Genome Biology and Evolution Advance Access published July 5, 2013 doi:10.1093/gbe/evt103

Genome Biology and Evolution

Cross-species analysis of genic GC_3 content and DNA methylation patterns

Tatiana Tatarinova[1,4,#,], Eran Elhaik[2,#], Matteo Pellegrini[3]*

1. *Laboratory of Applied Pharmacokinetics and Bioinformatics, University of Southern California, Los Angeles, CA, USA*
2. *Department of Mental Health, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA*
3. *Molecular, Cell, and Developmental Biology, University of California, Los Angeles, Los Angeles, CA, USA*
4. *Genomics and Computational Biology Research Group, University of South Wales, Pontypridd, United Kingdom*

[#]joint first authors

**Author for Correspondence: Tatiana Tatarinova, tatiana.tatarinova@lapk.org*

Abstract

Background. The GC-content in the third codon position (GC_3) exhibits a unimodal distribution in many plant and animal genomes. Interestingly, grasses and homeotherm vertebrates exhibit a unique bimodal distribution. High GC_3 was previously found to be associated with variable expression, higher frequency of upstream TATA boxes, and an increase of GC_3 from 5' to 3'. Moreover, GC_3-rich genes are predominant in certain gene classes and are enriched in CpG dinucleotides that are potential targets for methylation. Based on the GC_3 bimodal distribution we hypothesize that GC_3 has a regulatory role involving methylation and gene expression. To test that hypothesis, we selected diverse taxa (rice, thale cress, bee, and human) that varied in the modality of their GC_3 distribution and tested the association between GC_3, DNA methylation and gene expression.

Results. We examine the relationship between cytosine methylation levels and GC_3, gene expression, genome signature, gene length, and other gene compositional features. We find a strong negative correlation (Pearson's correlation coefficient r=-0.67, p-value <0.0001) between GC_3 and genic CpG methylation. The comparison between 5'-3' gradients of CG_3-skew and genic methylation for the taxa in the study suggests interplay between gene-body methylation and transcription-coupled cytosine deamination effect.

Conclusions. Compositional features are correlated with methylation levels of genes in rice, thale cress, human, bee and fruit fly (which acts as an unmethylated control). These patterns allow us to generate evolutionary hypotheses about the relationship between GC_3 and methylation and how these affect expression patterns. Specifically, we propose that the opposite effects of methylation and compositional gradients along coding regions of GC_3-poor and GC_3-rich genes are the products of several competing processes.

Keywords

DNA methylation, gene expression, GC_3, grasses, homeotherms, *Oryza sativa, Apis mellifera, Homo sapiens, Arabidopsis thaliana*

Introduction

The term "*epigenetics*" was coined in 1957 by Conrad Hal Waddington (Slack 2002). It is defined as the study of changes in gene expression due to mechanisms other than alterations to the DNA sequence; that is expression modifications are not *hard coded* into the nucleotide sequence. Consequently, epigenetics explains phenomena, which do not result from standard genetic mutations, like hereditary changes in gene expression under the influence of environmental factors. DNA methylation is one of the most studied epigenetic mechanisms modulating gene expression and has important health implications. For example, the gain or loss of DNA methylation can produce loss of genomic imprinting and results in diseases such as Beckwith-Wiedemann, Prader-Willi and Angelman syndromes (Adams 2008). Changes in the patterns of DNA methylation are also commonly seen in human tumors. Both genome wide hypomethylation (insufficient methylation) and region-specific hypermethylation (excessive methylation) have been thought to play a role in carcinogenesis (Lengauer 2007). DNA hypomethylation contributes to cancer development through an increase in genomic instability, reactivation of transposable elements, and loss of imprinting (Esteller 2002). Hypermethylation-induced silencing of primary transcripts through their CpG island promoters is a common cause of the loss of tumor suppressor miRNAs in cancer (Lengauer 2007; Lopez-Serra & Esteller 2012; Sonkin et al 2013).

Methylation occurs by the addition of a chemical methyl group (-CH3) through a covalent bond to the cytosine bases of the DNA backbone and tends to be more abundant at Cytosine-phosphate-Guanine-(CpG) dinucleotides (Sadikovic 2008). However, methylation can also happen in CHG and CHH contexts (where H indicates any nucleotide other than G). DNA methylation is common in humans and other mammals, where 70 to 80% of CpG dinucleotides are methylated. Interestingly, in some model organisms, such as yeast and fruit fly, there is little or no DNA methylation. Also, DNA methylation in mammals differs from that in plants as it targets CpG sites. In humans and mice, CpG dinucleotides account for roughly three quarters of the total DNA methylation content in their cells (Ziller et al. 2011).

In vertebrates, the methylation process is being catalyzed by members of the enzyme family of DNA methyltransferases (DNMTs), which recognize palindromic sequences with CpG dinucleotides. Thus far, three active DNMTs have been identified in mammals: DNMT1, DNMT3A, and DNMT3B. A fourth similar enzyme (DNMT2 or TRDMT1) is structurally similar to the other DMNTs. However it does not methylate DNA but rather transfers RNA (Goll et al. 2006). DNA methylation of CpG dinucleotides is essential for plant and mammalian development. Methylation mediates the expression of genes and plays a key role in chromosome X inactivation, genomic imprinting, embryonic development, chromosome stability, chromatin structure, and may also be involved in the immobilization of transposons and the control of tissue-specific gene expression (Li et al. 2008).

The relationship between gene expression, nucleotide composition and gene length were the subject of several studies in the past decades. Oliver and Marin (1996) associated the expected length of a reading frame to the CG composition using the property that stop codons (TAG, TAA, and TGA) are biased towards low GC content. They suggested that the longest coding sequences/exons in vertebrates are

GC-rich, while the shortest ones are GC-poor. Subsequently, Xia et al. (2003) described positive correlations between GC content and coding regions (CDS) lengths in 68 genomes. It was later shown that highly expressed rice and human GC-rich genes have significantly more and longer introns than lowly expressed genes, whereas their average exon length per gene is significantly lower. By contrast, GC-poor genes were shown to exhibit similar compactness between highly and lowly expressed genes (Mukhopadhyay and Ghosh 2010).

The relationship between gene-body methylation and gene expression was studied in a number of organisms, and a positive linear correlation was reported (Zemach et al, 2010; Xiang et al 2010). Anastasiadou et al. (2011) reported the relationship between splicing and methylation in the human genome as well as a positive relationship between alternative splicing and methylation. Recently, Flores et al (2012) reported a positive relationship between exon-level DNA methylation and mRNA expression in the honeybee. They also found that methylated genes are enriched for alternative splicing; therefore suggesting that gene-body DNA methylation positively influences exon inclusion during transcription. The authors proposed that DNA methylation and alternative splicing contribute to a longer gene length and a slower rate of gene evolution. However, none of these studies, considered the potential regulatory role of GC_3.

Several studies focused on coding regions that are enriched in methylation targets (CpG-rich). For example, Nanty et al. (2011) found an evolutionarily conserved feature in invertebrate genomes separating CpG-poor and CpG-rich genes: CpG-poor genes were associated with basic biological processes, while the latter with more specialized functions. Gavery and Roberts (2010) found that hypo- and hyper- methylated genes differ in both biological function and in the ratio between observed and expected CpG dinucleotides. Coding regions enriched in CpG dinucleotides also exhibit a higher frequency of G or C in the third codon position (GC_3). Because mutations in this position lead primarily to synonymous substitutions, the selective pressures affecting its composition are different from those acting on the first two codon positions, making it a valuable tool to study evolution. To name a few, it has been previously shown (Tatarinova et al. 2010; Sablok et al. 2011; Ahmad et al. 2013) that dicot and monocot plant genes with high GC_3 have distinctly different properties from genes with low GC_3: they contain more targets for methylated GC_3-rich genes, and also exhibit more variable expression, possess more upstream TATA boxes, are enriched for certain classes of genes (e.g. stress responsive genes), and have a GC_3 content that increases from 5' to 3' (Tatarinova et al. 2010). GC_3-rich genes were also shown to be inducible while the GC_3-poor are ubiquitously active (Tatarinova et al. 2010). Thus we speculate that GC_3 has evolved to be interdependent with gene-body methylation and gene expression so that genes that are GC_3-rich or -poor have different expression patterns.

Here, we tested the hypothesis of the regulatory role of GC_3 by studying the relationship between GC_3, gene-body methylation, and related genomic features in four taxa: rice, arabidopsis, bee, and human. These particular species were chosen because they have well-annotated genomes, rich collections of gene expression measurements, and genome-wide methylation measurements. Comparison with the fruit fly allows us to separate methylation-related effects from other factors. We show that GC_3 is

inversely correlated with gene methylation in these four organisms and propose an evolutionary theory to explain these patterns.

Materials and Methods

Gene models were taken from: MSU (version 6.1) for *Oryza sativa*; TAIR version 7 for *Arabidopsis thaliana*; BeeBase (www.beebase.org) annotation for *Apis mellifera*; NCBI GenBank for *Homo sapiens* (hg18) ; and dmel_hetr31 from FlyBase (www.flybase.org) as well as Release 5 from Berkeley Drosophila Genome Project (www.fruitfy.org) for *Drosophila melanogaster.*

Gene expression data were obtained from the NCBI GEO collection (GSE9415, GSE24177, GSE5624, GSE1647, GSE19700, GSE9646-GPL10978, GSE9646-GP10977, GSE16474, GSE34029, GSE34293, GSM846863, GSE25161, GSE34029, GSE34293, GSE42255, GSE5147, GSE1643, GSE7567, GSE16144, GSE21009-GPL10237).

