The DNA methylation Haplotype (mHap) format and mHapTools
Supplementary Methods and Tables

Materials and Methods

1. mHap format and mHapTools functions

1.1 Overview of the mHap format

The mHap format file was developed for efficient storage and manipulation of DNA methylation haplotypes. It is a genomic position-based format with six fields, including the genomic region specified by the first and last cytosines in the first three fields, the DNA methylation haplotype, and the numbers of reads with given methylation haplotypes and strand (**Fig. 1A**). In the fourth field, DNA methylation haplotypes are represented as binary strings with 0 and 1, where 1 indicates methylated CpG site and 0 indicates un-methylated CpG site. An mHap file can be indexed by *Tabix* [1] to achieve random and fast retrieval of haplotypes overlapping with a specified chromosomal region. In most cases, only one seek call is needed to retrieve haplotypes in a region.

1.2 Converting SAM/BAM to mHap files

This is the main function of mHapTools: it takes an indexed BS-seq BAM/SAM and a CpG position file as input to extract DNA methylation haplotypes, the resulting output is saved as a mHap file.

1.3 Merging multiple mHap files

This module merges multiple sorted mHap files to produce a single sorted mHap file that contains all input records.

1.4 Summary of CpG site-level methylation

This module outputs summary information of CpG site-level methylation data from mHap files. It is similar to the Bismark DNA methylation caller but uses a mHap file as input.

1.5 Summary of DNA methylation statistics

The *summary* function of mHapTools computes the methylated CpG sites, total CpG sites, discordant reads, methylated reads and total number of reads. The analysis can be performed for the entire genome or specific regions by the " -g " or “-b” options. PDR and CHALM are calcuated as the proportion of discordant reads and methylated reads among all reads that cover at least 4 CpG sites [2].

2. Differential methylation analysis

Differential methylation analysis was performed by comparing two groups of samples using Student's t-test, the resulting p-values were adjusted with the Benjamini–Hochberg procedure to obtain false discovery rates (FDR). This analysis was performed on genomic features, such as CpG islands, downloaded from the UCSC genome browser, and gene promoters, defined as 2,000 bp regions centered at the transcription start site (TSS). Similarly, changes in discordant methylation were tested by comparing PDR between different conditions using Student's t-test. Significantly

differentially methylated CpG islands and gene promoters were determined with FDR cutoff of 5% plus absolute changes of mean methylation or PDR of 0.1

3. Pathway enrichment

Gene ontology (GO), KEGG pathway and tumor Hallmarks enrichment were performed by a hypergeometric test using the clusterProfiler [3] and GSEA online tool, respectively. The P value was adjusted for multiple hypothesis testing according to Benjamini and Hochberg, with 5% as a cutoff.

4.The mHap database

Using mHapTools, we constructed a DNA methylation haplotype database, mHapDB, which includes 3,731 publicly available human RRBS samples. All of them were downloaded from the Gene Expression Omnibus (GEO), including 917 cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) [4]. The R package SRAdb [5] was used to query all publicly available human RRBS samples in GEO, with the following parameters : library strategy = "Bisulfite-Seq", platform = "ILLUMINA", library selection = "Reduced Representation" and taxon id = "9606". For all samples, the full names, cell types, tissue of origin, disease categories and other metadata were curated. Raw SRA files were downloaded and converted to FastQ files with the SRA Toolkit (version 2.9.1), followed by quality control using FastQC (version: 0.11.8). Sequence adapters were trimmed by trim_galore (version 0.6.2) in the RRBS mode, thus removing an additional 3 bp at the 3'-end when adapters were detected. Comprehensive quality control reports were generated using Bismark [6] with the first 1 million reads or read pairs. Trimmed sequences were aligned to the human genome version hg19 using BSMAP [7], with options " -q 20 -f 5 -r 0 -v 0.05 -s 16 -S 1". Mean CpG methylation levels were extracted using MethylDackel [8]. DNA methylation haplotypes were extracted using mHapTools and saved as mHap files.

