

## Research

# Comparative methylomics reveals gene-body H3K36me3 in *Drosophila* predicts DNA methylation and CpG landscapes in other invertebrates

Lisa Nanty,<sup>1,4</sup> Guillermo Carbajosa,<sup>1,4</sup> Graham A. Heap,<sup>1,4</sup> Francis Ratnieks,<sup>2</sup> David A. van Heel,<sup>1</sup> Thomas A. Down,<sup>3</sup> and Vardhman K. Rakan<sup>1,5</sup>

<sup>1</sup>The Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, United Kingdom; <sup>2</sup>Laboratory of Apiculture and Social Insects, Department of Biological and Environmental Science, University of Sussex, Brighton BN1 9QG, United Kingdom; <sup>3</sup>The Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge CB2 1QN, United Kingdom

In invertebrates that harbor functional DNA methylation enzymatic machinery, gene-bodies are the primary targets for CpG methylation. However, virtually all other aspects of invertebrate DNA methylation have remained a mystery until now. Here, using a comparative methylomics approach, we demonstrate that *Nematostella vectensis*, *Ciona intestinalis*, *Apis mellifera*, and *Bombyx mori* show two distinct populations of genes differentiated by gene-body CpG density. Genome-scale DNA methylation profiles for *A. mellifera* spermatozoa reveal CpG-poor genes are methylated in the germline, as predicted by the depletion of CpGs. We find an evolutionarily conserved distinction between CpG-poor and CpG-rich genes: The former are associated with basic biological processes, the latter with more specialized functions. This distinction is strikingly similar to that recently observed between euchromatin-associated genes in *Drosophila* that contain intragenic histone 3 lysine 36 trimethylation (H3K36me3) and those that do not, even though *Drosophila* does not display CpG density bimodality or methylation. We confirm that a significant number of CpG-poor genes in *N. vectensis*, *C. intestinalis*, *A. mellifera*, and *B. mori* are orthologs of H3K36me3-rich genes in *Drosophila*. We propose that over evolutionary time, gene-body H3K36me3 has influenced gene-body DNA methylation levels and, consequently, the gene-body CpG density bimodality characteristic of invertebrates that harbor CpG methylation.

[Supplemental material is available for this article.]

Methylation of CpG sites is a key epigenetic feature of many eukaryotic genomes, being involved in a variety of essential processes such as gene regulation, parental imprinting, and silencing of repeat elements. Traditionally, the role of CpG methylation has been studied in the context of well-characterized genomic features such as promoters, CpG islands (CGIs), and imprinted regions in human, mouse, and *Arabidopsis*. More recently though, advances in genome-wide DNA methylation profiling methods have allowed less biased, discovery-based, whole-genome studies (i.e., DNA methylomics) in many different species (Illingworth et al. 2008; Meissner et al. 2008; Rakan et al. 2008; Lister et al. 2009; Feng et al. 2010; Li et al. 2010; Xiang et al. 2010; Zemach et al. 2010). We now know that DNA methylation landscapes are far more complex than previously appreciated, and DNA methylation dynamics are likely to have a functional impact at a wide variety of genomic elements, including nonpromoter CGIs, enhancers, and gene-bodies.

Some of these recent studies also delineated, for the first time, CpG methylation profiles in a range of invertebrate species (Feng et al. 2010; Xiang et al. 2010; Zemach et al. 2010). Analysis of these profiles has led to suggestions that CpG methylation is an ancient epigenetic regulatory mechanism of eukaryotic cells that has been lost in various invertebrate lineages, rather than having arisen multiple times (Zemach and Zilberman 2010). Interestingly, in in-

vertebrates that contain functional orthologs of mammalian DNA methyltransferase families 1–3, gene-bodies are the primary targets for CpG methylation (Suzuki et al. 2007; Feng et al. 2010; Zemach et al. 2010). In fact, although overall genomic DNA methylation landscapes vary significantly among vertebrates, plants, and invertebrates, the targeting of methylation to gene-bodies is evolutionarily conserved. However, apart from this observation, virtually all other aspects of invertebrate DNA methylation have remained a mystery until now. We therefore reasoned that a “comparative methylomics” approach, which integrates DNA methylomic and genomic data from a range of invertebrate and other eukaryotic species, could uncover key evolutionarily conserved properties of invertebrate gene-body methylation and thus help provide important biological and evolutionary insights into the DNA methylation system in invertebrates.

## Results

### Invertebrate genomes show two distinct populations of genes differentiated by gene-body CpG density

We considered eight different eukaryotic species for which DNA methylomes have recently been generated: invertebrate species *Nematostella vectensis* (sea anemone), *Ciona intestinalis* (sea squirt), *Apis mellifera* (honey bee), and *Bombyx mori* (silkworm); plant species *Oryza sativa* (rice) and *Arabidopsis thaliana*; and vertebrate species *Tetraodon nigroviridis* (puffer fish) and *Homo sapiens* (Lister et al. 2009; Zemach et al. 2010). It is known that CpG methylation can influence the underlying sequence composition as methylated

<sup>4</sup>These authors contributed equally to this work.

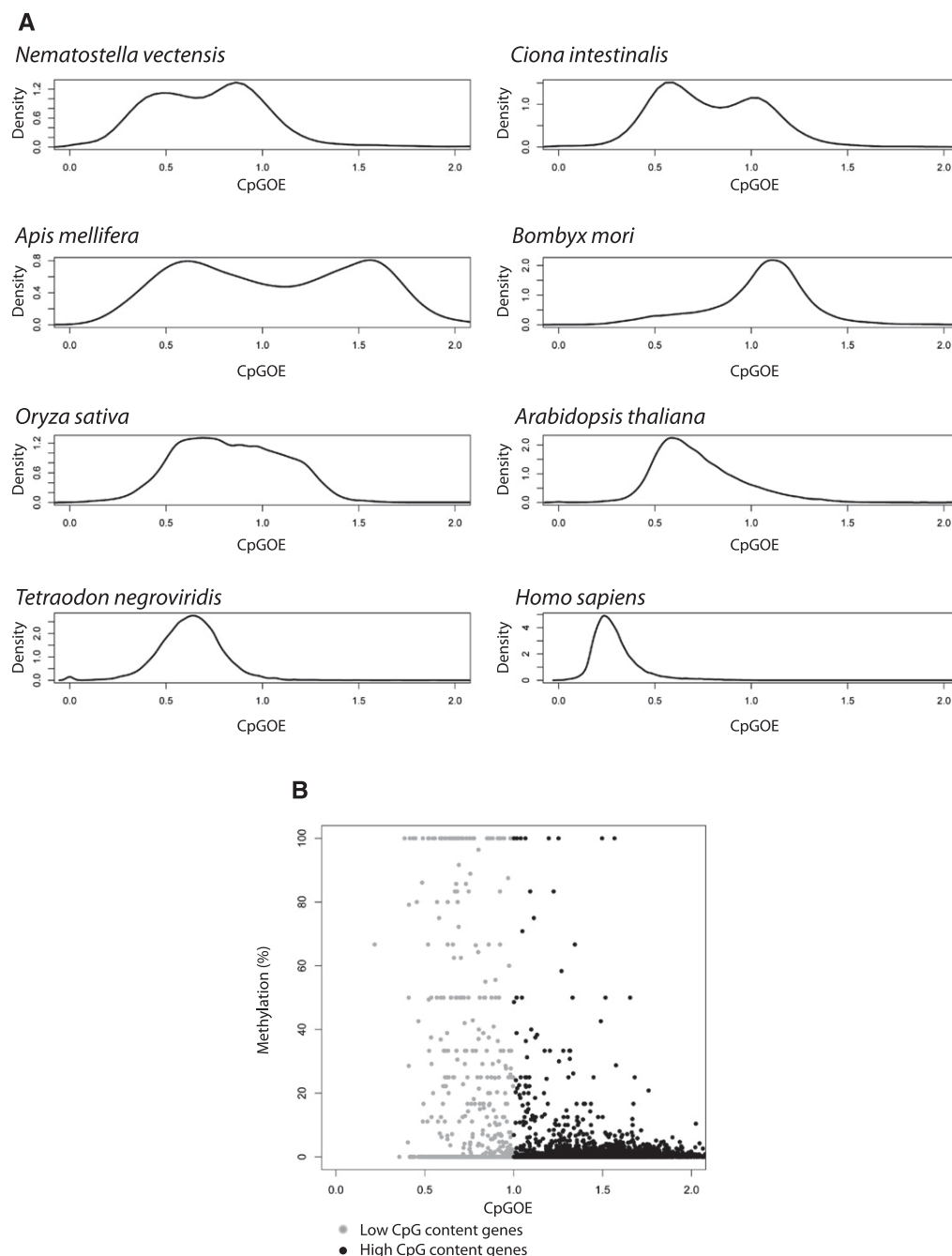
<sup>5</sup>Corresponding author.

