BMC Genomics

RESEARCH ARTICLE Open Access

Effects of cytosine methylation on transcription factor binding sites

Yulia A Medvedeva[1], Abdullah M Khamis[1], Ivan V Kulakovskiy[2,3], Wail Ba-Alawi[1], Md Shariful I Bhuyan[1], Hideya Kawaji[4,5,6], Timo Lassmann[4,5], Matthias Harbers[4], Alistair RR Forrest[4,5], Vladimir B Bajic[1*] and The FANTOM consortium

Abstract

Background: DNA methylation in promoters is closely linked to downstream gene repression. However, whether DNA methylation is a cause or a consequence of gene repression remains an open question. If it is a cause, then DNA methylation may affect the affinity of transcription factors (TFs) for their binding sites (TFBSs). If it is a consequence, then gene repression caused by chromatin modification may be stabilized by DNA methylation. Until now, these two possibilities have been supported only by non-systematic evidence and they have not been tested on a wide range of TFs. An average promoter methylation is usually used in studies, whereas recent results suggested that methylation of individual cytosines can also be important.

Results: We found that the methylation profiles of 16.6% of cytosines and the expression profiles of neighboring transcriptional start sites (TSSs) were significantly negatively correlated. We called the CpGs corresponding to such cytosines "traffic lights". We observed a strong selection against CpG "traffic lights" within TFBSs. The negative selection was stronger for transcriptional repressors as compared with transcriptional activators or multifunctional TFs as well as for core TFBS positions as compared with flanking TFBS positions.

Conclusions: Our results indicate that direct and selective methylation of certain TFBS that prevents TF binding is restricted to special cases and cannot be considered as a general regulatory mechanism of transcription.

Keywords: DNA methylation, Transcription factor binding sites, Transcriptional regulation, CAGE, RRBS, CpG "traffic lights", Bioinformatics, Computational biology

Background

DNA methylation is one of the most studied epigenetic modifications. In differentiated cells in higher animals, methylated cytosine is almost always followed by guanine, associating methylation of 60-90% of all cytosines in a CpG context [1,2]. Although recent evidence showed that cytosine methylation in embryonic stem cells may also occur as CpHpG and CpHpH (where H corresponds to A, C, or T) [3-5], genome-wide distributions of cytosine methylation in CpHpG and especially in CpHpH have great variability between individuals, contrary to methylation in the CpG context, which demonstrates stable cell-type-specific methylation [4]. Thus, cell-type-specific regulatory patterns most likely depend on methylation in the CpG context.

Various methodologies have been developed to study DNA methylation at different genomic scales (for a review, see, for example, [6-8]) with direct sequencing of bisulfite-converted DNA [9] continuing to be the method of choice. However, the analysis of a single CpG site or a few CpG sites as surrogate indicators of DNA methylation status of the surrounding region is the most prevalent strategy in epigenetic studies at different scales, due to the assumption of the relatively homogeneous distribution of DNA methylation within genomic regions. This assumption is supported by multiple pieces of evidence of unmethylated CpGs closely co-located within CpG islands (CGIs) and methylated CpGs in repetitive elements. In addition, the level of methylation of the *HpaII* sites (CCGG) within CGIs demonstrates a correlation with

* Correspondence: vladimir.bajic@kaust.edu.sa
[1]Computational Bioscience Research Center, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
Full list of author information is available at the end of the article

average CGI methylation levels [10]. At the same time, methylated CpGs have been found in unmethylated CGIs [4]. It was also shown that a single differentially methylated CpG might affect transcription of the ESR1 gene [11]. Moreover, it was hypothesized that DNA methylation of CpG-rich and CpG-poor regions might be involved in different regulatory programs [12]. In short, whether or not the distinct methylation status of a single CpG affects specific transcription-related functions remains an open question.

It is widely accepted that cytosine methylation is a crucial regulatory mechanism in both normal and pathological processes. DNA methylation is involved in development [13,14], cellular differentiation [15], maintaining cellular identity [16], pluripotency [17], aging [18,19], memory formation [20], responses to environmental changes [21,22] and reactions to diet [23]. Several pathological conditions, including cancer [22,24], diabetes [25], Alzheimer's and Parkinson's diseases [26], also show aberrant DNA methylation. Profiles of DNA methylation can be inherited through cell division [16] and in some cases through generations [21]. However, recent studies of dynamic DNA methylation/de-methylation *in vivo* [27,28] challenge the conventional view that DNA methylation is a permanent epigenetic mark and suggest the possibility of exploring DNA methylation as a promising target for non-invasive therapies for diseases linked with aberrant methylation.

DNA methylation of gene promoters is tightly associated to the repression of transcription, yet the mechanisms are still unclear [29]. In the last four decades, multiple studies have shown that the level of DNA methylation in promoters is negatively correlated with the expression of downstream genes [30-35]. It was also hypothesized that ubiquitous, low-density cytosine methylation in vertebrate genomes can contribute to reduction of the transcriptional "noise" from inappropriate promoters [36]. Recently, multiple pieces of evidence arguing against the paradigm that DNA methylation always represses transcription have started to appear. Transcription of some genes was found to be independent of methylation [37]. Promoters with low CpG content are usually methylated, yet they still may be transcriptionally active [38,39]. Although intergenic and gene terminal CGIs are frequently methylated, they demonstrate a pervasive transcription [40]. Sparse DNA methylation of promoters may repress transcription, but this effect could be overcome by an enhancer [41]. Genes exhibiting high levels of promoter methylation during normal development remain suppressed in Dnmt1-deficient mouse embryos, suggesting that developmental gene control does not globally rely on cytosine methylation and that the effects of DNA methylation are limited to specialized processes such as imprinting and mobile elements repression [29]. Alternative promoter usage in different regions of the aged brain seems to be independent of promoter methylation [42]. Promoter sequences are able to recapitulate correct DNA methylation autonomously and demonstrate proper *de novo* methylation during differentiation in pluripotent cells independently of the transcriptional activity of corresponding downstream promoters [43]. Furthermore, in some cases, methylation is required for activation of transcription and therefore is positively correlated with gene expression [44].

Despite the various controversies, evidence that DNA methylation as an important step in regulation remains solid. The mechanisms of the interplay between methylation and expression are therefore critically important. It remains unclear whether DNA methylation is the cause or the consequence of altered gene expression. If DNA methylation causes gene repression, then there are several possible outcomes (Figure 1a). Cytosine methylation may directly affect the affinity of transcription factors (TFs) towards their binding sites (TFBSs) [45]. Non-systematic experimental evidence that DNA methylation can prevent binding of some TFs to particular TFBSs [45,46] supports this hypothesis. For example, methylation of the E-box (CACGTG) prevents n-Myc from binding to promoters of EGFR and CASP8 in a cell-specific manner [47]; methylation of the YY1-binding site in the promoter of the Peg3 gene represses the binding activity of YY1 *in vitro* [48]. It is also worth noting that experimentally determined TFBSs usually show low levels of DNA methylation [4,49,50] and that TF-TFBS recognition is often associated with the lack of methylation [51,52]. Furthermore, certain positions within CTCF binding sites are more sensitive to methylation than are others [53]. Methylated cytosine can also attract TFs, both activators [44,54] and repressors [55]. Methylation of the CRE sequence enhances the DNA binding of C/EBPα, which in turn activates a set of promoters specific for adipocyte differentiation [44,54]. Methyl-binding domain (MBD) proteins bind methylated CpG dinucleotide and induce histone deacetylation, subsequent chromatin condensation and gene repression [55].

