The Evolution of Invertebrate Gene Body Methylation

Shrutii Sarda, Jia Zeng, Brendan G. Hunt, and Soojin V. Yi*

School of Biology, Georgia Institute of Technology, Atlanta, Georgia

*Corresponding author: E-mail: soojinyi@gatech.edu.

Abstract

DNA methylation of transcription units (gene bodies) occurs in the genomes of many animal and plant species. Phylogenetic persistence of gene body methylation implies biological significance; yet, the functional roles of gene body methylation remain elusive. In this study, we analyzed methylation levels of orthologs from four distantly related invertebrate species, including the honeybee, silkworm, sea squirt, and sea anemone. We demonstrate that in all four species, gene bodies distinctively cluster to two groups, which correspond to high and low methylation levels. This pattern resembles that of sequence composition arising from the mutagenetic effect of DNA methylation. In spite of this effect, our results show that protein sequences of genes targeted by high levels of methylation are conserved relative to genes lacking methylation. Our investigation identified many genes that either gained or lost methylation during the course of invertebrate evolution. Most of these genes appear to have lost methylation in the insect lineages we investigated, particularly in the honeybee. We found that genes that are methylated in all four invertebrate taxa are enriched for housekeeping functions related to transcription and translation, whereas the loss of DNA methylation occurred in genes whose functions include cellular signaling and reproductive processes. Overall, our study helps to illuminate the functional significance of gene body methylation and its impacts on genome evolution in diverse invertebrate taxa.

Key words: gene body methylation, sequence evolution, functional enrichment, gene length.

Introduction

DNA methylation is a phylogenetically widespread, evolutionarily ancient epigenetic modification (Colot and Rossignol 1999; Ponger and Li 2005; Suzuki and Bird 2008). In most animals studied, DNA methylation occurs predominantly at cytosines followed by guanines or “CpG dinucleotides.” Despite the conserved units of DNA methylation, the “patterns” of genomic DNA methylation are highly variable among animal taxa. In particular, they are fundamentally different between vertebrates and invertebrates (Suzuki et al. 2007; Feng et al. 2010; Zemach et al. 2010). Vertebrate genomes are heavily methylated at most CpGs in most developmental stages and tissues (Ehrlich et al. 1982; Gama-Sosa et al. 1983), whereas invertebrate genomes generally exhibit reduced levels of DNA methylation (Suzuki and Bird 2008; Glastad et al. 2011). Most incidences of DNA methylation in invertebrate genomes occur in the form of “gene body methylation,” which refers to methylation of transcription units, including exons and introns (Suzuki et al. 2007; Feng et al. 2010; Zemach et al. 2010). From the species investigated so far, it appears that only subsets of genes are targeted by DNA methylation in invertebrates (Elango and Yi 2008; Elango et al. 2009; Feng et al. 2010; Gavery and Roberts 2010; Walsh et al. 2010; Zemach et al. 2010; Smith et al. 2011a, b; Wurm et al. 2011).

Gene body methylation is garnering increasing support as the ancestral pattern of DNA methylation in animal genomes (Suzuki et al. 2007; Elango and Yi 2008; Feng et al. 2010; Zemach et al. 2010). Interestingly, the impact of gene body methylation on gene expression appears to be diametrically different from that of gene promoter methylation. Although promoter methylation is generally associated with repression of transcription, gene body methylation is often associated with active transcription in humans and other animals (Hellman and Chess 2007; Ball et al. 2009; Maunakea et al. 2010; Xiang et al. 2010; Zemach et al. 2010). Remarkably, the association of gene body methylation with active transcription is conserved in plants, despite over a billion or more years of divergence between animals and plants (Zhang et al. 2006; Zilberman et al. 2007; Zemach et al. 2010). Elucidating the functions and evolutionary patterns of gene body methylation will have manifold consequences on our understanding of the biological significance of DNA methylation.

Until recently, empirical data on genome-wide patterns of DNA methylation were limited. In the absence of direct methylation data, a commonly used tool to investigate the patterns of genomic DNA methylation was to infer the degree of methylation from DNA sequence composition. This approach is based upon the highly mutagenetic nature of DNA methylation. Specifically, methylated cytosines are subject to frequent spontaneous deamination, which converts methyl-cytosines to thymines (Coulondre et al. 1978; Duncan and Miller 1980). In other words, DNA methylation increases the frequencies of transition mutations from CpG to TpG, thus gradually depleting CpGs from DNA sequences (Bird 1980). A measure to describe the relative deficiency of CpGs by estimating “normalized CpG frequency” (also referred to as “CpG O/E”) was proposed by Bird (1980), which has been utilized by many studies to infer the pattern of genomic DNA methylation (Suzuki et al. 2007; Elango and Yi 2008; Elango et al. 2009; Gavery and Roberts 2010; Okamura et al. 2010; Park et al. 2011).

