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[5,6,7], Vladimir B. Bajic[1], Yulia A. Medvedeva[2,6,8]

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, Russian Federation

3 Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow 119991, Russian Federation

4 Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Moscow 127051, Russian Federation

5 Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, California 90095, USA

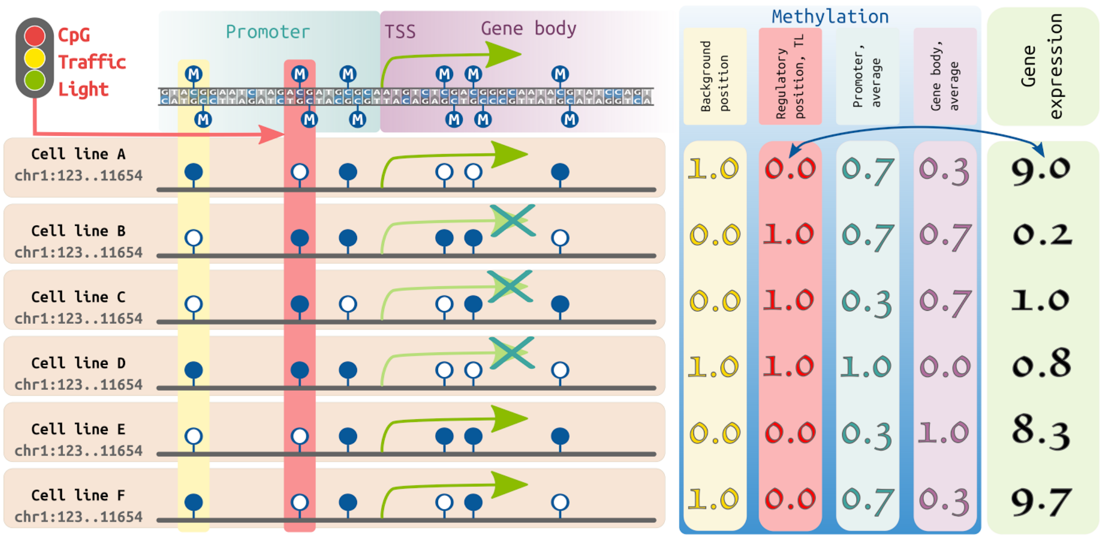
6 Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region 141701, Russian Federation

7 Immanuel Kant Baltic Federal University, Kaliningrad 236041, Russian Federation

8 Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow 119991, Russian Federation

Epigenetic regulation of gene expression attracts a lot of research attention with cytosine methylation being the most well-investigated mechanism. Contemporary methods based on bisulfite sequencing allow one to study DNA methylation with single cytosine resolution. However, at the step of downstream bioinformatic analysis, methylation levels of several dozens of cytosines are usually averaged with the aim to increase statistical power. Yet, experimental evidence show that the methylation levels of single CpG dinucleotides can affect the expression, for example in case of ESR1 gene. [1] Recently, we have shown that methylation levels of particular single CpGs are tightly linked to expression for specific cases. [2] We have called such positions CpG traffic lights (CpG TL).

Consider a specific locus containing one gene and promoter area for 6 cell lines (Fig.1). 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. A yellow column on the left panel 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 CpG TL (shown in red), methylation level significantly correlates with gene expression.



	Background position	Regulatory position, TL	Promoter, average	Gene body, average	Gene expression
Cell line A chr1:123..11654	1.0	0.0	0.7	0.3	9.0
Cell line B chr1:123..11654	0.0	1.0	0.7	0.7	0.2
Cell line C chr1:123..11654	0.0	1.0	0.3	0.7	1.0
Cell line D chr1:123..11654	1.0	1.0	1.0	0.0	0.8
Cell line E chr1:123..11654	0.0	0.0	0.3	1.0	8.3
Cell line F chr1:123..11654	1.0	0.0	0.7	0.3	9.7

Fig. 1 CpG traffic lights determination.

In our analysis we used 40 different human cell types for which whole genome bisulfite sequencing (WGBS) and RNA-seq were available. We define CpG traffic lights as CpG dinucleotides with significant Spearman correlation coefficient between DNA methylation and expression levels of a neighbouring gene (multiplicity testing correction was performed by Benjamini-Hochberg procedure for the total number of position-gene pairs). We also calculated a causality score between methylation and expression profiles to computationally assess the pairwise causal direction between these two variables.Moreover, we aimed to explore enrichment with CpG TL inside various genomic regions. For every CpG TL position we selected a random background CpG position (CpG BG) with not more than 5% difference for both GC- and CpG contents, as some genomic annotations are sensitive to these contents. We annotated all CpG positions with genomic features. For each feature we calculated the number of CpG TL and background positions located within the annotation. To test the significance of the overrepresentation we used the exact Fisher test.

We show that the average methylation level of promoter/gene body less frequently demonstrate a significant correlation with expression compared to the methylation level of CpG TL, even after a proper multiplicity correction.

We find that the CpG TL are intermediately methylated in many cell types and the analysis of hydroxymethylcytosine (5hmC) levels supports the idea of dynamic methylation of CpG TL.

The causality analysis show that those CpG TL where expression determines methylation are enriched with 5hmC, suggesting a positive feedback loop of the active transcription that activates DNA demethylation.

To address functionality of CpG TL, we investigate their evolutionary conservation. CpG TL are enriched with conserved positions both in mammals and in primates. To specify the functional role of CpG TL we tested different genomic markups for the overrepresentation. CpG TL are enriched in all known promoters types and in corresponding chromatin states, spiking at exact transcription start site, measured by CAGE. Among all promoter types, bivalent/poised promoters are especially enriched for CG TL. Our data also show that CpG TL are over-represented in regulatory regions, yet the strongest enrichment is observed in enhancers, determined both by CAGE and by chromatin modifications.

In this work we demonstrate that CpG Traffic Lights are enriched in regulatory regions, including poised/bivalent promoters and enhancers.The mechanism of CpG traffic lights provide a promising insight into enhancer activity and gene regulation linking methylation of single CpG to expression.

1. 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.

2. 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.