DNA Methylation (DM) data format and DMtools for efficient DNA methylation data storage and analysis

Qiangwei Zhou[1,2,†,*], Cong Zhou[1,2,†], Zhixian Zhu[1,2], Yuanhui Sun[1,2], Guoliang Li[1,2,*]

[1] National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan 430070, China

[2] Agricultural Bioinformatics Key Laboratory of Hubei Province, Hubei Engineering Technology Research Center of Agricultural Big Data, Key Laboratory of Smart Farming Technology for Agricultural Animals, 3D Genomics Research Center, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

† These authors contributed equally to this work.
* Corresponding author. Guoliang Li and Qiangwei Zhou
Email: guoliang.li@mail.hzau.edu.cn; qwzhou@mail.hzau.edu.cn

Abstract

DNA methylation is a critical epigenetic mechanism that plays a pivotal role in various biological processes. Currently, larger datasets from whole-genome bisulfite sequencing for DNA methylation pose challenges throughout the computational analysis pipeline, including storage and memory constraints. Unfortunately, storage formats and analysis tools have not kept pace with these increased resource demands. In this study, we present a new and efficient design for storing DNA methylation (DM) data after mapping in compressed binary indexed DM format. Our format significantly reduces storage space by 80%-95% compared to commonly used file formats for DNA methylation data after mapping. To enhance the processing of DNA methylation data in DM format, we have developed DMtools, a comprehensive toolkit that offers utilities such as rapid and random access, computation of DNA methylation profiles across genes, and analysis of differential DNA methylation. The analysis speed is improved by over 100 times compared to existing methods. Furthermore, we have created pyDMtools, a Python package that efficiently processes DM format files for Python users. The integration of the DM format and its associated tools represents

significant progress in handling and exploring DNA methylation data, offering the potential to significantly reduce storage needs and improve downstream analysis capabilities.

Introduction

DNA methylation is a critical epigenetic mechanism that plays a pivotal role in various biological processes, such as cell differentiation (1,2), development (3), aging (4), genomic imprinting (5), X chromosome inactivation (6) and other important biological processes (7). Aberrant DNA methylation can lead to tumor formation (8,9). Whole Genome Bisulfite Sequencing (WGBS) is a powerful method for precisely detecting DNA methylation at a single-base resolution across the entire genome. It achieves this by converting nonmethylated cytosines into thymines, significantly advancing the study of DNA methylation (10).

Currently, widely-utilized tools in DNA methylation data analysis include BatMeth2, Bismark, BS-Seeker, and BSMAP (11-14). Although efficient in handling DNA methylation data alignment and methylation level calculations, these tools store output files in various text formats. This storage format poses challenges in terms of extensive storage space, slow downstream analysis speed, and difficulty in extracting specific methylation levels for a chromosome region or gene, as well as computing methylation profiles for particular genes. This problem becomes more marked when dealing with multiple-sample data, leading to prolonged analysis times and increased storage pressure. For example, the National Center for Biotechnology Information (NCBI) already possesses over 5500 sets of WGBS data, and there are also several DNA methylation databases such as ASMdb (15) and MethBank (16) that include thousands of bisulfite sequencing (BS-Seq) datasets, imposing substantial computational and storage burdens. Therefore, there is an urgent need for a universal compressed binary index format and a downstream analysis tool that can support all DNA methylation alignment software, minimize storage space requirements, and accelerate downstream DNA methylation data analysis.

In the realm of DNA methylation data analysis, our introduced DNA Methylation (DM) data format and the accompanying DMtools tackle pivotal challenges related to file size, downstream analysis speed, and functionality. Notably, DMtools is a pioneering manipulation tool tailored specifically for the binary-indexed DNA methylation data format. The DM format files produced by our approach demonstrate a significantly reduced size, approximately 5%-20% compared to commonly used DNA methylation formats. While recent tools like CGmaptools(17) and BAllCools(18) have proposed CGbz and BAllC compression formats, respectively, it is noteworthy that CGmaptools still relies on downstream analysis provided by text formats, and BAllCools lacks downstream analysis functionality. In contrast, our work achieves superior compression ratios, accelerates downstream computation by over a hundredfold, and provides a comprehensive solution for efficient storage and analysis of DNA methylation data.