© The Author 2012. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

However, recent advances in DNA sequencing technology have enabled researchers to perform whole-genome nucleotide resolution analysis of DNA methylation by sequencing bisulfite converted genomic DNA (Grunau et al. 2001). In particular, data on gene body methylation from several invertebrate species have recently become available (Zemach et al. 2010). Such data makes it possible to directly analyze evolutionary patterns of gene body methylation among distantly related invertebrate genomes.

Here, we investigate the patterns of gene body methylation in four distantly related invertebrate taxa. We ask whether genes are targeted by DNA methylation in a lineage-specific manner, and whether divergence in patterns of gene body methylation mirrors divergence in protein sequence among taxa. We further investigate whether several previously reported findings based on limited taxa hold true in distantly related invertebrate lineages, including the putative relationships of DNA methylation with sequence conservation (Hunt et al. 2010; Park et al. 2011; Takuno and Gaut 2011) and gene lengths (Zeng and Yi 2010). By investigating conserved and diverged evolutionary features of gene body methylation, we gain insight into its functional significance and its impact on genome evolution.

Materials and Methods

Gene Sequences and Annotations

We used data on gene body DNA methylation from four invertebrate genomes generated by Zemach et al. (2010). These include the sea anemone (Nematostella vectensis), the sea squirt (Ciona intestinalis), the honeybee (Apis mellifera), and the silkworm (Bombyx mori). We used data generated by a single study (Zemach et al. 2010) to reduce variation caused by differences in experimental conditions among similar data sets (Feng et al. 2010; Xiang et al. 2010). Following Zemach et al. (2010), levels of DNA methylation per transcription unit are represented as “fractional methylation,” which is the number of methylated cytosines divided by the total number of cytosines present in each transcription unit (gene body) (Zemach et al. 2010). Since most, if not all, of the DNA methylation in these species occurs at CpG sites (Zemach et al. 2010), fractional methylation was estimated as \[ \frac{n_{\text{CpG}}}{n_{\text{CpG}} + n_{\text{C}} + n_{\text{G}}} \] where \( n_{\text{CpG}} \) stands for methylated CpGs. We note that the fold coverage of sequencing reads per species ranged from 4.7× to 15.3× (C. intestinalis), suggesting that these measures of fractional methylation are robust (Zemach et al. 2010). Indeed, similar biological inferences to those made by Zemach et al. (2010) were obtained in a study with substantially lower sequencing coverage (Feng et al. 2010).

Measurement and Classification of CpG O/E Distribution

CpG O/E or “normalized CpG content” measures the depletion of CpG dinucleotides for genomic regions of interest (Bird 1980; Elango and Yi 2008). It is defined as

\[
\text{CpG O/E} = \frac{P_{\text{CpG}}}{P_{c} \times P_{G}},
\]

where \( P_{\text{CpG}}, P_{c}, \) and \( P_{G} \) are the frequencies of CpG dinucleotides, C nucleotides, and G nucleotides, respectively. CpG O/E was calculated for each gene, using data from gene bodies.

Model Fitting and Clustering Analyses

The density distributions of CpG O/E and fractional methylation were drawn using R (www.r-project.org), and the number of components in a mixture distribution was estimated using model-based clustering (“mclust” package in R, Fraley and Raftery 2003), following the method described in Park et al. (2011). Briefly, we estimated the number of components under the Gaussian Mixture Model, described as

\[
\sum_{i=1}^{k} p_{i}N(\mu_{i}, \sigma_{i}),
\]

where the function \( N \) is a Gaussian model containing unknown parameters \( \mu_{i} \) (mean of each component) and \( \sigma_{i} \) (standard deviation of each component), \( k \) is the number of components in the mixture model, and \( p_{i} \) is the proportion of each Gaussian model component in the mixture model. The expectation–maximization (EM) algorithm was used for this process. This approach is an improvement over the likelihood ratio test approach used previously (Elango and Yi 2008; Park et al. 2011).

Classification of Genes According to High and Low Methylation Levels

We used two different approaches to distinguish genes with high and low methylation according to fractional methylation levels. We first used a model-based clustering approach to fit mixture distributions to the observed data (above) (Park et al. 2011). This method revealed that methylation levels of all four invertebrates could be unambiguously described as mixtures of two distributions (“components”) (supplementary fig. S1, Supplementary Material online). In the second approach, we divided gene bodies into two groups, representing high and low methylation levels, excluding genes that could not be clearly classified as belonging to either group (fig. 1 and table 1). These cutoff values were chosen manually, based on observed distributions of the data (fig. 1). The use of a binomial classification of genes into low and high methylated groups is a potential caveat of our study. However, it is unlikely that this classification has caused bias in our interpretation for the following reasons. First, analyses conducted without assuming the binary classification of DNA methylation observed the same results (e.g., table 2). Second, using an automated clustering approach, instead of manually chosen cutoff values led to the same conclusions (supplementary material, Supplementary Material online). In the current manuscript, we present results based upon our manual cutoff strategy.