The opposite scenario implies that chromatin modifications [56-58] reduce the accessibility of TFs and the transcriptional machinery to gene promoters, therefore leading to gene repression. DNA methylation in this model is not a cause, but a consequence of repression and serves to fix the repressed state of the chromatin (Figure 1b). In this case, cytosine methylation accumulates passively as a consequence of the independent absence of TF binding [50,53] or it appears as a result of direct DNA methyltransferase recruitment by transcription repression proteins such as the Polycomb group (PcG) protein EZH2 [59]. This model is supported by negative correlation of TF expression and average methylation

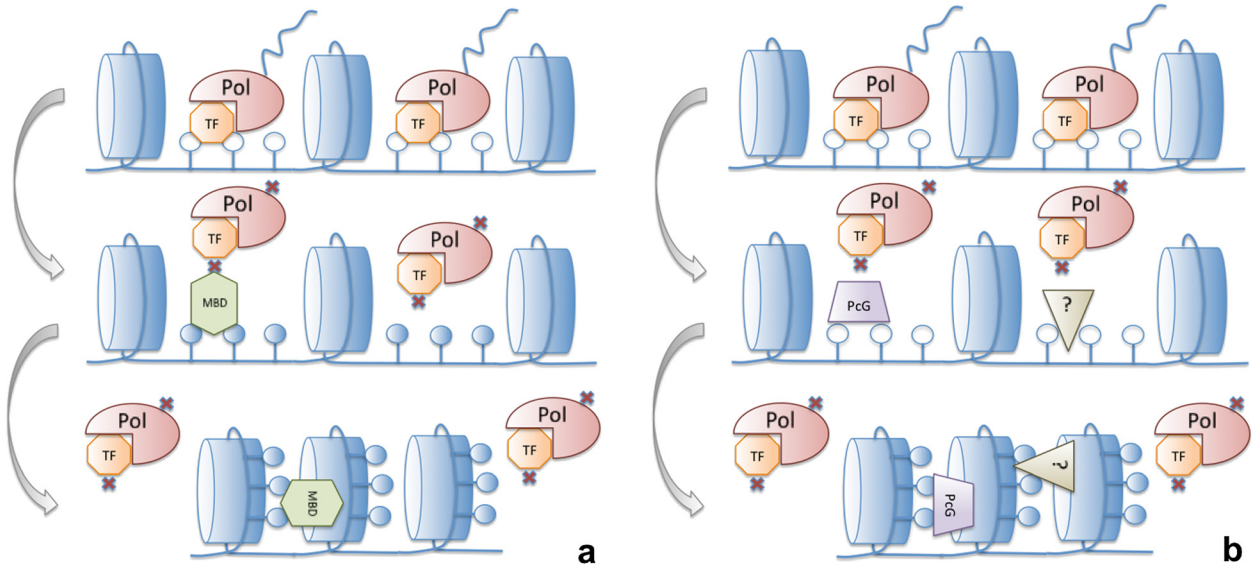

Figure 1 Schematic representation of the interaction between promoter methylation and transcription of the gene. In the absence of DNA methylation, TFs can bind DNA allowing RNA polymerase to bind and to start the transcription. Panel **a** shows the following scenario: if DNA becomes methylated, TFs are blocked from binding to DNA and therefore RNA polymerase is unable to bind and to initiate transcription. Panel **b** shows the following scenario: chromatin modifications reduce the ability of TFs to bind DNA and therefore RNA polymerase is unable to bind; the repressed condition of the chromatin is maintained by subsequent DNA methylation. PolII is shown as a maroon pie; nucleosome is shown as a blue cylinder. Plain (solid) lollipops represent unmethylated (methylated) cytosines. TF is shown as an orange octagon. The green hexagon and purple trapezoid are a methyl-binding domain and Policomb-group proteins, respectively. The brown triangle represents an unknown repressor.

of their TFBSs [50]. Besides, it was reported that binding of some TFs, including Sp1 and CTCF, is sufficient for maintaining a local unmethylated state [60-65]. Nevertheless, this scenario (Figure 1b) does not explain the sensitivity of certain TFs to methylation of their TFBSs.

In this study, we explore the evidence that supports one of these two scenarios. To achieve this, we first test whether methylation of a particular cytosine correlates with transcription. This effect may provide a basis for regulation of transcription through methylation of specific TFBSs. Second, we investigate whether some TFs are more sensitive than others to the presence of such cytosines in their TFBSs and what features of TFBSs can be associated with this sensitivity. To this end, we employed ENCODE [66] data on DNA methylation obtained by reduced representation bisulfite sequencing (RRBS) [67]. RRBS allows us to identify both methylated and unmethylated cytosines quantitatively at a single base pair resolution in the CCGG context in regions with high densities of rarely methylated cytosines, usually co-located within gene promoters [68]. To evaluate genome-wide expression across different cell types, we used FANTOM5 [69] data obtained by cap analysis of gene expression (CAGE) [70]. FANTOM5 provides quantitative estimation of expression in several hundreds of different cell types.

Our study shows that a fraction of single CpGs within promoters exhibits a significant negative correlation of their methylation profiles with the expression profiles of neighboring transcriptional start sites (TSSs) considered across various samples. Moreover, we observe a strong negative selection against the presence of such cytosines within TFBSs, especially in their core positions. Interestingly, we find that repressors are more sensitive to the presence of such cytosines in their binding sites.

This work is part of the FANTOM5 project. Data downloads, genomic tools and co-published manuscripts are collected at http://fantom.gsc.riken.jp/5/.

Results and discussion

Only a fraction of cytosines exhibits significant correlation between methylation and expression profiles of a corresponding TSS

It is well known that the level of cytosine methylation of promoters is negatively correlated with gene expression [71]; the role of methylation of particular CpGs in the regulation of gene expression has been demonstrated in the case of ESR1 [11]. The crucial role of the location of methylated regions relative to TSSs is also widely accepted. The question whether methylation of a particular cytosine may affect expression remains unanswered.

As the first step of this study, we studied whether the methylation level of a particular cytosine within a promoter region is correlated with the expression of the corresponding TSS, since such cytosines may serve as a basis for the regulation of transcription through TF binding. Table 1 demonstrates that among 237,244 cytosines analyzed in the study, only 16.6% (0.8%) have significantly (P-value ≤ 0.01) negative or positive Spearman Correlation Coefficients ($SCC_{M/E}$) between methylation and expression profiles of a closely located TSS (see Methods). This sheds different light on the common perception of a link between methylation and gene expression. We call cytosines demonstrating significantly negative $SCC_{M/E}$ CpG "traffic lights" (see Methods). In this study, we mostly focus on such cytosines.

Out of 50 cell types analyzed in this study, 14 were malignant. Genome-wide DNA methylation in cancer cells is dramatically different from that in normal cells (for the review see, for example [72-75]). Although we believe that the basic mechanism of interaction between DNA methylation and expression should be the same in cancer and non-cancer cells, we repeated the experiments on the 36 normal cell types and obtained similar results (Additional file 1): only a small fraction (9.5% and 1.5%) of cytosines have significant (P-value ≤ 0.01) negative and positive $SCC_{M/E}$, respectively.

CAGE tags are often found within gene bodies [76] and methylation of a gene body may have a positive correlation with gene expression [77-79]. It was also suggested that the cytosines within gene bodies are often not methylated (5mC) but hydroxymethylated (5hmC) [80]. However, bisulfite-based methods of cytosine modification detection (including RRBS) are unable to distinguish these two types of modifications [81]. The presence of 5hmC in a gene body may be the reason why a fraction of CpG dinucleotides has a significant positive $SCC_{M/E}$ value. Unfortunately, data on genome-wide distribution of 5hmC in humans is available for a very limited set of cell types, mostly developmental [82,83], preventing us from a direct study of the effects of 5hmC on transcription and TFBSs. At the current stage the 5hmC data is not available for inclusion in the manuscript. Yet, we were able to perform an indirect study based on the localization of the studied cytosines in various genomic regions. We tested whether cytosines demonstrating various $SCC_{M/E}$ are co-located within different gene regions (Table 2). Indeed, CpG "traffic lights" are located within promoters of GENCODE [84] annotated genes in 79% of the cases, and within gene bodies in 51% of the cases, while cytosines with positive $SCC_{M/E}$ are located within promoters in 56% of the cases and within gene bodies in 61% of the cases. Interestingly, 80% of CpG "traffic lights" are located within CGIs, while this fraction is smaller (67%) for cytosines with positive $SCC_{M/E}$. This observation allows us to speculate that CpG "traffic lights" are more likely methylated, while cytosines demonstrating positive $SCC_{M/E}$ may be subject to both methylation and hydroxymethylation. Cytosines with positive and negative $SCC_{M/E}$ may therefore contribute to different mechanisms of epigenetic regulation. It is also worth noting that cytosines with insignificant (P-value > 0.01) $SCC_{M/E}$ are more often located within the repetitive elements and less often within the conserved regions and that they are more often polymorphic as compared with cytosines with a significant $SCC_{M/E}$, suggesting that there is natural selection protecting CpGs with a significant $SCC_{M/E}$.

