



OPEN

# A Method for Identification of the Methylation Level of CpG Islands From NGS Data

Leonid A. Uroshlev<sup>1,2✉</sup>, Eldar T. Abdullaev<sup>3</sup>, Iren R. Umarova<sup>4</sup>, Irina A. Il'icheva<sup>1</sup>, Larisa A. Panchenko<sup>4</sup>, Robert V. Polozov<sup>5</sup>, Fyodor A. Kondrashov<sup>6</sup>, Yuri D. Nechipurenko<sup>1✉</sup> & Sergei L. Grokhovsky<sup>1,7</sup>

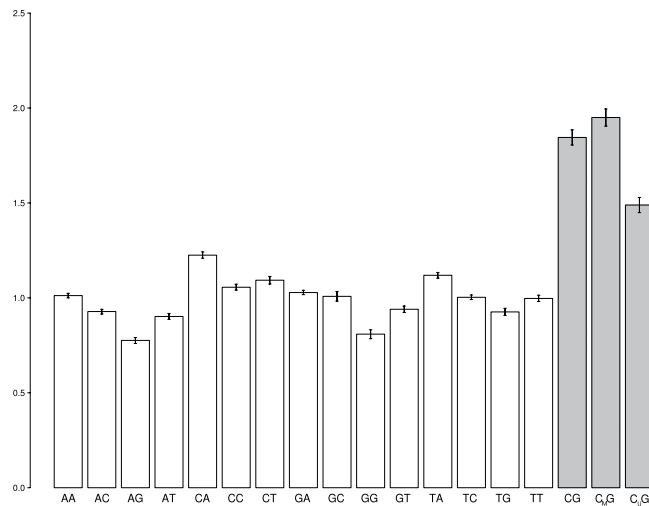
In the course of sample preparation for Next Generation Sequencing (NGS), DNA is fragmented by various methods. Fragmentation shows a persistent bias with regard to the cleavage rates of various dinucleotides. With the exception of CpG dinucleotides the previously described biases were consistent with results of the DNA cleavage in solution. Here we computed cleavage rates of all dinucleotides including the methylated CpG and unmethylated CpG dinucleotides using data of the Whole Genome Sequencing datasets of the 1000 Genomes project. We found that the cleavage rate of CpG is significantly higher for the methylated CpG dinucleotides. Using this information, we developed a classifier for distinguishing cancer and healthy tissues based on their CpG islands statuses of the fragmentation. A simple Support Vector Machine classifier based on this algorithm shows an accuracy of 84%. The proposed method allows the detection of epigenetic markers purely based on mechanochemical DNA fragmentation, which can be detected by a simple analysis of the NGS sequencing data.

DNA methylation level of CpG islands, genomic sequences with a high occurrence of methylated CpG dinucleotides, is an important regulator of gene expression. The level of gene expression can increase or decrease, depending on the methylation level in the CpG sites inside a CpG island. Analysis of methylation levels in regulatory regions of various genes can provide information on their involvement in the development of various diseases, including cancer.

Previously we showed that sonication of restriction DNA fragments leads to sugar-phosphate backbone breaks, which depend on the nucleotide sequence. Breaks in CpG dinucleotides occur about 1.5 times more often than in other dinucleotides<sup>1</sup>. Then we analyzed the genomic reads from Next-Generation Sequencing (NGS) data and showed that fragmentation methods based on the action of the hydrodynamic forces on DNA produces a similar bias in the cleavage rates<sup>2</sup> (relative frequencies of dinucleotide breaks denoted here as a cleavage rates). The only discrepancy is in the cleavage rate of CpG dinucleotides. We assume that this result can be explained by methylation of cytosines<sup>2</sup>. Recently it was shown that in the honeybee genome methylated CpG dinucleotides break more frequently than unmethylated ones<sup>3</sup>.

In this work the cleavage rates for methylated and unmethylated CpG dinucleotides of human genome were estimated. We found that the cleavage rate for methylated CpG dinucleotides is about 1.5 times higher than that for unmethylated ones. On the basis on this observation one can estimate the CpG methylation level in CpG islands without any experimental data on DNA methylation. Further, we show that tumor and healthy tissues differ significantly in the methylation status of CpG islands. In human somatic cells approximately 80% of CpG dinucleotides are methylated. Bisulfite sequencing was the first method for detection of cytosine methylation<sup>4</sup> (see also<sup>5,6</sup>). However, sample preparation for bisulfite sequencing and subsequent data processing is overly expensive of time. On the basis of our results, the criterion can be developed, by which the NGS data can be used for defining the total level of the CpG methylation in the given cell type without any additional experiments.