Ortholog Determination

To identify orthologs, complete sets of amino acid sequences for each study species were downloaded from NCBI.
We then performed pairwise FASTA searches (Pearson 1990) to identify reciprocal best hits. We identified significant hits as those satisfying the following criteria: E-value \( \leq 10^{-3} \) and the aligned segments covering at least 60% of the sequence length of the hit. We then identified three-way orthologs and four-way orthologs shared among all pairwise comparisons of the four invertebrate species. We also used BLASTP comparisons to identify orthologs (Altschul et al. 1997). The results from BLASTP and FASTA searches were highly similar. In the main text, we present data from FASTA.

**Sequence Divergence**

In order to compare the evolutionary divergence of orthologs and assess the relationship between evolutionary distance at the sequence level and the level of DNA methylation, we calculated pairwise sequence identity (exact amino acid matches/length of aligned segment) and pairwise protein distance \( d \) (number of amino acid replacements per site). We estimated protein distance following

**Fig. 1.** Distributions of CpG depletion (CpG O/E) and empirically measured methylation levels in the four invertebrate taxa. (A) The distribution of CpG O/E in the honeybee (Apis mellifera) exhibits a characteristic “bimodal” pattern. (B) Levels of experimentally measured methylation (fractional methylation levels) in the honeybee also indicate two peaks, one with low levels of DNA methylation and the other with high levels of DNA methylation. (C) CpG O/E and fractional methylation in the honeybee are highly negatively correlated. Similar patterns are found in (D–F) the silkworm (Bombyx mori), (G–I) the sea anemone (Nematostella vectensis), and (J–L) the sea squirt (Ciona intestinalis). These results indicate that genes in these species can be classified into highly methylated/CpG-depleted genes versus lowly methylated/CpG-rich genes. We used manual cutoff values to distinguish genes with low and high methylation levels, excluding those with intermediate methylation levels, in some analyses. The cutoff values used for each species are shown.
of DNA methylation. Using this matrix of DNA methylation as a binary trait, we generated ML trees using PAUP.

Analyses of Functional Enrichment Using Gene Ontology

Gene ontology (GO) annotations of orthologs in *Drosophila melanogaster* and *Homo sapiens* were analyzed following ortholog determination as described above. GO biological process term enrichment was determined by comparing groups of genes with distinct methylation profiles across taxa (described in Results and Discussion) to a background set of genes using the DAVID bioinformatics database functional annotation tool (Huang et al. 2008). For genes methylated in all taxa, full gene sets from *D. melanogaster* and *H. sapiens* were used as background gene sets. For analyses of lineage-specific methylation patterns, all available *D. melanogaster* or *H. sapiens* orthologs of genes with four-way invertebrate orthology and methylation data were used as background sets. We used “GO FAT” annotations from DAVID, which describes a subset of GO annotations that filters broad redundant terms. Default DAVID P-values and P-values following Benjamini multiple testing correction (Benjamini and Hochberg 1995) are alternately presented and demarcated.

Results and Discussion

Bimodal Patterns of Gene Body DNA Methylation Parallel Sequence Characteristics

We investigated the patterns of gene body DNA methylation in four invertebrates: the sea anemone *N. vectensis*, the sea squirt *C. intestinalis*, the honeybee *A. mellifera*, and the silkworm *B. mori* (Zemach et al. 2010). Among these species, the honeybee and the silkworm are the most closely related, yet diverged approximately 300 Ma (Douzery et al. 2004). The divergence of these arthropods from the sea squirt is estimated to have occurred around 900 Ma (Hedges et al. 2006). The divergence between these Bilateria taxa and the sea anemone (Cnidaria) is close to 1 billion years (Hedges et al. 2006). Thus, these four invertebrates encompass an exceptionally large swath of evolutionary time.