E-mail [v.rakan@qmul.ac.uk](mailto:v.rakan@qmul.ac.uk).

Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.121640.111>.

cytosines that spontaneously deaminate to thymine in the germline are gradually lost over evolutionary time (Saxonov et al. 2006). We therefore analyzed gene-body  $\text{CpG}_{\text{observed/expected}}$  ( $\text{CpG}_{\text{O/E}}$ ) profiles and found that vertebrate and plant species, with the exception of *O. sativa*, show essentially unimodal gene-body  $\text{CpG}_{\text{O/E}}$  profiles

(Fig. 1A). On the other hand, *N. vectensis*, *C. intestinalis*, and *A. mellifera* show clearly bimodal gene-body  $\text{CpG}_{\text{O/E}}$  profiles. This has been noted previously for *C. intestinalis* (Suzuki et al. 2007) and *A. mellifera* (Elango et al. 2009), and these studies also demonstrated that this bimodality is a CpG-specific effect as G + C content



**Figure 1.** Gene-body CpG content and methylation in various eukaryotic species. (A)  $\text{CpG}_{\text{O/E}}$  ratio distribution kernel density estimates for gene bodies for different eukaryotic genomes. Gene bodies were defined as the sequence between the transcriptional start and termination sites annotated in the latest genome build for each species (see Methods). (B) We generated RRB-seq data for *A. mellifera* mature spermatozoa (from a pool of six drones). Average gene-body  $\text{CpG}_{\text{O/E}}$  and DNA methylation levels were calculated for each gene (see Methods). Plotted are 933 low-CpG genes (light gray) and 4537 high-CpG genes (dark gray). Please also refer to Supplemental Figure S1 in which the similar plots are shown but using only those genes for which we had data from at least three CpG sites (with similar conclusions).

does not show concomitant bimodality, and that bimodal gene-body CpG<sub>o/e</sub> profiles are observed even when introns and exons are considered separately. Therefore, for *N. vectensis*, *C. intestinalis*, and *A. mellifera*, it is possible to define two gene categories: low-CpG genes (CpG<sub>o/e</sub> < 0.7 for *N. vectensis* and *C. intestinalis*, CpG<sub>o/e</sub> < 1.0 for *A. mellifera*) and high-CpG genes (CpG<sub>o/e</sub> ≥ 0.7 for *N. vectensis* and *C. intestinalis*, CpG<sub>o/e</sub> ≥ 1.0 for *A. mellifera*).

The gene-body CpG density profiles yielded a number of other interesting observations: (1) The median gene-body CpG<sub>o/e</sub> varied widely among the different species (regardless of bimodality); (2) *B. mori* seemed to display a small peak at CpG<sub>o/e</sub> ≈ 0.5; and (3) the profile for *O. sativa* was more complex than any other species. In order to explore these issues further, we integrated the gene-body CpG<sub>o/e</sub> data with DNA methylomic profiles, as described below.

### Low- and high-CpG genes in invertebrates are differentially methylated in somatic and germline tissues

The above observations predict that low-CpG genes should be significantly hypermethylated relative to high-CpG genes in the germline. To prove this, at least for *A. mellifera*, we generated genome-scale DNA methylation profiles for *A. mellifera* mature spermatozoa using reduced representation bisulfite sequencing (RRBseq) (Meissner et al. 2008). Of the genes included in the analysis, we found an order of magnitude difference in the proportion of low-CpG genes (210/933 i.e., 23%) compared with high CpG-genes (109/4537 i.e., 2.4%) that showed >10% methylation (Fig. 1B; Supplemental Fig. S1). Reanalysis of published random shotgun bisulfite sequencing (BS-seq)-based DNA methylomes for *N. vectensis*, *C. intestinalis*, and *A. mellifera* (albeit somatic tissues only) confirmed a strong inverse correlation between gene-body CpG content and DNA methylation levels (Fig. 2). So, although it has been previously noted that these invertebrates harbor gene-body methylation, our analysis shows that this is mostly limited to low-CpG genes in the soma and germline and that the reduced CpG density in low-CpG genes is most likely due to the mutagenic effect of DNA methylation in the germline. However, the possibility that a degree of selection may also be leading to CpG depletion cannot be ruled out, as we observed a number of unmethylated low-CpG and methylated high-CpG genes in *A. mellifera* sperm (Fig. 1B).

Integration with DNA methylomic and gene-length data also allowed us to address some of the outstanding issues from the previous section. First, the large variation in median gene-body CpG densities among the different species, regardless of bimodality, is very strongly correlated with the predominant intragenic methylation state in each species. For example, human genes are mostly methylated and relatively CpG-poor (median CpG<sub>o/e</sub> ≈ 0.2), whereas *B. mori* genes are mostly unmethylated (with the exception of a relatively small subset of genes discussed below) and CpG-rich (median CpG<sub>o/e</sub> ≈ 1.1). Hence the median CpG density of genes in any given eukaryotic genome is most likely a consequence of evolutionary effects of DNA methylation (most likely in the germline). Although it has been suggested that there is a correlation between DNA methylation and gene length (Zilberman et al. 2007) and this could be true in some species (e.g., methylated genes in insects appear to be smaller), Figure 2 demonstrates that this is not an evolutionarily conserved relationship either among invertebrates or, generally, among eukaryotic genomes that harbor DNA methylation.