<sup>1</sup>Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia. <sup>2</sup>Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia. <sup>3</sup>Max Planck Institute for Molecular Genetics, Berlin, Germany. <sup>4</sup>Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russia. <sup>5</sup>Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, Puschino, Russia. <sup>6</sup>Institute of Science and Technology Austria, Klosterneuburg, Austria. <sup>7</sup>Sergei L. Grokhovsky is deceased. ✉e-mail: [leoniduroshlev@gmail.com](mailto:leoniduroshlev@gmail.com); [yurii.nechipurenko@gmail.com](mailto:yurii.nechipurenko@gmail.com)



**Figure 1.** Average cleavage rates  $r(XY)$  of 18 dinucleotides (namely, 16 common dinucleotides along with methylated CpGs ( $C_MG$ ) and unmethylated CpGs ( $C_UG$ )). Cleavage rates are plotted on Y axis, dinucleotides are listed on X axis. CpG methylation results in a substantial increase of the cleavage rates in comparison with unmethylated CpGs and other dinucleotides.

## Results

We randomly selected 100 whole genome sequencing datasets from the 1000 Genomes project that contained reads mapped on a reference genome (GRCh37). According to<sup>7</sup>, these samples were prepared by Covaris ultrasonic DNA shearing. On the basis of the bisulfite-sequencing data obtained from NGSmethDB<sup>8</sup>, we identified the methylation status for all CpG dinucleotides in lymphoblastoid cell line. We filtered only those CpG dinucleotides that were reliably classified as methylated or not (see Methods). On the basis of 5'-read coordinates of each read in each dataset, cleavage rates for all dinucleotides were computed in the following way:

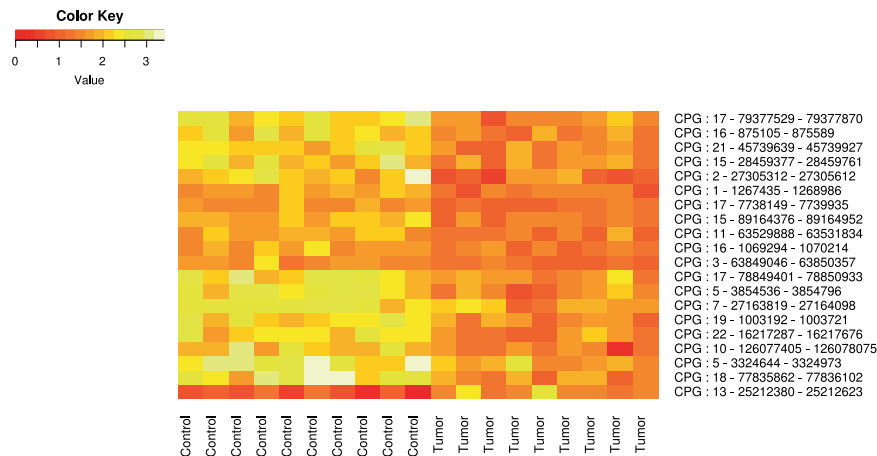
$$r(XY) = \frac{n(XY)}{N * p(XY)}, \quad (1)$$

where  $X, Y = \{A, T, G, C_U \text{ (unmethylated cytosine), } C_M \text{ (methylated cytosine)}\}$ ;  $n(XY)$  is an overall number of reads in the dataset with the 5'-end nucleotide, in which  $X$  precedes  $Y$  in the reference genome;  $N$  is an overall number of reads;  $p(XY)$  is an average fraction of  $XY$  dinucleotides in the genome fragment of length 200 bp around read start ( $i$ ) position<sup>2</sup>. Besides, we computed the average cleavage rates of methylated and unmethylated CpG dinucleotides separately (see Supplementary Tables 3 and 4). As a result,  $r(XY)$  values, where  $XY$  denotes 16 common dinucleotides + methylated CpGs ( $C_MG$ ) and unmethylated CpGs ( $C_UG$ ), were calculated on the bases of 100 randomly selected datasets (Fig. 1).