DM format and DMtools

The DM format is designed for efficient storage and swift querying of DNA methylation data. It encompasses DNA methylation details, and indexing information, as illustrated in Figure 1A. This format effectively stores genome-wide DNA methylation levels at single base/region resolution, providing comprehensive information (detailed in Table 1). The inclusion of genome location indexing enables rapid retrieval of DNA methylation results for any genomic region. These features ensure quick localization of target regions with less time complexity, streamlining DNA methylation queries through expeditious data parsing and access.

We have further developed DMtools, a powerful C++ tool designed to facilitate the manipulation of DNA methylation files in the DM format (Figure 1B). It offers seamless operations on DM files and a comprehensive range of downstream analysis functions. DMtools can convert data from popular analysis software like BatMeth2, Bismark, BSMAP and bedMethyl into the DM format (Figure 1C), ensuring compatibility and flexibility. DMtools also offers a comprehensive range of downstream analysis functions. These include the assessment of bisulfite conversion

rate, rapid calculation of DNA methylation levels for genomic regions and gene methylation profiles, as well as robust differential methylation analysis (Figure 1B).

Table 1. Basic information in DM format and DMtools

Order	Name	Description	Type
1	CHROM	Reference name	unit32_t
2	START	1-based start of chromosome	unit32_t
3	END (Optional)	End of the cytosine or region	unit32_t
4	ML	DNA methylation level [0-1]	unit32_t
5	STRAND	Strand for cytosine DNA methylation [+-.]	unit8_t
6	CONTEXT	DNA methylation Context [mC/CG/CHG/CHH]	unit8_t
7	COVERAGE	Coverage of the cytosine or region	unit16_t
8	ID (Optional)	Single-cell/Gene ID	unit32_t

Compared to traditional tab-delimited text file formats for DNA methylation data, such as BatMeth2 format, Bismark format, BSMAP format and bedMethyl, the DM format takes up less storage space (reduces about 80-95%) (Figure 1D and Supplementary Fig. S1). For querying and computing in any specific region, as shown in Figure 1A, DM format determine the precise location of the queried DNA methylation information in the file based on the index (Step 1). It then retrieves the detailed DNA methylation information (Steps 2-3) and calculates the DNA methylation results for the region. With the DM format, query and computation speed can be significantly accelerated. Notably, the utilization of DMtools for computing regional DNA methylation levels and gene profiles has achieved a remarkable speed improvement of over a hundred-fold (Figure 1E).

Except for its excellent analytical capabilities, DMtools offers a variety of DNA methylation calculation and visualization features. These include the sample correlation analysis (Supplementary Fig. S2, S3), calculation of coverage information (Supplementary Fig. S4), categorization of DNA methylation level (Supplementary Fig. S5), generation of DNA methylation profiles upstream and downstream of genes or transcription start sites (Supplementary Fig. S6-S8), and the generation of heatmaps displaying DNA methylation for all genes (Supplementary Fig. S9-S11). Moreover, for differential DNA methylation analysis (Supplementary Fig. S12),

DMtools require only about 10% of the memory compared to methylKit (19), making it a reliable and efficient choice for DNA methylation research (Supplementary Fig. S13-S16). To facilitate the efficient handling of DM format files by users, we also provide a Python package called pyDMtools.

DMtools software and datasets used for demonstration

DMtools is a C++ package that utilize the htslib and libBigWig libraries to parse and manipulate DNA methylation levels in the DM format. With DMtools, users can convert sorted BAM files of read alignments, as well as other DNA methylation formats like Bismark or bedMethyl methylation formats, into DM format. Additionally, we have developed a Python package, pyDMtools, based on DMtools that enables users to read and process DM files in Python. Detailed instructions on how to use DMtools and the pyDMtools Python package for reading and processing DM files are readily available on the DMtools website (https://dmtools-docs.readthedocs.io/en/latest/index.html). Installing pyDMtools is a straightforward process using either 'pip3 install pyDMtools' or 'conda install -c bioconda pyDMtools'.