Selection against TF binding sites overlapping with CpG "traffic lights"

We hypothesize that if CpG "traffic lights" are not induced by the average methylation of a silent promoter, they may affect TF binding sites (TFBSs) and therefore may regulate transcription. It was shown previously that cytosine methylation might change the spatial structure of DNA and thus might affect transcriptional regulation by changes in the affinity of TFs binding to DNA [47-49]. However, the answer to the question of if such a mechanism is widespread in the regulation of transcription remains unclear. For TFBSs prediction we used the remote dependency model (RDM) [85], a generalized version of a position weight matrix (PWM), which eliminates an assumption on the positional independence of nucleotides and takes into account possible correlations of nucleotides at remote positions within TFBSs. RDM was shown to decrease false positive rates effectively as compared with the widely used PWM model.

Our results demonstrate (Additional file 2) that from the 271 TFs studied here (having at least one CpG "traffic light" within TFBSs predicted by RDM), 100 TFs had a significant underrepresentation of CpG "traffic lights" within their predicted TFBSs (P-value < 0.05, Chi-square test, Bonferoni correction) and only one TF (OTX2) had

Table 1 Total numbers of CpGs with different $SCC_{M/E}$ between methylation and expression profiles

$SCC_{M/E}$ sign	$SCC_{M/E}$, P-value ≤ 0.05	$SCC_{M/E}$, P-value ≤ 0.01	$SCC_{M/E}$, P-value ≤ 0.001	$SCC_{M/E}$, P-value ≤ 0.05, fraction	$SCC_{M/E}$, P-value ≤ 0.01, fraction	$SCC_{M/E}$, P-value ≤ 0.001, fraction
Negative	73328	39414	17031	0.309	0.166	0.072
Positive	5750	1832	479	0.024	0.008	0.002

The total number of CpGs in the study is 237,244.

Table 2 Fraction of cytosines demonstrating different $SCC_{M/E}$ within genome regions

	CGI	Gene promoters	Gene bodies	Repetitive elements	Conserved regions	SNP	DNase sensitivity regions
CpG "traffic lights"	0.801	0.793	0.507	0.095	0.203	0.008	0.926
$SCC_{M/E} > 0$	0.674	0.556	0.606	0.095	0.210	0.009	0.829
$SCC_{M/E}$ insignificant	0.794	0.733	0.477	0.128	0.198	0.010	0.897

a significant overrepresentation of CpG "traffic lights" within the predicted TFBSs. Similar results were obtained using only the 36 normal cell lines: 35 TFs had a significant underrepresentation of CpG "traffic lights" within their predicted TFBSs (*P*-value < 0.05, Chi-square test, Bonferoni correction) and no TFs had a significant overrepresentation of such positions within TFBSs (Additional file 3). Figure 2 shows the distribution of the observed-to-expected ratio of TFBS overlapping with CpG "traffic lights". It is worth noting that the distribution is clearly bimodal with one mode around 0.45 (corresponding to TFs with more than double underrepresentation of CpG "traffic lights" in their binding sites) and another mode around 0.7 (corresponding to TFs with only 30% underrepresentation of CpG "traffic lights" in their binding sites). We speculate that for the first group of TFBSs, overlapping with CpG "traffic lights" is much more disruptive than for the second one, although the mechanism behind this division is not clear.

To ensure that the results were not caused by a novel method of TFBS prediction (i.e., due to the use of RDM), we performed the same analysis using the standard PWM approach. The results presented in Figure 2 and in Additional file 4 show that although the PWM-based method generated many more TFBS predictions as compared to RDM, the CpG "traffic lights" were significantly underrepresented in the TFBSs in 270 out of 279 TFs studied here (having at least one CpG "traffic light" within TFBSs as predicted by PWM), supporting our major finding.

We also analyzed if cytosines with significant positive $SCC_{M/E}$ demonstrated similar underrepresentation within TFBS. Indeed, among the tested TFs, almost all were depleted of such cytosines (Additional file 2), but only 17 of them were significantly over-represented due to the overall low number of cytosines with significant positive $SCC_{M/E}$. Results obtained using only the 36 normal cell lines were similar: 11 TFs were significantly depleted of such cytosines (Additional file 3), while most of the others were also depleted, yet insignificantly due to the low number of total predictions. Analysis based on PWM models (Additional file 4) showed significant underrepresentation of such

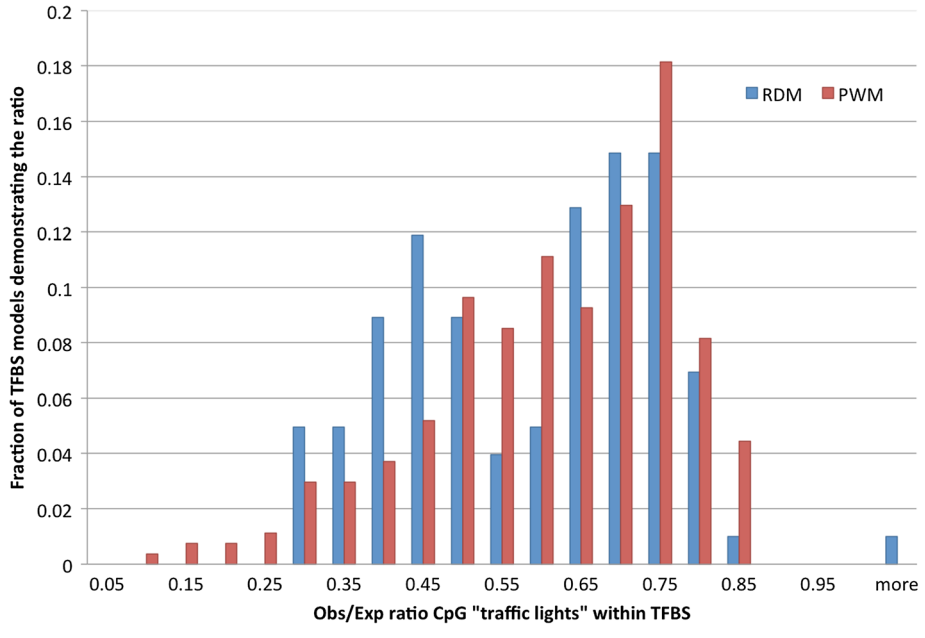

Figure 2 Distribution of the observed number of CpG "traffic lights" to their expected number overlapping with TFBSs of various TFs. The expected number was calculated based on the overall fraction of significant (*P*-value < 0.01) CpG "traffic lights" among all cytosines analyzed in the experiment.

cytosines for 229 TFs and overrepresentation for 7 (DLX3, GATA6, NR1I2, OTX2, SOX2, SOX5, SOX17). Interestingly, these 7 TFs all have highly AT-rich binding sites with very low probability of CpG.

It was previously shown that cytosine methylation can prevent binding of several TFs (such as Sp1 [60], CTCF [53] and others) and, therefore, methylation may serve as a global regulatory mechanism for cell-specific TF binding. Yet, we observe that most of TFs avoid CpG "traffic lights" in their binding sites, suggesting a potentially damaging effect of CpG "traffic lights" to TFBS and therefore a natural selection against TFBS overlapping with CpG "traffic lights".