The results of the ANOVA, Kruskal-Wallis test analysis suggest that the effect of the dinucleotide type on the relative frequency of its ultrasonic cleavage is statistically significant at the level  $p < \alpha = 0.05$  (see Supplementary Table 4). The results of the multiple comparisons, Duncantest and Kruskal-Wallis test, are shown in the Tables (see Supplementary Tables 2 and 3). The sample means in the Table 2 Suppl. are combined into subsets so that the means from different subsets (columns 1, 2, 7, 8, 9, 10) differ statistically significantly at the  $p < 0.05$ . Therefore the mean of  $r(C_MG)$  is the largest of all three values.

According to<sup>9</sup>, tumor cells have a very specific landscape of methylation, especially in apoptosis-related genes. It is noteworthy that CpG methylation is a stochastic process; therefore, only an average methylation of CpG islands matters, not a single dinucleotide status<sup>10</sup>. We can compute the cleavage rates in each CpG island separately and predict if CpG island is hypo- or hypermethylated. Comparison of the CpG methylation levels in CpG islands of specific genes may allow one to distinguish between healthy and cancerous tissues.

We intended to develop an approach that will make it possible to distinguish tumor tissue samples from normal ones by the dinucleotide cleavage characteristics. Our approach is based on the observation that methylated CpG islands have higher cleavage rates. Namely, the reads with the 5'-end nucleotide G preceded by C in genomic sequence will occur more frequently if C is methylated. For practical evaluation of this method, the data from EGA datasets was used (24 datasets of T-cell lymphoma, 15 high-coverage datasets of prostate cancer, big dataset ( $> 300$  samples) of breast cancer, 24 datasets of hepatocarcinoma, and 37 datasets of medulloblastoma (average coverage  $\sim 40\times$ ) from ICGC database, as well as control datasets of healthy tissues). We selected a subset of EGA T-cell datasets with a good coverage ( $> 40\times$ ). Our approach allows the prediction of the difference between methylation statuses of CpG islands in healthy and tumor tissues. Figure 2 shows the CpG islands with the biggest absolute difference in their mean cleavage levels. In almost all cases we observed demethylation effect in the CpG islands of tumor tissues with one exception: the CpG island with coordinates chr13:-25,212,380–25,212,623, which could be associated with promoter region of the pseudogene TNFRSF1A. This gene is a well known oncogene involved in tumorigenesis<sup>11</sup>. We suggest that the unmethylated CpG island in the promoter region could lead to expression of the pseudogene thus promoting oncogenesis; however we have not validated this mechanism experimentally.



**Figure 2.** Colormaps for the CpG islands with highest difference of cleavage rates between normal and tumor tissues in medulloblastoma. X axis corresponds to samples, and Y axis, to CpG islands with their coordinates in the genome. The intensity of the color scale corresponds to the mean cleavage rate for CpG in every CpG island, which was calculated according to (1).

Cancer type	EGA ID	Percent of true positive cases	Percent of true negative cases	Percent of false positive cases	Percent of false negative cases
Breast Cancer	EGAD00001000126	79	60	21	40
Medulloblastoma	EGAD00001000816	76	96	24	4
T-Cell Lymphoma	EGAD00001002738	84	82	16	18
Prostate Cancer	EGAD00001000263	70	97	30	3
Hepatocarcinoma	EGAD00001001881	87	71	13	29

**Table 1.** Results for of the prediction of tumor/normal status for different cancer types calculated by the SVM model described in the main text. Percentage of positive cases and negative cases are calculated separately.

The detection of the cancer-specific methylation of particular CpG islands is a complicated task even in the presence of qualitative results of bisulfite-sequencing experiments. Therefore, for effective prediction of disease status from dinucleotide cleavage rates we employed some machine learning algorithms. At first, for each CpG island in each dataset (control and tumor) we computed an average CpG dinucleotide cleavage rate and turned every set into high dimension vector. Each element of this vector is an average CpG dinucleotide cleavage rate in a specific CpG island. These vectors were used for training of binary classifier on the basis of the support vector machine (SVM) method with linear kernel. This SVM classifier was used to predict the state of specific sample (cancer or normal). For training we used the same number of control and tumor datasets for each cancer type. In our research we applied SVM realization in `e1071` package<sup>12</sup> of R programming language. Corresponding script for samples classification based on cleavage rates of CpG islands and computation results are available on the GitHub page: <https://github.com/LeonidU/DNA-Segmentation>.