The four invertebrate species we examined encompass highly variable levels of overall gene body DNA methylation. Among these taxa, the silkworm exhibits the lowest fractional methylation levels in gene bodies (mean = 0.076 ± 0.017), followed by the honeybee (mean = 0.161 ± 0.044), sea anemone (mean = 0.276 ± 0.085), and the sea squirt (mean = 0.451 ± 0.108). However, the distributions of methylation levels in these species are not normally distributed. Instead, we identified a striking and consistent pattern among the four species. Levels of DNA methylation in all four invertebrate can be described as “bimodal,” in which low- and highly methylated gene bodies can be distinguished (fig. 1 and table 1). For example, in the honeybee, we can divide genes into two distinctive groups with respect to the fractional methylation, representing high and low methylation levels (fig. 1B). Similarly, the levels of gene

| Table 1. Levels of Gene Body DNA Methylation in Four Invertebrate Taxa. |
|-------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                        | All Genes       | Low Methylation | High Methylation |
|                        | Mean            | Median          | Mean            | Median          | Mean            | Median          |
| Silkworm                | 0.0764          | 0.0041          | 0.0018          | 0.0014          | 0.1624          | 0.1058          |
| Honeybee                | 0.3609          | 0.0487          | 0.0012          | 0.001           | 0.2983          | 0.2473          |
| Sea anemone             | 0.2758          | 0.1745          | 0.0039          | 0.0026          | 0.4515          | 0.4244          |
| Sea squirt              | 0.4505          | 0.5903          | 0.0038          | 0.0026          | 0.6343          | 0.6995          |

**Table 2. Correlations between Protein Distance and Levels of DNA Methylation in Pairwise Comparisons among Taxa.**

<table>
<thead>
<tr>
<th>Species Pair</th>
<th>Spearman’s Correlation Coefficient</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honeybee–silkworm</td>
<td>−0.135</td>
<td>5.3 × 10⁻²⁸</td>
</tr>
<tr>
<td>Honeybee–sea squirt</td>
<td>−0.092</td>
<td>6.7 × 10⁻³</td>
</tr>
<tr>
<td>Silkworm–sea squirt</td>
<td>−0.145</td>
<td>5.4 × 10⁻⁷</td>
</tr>
<tr>
<td>Honeybee–sea anemone</td>
<td>−0.115</td>
<td>4.7 × 10⁻⁷</td>
</tr>
<tr>
<td>Silkworm–sea anemone</td>
<td>−0.167</td>
<td>8.5 × 10⁻¹⁵</td>
</tr>
<tr>
<td>Sea squirt–sea anemone</td>
<td>−0.145</td>
<td>1.4 × 10⁻¹¹</td>
</tr>
</tbody>
</table>
body methylation in the silkworm (fig. 1E), the sea anemone (fig. 1H), and the sea squirt (fig. 1K) exhibit two clusters, where some genes are sparsely methylated and others exhibit relatively high levels of DNA methylation.

The measures of actual DNA methylation from four distantly related invertebrate species provide an opportunity to examine the relationship between CpG O/E and DNA methylation. In all four invertebrates, the levels of fractional methylation are highly negatively correlated with the values of CpG O/E calculated for each locus ($P < 10^{-15}$ for all cases; fig. 1). Thus, the clustering of gene bodies into groups with high and low methylation levels mirrors the clustering of gene sequences into two distinctive CpG O/E groups (fig. 1). These results indicate that deamination of methyl-cytosines appears to be mutagenetic in distantly related invertebrate species and that methylation is targeted to gene bodies of a subset of genes. Our observation also supports the premise that, in the absence of actual methylation data, CpG O/E provides a highly usable first approximation to infer the level of gene body DNA methylation.

### High Methylation Levels Are Linked to Conservation of Protein Sequences

As shown above, the widespread mutagenetic property of DNA methylation leads to a reduction of the occurrence of CpG dinucleotides in distantly related invertebrates. Based upon this observation, one might hypothesize that highly methylated genes accumulate more mutations than unmethylated genes and should thus show reduced sequence conservation. Intriguingly, recent studies from insects including the pea aphid, the honeybee, and the parasitoid wasp (Hunt et al. 2010; Park et al. 2011), as well as the model plant Arabidopsis (Takuno and Gaut 2011), contradict this conjecture. These studies found that methylated genes are more conserved than sparsely methylated or nonmethylated genes. We sought to determine if these observations represent a common pattern among distantly related invertebrate animals. Thus, we examined the relations between methylation status and protein distance in orthologs among the four invertebrate species in our study. We found that mean fractional methylation levels and protein distances were negatively correlated in all possible pairwise comparisons (table 2), suggesting a universal association between DNA methylation and protein conservation in invertebrates.

To examine the relationship between methylation status and sequence conservation more deeply, we divided pairwise orthologs into the following categories: genes that exhibit high methylation levels in both species (conserved high methylation), genes that exhibit low methylation levels in both species (conserved low methylation), and genes that exhibit high methylation levels in one species and low methylation levels in the other (unconserved methylation). We found that the rate of ortholog detection is higher for genes that are highly methylated in both species when compared with genes with conserved low methylation or unconserved methylation (table 3). We then compared the protein distances of genes within each category. We found that “conserved high methylation” genes were significantly more conserved at the level of protein sequence than genes belonging to “unconserved methylation” or “conserved low methylation” groups (fig. 2).

