0016070	RNA metabolic process	1.32	Yes	5.81×10^{-09}
	GO:0008152	Metabolic process	1.09	Yes	1.66×10^{-08}
	GO:0043170	Macromolecule metabolic process	1.12	Yes	3.65×10^{-08}
	GO:0006139	Nucleobase, nucleoside, nucleotide, and nucleic acid metabolic process	1.20	Yes	4.72×10^{-08}
	GO:0009987	Cellular process	1.06	Yes	3.62×10^{-07}
	GO:0009057	Macromolecule catabolic process	1.45	No	3.83×10^{-07}
	GO:0044265	Cellular macromolecule catabolic process	1.46	No	4.63×10^{-07}
	GO:0030163	Protein catabolic process	1.47	No	4.58×10^{-06}
High	GO:0007186	G protein–coupled receptor protein signaling pathway	1.72	No	2.48×10^{-05}
	GO:0007165	Signal transduction	1.28	Yes	0.0035
	GO:0007610	Behavior	1.40	No	0.0074
	GO:0003008	System process	1.30	No	0.0179
	GO:0050890	Cognition	1.43	No	0.0267
	GO:0050877	Neurological system process	1.29	No	0.0279
	GO:0032501	Multicellular organismal process	1.12	Yes	0.0280
	GO:0009581	Detection of external stimulus	1.77	No	0.0492
	GO:0009582	Detection of abiotic stimulus	1.77	No	0.0492
	GO:0006811	Ion transport	1.39	No	0.0565

[a] According to Elango et al. (2009).
[b] Benjamini multiple-testing correction of the EASE score (a modified Fisher exact *P* value).

Table 4
Correlations between *Acyrthosiphon pisum* Differential Gene Expression and $CpG_{O/E}$

	Pearson Product-Moment Correlation with $CpG_{O/E}$
Mean |$\log_2$ expression ratio| for all comparisons	0.0996***
Mean |$\log_2$ expression ratio| for developmental stages	0.1091****
Mean |$\log_2$ expression ratio| for female wing polyphenism	0.0905***
Mean |$\log_2$ expression ratio| for sexes	0.0660*
Mean |$\log_2$ expression ratio| for male wing polymorphism	0.0144

*$P < 0.05$, ***$P < 0.001$, ****$P < 0.0001$.

phenotypes in both *A. pisum* and *Ap. mellifera* (fig. 4*A* and *B*; Elango et al. 2009). Genes with ubiquitous expression among tissues are also preferentially targeted by DNA methylation in both *Ap. mellifera* (fig. 4*C*; Foret et al. 2009) and the silkworm, *Bombyx mori*, even though *B. mori* possesses only a partial complement of DNA methylation enzymes (Xiang et al. 2010). By comparison, genes with tissue-specific expression in *Ap. mellifera* (fig. 4*C*; Foret et al. 2009) and *B. mori* (Xiang et al. 2010), with caste-specific expression in *Ap. mellifera* (fig. 4*B*; Elango et al. 2009), and with differential expression between developmental stages, sexes, and polyphenic wing morphs in *A. pisum*, all exhibit lower levels of DNA methylation than their ubiquitously expressed counterparts (fig. 4*A*). Thus, sparse levels of DNA methylation are associated with flexibility in gene expression, either between polyphenic forms or different tissues.

Our results reveal that complex gene regulation is associated with low levels of DNA methylation in disparate insects. This finding may appear to contrast with the idea that DNA methylation plays an important role in the epigenetic regulation of phenotypic plasticity (Jaenisch and Bird 2003; Kucharski et al. 2008; Maleszka 2008). Indeed, our observations suggest that the primary targets of DNA methylation are those genes least likely to be implicated as leading to phenotypic variation. However, we cannot rule out the cooption of DNA methylation