Computational prediction of TFBSs identifies DNA regions of potential binding, which may not be available for a TF in a particular cell type due to chromatin modifications. To avoid a bias caused by potential TFBSs that are not functional in particular cell types, we used experimentally obtained regions of TF binding. Chromatin immunoprecipitation followed by parallel DNA sequencing (ChIP-seq) is an effective experimental technique for the identification of regions for DNA-protein interaction [86]. Yet, regions where TFs most likely bind DNA (ChIP-seq peaks) in a particular cell type are relatively long, usually longer than several hundreds of base pairs, while real TFBSs are on average a dozen base pairs long. Therefore, we combined experimental and computational approaches and filtered out the predictions of TFBSs outside of ChiP-seq peak regions. We tested our results on ChIP-seq data for CTCF as it is the only TF in ENCODE with experimental binding information in as many as 22 cell types out of the 50 cell types we used in our study (14 of the 22 were normal cell types). Results in Additional file 5 support our initial finding: CTCF binding sites avoid CpG "traffic lights". ChIP-seq data for other TFs are available only for the cancer cell lines included in our study, making it impossible to draw conclusions about normal cell functioning. At the current stage the ChiP-seq data for other TFs is not available for inclusion in the manuscript. Our findings suggest that changing a TF's affinity to DNA or even blocking TF binding sites by direct and selective methylation is limited to certain TFBSs within a few promoters and thus is not likely to be a general mechanism of methylation-dependent regulation of gene expression.

TFBSs of repressors are especially sensitive to the presence of CpG "traffic lights"

Overlapping of TFBS with CpG "traffic lights" may affect TF binding in various ways depending on the functions of TFs in the regulation of transcription. There are four possible simple scenarios, as described in Table 3. However, it is worth noting that many TFs can work both as activators and repressors depending on their cofactors. Moreover, some TFs can bind both methylated and unmethylated DNA [87]. Such TFs are expected to be less sensitive to the presence of CpG "traffic lights" than are those with a single function and clear preferences for methylated or unmethylated DNA.

Table 3 Expected sign of $SCC_{M/E}$ depending on TF binding preferences and function

	TF binding preferences		
TF function	Unmethylated DNA	Methylated DNA	Both
Activator	(1) negative $SCC_{M/E}$	(2) positive $SCC_{M/E}$	insignificant $SCC_{M/E}$
Repressor	(3) positive $SCC_{M/E}$	(4) negative $SCC_{M/E}$	insignificant $SCC_{M/E}$
Both	insignificant $SCC_{M/E}$	insignificant $SCC_{M/E}$	

There are four possible scenarios of interaction of DNA methylation and TF functions:
(1) TF can bind unmethylated DNA and cannot bind methylated DNA. TF acts as a transcription activator. The methylation profile of cytosines within TFBS should be negatively correlated with TSS expression.
(2) TF can bind methylated DNA and cannot bind unmethylated DNA. TF acts as a transcription activator. The methylation profile of cytosines within TFBS should be positively correlated with TSS expression.
(3) TF can bind unmethylated DNA and cannot bind methylated DNA. TF acts as a transcription repressor. The methylation profile of cytosines within TFBS should be positively correlated with TSS expression.
(4) TF can bind methylated DNA and cannot bind unmethylated DNA. TF acts as transcription repressor. The methylation profile of cytosines within TFBS should be negatively correlated with TSS expression.

Using information about molecular function of TFs from UniProt [88] (Additional files 2, 3, 4 and 5), we compared the observed-to-expected ratio of TFBS overlapping with CpG "traffic lights" for different classes of TFs. Figure 3 shows the distribution of the ratios for activators, repressors and multifunctional TFs (able to function as both activators and repressors). The figure shows that repressors are more sensitive (average observed-to-expected ratio is 0.5) to the presence of CpG "traffic lights" as compared with the other two classes of TFs (average observed-to-expected ratio for activators and multifunctional TFs is 0.6; t-test, P-value < 0.05), suggesting a higher disruptive effect of CpG "traffic lights" on the TFBSs of repressors. Although results based on the RDM method of TFBS prediction show similar distributions (Additional file 6), the differences between them are not significant due to a much lower number of TFBSs predicted by this method. Multifunctional TFs exhibit a bimodal distribution with one mode similar to repressors (observed-to-expected ratio 0.5) and another mode similar to activators (observed-to-expected ratio 0.75). This suggests that some multifunctional TFs act more often as activators while others act more often as repressors. Taking into account that most of the known TFs prefer to bind unmethylated DNA, our results are in concordance with the theoretical scenarios presented in Table 3.

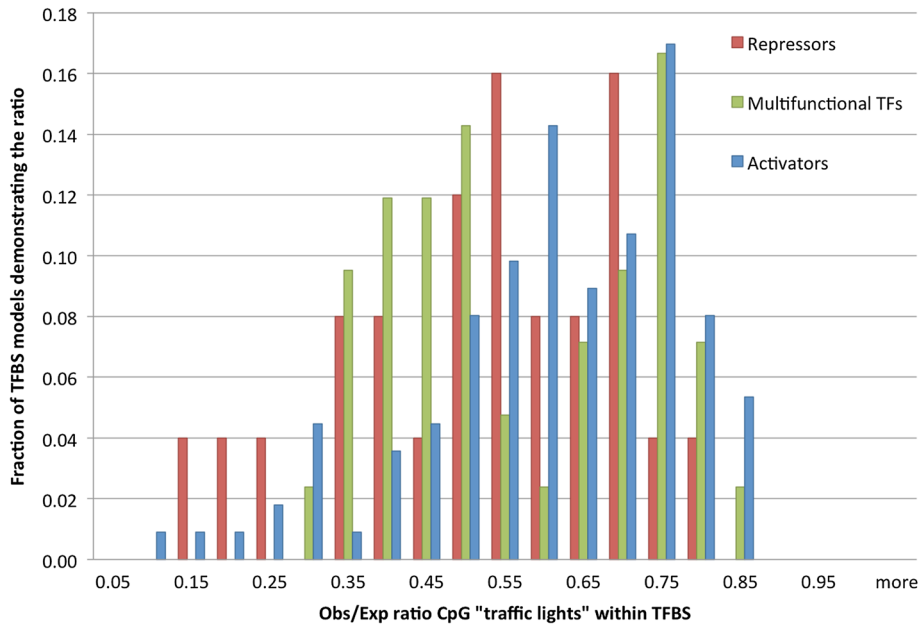

Figure 3 Distribution of the observed number of CpG "traffic lights" to their expected number overlapping with TFBSs of activators, repressors and multifunctional TFs. The expected number was calculated based on the overall fraction of significant (*P*-value < 0.01) CpG "traffic lights" among all cytosines analyzed in the experiment.

"Core" positions within TFBSs are especially sensitive to the presence of CpG "traffic lights"

We also evaluated if the information content of the positions within TFBS (measured for PWMs) affected the probability to find CpG "traffic lights" (Additional files 7 and 8). We observed that high information content in these positions ("core" TFBS positions, see Methods) decreases the probability to find CpG "traffic lights" in these positions supporting the hypothesis of the damaging effect of CpG "traffic lights" to TFBS (t-test, *P*-value < 0.05). The tendency holds independent of the chosen method of TFBS prediction (RDM or RWM). It is noteworthy that "core" positions of TFBS are also depleted of CpGs having positive $SCC_{M/E}$ as compared to "flanking" positions (low information content of a position within PWM, (see Methods), although the results are not significant due to the low number of such CpGs (Additional files 7 and 8).

Conclusions

We found that the methylation profiles and expression profiles in 16.6% of single CpG dinucleotides in CAGE-derived promoters were significantly negatively correlated with neighbouring TSS, supporting the argument that single cytosine methylation is involved in the regulation of transcription. In a way, the current common perception of the link between methylation and gene expression is seen in a different light. Unexpectedly, we observed a strong selection against the presence of CpG "traffic lights" within the TFBSs of many TFs. We demonstrated that the selection against CpG "traffic lights" within TFBS is even more pronounced in the case of "core" positions within TFBSs as compared to "flanking" positions. These observations allow us to suggest that blocking of TFBSs by selective methylation is unlikely to be a general mechanism of methylation-dependent transcription regulation and that such a mechanism is limited to special cases. We conclude that the regulation of expression via DNA methylation and via TF binding are relatively independent regulatory mechanisms; both mechanisms are thus not in a direct causal relationship. Known cases of interaction between these mechanisms appear mostly because they operate on the same target regions (promoters) and require intermediate partners, for example, modification of chromatin.

