CpG traffic lights are markers of regulatory regions in humans.

Abdullah M. Khamis[1,*], Anna V. Lioznova[2,*], Artem V. Artemov[2,3,4], Vasily Ramensky[5], Vladimir B. Bajic[1], Yulia A Medvedeva[2,6,7,#]

[1] King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, Thuwal 23955-6900, Saudi Arabia
[2] Institute of Bioengineering, Research Center of Biotechnology, Russian Academy of Sciences, Moscow 119071, Russia
[3] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow 119991, Russia
[4] Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow 127051, Russia
[5] Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, California 90095, USA
[6] Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow 119991, Russia
[7] Moscow Institute of Physics and Technology, Dolgoprudny 141701, Russia

*** Authors contributed equally to the work**
# Correspondence should be addressed to ju.medvedeva@gmail.com

Abstract

DNA methylation is involved in regulation of gene expression. Although modern methods profile DNA methylation at single CpG sites, methylation levels are usually averaged over genomic regions in the downstream analyses. In this study we demonstrate that single CpG methylation can serve as a more accurate predictor of gene expression compared to average promoter / gene body methylation. CpG positions with significant correlation between methylation and expression of a gene nearby (called CpG traffic lights) are evolutionary conserved and enriched for exact TSS positions and active enhancers. Among all promoter types, CpG traffic lights are especially enriched in poised promoters. Genes that harbor CpG traffic lights are associated with development and signal transduction. Methylation levels of individual CpG traffic lights vary between cell types dramatically with the increased frequency of intermediate methylation levels, indicating cell population heterogeneity in CpG methylation levels Being in line with the concept of the inherited stochastic epigenetic variation, methylation of such CpG positions might contribute to transcriptional regulation. Alternatively, one can hypothesize that traffic lights are markers of absent gene expression resulting from inactivation of their regulatory elements. In any case, CpG traffic light mechanism provide a promising insight into enhancer activity and gene expression, important from both fundamental and practical points of view.

Keywords

Regulation of transcription, DNA methylation, enhancers, CAGE, chromatin states, CpG traffic lights

Introduction

Epigenetic regulation of gene expression attracts a lot of research attention over the last decade with cytosine methylation being probably the most well-investigated mechanism. DNA methylation is linked to many normal and pathological biological processes: organism development, cell differentiation, cell identity and pluripotency maintenance (reviewed in [20,31,45]), aging [4], memory formation [12,32], responses to environmental exposures, stress and diet [22, 27, 33] There is an increasing evidence of abnormalities in DNA methylation present in various diseases, including metabolic [7], cardiovascular [50], neurodegenerative [39, 48] diseases and cancers (reviewed in [2]. For about a decade, DNA demethylating drugs (Decitabine, Azacytidine) are used in clinic for the treatment of acute myeloid leukemia and myelodysplastic syndrome [6]. Recent advances in site-specific editing of DNA methylation [40] suggest the possibility of exploring DNA methylation as a promising target for non-invasive therapies against many diseases linked with aberrant methylation.

Functionally, DNA methylation of promoter regions is tightly associated with the repression of transcription initiation, while methylation of the gene body is proportional to expression intensity (reviewed in [23]). Enhancers, distant regulatory regions that contribute to the establishment of the correct temporal and cell-type-specific gene expression pattern, have been shown to initiate transcription of short RNAs by PolII [24]. Therefore, it is no surprise that DNA methylation might also regulate the enhancer function as well [16, 25, 34]. Recent studies support the role of DNA methyltransferase in enhancer-associated transcription [36]. The enhancers locations are more difficult to determine genome-wide than those of genes. Some progress in this direction has been made with the use of histone modifications profiles, transcription factor binding or DNase I hypersensitive sites (DHSs) (reviewed in [41]) or the presence of balanced bidirectional capped transcripts (CAGE) [1]. Yet, due to the difficulties in localization of enhancers, the role of their methylation is not completely clear.

It is important to emphasize that epigenetic profiles vary between cells that belong to the same organism and therefore share the same genetic background. The majority of these epigenetic differences are established during development and can be explained by cell types and tissues in a multicellular organism. Yet, an epigenetic heterogeneity has been observed in the normal tissues of inbred laboratory mice [19] and at the level of single cells [10], suggesting stochasticity in the epigenetic profiles intrinsic to some genome loci but not others [11]. The effect of genome-wide epigenetic stochasticity for gene expression has not been addressed so far in details [10].

Contemporary methods to study DNA methylation based on bisulfite sequencing allow detection of single cytosine methylation. Yet, at the step of downstream bioinformatic analysis, methylation levels of several dozens of cytosines are usually averaged with the aim to increase statistical power [3]. However, several examples show that changes in methylation of a single CpG affect gene transcription [29]. Recently, we have shown that methylation levels of particular single CpGs are tightly linked to expression for specific cases [30]. We have called such positions CpG traffic lights (TL) and have demonstrated a strong negative selection against them in transcriptional factor binding sites. In this study we show enrichment of TLs in transcriptional start sites (TSS), in particular, in poised promoters, enhancers and regions with active chromatin marks, suggesting another mechanism of transcriptional regulation. Also, a study of methylation at the level of a single CpG dinucleotide allows one to address the issue of methylation heterogeneity. Although, allele-specific methylation has been reported to affect up to 10% of human genes [49] it is usually linked to genetic polymorphisms [42, 49], therefore reducing the contribution of allele-specific methylation when samples from different individuals are investigated. So technical errors aside, intermediate values of methylation, if observed in the same location but in different samples, show regions of high cell population heterogeneity. Here, we report a high level of cell population heterogeneity of methylation levels in TLs suggesting a novel flexible yet abundant mechanism of transcriptional regulation.

Results

CpG traffic lights determination

As has been shown many times, DNA methylation of a promoter can repress expression of a corresponding gene. Nevertheless, correlation between gene expression and methylation of its promoter or body is not straightforward, suggesting the need to deconvolute DNA methylation profiles into the regions smaller than promoters. For this purpose, we focus on a methylation level of particular CpGs to investigate the link between methylation and expression. Following the logic previously reported in our work [30] where we used the reduced set of CpGs in the RRBS data, we expanded our previous approach and use whole-genome DNA methylation data (bisulfite sequencing, WGBS) and expression (RNA-seq) levels for 40 normal human primary cells and tissues from the Roadmap Epigenomics Project. We define CpG traffic lights as CpG dinucleotides with significant Spearman correlation coefficient (SCC) between DNA methylation and expression levels of a neighbouring gene (FDR <0.1, Fig. 1).

