DNA Methylation is Associated with an Increased Level of Conservation at Nondegenerate Nucleotides in Mammals

Trees-Juen Chuang*[,1] and Feng-Chi Chen*[,2,3,4]

[1]Physical and Computational Genomics Division, Genomics Research Center, Academia Sinica, Taipei, Taiwan

[2]Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Miaoli County, Taiwan

[3]Department of Life Science, National Chiao-Tung University, Hsinchu, Taiwan

[4]Department of Dentistry, China Medical University, Taichung, Taiwan

***Corresponding author:** E-mail: trees@gate.sinica.edu.tw; fcchen@nhri.org.tw.

Associate editor: Takashi Gojobori

Abstract

DNA methylation at CpG dinucleotides can significantly increase the rate of cytosine-to-thymine mutations and the level of sequence divergence. Although the correlations between DNA methylation and genomic sequence evolution have been widely studied, an unaddressed yet fundamental question is how DNA methylation is associated with the conservation of individual nucleotides in different sequence contexts. Here, we demonstrate that in mammalian exons, the correlations between DNA methylation and the conservation of individual nucleotides are dependent on the type of exonic sequence (coding or untranslated), the degeneracy of coding nucleotides, background selection pressure, and the relative position (first or nonfirst exon in the transcript) where the nucleotides are located. For untranslated and nonzero-fold degenerate nucleotides, methylated sites are less conserved than unmethylated sites regardless of background selection pressure and the relative position of the exon. For zero-fold degenerate (or nondegenerate) nucleotides, however, the reverse trend is observed in nonfirst coding exons and first coding exons that are under stringent background selection pressure. Furthermore, cytosine-to-thymine mutations at methylated zero-fold degenerate nucleotides are predicted to be more detrimental than those that occur at unmethylated nucleotides. As zero-fold and nonzero-fold degenerate nucleotides are very close to each other, our results suggest that the "functional resolution" of DNA methylation may be finer than previously recognized. In addition, the positive correlation between CpG methylation and the level of conservation at zero-fold degenerate nucleotides implies that CpG methylation may serve as an "indicator" of functional importance of these nucleotides.

Key words: **DNA methylation, methylation-associated mutation, single-nucleotide evolution, degeneracy of nucleotide, genomics.**

Introduction

The mutation of nucleotide is the ultimate driving force of evolution. In exonic regions, the majority of mutations, especially those that occur in coding exons, are destined to selective elimination because of their deleterious effects (Kimura 1983; Nei et al. 2010). Therefore, the currently observable nucleotide substitutions represent past mutations that have been screened and retained by natural selection.

The rate of mutation varies significantly across genomic regions because of differences in such features as the type of genomic sequence, G + C content, chromosomal locations, presence of repetitive elements, recombination rate, and the level of DNA methylation (Ellegren et al. 2003; Ananda et al. 2011; Hodgkinson and Eyre-Walker 2011). Among these genomic features, DNA methylation is arguably the most influential because it can lead to spontaneous cytosine-to-thymine (C-to-T) transitions at a rate ten times higher than the genome-wide average (Coulondre et al. 1978; Bird 1980; Holliday and Grigg 1993). Interestingly, however, mutagenesis is not the only biological role of DNA methylation. DNA methylation is also important for a variety of biological functions, including genomic imprinting (Li et al. 1993), X-chromosome inactivation (Heard et al. 1997), DNA–protein interactions (Mancini et al. 1999; Hark et al. 2000; Lister et al. 2009), the silencing of transposable elements (Walsh et al. 1998), tumorigenesis (Feinberg and Tycko 2004), embryogenesis and development (Reik 2007; Laurent et al. 2010), and regulations of gene expression and mRNA splicing (Anastasiadou et al. 2011; Shukla et al. 2011; Jones 2012). Recent studies on single-base DNA methylomes have shown that DNA methylation is unevenly distributed within the gene body (Lister et al. 2009; Schwartz et al. 2009; Laurent et al. 2010). Furthermore, DNA methylation was suggested to be a strong marker of exonic regions and exon–intron boundaries and possibly be important for regulating mRNA splicing (Laurent et al. 2010; Lyko et al. 2010; Anastasiadou et al. 2011; Gelfman et al. 2013). As regulations of mRNA splicing and gene expression are heavily targeted by natural selection, methylated exonic nucleotides may be functionally important and selectively constrained. Consistent with this notion,

densely methylated genes have been reported to be more conserved than sparsely methylated genes (Hunt et al. 2010; Lyko et al. 2010; Park et al. 2011). Meanwhile, DNA methylation is also related to within-gene variations in evolutionary rate. We have previously shown that DNA methylation (measured by the density of methylated CpG dinucleotides) has fairly complicated correlations with exonic evolutionary rates, and that the relative position of the exon (first or nonfirst exon) in the transcript appears to be an important determinant of such correlations (Chuang et al. 2012). In addition, the rates of nonsynonymous and synonymous substitution differ significantly from each other in their correlations with exonic level of DNA methylation (Chuang et al. 2012). Notably, however, these previous studies focused on the evolutionary rates of genes or exons as a whole. How DNA methylation is associated with the conservation level of individual nucleotides in exonic regions remains underexplored.

An in-depth understanding of the correlation between DNA methylation and single-nucleotide evolution has important biological implications because 1) the exon-level (and gene-level) sequence evolution is the combinatorial result of single-nucleotide substitutions; 2) individual nucleotides within the same exonic regions are subject to different selective constraints depending on their biological roles (e.g., untranslated region [UTR] vs. coding sequence [CDS]; zero-fold vs. nonzero-fold degenerate sites); 3) substitutions that occur at coding nucleotides (particularly zero-fold and two-fold degenerate sites) can lead to various types of amino acid substitutions and thus may have different fitness effects; 4) the fitness effects of amino acid substitutions are dependent on the local sequence contexts and the functional importance of the peptides; and 5) the biological significance of DNA methylation in the gene body has remained unclear. Therefore, exploring how DNA methylation is correlated with single-nucleotide conservation can not only push the resolution of methylation-related sequence evolution to the very basic unit but also help clarify the role of DNA methylation in the gene body.

