*Bioinformatics*, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

*Data and Text Mining*

# mHapTk: A comprehensive toolkit for the analysis of DNA methylation haplotypes

Yi Ding[1,#], Kangwen Cai[2,#], Leiqin Liu[1], Zhiqiang Zhang[1], Xiaoqi Zheng[3,*] and Jiantao Shi[1,*]

[1]State Key Laboratory of Molecular Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China [2]Department of Mathematics, Shanghai Normal University, Shanghai 200234, China, [3]Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

#Equal contribution

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Summary:** Bisulfite sequencing (BS-seq) remains the gold standard technique to detect DNA methylation profiles at single-nucleotide resolution. The DNA methylation status of CpG sites on the same fragment represents a discrete methylation haplotype (mHap). The mHap-level metrics were demonstrated to be promising cancer biomarkers and explain more gene expression variation than average methylation. However, most existing tools focus on average methylation and neglect mHap patterns. Here, we present mhapTk, a comprehensive python toolkit for the analysis of DNA methylation haplotypes. It calculates eight mHap-level summary statistics in predefined regions or across individual CpG in a genome-wide manner. It identifies methylation haplotype blocks (MHBs), in which methylations of pairwise CpGs is tightly correlated. Furthermore, mHap patterns can be visualized with the built-in functions in mHapTk or external tools such as IGV and deepTools.

**Availability:** https://jiantaoshi.github.io/mhaptk/index.html

**Contact:** jtshi@sibcb.ac.cn, xqzheng@shnu.edu.cn

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

DNA methylation is an essential epigenetic regulatory mechanism that plays critical roles in many biological processes, including embryonic development (Greenberg & Bourc'his, 2019), tumorigenesis (Blewitt, Skvortsova, Stirzaker, & Taberlay, 2019), and aging (Unnikrishnan et al., 2019). Mammalian DNA methylation predominantly occurs at CpG sites. Bisulfite sequencing (BS-seq), including whole genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS), is the gold standard technique to detect DNA methylation profiles at single-nucleotide resolution. The DNA methylation status of CpG sites on the same fragment represents a discrete methylation haplotypes (mHap) (Shoemaker, Deng, Wang, & Zhang, 2010). The mHap-level metrics characterize DNA methylation patterns rather than average methylation. Based on mHap-level patterns, methylation entropy was defined to assess the variability of DNA methylation (Xie et al., 2011). Proportion of Discordant Reads (PDR) was proposed to measure intra-sample heterogeneity (Landau et al., 2014). Recently, Cellular Heterogeneity-Adjusted cLonal Methylation (CHALM) (Xu et al., 2021) and methylation concurrence ratio (MCR) (Shi et al., 2021) were demonstrated to explain gene expression variation better than average methylation. Furthermore, the mHap-level patterns show promising translational potentials. For example, DNA methylations of adjacent CpG sites were found to be co-methylated and form methylation haplotype blocks (MHBs) (S. Guo et al., 2017). This co-methylation pattern can be quantified by methylated haplotype load (MHL) and methylation block score (MBS) (Liang et al., 2021), both of which preserve higher signal to noise ratio than average methylation in early cancer detection.

However, the tools for analyzing DNA methylation haplotypes are limited. One tool of this kind is RLM, but it only calculates PDR and entropy, and does not support plot interface. Besides, it takes aligned BAM/SAM files as input, which is prohibitively large in size for large-scale BS-seq analysis (Hetzel, Giesselmann, Reinert, Meissner, & Kretzmer, 2021). Previously, we have developed a novel mHap format, which reduces the size of a BAM file by up to 140-fold while keeps all mHap-level information (Zhang et al., 2021). It is also compatible with the Tabix tool for random and fast access (H. Li, 2011). Furthermore, the mHap format file contains no genetic information and can be shared as the CpG-level mean methylation file, which poses minimal risk to an individual's privacy. Here, we further developed mHapTk, a comprehensive toolkit for the analysis of DNA methylation haplotypes based on mHap format.

## 2 mHapTk description

mHapTk takes mHap files as standard input, which can be converted from BAM files using mHapTools (Zhang et al., 2021). Functions in mHapTk can be assigned into three categories, visualization, MHB discovery, and calculation of mHap-level summary metrics (**Fig. 1**, **Supplementary Fig. 1**).