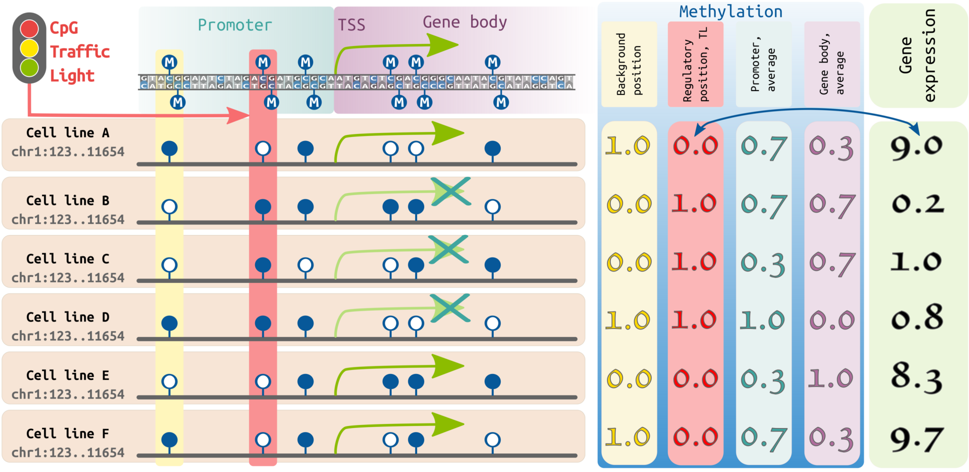

Figure 1. Schematic representation of a CpG traffic light determination. Left panel. Suppose we analyze a particular genomic region (chr1:123..11654), which contains for simplicity one gene, for 6 cell lines. For each CpG in this region and the gene we have methylation and expression vectors, respectively. CpG positions are represented by dark blue lollipops (filled: methylated CpG, empty: unmethylated CpG). First three CpGs are located within the promoter region, while the last three are located in gene body. Gene expression or lack of it is represented by green arrows. **Right panel.** A yellow column shows methylation of a random CpG (used as a background), methylation vector of this CpG demonstrates low correlation with gene expression (green box on the right, in RPKM). Correlation between an average promoter/gene body methylation (shown in light blue and light purple columns, respectively) and the corresponding gene expression is also low. However, for TLs (shown in red), methylation level significantly correlates with gene expression.

Here we state that the average methylation of promoter/gene body region has weaker correlation with gene expression genome wide, as compared to the methylation of TLs. In particular, at the level of FDR<0.1 we find only 44/58 genes for which average promoter/gene body methylation vectors correlate with expression vectors, while at the same level of significance we observe 6,153 genes to correlate well with methylation levels of TLs. Other levels of significance demonstrate similar tendency (Table 1).

Table 1. Number of genes which have significant correlation between expression and methylation. Note: for multiplicity testing correction the number of genes was used in (1) and (2), while number of all CpG positions in each studied gene was used for the same purpose in (3). The (TTS) refers to the Transcript Termination Site.

FDR-corrected p-value (significance level)	Total number of genes, which have significant correlations between gene expression and methylation		
	average methylation of promoter region (-1000..500) (1)	average methylation of gene body (+500..TTS) (2)	methylation of CpG traffic light (3)
0.05	0	11	2,706
0.1	44	58	6,153
0.2	300	406	12,040

Among TLs, defined above (FDR<0.1), the majority of those located in promoters demonstrate negative SCC, while the majority of those located in intron demonstrate positive SCC, and TLs in exons demonstrate similar number of both positive and negative SCC (Fig. 2). TL are uniformly distributed along the genome (Manhattan plot, Supplementary Figure S1).

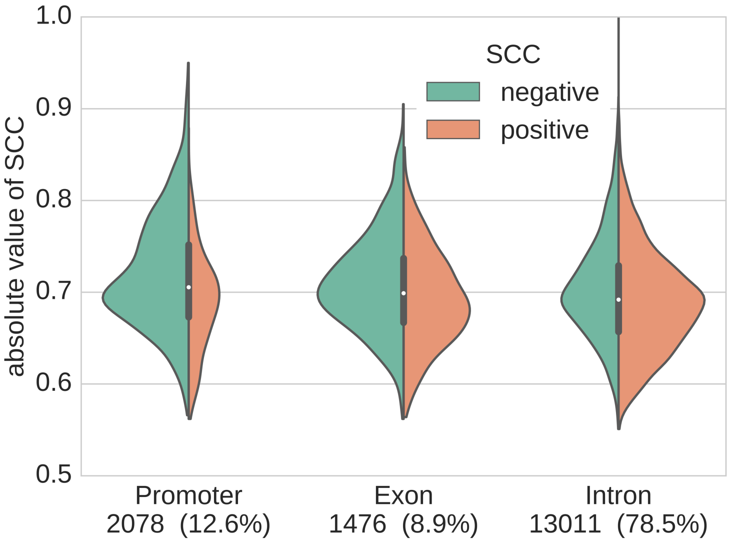

Figure 2. The distribution of SCC in the TLs. The total number of TLs in promoters, exons and introns are present at the bottom. Green (left) / pink (right) parts of the violin plots show the distribution of positive and negative SCC, respectively.

CpG traffic lights are associated with highly heterogeneous genomic regions

A single CpG position can be either methylated or not, resulting in a 0 or 1 methylation levels, in a diploid cell, allele specific methylation for some CpG positions can result in the methylation levels of 0.5. Since the allele specific methylation is usually linked to SNPs, intermediate methylation levels reported at the same genomic locations for several genetically unrelated samples usually means heterogeneity of methylation levels among individual cells at a given CpG position. For the majority of CpG positions not detected as TLs (background CpGs, see Methods section for details), the levels of methylation were close either to 0 or 1 in all studied cell types (Fig. 3a,b), demonstrating homogeneity of the methylation levels in the cell population. At the same time the TLs with negative SCC between expression and methylation, both located in promoters and gene body, are intermediately methylated in many cell types (Fig. 3c,d). The similar tendency was observed for TLs with positive SCC (Supplementary figure S2).