To address this issue, we collected base-resolution DNA methylation data from six human cell types (including both normal cells and cell lines) and systematically examined the correlations between DNA methylation and the level of single-nucleotide conservation in mammals. Our results indicate that such correlations differ not only between different types of genomic sequences (i.e., UTR vs. CDS) and different positions of exons (i.e., first vs. nonfirst exons) but also between coding nucleotides of different levels of degeneracy (i.e., zero-fold vs. nonzero-fold degenerate sites). Intriguingly, for specific groups of nucleotides, such as zero-fold degenerate (also known as "nondegenerate") nucleotides, methylated sites have an increased level of conservation. In addition, the C-to-T mutations at methylated zero-fold degenerate nucleotides are predicted to be more damaging than those that occur at unmethylated nucleotides. These observations appear to suggest that CpG methylation serves as a "signature" of biological importance at certain nucleotides, where mutations may be detrimental and thus are disfavored by natural selection.

Results

Different Exonic Regions Vary Significantly in Epigenetic and Evolutionary Properties

To investigate the correlations between DNA methylation and the level of single-nucleotide conservation in human exonic sequences, we retrieved single-base-resolution DNA methylation data from six different human cell types, including nondifferentiated (embryonic stem cells [ESCs]), lowly differentiated (lung and skin fibroblasts, fibroblastic derivatives of ESCs), and differentiated cells (blood mononuclear cells) (see Materials and Methods; table 1). It has been reported that DNA methylation patterns may vary considerably between cell types (Lister et al. 2009; Laurent et al. 2010; Chuang et al. 2012). Therefore, if the analysis results derived from different cell types are consistent with one another, such results may be considered as robust against variations in DNA methylation patterns. Table 1 shows the number of exons and methylated/unmethylated nucleotides examined in each data set. On average, more than 60% of the CpGs were sampled for the examined exons. The CpG dinucleotides in the examined exons therefore appear to be adequately sampled.

As natural selection and DNA methylation may differentially affect different types of exonic regions (Chuang et al. 2012), we classified human exonic regions into six groups according to whether they are translated and their relative positions in the transcript: 5′UTR-First (5U-F), 5′UTR-nonFirst (5U-nonF), CDS-First (CDS-F), CDS-nonFirst (CDS-nonF), 3′UTR-First (3U-F), and 3′UTR-nonFirst (3U-nonF) (see Materials and Methods; fig. 1*A*). We then examined the basic sequence features of these six groups of exonic regions. As expected, the DNA methylation levels (measured by the "mCG density"; see Materials and Methods) vary significantly among different groups of exonic regions, with 5′UTRs (whether they are located at the first exons or nonfirst exons) and first exons (whether they are UTRs or CDSs) having significantly reduced average mCG density as compared with the other two regions (supplementary fig. S1*A* and *B*, Supplementary Material online). In addition, the first exonic regions (5U-F, CDS-F, and 3U-F) have higher G + C contents and CpG densities than the corresponding nonfirst exonic regions (5U-nonF, CDS-nonF, and 3U-nonF, respectively) (supplementary fig. S1*C*, Supplementary Material online). These observations are consistent with previous findings that the level of DNA methylation is negatively correlated with CpG density (Lister et al. 2009; Laurent et al. 2010; Chuang et al. 2012).

Next, we examined the correlations between exonic mCG density and the $CpG_{O/E}$ ratio (see Materials and Methods). As a high level of C-to-T mutation (which leads to a low $CpG_{O/E}$ ratio) has been suggested to result mostly from DNA methylation (Bird and Taggart 1980; Park et al. 2011; Park et al. 2012), a negative correlation between $CpG_{O/E}$ ratio and mCG density is expected (Bird and Taggart 1980; Zemach et al. 2010; Park et al. 2011). A higher level of negative correlation between the two measurements was suggested to reflect a higher proportion of methylated CpGs having undergone

Table 1. The Base-Resolution DNA Methylation Data Sets Used in this Study.

Data Set	Description (Ref.)	No. of Exons (Average CpG Coverage)[a]	No. of Methylated Sites	No. of Unmethylated Sites
S1	Peripheral blood mononuclear cells (Li et al. 2010)	29,865 (63.19%)	305,364	204,848
S2	H1 human ESCs (Lister et al. 2009)	48,833 (84.75%)	841,279	244,795
S3	IMR90 fetal lung fibroblasts (Lister et al. 2009)	56,614 (89.42%)	721,143	515,777
S4	WA09 human ESCs (Laurent et al. 2010)	53,476 (90.09%)	691,427	883,790
S5	Fibroblastic differentiated derivatives of WA09 hESCs (Laurent et al. 2010)	50,222 (86.82%)	696,576	809,697
S6	Neonatal human foreskin fibroblasts (Laurent et al. 2010)	50,776 (87.43%)	544,756	900,590

[a]Coverage of the CpG dinucleotides for each exon = (number of the sampled CpG dinucleotides)/(number of the sampled CpG dinucleotides + number of the nonsampled CpG dinucleotides).