Next, although there seemed to be a small peak at CpG<sub>o/e</sub> ≈ 0.5 in the *B. mori* profiles in Figure 1A, it was not clear whether this was evidence of gene-body CpG content bimodality as observed for the

other invertebrates. By separating out genes based on DNA methylation and length, a degree of bimodality becomes evident for *B. mori*. However, because the number of methylated or partially methylated genes is relatively low in this species (81 genes), we used a different statistical approach, compared with that used for the other invertebrates, to test if bimodality is associated with gene-body methylation status. We classified *B. mori* genes as methylated if they displayed an average gene-body DNA methylation ≥ 33%, and as unmethylated if the methylation was <33%. Then, we further categorized the genes as low-CpG genes if the gene-body CpG<sub>o/e</sub> was <1 or otherwise as high-CpG genes. Of the 11,048 unmethylated genes, 3790 (34%) were low-CpG and 7258 (66%) were high-CpG genes. On the other hand, of the 81 methylated genes, 47 (58%) were low-CpG genes and 34 (42%) were high-CpG genes. A Pearson's  $\chi^2$  test with Yates' continuity correction revealed the difference between these proportions to be significant ( $P < 0.001$ ). So even though *B. mori* genes are mostly methylation-free, consistent with the fact that its genome harbors only a partial DNA methylation system compared to the other invertebrates considered here, a small degree of gene-body CpG content bimodality does exist and may be the vestiges of a much more pronounced bimodality that existed earlier in the evolutionary history of *B. mori*.

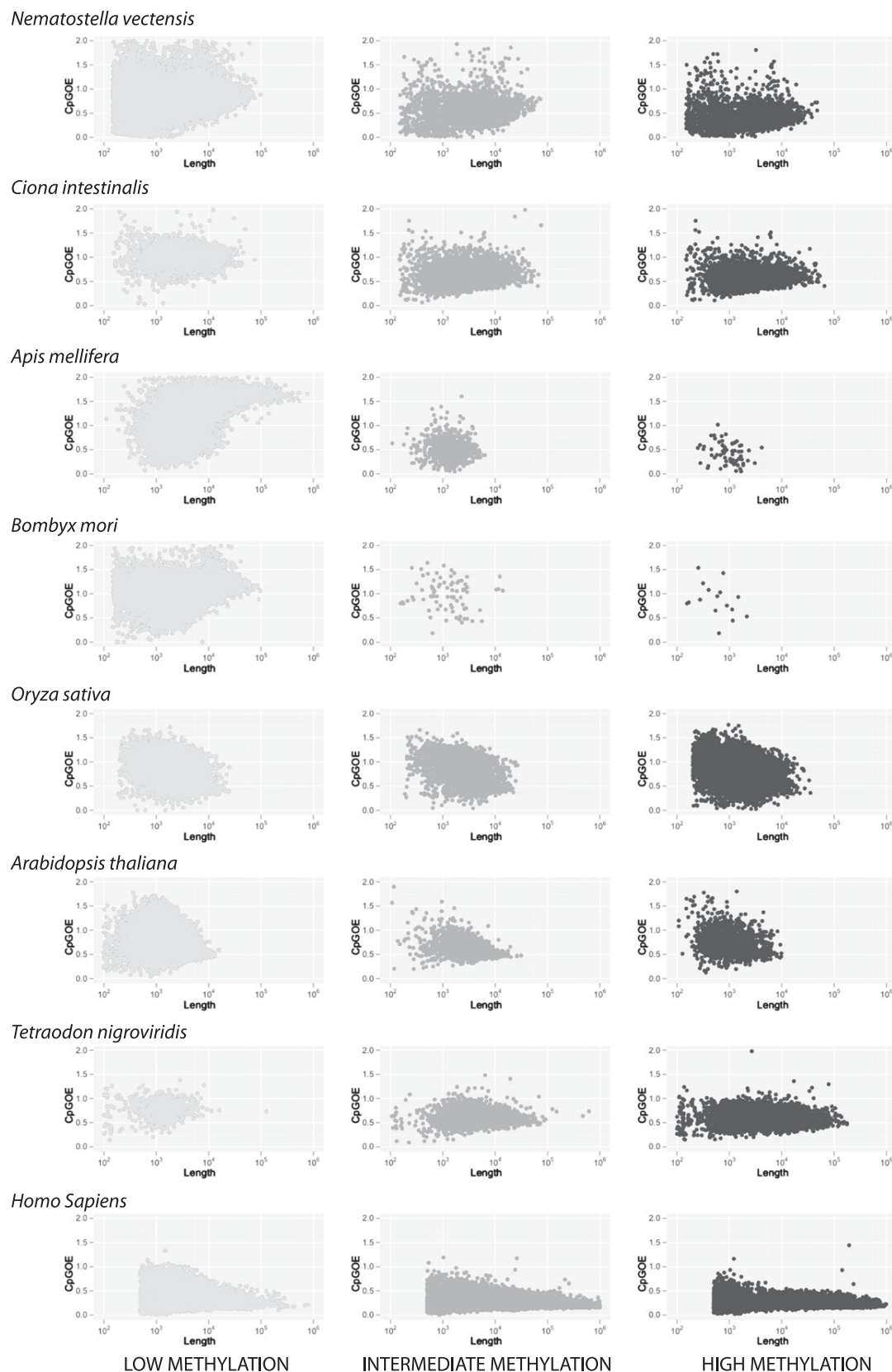
As noted above, the gene-body CpG<sub>o/e</sub> profiles for *O. sativa* were not as clear-cut as the others. Unfortunately, neither separation by gene length nor DNA methylation state could shed any further light on the reasons for the distinctive gene-body CpG<sub>o/e</sub> profiles for this organism.

### Exons in invertebrate genomes display characteristic peaks of CpG methylation

Gene-body DNA methylation in mammals is characterized by "peaks" of DNA methylation within exons relative to the surrounding introns (Lister et al. 2009). We found similar exonic methylation peaks in *N. vectensis*, *C. intestinalis*, and *A. mellifera*, most prominently in low-CpG genes (Fig. 3; it was not possible to generate a similar plot for *B. mori* due to the very low levels of methylation). The functional consequence of exonic methylation peaks is unclear, although it has been suggested that it could be involved in alternative splicing (Zemach et al. 2010). Whatever the true function of exonic methylation might be, it seems that it is an evolutionarily conserved feature in organisms that contain gene-body DNA methylation.

### Gene expression differences between low- and high-CpG genes are not evolutionarily conserved

It is well established that mammalian genomes display CpG content bimodality at promoters (Saxonov et al. 2006). The functional impact at high- and low-CpG promoters is thought to vary though: Methylation at high-CpG promoters is associated with gene silencing, whereas methylation at low-CpG promoters is thought to have less of an influence (Weber et al. 2007). It has also been reported that gene-body methylation correlates with expression levels (Zilberman et al. 2007). Given these previous observations, we decided to investigate whether there is an evolutionarily conserved relationship among gene-body CpG density, methylation, and expression levels in *N. vectensis*, *C. intestinalis*, and *A. mellifera* (Fig. 4; it was not possible to perform a similar analysis for *B. mori* due to the very low levels of methylation). The analyses were performed using previously published whole-genome mRNA-seq data from Zemach et al. 2010. Even though the size of the error bars

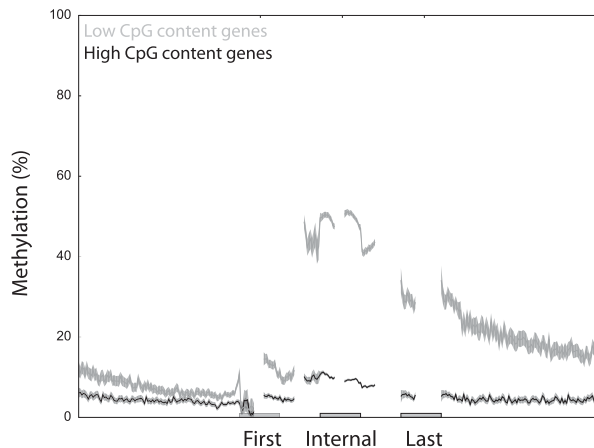


**Figure 2.** Relationship among gene-body  $CpG_{O/E}$ , average gene-body DNA methylation levels, and gene length. The average methylation value for each gene was obtained by averaging the methylation values over all CpG sites (for which data were available) contained within the gene-body. Raw BS-seq data are from Cokus et al. (2008), Zemach et al. (2010), and Lister et al. (2009).