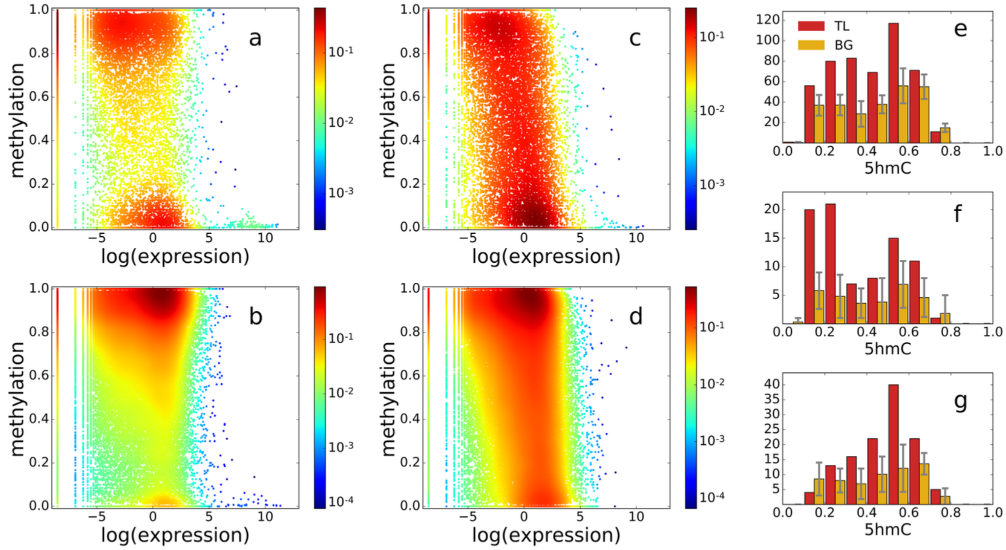

Figure 3. The distribution of CpG methylation and corresponding gene expression for TLs and background (negative SCC). The color represents the density of points in logarithmic scale. The distribution is shown for **(a)** random background CpG (BG) in promoters (the number is equal to the number of TL points), **(b)** random background CpG (BG) in gene bodies, **(c)** TL in promoters (-1000. . . +500), **(d)** TL in gene bodies (+500...TTS), **(e)** levels of 5hmC in TL and BG, **(f)** levels of 5hmC for TL with positive and **(g)** negative causality score between DNA methylation and gene expression. Whiskers represent minimum/maximum out of the 10 random background samples.

Since methylation levels of TLs are clearly more heterogeneous than that of background CpGs, we decided to test whether methylation of these positions is also more dynamic in time. As a proxy of methylation dynamics we used levels of hydroxymethylcytosine (5hmC). Although the functional role of 5hmC is not fully elucidated, one of the most supported hypothesis is that 5hmC is an intermediate product of active DNA demethylation [14]. In standard bisulfite conversion experiments 5hmC cannot be distinguished from its precursor 5mC [21]. To compensate for that we used Illumina 450K oxBS-array data [13]. We report that TLs are enriched for 5hmC as compared to the background CpG, supporting the idea of dynamic methylation in TLs (Fig 3e). This dynamic methylation of the TLs supports the heterogeneity observed among them.

As a next step we divided TLs into subgroups based on causality scores, which allows one to computationally determine which of the vectors (methylation or expression) is the causal variable (see the Methods section for details). In our case, positive causality scores reflect cases where changes in DNA methylation cause the change in expression, whereas negative values of causality score correspond to CpG positions for which levels of methylation are a consequence of expression level. Surprisingly, it is mostly TLs with negative causality scores that demonstrate enrichment of high concentrations of 5hmC per site (Fig. 3g), which may suggest a positive feedback loop of the active transcription that activates DNA demethylation.

CpG traffic lights are conserved across mammals and primates

To address functionality of TLs, we first investigate their evolutionary conservation. By comparing TLs with negative SCC with ten random CpG background sets of the same size and of the same GC/CpG content (see Methods) we demonstrate that the TLs are enriched with conserved positions both in mammals and in primates, estimated by GERP RS and PhyloP conservation scores, respectively (Fig. 4ab). Also, TLs are depleted in polymorphisms from ExAC (Fig 4c), as well as in repetitive sequences determined by both chromatin states (chromHMM, Fig. 4e) and repeatMasker (Fig 5a). This is in agreement with Eigen non-coding scores being significantly higher for TLs (Fig 4d). Moreover, gene enrichment analysis (Table 2) for GO terms shows that TLs are linked to genes involved in development, cell-to-cell communication and apoptosis. Taken together, these results clearly suggest the functional role of TLs in the genome.

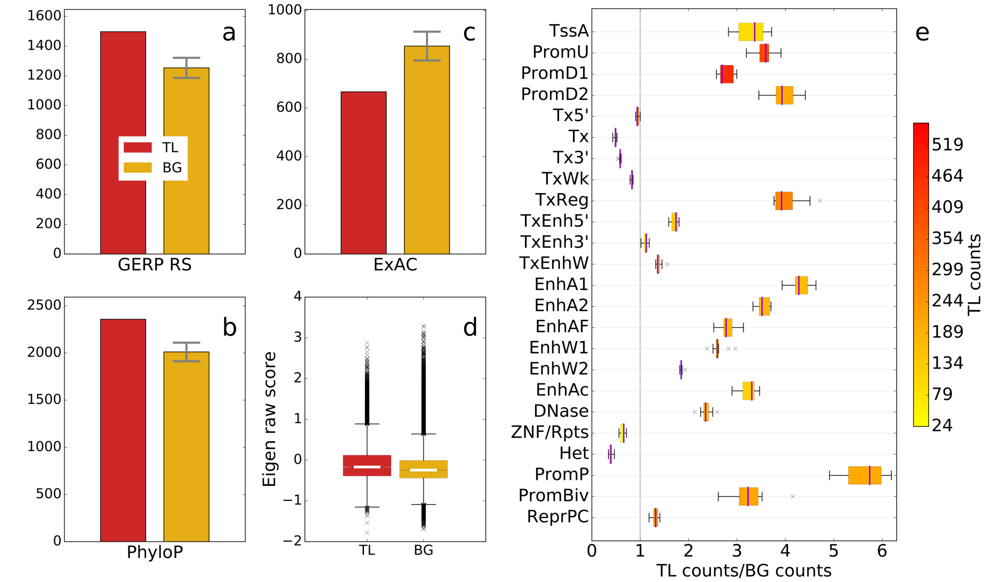

Figure 4. Number of TL and BG sites demonstrating evolutionary conservation (a) in mammals and **(b)** in primates, **(c)** polymorphisms from ExAC, **(d)** Eigen non-coding functionality score, **(e)** averaged across 127 cell types ratio of TL / BG in chromatin states determined by chromHMM. Whiskers (abc) represent minimum / maximum out of the 10 random background samples. The color (e) reflects absolute number of TL located in the given chromatin state.