Filtering. We selected gene sets where gene expression, methylation and high-quality annotation data were available: there were 12,577 such genes in *Arabidopsis thaliana*, 14,069 in *Homo sapiens*, 9,607 genes in *Oryza sativa,* and 15,381 genes in *Apis mellifera*. For *Drosophila melanogaster* we used 18,731 coding sequences.

Methylation bisulfite sequencing measurements for the four organisms were obtained from previously published studies (Chodavarapu et al. 2010; Foret et al. 2012; Bernal et al. 2012; Chodavarapu et al. 2012). We required a minimum of five reads to call the methylation state of a cytosine. The DNA methylation level was estimated from the fraction of cytosines that failed to undergo bisulfite conversion. Therefore, for each cytosine, the methylation level ranged from 0 to 1. When we computed average gene-body methylation for a given context, we calculated the average methylation for all coding regions, using appropriate gene models for each organism. For *H. sapiens* we used H1 embryonic stem cell line methylation profile. The distributions of gene-body methylation levels are shown in Figure 1B.

GC_3. For every open reading frame GC_3 was computed as $GC_3 = \frac{C_3 + G_3}{\left(L/3\right)}$, where C_3 and G_3 are counts of cytosines and guanines in the third position of the codon and *L* is the length of the coding sequence.

GC_3 distributions are obtained from a histogram of GC_3 values, where GC_3 values were rounded to hundredths (Figures 2 and 4) or tenth (Figures 5 and 6). We require that all points on the graph were supported by at least 100 observations, criteria which determined the choice of the bin size.

Standardization of gene expression (*Z*-statistic). For a gene g, $Z_g(Expression) = \frac{\overline{(E_g)} - e}{\sigma}$, where $\bar{E}_g$ is the average expression of the gene g across N experiments, e is average expression of all genes, and σ is the standard deviation of gene expression. All expression levels were log-transformed. The genes were

divided into three groups based on their expression level, namely: $Z_g(Expression) > 1$, $-1 < Z_g(Expression) \leq 1$, and $Z_g(Expression) \leq -1$.

The genome signature (ρ_{CG}) is defined as the relative abundance of the frequency of di-nucleotides in the genome, so that $\rho_{CG} = \frac{f_{CG}}{f_C f_G}$, where f_x is the frequency of a (di) nucleotide. Genomes or genes can thus be compared with respect to their relative abundance of methylation targets and GC_3 richness.

CG_3-skew. Following (Tatarinova et al. 2003), CG_3-skew was defined as $CG_{3skew} = \frac{C_3 - G_3}{C_3 + G_3}$. We calculated the 5'-3' CG_3-skew gradient patterns in arabidopsis, rice, bee, fruit fly and human by counting the number of Cs and Gs in the third position of codons in the first 200 codons of GC_3-rich and GC_3-poor genes.

Expression measures. We use mean expression value across all collected experiment for every gene, standard deviation of gene expression values across all conditions and coefficient of variation (CV), defined as a ratio or standard deviation and mean gene expression.

Distinguishing GC_3-rich from GC_3-poor genes. Since GC_3 varies between organisms, such definitions are organism-specific and depend on the shape of its distribution which can be either unimodal or bimodal (Figure 1A). In the case of unimodal bell-shaped distribution, common to many plant and animal species, the extreme 5% of the genes from the tails of the distributions are denoted as "GC_3-rich" and "GC_3-poor" genes (Sablok et al. 2011; Ahmad et al., 2013). By contrast, for bimodal distributions that are common to grasses and homeotherm vertebrates (Elhaik et al. 2009; Elhaik & Tatarinova, 2012), the GC_3 cutoff is determined based on the position of the "valley" between the two peaks.

Gene Ontology (GO) annotation. GO annotations were obtained from www.geneontology.org, TAIR (www.arabidopsis.org), and Michigan State University (ftp.plantbiology.msu.edu). Upon division of genes into GC_3-rich and –poor classes, we computed $\chi^2 = \frac{(O-E)^2}{E}$ statistic for each GO category (Tables S4, S5).

Results

GC_3, body methylation and gene expression

Of the guanine and cytosine (GC) content at each codon position (GC_1,GC_2,GC_3), the last measure represents the fraction of GC content in the codon's wobble position that has the most freedom to change without altering amino acid sequence of the gene. GC_3 exhibits the strongest Pearson's correlation with gene-body methylation (r_{GC1}=-0.47, r_{GC2}=-0.35, r_{GC3}=-0.67) and variability of gene expression (r_{GC1}=0.1, r_{GC2}=0.14, r_{GC3}=0.21) and is correlated with the gene's GC content (e.g. in rice correlation between genic GC and GC_3 is 0.94).

Due to the different shapes of the GC_3 distributions in the studied taxa (Figure 1A) we hypothesized that the GC_3 content has a regulatory role and should be correlated with both CpG methylation and gene expression which, in turn, should also be correlated with one another. To test our hypothesis, we carried detailed analyses of the relationship between GC_3 composition, gene body methylation, and gene expression in rice, arabidopsis, honey bee, and human. As expected, in all four species, GC_3 and genic CpG methylation were negatively correlated and CpG methylation had a consistently negative effect on the variability of gene expression (Table 1). The relationship between GC_3 and average gene expression is nonlinear and saddle-like for all four organisms (Figure 2), but the strength of the dependencies varies from organism to organism.

We compared full and partial correlation coefficients, calculated as in (Kim and Yi, 2007), between GC_3, gene expression variability, and gene body methylation (Table 1). We found that the relationship between gene-body methylation and GC_3 is approximately the same, when controlling for variability of gene expression as compared to the full correlation coefficient. Partial correlations between gene expression variability and methylation and between GC_3 and gene expression variability are much smaller than the full correlation coefficients. These results suggest that the relationship between GC_3 and gene body methylation is the driving force and confounds the two other correlations.

In the following sections, we describe the relationship between GC_3, gene-body methylation and gene expression for each of the four organisms we investigated.

Oryza sativa

In rice, distributions of GC_3 and gene-body methylation are both clearly bimodal (Figure 1). Genes can be divided into GC_3-rich and -poor classes using the position of the valley between the two peaks (at $GC_3 \approx 0.8$) and, similarly, into highly methylated and lowly methylated classes (gene-body methylation ≈0.0178). We have previously shown (Tatarinova et al. 2010) that GC_3-rich genes in rice have more methylation targets (ρ_{CG}) that can be used to modulate tissue-specific expression: $\rho_{CG}(\text{poor}) = 0.55$ and $\rho_{CG}(\text{rich}) = 1.15$.

To estimate the regulatory effects of GC_3 we first calculated its correlation with different genic measures including intron density, the number of introns per 1000 bases and intron fraction, defined as the ratio of intron length to gene length, for GC_3-poor and -rich genes that are highly and lowly expressed (Table 2). Compared to lowly expressed genes, highly expressed genes have an intron density approximately twice as high; with both the average number of exons and average intron length being 1.5 times higher. Remarkably, genic measures for highly and lowly expressed genes varied markedly when compared between GC_3-poor and -rich genes (Table 2). For instance, GC_3-poor genes with high (E>1) and low (E<-1) expression values differ in their intron density (6.296 and 3.090, respectively) and number of exons (9.60 and 5.41, respectively). We also found that GC_3 is negatively associated with intron density (r=—0.36 p-value <0.0001) and with intron fraction (r=—0.40, p-value <0.0001).

We found a significant association between methylation and GC_3-richness (Table 3) in agreement with previous studies that described a positive correlation between GC_3 content and the variability of gene expression in grasses (Tatarinova et al. 2010). Studying the triangular relationship between methylation, gene expression, and GC_3 (Figures 2 and 3) we observe that GC_3-rich genes tend to have more variable gene expression and lesser gene-body methylation levels than the GC_3-poor genes. Moreover, methylation of CpG in coding regions has a nonlinear relationship with gene expression. Both the most lowly and highly expressed genes have low levels of methylation while medium-expressed genes are more methylated, in agreement with the trends reported by Jjingo et al. (2012). These observations suggest the interplay of two or more forces that affect gene expression. GC_3 exhibits a trend from high GC_3 and low methylation to low GC_3 and high methylation. Highly methylated genes, associated with development, genomic imprinting, or silencing of trans-genes, exhibit low expression levels. These results are consistent with the notion that methylated genes can undergo 5-methylcytosine deamination where mC→T. In such cases, the third position can often undergo cytosine deamination reducing GC_3 without affecting the protein sequence, whereas the first two nucleotides in the codon are less likely to mutate due to selective pressures to conserve amino acid sequences. Hence, methylated genes are expected to be GC_3-poor. Consequently, low-methylated genes have high GC_3 values and low average expression (Figure 3), where an increase of CpG methylation and high deamination rate lead to a drop in GC_3 values; at the same time the average expression reaches the maximum for the broadly expressed genes. A further increase in methylation does not affect GC_3, but rather reduces gene expression, leading to a repression of the gene (see Supplementary Materials and Figure S5 for further details).