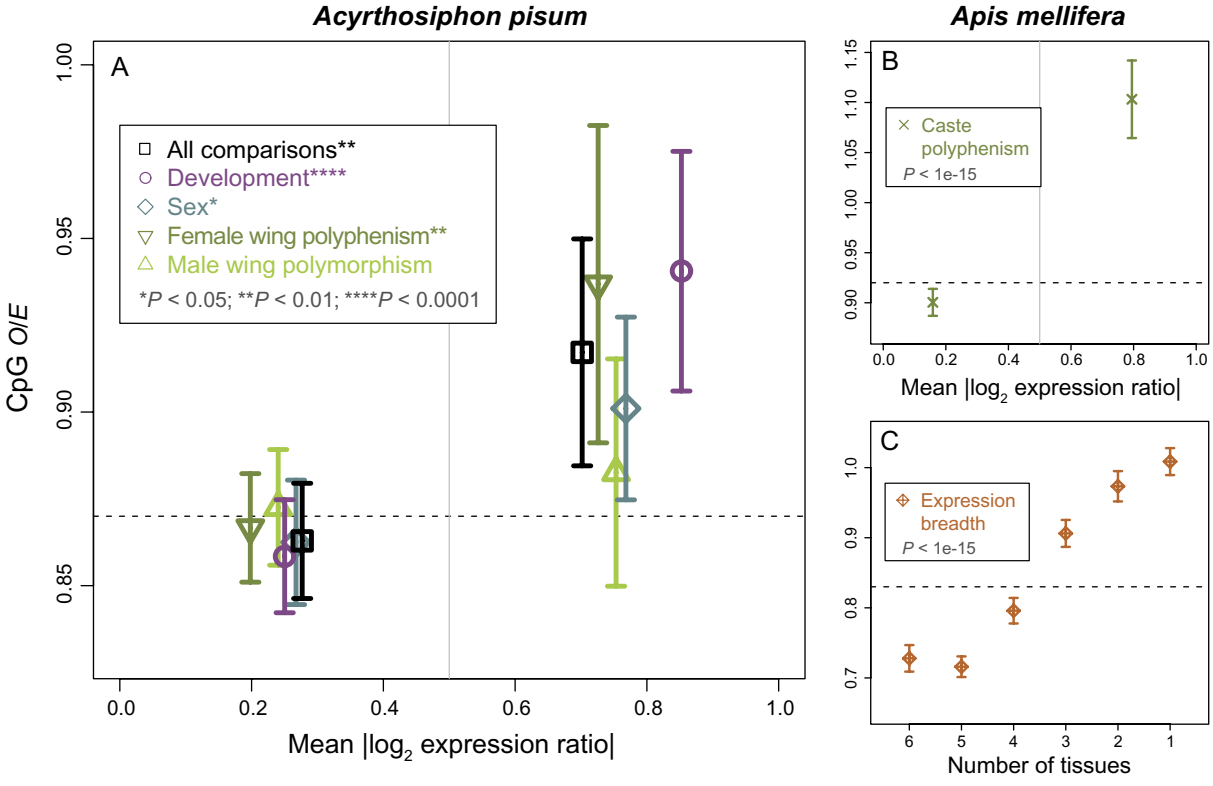

Fig. 4.—Ubiquitously expressed genes exhibit higher levels of DNA methylation than genes with condition-specific expression. (*A*) Genes with a high degree of differential expression between groups exhibit significantly higher $CpG_{O/E}$ than genes with ubiquitous expression in *Acyrthosiphon pisum*. This relationship also holds true for (*B*) differential expression between *Apis mellifera* queen and worker castes (adapted from Elango et al. 2009). (*C*) Similarly, genes with a high degree of tissue specificity exhibit significantly higher $CpG_{O/E}$ than genes with ubiquitous expression among tissues in *Ap. mellifera* (adapted from Foret et al. 2009). Significance values represent Wilcoxon signed-rank tests in panels *A* and *B* and a Kruskal-Wallis rank sum test in panel *C*. Means and 95% confidence intervals are plotted. Horizontal dashed lines represent the mean $CpG_{O/E}$ for all genes in a given panel. Vertical gray lines represent bin cutoffs for classification of genes according to mean |$\log_2$ expression ratio|.

for complex regulatory roles operating on a smaller number of loci.