Table 2. Enrichment of TLs in biological processes (www.pantherdb.org)

PANTHER GO-Slim Biological Process	Homo sapiens (REF)	# genes with TL	# genes expected	Fold Enrichment	+/-	P value
developmental process	1938	627	451.14	1.39	+	2.11E-14
cellular process	8199	2160	1908.62	1.13	+	3.24E-11
cell communication	2674	787	622.47	1.26	+	1.21E-09
cell adhesion	481	190	111.97	1.7	+	1.57E-09
biological adhesion	481	190	111.97	1.7	+	1.57E-09
system development	1065	358	247.92	1.44	+	2.05E-09
signal transduction	2390	702	556.36	1.26	+	3.16E-08
nervous system development	668	238	155.5	1.53	+	5.81E-08
cell-cell adhesion	305	124	71	1.75	+	1.43E-06
intracellular signal transduction	991	312	230.69	1.35	+	2.46E-05
mesoderm development	447	159	104.06	1.53	+	5.96E-05
ectoderm development	405	142	94.28	1.51	+	5.32E-04
heart development	143	62	33.29	1.86	+	1.23E-03
cellular component movement	413	137	96.14	1.42	+	1.03E-02
mitosis	372	122	86.6	1.41	+	4.05E-02
induction of apoptosis	85	38	19.79	1.92	+	4.12E-02

CpG traffic lights are enriched in transcription start sites, promoters and enhancers

To specify the functional role of TLs we tested various different genomic markups for the overrepresentation. We observe that TL are enriched in all promoter types, determined by chromHMM, including active, bivalent and poised promoters (Fig. 4e). Interestingly, the strongest enrichment was observed in poised promoters (¿3.5 times). Since poised or bivalent chromatin is thought to be able to easily switch between active and repressed states [28], such enrichment may suggest TLs as a possible mechanism contributing to maintenance of the bivalent state of chromatin.

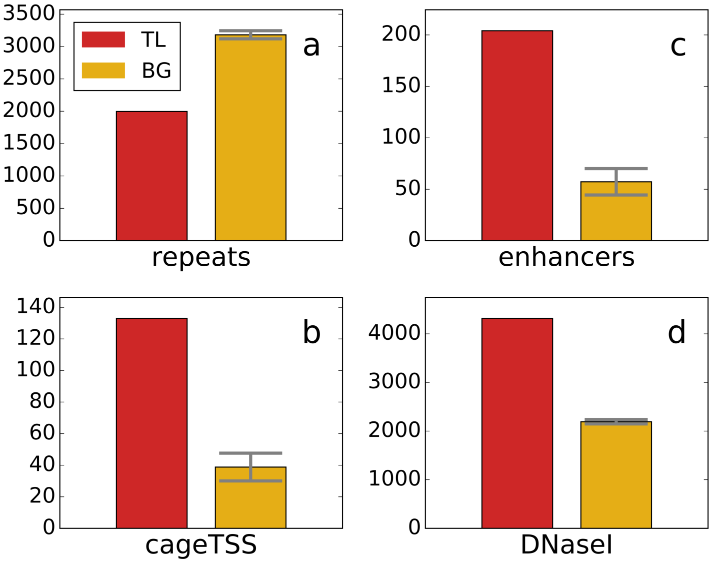

Figure 5. Frequency of TL and BG CpGs in repeats **(a)**, TSS determined by CAGE **(b)**, enhancers **(c)** and DNase hypersensitive sites **(d)**. Whiskers represent minimum/maximum out of the 10 random background samples. All differences are significant (p-value <5E-5).

We also notice that TL are highly enriched in the chromatin state, corresponding to transcriptional start site (TSS) per se. To dig deeper, we use TSS, determined by CAGE (Cap Analysis of Gene Expression), currently the most accurate technique to determine exact locations of TSS [9]. We determine 3.5-fold overrepresentation of TL in the exact TSS position (Fig 5b). It should be noted that among all groups of TL located in TSS, the biggest group with the most pronounced overrepresentation over the background, has negative correlation and causality scores (Supplementary Figure S3). Negative causality score represents that levels of expression are the cause of the methylation levels, suggesting that for TSS regions methylation of a TL is a marker, not the cause of expression.

Our data also show that TL are enriched in various regulatory regions, yet the strongest enrichment is observed in enhancers, determined by CAGE bi-directional transcription (Fig 5a) and chromatin states (Fig 4). Although all the enhancers are enriched for TLs, some types of enhancers are more prone to harbour them. We detect that among all enhancer categories the most enriched are hematopoietic and stem cell enhancers (Supplementary table 1). All open chromatin regions determined as regions sensitive to DNaseI are also enriched for TL (Fig 4e, 5d). On the other hand, as we reported before [30], TL are not enriched in TFBS if TFBS prediction is performed in the DNaseI sensitive regions (Supplementary figure S4).

Discussion

DNA methylation is tightly involved in regulation of gene expression in various normal and pathological processes. Therefore, it is an attractive target for therapies of the diseases with epigenetic abnormalities (reviewed in [38]). Modern technologies based on bisulfite sequencing allow for detection of DNA methylation with a single CpG dinucleotide resolution. Yet, at the stage of the downstream analysis methylation levels are averaged over the large regions. In this work we demonstrate that methylation profiles of particular single CpG dinucleotides (TLs) are stronger correlated with gene expression as compared to average promoter / gene body methylation even if for the multiplicity testing total number of CpG-gene pairs is used. It is a surprising observation, since it is widely accepted that DNA methyltransferases once bound to DNA move along [15] it or multimerize [43]methylating all neighbouring CpGs unless a boundary protein, such as Sp1, is in their way (reviewed in [46]. Yet, only a small fraction of TLs are located within the promoter and body of the same gene. We speculate that local change in DNA methylation can be achieved rather through active DNA demethylation, probably with the help of TET proteins, since byproduct of active demethylation, 5hmC is found to be overrepresented in TLs. However, a direct experiment, probably with the use of CRISPR/TALEN-based technology, is required to validate this hypothesis.