We demonstrated the capabilities of DMtools with genome-wide DNA methylation data results for human NeuN control, NeuN Schizo, and AZA-AML3 (20) to assess the file size, differential methylation analysis, and visualization. Additionally, we obtained WGBS-Seq data for Arabidopsis Col-0, met1-3, ddcc, and mddcc (21), and used DMtools to analyze and visualize the DNA methylation levels. Please refer to the supplementary file for details of RRBS and target-BS data.

Conclusion

With the application of next-generation sequencing technologies and their decreasing costs, more and more WGBS and RRBS DNA methylation data are becoming available. Efficient storage and processing pipelines are necessary to handle such

large datasets. While existing methods such as GZIP, CGbz and BAllC are available for compressing DNA methylation results, they lack suitable tools for subsequent analysis of DNA methylation data. Additionally, GZIP is not well-suited for effective compression of DNA methylation results (22).

In this work, we have presented an innovative binary index DM format for DNA methylation data after mapping, which includes methylation level, coverage, context, and additional relevant information. Compared to traditional DNA methylation formats, the DM format offers several advantages, such as smaller file sizes and efficient query capabilities for DNA methylation levels in any genomic region. We have developed DMtools, a comprehensive toolkit that can generate DM format files from sorted BAM files, convert traditional DNA methylation files to DM format, process DM format files, and perform various DNA methylation analyses, including DNA methylation calculation, differential methylation analysis, and DNA methylation visualization.

We believe that the DM format and DMtools have the potential to make the analysis of WGBS and RRBS DNA methylation data more efficient. We are committed to continuously improving the software and incorporating new features based on user feedback in future.

Data availability

The software and usage documents can be accessed via the following links:

Software documentation: https://dmtools-docs.rtfd.io/

GitHub repository for DMtools: https://github.com/ZhouQiangwei/DMtools

GitHub repository for pyDMtools: https://github.com/ZhouQiangwei/pyDMtools

Acknowledgements

We thank the group members for providing feedback on the software.

Funding

National Key Research and Development Program of China [2021YFC2701201]; National Natural Science Foundation of China [31970590].

References

1. Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J. *et al.* (2010) Dynamic changes in the human methylome during differentiation. *Genome Res.*, **20**, 320-331.
2. Bock, C., Beerman, I., Lien, W.H., Smith, Z.D., Gu, H., Boyle, P., Gnirke, A., Fuchs, E., Rossi, D.J. and Meissner, A. (2012) DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. *Mol. Cell*, **47**, 633-647.
3. Luo, Y., Lu, X. and Xie, H. (2014) Dynamic Alu methylation during normal development, aging, and tumorigenesis. *Biomed Res Int*, **2014**, 784706.
4. Jones, M.J., Goodman, S.J. and Kobor, M.S. (2015) DNA methylation and healthy human aging. *Aging Cell*, **14**, 924-932.
5. Tucci, V., Isles, A.R., Kelsey, G., Ferguson-Smith, A.C. and Erice Imprinting, G. (2019) Genomic Imprinting and Physiological Processes in Mammals. *Cell*, **176**, 952-965.
6. Luijk, R., Wu, H., Ward-Caviness, C.K., Hannon, E., Carnero-Montoro, E., Min, J.L., Mandaviya, P., Muller-Nurasyid, M., Mei, H., van der Maarel, S.M. *et al.* (2018) Autosomal genetic variation is associated with DNA methylation in regions variably escaping X-chromosome inactivation. *Nat Commun*, **9**, 3738.
7. Yang, X., Han, H., De Carvalho, D.D., Lay, F.D., Jones, P.A. and Liang, G. (2014) Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell*, **26**, 577-590.
8. Saghafinia, S., Mina, M., Riggi, N., Hanahan, D. and Ciriello, G. (2018) Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors. *Cell Rep*, **25**, 1066-1080 e1068.
9. Cheng, Y., He, C., Wang, M., Ma, X., Mo, F., Yang, S., Han, J. and Wei, X. (2019) Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. *Signal Transduct Target Ther*, **4**, 62.
10. Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L. and Paul, C.L. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A*, **89**, 1827-1831.
11. Zhou, Q., Lim, J.Q., Sung, W.K. and Li, G. (2019) An integrated package for bisulfite DNA methylation data analysis with Indel-sensitive mapping. *BMC Bioinformatics*, **20**, 47.
12. Krueger, F. and Andrews, S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571-1572.
13. Huang, K.Y.Y., Huang, Y.J. and Chen, P.Y. (2018) BS-Seeker3: ultrafast pipeline for bisulfite sequencing. *BMC Bioinformatics*, **19**, 111.
14. Xi, Y. and Li, W. (2009) BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, **10**, 232.
15. Zhou, Q., Guan, P., Zhu, Z., Cheng, S., Zhou, C., Wang, H., Xu, Q., Sung, W.K. and Li, G.