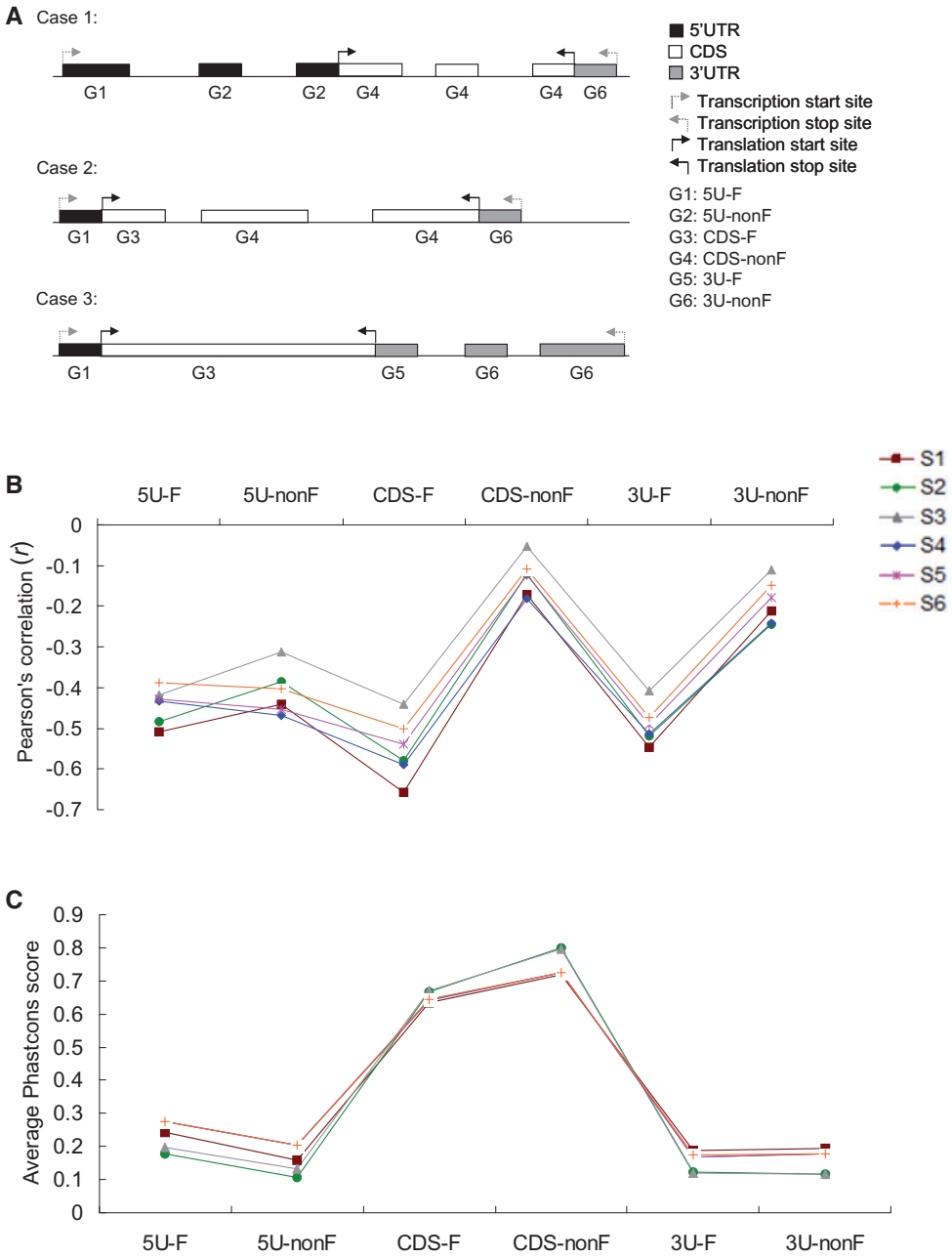

Fig. 1. Comparisons of evolutionary properties among different types of exonic regions for the six cell types examined. (*A*) Examples of six groups of exonic regions categorized: G1, 5U-F; G2, 5U-nonF; G3, CDS-F; G4, CDS-nonF; G5, 3U-F; and G6, 3U-nonF. (*B*) Pearson's coefficient of correlation (*r*) between $CpG_{O/E}$ and mCG density. (*C*) Average conservation scores (phastCons scores) of exons for the six exonic groups.

mutation (Chuang et al. 2012). As shown in figure 1*B*, we find that $CpG_{O/E}$ ratios are indeed significantly negatively correlated with mCG density, regardless of the type of exonic region (all P values $<10^{-15}$). Interestingly, the profiles of the mCG density–$CpG_{O/E}$ correlations are similar to those of the mCG densities among different exonic regions (fig. 1*B* and supplementary fig. S1*A*, Supplementary Material online), with CDS-nonF and 3U-nonF exonic regions having the weakest correlations. This observation suggests that DNA methylation in CDS-nonF and 3U-nonF regions may be less effective in causing mutations. By contrast, the mutagenic effect of DNA methylation is stronger in the other four groups of regions (i.e., 5′UTRs and the first exons) despite the lower levels of DNA methylation in these exonic regions (supplementary fig. S1*A*, Supplementary Material online). This observation implies that other factors (presumably selection-related forces) may be involved in shaping the substitution profiles of human exonic regions. Furthermore, we compared the conservation levels (measured by the average phastCons score (Siepel et al. 2005) of the six groups of exonic regions. Figure 1*C* shows that, as expected, CDS-nonF regions have the highest average phastCons scores, with CDS-F regions following closely. All of the four UTR regions have significantly lower phastCons scores when compared with CDS (supplementary table S1, Supplementary Material online), consistent with our understanding of sequence conservation in exonic regions (Graur and Li 2000). The differences in basic properties between exonic regions thus support the necessity of classification while investigating the correlations between exonic DNA methylation and single-nucleotide conservation.

Methylated Nucleotides Evolve More Slowly Than Unmethylated Nucleotides in Highly Conserved Coding Regions

We then examined how DNA methylation may affect the conservation of individual nucleotides (measured by the PhyloP score [Pertea et al. 2011]) in the six groups of exonic regions. Theoretically, the mutagenic effect of DNA methylation should decrease the level of nucleotide conservation. In other words, methylated nucleotides are expected to have lower PhyloP scores than unmethylated nucleotides. Interestingly, however, although this expected correlation is observed in five of the six groups of exonic regions (5U-F, 5U-nonF, CDS-F, 3U-F, and 3U-nonF), the reverse is true for CDS-nonF exonic regions (fig. 2*A* and supplementary fig. S2*A*, Supplementary Material online). This observation is remarkable considering that CDS-nonF exonic regions actually have the highest average mCG density among the six groups (supplementary fig. S1*A*, Supplementary Material online). This result also echoes our above finding that in CDS-nonF regions, the mutagenic effect of DNA methylation tends to be weak (fig. 1*B*). However, even if DNA methylation does not cause