Our results suggest that sequence conservation of highly methylated genes is a common feature of invertebrate genome evolution. This relationship may arise through the preferential targeting of conserved genes by methylation (Suzuki et al. 2007) or, alternatively, through the repression of non-CpG mutations in methylated DNA (Park et al. 2011; Takuno and Gaut 2011).

#### Phylogenetic Tree of Gene Body Methylation Captures Species Relationships and Lineage-Specific Changes in DNA Methylation

We utilized the observation that genes can be assigned into categories based on low or high methylation status in each species to investigate the evolutionary relationships of taxa according to gene body methylation. Specifically, we compared phylogenies constructed based on the evolutionary patterns of gene body methylation to phylogenies constructed based on protein sequence evolution.

Figure 3A illustrates a consensus ML tree of the four species generated from the alignment of amino acid sequences representing 563 orthologous proteins in each species (supplementary table S2, Supplementary Material online). As expected, the two insect species, the honeybee and...
silkworm, group together. The other two invertebrates, the sea anemone and the sea squirt, form another group. Note that this is a paraphyletic group and the actual root of tree lies somewhere along the branch leading to the sea anemone (Hedges et al. 2006).

We observed notable variation in protein evolutionary rates between species (fig. 3A): The branch leading to the silkworm *B. mori* is longer than that leading to the honeybee *A. mellifera*, as shown previously (Zdobnov and Bork 2007). The branch leading to *C. intestinalis* is longer than that leading to *N. vectensis* (fig. 3A), even though the root of the tree would be located on the sea anemone branch. This observation agrees with a previous study, demonstrating that the sea squirt lineage is evolving faster than the sea anemone lineage (Putnam et al. 2008). Thus, the amino acid sequence phylogeny captures the known species relationships and evolutionary rate variation among taxa. It is also notable that the distance between the two insects is comparable with the distance between the two other invertebrates (fig. 3A), even though the actual divergence time between the insect species is approximately one-third of that separating the other two invertebrates (Douzery et al. 2004; Hedges et al. 2006).

Figure 3B illustrates a consensus ML tree constructed from gene body methylation patterns in each of the four invertebrate species. This tree demonstrates a parallel pattern to the species relationship shown in the sequence tree: The two insect species group together, whereas the other two distant invertebrates form another group (fig. 3B). Apart from this similarity, the two trees differ dramatically. For example, the branch connecting the arthropod common ancestor to the other two invertebrates is much longer in the gene body methylation tree (fig. 3B) as compared with the sequence tree (fig. 3A). Another clear distinction is the honeybee branch, which in comparison to the other taxa retains a long length in the gene body methylation tree (fig. 3B).

To examine the difference between these two trees further, we evaluated a tree distance statistic, the $K$ tree score (Soria-Carrasco et al. 2007). The $K$ tree score increases as two trees differ in terms of relative branch lengths (Soria-Carrasco et al. 2007). The distribution of the $K$ tree scores between the observed amino acid sequence tree and 10,000 bootstrapped sequence trees, as well as the $K$ tree score between the observed amino acid sequence tree and gene body methylation tree, are shown in figure 3C. We found that the tree generated from gene body methylation exhibits a highly significantly different evolutionary pattern than the sequence tree. In particular, the insect lineages (especially the honeybee lineage) show evidence of enriched evolution of gene body methylation relative to other taxa.

### Functional Significance of Methylated Genes in Invertebrates

Among the 563 four species orthologs we identified, 429 are classified as highly methylated in all species (supplementary table S2, Supplementary Material online). This is reflective of the trend that highly methylated genes tend to be more conserved at the sequence level (see above). Thus, most evolutionary conserved genes are highly methylated among distantly related invertebrate species. We investigated whether these genes belong to specific functional categories by examining enrichment of specific GO classifications. Among the 429 highly methylated genes, 147 genes have GO annotations from *D. melanogaster* and 185 have GO annotations from *H. sapiens*. Table 4 presents...
lists the top 10 GO biological processes enriched in both fly and human orthologs. The enriched categories, as compared with the genomic background, include housekeeping functions, such as translation, ribosome biogenesis, RNA splicing, and protein localization. Previous studies demonstrated that genes with housekeeping functions tend to exhibit signatures of high levels of DNA methylation (Elango et al. 2009; Hunt et al. 2010). Our observation adds to these findings and suggests that housekeeping genes have been heavily methylated throughout invertebrate evolution.