To examine the effect of alternative splicing on the correlation between methylation and GC_3, we next considered the relationship between GC_3 and gene-body methylation for intron-containing and intron-less genes. Lyko et al. (2010) discovered that clusters of methylated cytosines are associated with alternatively-spliced exons and that intron containing genes are more methylated than intron-less genes. Intron-less genes are, obviously, not subject to alternative splicing while genes with introns may be alternatively spliced. There are 2,648 intron-less genes in the dataset; for these the average values of GC_1=0.63, GC_2=0.51, and GC_3=0.77, compared to 6,959 intron-containing rice genes with average GC_1=0.58, GC_2=0.47, and GC_3=0.61. Indeed, intron-containing genes are twice more methylated than intron-less genes (0.18 vs. 0.09). As expected, intron-containing genes also exhibit a stronger positive relationship between the average methylation and expression, between the CV of gene expression and GC_3, and stronger negative correlation between the CV of gene expression and methylation (Table 4). Interestingly, we observed only a small difference in the correlations between the average methylation and GC_3 between intron-less (r=-0.6) and intron-containing (r=-0.67) genes. Therefore, splicing influences the relationship between methylation, expression and nucleotide composition.

Traditional microarray measurements, which ignore alternative splicing, are not able to fully measure variability of gene expression. This may partially explain why when comparing intron-containing with intron-less genes, the first have higher average expression (1.41 vs. 1.13, respectively) and lower CV of gene expression (0.92 vs. 1.28, respectively). We hypothesize that apparent constitutive expression of hyper-methylated, intron-containing genes can be a complex phenomenon, with different splicing forms

expressed at different developmental stages, tissue types and external conditions. We hypothesize that gene expression variability of hypo-methylated, intron-less genes is achieved by transcriptional regulation. Overall, alternative splicing evens may explain the differences in methylation and expression levels between intron-less and intron-containing genes, but not the differences between GC_3 and gene-body methylation.

A more general explanation of the relationship between gene expression and methylation involving the nucleosome was recently proposed by (Jjingo et al. 2012). The authors pointed out that CpG sites occur frequently across gene bodies and that in genes with low levels of expression, methylation is prevented by dense nucleosome packing. By contrast, in genes with average levels of expression these sites are accessible to DNMTs and hence are more likely to be methylated. When expression is high, polymerases and DNMTs compete for the access to the same sites and hence methylation is suppressed again.

Arabidopsis thaliana

Arabidopsis has a narrow and unimodal distribution of GC_3 and a bimodal distribution of methylation levels (Figure 1). Despite of the apparent unimodality of the GC_3 distribution, Arabidopsis genes with $GC_3>0.5$ are significantly less methylated than genes with $GC_3\leq0.5$: $P(\text{methylation}<0.016|GC_3>0.5)=0.72$ and $P(\text{methylation}<0.016|GC_3\leq0.5)=0.33$, suggesting a relationship between GC_3 and methylation. More specifically, the increase of GC_3 composition is negatively correlated with gene body methylation levels (Figure 4A). Of the three methylation contexts, the most pronounced effect is observed for CpG methylation (r=-0.27, p-value <0.0001) (Figure 4), while CHG and CHH methylation levels appear to be less affected by GC_3 composition.

In thale cress, the average relative abundance of the frequency of CG di-nucleotide (genome signature, ρ_{CG}) for all genes is 0.73. The relative abundance of methylation targets depends on GC_3 richness, for genes with $GC_3>0.5$ (mean $\rho_{CG}=0.91$) and the remaining genes (average $\rho_{CG}=0.71$). There is also a relationship between methylation levels and ρ_{CG}: $\rho_{CG}(\text{methylation}<0.016)=0.84$ while $\rho_{CG}(\text{methylation}\geq0.016)=0.67$. Hence, GC_3-rich genes have more methylation targets but are less methylated. Therefore, in spite of the unimodality of the GC_3 distribution in *A. thaliana*, the relationship between methylation and GC_3 is similar to the pattern observed for rice. That is, like the other taxa, arabidopsis exhibits a nonlinear, saddle-like dependence, between the strength of gene expression and GC_3 (Figure 2A), but its gene expression variability grows almost linearly with GC_3 (Figure 2B).

To further study the relationship between tissue-specific gene body methylation and tissue-specific expression, we next examined tissue specific patterns across shoots and roots as these exhibit differences in morphology, gene expression activity and function. We investigated 1000 genes from the two tails of the log(shoots/roots) expression distribution (see *Materials and Methods*) and compared the differences between shoot and root body methylation levels for the two gene groups. We found that the average genic methylation was similar for shoots and roots (0.063 in shoots vs. 0.057 in roots). However, for genes over-expressed in shoots, there was a negligible difference between shoot and root

methylation, whereas for genes over-expressed in roots, on average, the "shoot" genes were 21% more methylated than the "root" genes (p-value =0.003). These results are again in agreement with Jjingo et al. (2012) and highlight the role of methylation in contributing to tissue-specific expression. Interestingly, differences between methylation levels in shoots and roots increase with GC_3 for all methylation types (Figure 4B).

In summary, GC_3 is positively correlated with both expression variability and variation in genic methylation. There is also an inverse relationship between gene-body tissue-specific methylation and tissue specific gene expression.

Apis mellifera

The GC_3 distribution of the European honey bee, *Apis mellifera*, is a unimodal right skewed distribution with a long tail of high GC_3 values (Figure 1). The honey bee is a GC_3–poor organism, but it has a surprising medium and high GC_3 tail, containing approximately 25% of its genes with GC_3>0.5. Based on the current annotation, 2.2% of all *Apis mellifera* genes encode receptors (such as *Metabotropic glutamate receptor, Toll-like receptor, Dopamine receptor type D2, D2-like dopamine receptor, Ephrin receptor, SIFamide receptor, Ecdysteroid receptor A isoform, Antennapedia protein, Nicotinic acetylcholine receptor alpha1 subunit, Alpha-glycosidase G-protein coupled receptor*, and others). Genes with GC_3>0.505 are significantly enriched in receptor encoding genes, which account for 5.6% of these compared to 1.3% in genes with GC_3<0.12 (p-value=2.6E-7). The frequency of CG dinucleotides differ between GC_3-rich genes (average $\rho_{CG} = 1.18$) and GC_3-poor genes (average $\rho_{CG} = 0.41$) To further study the relationships between GC_3-richness, receptor genes and methylation, we compared data from Queen and Worker bees.

Queen and Worker bees share the same genome but differ in size, appearance and life span. While there is little difference between the whole-genome methylation level of the Worker and Queen bees (around 1% of cytosines in CpG contexts are methylated in both), their genes differ significantly in methylation levels (Figures 5A-B). This finding in is in agreement with a previous report that Worker and Queen bees differ in the methylation of ~550 genes (Lyko et al. 2010). Lyko et al. (2010) also reported that unmethylated genes are enriched in receptors. The methylated genes encode proteins showing a higher degree of conservation than proteins encoded by non-methylated genes (Foret et al 2009). Of the three methylation contexts, we observed that the average fraction of CG methylation per gene was associated with GC_3 composition (Figures 5C-D) in support of a putative GC_3 regulatory role. In other words, increases in GC_3 in bees are associated with a decrease in gene-body methylation levels, which are enriched for receptor encoding genes. Follow up analyses of bee methylation patterns can be found elsewhere (Lyko et al. 2010; Foret et al. 2012).

In addition to these relationships, we also found that differences between methylation levels in Worker and Queen bees depend on the nucleotide composition of coding regions. We analyzed the relationship between gene body-methylation and GC_3 for Queen and Worker bees. The relative difference between

gene-body methylation in queen and worker bees, defined as $\frac{(Q-W)}{Q+W}$, depends on the methylation context (CpG, CHH or CHG) (Figure 5A). The relative difference in CpG and in total methylation is low for the GC_3-poor genes and increases substantially when GC_3 approaches 0.4 after which methylation stays roughly the same for genes with GC_3>0.4 (Figure 5A). Relative difference between CHH and CHG methylation decreases with the increase of GC_3. The transition between the compositional environments may be related to changes in the regulatory role of each region. The difference between CpG methylation levels between Queen and Worker is negative for low GC_3 genes (Queen bee is less methylated) and becomes positive with an increase of GC_3 (Figure 5B). Overall, GC_3 poor genes are more methylated than GC_3 rich genes (Figures 5C and 5D). As compared to the worker bees, queen bees have lower body methylation levels for GC_3-poor class (enriched for ubiquitously expressed genes) and higher for GC_3-rich class (enriched for receptor-encoding genes) (Figure 5B). Since queen and worker bees play drastically different roles in the beehive, they activate and rely onto different sets of genes (Aamodt, RM 2009). Higher social role of the queen bee may require more elaborate interaction with environment, which necessitates more regulation of the GC_3-rich receptor-encoding genes through methylation. Our observations agree with Foret et al (2009) and Elango el al (2009), who pointed out that ubiquitously expressed critical genes are methylated at the germ-line, while cast-specific genes lack methylation. Caste-specific genes remain unmethylated to allow for grater epigenetic flexibility and regulatory control (Elango et al., 2009). Greater degree of flexibility is important for certain classes of genes in other invertebrates: according to Gavery and Roberts (2010) and Roberts and Gavery (2012), the ubiquitously expressed housekeeping genes tend to be hyper-methylated while tissue-specific and inducible genes are hypomethylated.