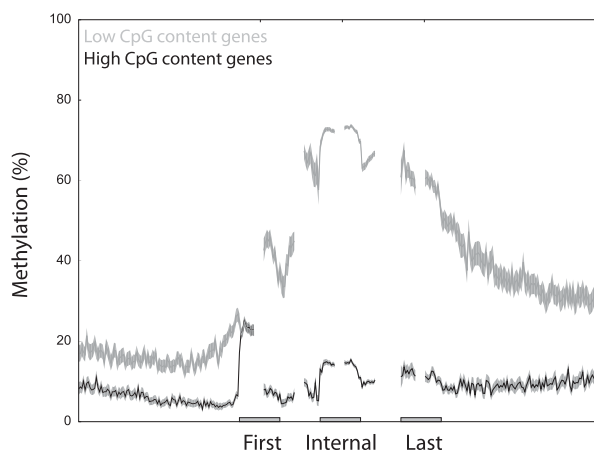
are relatively large, it is clear that low-CpG genes are generally expressed at higher levels relative to high-CpG genes in *N. vectensis*; there is no detectable difference between the two gene categories in *C. intestinalis*; and low-CpG genes are expressed at higher levels relative to high-CpG genes in *A. mellifera*. Furthermore, there was

no evidence that the correlation between methylation and expression levels was significantly different between low- and high-CpG genes for any of the three species. Therefore, we were unable to find an evolutionarily conserved relationship among gene-body CpG content, methylation, and gene expression levels that can distinguish low- from high-CpG genes in the various invertebrates under consideration here.

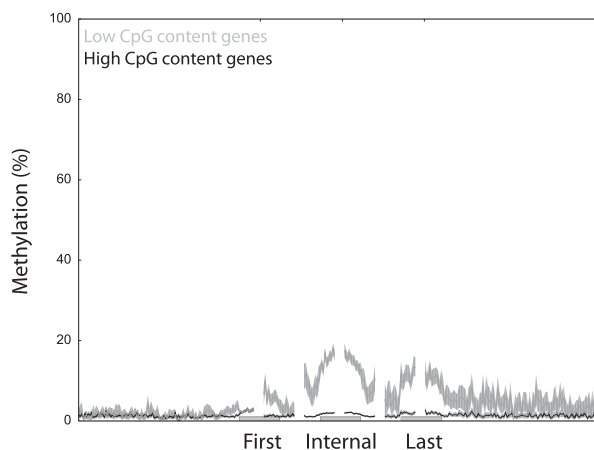
#### *Nematostella vectensis*



#### *Ciona intestinalis*



#### *Apis mellifera*



#### CpG-poor genes in *N. vectensis*, *C. intestinalis*, and *A. mellifera* are orthologs of H3K36me3-rich genes in *Drosophila*

It has been reported that low- and high-CpG genes are associated with different functional classes of genes in *A. mellifera*: Low-CpG genes are associated with basic biological processes, whereas high-CpG genes with developmental processes and signaling pathways (Elango et al. 2009). We found that similar ontological distinctions also exist between high- and low-CpG genes in *N. vectensis*, *C. intestinalis*, and *B. mori* (Table 1).

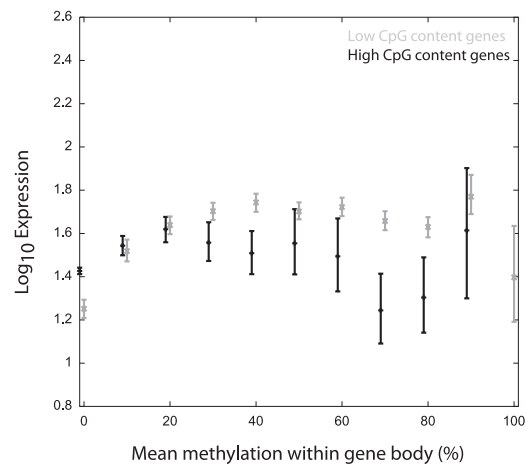
More strikingly, the difference in GO term enrichment between low- and high-CpG genes is reminiscent of that recently observed between “YELLOW” and “RED” chromatin-associated genes in *Drosophila melanogaster* (Filion et al. 2010), an organism that does not display bimodal gene-body CpG<sub>0/e</sub> profiles or CpG methylation. The *Drosophila* genome can be segmented into five principal chromatin types defined by combinations of chromatin-associated factors (Filion et al. 2010). Two of these types, termed YELLOW and RED, are associated with transcriptionally active euchromatin but with a key difference between the two categories: YELLOW chromatin-associated genes harbor significant H3K36me3 (histone 3 lysine 36 trimethylation) within gene bodies, whereas RED chromatin-associated genes contain very little H3K36me3. Until recently, it was generally assumed that gene-body H3K36me3 is a universal marker of active transcription, but Filion et al. (2010) found that the range of expression levels associated with YELLOW and RED genes is similar. Another major distinction noted by the investigators is that YELLOW genes display broad expression patterns and are associated with basic biological processes (e.g., enriched for GO terms such as “ribosome” and “DNA repair”), whereas RED genes display restricted expression patterns and are enriched for GO terms such as “transcription factor activity” and “signal transduction.” We therefore wondered if there is a correlation between low- and high-CpG genes in *N. vectensis*, *C. intestinalis*,

**Figure 3.** Exonic methylation peaks are prominent in low-CpG genes. DNA methylation values were calculated for 2-kb windows centered on the transcriptional start site, “internal” exons (i.e., not first or last exons), and last exons. Trend lines represent the median methylation values for all genes within that category for which data were available, and vertical, faint gray lines show the standard deviations. (Light gray) Low-CpG genes; (dark gray) high-CpG genes. The trend lines are interrupted because exon sizes vary within any genome, and it is not straightforward to generate a composite plot that accurately represents methylation profiles over the entire exon. Our plots show methylation levels 100 bp upstream of and downstream from exon–intron junctions. Raw BS-seq data are from Zemach et al. (2010). Number of genes plotted: 10,827 low-CpG and 14,566 high-CpG genes (*N. vectensis*), 3821 low-CpG and 4336 high-CpG genes (*C. intestinalis*), and 4064 low-CpG and 4446 high-CpG genes (*A. mellifera*). High-CpG genes in *C. intestinalis* seem to be associated with a small peak of methylation at the transcriptional start site (TSS). However, this is simply due to the cut-off between low and high-CpG genes not being perfect; that is, the high CpG gene category is likely to contain some low CpG genes. Also, incorrect genomic annotation (i.e., incorrectly marked TSSs) is also likely to have an influence. None of this appreciably affects our overall conclusions since the vast majority of the “signal” is from the appropriate gene category.