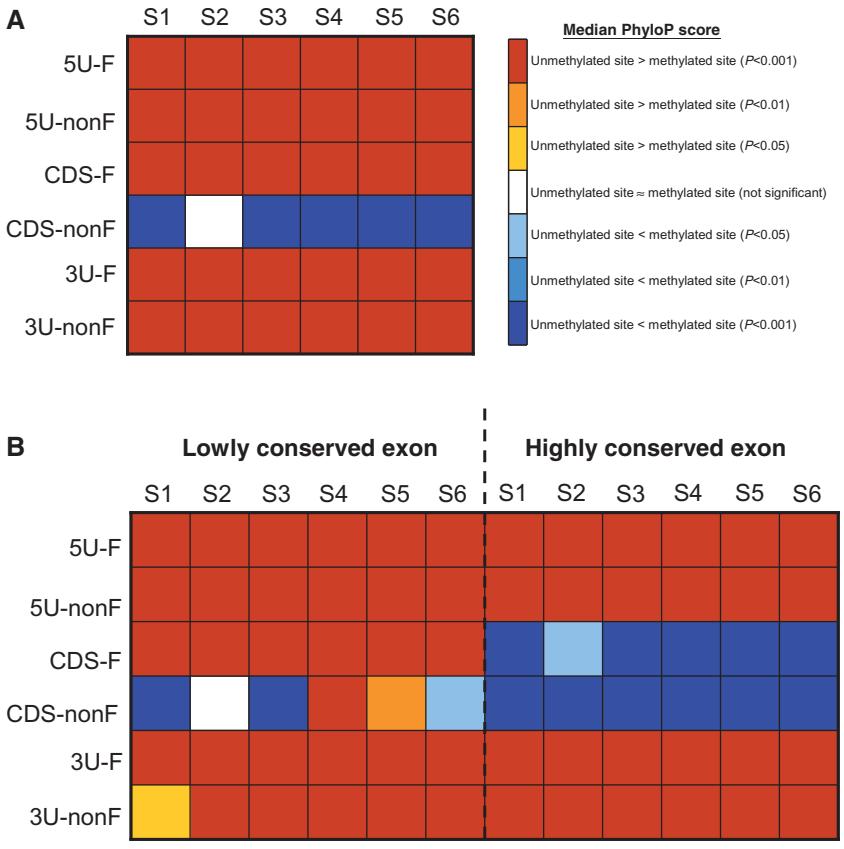

Fig. 2. Comparisons of the conservation scores (the median PhyloP scores) at methylated and unmethylated sites in six groups of exonic regions across six cell types examined (S1–S6) (*A*) before and (*B*) after dividing each of the exonic region group into lowly conserved (left column) and highly conserved regions (right column). The statistical significance was evaluated by using the two-tailed Wilcoxon rank-sum test.

any mutations at the methylated cytosines (i.e., DNA methylation has no mutagenic effects at all), the methylated sites should at best be as conserved as (but not more conserved than) the unmethylated sites. Furthermore, the weak mutagenic effect is also observed for 3U-nonF exonic regions (fig. 1*B*), but in these regions methylated nucleotides are less conserved than the unmethylated ones (fig. 2*A* and supplementary fig. S2*A*, Supplementary Material online). Therefore, the cause of the higher-than-expected conservation level at methylated CDS-nonF nucleotides is worth further explorations.

As CDS-nonF regions have a higher level of conservation than the other five groups of exonic regions (fig. 1*C*), we then ask whether the background selective constraints might have affected the correlation between DNA methylation and single-nucleotide conservation. We divided each of the six groups of exonic regions into highly conserved and lowly conserved exonic regions (see Materials and Methods) and reexamined the correlations between DNA methylation and PhyloP score. Interestingly, as illustrated in figure 2*B* and supplementary figure S2*B*, Supplementary Material online, first, for UTRs (5′UTRs and 3′UTRs), a negative methylation–PhyloP correlation is still observed regardless of background selective constraints or relative position (first or nonfirst) of the exonic region. Second, for CDS-F regions, negative methylation–PhyloP correlations are observed when these regions are lowly conserved, but the correlations turn positive when the regions are highly conserved. This latter observation differs from what is shown in figure 2*A*. Third, for CDS-nonF regions, the positive methylation–PhyloP correlation occurs in highly conserved regions. However, there is no clear trend in lowly conserved regions, with both positive and negative correlations being observed (fig. 2*B* and supplementary fig. S2*B*, Supplementary Material online). These results suggest that background selection pressure may affect the methylation–PhyloP correlations in coding regions but not in UTRs.

Degeneracy Significantly Affects the Conservation Level of Methylated Coding Nucleotides

We have shown that the difference in conservation level between methylated and unmethylated coding nucleotides is associated with the background selection pressure. As coding nucleotides of different levels of degeneracy are under different levels of selective constraints, we separated the analyzed coding nucleotides into zero-fold ($i = 0$), two-/three-fold ($i = 2$ or 3), and four-fold ($i = 4$) degenerate sites and reexamined the conservation scores of methylated/unmethylated nucleotides for each type of degenerate sites. Interestingly, for nonzero-fold ($i = 2$–4) degenerate sites, the unmethylated are more conserved than the methylated ones, regardless of the relative position of the exon (fig. 3*A* and supplementary fig. S3*A*, Supplementary Material online). By contrast, for zero-fold ($i = 0$) degenerate sites, the methylated are more conserved than the unmethylated ones in