Homo sapiens

Coding regions of *Homo sapiens* have a broad bimodal distribution of GC_3 values (Figure 1A) and a unimodal distribution of genic methylation levels, with a long tail towards low methylation levels (Figure 1B) (Chodavarapu et al. 2010). As in the other three species, the relative abundance of CpG dinucleotides differs for GC_3-rich and poor genes: $\rho_{CG}(rich) = 0.68$ and $\rho_{CG}(poor) = 0.29$. Overall, the *H. sapiens* genome is more methylated than bee, rice, and arabidopsis (Figure 1B). Although the nonlinear dependence between GC_3 and gene expression is apparent (Table 1, Figure 2A and Figure 2B), its shape differs compared to the other three species we analyzed. In human, CpG methylation is negatively correlated with GC_3 and CHH and has no significant correlation with CHG methylation (Table 1 and

Figure **6**). The weak correlation between GC_3, expression and methylation suggests the existence of other evolutionary forces affecting gene expression in the human genome.

The compositional environment and gene-body methylation paradox

A pronounced pattern that emerged from all our analyses is that GC_3-rich genes are, on average, under methylated, despite their enrichment of CpG dinucleotides. To further illustrate this trend, we compared the GC_3 gradient (Figure S1) and the CG_3-skew (Figure S2) across all tested taxa with gradients of methylation levels using the same groups of GC_3-rich and GC_3-poor genes (Figure S3). The positive 5'-3' gradient of body-methylation, where methylation increases toward the mid-portion of the transcribed part of the gene can be attributed to a gene experiencing "boundary effects" from the attachment of transcriptional and translational machinery. At the 5'-end methylation needs to be low to enable attachment of proteins. Deamination of methylated cytosines in broadly expressed and highly methylated GC_3-poor genes leads to the decrease in C nucleotides and negative CG_3-skew in the middle of the gene (Figure S2). Although GC_3-rich genes are enriched in methylation targets, they are under-methylated compared to GC_3-poor genes. In fact, GC_3-rich genes were so hypomethylated that we had to log-transform the methylation levels to be able to plot the two trends on the same figure. Additional evidence of the different regulatory roles GC_3-poor and GC_3-rich genes assume in methylation can be found by looking at the competing process of cytosine deamination reducing methylation targets.

GC_3-rich and GC_3-poor genes exhibit different body-methylation levels and different gradients of methylation in coding regions (see Supplementary Materials, Figures S1-S3). The variation in compositional gradients may explain the under methylation observed in GC_3-rich genes. Methylation level of GC_3-poor genes experiences steep growth in the first 100 codons (300 nucleotides) and then stays approximately constant (Figure S3). With the exception of *H. sapiens* H1 cell line, methylation levels of GC_3-rich genes are position-independent. As shown by Tatarinova et al. (2010) and by Sablok et al. (2011), towards the middle of the gene, GC_3-rich genes continuously become more C-rich (positive CG_3-skew), whereas GC_3-poor genes become G-rich (negative CG-skew); GC_3-rich genes become even more GC_3-rich towards the middle of the gene, and GC_3-poor genes become more GC_3-poor. We hypothesize that for the broadly expressed and highly methylated GC_3-poor genes, the decrease in C nucleotides may be due to cytosine deamination (mC→T transitions).

To this end, we next looked at genes of *Drosophila melanogaster*, which belongs to the so-called "Dnmt2 only" organisms that do not contain any of the canonical DNA methyltransferases (Dnmt1 and Dnmt3) (Krauss and Reuter 2011). The levels of DNA methylation in the fruit fly are significantly lower than in other organisms (Lyko et al. 2000). In the fruit fly GC_3 content is positively associated with strength of gene expression (Figure S4D). For the 300 genes with GC_3<0.55, average expression across 71 conditions is 2.12 on the log10 scale, vs. average expression of 3.62 for the 300 genes with GC_3>0.8. Surprisingly, variability of fruit fly gene expression does not seem to be affected by GC_3. In addition, average genome signatures , for both GC_3-rich and -poor fly genes are even (ρ_{CG}= 0.9).

We compared the 5' to 3' gradients of CG skew in bee, thale cress, rice and human (where a significant degree of gene body methylation exists) to those in the fruit fly. In the first four taxa, we observed drastically different 5' to 3' gradients of CG skew in both GC_3 rich and GC_3 poor genes (Figure S2), whereas in the fruit fly these trends are absent (Figure S4C). Decreased methylation of 5' regions of was previously described by Roberts and Gavery (2013). In other words, the unmethylated fly genes exhibit similar GC_3 5'-3' gradients (Figure S4B) to those of the other taxa. However, due to the absence of cytosine deamination there are even levels of Cs and Gs for both fly GC_3-rich and -poor genes, whereas in the other taxa cytosine deamination reduces the number of Cs for the highly methylated GC_3-poor genes in a position-specific manner (Figure S2).

DISCUSSION

Gene-body methylation and gene expression exhibit complex relationships with one another and with sequence composition. For example, DNA methylation in coding and non-coding regions have opposite effects on gene expression: in promoters, cytosine methylation often makes transcription factor binding sites inaccessible to transcription factors and is responsible for transcriptional repression in *A. thaliana* (Chan et al. 2005) while gene-body methylation is reported to be positively correlated with gene expression in *H. sapiens* (Hellman and Chess 2007). Generally, these relationships exhibit similarity across diverse taxa, but may vary for particular genes. For example, Aceituno et al. (2008) noted that in *A. thaliana* housekeeping genes that have broad and steady expression levels were more body-methylated than expected based on whole genome methylation levels (p=1.5E-35). Only 8% of the hyper-variable genes, such as stress response or tissue specific genes with high values of gene expression coefficient of variation) were found to be body-methylated. Aceituno et al. (2008) also reported that gene body-methylation is negatively correlated (r=-0.89) with the variability of gene expression on a genome-wide scale, implying that housekeeping genes having low expression variability have higher methylation levels and vice versa. This report follows Bird et al.'s (1995) hypothesis that gene-body methylation could be responsible for the repression of spurious transcription within genes and hence lead to more reliable transcription, which results in a positive correlation between gene expression and gene-body methylation. This relationship was previously described as exhibiting a bell-shaped distribution (Zilberman et al. 2007; Zemach et al. 2010).

To better understand the regulatory role of gene-body methylation and its relationship with sequence composition, we studied the role of GC_3 in four taxa: rice, thale cress, bee, and human. We showed that GC_3 richness and methylation are negatively correlated, which leads to a seeming paradox: if GC_3-rich genes are enriched in methylation targets, why are they under methylated compared to GC_3-poor genes? One reason for this negative correlation may be due to the prevalence of ubiquitously expressed genes in the GC_3-poor class that use body-methylation as one of the mechanisms to maintain broad expression. Association between alternative splicing, gene expression and methylation allows us to hypothesize that the alternatively spliced intron-containing genes and oppositely, the intron-less achieve gene expression variability via different mechanisms. Hypo-methylation of intron-less, high GC_3 genes and abundance of methylation targets allows achieving higher regulatory control. Hyper-

methylated, intron-containing, low GC_3 genes can express different spicing forms and be expressed at different developmental stages, tissue types and external conditions. It is thus not surprising that GC_3-rich, hypo-methylated genes have higher genetic diversity as compared to the GC_3-poor, hyper-methylated genes (Tatarinova et al 2009; Lyko et al., 2010; Roberts and Gavery, 2012).

We propose that the opposite effects of methylation and compositional gradients along CDS of GC_3-poor and GC_3-rich genes (Figure S3), are the products of two or more competing processes. The first driver is transcriptional efficiency. There may be a "universal pressure" to increase the fraction of C-ending codons from the 5' to the 3' end of the gene that can be explained by the need to increase the speed of transcription in this direction. This is especially important for stress-specific genes (that are frequently GC_3-rich) (Tatarinova et al. 2010), since they are expressed as a response to a certain environmental condition, likely at a high level, for a limited amount of time resulting in a large number of RNA polymerases (RNAPs) that move simultaneously along the same track. Hence, it is necessary to avoid RNAP congestion and increase the speed of transcription. There is no such pressure for ubiquitously expressed genes (frequently GC_3-poor), since RNAP congestion effects are not likely to occur.

The competing process may be cytosine deamination, which affects more methylated genes and genes that are expressed at relatively constant levels across tissues. GC_3-rich genes are less methylated and are likely to have limited tissue-specific and stress-specific expression patterns that require less time in the transcriptional bubble. Therefore, the effect of cytosine deamination is less pronounced in GC_3-rich genes. For GC_3-rich genes, transcriptional kinetics is the winning driver.