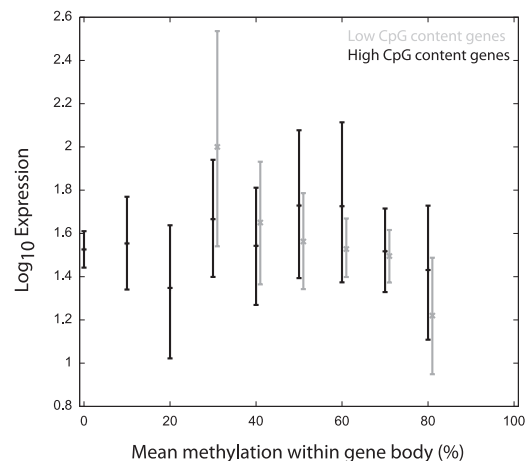


*A. mellifera*, and *B. mori* and YELLOW and RED genes in *Drosophila*. Indeed, orthologs of YELLOW genes displayed significantly lower gene-body CpG density than orthologs of RED genes in all four

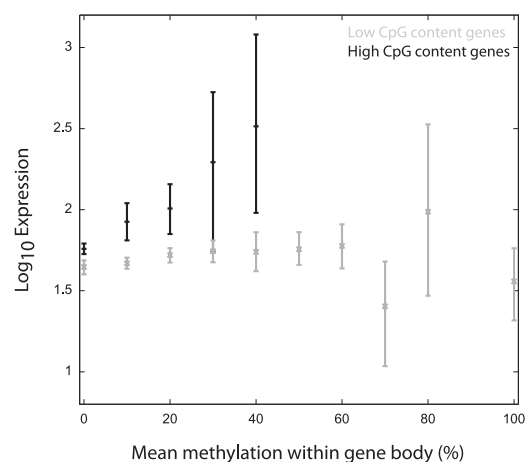
#### *Nematostella vectensis*



#### *Ciona intestinalis*



#### *Apis mellifera*



species (Fig. 5). Consistent with this finding, YELLOW gene orthologs display significantly higher methylation levels than RED gene orthologs in *N. vectensis*, *C. intestinalis*, and *A. mellifera* (Fig. 5; this analysis could not be performed for *B. mori* due to the very small number of methylated genes). We also found that YELLOW and RED genes in *Drosophila* do not show significant differences in gene-body CpG density (Supplemental Fig. S2). Interestingly, re-analysis of published RNA-seq data revealed that RED gene orthologs in *N. vectensis* and *A. mellifera* were expressed at slightly higher levels compared with YELLOW gene orthologs, in contrast to what is observed in *Drosophila* (in which RED and YELLOW genes are expressed at roughly similar levels). This analysis could not be performed for *C. intestinalis* or *B. mori* due to insufficient expression data for RED gene orthologs.

## Discussion

Based on the above observations and assuming that YELLOW and RED gene orthologs in *N. vectensis*, *C. intestinalis*, *A. mellifera*, and *B. mori* are also differentially marked by H3K36me3, we propose that over evolutionary time, gene-body H3K36me3 bimodality has influenced DNA methylation levels and, consequently, helped shape gene-body CpG density profiles in these invertebrates. This is supported by the following observations: (1) H3K36me3 has been found in every eukaryotic species thus far examined for this mark, whereas CpG methylation occurs in only some of these species. Therefore, H3K36me3 can be targeted to gene bodies, including distinctive peaks of enrichment within internal exons (Kolasinska-Zwierz et al. 2009), in the absence of CpG methylation. (2) Recently, a comparison of gene-body CpG<sub>o/e</sub> and gene length in *A. mellifera* and *Acyrthosiphon pisum* (pea aphid) revealed that low-CpG genes tend to be small, leading the investigators to suggest a correlation between DNA methylation and gene length (Fig. 1B; Hunt et al. 2010). The *A. pisum* methylome was not actually analyzed by Hunt et al. (2010), but rather inferred. In the analysis presented here, we observed that low-CpG genes in *B. mori* are also smaller than high-CpG genes. However, Hunt et al. (2010) also noted that gene lengths between *A. mellifera* and *Drosophila* display a very strong correlation. We therefore hypothesized that the correlation of gene-lengths among *A. mellifera*, *A. pisum*, *B. mori*, and *Drosophila* is more likely due to differences in H3K36me3 gene body marking. Comparison of gene lengths between YELLOW and RED chromatin-associated genes in *Drosophila* revealed that YELLOW genes (whose orthologs in *A. mellifera* are mostly low-CpG, methylated, and relatively short) are significantly smaller than RED genes (80% of YELLOW genes are <4.2 kb, whereas only 55% of RED genes are <4.2 kb, with 4.2 kb being the gene-size at which the two distributions intersect) (Fig. 6). Consistent with this finding, there is no correlation between gene-body DNA methylation and gene length in either *N. vectensis* or *C. intestinalis* (Fig. 2). So although there is a relationship between DNA methylation and gene-length in the DNA methylation-containing insects considered here, this relationship is not conserved among invertebrates. (3) Jeltsch and colleagues

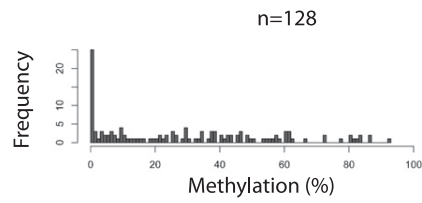
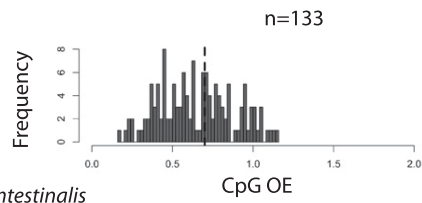
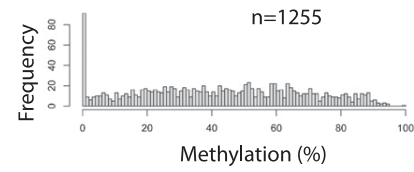
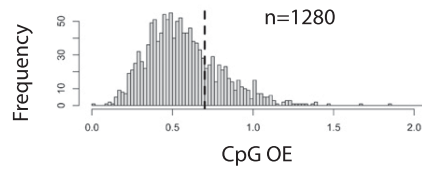
**Figure 4.** Expression levels of low- and high-CpG genes. Plotted are  $\log_{10}$  RPKM values obtained from Zemach et al. (2010). The x-axis represents the average expression values within 10% methylation bins spanning 0%–100%. Error bars, SD. For *A. mellifera*, there were very few data points available for high-CpG genes with >40% gene-body methylation. Plotted are 9442 low-CpG and 12,116 high-CpG genes (*N. vectensis*), 448 low-CpG and 561 high-CpG genes (*C. intestinalis*), and 3077 low-CpG and 2838 high-CpG genes (*A. mellifera*). (Light gray) Data for low-CpG genes; (dark gray) high-CpG genes.