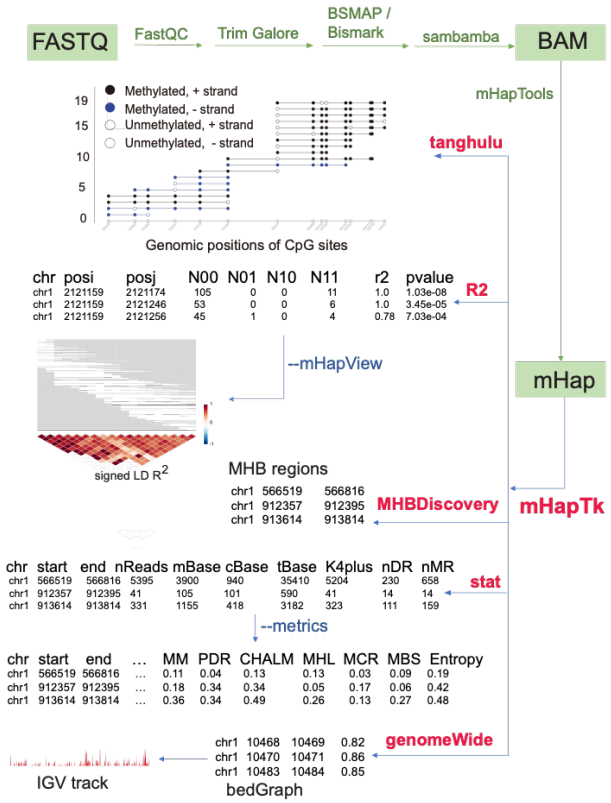


**Fig. 1** A schematic diagram of mHapTk. The preprocessing steps are shown in green, which output mHap files that are used as standard input in mHapTk. There are 5 sub-commands in mHapTk, including 'tanghulu', 'R2', 'MHBDiscovery', 'stat' and 'genomeWide'. The example outputs of each command are shown in a concise way.

For a given region, mHapTk visualizes the read-level methylation statuses as a tanghulu plot (W. Guo et al., 2018) (**Supplementary Fig. 2A**). Reads with the same methylation pattern can be optionally stacked with its occurrence number shown beside (**Supplementary Fig. 2B**). For a region with large number of reads, a given number (20 by default) of mHaps can be simulated to maximize the likelihood given the observed sequencing reads (**Supplementary Fig. 2C**). Alternatively, mHap-level information can be shown as a heatmap (**Supplementary Fig. 2D, upper panel**). The co-methylation levels of pairwise CpGs are measure by linkage disequilibrium (LD) $R^2$, calculated from individual reads rather than mean methylation (S. Guo et al., 2017). Note that we used signed $R^2$ to distinguish positive and negative correlations (**Supplementary Fig. 2D, lower panel**). The combination of these two plots is termed mHapView in mHapTk. It also implemented a *de novo* MHB discovery tool that identifies locally co-methylated regions across the genome. Using a public dataset of esophageal squamous cell carcinoma as an example (Cao et al., 2020), mHapTk identified 11,112 MHBs, which can be potentially used for non-invasive cancer detection (**Supplementary Fig. 3**). For a set of regions, typical defined by a BED file, mHapTk calculates eight mHap-level summary statistics, i.e., average methylation, CHALM, PDR, MHL, MCR, Entropy, MBS, and signed LD $R^2$ (**Supplementary Table 1**). Furthermore, the above mHap-level metrics can also be calculated in terms of individual CpG sites across the genome, resulting in files in bedGraph format, which can be used in combination with IGV (Thorvaldsdottir, Robinson, & Mesirov, 2013), pyGenomeTracks (Lopez-Delisle et al., 2021) and WashU browser (D. Li et al., 2022) for visualization. For instance, CASC9 is upregulated in esophageal cancer, which is potentially explained by decreased CHALM, PDR, and MBS, as well as the presence of MHBs in the promoter region (**Supplementary Fig. 4**). The bedGraph files generated by mHapTk can be converted to bigWig files and used by deepTools (Ramirez, Dundar, Diehl, Gruning, & Manke, 2014) for visualization (**Supplementary Fig. 5**). Example outputs of mHapTk have been described with more details in **Supplementary Table 1-6.**

## 3 Application to real datasets

We used mHapTk to explore potential association between DNA methylation patterns and gene expression in lung cancer cell lines from the CCLE dataset. we focused on promoters with significant changes only in mHap-level metrics but not mean methylation. For instance, promoters were assigned into different groups according to changes of mean methylation and changes of DNA methylation entropy between two subtypes of lung cancer, i.e., non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) (**Supplementary Fig. 6**). Specifically, four groups were defined: gene promoters with significant changes in both mean methylation and entropy (**Supplementary Fig. 6A**), significant changes in entropy only (**Supplementary Fig. 6B**), significant changes in mean methylation only (**Supplementary Fig. 6C**) and those with no significant changes in either mean methylation or entropy (**Supplementary Fig. 6D**). Interestingly, association between DNA methylation entropy and gene expression are statistically significant regardless the change of mean methylation (Odd ration = 610.64, p-value < 2.2e-16). Besides entropy, PDR, CHALM, and MBS also explain gene expression variation independent of mean methylation (**Supplementary Fig. 7-12**). These results demonstrate that mHapTk has the potential to uncover novel association between DNA methylation patterns and gene expression. Finally, we benchmarked the running time of mHapTk and showed that it was computationally efficient for typical WGBS samples (**Supplementary Fig. 13**).