Takuno and Gaut (2012) hypothesized that "body-methylated genes would be both longer and more functionally important than unmethylated genes." The authors suggested that methylation has a functional role, such as maintaining transcriptional accuracy and splicing efficiency, thus explaining why the GC_3-poor housekeeping genes are overall highly methylated. This agrees with our findings (Table 2) that GC_3 poor genes are longer (e.g., in rice, GC_3-rich genes are on average 1031 nt long and GC_3-poor genes are on average 1648 nt long) and have more exons (e.g., in rice, GC_3-rich genes have on average 2.38 exons and GC_3-poor genes have on average 8.57 exons). Takuno and Gaut (2012) also found that "body-methylated genes evolve more slowly than unmethylated genes, despite the potential for increased mutation rates in methylated CpG dinucleotides." This is also consistent with our observation (Tatarinova et al. 2010) of faster evolution of unmethylated GC_3-rich genes as compared to methylated GC_3-poor genes. Finally, we have shown that methylated genes have a lower proportion of CpG nucleotides, which supports the deamination hypothesis.

Overall, our work supports and expands recent findings by Takuno and Gaut (2012) and Roberts and Gavery (2012). We propose several possible explanations to the question of why GC_3-rich genes are enriched in CpG dinucleotides compared to GC_3-poor genes : first, these sites may have played a regulatory role in the past and are maintained in the genome to allow phenotypic plasticity by increasing the number of transcriptional opportunities (Roberts and Gavery, 2012). Second, these sites may have an active regulatory role that has yet to be determined. Third, we suggest considering the problem from

a different angle –that while GC_3-poor genes have less CpG sites than GC_3-rich genes, they are more body-methylated because as methylation increases in the 5'→3' direction, there is more chance for mC→T mutation towards the middle of the gene. Most of the GC_3-poor genes are ubiquitously expressed; therefore, the sense strand spends more time unprotected during transcription (Tatarinova et al. 2003). The cytosines are therefore lost in the deamination processes and the CG_3-skew value is reduced. Since the third position in the codon is not under pressure to conserve the protein sequence, the mC→T mutations are manifested as gene's GC_3-poorness. In support of this view, the 5' end of genes has a lower level of methylation and positive gradient of CG_3-skew for both GC_3-rich and GC_3-poor genes, which can be explained by transcription/translation initiation requirements.

If methylation is associated with transcription then the ubiquitously active genes should lose GC_3 due to deamination while the inducible ones should not. Looking at the gene body methylation and GC_3 composition as a function of the normalized average gene expression in rice (Figure S5), methylation and GC_3 have opposing trends: where GC_3 increases, methylation decreases and vice versa. Normalized gene expression between -1 and +1 contains many of the ubiquitously expressed genes, and in this region a decrease in GC_3 is accompanied by an increase in methylation. Methylation and GC_3 of inducible genes, having low average exprerssion (below -1 in Figure S5) are not affected by the change in gene expression.

Our observation that the unmethylated fly genes exhibit similar GC_3 5'-3' gradients to those of the other taxa but different patterns of CG_3-skew supports the significance of cytosine deamination. In the fruit fly, due to the absence of cytosine deamination, levels of Cs and Gs for both GC_3-rich and -poor genes are approximately the same, whereas in the other taxa cytosine deamination reduces the number of Cs for the highly methylated GC_3-poor genes.

We note that in addition to the processes described here, there are two major forces affecting GC_3. One is GC-biased gene conversion (BGC) (Duret 2008), which is common to all our model species (Duret and Arndt 2008; Duret and Galtier 2009; Katzman et al. 2011; Kent et al. 2012; Günther et al. 2012; Muyle et al. 2011). The other is selection on codon usage, which has been shown to occur in Arabidopsis (Muyle et al. 2011; Günther et al. 2012). It has been suggested that recombination hotspots can create strong substitution hotspots that are correlated with gene density that drive the evolution of GC content (Duret and Arndt 2008; Tatarinova et al., 2010). Affecting both coding and non-coding regions, BGC may lead to enrichment in GC-content in genomic regions of high recombination compared to regions of low recombination and may explain the patterns observed in human. Coding regions may also be susceptible to codon usage bias that directly affects the frequency of GC_3. The complex interplay between these forces and their relative effect on methylation and gene expression in different species remains unclear and provides a fertile area for future studies.

Conclusions

We report strong negative correlations between CpG methylation and the GC_3 content of genes in rice, bees, Arabidopsis and humans. We propose several explanations for the triangular relationship between GC_3, methylation, and expression patterns. The negative correlation between GC_3 and methylation can be explained by the prevalence of ubiquitously expressed genes in the GC_3-poor class that use body-methylation as one of the mechanisms to maintain broad expression. Positive 5'-3' gradient of body-methylation, where methylation levels rise toward the mid-portion of the transcribed part of the gene, can be attributed to a gene experiencing "boundary effects" from the attachment of transcriptional and translational machinery. We propose that the opposite effects of methylation and compositional gradients along CDS of GC_3-poor and GC_3-rich genes are the products of two or more competing processes. The first driver is transcriptional efficiency. The competing process may be cytosine deamination, which affects more methylated genes and genes that are expressed at relatively constant levels across tissues. GC_3-rich genes may be enriched in CpG dinucleotides as compared to GC_3-poor genes for a number of reasons: firstly, these sites may have played a regulatory role in the past and are maintained in the genome to allow phenotypic plasticity. Secondly, these sites may have an active regulatory role that has yet to be determined. Thirdly, cytosine deamination may reduce the frequency of CpG dinucleotides in ubiquitously expressed (GC_3-poor) genes.

List of abbreviations

GC_3 - the percentage of cytosines and guanines in third codon positions in a gene.

CDS - coding DNA sequence.

CpG, CpHpG, and CpHpH – motifs consisting of Cysteine (C), phosphate (p), Guanine (G), and H any nucleotide except guanine (H).

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

TT and EE designed the study and carried out all analyses. MP conceived of the study, and participated in its implementation. All authors were involved in preparation of the manuscript; they read and approved the final version of it.

Authors' information

TT is the founder of Genomics and Computational Biology Group at the University of South Wales, UK and an Associate Professor of Research Pediatrics at the Children's Hospital Los Angeles, University of Southern California, USA. EE is a post-doctorate fellow in the Department of Mental Health, Johns

Hopkins University Bloomberg School of Public Health. MP is a professor at Department of Molecular, Cell and Developmental Biology, UCLA.

Acknowledgements

EE work was supported in part by NIH training grant T32MH014592. The authors would like to thank Professor Roger Jelliffe, USC, for proofreading the manuscript and two anonymous reviewers for their helpful suggestions.

References

Aamodt RM. 2009. Age-and caste-dependent decrease in expression of genes maintaining DNA and RNA quality and mitochondrial integrity in the honeybee wing muscle. Exp Gerontol. 2009 Sep;44(9):586-93.

Aceituno, F.F., Moseyko, N., Rhee, S.Y. & Gutiérrez, R.A., 2008. The rules of gene expression in plants: Organ identity and gene body methylation are key factors for regulation of gene expression in Arabidopsis thaliana. BMC Genomics, 9(438).

Adams, J., 2008. Imprinting and genetic disease: Angelman, Prader-Willi and Beckwith-Weidemann syndromes. Nature Education, 1(1).

Ahmad, T. et al., 2013. Evaluation of Codon Biology in Citrus and Poncirus trifoliata Based on Genomic Features and Frame Corrected Expressed Sequence Tags. DNA research.

Anastasiadou, C., Malousi, A., Maglaveras, N., Kouidou, S. 2011, Human epigenome data reveal increased CpG methylation in alternatively spliced sites and putative exonic splicing enhancers. DNA Cell Biol. 2011 May;30(5):267-75.

Bell, J. et al., 2011. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biology, 12(1).

Bernal, M. et al., 2012. Transcriptome Sequencing Identifies SPL7-Regulated Copper Acquisition Genes FRO4/FRO5 and the Copper Dependence of Iron Homeostasis in Arabidopsis. Plant Cell..

Bhasin, M., Zhang, H., Reinherz, E. & Reche, P., 2005. Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS letters, 579(20), pp.4302-08.

Bird, A. et al., 1995. Studies of DNA methylation in animals. J Cell Sci Suppl., 19, pp.37-39.

Bock, C. et al., 2006. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. PLoS Genet, 2(e26).

Bock, C. et al., 2010. Web-based analysis of (Epi-) genome data using EpiGRAPH and Galaxy. Methods Mol Biol., 628, pp.275-96.

Chan, S., Henderson, I. & Jacobsen, S., 2005. Gardening the genome: DNA methylation in Arabidopsis thaliana. Nat. Rev Genet. , 6(5), pp.351-60.

Chodavarapu, R. et al., 2010. Relationship between nucleosome positioning and DNA methylation. Nature.

Chodavarapu, R.K. et al., 2012. Transcriptome and methylome interactions in rice hybrids. PNAS.

Duret, L., 2008. Neutral Theory: The Null Hypothesis of Molecular Evolution. Nature Education, 1(1).

Duret, L. and Arndt, P.F., 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genetics, 4(e1000071).

Duret, L. and Galtier, N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet, 2009, 10: 285-311.

Eckhardt, F.L.J. et al., 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. Nature Genet., 38, p.1378–1385.

Elhaik, E., Landan, G. and Graur, D., 2009. Can GC Content at Third-Codon Positions Be Used as a Proxy for Isochore Composition?. Mol. Biol. Evol., 26, pp.1829-33.

Elhaik, E. and Tatarinova, T., 2012. GC3 Biology in Eukaryotes and Prokaryotes. In T. Tatarinova & O. Kerton, eds. DNA Methylation - From Genomics to Technology. InTech.