The sea squirt <i>C. intestinalis</i> and the sea anemone <i>N. vectensis</i> exhibit very few lineage-specific changes of gene body DNA methylation following their divergence from a common ancestor. In fact, there are only three genes with high levels of methylation in <i>C. intestinalis</i> and low levels of methylation in the other three species (supplementary table S2, Supplementary Material online), and no genes are present which exhibit the opposite pattern. Similarly, only three genes exhibit low methylation in sea anemone that are highly methylated in the other three species (supplementary table S2, Supplementary Material online), and no genes are present which exhibit the opposite pattern. Some of these genes have orthologs in humans with GO annotations, but the small numbers of genes prevent us from inferring meaningful functional trends.

The branch of the gene body methylation tree connecting <i>N. vectensis</i> and <i>C. intestinalis</i> to the two arthropods is very long, representing many changes between the two insect species and the other two invertebrates (fig. 3B). Indeed, there are 32 genes that exhibit high methylation in the sea squirt and sea anemone but low methylation in the honeybee and the silkworm (interestingly, no gene exhibits the opposite pattern). This could either be caused by changes from low to high methylation in the sea squirt and the sea anemone lineages or from high to low methylation in the lineage leading to insects. Since the root of the tree lies on the <i>N. vectensis</i> branch, it is more parsimonious to infer that the ancestral pattern of gene body methylation was high methylation for these genes and that these genes changed to low methylation in the lineage leading to the insects. For 15 of these genes, GO annotations in <i>Drosophila</i> orthologs were available. These genes tend to function in cellular signaling pathways, including phosphorylation of proteins, synaptic transmission, and cell–cell signaling (table 5). At face value, this observation suggests that the role of DNA methylation in signal transmission has changed between the insect lineages and the other invertebrates. However, due to small sample size, this finding is speculative at present.

Within insects, there are many lineage-specific changes in DNA methylation status. In the lineage leading to the silkworm, there are 30 genes that are highly methylated in other lineages and sparsely methylated in the silkworm. <i>Drosophila</i> orthologs of these genes with GO annotations (nine genes) show functional terms in development, immune response, defense response, cellular respiration, and phosphorylation. Finally, there are 61 genes that exhibit low methylation levels in the honeybee and high methylation levels in the other three species. This reflects the long length of the honeybee branch in the gene body methylation phylogeny (fig. 3B). Available GO annotations of <i>Drosophila</i> orthologs (15 genes) show that these genes are enriched for reproductive processes (table 5).

There have been frequent lineage-specific changes in gene body DNA methylation during invertebrate evolution and in insects in particular. Our examination of functional categories of orthologous genes in <i>Drosophila</i> and humans suggests that some of these changes were targeted to specific functional categories. It is tempting to hypothesize that genes belonging to specific functional groups have changed between low and high methylation levels in
response to lineage-specific evolutionary events, such as adaptation to new environments or accommodation of newly evolved developmental pathways. However, lineage-specific changes are represented by only a small number of genes in the present data and functional enrichment cannot be statistically substantiated. Nevertheless, these findings warrant future studies of lineage-specific changes in gene body methylation.

Insect-Specific Relationship between Gene Lengths and DNA Methylation

In a previous study, we observed that in the honeybee, genes harboring signatures of low DNA methylation were significantly longer than those with high methylation (Zeng and Yi 2010). By using empirical methylation data from distantly related invertebrates, we sought to determine whether a consistent relationship between gene lengths and DNA methylation is common among invertebrate taxa.

Figure 4 illustrates mean gene lengths of genes with low and high methylation levels in each species. In the honeybee, we confirm the pattern observed previously (Zeng and Yi 2010), where genes that exhibit low levels of DNA methylation are significantly longer than highly methylated genes ($P < 10^{-15}$). Interestingly, we observe a qualitatively similar pattern in the silkworm. Genes with low methylation levels are significantly longer than highly methylated genes ($P < 10^{-15}$), although the degree of bias is substantially reduced in the silkworm compared with the honeybee (fig. 4). In contrast, we observe a reversal of the relationship between gene length and DNA methylation in the other two invertebrates, where genes with low methylation levels are significantly shorter than those with high methylation levels (fig. 4, $P < 10^{-15}$ in both cases) (Nanty et al. 2011).

### Table 4. Characteristics of Genes that Are Heavily Methylated in All Four Species (top 10 enriched terms for each of two ortholog sets).