TLs are evolutionary conserved in both mammal and primate lineage, suggesting possible selection constraint, as well as depleted in SNPs, repeats and heterochromatin regions, supporting the hypothesis of TL functionality. Genes that harbor TLs are associated with fundamental biological processes, such as development and signal transduction. TLs are also enriched in open chromatin and various regulatory regions, in particular in the exact TSS positions and active enhancers, especially those detected by bi-directional CAGE transcription [1]. This observation is in line with the recent reports that DNMT3a/b are associated with enhancers and are important regulators of enhancer RNA production in hematopoietic stem cells [36]

In the light of overrepresentation in regulatory regions, depletion of TLs within TFBS is puzzling. One possible explanation would be that TLs are CpG dinucleotides located within the enhancers but outside the sites of regulatory protein binding. In this case, cytosine methylation accumulates as a consequence of the absence of TF binding [44, 47], which makes methylation of TL not a primary cause, but just a ”passive” marker of absent gene expression resulting from inactivation of its regulatory element. Still, CpG traffic light methylation is a reliable marker of enhancer activity and gene expression, and can be used for practical applications.

Methylation levels of TLs vary between cell types dramatically and are characterized by increased frequency of intermediate methylation levels, indicating that only a fraction of cells within the same tissue have a certain CpG traffic light methylated. This variation cannot be attributed to genetic polymorphisms, since for our study we used samples from genetically different subjects, so it would be highly unlikely to have the same allele in a given position in the huge fraction of samples in the study. The more probable explanation is heterogeneity at the cell population level, which is indirectly supported by methylation dynamics in the form of increased levels of 5hmC. This heterogeneity is most likely stochastic, suggesting the ”active” role for TLs as a novel highly dynamic yet abundant mechanism of transcriptional regulation. This hypothesis is supported by the observation that the TLs are highly overrepresented in poised promoters, suggesting their contribution into dynamics of expression.

Methods

DNA methylation and expression data processing

We selected 40 tissues and cell types (see Supplementary table 2) for which both WGBS and RNA-seq data were available in Roadmap Epigenomics Project (NCBI). For WGBS data for each cell type we used three replicates with the highest number of reads and the best genome-mapping ratio. For 28 cell types 3 RNA-seq replicates were available, while for 7/5 cell types only 2/1 replicate were available respectively. All WGBS data and the majority (95 out of 103) RNA-seq files were obtained by Illumina, while 8 RNA-seq files were obtained by SOLiD. The quality of all files were checked with FASTQC (bioinformatics.bbsrc.ac.uk/projects/fastqc). For all files sequenced by Illumina read trimming and adapter removal were performed by Trimmomatic (adapters from NCBI; up to 2 mismatches between an adapter and a read sequence; 5bp sliding window; quality threshold of 20; removing sequences if their length after trimming becomes less than 20 bp) For the SOLiD samples we used Cutadapt (adapters from NCBI, up to 10% error rate relative to the length of the matching region; quality threshold of 20; removing sequences if their length after trimming becomes less than 20 bp).

We mapped WGBS data to the genome (assembly GRCh38-Ensembl 78) with Bismark (zero mismatches permitted in the seed, 20bp seed length, 0/500bp the min/max insert size for valid paired-end alignments). We used only methylated cytosines in CpG context, covered with not less than 4 reads. For each CpG position in each of the 40 samples, the methylation values were averaged from the three replicates per sample. We removed a CpG position if it has values in less than 20 samples. We also removed a CpG position if it had the same value for all the samples (i.e. all zeros, all ones, etc.), as this position did not vary across samples .

We mapped RNA-seq data to the genome (assembly GRCh38-Ensembl 78) with Tophat v2.0.13 (allowing for up to 2 mismatches and 2 gaps per read, reporting read alignments for paired-end reads only if both reads in a pair can be mapped). We generated expression matrix using the FeatureCount tool, while the expression profiles were normalized to RPKM values. Genes with zero reads in all samples were removed. The expression profiles were normalized to a range [0, 1] [y = (x-xmin)/(xmax-xmin)] to match the range with the one of the methylation profile.

CpG traffic lights detection

To determine TLs we considered all pairs of genes and CpGs located within 1000 bp upstream of gene's TSS to its 3' end (genome assembly GRCh38-Ensembl 78). One CpG might be associated with multiple genes, similarly, one gene might be associated with multiple CpGs. For each CpG-gene pair we created two 20-40-dimensional vectors of methylation levels [0, 1] and normalized gene expression [0, 1], we further refer to each of the two vectors as a methylation and expression profiles. In total we had 1,774,602 CpGs associated with 46,692 genes (which gives 1,963,205 pairs).

For each CpG position, we calculated SCC between the methylation and expression profiles for all available samples. FDR was performed by Benjamini-Hochberg procedure for correction for multiplicity testing for the total number of position-gene pairs. We called a CpG position a CpG traffic light (TL) if it had a significant correlation coefficient between methylation and expression profiles at the level of FDR<0.1 (unless explicitly mentioned otherwise). We found 16,178 such TLs (0.9% of the original number of CpGs) that correspond to 6,153 genes.

We also calculated a causality score between methylation and expression profiles to computationally assess the pairwise causal direction between these two variables. We used a pairwise linear non-Gaussian acyclic model, LINGAM [17] to calculate the likelihood ratio defined as follows:

$$R(Meth, Expr) = \log L(Meth \rightarrow Expr) - \log L(Expr \rightarrow Meth)$$

The positive causality means that the change in methylation is expected to cause the expression change, and vice versa for the negative causality values: expression determines methylation. It should be noted that the range for possible causality scores depends on the number of samples. Since for different CpG positions we used various numbers of samples (20-40), we normalized causality scores to the normal distribution N(0, 1). To make the causality scores directly comparable between CpG positions, we performed this normalization independently for each group of CpGs that have the same profile length.

To avoid noise in the causality scores, we did not consider values close to 0 (between -1 and +1) and for simplicity we call “positive” the values that are higher than 1 and “negative” the values that were smaller than -1.