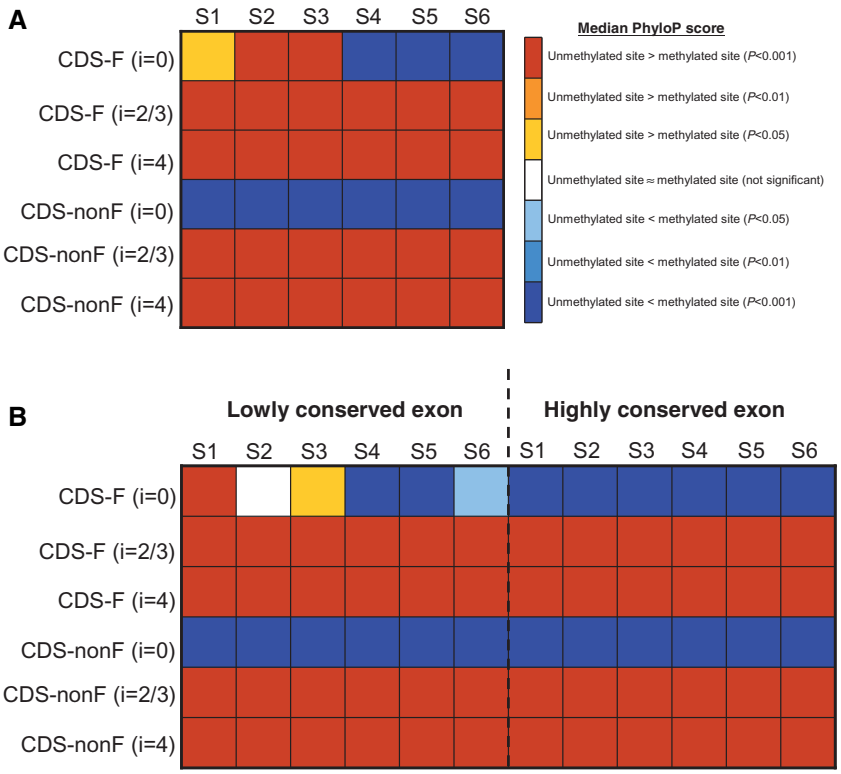

Fig. 3. Comparisons of the median PhyloP scores of zero-fold ($i = 0$), two-/three-fold ($i = 2$ or 3), and four-fold ($i = 4$) degenerate nucleotides at methylated and unmethylated sites in the coding exonic regions (including CDS-F and CDS-nonF regions) (*A*) before and (*B*) after dividing each of the exonic region group into lowly conserved (left column) and highly conserved regions (right column). The statistical significance was evaluated by using the two-tailed Wilcoxon rank-sum test.

CDS-nonF regions, whereas no clear trend is observed in CDS-F regions (fig. 3*A* and supplementary fig. S3*A*, Supplementary Material online). These observations suggest that the positive methylation-PhyloP correlation in CDS-nonF regions (fig. 2*A*) is probably dominated by zero-fold degenerate nucleotides, for neither two-/three-fold nor four-fold degenerate sites in the same regions show such a positive correlation. Similar comments may also apply to CDS-F regions—although the correlations between DNA methylation and PhyloP score vary between cell types at zero-fold degenerate sites, at nonzero-fold degenerate sites the correlations remain negative across the six cell types (fig. 3*A*). These results indicate that the degeneracy of nucleotides can significantly affect the level of nucleotide conservation at methylated sites. Of note, zero-fold degenerate sites are generally thought to be under more stringent selective constraints than nonzero-fold degenerate sites (Graur and Li 2000). Therefore, the association between degeneracy and the methylation–PhyloP correlation suggests the involvement of selective constraints in shaping the conservation profiles of methylated and unmethylated coding nucleotides.

As background selection pressure can remarkably affect the level of conservation of methylated coding nucleotides (fig. 2*B*), we are interested to know whether this factor can also affect the methylation–PhyloP correlation specifically at zero-fold degenerate nucleotides. Intriguingly, as shown in figure 3*B* and supplementary figure S3*B*, Supplementary Material online, highly conserved exonic regions show a positive methylation–PhyloP correlation in both CDS-F and CDS-nonF regions. Similar trends are also observed for lowly conserved CDS-nonF regions (fig. 3*B*). However, lowly conserved CDS-F regions show inconsistent trends in different cell types (fig. 3*B*). In contrast, for nonzero-fold ($i = 2–4$) degenerate nucleotides, negative methylation–PhyloP correlations are observed, regardless of the relative position of the exon and background selection pressure (fig. 3*B* and supplementary fig. S3*B*, Supplementary Material online). These results indicate that the effect of degeneracy on the methylation–PhyloP correlation is dependent on the sequence context, namely the relative position of the exon and background selection pressure.

C-to-T Mutations Tend to be More Damaging at Methylated Than at Unmethylated Zero-Fold Nucleotides

We have shown that for highly conserved regions, methylated zero-fold degenerate nucleotides have a higher level of conservation than those that are unmethylated. We are thus curious about whether these methylated nucleotides are biologically more important and are subject to more stringent selective constraints than unmethylated nucleotides within the corresponding exonic regions. To examine this possibility, we used SIFT (Ng and Henikoff 2003) and PolyPhen-2 (Adzhubei et al. 2010), two well-developed tools for measuring the functional consequences of nonsynonymous variants, to predict the fitness effects of potential C-to-T mutations at methylated and unmethylated zero-fold degenerate nucleotides. As shown in figure 4*A* and *B*, for both CDS-F and CDS-nonF regions, both SIFT and PolyPhen-2 predicted that C-to-T mutations at methylated zero-fold degenerate sites are generally more damaging than those that occur at unmethylated zero-fold degenerate sites (all P values <0.01 except for S2/S3 in some SIFT/PolyPhen-2 predictions, by the two-tailed Fisher's exact test). Furthermore, for both CDS-F and CDS-nonF regions, the odds ratios of damaging C-to-T mutations occurring at methylated sites over those occurring at unmethylated sites are significantly larger in highly conserved regions than in lowly conserved regions according to the predictions of SIFT and PolyPhen-2 across the six cell types (all P values <0.05 by the paired t-test). Our results thus indicate that methylation-related mutations at zero-fold degenerate nucleotides are likely to be deleterious (especially in highly conserved exonic regions). These observations imply that the positive methylation–PhyloP correlations at zero-fold degenerate nucleotides are probably associated with the selective constraints specifically imposed on these sites, and that CpG methylation may serve more as a "mark" of biological importance than a mutation driver in this context.

Discussion

In this study, we examined the correlations between DNA methylation and the conservation of individual nucleotides in exonic regions. We demonstrated that in the human genome, the level of conservation of individual exonic nucleotides is differentially correlated with DNA methylation in a sequence context-dependent manner. The sequence features that may affect such correlations include the degeneracy of the nucleotide, the background selection pressure, and the type of exonic regions where the nucleotide is located. These findings have several important implications for genomic and evolutionary studies.