Esteller, M., 2002. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. Oncogene, 21(35), pp.5427-40.

Fang, F., Fan, S., Zhang, Z. & Zhang, M., 2006. Predicting methylation status of CpG islands in the human brain. Bioinformatics, 22(18), pp.2204-09.

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

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

Foret, S. et al., 2012. DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. PNAS.

Gavery MR, Roberts SB. 2010. DNA methylation patterns provide insight into epigenetic regulation in the Pacific oyster (*Crassostrea gigas*). BMC Genomics. 11:483.

Goll, M. et al., 2006. Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. Science, 311(5759), pp.395-8.

Günther, T., Lampei, C., and Schmid, KJ. 2012, Mutational Bias and Gene Conversion Affect the Intraspecific Nitrogen Stoichiometry of the Arabidopsis thaliana Transcriptome, Mol Biol Evol, Oct 31 2012.

Hellman, A. & Chess, A., 2007. Gene body-specific methylation on the active X chromosome. Science, 315(5815), pp.1141-43.

Jjingo, D. et al., 2012. On the presence and role of human gene-body DNA methylation. Oncotarget, 3(4).

Katzman, S, Capra, JA, Haussler, D, and Pollard, KS. 2011, Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. Genome Biol Evol. 2011;3:614-26.

Kent, CF., Minaei, S., Harpur, BA. and Zayed, A, 2012, Recombination is associated with the evolution of genome structure and worker behavior in honey bees, PNAS, vol. 109 no. 44.

Kim, SH. and Yi, S. 2007, Understanding relationship between sequence and functional evolution in yeast proteins. Genetica, 131: 151.

Krauss, V. and Reuter, G., 2011. DNA methylation in Drosophila-a critical evaluation. Prog Mol Biol Transl Sci., 101, pp.177-91.

Lengauer, C., 2007. DNA Methylation. McGraw-Hill Encyclopedia of Science & Technology ed. New York: McGraw-Hill.

Li, D. et al., 2008. CpG methylation plays a vital role in determining tissue- and cell-specific expression of the human cell-death-inducing DFF45-like effector A gene through the regulation of Sp1/Sp3 binding. Nucleic Acids Res., 36(1), p.330–341.

Lisch, D., 2009. Epigenetic regulation of transposable elements in plants. Annu Rev Plant Biol, 60, pp.43-66.

Li, Z. et al., 2008. High-Resolution Mapping of Epigenetic Modifications of the Rice Genome Uncovers Interplay between DNA Methylation, Histone Methylation, and Gene Expression. The Plant Cell, 20(2), pp.259-76.

Lopez-Serra, P. & Esteller, M., 2012. DNA methylation-associated silencing of tumor-suppressor microRNAs in cancer. Oncogene, 31(13), p.1609–1622.

Lyko, F. et al., 2010. The Honey Bee Epigenomes: Differential Methylation of Brain DNA in Queens and Workers.. PLoS Biology.

Lyko, F., Ramsahoye, B. & Jaenisch, R., 2000. DNA methylation in Drosophila melanogaster.. Nature , 408, pp.538-40.

Ma, A., Zhou, X.-J. & Wang, Y.-X., 2010. Histone deacetylation directs DNA methylation in survivin gene silencing. BBRC, 404, pp.268-72.

Mukhopadhyay, P. & Ghosh, T.C., 2010. Relationship between gene compactness and base composition in rice and human genome. J Biomol Struct Dyn., 27, pp.477-88.

Muyle, A., Serres-Giardi, L., Ressayre, A., Escobar, J. Glémin, S. 2011.GC-Biased Gene Conversion and Selection Affect GC Content in the Oryza Genus (rice). Molecular Biology and Evolution , Volume: 28, Issue: 9, Publisher: SMBE, Pages: 2695-2706.

Nagai, M. & Meguro-Horike, M.a.H.S.-i., 2012. Epigenetic Defects Related Reproductive Technologies: Large Offspring Syndrome (LOS). In Kerton, T.T.a.O. DNA Methylation - From Genomics to Technology. InTech.

Nanty, L. et al., 2011. Comparative methylomics reveals gene-body H3K36me3 in Drosophila predicts DNA methylation and CpG landscapes in other invertebrates. Genome Res. , 21(11), pp.1841-50.

Oliver, J. & Marin, A., 1996. A relationship between GC content and coding-sequence length. J Mol Evol., 3(2), pp.216-23.

Rakyan, V.K. et al., 2004. DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. PLoS Biol., 2(e405).

Roberts, SB, Gavery, MR. 2012. Is There a Relationship between DNA Methylation and Phenotypic Plasticity in Invertebrates? Front Physiol. 2:116.

Sablok, G., Nayak, K., Vazquez, F. & Tatarinova, T., 2011. Synonymous codon usage, GC(3), and evolutionary patterns across plastomes of three pooid model species: emerging grass genome models for monocots. Mol Biotechnol., 49(2), pp.116-28.

Sadikovic, B.et.al., 2008. Cause and Consequences of Genetic and Epigenetic Alterations in Human Cancer. Current Genomics, 9, pp.394-408.

Sadikovic, B., Al-Romaih, K., Squire, J.A. & Zielenska, M., 2008. Cause and Consequences of Genetic and Epigenetic Alterations in Human Cancer. Current Genomics, 9(6), pp.394-408.

Slack, J.M.W., 2002. Conrad Hal Waddington: the last Renaissance biologist? Nature Reviews: Genetics, 3, pp.889-95.

Sonkin, D., Hassan, M., Murphy, D. & Tatarinova, T. 2013. Tumor Suppressors Status in Cancer Cell Line Encyclopedia, Molecular Oncology, 10.1016/j.molonc.2013.04

Spannoff, A. et al., 2011. Histobe deacetylase inhibitor activity in royal jelly might facilitate caste switching in bees. EMBO reports, 12, pp.238-43.

Suzuki, M. & Bird, A., 2008. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet., 9(6), pp.465-76.

Takuno, S. & Gaut, B.S., 2012. Body-Methylated Genes in Arabidopsis thaliana Are Functionally Important and Evolve Slowly. Mol Biol Evol.

Tatarinova, T., Alexandrov, N., Bouck, J. & Feldmann, K., 2010. GC3 biology in corn, rice, sorghum and other grasses. BMC Genomics, 11(308).

Tatarinova, T., Brover, V., Troukhan, M. & Alexandrov, N., 2003. Skew in CG content near the transcription start site in Arabidopsis thaliana. Bioinformatics, 19, pp.313-14.

Tatarinova, T. & Kerton, O., 2012. DNA Methylation - From Genomics to Technology. InTech.

Xia, X., Xie, Z. & Li, W., 2003. Effects of GC content and mutational pressure on the lengths of exons and coding sequences. J Mol Evol. , 56(3), pp.362-70.

Xiang H, Zhu J, Chen Q, Dai F, Li X, Li M, Zhang H, Zhang G, Li D, Dong Y, Zhao L, Lin Y, Cheng D, Yu J, Sun J, Zhou X, Ma K, He Y, Zhao Y, Guo S, Ye M, Guo G, Li Y, Li R, Zhang X, Ma L, Kristiansen K, Guo Q, Jiang J, Beck S, Xia Q, Wang W, Wang J: Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat Biotechnol 2010, 28:516-520.

Yan, C., Kim, Y.-W., Ha, Y.-S. & al, e., 2011. RUNX3 methylation as a predictor for disease progression in patients with non-muscle-invasive bladder cancer. Journal of Surgical Oncology, 105(4), pp.425-30.

Zemach, A., McDaniel, I., Silva, P. & Zilberman, D., 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science, 328(5980), pp.916-19.

Zhou, X., Li, Z., Dai, Z. & Zou, X., 2012. Prediction of methylation CpGs and their methylation degrees in human DNA sequences. Computers in Biology and Medicine, 42, pp.408-13.

Zilberman, D. et al., 2007. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nature genetics, 39(1), pp.61-69.

Ziller, M.J. et al., 2011. Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types. PLoS Genet., 7(12), p.e1002389.

Figures

Figure 1: Distributions of GC_3 content for rice, Arabidopsis, bee and human.

Figure 1: Distributions of GC_3 content for rice, Arabidopsis, bee and human.