<table>
<thead>
<tr>
<th>Ortholog Set</th>
<th>GO Biological Process</th>
<th>Accession</th>
<th>Fold Enrichment</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drosophila melanogaster</td>
<td>Translation GO:0006412</td>
<td>GO:0006412</td>
<td>3.11</td>
<td>4.04 x 10^{-5}</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>Spindle elongation GO:0001531</td>
<td>GO:0001531</td>
<td>7.52</td>
<td>4.04 x 10^{-4}</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>Mitotic spindle elongation GO:0000022</td>
<td>GO:0000022</td>
<td>7.61</td>
<td>5.38 x 10^{-4}</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>Intracellular protein transport GO:0006886</td>
<td>GO:0006886</td>
<td>4.03</td>
<td>0.015</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>Cellular protein localization GO:0003463</td>
<td>GO:0003463</td>
<td>3.92</td>
<td>0.016</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>Protein transport GO:00015031</td>
<td>GO:00015031</td>
<td>3.11</td>
<td>0.016</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>Mitotic spindle organization GO:0007052</td>
<td>GO:0007052</td>
<td>3.62</td>
<td>0.017</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>Protein localization in organelle GO:00033365</td>
<td>GO:00033365</td>
<td>5.34</td>
<td>0.017</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>Establishment of protein localization GO:0006605</td>
<td>GO:0006605</td>
<td>4.95</td>
<td>0.018</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>Translational elongation GO:0006414</td>
<td>GO:0006414</td>
<td>13.03</td>
<td>2.97 x 10^{-11}</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>Translation GO:0006412</td>
<td>GO:0006412</td>
<td>5.74</td>
<td>1.73 x 10^{-9}</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>Ribonucleoprotein complex biogenesis GO:00022613</td>
<td>GO:00022613</td>
<td>7.72</td>
<td>1.56 x 10^{-8}</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>Ribosome biogenesis GO:00042254</td>
<td>GO:00042254</td>
<td>8.99</td>
<td>3.04 x 10^{-7}</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>RNA processing GO:0006396</td>
<td>GO:0006396</td>
<td>3.61</td>
<td>5.67 x 10^{-6}</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>rRNA processing GO:0006364</td>
<td>GO:0006364</td>
<td>7.95</td>
<td>8.15 x 10^{-4}</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>RNA splicing GO:0008380</td>
<td>GO:0008380</td>
<td>4.12</td>
<td>9.81 x 10^{-4}</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>rRNA metabolic process GO:0016072</td>
<td>GO:0016072</td>
<td>7.62</td>
<td>9.94 x 10^{-4}</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>RNA splicing, via transesterification reactions GO:0000375</td>
<td>GO:0000375</td>
<td>5.74</td>
<td>0.001</td>
</tr>
</tbody>
</table>

### Table 5. Characteristics of Genes that Show Insect- and Honeybee-Specific Methylation Loss (all significantly enriched genes prior to multiple test correction are listed).