To begin with, our results indicate that the methylation–PhyloP correlations differ between CDSs and UTRs and also between CDS-F and CDS-nonF regions. These observations suggest that the biological roles of DNA methylation may vary with the background selection pressure and the type of exonic regions. Of note, for CDSs (particularly for CDS-F regions), the correlations differ with the levels of exonic sequence conservation. The difference between CDSs and UTRs is expected, and the difference between CDS-F and CDS-nonF regions is also understandable (Chuang et al. 2012). However, the difference in the methylation-PhyloP correlation between highly and lowly conserved CDS-F regions and that between zero-fold and nonzero-fold degenerate nucleotides are unexpected. These differences are perhaps ascribable to three nonmutually exclusive factors. First, some CDS-F regions are close to promoter regions, whereas other CDS-F regions are not. The CDS-F regions that are closer to promoter regions may have been somewhat "homogenized" by the background sequences and become less selectively constrained than the CDS-F regions that are far away from promoters. This divergence in the distance from promoters may have divided CDS-F regions into two groups with different methylation–PhyloP correlations. By contrast, CDS-nonF regions are usually far away from promoter regions. They are

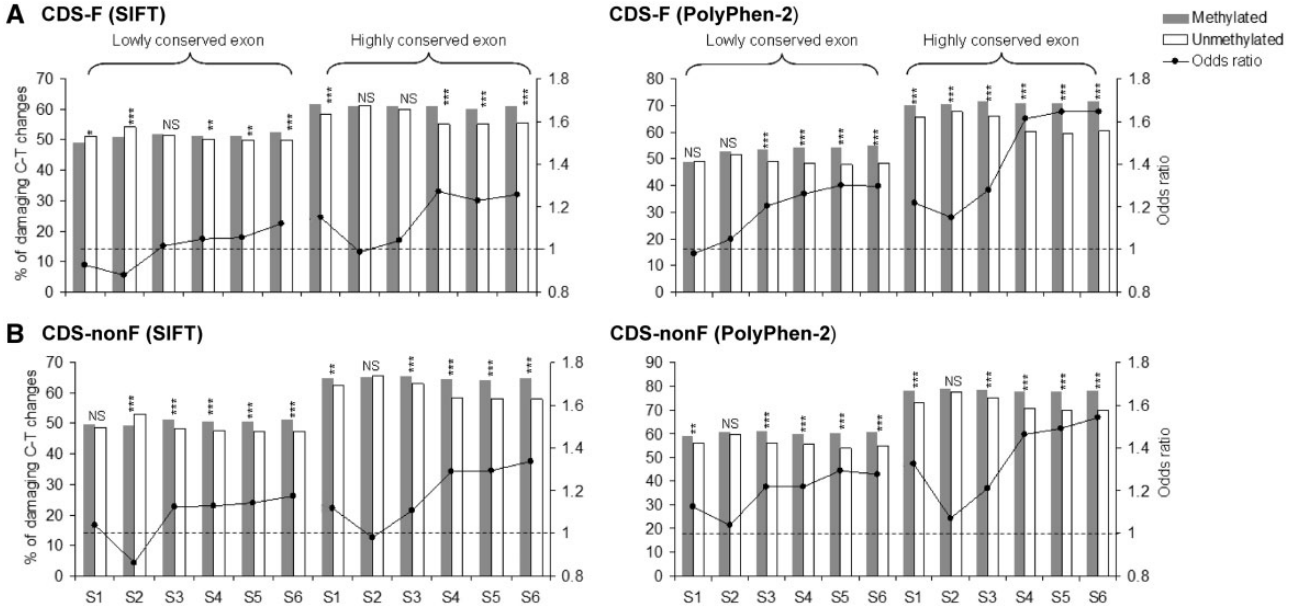

FIG. 4. The percentages of damaging C-to-T mutations at methylated and unmethylated zero-fold-degenerate sites (measured by the SIFT and PolyPhen-2 predictions) and the odds ratios of such substitutions occurring at methylated sites being predicted to be damaging over those occuring at unmethylated sites in the lowly (left column) and highly (right column) conserved regions of the (*A*) CDS-F and (*B*) CDS-nonF regions. The statistical significance was evaluated by using the two-tailed Fisher's exact test: $^{*}P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$. NS, not significant.

therefore less influenced by the distance from promoter regions and form a single group in this regard. Second, in highly conserved CDS regions, and especially at zero-fold degenerate nucleotides, the selective constraints are so strong that these nucleotides are virtually invulnerable to the mutagenic effect of DNA methylation. By comparison, the less selectively constrained lowly conserved regions and nonzero-fold degenerate nucleotides may be more tolerant for the methylation-related mutations. Therefore, the negative methylation–PhyloP correlations can be observed in the latter regions. Third, we speculate that at zero-fold degenerate nucleotides (particularly those located in CDS-nonF regions), DNA methylation may be an indicator of functional importance. To be sure, DNA methylation can be more than just an "indicator." It has been reported that in *Arabidopsis thaliana*, methylation at specific coding nucleotides are sufficient to silence a gene (Rangani et al. 2012). Therefore, DNA methylation at individual nucleotides can play important regulatory roles. It is likely that these methylated nucleotides, with their importance in regulations, are subject to more stringent selective constraints than unmethylated nucleotides. By examining the potential fitness effects of C-to-T mutations at zero-fold degenerate nucleotides, we demonstrated that these mutations tend to be more damaging at methylated than at unmethylated nucleotides (fig. 4). Furthermore, these mutations are more damaging in highly conserved regions than in lowly conserved regions (fig. 4). These observations further support the hypothesis of CpG methylation as an indicator of biological importance.

Still another implication of our results is that the "functional resolution" of DNA methylation may be finer than previously thought. The observation that zero-fold and nonzero-fold degenerate nucleotides have opposite methylation–PhyloP correlations indicates differences in the biological role of DNA methylation at these two types of nucleotides. This is somewhat surprising considering that zero-fold and nonzero-fold nucleotides are no more than two nucleotides away from each other (because all of the second sites in the codon triplets are zero-fold degenerate sites). Therefore, the functions of DNA methylation may differ not only between exonic and intronic regions as previously reported (Choi et al. 2009; Flores et al. 2012; Jones 2012; Sati et al. 2012) but also between individual nucleotides within coding regions.