Methods

Cell types

We manually selected 137 FANTOM5 samples (cell types) matching 50 ENCODE samples. We grouped them into 50 classes of identical or similar biological cell types. To reduce the noise coming from inexact matching of cell types between FANTOM5 and ENCODE data, we averaged the expression/methylation values for different technical or biological replicas, donors and cell types within the same class. Detailed information is provided in Additional file 9.

All human samples used in the FANTOM5 project were either exempted material (available in public collections or commercially available), or provided under informed consent. All non-exempt material is covered

under RIKEN Yokohama Ethics applications (H17-34 and H21-14) and collected in compliance with the Helsinki Declaration.

TSSs and promoter regions

We used TSSs found by the CAGE method in FANTOM5. The relative log expression normalization method (RLE [89]) was applied to CAGE-tags in each sample [69]. For a particular TSS, we referred to a set of expression values across the selected 50 classes of cell types as an expression profile. Low expressed CAGE-tag clusters may be non-robust to sequencing errors or heterogeneity of the cell population. To reduce the effect of such CAGE-tag clusters, we excluded TSSs with all RLE-normalized expression values less than 1. For each CAGE-tag cluster, we selected a promoter region of 1500 bp upstream and 500 bp downstream of the ends of reported CAGE-tag clusters. Overlapping promoters were considered independently.

Cytosine methylation data

We used cytosine methylation data obtained by RRBS (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeHaibMethylRrbs). All data included cytosine methylation only in the CCGG context. We excluded cytosines covered by less than 10 reads. For a particular cytosine, we referred to a set of methylation values (the proportion of methylated reads relative to all reads) across the selected 50 cell types as a methylation profile. We excluded cytosines having methylation data for less than 50% of samples (25 when using all 50 cell types and 18 when using the 36 normal cell types) in the methylation profiles.

While each particular cytosine may be either methylated or unmethylated, the RRBS technique measures the average methylation of a particular cytosine in the cell population, which results in a 0 to 100% range of values. Although methylation values of most of the cytosines tend to be 0 or 100%, intermediate values are also possible. Low (but not 0) levels of cytosine methylation may appear as a result of experimental errors, and these levels can affect further analysis. To avoid any bias caused by such cytosines, we used only positions differentially methylated between cell types. We defined a CpG as differentially methylated if the amplitude (the difference between the maximum and minimum values in the normalized profile) of the methylation profile for a particular CpG was greater than 50%.

Correlation of cytosine methylation and TSS expression

For all cytosines located within promoter regions, we calculated the Spearman Correlation Coefficient between methylation profiles of the cytosine and the expression profiles of the corresponding TSS (referred to as $SCC_{M/E}$). We estimated the statistical significance of $SCC_{M/E}$ based on transformation to a Student's t-test distribution:

$$t = SCC_{M/E}\sqrt{\frac{n-2}{1-SCC_{M/E}^2}}$$

Here n is the length of the methylation/expression profile for a given position. In our analysis (if not stated otherwise), we referred to positions with P-values ($SCC_{M/E}$) ≤ 0.01 as positions with significantly negative or positive correlations between the methylation and the expression profiles. It is noteworthy that due to the overlapping of promoter regions for different TSSs, one cytosine may have several $SCC_{M/E}$. In the case of overlapping promoters, it is difficult to estimate which TSS is affected by the methylation of a particular cytosine. We therefore considered that a particular CpG affects transcription if it has at least one $SCC_{M/E}$ above (or below) the significance level (see Table 1).

CpG "traffic lights"

To avoid bias in estimating $SCC_{M/E}$ for low methylated cytosines caused by experimental errors, we introduced differentially methylated cytosines based on the difference between the highest and lowest value (amplitude) in the normalized methylated profile when it was greater than 50% of the maximum possible value. In the analysis of TFBSs affected by cytosine methylation, we considered only CpGs differentially methylated across cell types. We introduced the term CpG "traffic lights" to describe differentially methylated cytosines with significantly (P-values ($SCC_{M/E}$) ≤ 0.01) negative $SCC_{M/E}$.

We also looked for co-localization of CpG "traffic lights" and several genomic features (data downloaded from http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/): known gene promoters (1500 bp upstream of TSS and 500 bp downstream) and gene bodies (500 bp downstream TSS to the end of the gene) (wgEncodeGencodeBasicV140); CpG islands (cpgIslandExt); DNase sensitivity regions (wgEncodeRegDnaseClusteredV2); repetitive elements (rmsk); SNPs (snp137Common); and conserved elements (phastConsElements46wayPrimates).

Prediction of TFBSs using the remote dependency models

To create RDMs, we used binding site alignments from HOCOMOCO [90]. This collection of TFBS models was selected due to the low level of redundancy of TFBS models per single TF. Binding sites having scores less than PWM thresholds were excluded. PWM thresholds were selected according to the P-value < 0.0005 (i.e., when 5 of 10,000 random words had scores no less than the thresholds). P-values were computed by the MACRO-APE software (http://autosome.ru/macroape) [90] that implements the strategy presented in the work of Touzet and Varre

[91]. Due to the large number of parameters in RDM models as compared to PWM models provided in HOCOMOCO, the minimal number of sequences in the alignment was increased from 8 to 15. Filtered alignments of fewer than 15 binding sites were discarded, which reduced the initial set of 426 TFBS models available in HOCOMOCO to 280 TFBS models (Additional file 4, column 1).

Using the frequency of each dinucleotide with one nucleotide being at position i and the other at position j, where $i = 1, \ldots, L\text{-}1$, $j = i + 1, \ldots, L$, in the set of aligned binding sites, the dinucleotide frequency matrix with remote dependencies was constructed and normalized similar to PWM normalization in Bajic *et al.* [92]:

$$RDM_{a,i,j} = \frac{f_{a,i,j}}{\sum_{i=1}^{L-1}\sum_{j=i+1}^{L} \max_a\left(f_{a,i,j}\right)}$$

Here $f_{a,i,j}$ is the frequency of dinucleotide a formed of nucleotides at positions i and j, and L is the length of the aligned TFBSs. We predicted TFBSs using the RDM models across the whole promoter set.

Prediction of TFBSs using position weight matrices

To check if the TFBS prediction method affects the results, we also predicted TFBS using widely accepted PWM models. We took the same PWMs from HOCOMOCO as used for RDM construction. PWM thresholds were selected according to the *P*-value of 0.0005 (Additional file 10).

TFBSs potentially affected by DNA methylation

We selected all cytosines for which $SCC_{M/E}$ were available and checked whether they were located within predicted TFBSs. The total number of predicted TFBSs is available in Additional files 2, 3 and 4 (column D). It is noteworthy that average GC-content of the RDM hits was undistinguishable from that of the binding sites in the initial alignments.

"Core" and "flanking" CpG positions within TFBS

If we consider all genome-wide hits of any TFBS model, we may find that CpG dinucleotides can appear almost in every position of TFBSs. However, some positions within binding sites contain CpG dinucleotide more often than do others, so we repeated the analysis for each type of binding site position separately. For a particular TFBS model, we selected CpG positions in the HOCOMOCO alignments according to the information content of the corresponding PWM columns. Information content is defined as DIC (Discrete Information Content [93]) separately for different types of binding site positions. For a particular TFBS model, we selected CpG positions in the HOCOMOCO alignments according to the information content of the corresponding PWM columns:

$$DIC_j = \frac{1}{N}\left(\sum\nolimits_{a\in(A,C,G,T)} \log\left(x_{a,j}!\right) - \log(N!)\right),$$

Here $x_{a,j}$ are elements of the position count matrix (i.e., nucleotide counts), N is the total number of aligned TFBS sequences. In contrast to classic information content [94], DIC is based on raw counts (instead of per-column nucleotide probabilities, which can be inaccurate for a small set of aligned sequences). We define two empirical DIC thresholds [95] *Th* and *th* (introduced in [96]). *Th* corresponds to the DIC of the column having only 3 (of 4 possible) nucleotides that have the same frequency, *th* corresponds to the DIC of the column having two nucleotides with the same frequency, *f*, and the other two nucleotides each with the frequency *2f*.