**Table 1.** Gene ontology (GO) analysis of low- and high-CpG genes

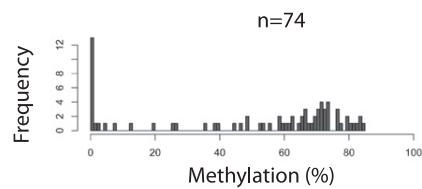
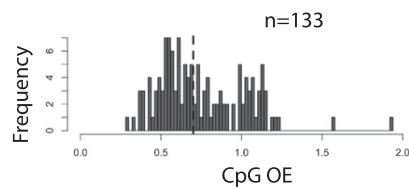
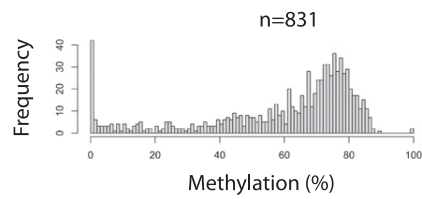
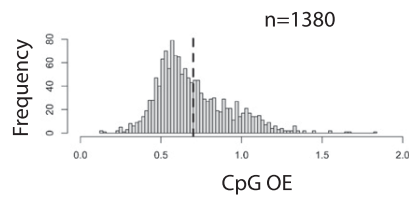
	High-CpG genes			Low-CpG genes		
	GO term	Function	P-value	GO term	Function	P-value
<i>N. vectensis</i>	7166	Cell surface receptor linked signaling pathway	$5.68 \times 10^{-90}$	8152	Metabolic process	$8.78 \times 10^{-92}$
	7186	G-protein coupled receptor protein signaling pathway	$2.63 \times 10^{-87}$	44237	Cellular metabolic process	$3.42 \times 10^{-78}$
	23033	Signaling pathway	$1.04 \times 10^{-61}$	44238	Primary metabolic process	$6.05 \times 10^{-53}$
	23052	Signaling	$1.27 \times 10^{-53}$	9987	Cellular process	$1.83 \times 10^{-52}$
	6323	DNA packaging	$9.48 \times 10^{-19}$	44260	Cellular macromolecule metabolic process	$5.95 \times 10^{-48}$
	65004	Protein-DNA complex assembly	$1.46 \times 10^{-18}$	44267	Cellular protein metabolic process	$7.60 \times 10^{-47}$
	34728	Nucleosome organization	$1.46 \times 10^{-18}$	43170	Macromolecule metabolic process	$5.34 \times 10^{-28}$
	6334	Nucleosome assembly	$1.46 \times 10^{-18}$	19538	Protein metabolic process	$1.50 \times 10^{-22}$
	31497	Chromatin assembly	$1.46 \times 10^{-18}$	46907	Intracellular transport	$5.87 \times 10^{-20}$
	6333	Chromatin assembly or disassembly	$8.62 \times 10^{-16}$	6886	Intracellular protein transport	$4.29 \times 10^{-19}$
<i>C. intestinalis</i>	23033	Signaling pathway	$1.90 \times 10^{-5}$	8152	Metabolic process	$3.52 \times 10^{-9}$
	7166	Cell surface receptor linked signaling pathway	$3.16 \times 10^{-5}$	44238	Primary metabolic process	$2.31 \times 10^{-8}$
	7186	G-protein coupled receptor protein signaling pathway	0.000191168	19538	Protein metabolic process	$6.26 \times 10^{-7}$
	23052	Signaling	0.000209058	44237	Cellular metabolic process	$2.34 \times 10^{-6}$
	50794	Regulation of cellular process	0.01420171	44267	Cellular protein metabolic process	$1.23 \times 10^{-5}$
	34621	Cellular macromolecular complex subunit organization	0.023818712	43170	Macromolecule metabolic process	$7.65 \times 10^{-5}$
	34622	Cellular macromolecular complex assembly	0.023818712	6091	Generation of precursor metabolites and energy	0.000152545
				43412	Macromolecule modification	0.000168429
				9056	Catabolic process	0.000192749
				9987	Cellular process	0.000352586
<i>A. mellifera</i>	32501	Multicellular organismal process	$9.88 \times 10^{-8}$	9987	Cellular process	$1.85 \times 10^{-21}$
	48513	Organ development	$2.01 \times 10^{-5}$	44260	Cellular macromolecule metabolic process	$1.05 \times 10^{-20}$
	48731	System development	$2.78 \times 10^{-5}$	44237	Cellular metabolic process	$6.88 \times 10^{-20}$
	7155	Cell adhesion	$5.93 \times 10^{-5}$	6139	Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	$8.82 \times 10^{-16}$
	22610	Biological adhesion	$5.93 \times 10^{-5}$	43170	Macromolecule metabolic process	$5.08 \times 10^{-15}$
	3008	System process	0.000206375	90304	Nucleic acid metabolic process	$1.18 \times 10^{-14}$
	7186	G-protein coupled receptor protein signaling pathway	0.000444507	16070	RNA metabolic process	$1.70 \times 10^{-14}$
	32502	Developmental process	0.000494197	10467	Gene expression	$7.14 \times 10^{-14}$
	7275	Multicellular organismal development	0.00052241	34641	Cellular nitrogen compound metabolic process	$1.29 \times 10^{-13}$
	48856	Anatomical structure development	0.001744484	44238	Primary metabolic process	$1.98 \times 10^{-11}$
<i>B. mori</i>	23052	Signaling	$2.22 \times 10^{-8}$	44237	Cellular metabolic process	$3.87 \times 10^{-22}$
	7165	Signal transduction	$4.37 \times 10^{-8}$	44260	Cellular macromolecule metabolic process	$3.38 \times 10^{-21}$
	7186	G-protein coupled receptor protein signaling pathway	$1.01 \times 10^{-6}$	6412	Translation	$8.49 \times 10^{-20}$
	7166	Cell surface receptor linked signaling pathway	$1.09 \times 10^{-6}$	44267	Cellular protein metabolic process	$5.29 \times 10^{-17}$
	50794	Regulation of cellular process	$3.70 \times 10^{-6}$	9987	Cellular process	$7.98 \times 10^{-17}$
	50789	Regulation of biological process	$5.04 \times 10^{-6}$	10467	Gene expression	$6.63 \times 10^{-12}$
	65007	Biological regulation	$6.06 \times 10^{-6}$	43170	Macromolecule metabolic process	$1.26 \times 10^{-9}$
	7155	Cell adhesion	$9.33 \times 10^{-5}$	44249	Cellular biosynthetic process	$3.78 \times 10^{-9}$
	22610	Biological adhesion	$9.33 \times 10^{-5}$	34645	Cellular macromolecule biosynthetic process	$4.03 \times 10^{-9}$
	32502	Developmental process	0.000260739	9059	Macromolecule biosynthetic process	$4.48 \times 10^{-9}$
<i>D. melanogaster</i>	Red genes function			Yellow genes function		
		Multicellular organismal development Plasma membrane Behavior Transcription factor activity Cellular component movement Receptor binding Extracellular region Proteinaceous extracellular matrix			DNA repair Ribosome Structural constituent of ribosome DNA metabolic process Structural molecule activity Nucleic acid metabolic process Nucleus Intracellular	

Over-representation of GO terms and adjusted *P*-values were calculated using R packages (Falcon and Gentleman 2007; [www.R-project.org](http://www.R-project.org)). Only the top 10 categories (ranked by *P*-value) are shown for each species, and for *C. intestinalis*, only seven categories yielded *P* < 0.05 (corrected) for high-CpG genes. For comparative purposes, the results of the GO analysis of YELLOW and RED genes in *Drosophila* performed by Filion et al. 2010 are reproduced here verbatim.