Of note, here we use DNA methylation data from post-fertilization cells rather than data from germ line cells. In principle, DNA methylation in germ line cells can be more accurately related to nucleotide substitutions because only the mutations in these cells can be transferred to the next generation. Interestingly, nevertheless, one previous report indicated that the global patterns of DNA methylation in ESCs are similar to those in the germ line cells (Zechner et al. 2009). Furthermore, although the ratios of methylated-to-unmethylated number of sites vary considerably (from 0.60 to 3.44; see table 1), the overall results are generally consistent across different cell types. Of note, the six analyzed data sets include both cells derived from natural sources (i.e., S1, peripheral blood mononuclear cells) (Wang et al. 2008; Li et al. 2010) and those that were cultivated in the laboratory (i.e., S2–S6). Therefore, our results appear to be robust against variations in methylation pattern between cell types. As DNA methylation patterns may vary considerably among species (especially for plants vs. animals) (Feng et al. 2010; Zemach et al. 2010), it will be interesting to

investigate whether lineage-specific methylation patterns may contribute to alterations in the methylation–PhyloP correlation when single-base-resolution methylome data from other species become available.

Recently, the ENCyclopedia Of DNA Elements (ENCODE) Consortium reported that a significant proportion of the human genome is methylated and suggested that the methylated genomic regions are probably functional (Bernstein et al. 2012). This proposition, however, has been disputable. For instance, Graur et al. (2013) commented that not all of the methylated CpG dinucleotides are equally functional—some of the CpG dinucleotides may have been methylated simply by chance. Interestingly, our results appear to echo Dr Graur's comments in that DNA methylation in different sequence contexts may play distinct roles, even for methylation that occurs at physically close CpG dinucleotides (e.g., degenerate vs. nondegenerate sites). This is understandable if a "neutral view" of DNA methylation is considered. In other words, akin to mutations, DNA methylation at CpG dinucleotides may have occurred spontaneously and is subsequently subject to context-dependent selective constraints. Although our study was not aimed for the resolution of the abovementioned dispute, it demonstrates that the biological roles of DNA methylation can vary significantly within a short genomic distance. The regulations and functions of DNA methylation thus appear to be more complicated than previously anticipated.

Materials and Methods

Collection of Single-Base-Resolution DNA Methylation Data

The base-resolution DNA methylation data from six human cell types (S1–S6, table 1) were downloaded from NGSmethDB (Hackenberg et al. 2011) at http://bioinfo5.ugr.es/NGSmethDB2/index.php (last accessed November 8, 2013). These data sets were generated with bisulfite (S1, S4, S5, and S6) or MethylC (S2 and S3) sequencing. To ensure the accuracy of the data, we considered only the CpG dinucleotides that were covered by $\geq$5 bisulfite/MethylC reads (such CpG dinucleotides are designated as "sampled CpGs"). A CpG site is defined as methylated (designated as "mCG") if $\geq$80% of the mapped reads support the methylation status at the CpG site, whereas a CpG site supported by <20% of methylation read is defined as unmethylated (Meissner et al. 2008; Laurent et al. 2010). To ensure that the examined exonic regions contain sufficient information for estimations of the methylation level, only the exonic regions that contained $\geq$10 sampled CpGs were considered.

Data Retrieval

The human gene annotations were downloaded from the Ensembl database at http://www.ensembl.org/ (last accessed November 8, 2013) (version 69). According to the type of genomic sequence (i.e., UTR or CDS) and the relative position of the exon (first or nonfirst exons) in the Ensembl-annotated genes, each retrieved exonic region was assigned to one of the six following groups: 1) 5′UTR-First (5U-F); 2) 5′UTR-nonFirst (5U-nonF); 3) CDS-First (CDS-F); 4) CDS-nonFirst (CDS-nonF); 5) 3′UTR-First (3U-F); and 6) 3′UTR-nonFirst (3U-nonF) (examples are given in fig. 1A). To avoid ambiguous classification, single-exon genes and the exonic regions that overlap with noncoding RNAs or pseudogenes were excluded. The exonic regions that are differentially annotated as CDSs and UTRs in different alternatively spliced transcripts were also excluded. The numbers of exonic regions, methylated sites, and unmethylated sites for each group are listed in table 1 and supplementary table S2, Supplementary Material online. The conservation levels of single nucleotides and exonic regions were measured by the PhyloP score (Pertea et al. 2011) and phastCons score (Siepel et al. 2005), respectively. Both the PhyloP and phastCons scores were downloaded from the UCSC genome browser at http://genome.ucsc.edu/ (last accessed November 8, 2013). The conservation level of an exonic region was measured by calculating the average phastCons scores of all nucleotides that are located within this exonic region. An exonic region is regarded as lowly conserved if the average phastCons score of this region is lower than the median of all regions in the corresponding exonic group. Otherwise, the exonic region is regarded as highly conserved. Degeneracy of coding nucleotides was determined on the basis of the Ensembl gene annotations (version 69). The nucleotides with ambiguous degeneracy (e.g., caused by overlapping genes or alternative mRNA splicing) were excluded. The human–mouse gene orthology assignments and d_N/d_S ratios were downloaded from the Ensembl database.

Measurement of CpG Dinucleotide Depletion ($CpG_{O/E}$)

The ratio of observed-to-expected number of CpG dinucleotides ($CpG_{O/E}$) is a measurement of CpG dinucleotide depletion because a low $CpG_{O/E}$ indicates a large fraction of CpG dinucleotides having being mutated (Bird and Taggart 1980; Park et al. 2011; Park et al. 2012). $CpG_{O/E}$ is defined as

$$CpG_{O/E} = \frac{P_{CpG}}{P_C \times P_G} = \frac{\text{number of CpGs} \times \text{length of the exonic region}}{\text{number of Cs} \times \text{number of Gs}},$$

where P_{CpG}, P_C, and P_G, respectively, represents the frequency of CpG dinucleotides, C nucleotides, and G nucleotides in the examined exonic region.

Measurement of the Methylation Level of an Exonic Region

The methylation level of an exonic region was measured by calculating the density of mCG per 100 CpG dinucleotides (designated as mCG density), which is defined as number of mCGs × 100/number of all CpGs sampled.