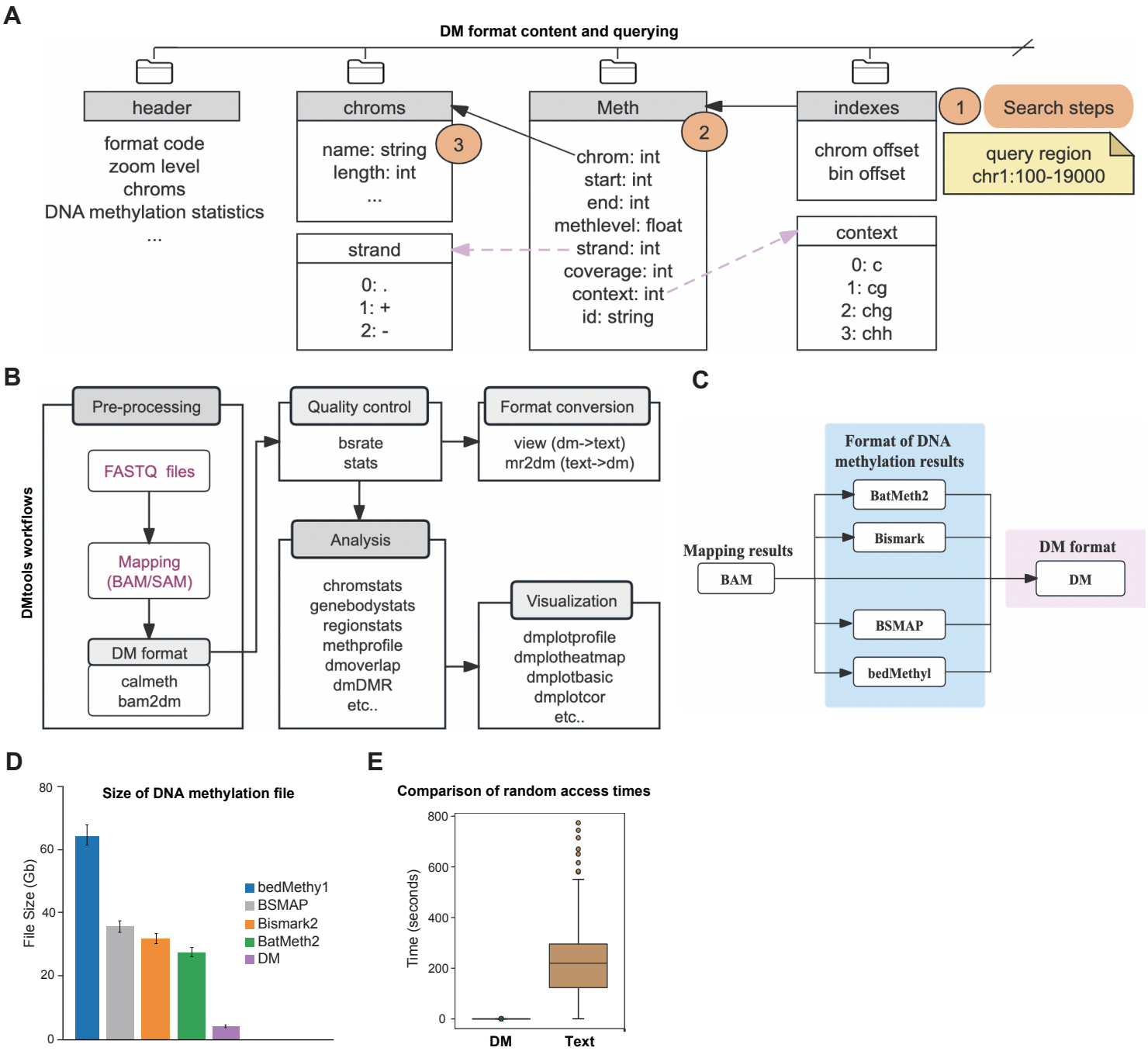
(2022) ASMdb: a comprehensive database for allele-specific DNA methylation in diverse organisms. *Nucleic Acids Res*, **50**, D60-D71.