Construction of background datasets

We aimed to explore enrichment with TLs inside various genomic regions. For this purpose we needed to have an equal size background set. For every TL position we selected a random background CpG position with not more than 5% difference for both GC- and CpG contents in 200bp window, as some genomic annotations are sensitive to GC- and CpG- content. We repeated the selection process 10 times to obtain 10 different independent background sets.

For heatmaps (Fig 2) we selected TL with negative SCC and an equal size random background set, split all the CpGs into promoter regions [TSS - 1000, TSS + 500] and gene body [TSS + 500, end of the gene] and created density plots using gaussian_kde from scipy.stats.

Genomic annotations

We annotated all CpG positions with overlapping genomic features. For each feature we calculated the number of TL and background positions located within the annotation. To test the significance of the overrepresentation we used the exact Fisher test. Additionally, we calculated the overrepresentation for TL with positive/negative SCC/causality scores separately.

For 5-hydroxymethylcytosine in human cerebellum we used oxidative-bisulfite (oxBS) assay data from GEO (GSE63179). We converted the coordinates to genomic ranges with the help of R Bioconductor ‘minfi’ package and to hg38 with liftOver. Four oxBS replicates were averaged.

We use repeats obtained by RepeatMasker for hg38 track from USCS Genome Browser hgdownload.soe.ucsc.edu/goldenPath/hg38/database/rmsk.txt.gz

We obtained the robust CAGE clusters [9] from fantom.gsc.riken.jp/5/data/ and the robust hg19 enhancers [1] from FANTOM5 from (enhancer.binf.ku.dk/presets/robust_enhancers.bed) and mapped them to hg38 with the liftOver.

The DNaseI hypersensitivity clusters were downloaded from UCSC Genome Browser (hgdownload.soe.ucsc.edu/goldenPath/hg38/database/wgEncodeRegDnaseClustered.txt.gz)

Conservation and Eigen scores

Conservation of TL and background sites in mammalian and primate lineages was assessed with UCSC Genome Browser GERP RS [5] and PhyloP [35] hg19 tracks, respectively. We calculated how many sites in each dataset have GERP RS score greater than 2, which we considered as conserved in mammals and PhyloP score greater than 0.5, which we considered conserved in primates. Overall functional scores for each site were calculated with Eigen, an approach to predict functionality of non-coding variants using different annotations [18]. Higher Eigen scores imply more likely functionality of respective genome sites.

TFBS

For transcriptional factor binding site (TFBS) prediction we used models provided in HOCOMOCO v10 [26]. PWM thresholds were selected according to the pre-calculated the P-value <0.0005 (i.e., when 5 of 10,000 random words had scores no less than the thresholds). Out of all predicted TFBS we considered only those present in DNaseI hypersensitivity regions.

ChromHMM

The Roadmap Epigenomics Consortium 25-state segmentation of 127 epigenomes predicted with ChromHMM [8, 37] was used to assess chromatin location of TL. The annotation in based on the imputed data for 12 chromatin marks (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H4K20me1, H3K79me2, H3K36me3, H3K9me3, H3K27me3, H2A.Z, and DNaseI). The annotations were downloaded from egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp.

Each of the TL/background datasets was characterized by an average frequency of a CpG from a dataset to be located in one of the 25 chromatine stated.

Authors contribution

AK processed the raw data and contributed to data analysis; AL contributed to data processing and performed the overrepresentation analysis; VR performed data analysis; AA contributed to statistical analysis; BVB contributed to study design; YAM designed the study and drafted the MS. All authors contributed to MS preparation.

Acknowledgments

This work was supported by RFBR grant 14-04-00180 to YAM. VBB is supported by the base research fund of the King Abdullah University of Science and Technology (KAUST).

References

1. R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jørgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller, FANTOM Consortium, A. R. R. Forrest, P. Carninci, M. Rehli, and A. Sandelin. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, 27 Mar. 2014.

2. S. B. Baylin and P. A. Jones. Epigenetic determinants of cancer. *Cold Spring Harb. Perspect. Biol.*, 8(9), 1 Sept. 2016.

3. C. Bock. Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.*, 13(10):705–719, Oct. 2012.

4. B. C. Christensen, E. A. Houseman, C. J. Marsit, S. Zheng, M. R. Wrensch, J. L. Wiemels, H. H. Nelson, M. R. Karagas, J. F. Padbury, R. Bueno, D. J. Sugarbaker, R.-F. Yeh, J. K. Wiencke, and K. T. Kelsey. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet.*, 5(8):e1000602, Aug. 2009.

5. E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, 6(12):e1001025, 2 Dec. 2010.

6. E. J. B. Derissen, J. H. Beijnen, and J. H. M. Schellens. Concise drug review: azacitidine and decitabine. *Oncologist*, 18(5):619–624, 13 May 2013.

7. M. Desai, J. K. Jellyman, and M. G. Ross. Epigenomics, gestational programming and risk of metabolic syndrome. *Int. J. Obes.*, 39(4):633–641, Apr. 2015.