To control the accuracy of this algorithm, the jack-knife method was used. On each round of jack-knife we excluded random subset for each cancer type and trained the model on the remaining sets. Each subset contained an equal number of control and tumor sets. Then we made a prediction on the excluded subsets and computed true and false positive rates (Table 1).

Our predictions were comparable with bisulfite sequencing based classifiers in the accuracy of prediction. The accuracy of the sample status prediction in<sup>13</sup> was ~84–85%, which is comparable with our results.

## Discussion

We analyzed raw WGS data and found that cytosine methylation strongly affects the cleavage rate of CpG dinucleotides in the human genome. This fact agrees with recent observations of another group<sup>3</sup>. In all studied datasets from 1000 Genomes project the cleavage rate of methylated CpG dinucleotides was higher than the cleavage rate of other dinucleotides including unmethylated CpGs<sup>1,2</sup>. This observation was then used as a basis of the method that allows one to predict changes in the gene promoter epigenetics associated with cancer.

Using our method, one can estimate the total level of the whole-genome base methylation and compare the methylation degree in different genomic regions with similar functions at different stages of the organism development. The inferred correlation between the probability of mechanochemical cleavage of the DNA sugar-phosphate backbone and local nucleotide sequence enabled us to find an important connection between structural parameters of specific nucleotide sequences of DNA and their biological functions. Moreover, it should be possible to conduct a comparative analysis of methylation profiles for cells during differentiation and during aging by estimating the total

DNA methylation level in various cells at least at the resolution of the CpG islands. The main problem for practical use of our method is a bad coverage of CpG islands and their heterogeneity in different cancer cells. For example, worst results of tumor/control separation were obtained for breast cancer with badly covered samples

On the basis of these results, the method can be developed for detecting the total level of the CpG methylation from the raw WGS data without any additional experiments. We hope that the developed method makes it possible to recognize base modifications in genomes other than methylation of cytosines in CpG dinucleotides<sup>14</sup>. There is a large space for further development of the designed method with regard both to the types of data analyzed and to the algorithm. Higher level of coverage of the genome could also increase a resolution of our predictions. Segmentation of DNA by sonication or other methods used in NGS provide tremendous amount of data concerning such an unusual mechanical property of DNA as the susceptibility to cleavage. Our approach opens up new possibilities for studies of both DNA physics and its relationships with double helix structure and epigenetics<sup>15</sup>.

## Methods

**Protocol of data processing.** We used the whole genome sequencing samples of human individuals from the 1000 Genomes project<sup>7</sup>. The BAM files with WGS read alignments on a reference genome were downloaded from the FTP server of the 1000 Genomes project (<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/>). SAMtools suit was used to filter reads that only map on a forward strand of all autosomes of the human genome (excluding X and Y chromosomes) (`samtools view -f 35 -F 4 input.bam {1..22} > filtered.bam`)<sup>15</sup>. We also tested whether our filtration criteria affect average cleavage rates: we tested several other scenarios of filtration on 5 random datasets of the 1000 Genomes project. We checked if filtering out duplicated reads, usage of reads that only map on reverse strand, or filtering out the reads that map on CpG islands would affect average cleavage rates. The last filtering criteria were used in order to test if an increased cleavage rate of methylated CpG dinucleotides, as compared with that of unmethylated CpGs, can be explained by the difference in the CpG dinucleotide distribution in the genome. Most unmethylated CpGs are clustered in CpG islands; however, methylated CpGs can be observed both inside and outside the CpG islands. Thus, by filtering out the reads from CpG islands we compared the cleavage rates of methylated and unmethylated CpG dinucleotides from the rest of the genome. We found that the average dinucleotide cleavage rates do not depend on the filtering strategy, so we assumed that different distribution of methylated and unmethylated CpG dinucleotides in the genome did not affect our results substantially.

GRCh37 was used as a reference genome. To be sure about the type of dinucleotides observed we masked all low sequence complexity regions by RepeatMasker<sup>16,17</sup>. These frequency of XY dinucleotides was calculated not in the whole genome but in the 200-bp-long genome fragment around the read-start positions. By doing that we took into account all the sequence biases associated with read mapping positions<sup>18</sup>.