16. Li, R., Liang, F., Li, M., Zou, D., Sun, S., Zhao, Y., Zhao, W., Bao, Y., Xiao, J. and Zhang, Z. (2018) MethBank 3.0: a database of DNA methylomes across a variety of species. *Nucleic Acids Res.*, **46**, D288-D295.
17. Guo, W., Zhu, P., Pellegrini, M., Zhang, M.Q., Wang, X. and 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**, 381-387.
18. Tian, W., Ding, W. and Ecker, J.R. (2023) BAllC and BAllCools: Efficient Formatting and Operating for Single-Cell DNA Methylation Data. *bioRxiv*.
19. Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A. and Mason, C.E. (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*, **13**, R87.
20. Lund, K., Cole, J.J., VanderKraats, N.D., McBryan, T., Pchelintsev, N.A., Clark, W., Copland, M., Edwards, J.R. and Adams, P.D. (2014) DNMT inhibitors reverse a specific signature of aberrant promoter DNA methylation and associated gene silencing in AML. *Genome Biol*, **15**, 406.
21. Zhao, L., Zhou, Q., He, L., Deng, L., Lozano-Duran, R., Li, G. and Zhu, J.-K. (2022) DNA methylation underpins the epigenomic landscape regulating genome transcription in Arabidopsis. *Genome Biol.*, **23**, 197.
22. Peng, J., Milenkovic, O. and Ochoa, I. (2018) METHCOMP: a special purpose compression platform for DNA methylation data. *Bioinformatics*, **34**, 2654-2656.

Figure legends

Fig. 1. DM format and related performance comparison. (A) Detailed information of the DM format and an example of querying in a random region. Step 1) Represents the search query region. Step 2) Searches and locates the query region based on the DM file index. Step 3) Retrieves and returns DNA methylation result information within the query region. **(B)** DMtools workflows and tools. Entry points for external data are highlighted in purple. **(C)** Conversion from other formats of DNA methylation files after mapping to DM format. **(D)** Comparison of storage space for different DNA methylation file formats. File formats are DM-format and tab-separated text DNA methylation results from BatMeth2, Bismark, BSMAP, and bedMethyl format. **(E)** Comparison of analysis speed for chromosome region DNA methylation level calculation with DM and tab-separated methylation file formats.

Figure 1


A
DM format content and querying
header
format code
zoom level
chroms
DNA methylation statistics
...
chroms
name: string
length: int
...
3
strand
0: .
1: +
2: -
Meth
chrom: int
start: int
end: int
methlevel: float
strand: int
coverage: int
context: int
id: string
2
indexes
chrom offset
bin offset
1
Search steps
query region
chr1:100-19000
context
0: c
1: cg
2: chg
3: chh
B
DMtools workflows
Pre-processing
FASTQ files
Mapping
(BAM/SAM)
DM format
calmeth
bam2dm
Quality control
bsrate
stats
Format conversion
view (dm->text)
mr2dm (text->dm)
Analysis
chromstats
genebodystats
regionstats
methprofile
dmoverlap
dmDMR
etc..
Visualization
dmplotprofile
dmplotheatmap
dmplotbasic
dmplotcor
etc..
C
Format of DNA
methylation results
BatMeth2
Bismark
BSMAP
bedMethyl
Mapping results
BAM
DM format
DM
D
Size of DNA methylation file
File Size (Gb)
80
60
40
20
0
bedMethy1
BSMAP
Bismark2
BatMeth2
DM
E
Comparison of random access times
Time (seconds)
800
600
400
200
0
DM
Text