8. J. Ernst and M. Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, 9(3):215–216, 28 Feb. 2012.

9. FANTOM Consortium and the RIKEN PMI and CLST (DGT), A. R. R. Forrest, H. Kawaji, M. Rehli, J. K. Baillie, M. J. L. de Hoon, V. Haberle, T. Lassmann, I. V. Kulakovskiy, M. Lizio, M. Itoh, R. Andersson, C. J. Mungall, T. F. Meehan, S. Schmeier, N. Bertin, M. Jørgensen, E. Dimont, E. Arner, C. Schmidl, U. Schaefer, Y. A. Medvedeva, C. Plessy, M. Vitezic, J. Severin, C. A. Semple, Y. Ishizu, R. S. Young, M. Francescatto, I. Alam, D. Albanese, G. M. Altschuler, T. Arakawa, J. A. C. Archer, P. Arner, M. Babina, S. Rennie, P. J. Balwierz, A. G. Beckhouse, S. Pradhan-Bhatt, J. A. Blake, A. Blumenthal, B. Bodega, A. Bonetti, J. Briggs, F. Brombacher, A. M. Burroughs, A. Califano, C. V. Cannistraci, D. Carbajo, Y. Chen, M. Chierici, Y. Ciani, H. C. Clevers, E. Dalla, C. A. Davis, M. Detmar, A. D. Diehl, T. Dohi, F. Drabløs, A. S. B. Edge, M. Edinger, K. Ekwall, M. Endoh, H. Enomoto, M. Fagiolini, L. Fairbairn, H. Fang, M. C. Farach-Carson, G. J. Faulkner, A. V. Favorov, M. E. Fisher, M. C. Frith, R. Fujita, S. Fukuda, C. Furlanello, M. Furino, J.-I. Furusawa, T. B. Geijtenbeek, A. P. Gibson, T. Gingeras, D. Goldowitz, J. Gough, S. Guhl, R. Guler, S. Gustincich, T. J. Ha, M. Hamaguchi, M. Hara, M. Harbers, J. Harshbarger, A. Hasegawa, Y. Hasegawa, T. Hashimoto, M. Herlyn, K. J. Hitchens, S. J. Ho Sui, O. M. Hofmann, I. Hoof, F. Hori, L. Huminiecki, K. Iida, T. Ikawa, B. R. Jankovic, H. Jia, A. Joshi, G. Jurman, B. Kaczkowski, C. Kai, K. Kaida, A. Kaiho, K. Kajiyama, M. Kanamori-Katayama, A. S. Kasianov, T. Kasukawa, S. Katayama, S. Kato, S. Kawaguchi, H. Kawamoto, Y. I. Kawamura, T. Kawashima, J. S. Kempfle, T. J. Kenna, J. Kere, L. M. Khachigian, T. Kitamura, S. P. Klinken, A. J. Knox, M. Kojima, S. Kojima, N. Kondo, H. Koseki, S. Koyasu, S. Krampitz, A. Kubosaki, A. T. Kwon, J. F. J. Laros, W. Lee, A. Lennartsson, K. Li, B. Lilje, L. Lipovich, A. Mackay-Sim, R.-I. Manabe, J. C. Mar, B. Marchand, A. Mathelier, N. Mejhert, A. Meynert, Y. Mizuno, D. A. de Lima Morais, H. Morikawa, M. Morimoto, K. Moro, E. Motakis, H. Motohashi, C. L. Mummery, M. Murata, S. Nagao-Sato, Y. Nakachi, F. Nakahara, T. Nakamura, Y. Nakamura, K. Nakazato, E. van Nimwegen, N. Ninomiya, H. Nishiyori, S. Noma, S. Noma, T. Noazaki, S. Ogishima, N. Ohkura, H. Ohimiya, H. Ohno, M. Ohshima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. A. Ovchinnikov, A. Pain, R. Passier, M. Patrikakis, H. Persson, S. Piazza, J. G. D. Prendergast, O. J. L. Rackham, J. A. Ramilowski, M. Rashid, T. Ravasi, P. Rizzu, M. Roncador, S. Roy, M. B. Rye, E. Saijyo, A. Sajantila, A. Saka, S. Sakaguchi, M. Sakai, H. Sato, S. Savvi, A. Saxena, C. Schneider, E. A. Schultes, G. G. Schulze-Tanzil, A. Schwegmann, T. Sengstag, G. Sheng, H. Shimoji, Y. Shimoni, J. W. Shin, C. Simon, D. Sugiyama, T. Sugiyama, M. Suzuki, N. Suzuki, R. K. Swoboda, P. A. C. 't Hoen, M. Tagami, N. Takahashi, J. Takai, H. Tanaka, H. Tatsukawa, Z. Tatum, M. Thompson, H. Toyodo, T. Toyoda, E. Valen, M. van de Wetering, L. M. van den Berg, R. Verado, D. Vijayan, I. E. Vorontsov, W. W. Wasserman, S. Watanabe, C. A. Wells, L. N. Winteringham, E. Wolvetang, E. J. Wood, Y. Yamaguchi, M. Yamamoto, M. Yoneda, Y. Yonekura, S. Yoshida, S. E. Zabierowski, P. G. Zhang, X. Zhao, S. Zucchelli, K. M. Summers, H. Suzuki, C. O. Daub, J. Kawai, P. Heutink, W. Hide, T. C. Freeman, B. Lenhard, V. B. Bajic, M. S. Taylor, V. J. Makeev, A. Sandelin, D. A. Hume, P. Carninci, and Y. Hayashizaki.

A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470, 27 Mar. 2014.

10. M. Farlik, N. C. Sheffield, A. Nuzzo, P. Datlinger, A. Schönegger, J. Klughammer, and C. Bock. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep.*, 10(8):1386–1397, 3 Mar. 2015.
11. A. P. Feinberg. Epigenetic stochasticity, nuclear structure and cancer: the implications for medicine. *J. Intern. Med.*, 276(1):5, July 2014.
12. J. Feng, Y. Zhou, S. L. Campbell, T. Le, E. Li, J. D. Sweatt, A. J. Silva, and G. Fan. Dnmt1 and dnmt3a maintain DNA methylation and regulate synaptic function in adult forebrain neurons. *Nat. Neurosci.*, 13(4):423–430, Apr. 2010.
13. S. F. Field, D. Beraldi, M. Bachman, S. K. Stewart, S. Beck, and S. Balasubramanian. Accurate measurement of 5-methylcytosine and 5-hydroxymethylcytosine in human cerebellum DNA by oxidative bisulfite on an array (OxBS-array). *PLoS One*, 10(2):e0118202, 23 Feb. 2015.
14. J. A. Hackett, R. Sengupta, J. J. Zylicz, K. Murakami, C. Lee, T. A. Down, and M. A. Surani. Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. *Science*, 339(6118):448–452, 25 Jan. 2013.
15. A. Hermann, R. Goyal, and A. Jeltsch. The dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. *J. Biol. Chem.*, 279(46):48350–48359, 12 Nov. 2004.
16. H. Heyn, E. Vidal, H. J. Ferreira, M. Vizoso, S. Sayols, A. Gomez, S. Moran, R. Boque-Sastre, S. Guil, A. Martinez-Cardus, C. Y. Lin, R. Royo, J. V. Sanchez-Mut, R. Martinez, M. Gut, D. Torrents, M. Orozco, I. Gut, R. A. Young, and M. Esteller. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol.*, 17, 2016.
17. A. Hyvaarinen. Pairwise measures of causal direction in linear Non-Gaussian acyclic models. *), Tokyo, Japan, No*, 8, 2010.
18. I. Ionita-Laza, K. McCallum, B. Xu, and J. D. Buxbaum. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.*, 48(2):214–220, Feb. 2016.
19. R. A. Irizarry, C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, and Others. The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, 41(2):178–186, 2009.
20. R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, 33 Suppl:245–254, Mar. 2003.
21. S.-G. Jin, S. Kadam, and G. P. Pfeifer. Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res.*, 38(11):e125, June 2010.
22. R. L. Jirtle and M. K. Skinner. Environmental epigenomics and disease susceptibility. *Nat. Rev. Genet.*, 8(4):253–262, Apr. 2007.
23. P. A. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, 13(7):484–492, 29 May 2012.