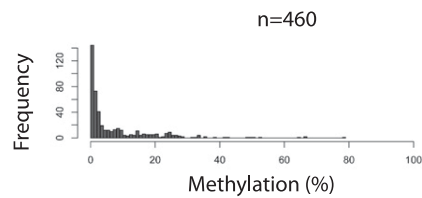
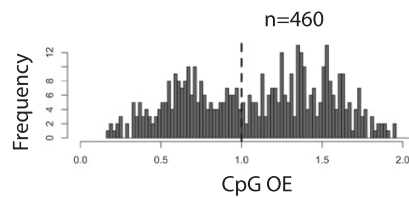
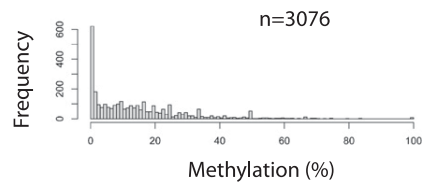
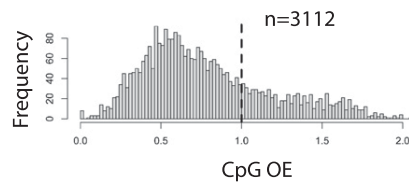
*Nematostella vectensis*



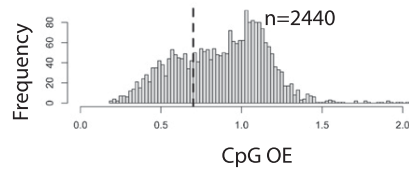
*Ciona intestinalis*




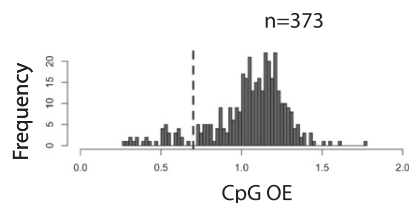
*Apis mellifera*




*Bombyx mori*



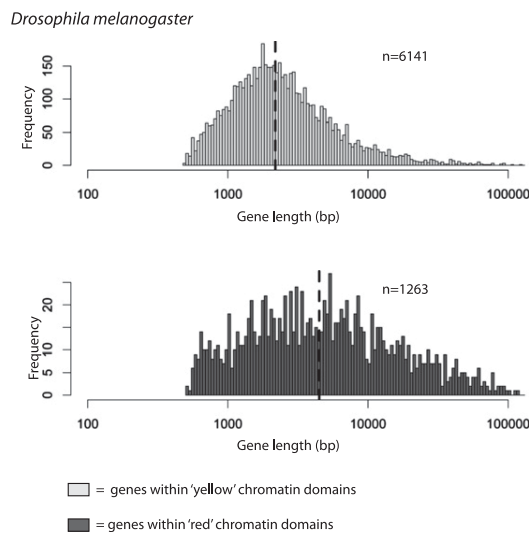
 = orthologs of *Drosophila melanogaster* genes within 'yellow' chromatin domains



 = orthologs of *Drosophila melanogaster* genes within 'red' chromatin domains

**Figure 5.** Gene-body CpG content and methylation profiles of YELLOW and RED gene orthologs. Plotted are frequency distributions of gene-body CpG density (left panels) and DNA methylation (right panels). The CpG density cut-off between low-CpG and high-CpG genes in the four organisms (0.7 for *N. vectensis*, *C. intestinalis*, and *B. mori*; 1.0 for *A. mellifera*) is indicated by a vertical dashed line. YELLOW gene orthologs (light gray) display significantly lower gene-body CpG density and higher gene-body CpG methylation compared with RED gene orthologs (dark gray) in all cases ( $P < 0.005$  for all except *C. intestinalis* gene-body CpG density,  $P = 0.034$ , Wilcoxon rank sum tests with continuity correction). The DNA methylation analysis could not be performed for *B. mori* due to the very small number of methylated genes.





**Figure 6.** Comparison of gene lengths of YELLOW and RED chromatin-associated genes in *Drosophila*. The two distributions intersect at ~4.2 kb. Plotted are 6141 YELLOW (light gray) and 1263 RED genes (dark gray).

recently used peptide arrays to show that the PWWP domain of mammalian DNA methyltransferase 3A recognizes the H3K36me3 mark (Dhayalan et al. 2010). Orthologs of DNMT3 are found in *N. vectensis*, *C. intestinalis*, and *A. mellifera* (Zemach and Zilberman 2010).

The correlation between YELLOW and RED genes in *Drosophila* with high- and low-CpG genes in *N. vectensis* or *C. intestinalis* and *A. mellifera* is not perfect. It should be kept in mind that we compared data derived from a relatively homogenous *Drosophila* cell line (Kc-167), with whole-organism DNA methylation profiles in the other four species. Our results are also likely to be influenced by incomplete genome annotation for *N. vectensis*, *C. intestinalis*, *A. mellifera*, and *B. mori*. It is also possible that targeting of DNA methylation to gene-bodies is not solely reliant on H3K36me3. Future experiments will need to independently manipulate H3K36me3 and DNA methylation levels in homogenous tissues from *N. vectensis*, *C. intestinalis*, or *A. mellifera*, coupled with whole-genome profiles of DNA methylation, transcription, and H3K36me3 (chromatin immunoprecipitation protocols for these four species have yet to be reported). It will also be important to investigate functional relationships between H3K36me3 and CpG methylation in mammals and plants and the impact of other histone modifications on intragenic DNA methylation, as H3K36me3 is not the only modification present in gene-bodies.

It is interesting to consider why genes associated with basic biological processes should be preferentially methylated, given the mutagenic effects of methylation in the germline. The functional role of gene-body DNA methylation is still a matter of debate, although it has previously been suggested that it serves to suppress spurious transcription of genes that are broadly expressed across tissues (Suzuki et al. 2007; Zilberman et al. 2007). Therefore, the genome must strike a balance between maintaining gene-body methylation (whatever its true role), and suppressing mutagenic effects on key CpG sites in the germline.

In summary, we show that gene-body H3K36me3 bimodality in *Drosophila* is predictive of gene-body CpG density and methylation in at least four different invertebrate species that harbor CpG methylation. We believe our data are most consistent with a model in which gene-body H3K36me3, over evolutionary time, has

influenced gene-body DNA methylation levels and, consequently, the gene-body CpG density bimodality characteristic of invertebrates that harbor a functional CpG methylation system. In other words, histone modifications could, indirectly, affect sequence change over evolutionary time.

## Methods

### Genome annotation and analysis

Genome sequences, gene and protein annotations were obtained from: *N. vectensis* (<http://genome.jgi-psf.org/Nemve1/Nemve1.home.html>), *C. intestinalis* (<http://genome.jgi-psf.org/Cioin2>), *A. mellifera* (<http://hymenopteragenome.org/beebase>), *B. mori* (<http://silkworm.genomics.org.cn>), *O. sativa* (<http://rice.plantbiology.msu.edu/index.shtml>), and *A. thaliana* (<http://www.arabidopsis.org>). For *H. sapiens* and *T. nigroviridis*, we used Ensembl annotation (<http://www.ensembl.org>), and the genome sequences were obtained via the UCSC Genome Browser (<http://genome.ucsc.edu/index.html>). In all cases, the latest genome build was used except for *A. mellifera*, for which we used the assembly 2 genome as the official annotation gene set is mapped to this build. CpG<sub>o/e</sub> ratios were calculated using the formula described in Elango et al. (2009):  $\text{CpG}_{o/e} = P_{\text{CpG}} / (P_C \times P_G)$ , where  $P_{\text{CpG}}$ ,  $P_C$ , and  $P_G$  are the frequencies of CpG dinucleotides, C nucleotides, G nucleotides, respectively, estimated from each gene.