Measurement of the Odds Ratio

To evaluate possible fitness outcomes of CpG methylation, we simulated C-to-T mutations at zero-fold degenerate cytosines that belong to a CpG dinucleotide. Whether the resulting

amino acid substitutions are damaging was then assessed by using SIFT (Ng and Henikoff 2003) and PolyPhen-2 (Adzhubei et al. 2010). The SIFT and PolyPhen-2 scores were queried through the Galaxy platform at https://main.g2.bx.psu.edu/ (last accessed November 8, 2013) and the PolyPhen sever (version 2.2.2) at http://genetics.bwh.harvard.edu/pph2/ (last accessed November 8, 2013), respectively. For the PolyPhen-2 prediction, "possibly" and "probably" damaging mutations were both considered as "damaging substitutions" in this study. We then calculated the odds ratio (OR) of damaging mutations occurring at methylated sites over those occurring at unmethylated sites. To this end, a two-way contingency table was generated, with rows containing the numbers of damaging and nondamaging mutations and columns containing the numbers of methylated and unmethylated sites. Two-tailed Fisher's exact test was then conducted to evaluate the statistical significance. If the OR is significantly larger than one, the amino acid mutations that occur at methylated sites are deemed to have a higher probability of causing damaging effects than those that occur at unmethylated sites.

Supplementary Material

Supplementary figs. S1–S3 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Dr Ben-Yang Liao for constructive comments. They also thank Dr Chia-Ying Chen and Ms. I-Ching Wu for their assistance in statistical tests and programming, respectively. This work was supported by the Genomics Research Center of Academia Sinica (to T.J.C.), the intramural funding of the National Health Research Institutes (to F.C.C.), and the National Science Council of Taiwan under contract numbers NSC 102-2621-B-001-003 (to T.J.C.) and NSC 102-2311-B-400-003 (to F.C.C.).

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods.* 7:248–249.

Ananda G, Chiaromonte F, Makova KD. 2011. A genome-wide view of mutation rate co-variation using multivariate analyses. *Genome Biol.* 12:R27.

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.* 30:267–275.

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

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

Bird AP, Taggart MH. 1980. Variable patterns of total DNA and rDNA methylation in animals. *Nucleic Acids Res.* 8:1485–1497.

Choi JK, Bae JB, Lyu J, Kim TY, Kim YJ. 2009. Nucleosome deposition and DNA methylation at coding region boundaries. *Genome Biol.* 10:R89.

Chuang TJ, Chen FC, Chen YZ. 2012. Position-dependent correlations between DNA methylation and the evolutionary rates of mammalian coding exons. *Proc Natl Acad Sci U S A.* 109:15841–15846.

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W. 1978. Molecular basis of base substitution hotspots in Escherichia coli. *Nature* 274:775–780.

Ellegren H, Smith NG, Webster MT. 2003. Mutation rate variation in the mammalian genome. *Curr Opin Genet Dev.* 13:562–568.

Feinberg AP, Tycko B. 2004. The history of cancer epigenetics. *Nat Rev Cancer.* 4:143–153.

Feng S, Cokus SJ, Zhang X, et al. (15 co-authors). 2010. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A.* 107:8689–8694.

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

Gelfman S, Cohen N, Yearim A, Ast G. 2013. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. *Genome Res.* 23:789–799.

Graur D, Li W-H. 2000. Fundamentals of molecular evolution. Sunderland (MA): Sinauer Associates.

Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol.* 5:578–590.

Hackenberg M, Barturen G, Oliver JL. 2011. NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data. *Nucleic Acids Res.* 39:D75–D79.

Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM. 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* 405:486–489.

Heard E, Clerc P, Avner P. 1997. X-chromosome inactivation in mammals. *Annu Rev Genet.* 31:571–610.

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet.* 12:756–766.

Holliday R, Grigg GW. 1993. DNA methylation and mutation. *Mutat Res.* 285:61–67.

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.

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

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge; New York: Cambridge University Press.

Laurent L, Wong E, Li G, et al. (11 co-authors). 2010. Dynamic changes in the human methylome during differentiation. *Genome Res.* 20: 320–331.

Li E, Beard C, Jaenisch R. 1993. Role for DNA methylation in genomic imprinting. *Nature* 366:362–365.

Li Y, Zhu J, Tian G, et al. (37 co-authors). 2010. The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol.* 8:e1000533.

Lister R, Pelizzola M, Dowen RH, et al. (18 co-authors). 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462:315–322.

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

Mancini DN, Singh SM, Archer TK, Rodenhiser DI. 1999. Site-specific DNA methylation in the neurofibromatosis (NF1) promoter interferes with binding of CREB and SP1 transcription factors. *Oncogene* 18:4108–4119.

Meissner A, Mikkelsen TS, Gu H, et al. (13 co-authors). 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454:766–770.

Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. *Annu Rev Genomics Hum Genet.* 11: 265–289.

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31:3812–3814.

Park J, Peng Z, Zeng J, Elango N, Park T, Wheeler D, Werren JH, Yi SV. 2011. Comparative analyses of DNA methylation and sequence evolution using Nasonia genomes. *Mol Biol Evol.* 28:3345–3354.

Park J, Xu K, Park T, Yi SV. 2012. What are the determinants of gene expression levels and breadths in the human genome? *Hum Mol Genet.* 21:46–56.

Pertea M, Pertea GM, Salzberg SL. 2011. Detection of lineage-specific evolutionary changes among primate species. *BMC Bioinformatics* 12:274.

Rangani G, Khodakovskaya M, Alimohammadi M, Hoecker U, Srivastava V. 2012. Site-specific methylation in gene coding region underlies transcriptional silencing of the Phytochrome A epiallele in *Arabidopsis thaliana*. *Plant Mol Biol.* 79:191–202.

Reik W. 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 447:425–432.

Sati S, Tanwar VS, Kumar KA, et al. (15 co-authors). 2012. High resolution methylome map of rat indicates role of intragenic DNA methylation in identification of coding region. *PLoS One* 7: e31621.

Schwartz S, Meshorer E, Ast G. 2009. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol.* 16:990–995.

Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S. 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* 479:74–79.

Siepel A, Bejerano G, Pedersen JS, et al. (16 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034–1050.

Walsh CP, Chaillet JR, Bestor TH. 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet.* 20: 116–117.

Wang J, Wang W, Li R, et al. (77 co-authors). 2008. The diploid genome sequence of an Asian individual. *Nature* 456:60–65.

Zechner U, Nolte J, Wolf M, Shirneshan K, Hajj NE, Weise D, Kaltwasser B, Zovoilis A, Haaf T, Engel W. 2009. Comparative methylation profiles and telomerase biology of mouse multipotent adult germline stem cells and embryonic stem cells. *Mol Hum Reprod.* 15:345–353.

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