The CpG positions have C and G as major nucleotides (with the highest frequency) in the neighbouring columns. High information content CpG ("core" TFBS positions) has both C and G columns with DIC greater than *Th*. The medium (or low) information content CpG ("flanking" TFBS positions) has both C- and G-column DIC between *Th* and *th* (or lower than *th*). The summary is presented in Additional files 4 and 5.

Additional files

Additional file 1: Contains the total number of analyzed CpGs as well as the count of CpG demonstrating $SCC_{M/E}$ above certain significance levels. These results were obtained using only the 36 normal cell types.

Additional file 2: Contains RDM-based predicted TFBSs based on 50 cell samples; tables containing information about the names of the TFBSs models used, their function in regulation (activator or repressor), the number of cytosines in our study (with any $SCC_{M/E}$) overlapping with TFBSs, the number of CpG "traffic lights" overlapping with TFBSs for each TF, the expected number of such overlaps and the statistical significance of the over-/underrepresentation of TFBS in CpG "traffic light" positions. Consistent information is given for positions with positive $SCC_{M/E}$.

Additional file 3: Contains RDM-based predicted TFBSs based on the 36 normal cell samples; tables containing information about the names of the TFBSs models used, their function in regulation (activator or repressor), the number of cytosines in our study (with any $SCC_{M/E}$) overlapping with TFBSs, the number of CpG "traffic lights" overlapping with TFBSs for each TF, the expected number of such overlaps and the statistical significance of the over-/underrepresentation of TFBS in CpG "traffic light" positions. Consistent information is given for positions with positive $SCC_{M/E}$.

Additional file 4: Contains PWM-based predicted TFBSs based on 50 cell samples; tables containing information about the names of the TFBSs models used, their function in regulation (activator or repressor), the number of cytosines in our study (with any $SCC_{M/E}$) overlapping with TFBSs, the number of CpG "traffic lights" overlapping with TFBSs for each TF, the expected number of such overlaps and the statistical significance of the over-/underrepresentation of TFBS in CpG "traffic light" positions. Consistent information is given for positions with positive $SCC_{M/E}$.

Additional file 5: Contains RDM-based predicted TFBS for CTCF supported with ChIP-seq peak data; tables containing information about the names of the TFBSs models used, their function in regulation (activator or repressor), the number of cytosines in our study (with any $SCC_{M/E}$) overlapping with TFBSs, the number of CpG "traffic lights"

overlapping with TFBSs for each TF, the expected number of such overlaps and the statistical significance of the over-/underrepresentation of TFBS in CpG "traffic light" positions. Consistent information is given for positions with positive $SCC_{M/E}$.

Additional file 6: Contains a figure showing the distribution of the observed to expected ratio of CpG "traffic lights" overlapping with TFBSs of activators, repressors and multifunctional TFs. TFBSs were predicted using RDM.

Additional file 7: Contains RDM-based predicted TFBSs. Tables containing the positions within TFBSs with high, medium and low IC, the number of cytosines in our study (with any $SCC_{M/E}$) overlapping with TFBS, and the number of CpG "traffic lights" overlapping with TFBS for each TF. Consistent information is given for positions with positive $SCC_{M/E}$.

Additional file 8: Contains PWM-based predicted TFBSs. Tables containing the positions within TFBSs with high, medium and low IC, the number of cytosines in our study (with any $SCC_{M/E}$) overlapping with TFBS, and the number of CpG "traffic lights" overlapping with TFBS for each TF. Consistent information is given for positions with positive $SCC_{M/E}$.

Additional file 9: Contains a table listing FANTOM5 samples (cell types) matching 50 ENCODE samples. We grouped them into 50 classes of identical or similar biological cell types. The ENCODE sample description is also provided. Normal/cancer cell types (36/14) are marked in the last column.

Additional file 10: Contains thresholds for PWM corresponding to the *P*-value < 0.0005 (i.e., when 5 of 10,000 random words have scores no less than the thresholds). *P*-values were computed by MACRO-APE software (http://autosome.ru/macroape).

Abbreviations

RRBS: Reduced representation bisulphite sequencing; CAGE: Cap analysis of gene expression; ChIP-seq: Chromatin immunoprecipitation followed by DNA sequencing; TSS: Transcription start site; TF: Transcription factor; TFBS: Transcription factor binding site; RDM: Remote dependency model; PWM: Position weight matrix; $SCC_{M/E}$: Spearman correlation coefficient between methylation and expression profiles; CGI: CpG island; DIC: Discrete information content.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

YAM designed the computational experiments, selected and preprocessed the data, performed statistical analysis and wrote the manuscript. AK performed most of the data analysis. WBA and MdSIB provided RDM models and tools for threshold estimation and mapping. TL was responsible for tag mapping. HK managed the data handling. ARRF was responsible for FANTOM5 management and its concept. IVK performed part of the analysis and contributed to the design of the experiments and writing of the manuscript. VBB contributed to the design of the experiments and writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by a Research Grant from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) to the RIKEN Center for Life Science Technologies; by a research grant for RIKEN Omics Science Center from MEXT to Yoshihide Hayashizaki; and by a research grant from MEXT through RIKEN Preventive Medicine and Diagnosis Innovation Program to YH. We would like to thank all members of the FANTOM5 consortium for contributing to the generation of samples and analysis of the dataset and to thank GeNAS for data production. We thank Prof. A.S. Kondrashov for computational facilities provided to IVK under Russian Ministry of Science and Education grant [11.G34.31.0008]. We also would like to thank Virginia Unkefer for careful editing of the manuscript.
YAM, AK, WBA, MdSIB and VBB were supported by the base research funds of VBB at KAUST. IVK was supported by the Dynasty Foundation, the Russian Foundation for Basic Research, 12-04-32082-mol_a and Russian Ministry of Science and Education State Contract [14.512.11.0092].
RIKEN Omics Science Center ceased to exist as of April 1st, 2013, due to RIKEN reorganization.

Author details

[1]Computational Bioscience Research Center, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. [2]Laboratory of Bioinformatics and Systems Biology, Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilov str. 32, Moscow 119991GSP-1, Russia. [3]Department of Computational Systems Biology, Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina str. 3, Moscow 119991, Russia. [4]RIKEN Omics Science Center, Yokohama, Kanagawa 230-0045, Japan. [5]RIKEN Center for Life Science Technologies, Division of Genomic Technologies, Yokohama, Kanagawa 230-0045, Japan. [6]RIKEN Preventive Medicine and Diagnosis Innovation Program, Wako, Saitama 351-0198, Japan.