5. Data availability

All the raw DNA methylation sequencing data (RRBS) was downloaded from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/gds/). The TPM normalized CCLE RNA-seq matrix data and samples annotation files were derived from CCLE Browse Datasets (https://portals.broadinstitute.org/ccle/data). The mouse extraembryonic ectoderm (ExE) and Epiblast data are accessible under GSE84236; and CLL and normal B lymphocytes from GSE118255, GSE71702, GSE109085, GSE66121.

Table S1. Comparison of output formats between mHapTools and other BS-seq processing programs

Software	Input format	Output format	Description of outputs
mHapTools	BAM/SAM-indexed	mHap	CpG methylation haplotypes
Bismark [6]	Fastq	SAM/BAM, tab-delimited text	BAM files with additional XM tags to represent methylation states for CpG and non-CpG sites
BSMAP [7]	Fastq, BAM	SAM/BAM, BSP, text	Standard SAM/BAM format
BS-Seeker [9]	Fastq	SAM/BAM, tab-delimited text	BAM files with additional XM tags to represent methylation states for CpG and non-CpG sites
BS-Seeker2 [10]	Fastq	SAM/BAM, tab-delimited text	BAM files with additional XM tags to represent methylation states for CpG and non-CpG sites
MethyQA [11]	Fastq	Tab-delimited text	Standard SAM/BAM format
BRAT [12]	Fastq	Tab-delimited text	Standard SAM/BAM format
BRAT-BW [13]	Fastq	Tab-delimited text	Standard SAM/BAM format
MethPipe [14]	Fastq, SAM, BAM	BAM, tab-delimited text	Standard SAM/BAM format
MethylCoder [15]	Fastq	Tab-delimited text	Summary statistics for single cytosine
MethylDackel [8]	SAM, BAM	Tab-delimited text	Summary statistics for single cytosine
METHCOMP [16]	bedMethyl	Compressed BedMethyl	Compressed BedMethyl
CGmapTools [17]	BAM	ATCGmap	Summary statistics for single cytosine

Table S2. Performance of mHapTools for processing different BS-seq Assay with human hg19 genome

Assay	Sample	Reads length (bp)	Depth Coverage	BAM (MB)	RAM (MB)	Running Time（min)	mHap (MB)	Compression ratio
RRBS	SRX1085030	100	18.85	506	375.96	4.6	14.89	33.98
	SRX1085031	100	24.24	764	412.62	6.67	20.82	36.7
	SRX1085026	100	42	1036	598.75	12.82	27.5	37.67
	SRX1070571	100	32.16	926.76	452.99	8.03	20.93	44.28
	SRX1075147	100	16.68	1056	497.84	7.07	24.61	42.91
	SRX1070580	100	45.11	3172.18	969.5	20.22	82.61	38.4
	Average	100	29.84	1243.49	551.28	9.9	31.89	38.99
WGBS	sample20A042139_S14_L002	150	16.03	23,017.4	2,450.53	85.6	349.77	65.81
	sample20A042163_S16_L002	150	16.55	25,408.6	3,225.26	91.1	349.5	72.7
	sample20A042220_S21_L003	150	13.35	21,902	2,478.04	72.88	325.77	67.23
	sample20A042245_S30_L004	150	20.88	31,186.4	3,588.05	114.53	432.38	72.13
	sample20A042317_S11_L002	150	14.72	22,953.8	2,478.05	82.63	341.06	67.3
	sample20A042444_S12_L002	150	13.97	21,670.8	2,565.02	77.78	324.05	66.87
	Average	150	15.92	24,356.5	2,797.49	87.42	353.76	68.67
Targeted-BS	Sample767501	150	332.187	23.34	305.71	0.53	0.2	117.67
	sample737449	150	627.275	39.9	329.2	0.78	0.27	147.8
	sample738397	150	485.214	38.18	300.38	0.72	0.25	152.72
	sample748185	150	527.614	41.06	349.38	0.78	0.27	152.07
	sample748560	150	701.374	57.91	364.23	0.95	0.38	152.4
	sample784846	150	457.764	30.93	322.3	0.63	0.19	162.79
	Average	150	521.9	38.55	328.53	0.73	0.26	147.58

Legends for Supplementary Figures

Figure S1. The comparison of mHap and BAM formats. (**A-B**) The BAM files generated by BSMAP and BISMARK, respectively. (**C**) The DNA methylation haplotypes, each of which is shown separately. (**D**) The mHap format, in which the same DNA methylation haplotypes are merged.