## 4 Conclusion

Here we present mHapTk, a novel software for manipulating mHap data. Using mHap format data as standard inputs, it separates the steps of pre-processing and data mining when dealing with BS-seq data. Coupled with mHapTools, it streamlines the analysis of DNA methylation haplotypes. Given the tools in mHapTk, it will have broad application in fields of gene regulation and biomarker discovery.

## Funding

J.S. is a recipient of the Hundred Talents Program Award of the Chinese Academy of Sciences. This study is supported by the Shanghai Pujiang Program (20PJ1414700 to J.S.) and National Natural Science Foundation of China (61972257 to X.Z.).

## References

Blewitt, M., Skvortsova, K., Stirzaker, C., & Taberlay, P. (2019). The DNA methylation landscape in cancer. *Essays in Biochemistry, 63*(6), 797-811. doi:10.1042/ebc20190037

Cao, W., Lee, H., Wu, W., Zaman, A., McCorkle, S., Yan, M., . . . Bivona, T. G. (2020). Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma. *Nat Commun, 11*(1), 3675. doi:10.1038/s41467-020-17227-z

Greenberg, M. V. C., & Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol, 20*(10), 590-607. doi:10.1038/s41580-019-0159-6

Guo, S., Diep, D., Plongthongkum, N., Fung, H. L., Zhang, K., & Zhang, K. (2017). Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat Genet, 49*(4), 635-642. doi:10.1038/ng.3805

Guo, W., Zhu, P., Pellegrini, M., Zhang, M. Q., Wang, X., & Ni, Z. (2018). CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. *Bioinformatics, 34*(3), 381-387. doi:10.1093/bioinformatics/btx595

Hetzel, S., Giesselmann, P., Reinert, K., Meissner, A., & Kretzmer, H. (2021). RLM: Fast and simplified extraction of Read-Level Methylation metrics from bisulfite sequencing data. *Bioinformatics*. doi:10.1093/bioinformatics/btab663

Landau, D. A., Clement, K., Ziller, M. J., Boyle, P., Fan, J., Gu, H., . . . Wu, C. J. (2014). Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell, 26*(6), 813-825. doi:10.1016/j.ccell.2014.10.012

Li, D., Purushotham, D., Harrison, J. K., Hsu, S., Zhuo, X., Fan, C., . . . Wang, T. (2022). WashU Epigenome Browser update 2022. *Nucleic Acids Res*. doi:10.1093/nar/gkac238

Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics, 27*(5), 718-719. doi:10.1093/bioinformatics/btq671

Liang, N., Li, B., Jia, Z., Wang, C., Wu, P., Zheng, T., . . . Zhang, Z. (2021). Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning. *Nat Biomed Eng, 5*(6), 586-599. doi:10.1038/s41551-021-00746-5

Lopez-Delisle, L., Rabbani, L., Wolff, J., Bhardwaj, V., Backofen, R., Gruning, B., . . . Manke, T. (2021). pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics, 37*(3), 422-423. doi:10.1093/bioinformatics/btaa692

Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A., & Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res, 42*(Web Server issue), W187-191. doi:10.1093/nar/gku365

Shi, J., Xu, J., Chen, Y. E., Li, J. S., Cui, Y., Shen, L., . . . Li, W. (2021). The concurrence of DNA methylation and demethylation is associated with transcription regulation. *Nat Commun, 12*(1), 5285. doi:10.1038/s41467-021-25521-7

Shoemaker, R., Deng, J., Wang, W., & Zhang, K. (2010). Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res, 20*(7), 883-889. doi:10.1101/gr.104695.109

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform, 14*(2), 178-192. doi:10.1093/bib/bbs017

Unnikrishnan, A., Freeman, W. M., Jackson, J., Wren, J. D., Porter, H., & Richardson, A. (2019). The role of DNA methylation in epigenetics of aging. *Pharmacol Ther, 195*, 172-185. doi:10.1016/j.pharmthera.2018.11.001

Xie, H., Wang, M., de Andrade, A., Bonaldo Mde, F., Galat, V., Arndt, K., . . . Soares, M. B. (2011). Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Res, 39*(10), 4099-4108. doi:10.1093/nar/gkr017

Xu, J., Shi, J., Cui, X., Cui, Y., Li, J. J., Goel, A., . . . Li, W. (2021). Cellular Heterogeneity-Adjusted cLonal Methylation (CHALM) improves prediction of gene expression. *Nat Commun, 12*(1), 400. doi:10.1038/s41467-020-20492-7

Zhang, Z., Dan, Y., Xu, Y., Zhang, J., Zheng, X., & Shi, J. (2021). The DNA methylation haplotype (mHap) format and mHapTools. *Bioinformatics*. doi:10.1093/bioinformatics/btab458