### BS-seq data

Human BS-seq data were from Lister et al. 2009, *Arabidopsis* BS-seq was from Cokus et al. 2008, and all other BS-seq data were from Zemach et al. 2010. For *A. mellifera*, we mapped Fastq files deposited in SRA (<http://www.ncbi.nlm.nih.gov/sra>, accession nos. SRX020137 and SRX020138) to the *A. mellifera* assembly 2 genome. This allowed the inclusion of 5000 genes that map to unassembled chromosomes but were not included in the processed data by Zemach et al. 2010. We calculated a single methylation value for each gene as an average of the methylation ratio of each CpG site within a given gene (C/C + T).

### *A. mellifera* sperm samples

*A. mellifera* drones were obtained in Sussex, United Kingdom. The sperm was collected from six mature drones as they were preparing for mating flights. Drone heads were crushed, provoking the eversion of the genitalia, and sperm was collected with a Pasteur pipette. Approximately 1  $\mu\text{L}$  of semen was obtained per drone and then immediately frozen prior to DNA extraction. DNA was extracted using a standard phenol/chloroform method.

### Reduced representation bisulfite sequencing

We performed two different digestions in parallel on 1  $\mu\text{g}$  of genomic DNA using MspI/BsoBI for the first and TaqαI for the second. Library construction for sequencing on an Illumina GAIIx analyzer was performed according to the protocol reported in Reference 7. MspI/BsoBI and TaqαI libraries were admixed at 1:3 ratio prior to sequencing. Raw reads were aligned using the MAQ methylation mode. Only CpG sites with at least 5 $\times$  coverage were retained for further analyses.

### RNA-seq analyses

Processed RNA-seq data for *A. mellifera*, *C. intestinalis*, and *N. vectensis* were obtained from Zemach et al. (2010) (GEO accession no. GSE19824). They assigned a score equal to the number of mapped

reads per 1000 bp of sequence to each cDNA model (RPKM). We transformed the RPKM values using log10 for plotting clarity purposes.

### Gene ontology analysis

We generated four lists of orthologs between *D. melanogaster* and each of *N. vectensis*, *C. intestinalis*, *A. mellifera*, and *B. mori* by running BLAST locally (version 2.2.22+) for the proteomes of the first against the last four. We consider a pair of proteins as the best bidirectional hit when the two proteins are the best hit of each other using a BLAST score threshold of  $1.0 \times 10^{-5}$ . Gene ontology (GO) annotations were downloaded from the websites of the sequencing consortiums as listed above, except in the case of *A. mellifera*, for which we used *D. melanogaster* annotation obtained from FlyBase (version FB2010\_09). We transferred the GO annotation from *D. melanogaster* to *A. mellifera* using bidirectional orthologs as described above.

### Comparative analyses of *Drosophila* YELLOW and RED chromatin-associated genes

*Drosophila* YELLOW and RED chromatin gene annotation was provided by Drs. Guillaume Filion and Bas van Steensel. We performed a “two-way” hit blast of genes in *N. vectensis*, *C. intestinalis*, and *A. mellifera* against the *Drosophila* orthologs and reciprocally blasted the *Drosophila* proteins against each of the other three invertebrate species. The two-way hit genes that scored significant *P*-values were kept and assigned to the chromatin category of their corresponding *Drosophila*.

### Data access

The data discussed in this manuscript have been deposited in the NCBI Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/>) under Series accession no. GSE29982.

### Acknowledgments

We thank Drs. Guillaume Filion and Bas van Steensel for providing the list of *Drosophila* YELLOW/RED genes, Drs. Elia Stupka and Eamonn Maher for helpful discussions during the early versions of this work, and Dr. Michelle Holland for helping to generate the figures. L.N., G.C., and V.K.R. are supported by the Barts and The London Charity; G.A.H. is funded by a Foulkes Foundation Fellowship; and T.A.D. is a Wellcome Trust Research Career Development Fellow (054523).

### References

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

- Dhayalan A, Rajavelu A, Rathert P, Tamas R, Jurkowska RZ, Ragozin S, Jeltsch A. 2010. The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *J Biol Chem* **285**: 26114–26120.
- 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* **106**: 11206–11211.
- Falcon S, Gentleman R. 2007. Using GStats to test gene lists for GO term association. *Bioinformatics* **23**: 257–258.
- 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* **107**: 8689–8694.
- Filion GJ, van Bommel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, et al. 2010. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* **143**: 212–224.
- Hunt BG, Brisson JA, Yi SV, Goodisman MA. 2010. Functional conservation of DNA methylation in the pea aphid and the honeybee. *Genome Biol Evol* **2**: 719–728.
- Illingworth R, Kerr A, Desousa D, Jørgensen H, Ellis P, Stalker J, Jackson D, Clee C, Plumb R, Rogers J, et al. 2008. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol* **6**: e22. doi: 10.1371/journal.pbio.0060022.
- Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahlinger J. 2009. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet* **41**: 376–381.
- Li Y, Zhu J, Tian G, Li N, Li Q, Ye M, Zheng H, Yu J, Wu H, Sun J, et al. 2010. The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol* **8**: e1000533. doi: 10.1371/journal.pbio.1000533.
- Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**: 315–322.
- Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**: 766–770.
- Rakyan VK, Down TA, Thorne NP, Flicek P, Kulesha E, Gräf S, Tomazou EM, Bäckdahl L, Johnson N, Herberth M, et al. 2008. An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res* **18**: 1518–1529.
- 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* **103**: 1412–1417.
- Suzuki MM, Kerr AR, De Sousa D, Bird A. 2007. CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res* **17**: 625–631.
- Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, Schübeler D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* **39**: 457–466.
- Xiang H, Zhu J, Chen Q, Dai F, Li X, Li M, Zhang H, Zhang G, Li D, Dong Y, et al. 2010. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nat Biotechnol* **28**: 516–520.
- Zemach A, Zilberman D. 2010. Evolution of eukaryotic DNA methylation and the pursuit of safer sex. *Curr Biol* **20**: R780–R785.
- Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**: 916–919.
- Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2007. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* **39**: 61–69.

Received January 28, 2011; accepted in revised form July 11, 2011.



## Comparative methylomics reveals gene-body H3K36me3 in *Drosophila* predicts DNA methylation and CpG landscapes in other invertebrates

Lisa Nanty, Guillermo Carbajosa, Graham A. Heap, et al.

*Genome Res.* 2011 21: 1841-1850 originally published online September 22, 2011  
Access the most recent version at doi:[10.1101/gr.121640.111](https://doi.org/10.1101/gr.121640.111)

---

**Supplemental Material** <http://genome.cshlp.org/content/suppl/2011/07/15/gr.121640.111.DC1>

**References** This article cites 20 articles, 7 of which can be accessed free at:  
<http://genome.cshlp.org/content/21/11/1841.full.html#ref-list-1>

### License

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

---

---

To subscribe to *Genome Research* go to:  
<http://genome.cshlp.org/subscriptions>

---