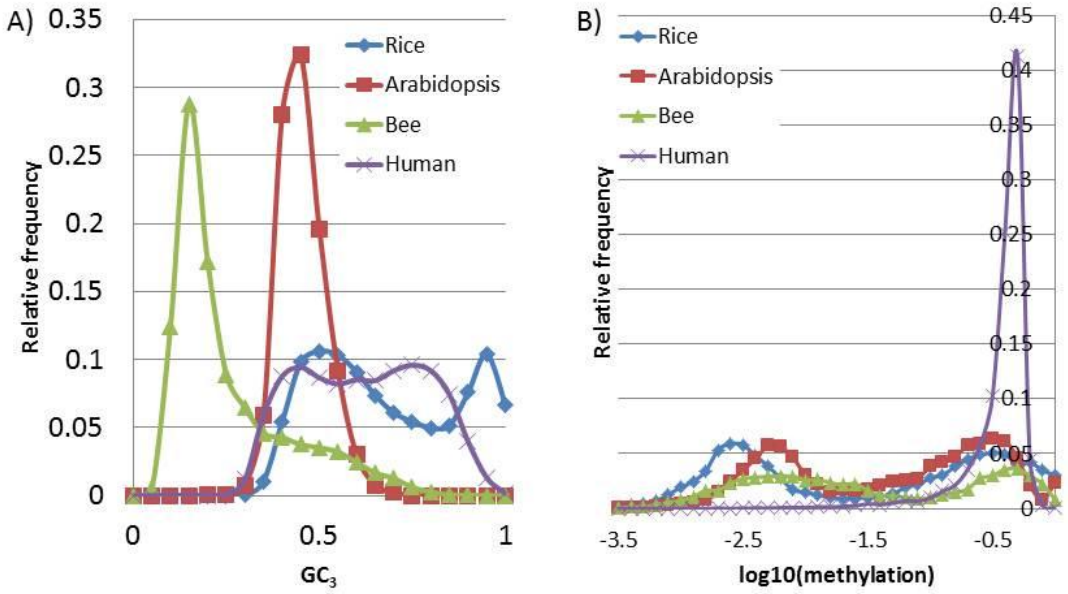


Figure 2: GC_3 vs. Expression for 4 organisms: Bee (green), Rice (blue), Arabidopsis (red), and Human (purple). (A) shows relationship between standardized values of GC_3 and average expression, (B) shows gene expression variability as a function of GC_3. Every point represents a mean across at least 100 genes and the standard error of the mean does not exceed 0.1 (plot A) and 0.06 (plot B).

Figure 2: GC_3 vs. Expression for 4 organisms: Bee (green), Rice (blue), Arabidopsis (red), Human (purple). (A) shows relationship between standardized values of GC_3 and average expression, (B) shows gene expression variability as a function of GC_3. Every point represents a mean across at least 100 genes and the standard error of the mean does not exceed 0.1 (plot A) and 0.06 (plot B).

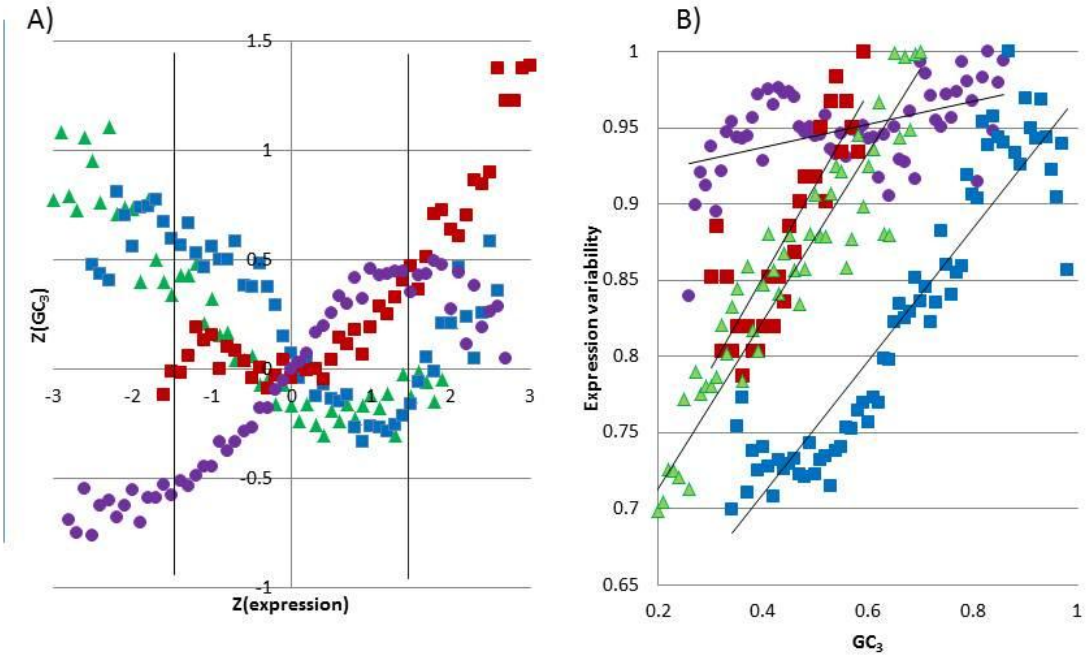


Figure 3: *Oryza sativa*: Relationship between GC_3 (purple triangles), gene expression strength (blue diamonds), expression variability (red squares) and methylation. Standard error of the mean is below 0.03 (GC_3), 0.11 (expression), 0.03 (expression variability).

Figure *3.Oryza sativa*: Relationship between GC_3 (purple triangles), gene expression strength (blue diamonds), expression variability (red squares) and methylation. Standard error of the mean is below 0.03 (GC_3), 0.11 (expression), 0.03 (expression variability).

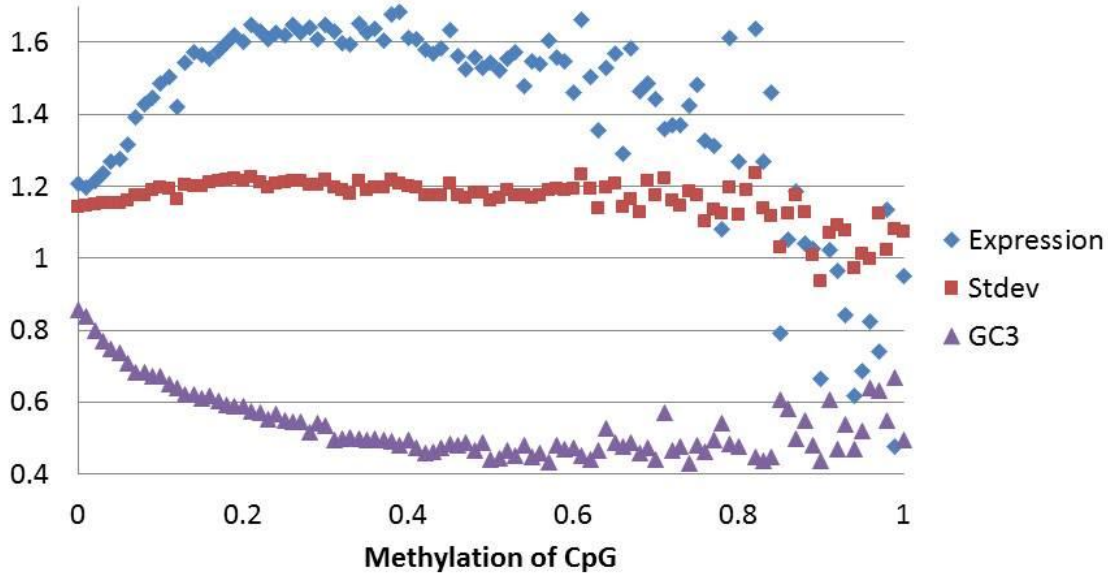


Figure 4: *Arabidopsis thaliana.* Methylation level in Arabidopsis as a function of GC_3 (A) and differential methylation between shoots and roots (B). Blue diamond: CG, Red square: C, Green triangle: CHG, violet cross (CHH). Every point represents an average across 100 or more genes. The absolute relative difference is calculated as $\frac{|Meth_{shoots}-Meth_{roots}|}{Meth_{shoots}+Meth_{roots}}$.Standard error of the mean does not exceed 0.05 for the mean methylation levels calculation.

Figure 4: Methylation level in Arabidopsis as a function of GC_3 (left) and differential methylation between shoots and roots (right). Blue diamond: CG, Red square: C, Green triangle: CHG, violet cross (CHH). Every point represents an average of at least 100 genes. Standard error of the mean does not exceed 0.05 for the mean methylation levels calculation.

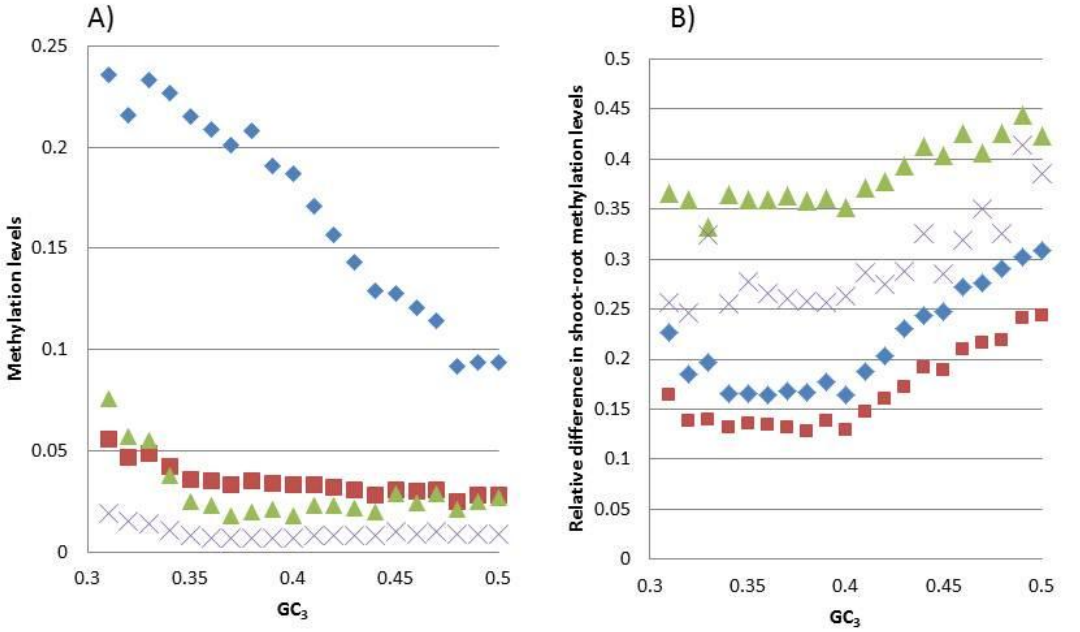


Figure 5: *Apis mellifera:* (A) Relative difference in gene body methylation levels $\frac{Q-W}{Q+W}$ as a function of GC_3 between worker and queen bee. (B) Difference in gene body methylation between worker and queen bee as a function of GC_3 (C) Queen and (D) Worker bee methylation as a function of GC_3. Every point represents an average of at least 228 genes. Standard error of the mean for methylation levels was below 0.006.