Steps toward a Unified View of Intragenic Methylation

Recently, a unified view of the functional role of intragenic (vs. intergenic or promoter) DNA methylation in vertebrates and invertebrates has begun to emerge. For example, methylation of gene bodies in many vertebrates and invertebrates is associated with moderate gene expression levels (Zemach et al. 2010). Our data, obtained from microarray analyses, do not directly address overall levels of gene expression but instead address expression breadth among tissues or alternate phenotypic classes. We find that genes with high $CpG_{O/E}$ measures possess an enriched aptitude for conditional expression associated with distinct tissues or alternate phenotypes. In contrast, genes with dense methylation exhibit a greater propensity for static levels of expression.

A recent mammalian study revealed that intragenic methylation limits the generation of alternate gene transcripts by masking intragenic promoters (Maunakea et al. 2010). This mechanism may explain why broadly expressed genes are subject to the highest levels of methylation in invertebrates: broadly expressed genes may be preferentially targeted by DNA methylation due to enhanced negative effects associated with alternate promoters at such loci. Importantly, the proposed link between intragenic methylation and the regulation of alternate transcription (Maunakea et al. 2010) suggests that different levels of methylation in distinct tissues or developmental stages could have important phenotypic consequences.

Finally, we note that our results do not apply to insect taxa that have heavily diminished methylation systems (Urieli-Shoval et al. 1982; Field et al. 2004). Instead, we suggest that DNA methylation is one of many tools that can be co-opted for the purposes of gene regulation in organisms that have retained a complete enzymatic toolkit for mediating DNA methylation.

Supplementary Material

Supplementary figure S1 and tables S1–S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Sylvain Foret and Christina Grozinger for sharing data from previous investigations and Jungsun Park from the Yi laboratory for sharing code used to analyze CpG depletion. We also thank three anonymous reviewers for comments which helped to improve this manuscript. This work was supported by the U.S. National Science Foundation (grant numbers MCB-0950896 to S.V.Y. and M.A.D.G., DEB-0640690 to M.A.D.G. and S.V.Y., and DEB-1011349 to B.G.H., S.V.Y., and M.A.D.G.) and National Institute of Environmental Health Sciences (grant number 4R00ES017367 to J.A.B.).

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Ashburner M, et al. 2000. Gene Ontology: tool for the unification of biology. Nat Genet. 25:25–29.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 8:1499–1504.

Brisson JA. 2010. Aphid wing dimorphisms: linking environmental and genetic control of trait variation. Philos Trans R Soc Lond B Biol Sci. 365:605–616.

Brisson JA, Davis GK, Stern DL. 2007. Common genome-wide patterns of transcript accumulation underlying the wing polyphenism and polymorphism in the pea aphid (*Acyrthosiphon pisum*). Evol Dev. 9:338–346.

Brisson JA, Stern DL. 2006. The pea aphid, *Acyrthosiphon pisum*: an emerging genomic model system for ecological, developmental and evolutionary studies. BioEssays 28:747–755.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Dennis G, et al. 2003. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol. 4:R60.

Elango N, Hunt BG, Goodisman MAD, 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, Kim SH, Vigoda E, Yi SV. NISC Comparative Sequencing Program. 2008. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. PLoS Comput Biol. 4:e1000015.

Evans JD, Wheeler DE. 2001. Gene expression and the evolution of insect polyphenisms. BioEssays 23:62–68.

Field LM, Lyko F, Mandrioli M, Prantera G. 2004. DNA methylation in insects. Insect Mol Biol. 13:109–115.

Foret S, Kucharski R, Pittelkow Y, Lockett G, Maleszka R. 2009. Epigenetic regulation of the honey bee transcriptome: unravelling the nature of methylated genes. BMC Genomics. 10:472.