Figure S2. Memory usage (RAM) and time consumptions in the processing of RRBS and targeted-BS data.

Figure S3. File size reduction of the mHap format compared to the BAM format. In order to explore how mHap format achieve its file size reduction, the effects of different components, including information reduction, aggregation, gzip compression and binarization, were tested. For information reduction, DNA methylation haplotypes were extracted from BAM files. In aggregation step, reads with the same haplotypes were merged and the count column was updated to reflect number of reads with the same haplotypes. Bgzip was used for compression. Datasets used in this testing were described in detail in Supplementary Table S2. File size reduction under each condition was shown for RRBS (**A**), WGBS (**B**) and targeted-BS (**C**).

Figure S4. Binarization slightly improve compression efficiency. The testing datasets include 30 RRBS samples which were randomly chosen from the mHap database (http://mhap.sciplus.cloud/mHapDB/). On average, binarization reduces file size by 7%.

Figure S5. PDR and CHALM quantify cell heterogeneity using DNA methylation haplotypes. Mouse extraembryonic ectoderm (ExE) (GSE98963) was used as a model to study cellular heterogeneity which was quantified by PDR or CHALM. The dependency between cellular heterogeneity and mean methylation were shown for PDR (**A**) and CHALM (**B**), respectively.

Figure S6. DNA methylation haplotype-level summary statistics are correlated with gene expression. Similar to mean methylation, haplotype-level summary statistics such as PDR and CHALM are both negatively correlated with gene expression in extraembryonic ectoderm (ExE) (GSE98963). Promoter are calculated as ±1kb of an annotated TSS. The dashed lines were fitted by lowess method.

Figure S7. Genes with DNA methylation changes were significantly enriched with biological pathways in ExE and Epiblast. (**A**) Assignment of promoters by changes of mean methylation and PDR between ExE and Epiblast. Specifically, 4 groups were defined: gene promoters with significant changes in both mean methylation and PDR (group 1, n = 843), only significant changes in PDR (group 2, n = 1394), only significant changes in mean methylation (group 3, n = 295) and those with no significant changes in both mean methylation and PDR (group 4, n = 10,763). The correlation between changes in mean methylation and PDR were shown for all genes in these 4 groups. (**B-D**) GO biological processes enrichment analysis of genes in group 1, 2 and 3.

Figure S8. Differential PDR is associated with differential gene expression in ExE and Epiblast. Genes were assigned into 4 groups, as described in **Figure S7.** Fisher's Exact test was used to

determine whether changes in PDR (**A** and **B)** or mean methylation (**C**) were associated with changes in gene expression. For gene expression, 2-fold change was used as a cutoff to define differentially expressed genes.

Figure S9. Disordered methylation is a hallmark of Chronic lymphocytic leukemia (CLL). (**A)** Proportion of discordant reads (PDR) is a haplotype-level summary statistics that quantifies disordered methylation. (**B-C)** Mean methylation patterns in CLL and B-cells. The analysis was limited to tiles or CpG islands that were covered with at least 30 reads. The boxplot shows the sample mean methylation patterns of normal B cells (n=59) and CLL cells (n=45) in genome-wide tiles (**B**) or CpG islands (**C**). (**D-E)** Patterns of disordered methylation in CLL and B-cells. The analysis was limited to tiles and CpG islands that were covered with greater than 10 reads and 4 or more CpG sites per read. (**F)** Mean intrasample CpG islands PDR variance. (**G)** The volcano plot showing the distribution of PDR changes and p-values for all covered CpG islands. Significant CpG islands (FDR < 5%, delta PDR > 0.1) were plotted as brown or purple points. For CpG islands with increased PDR in CLL, the enriched genomic features (**H**) and gene signatures (**I-K**) were shown.