Figure 5: (A) Relative difference in gene body methylation between worker and queen bee. (B) Difference in gene body methylation between worker and queen bee as a function of GC_3 (C) Queen and (D) Worker bee methylation as a function of GC_3. Every point represents an average of at least 228 genes. Standard error of the mean for methylation levels was below 0.006.

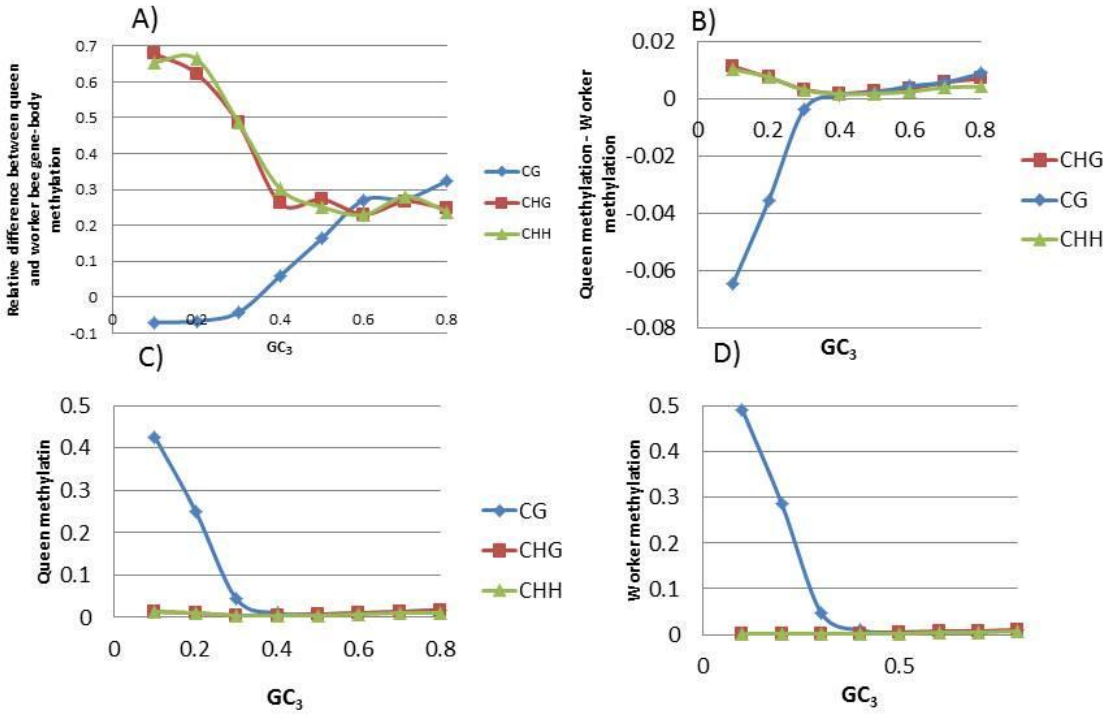
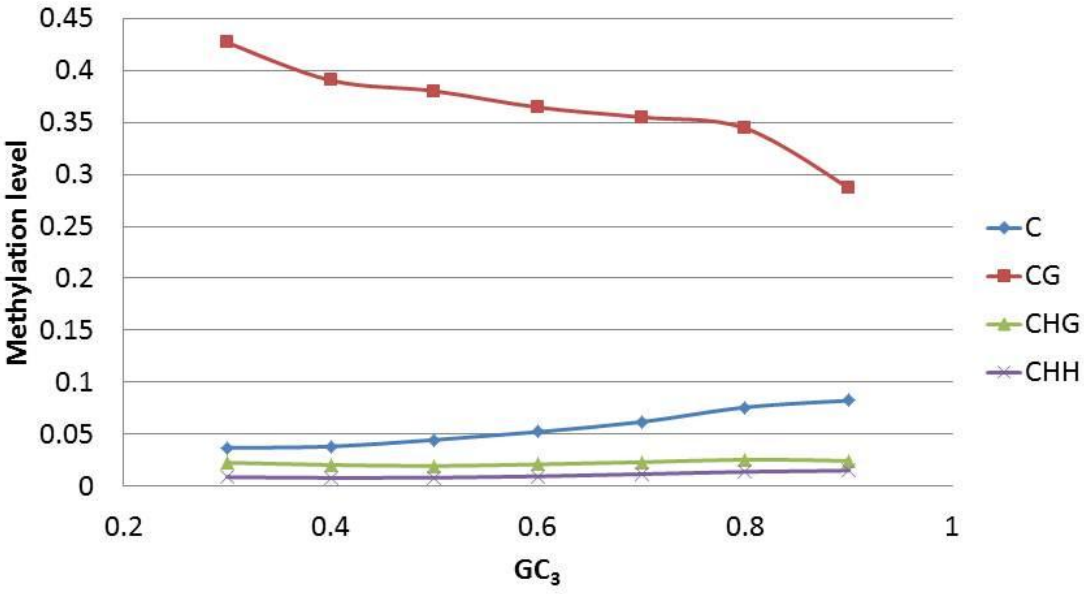


Figure 6: *Homo sapiens*: Methylation as a function of GC_3. Every point represents a mean across at least 100 genes, with standard error of the mean not exceeding 10% of the mean. Methylation as a function of GC_3

Figure 6: Homo sapiens: Methylation as a function of GC_3. Every point represents a mean across at least 100 genes, with standard error of the mean not exceeding 10% of the mean.



Tables

Table 1: Pearson's correlation coefficients between CpG methylation, GC_3 and gene expression variability for *O. sativa, A. thaliana, A. mellifera,* and *H. sapiens.* Top numbers in each cell represents Pearson's correlation coefficients and bottom numbers represent partial correlation coefficients.

Correlation between	*O. sativa*	*A. thaliana*	*A. mellifera*	*H. sapiens*
CpG methylation and GC_3	-0.67	-0.27	-0.65	-0.23
	-0.65	-0.23	-0.62	-0.23
CpG methylation and gene expression variability (CV)	-0.18	-0.18	-0.24	-0.02
	-0.06	-0.13	-0.04	-0.06
Gene expression variability (CV) and GC_3	0.21	0.16	0.34	-0.16
	0.12	0.12	0.22	-0.16

Table 2: Compactness of rice genes, stratified by expression and GC_3

GC_3	EXON LENGTH	EXONS	INTRON DENSITY (per 1000nt)	INTRON LENGTH	INTRON FRACTION (length)	NUMBER OF ORFs	EXPRESSION (standardized)
GC_3>0.800	767	2.47	2.301	1683	62.4%	428	E>1
	1132	2.21	1.132	1085	41.9%	1215	E<-1
GC_3<0.491	1503	9.60	6.296	4249	73.3%	924	E>1
	1587	5.41	3.090	3116	60.1%	386	E<-1

Table 3: Four classes of (n=9,607) rice genes by GC_3 and methylation. Yates's χ^2= 4267.237

	GC_3-rich	GC_3-poor
High methylation	289	4787
Low methylation	3161	1370

Table 4: Pearson's correlation coefficients between GC_3, methylation (AVG_MET**), average expression (**AVG _EXP**), standard deviation of gene expression (**STD_EXP**), coefficient of variation of gene expression (**CV_EXP**) and gene length (**LENGTH**) for intron-less genes and genes with introns Calculated for 2,648 intron-less genes and 6,959 intron-containing genes. 95% CI is shown in square brackets below each correlation value.**

Type	AVG_MET and GC_3	AVG _EXP and GC_3	STD_EXP and GC_3	CV_EXP and GC_3	LENGTH and GC_3	AVG_MET and AVG _EXP	AVG_MET and CV_EXP
Intron-less genes	-0.602 [-0.626,-0.577]	0.103 [0.065, 0.141]	0.187 [0.149, 0.223]	-0.075 [-0.113,-0.037]	-0.235 [-0.270,-0.198]	0.038 [-0.001, 0.075]	-0.017 [-0.055, 0.021]
Intron-containing genes	-0.671 [-0.684,-0.657]	-0.230 [-0.252,-0.208]	0.000 [-0.023,0.024]	0.245 [0.222,0.267]	-0.307 [-0.328,-0.286]	0.233 [0.211,0.255]	-0.209 [-0.231,-0.186]