Gaunt MW, Miles MA. 2002. An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. Mol Biol Evol. 19:748–761.

Grozinger CM, Fan YL, Hoover SER, Winston ML. 2007. Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (*Apis mellifera*). Mol Ecol. 16:4837–4848.

Hendrich B, Tweedie S. 2003. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. Trends Genet. 19:269–277.

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

Hosack DA, Dennis G, Sherman BT, Lane HC, Lempicki RA. 2003. Identifying biological themes within lists of genes with EASE. Genome Biol. 4:P4.

The International Aphid Genomics Consortium. 2010. Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol. 8:e1000313.

Jaenisch R, Bird A. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 33:245–254.

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, Hedges SB. 1998. A molecular timescale for vertebrate evolution. Nature 392:917–920.

Maleszka R. 2008. Epigenetic integration of environmental and genomic signals in honey bees. Epigenetics 3:188–192.

Maunakea AK, et al. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466: 253–257.

Müller CB, Williams IS, Hardie J. 2001. The role of nutrition, crowding and interspecific interactions in the development of winged aphids. Ecol Entomol. 26:330–340.

The Nasonia Genome Working Group. 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science 327:343–348.

Ott J. 1979. Detection of rare major genes in lipid-levels. Hum Genet. 51:79–91.

R Development Core Team. 2010. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Regev A, Lamb M, Jablonka E. 1998. The role of DNA methylation in invertebrates: developmental regulation or genome defense? Mol Biol Evol. 15:880–891.

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

Saxonov S, Berg P, Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A. 103:1412–1417.

Simmen MW, et al. 1999. Nonmethylated transposable elements and methylated genes in a chordate genome. Science 283: 1164–1167.

Smith MAH, MacKay PA. 1989. Genetic variation in male alary dimorphism in populations of pea aphid, *Acyrthosiphon pisum*. Entomol Exp Appl. 51:125–132.

Stajich JE, et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12:1611–1618.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609–W612.

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

Suzuki MM, Kerr ARW, De Sousa D, Bird A. 2007. CpG methylation is targeted to transcription units in an invertebrate genome. Genome Res. 17:625–631.

Suzuki Y, Gojobori T, Kumar S. 2009. Methods for incorporating the hypermutability of CpG dinucleotides in detecting natural selection operating at the amino acid sequence level. Mol Biol Evol. 26:2275–2284.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Urieli-Shoval S, Gruenbaum Y, Sedat J, Razin A. 1982. The absence of detectable methylated bases in *Drosophila melanogaster* DNA. FEBS Lett. 146:148–152.

Walsh TK, et al. 2010. A functional DNA methylation system in the pea aphid, *Acyrthosiphon pisum*. Insect Mol Biol. 19:215–228.

Wang Y, et al. 2006. Functional CpG methylation system in a social insect. Science 314:645–647.

Wang Y, Leung F. 2009. In silico prediction of two classes of honeybee genes with CpG deficiency or CpG enrichment and sorting according to gene ontology classes. J Mol Evol. 68:700–705.

Weber M, et al. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet. 39:457–466.

Wheeler DE. 1986. Developmental and physiological determinants of caste in social Hymenoptera: evolutionary implications. Am Nat. 128:13–34.

Wilson A, et al. 2006. A dual-genome microarray for the pea aphid, *Acyrthosiphon pisum*, and its obligate bacterial symbiont, *Buchnera aphidicola*. BMC Genomics. 7:50.

Wilson EO. 1971. The insect societies. Cambridge: Harvard University Press.

Wolffe AP, Matzke MA. 1999. Epigenetics: regulation through repression. Science 286:481–486.

Xiang H, et al. 2010. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat Biotechnol. 28:516–520.

Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Yi SV, Goodisman MAD. 2009. Computational approaches for understanding the evolution of DNA methylation in animals. Epigenetics 4:551–556.

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.

Associate editor: George Weinstock