Received: 16 April 2013 Accepted: 16 August 2013
Published: 26 March 2014

References

1. Ehrlich M, Gama-Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA, Gehrke C: **Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells.** *Nucleic Acids Res* 1982, **10:**2709–2721.
2. Tucker KL: **Methylated cytosine and the brain: a new base for neuroscience.** *Neuron* 2001, **30:**649–652.
3. Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R: **Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a.** *Proc Natl Acad Sci USA* 2000, **97:**5237–5242.
4. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR: **Human DNA methylomes at base resolution show widespread epigenomic differences.** *Nature* 2009, **462**(7271):315–322.
5. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J, Wei CL: **Dynamic changes in the human methylome during differentiation.** *Genome Res* 2010, **20:**320–331.
6. Suzuki MM, Bird A: **DNA methylation landscapes: provocative insights from epigenomics.** *Nat Rev Genet* 2008, **9:**465–476.
7. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y, Olshen A, Ballinger T, Zhou X, Forsberg KJ, Gu J, Echipare L, O'Geen H, Lister R, Pelizzola M, Xi Y, Epstein CB, Bernstein BE, Hawkins RD, Ren B, Chung WY, Gu H, Bock C, Gnirke A, Zhang MQ, Haussler D, *et al*: **Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications.** *Nat Biotechnol* 2010, **28**(10):1097–1105.
8. Laird PW: **Principles and challenges of genomewide DNA methylation analysis.** *Nat Rev Genet* 2010, **11:**191–203.
9. Clark SJ, Statham A, Stirzaker C, Molloy PL, Frommer M: **DNA methylation: bisulphite modification and analysis.** *Nat Protoc* 2006, **1:**2353–2364.
10. Barrera V, Peinado MA: **Evaluation of single CpG sites as proxies of CpG island methylation states at the genome scale.** *Nucleic Acids Res* 2012, **40:**11490–11498.
11. Furst RW, Kliem H, Meyer HH, Ulbrich SE: **A differentially methylated single CpG-site is correlated with estrogen receptor alpha transcription.** *J Steroid Biochem Mol Biol* 2012, **130:**96–104.
12. Schubeler D: **Molecular biology. Epigenetic islands in a genetic ocean.** *Science* 2012, **338:**756–757.
13. Jaenisch R, Bird A: **Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.** *Nat Genet* 2003, **33**(Suppl):245–254.
14. Borgel J, Guibert S, Li Y, Chiba H, Schubeler D, Sasaki H, Forne T, Weber M: **Targets and dynamics of promoter DNA methylation during early mouse development.** *Nat Genet* 2010, **42:**1093–1100.
15. Farthing CR, Ficz G, Ng RK, Chan CF, Andrews S, Dean W, Hemberger M, Reik W: **Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes.** *PLoS Genet* 2008, **4:**e1000116.
16. Oda M, Yamagiwa A, Yamamoto S, Nakayama T, Tsumura A, Sasaki H, Nakao K, Li E, Okano M: **DNA methylation regulates long-range gene silencing of an X-linked homeobox gene cluster in a lineage-specific manner.** *Genes Dev* 2006, **20:**3382–3394.
17. Tomazou EM, Meissner A: **Epigenetic regulation of pluripotency.** *Adv Exp Med Biol* 2010, **695:**26–40.

18. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, Nelson HH, Karagas MR, Padbury JF, Bueno R, Sugarbaker DJ, Yeh RF, Wiencke JK, Kelsey KT: **Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context.** *PLoS Genet* 2009, **5**(8):e1000602.
19. Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, Beyan H, Whittaker P, McCann OT, Finer S, Valdes AM, Leslie RD, Deloukas P, Spector TD: **Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains.** *Genome Res* 2010, **20**(4):434–439.
20. Miller CA, Sweatt JD: **Covalent modification of DNA regulates memory formation.** *Neuron* 2007, **53:**857–869.
21. Jirtle RL, Skinner MK: **Environmental epigenomics and disease susceptibility.** *Nat Rev Genet* 2007, **8:**253–262.
22. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suner D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu YZ, Plass C, Esteller M: **Epigenetic differences arise during the lifetime of monozygotic twins.** *Proc Natl Acad Sci USA* 2005, **102**(30):10604–10609.
23. Kucharski R, Maleszka J, Foret S, Maleszka R: **Nutritional control of reproductive status in honeybees via DNA methylation.** *Science* 2008, **319:**1827–1830.
24. Esteller M, Fraga MF, Paz MF, Campo E, Colomer D, Novo FJ, Calasanz MJ, Galm O, Guo M, Benitez J, Herman JG: **Cancer epigenetics and methylation.** *Science* 2002, **297:**1807–1808. discussion 1807–1808.
25. Gilbert ER, Liu D: **Epigenetics: the missing link to understanding beta-cell dysfunction in the pathogenesis of type 2 diabetes.** *Epigenetics* 2012, **7:**841–852.
26. Kwok JB: **Role of epigenetics in Alzheimer's and Parkinson's disease.** *Epigenomics* 2010, **2:**671–682.
27. Kangaspeska S, Stride B, Metivier R, Polycarpou-Schwarz M, Ibberson D, Carmouche RP, Benes V, Gannon F, Reid G: **Transient cyclical methylation of promoter DNA.** *Nature* 2008, **452:**112–115.
28. Metivier R, Gallais R, Tiffoche C, Le Peron C, Jurkowska RZ, Carmouche RP, Ibberson D, Barath P, Demay F, Reid G, Benes V, Jeltsch A, Gannon F, Salbert G: **Cyclical DNA methylation of a transcriptionally active promoter.** *Nature* 2008, **452**(7183):45–50.
29. Walsh CP, Bestor TH: **Cytosine methylation and mammalian development.** *Genes Dev* 1999, **13:**26–34.
30. Shen L, Kondo Y, Guo Y, Zhang J, Zhang L, Ahmed S, Shu J, Chen X, Waterland RA, Issa JP: **Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters.** *PLoS Genet* 2007, **3:**2023–2036.
31. Hsieh CL: **Dependence of transcriptional repression on CpG methylation density.** *Mol Cell Biol* 1994, **14:**5487–5494.
32. Jones PA, Takai D: **The role of DNA methylation in mammalian epigenetics.** *Science* 2001, **293:**1068–1070.
33. Wolffe AP, Matzke MA: **Epigenetics: regulation through repression.** *Science* 1999, **286:**481–486.
34. Riggs AD: **X inactivation, differentiation, and DNA methylation.** *Cytogenet Cell Genet* 1975, **14:**9–25.
35. Holliday R, Pugh JE: **DNA modification mechanisms and gene activity during development.** *Science* 1975, **187:**226–232.
36. Bird AP: **Functions for DNA methylation in vertebrates.** *Cold Spring Harb Symp Quant Biol* 1993, **58:**281–285.
37. Bird AP: **DNA methylation versus gene expression.** *J Embryol Exp Morphol* 1984, **83**(Suppl):31–40.
38. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S: **DNA methylation profiling of human chromosomes 6, 20 and 22.** *Nat Genet* 2006, **38**(12)**:**1378–1385.
39. Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D: **Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome.** *Nat Genet* 2007, **39:**457–466.
40. Medvedeva YA, Fridman MV, Oparina NJ, Malko DB, Ermakova EO, Kulakovskiy IV, Heinzel A, Makeev VJ: **Intergenic, gene terminal, and intragenic CpG islands in the human genome.** *BMC Genomics* 2010, **11:**48.
41. Boyes J, Bird A: **Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein.** *EMBO J* 1992, **11:**327–333.
42. Pardo LM, Rizzu P, Francescatto M, Vitezic M, Leday GG, Sanchez JS, Khamis A, Takahashi H, van de Berg WD, Medvedeva YA, van de Wiel MA, Daub CO, Carninci P, Heutink P: **Regional differences in gene expression and promoter usage in aged human brains.** *Neurobiol Aging* 2013.
43. Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schubeler D: **Identification of genetic elements that autonomously determine DNA methylation states.** *Nat Genet* 2011, **43:**1091–1097.
44. Rishi V, Bhattacharya P, Chatterjee R, Rozenberg J, Zhao J, Glass K, Fitzgerald P, Vinson C: **CpG methylation of half-CRE sequences creates C/EBPalpha binding sites that activate some tissue-specific genes.** *Proc Natl Acad Sci USA* 2010, **107:**20311–20316.
45. Tate PH, Bird AP: **Effects of DNA methylation on DNA-binding proteins and gene expression.** *Curr Opin Genet Dev* 1993, **3:**226–231.
46. Choy MK, Movassagh M, Goh HG, Bennett MR, Down TA, Foo RS: **Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated.** *BMC Genomics* 2010, **11:**519.
47. Perini G, Diolaiti D, Porro A, Della Valle G: **In vivo transcriptional regulation of N-Myc target genes is controlled by E-box methylation.** *Proc Natl Acad Sci USA* 2005, **102:**12117–12122.
48. Kim J, Kollhoff A, Bergmann A, Stubbs L: **Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3.** *Hum Mol Genet* 2003, **12:**233–245.
49. Mukhopadhyay R, Yu W, Whitehead J, Xu J, Lezcano M, Pack S, Kanduri C, Kanduri M, Ginjala V, Vostrov A, Quitschke W, Chernukhin I, Klenova E, Lobanenkov V, Ohlsson R: **The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide.** *Genome Res* 2004, **14**(8)**:**1594–1602.
50. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, *et al*: **The accessible chromatin landscape of the human genome.** *Nature* 2012, **489**(7414)**:**75–82.
51. Straussman R, Nejman D, Roberts D, Steinfeld I, Blum B, Benvenisty N, Simon I, Yakhini Z, Cedar H: **Developmental programming of CpG island methylation profiles in the human genome.** *Nat Struct Mol Biol* 2009, **16:**564–571.
52. Gebhard C, Benner C, Ehrich M, Schwarzfischer L, Schilling E, Klug M, Dietmaier W, Thiede C, Holler E, Andreesen R, Rehli M: **General transcription factor binding at CpG islands in normal cells correlates with resistance to de novo DNA methylation in cancer cells.** *Cancer Res* 2010, **70:**1398–1407.
53. Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, Thurman RE, Kaul R, Myers RM, Stamatoyannopoulos JA: **Widespread plasticity in CTCF occupancy linked to DNA methylation.** *Genome Res* 2012, **22**(9)**:**1680–1688.
54. Chatterjee R, Vinson C: **CpG methylation recruits sequence specific transcription factors essential for tissue specific gene expression.** *Biochim Biophys Acta* 1819, **2012:**763–770.
55. Nan X, Ng HH, Johnson CA, Laherty CD, Turner BM, Eisenman RN, Bird A: **Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex.** *Nature* 1998, **393:**386–389.
56. Feldman N, Gerson A, Fang J, Li E, Zhang Y, Shinkai Y, Cedar H, Bergman Y: **G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis.** *Nat Cell Biol* 2006, **8:**188–194.
57. Tachibana M, Matsumura Y, Fukuda M, Kimura H, Shinkai Y: **G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription.** *EMBO J* 2008, **27:**2681–2690.
58. Strunnikova M, Schagdarsurengin U, Kehlen A, Garbe JC, Stampfer MR, Dammann R: **Chromatin inactivation precedes de novo DNA methylation during the progressive epigenetic silencing of the RASSF1A promoter.** *Mol Cell Biol* 2005, **25:**3923–3933.
59. Vire E, Brenner C, Deplus R, Blanchon L, Fraga M, Didelot C, Morey L, Van Eynde A, Bernard D, Vanderwinden JM, Bollen M, Esteller M, Di Croce L, de Launoit Y, Fuks F: **The Polycomb group protein EZH2 directly controls DNA methylation.** *Nature* 2006, **439**(7078)**:**871–874.
60. Macleod D, Charlton J, Mullins J, Bird AP: **Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island.** *Genes Dev* 1994, **8:**2282–2292.
61. Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, Temper V, Razin A, Cedar H: **Sp1 elements protect a CpG island from de novo methylation.** *Nature* 1994, **371:**435–438.