24. T.-K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman, and M. E. Greenberg. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187, 14 Apr. 2010.

25. A. Kozlenkov, P. Roussos, A. Timashpolsky, M. Barbu, S. Rudchenko, M. Bibikova, B. Klotzle, W. Byne, R. Lyddon, A. F. Di Narzo, Y. L. Hurd, E. V. Koonin, and S. Dracheva. Differences in DNA methylation between human neuronal and glial cells are concentrated in enhancers and non-CpG sites. *Nucleic Acids Res.*, 42(1):109, 1 Jan. 2014.

26. I. V. Kulakovskiy, I. E. Vorontsov, I. S. Yevshin, A. V. Soboleva, A. S. Kasianov, H. Ashoor, W. Ba-Alawi, V. B. Bajic, Y. A. Medvedeva, F. A. Kolpakov, and V. J. Makeev. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res.*, 44(D1):D116–25, 4 Jan. 2016.

27. C. Ladd-Acosta and M. D. Fallin. The role of epigenetics in genetic and environmental epidemiology. *Epigenomics*, 8(2):271–283, Feb. 2016.

28. B. J. Lesch and D. C. Page. Poised chromatin in the mammalian germ line. *Development*, 141(19):3619, Oct. 2014.

29. S. Mamrut, H. Harony, R. Sood, H. Shahar-Gold, H. Gainer, Y.-J. Shi, L. Barki-Harrington, and S. Wagner. DNA methylation of specific CpG sites in the promoter region regulates the transcription of the mouse oxytocin receptor. *PLoS One*, 8(2):e56869, 18 Feb. 2013.

30. Y. A. Medvedeva, A. M. Khamis, I. V. Kulakovskiy, W. Ba-Alawi, M. S. I. Bhuyan, H. Kawaji, T. Lassmann, M. Harbers, A. R. R. Forrest, and V. B. Bajic. Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics*, 15:119, 2014.

31. D. M. Messerschmidt, B. B. Knowles, and D. Solter. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes Dev.*, 28(8):812–828, 2014.

32. C. A. Miller and J. D. Sweatt. Covalent modification of DNA regulates memory formation. *Neuron*, 53(6):857–869, 15 Mar. 2007.

33. F. Pacchierotti and M. Spanò. Environmental impact on DNA methylation in the germline: State of the art and gaps of knowledge. *Biomed Res. Int.*, 2015:123484, 3 Aug. 2015.

34. C. J. Petell, L. Alabdi, M. He, P. S. Miguel, R. Rose, and H. Gowher. An epigenetic switch regulates de novo DNA methylation at a subset of pluripotency gene enhancers during embryonic stem cell differentiation. *Nucleic Acids Res.*, 44(16):7605, 19 Sept. 2016.

35. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, 20(1):110–121, Jan. 2010.

36. L. Rinaldi, D. Datta, J. Serrat, L. Morey, G. Solanas, A. Avgustinova, E. Blanco, J. I. Pons, D. Matallanas, A. Von Kriegsheim, L. Di Croce, and S. A. Benitah. Dnmt3a and dnmt3b associate with enhancers to regulate human epidermal stem cell homeostasis. *Cell Stem Cell*, 19(4):491–501, 6 Oct. 2016.

37. Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 19 Feb. 2015.

38. K. D. Robertson. DNA methylation and human disease. *Nat. Rev. Genet.*, 6(8):597–610, Aug. 2005.

39. J. V. Sanchez-Mut and J. Gräff. Epigenetic alterations in alzheimer's disease. *Front. Behav. Neurosci.*, 9:347, 17 Dec. 2015.

40. X. Shawn Liu, H. Wu, X. Ji, Y. Stelzer, X. Wu, S. Czauderna, J. Shu, D. Dadon, R. A. Young, and R. Jaenisch. Editing DNA methylation in the mammalian genome. *Cell*, 167(1):233–247.e17, 22 Sept. 2016.

41. D. Shlyueva, G. Stampfel, and A. Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, 15(4):272–286, Apr. 2014.

42. R. Shoemaker, J. Deng, W. Wang, and K. Zhang. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.*, 20(7):883–889, July 2010.

43. P. Stepper, G. Kungulovski, R. Z. Jurkowska, T. Chandra, F. Krueger, R. Reinhardt, W. Reik, A. Jeltsch, and T. P. Jurkowski. Efficient targeted DNA methylation with chimeric dCas9-Dnmt3a-Dnmt3L methyltransferase. *Nucleic Acids Res.*, 29 Nov. 2016.

44. R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B.-K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 6 Sept. 2012.

45. E. M. Tomazou and A. Meissner. Epigenetic regulation of pluripotency. In *Advances in Experimental Medicine and Biology*, pages 26–40. 2010.

46. M. S. Turker. Gene silencing in mammalian cells and the spread of DNA methylation. *Oncogene*, 21(35):5388–5393, 12 Aug. 2002.

47. H. Wang, M. T. Maurano, H. Qu, K. E. Varley, J. Gertz, F. Pauli, K. Lee, T. Canfield, M. Weaver, R. Sandstrom, R. E. Thurman, R. Kaul, R. M. Myers, and J. A. Stamatoyannopoulos. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.*, 22(9):1680–1688, Sept. 2012.

48. U. Wüllner, O. Kaut, L. deBoni, D. Piston, and I. Schmitt. DNA methylation in parkinson's disease. *J. Neurochem.*, 139 Suppl 1:108–120, Oct. 2016.

49. Y. Zhang, C. Rohde, R. Reinhardt, C. Voelcker-Rehage, and A. Jeltsch. Non-imprinted allele-specific DNA methylation on human autosomes. *Genome Biol.*, 10(12):R138, 3 Dec. 2009.

50. J. Zhong, G. Agha, and A. A. Baccarelli. The role of DNA methylation in cardiovascular risk and disease: Methodological aspects, study design, and data analysis for epidemiological studies. *Circ. Res.*, 118(1):119–131, 8 Jan. 2016.