We used bisulfite-sequencing data from NGSmethDB database<sup>19</sup> for the lymphoblastoid cell line samples (<https://bioinfo2.ugr.es/NGSmethDB/methylation-maps/>) to predict methylation status of cytosine in CpG dinucleotides in the cell line that was used in the 1000 Genome project. We used only those CpG dinucleotides that were covered by more than 10 reads in bisulfite-sequencing data. We defined CpG dinucleotide as methylated or unmethylated if >90% of reads covering a specific dinucleotide contained C<sub>M</sub> or C<sub>U</sub> respectively at a corresponding position. CpG dinucleotides with some intermediate C<sub>M</sub>/C<sub>U</sub> ratios were excluded from further analysis.

Received: 30 September 2019; Accepted: 4 May 2020;

Published online: 25 May 2020

## References

- Grokhovskiy, S. L. *et al.* Sequence-specific ultrasonic cleavage of DNA. *Biophys. J.* **100**, 117–125 (2011).
- Poptsova, M. S. *et al.* Non-random DNA fragmentation in next-generation sequencing. *Sci. Rep.* **4**, 4532 (2014).
- Garafutdinov, R. R., Galimova, A. A. & Sakhabutdinova, A. R., The influence of CpG (5'-d(CpG)-3' dinucleotides) methylation on ultrasonic DNA fragmentation. *J. Biomol. Struct. Dyn.*, <https://doi.org/10.1080/07391102.2018.1533888> (2018)
- Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477 (2013).
- Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72–e72 (2012).
- Ehrlick, M. & Wang, R. Y. H. 5-Methylcytosine in eukaryotic DNA. *Science* **212**, 1350–1357 (1981).
- 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68 (2015).
- Hackenberg, M., Barturen, G. & Oliver, J. L. NGSmethDB: A database for next-generation sequencing single-cytosine - resolution DNAmethylation data. *Nucleic Acids Res.* **39**, D75–D79 (2011).
- Esteller, M. CpG island hypermethylation and tumor suppressor genes: A booming present, a brighter future. *Oncogene* **21**, 5427 (2002).
- Landan, G. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.* **44**, 1207 (2012).
- Egusquiaguirre, S. P. *et al.* The STAT3 target gene TNFRSF1A modulates the NF-κB pathway in breast Cancer cells. *Neoplasia*. **20**, 489–498 (2018).
- Dimitriadou, E. *et al.* Misc functions of the Department of Statistics (e1071). *R package*. **1**, 5–24 (2008).
- Yegnasubramanian, S. *et al.* Hypermethylation of CpG Islands in Primary and Metastatic Human Prostate Cancer. *Cancer Res.* **64**, 1975–1986 (2004).
- Semyonov D., Nechipurenko Y. Non-Canonical GC Base Pairs and Mechanochemical Cleavage of DNA <https://arxiv.org/abs/2001.03561> (2009).
- Nechipurenko, D. I. *et al.* Modeling of mechanochemical DNA cleavage by action of ultrasound. *Biofizika*. **59**, 1061–1070 (2014).
- Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
- Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0.6 2013-2015. Available at, <http://www.repeatmasker.org>. (Accessed: 20th January 2019).
- Il'icheva, I. *et al.* Structural features of DNA that determine RNA polymerase II core promoter. *BMC genomics* **17**, 973 (2016).
- Lebrón, R. *et al.* NGSmethDB 2017: Enhanced methylomes and differential methylation. *Nucleic Acids Res.* **45**, 97–103 (2017).

## Acknowledgements

Authors thank V.J.Makeev, D.Y. Nechipurenko, M.V. Khodykov, and M.V.Fridman. This work was supported by project no. 0112-2019-0001 and RFFI grants 18-07-00354 A; 17-07-01331 A, program of the Presidium of the Russian Academy of Sciences for Molecular and Cellular Biology no. 01201456592, the Russian Foundation for Fundamental Investigations project no. 14-04-01269, Program of fundamental research for state academies for 2013-2020 years, no. 01201363818.

## Author contributions

S.L.G., I.A.I. and Yu.D.N conceived the project, F.A.K. and R.V.P. designed the experiments, E.T.A., L.A.U. and I.R.U. conducted all simulations and analysis, L.A.P performed statistical analysis of the obtained data, S.L.G., I.A.I, Yu.D.N., R.V.P, E.T.A., L.A.U., L.A.P. and F.A.K. wrote and prepared the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-65406-1>.

**Correspondence** and requests for materials should be addressed to L.A.U. or Y.D.N.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020