<table>
<thead>
<tr>
<th>Taxon-Specific Low Methylation (high in others)</th>
<th>GO Biological Process</th>
<th>Accession</th>
<th>Fold Enrichment</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apis mellifera and Bombyx mori</td>
<td>Protein amino acid phosphorylation GO:0006468</td>
<td>GO:0006468</td>
<td>5.69</td>
<td>0.020</td>
</tr>
<tr>
<td>A. mellifera and B. mori</td>
<td>Synaptic transmission GO:0007268</td>
<td>GO:0007268</td>
<td>7.68</td>
<td>0.044</td>
</tr>
<tr>
<td>A. mellifera and B. mori</td>
<td>Transmission of nerve impulse GO:0019226</td>
<td>GO:0019226</td>
<td>7.68</td>
<td>0.044</td>
</tr>
<tr>
<td>A. mellifera and B. mori</td>
<td>Cell–cell signaling GO:0007267</td>
<td>GO:0007267</td>
<td>7.68</td>
<td>0.044</td>
</tr>
<tr>
<td>A. mellifera only</td>
<td>Reproductive cellular process GO:0048610</td>
<td>GO:0048610</td>
<td>3.95</td>
<td>0.023</td>
</tr>
<tr>
<td>A. mellifera only</td>
<td>Sexual reproduction GO:0019953</td>
<td>GO:0019953</td>
<td>3.06</td>
<td>0.027</td>
</tr>
<tr>
<td>A. mellifera only</td>
<td>Regulation of protein kinase activity GO:0045889</td>
<td>GO:0045889</td>
<td>8.68</td>
<td>0.033</td>
</tr>
<tr>
<td>A. mellifera only</td>
<td>Regulation of phosphorus metabolic process GO:0051174</td>
<td>GO:0051174</td>
<td>8.68</td>
<td>0.033</td>
</tr>
<tr>
<td>A. mellifera only</td>
<td>Regulation of kinase activity GO:0043549</td>
<td>GO:0043549</td>
<td>8.68</td>
<td>0.033</td>
</tr>
<tr>
<td>A. mellifera only</td>
<td>Regulation of phosphorous metabolic process GO:0042325</td>
<td>GO:0042325</td>
<td>8.68</td>
<td>0.033</td>
</tr>
<tr>
<td>A. mellifera only</td>
<td>Regulation of transferase activity GO:00051338</td>
<td>GO:00051338</td>
<td>8.68</td>
<td>0.033</td>
</tr>
<tr>
<td>A. mellifera only</td>
<td>Multicellular organism reproduction GO:0032504</td>
<td>GO:0032504</td>
<td>2.89</td>
<td>0.034</td>
</tr>
<tr>
<td>A. mellifera only</td>
<td>Reproductive process in a multicellular organism GO:0048609</td>
<td>GO:0048609</td>
<td>2.89</td>
<td>0.034</td>
</tr>
</tbody>
</table>
and gene lengths changed dramatically between insects able if the relationship between gene expression breadths our results suggest that this hypothesis would only be vi-
specific genes tend to be longer than housekeeping genes
methylations modulate the relationship between gene lengths
and DNA methylation), we find that many genes have
methylated genes are more conserved at the sequence level
methylation tend to represent "housekeeping" functions.
our study reveals features of gene body DNA methylation
that are conserved among distantly related invertebrate
species. First, we show that genes targeted by DNA
methylation tend to represent "housekeeping" functions. Second, our study confirms previous reports that highly
methylated genes are more conserved at the sequence level
than genes with low methylation levels. Although most of
the genes we identify as orthologs are highly methylated
(due to the above relation between sequence evolution
and DNA methylation), we find that many genes have
undergone changes in DNA methylation in insect lineages.
Interestingly, most of these changes are from high to low
DNA methylation levels and some of these genes exhibit
enrichment for specific functional categories. Investigating
more closely related insect taxa in detail will provide oppor-
tunities to test the validity of these putative functional
associations. Our study demonstrates that evolutionary
analysis of gene body methylation is a powerful approach
to investigate the functional roles of DNA methylation
and the impact of epigenetic modifications on genome
evolution.

Supplementary Material
Supplementary tables and figures are available at Molecular
Biology and Evolution online (http://www.mbe.oxfordjournals. org/).

Acknowledgments
We thank the members of the Zilberman group for making
data on gene body methylation available. We also thank
Eric Gaucher and Ziming Zhao for help with phylogenetic
analyses. This study is supported by the Bioinformatics
Masters program and National Science Foundation grants
to S.Y. (MCB-0950896, BCS-0751481).

References
generation of protein database search programs. Nucleic Acids
Res. 25:3389–3402.
Church GM. 2009. Targeted and genome-scale strategies reveal
27:361–368.
Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate:
a practical and powerful approach to multiple testing. J R Statist
Bird A. 1980. DNA methylation and the frequency of CpG in animal
DNA. Nucleic Acids Res. 8:1499–1504.
Colot V, Rossignol J-L. 1999. Eukaryotic DNA methylation as an
basis of base substitution hotspots in Escherichia coli. Nature
274:775–780.
Douzery EJP, Snell EA, Bapteste E, Fdr Delsuc, Philippe H. 2004. The
timing of eukaryotic evolution: does a relaxed molecular clock
reconcile proteins and fossils? Proc Natl Acad Sci U S A.
101:15386–15391.
Duncan BK, Miller JH. 1980. Mutagenic deamination of cytosine
Ehrlich M, Gama-Sosa MA, Huang L-H, Midgett RM, Kuo KC,
McCune RA, Gehrke C. 1982. Amount and distribution of 5-
methylcytosine in human DNA from different types of tissues or
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
Elango N, Yi SV. 2008. DNA methylation and structural and
and divergence of methylation patterning in plants and animals.
estimation, and discriminant analysis software: MCLUST. J
Gama-Sosa MA, Midgett RM, Slagel VA, Githens S, Kuo KC,
methylation in various mammals. Biochim Biophys Acta. 740:
212–219.
Gavery MR, Roberts SB. 2010. DNA methylation patterns provide
insight into epigenetic regulation in the Pacific oyster
Glastad J, Hunt BG, Yi SV, Goodisman MAD. 2011. DNA
methylation in insects: on the brink of the epigenomic era.
Insect Mol Biol. 20:553–565.

Fig. 4. Comparisons of the lengths of genes with low and high
methylation levels in four invertebrate species. In the honeybee
and the silkworm, genes with high methylation levels are significantly
shorter than genes with low methylation levels. In contrast, genes
with high methylation levels are significantly longer than genes with
low methylation levels in the sea squirt and the sea anemone.