62. Mummaneni P, Yates P, Simpson J, Rose J, Turker MS: **The primary function of a redundant Sp1 binding site in the mouse aprt gene promoter is to block epigenetic gene inactivation.** *Nucleic Acids Res* 1998, **26:**5163–5169.
63. Maurano MT, Wang H, Kutyavin T, Stamatoyannopoulos JA: **Widespread site-dependent buffering of human regulatory polymorphism.** *PLoS Genet* 2012, **8:**e1002599.
64. Lin IG, Tomzynski TJ, Ou Q, Hsieh CL: **Modulation of DNA binding protein affinity directly affects target site demethylation.** *Mol Cell Biol* 2000, **20:**2343–2349.
65. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schubeler D: **DNA-binding factors shape the mouse methylome at distal regulatory regions.** *Nature* 2011, **480**(7378)**:**490–495.
66. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee BK, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Dunham I, *et al*: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**(7414)**:**57–74.
67. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R: **Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis.** *Nucleic Acids Res* 2005, **33:**5868–5877.
68. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A: **Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling.** *Nat Protoc* 2011, **6:**468–481.
69. Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, Andersson R, Mungall CJ, Meehan TF, Schmeier S, Bertin N, Jørgensen M, Dimont E, Arner E, Schmidl C, Schaefer U, Medvedeva YA, Plessy C, Vitezic M, Severin J, Semple CA, Ishizu Y, Young RS, Francescatto M, Alam I, Albanese D, *et al*: **A promoter level mammalian expression atlas.** *Nature* 2014, http://dx.doi.org/10.1038/nature13182.
70. Salimullah M, Sakai M, Plessy C, Carninci P: **NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes.** *Cold Spring Harb Protoc* 2011. pdb.prot5559.
71. Stein R, Razin A, Cedar H: **In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells.** *Proc Natl Acad Sci USA* 1982, **79:**3418–3422.
72. Feinberg AP, Tycko B: **The history of cancer epigenetics.** *Nat Rev Cancer* 2004, **4:**143–153.
73. Esteller M: **Cancer epigenomics: DNA methylomes and histone-modification maps.** *Nat Rev Genet* 2007, **8:**286–298.
74. Ehrlich M: **DNA hypomethylation in cancer cells.** *Epigenomics* 2009, **1:**239–259.
75. Ehrlich M: **DNA methylation in cancer: too much, but also too little.** *Oncogene* 2002, **21:**5400–5413.
76. Kanamori-Katayama M, Itoh M, Kawaji H, Lassmann T, Katayama S, Kojima M, Bertin N, Kaiho A, Ninomiya N, Daub CO, Carninci P, Forrest AR, Hayashizaki Y: **Unamplified cap analysis of gene expression on a single-molecule sequencer.** *Genome Res* 2011, **21**(7)**:**1150–1159.
77. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES: **Genome-scale DNA methylation maps of pluripotent and differentiated cells.** *Nature* 2008, **454**(7205)**:**766–770.
78. Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ, Church GM: **Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells.** *Nat Biotechnol* 2009, **27:**361–368.
79. Jones PA: **The DNA methylation paradox.** *Trends Genet* 1999, **15:**34–37.
80. Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, Balasubramanian S: **Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution.** *Science* 2012, **336:**934–937.
81. Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A: **The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing.** *PLoS One* 2010, **5:**e8888.
82. Wang T, Pan Q, Lin L, Szulwach KE, Song CX, He C, Wu H, Warren ST, Jin P, Duan R, Li X: **Genome-wide DNA hydroxymethylation changes are associated with neurodevelopmental genes in the developing human cerebellum.** *Hum Mol Genet* 2012, **21:**5500–5510.
83. Stroud H, Feng S, Morey Kinney S, Pradhan S, Jacobsen SE: **5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells.** *Genome Biol* 2011, **12:**R54.
84. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, *et al*: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res* 2012, **22**(9)**:**1760–1774.
85. Siddharthan R: **Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix.** *PLoS One* 2010, **5:**e9722.
86. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A: **Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.** *Nat Methods* 2008, **5:**829–834.
87. Daniel JM, Spring CM, Crawford HC, Reynolds AB, Baig A: **The p120(ctn)-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides.** *Nucleic Acids Res* 2002, **30:**2911–2919.
88. Consortium U: **Update on activities at the Universal Protein Resource (UniProt) in 2013.** *Nucleic Acids Res* 2013, **41:**D43–47.
89. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11:**R106.
90. Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, Makeev VJ: **HOCOMOCO: a comprehensive collection of human transcription factor binding sites models.** *Nucleic Acids Res* 2013, **41:**D195–202.
91. Touzet H, Varre JS: **Efficient and accurate P-value computation for Position Weight Matrices.** *Algorithms Mol Biol* 2007, **2:**15.
92. Bajic VB, Seah SH, Chong A, Krishnan SP, Koh JL, Brusic V: **Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates.** *J Mol Graph Model* 2003, **21:**323–332.
93. Kulakovskiy IV, Favorov AV, Makeev VJ: **Motif discovery and motif finding from genome-mapped DNase footprint data.** *Bioinformatics* 2009, **25:**2318–2325.
94. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A: **Information content of binding sites on nucleotide sequences.** *J Mol Biol* 1986, **188:**415–431.
95. Kulakovskiy IV, Makeev VJ: **Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources.** *Biophysics* 2009, **54:**667–674.
96. Kulakovskii IV, Makeev V: **Integration of data obtained by different experimental methods to determine the motifs in DNA sequences recognized by transcription-regulating factors.** *Biofizika* 2009, **54:**965–974.

doi:10.1186/1471-2164-15-119
Cite this article as: Medvedeva *et al.*: Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics* 2013 **15**:119.