Figure S10. Locally disordered methylation is associated with gene regulation in lung cancer. (**A)** Global discordant methylation levels measured by PDR are significantly different between NSCLC (n = 122) and SCLC (n = 49) cell lines. Averaged PDR for each cell line is calculated over promoter regions (±TSS 1000 bp) of 12,673 high coverage genes in all 917 CCLE RRBS samples. (**B)** Mean promoter PDR and gene expression are negatively correlated in lung cancer cell lines. (**C)** Assignment of promoters by changes of mean methylation and changes of disordered methylation between NSCLC and SCLC cell lines. Specifically, 4 groups were defined: gene promoters with significant changes in both mean methylation and PDR (group 1, n = 862), only significant changes in PDR (group 2, n = 236), only significant changes in mean methylation (group 3, n = 780) and those with no significant changes in both mean methylation and PDR (group 4, n = 10,795). The correlation and corresponding p-value measure all genes in the 4 groups.

Figure S11: Differential PDR is associated with differential gene expression in lung cancer. Genes were assigned into 4 groups, as described in **Figure S10.** Fisher's Exact test was used to determine whether changes in PDR (**A** and **B**) or mean methylation (**C**) were associated with changes in gene expression. For gene expression, 2-fold change was used as a cutoff to define differentially expressed genes.

Figure S12. Cancer-specific PDR changes in the CCLE dataset. To identify cancer type-specific promoters, one cancer type was compared to all other cancer types in CCLE dataset (FDR < 0.05, delta PDR > 0.1, delta mean methylation > 0.1). For visualization purposes, only cancer types with more than 30 PDR promoters were shown. (**A)** Promoters with up-regulated PDR. The right panel shows the proportion of gene promoters with significant changes in discordant methylation but not mean methylation. (**B)** Promoters with down-regulated PDR. The right panel shows the proportion of gene promoters with significant changes in discordant methylation but not mean methylation.

Reference

1. Li, H., *Tabix: fast retrieval of sequence features from generic TAB-delimited files*. Bioinformatics, 2011. **27**(5): p. 718-9.
2. Landau, D.A., et al., *Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia*. Cancer Cell, 2014. **26**(6): p. 813-825.
3. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters*. OMICS, 2012. **16**(5): p. 284-7.
4. Ghandi, M., et al., *Next-generation characterization of the Cancer Cell Line Encyclopedia*. Nature, 2019. **569**(7757): p. 503-508.
5. Zhu, Y., et al., *SRAdb: query and use public next-generation sequencing data from within R*. BMC Bioinformatics, 2013. **14**: p. 19.
6. Krueger, F. and S.R. Andrews, *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications*. Bioinformatics, 2011. **27**(11): p. 1571-2.
7. Xi, Y. and W. Li, *BSMAP: whole genome bisulfite sequence MAPping program*. BMC Bioinformatics, 2009. **10**: p. 232.
8. Ryan, D. *MethylDackel*. 2017; Available from: https://github.com/dpryan79/MethylDackel.
9. Chen, P.Y., S.J. Cokus, and M. Pellegrini, *BS Seeker: precise mapping for bisulfite sequencing*. BMC Bioinformatics, 2010. **11**: p. 203.
10. Guo, W., et al., *BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data*. BMC Genomics, 2013. **14**: p. 774.
11. Sun, S., A. Noviski, and X. Yu, *MethyQA: a pipeline for bisulfite-treated methylation sequencing quality assessment*. BMC Bioinformatics, 2013. **14**: p. 259.
12. Harris, E.Y., et al., *BRAT: bisulfite-treated reads analysis tool*. Bioinformatics, 2010. **26**(4): p. 572-3.
13. Harris, E.Y., et al., *BRAT-BW: efficient and accurate mapping of bisulfite-treated reads*. Bioinformatics, 2012. **28**(13): p. 1795-6.
14. Song, Q., et al., *A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics*. PLoS One, 2013. **8**(12): p. e81148.
15. Pedersen, B., et al., *MethylCoder: software pipeline for bisulfite-treated sequences*. Bioinformatics, 2011. **27**(17): p. 2435-6.
16. Peng, J., O. Milenkovic, and I. Ochoa, *METHCOMP: a special purpose compression platform for DNA methylation data*. Bioinformatics, 2018. **34**(15): p. 2654-2656.
17. Guo, W., et al., *CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data*. Bioinformatics, 2018. **34**(3): p. 381-387.