Nucleic Acids Research, 2024, **52**, D929–D937
https://doi.org/10.1093/nar/gkad881
Advance access publication date: 13 October 2023
Database issue

mHapBrowser: a comprehensive database for visualization and analysis of DNA methylation haplotypes

Yuyang Hong[1,†], Leiqin Liu[1,†], Yan Feng[1,†], Zhiqiang Zhang[2,3,†], Rui Hou[1], Qiong Xu[4,*] and Jiantao Shi[1,*]

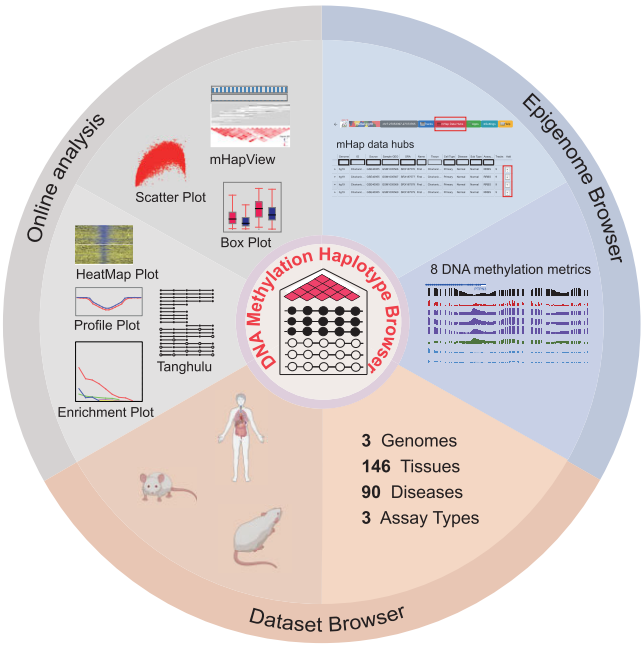
[1]Key Laboratory of RNA Science and Engineering, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China
[2]Shanghai Institute of Hematology, State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine at Shanghai, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China
[3]School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200030, China
[4]Department of Respiratory Disease, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200127, China

[*]To whom correspondence should be addressed. Tel: +86 54921122; Email: jtshi@sibcb.ac.cn
Correspondence may also be addressed to Qiong Xu. Email: drxuqiong@gmail.com
[†]The authors wish it to be known that, in their opinion, the first four authors should be regarded as Joint First Authors.

Abstract

DNA methylation acts as a vital epigenetic regulatory mechanism involved in controlling gene expression. Advances in sequencing technologies have enabled characterization of methylation patterns at single-base resolution using bisulfite sequencing approaches. However, existing methylation databases have primarily focused on mean methylation levels, overlooking phased methylation patterns. The methylation status of CpGs on individual sequencing reads represents discrete DNA methylation haplotypes (mHaps). Here, we present mHapBrowser, a comprehensive database for visualizing and analyzing mHaps. We systematically processed data of diverse tissues in human, mouse and rat from public repositories, generating mHap format files for 6366 samples. mHapBrowser enables users to visualize eight mHap metrics across the genome through an integrated WashU Epigenome Browser. It also provides an online server for comparing mHap patterns across samples. Additionally, mHap files for all samples can be downloaded to facilitate local processing using downstream analysis toolkits. The utilities of mHapBrowser were demonstrated through three case studies: (i) mHap patterns are associated with gene expression; (ii) changes in mHap patterns independent of mean methylation correlate with differential expression between lung cancer subtypes; and (iii) the mHap metric MHL outperforms mean methylation for classifying tumor and normal samples from cell-free DNA. The database is freely accessible at http://mhap.sibcb.ac.cn/.

Graphical abstract



Received: August 11, 2023. **Revised:** September 13, 2023. **Editorial Decision:** September 29, 2023. **Accepted:** September 29, 2023

Introduction

DNA methylation is an important epigenetic modification that plays a crucial role in gene regulation, genome stability, cellular differentiation and disease processes (1). Sequencing-based methodologies, such as whole-genome bisulfite sequencing (WGBS), reduced representation bisulfite sequencing (RRBS) and targeted bisulfite sequencing (BS-seq), enable characterization of DNA methylation patterns at a single-nucleotide resolution. The continuous accumulation of extensive experiments and datasets presents significant challenges for integrating and mining DNA methylation data. To address this, several databases have been established, including MethDB (2), MethBase (3), DNMIVD (4), NGSmethDB (5) and MethBank (6). However, these databases primarily focus on mean methylation levels at each CpG site, overlooking phased DNA methylation patterns. Even when derived from bulk data, CpG methylation within an individual read fragment can be traced back to a single chromosome in a single cell. Thus, the methylation pattern of CpGs on each fragment represents a discrete methylation haplotype (mHap). Several DNA methylation metrics have been proposed to quantify mHap-level patterns in bulk BS-seq data (7). Notably, mHap-level metrics, including proportion of discordant reads (PDR) (8), cellular heterogeneity-adjusted clonal methylation (CHALM) (9) and methylation concurrence ratio (MCR) (10), have demonstrated greater ability to explain gene expression variation compared to mean methylation. Additionally, methylated haplotype load (MHL) (11) and methylation block score (MBS) (12) show promise for noninvasive early cancer detection.

To improve accessibility of mHaps for the research community, we comprehensively processed BS-seq data from public databases. The results are presented as a novel, comprehensive database called the DNA methylation haplotype browser (mHapBrowser). Within this database, users can access a full-featured WashU Epigenome Browser (13–16) to visualize eight DNA methylation metrics. Additionally, this database offers convenient functionalities to visualize DNA mHaps and summarize related metrics, facilitating cross-sample comparisons. mHap files for all samples can be freely downloaded to enable local processing with our mHap toolkit mHapSuite, a Java implementation of mHapTk (17), along with the visualization package deepTools (18).

Materials and methods

Data source

The mHapBrowser database compiled and systematically curated DNA methylation sequencing data from publicly accessible databases, including the NCBI Gene Expression Omnibus (GEO), Sequence Read Archive (SRA) and ArrayExpress. Sample annotations were automatically retrieved and manually reviewed. Whenever possible, FASTQ files were downloaded for downstream processing.

Data processing

For preprocessing of raw BS-seq data from GEO, SRA files were obtained from the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra) and converted into FASTQ format with SRA-Toolkit (v3.0.2). Subsequently, Trim Galore (v0.6.2) was utilized to remove adapters from the BS-seq reads. For WGBS and targeted BS-seq data, the optimal value for the clip_R1 parameter, determining the number of bases trimmed from the 5′ end, was identified by analyzing the top 250 000 reads. Values ranging from 0 to 20 were tested, and the one yielding the highest mapping rate was selected. Then, the mapping rate with the selected clip_R1 value was compared to that with clip_R1 set to 0. If the difference was under 5%, clip_R1 was set to 0 to retain more data with adequate mapping. Analogous procedures determined clip_R2 settings. For RRBS data, the '-rrbs' flag was incorporated into the command line. Trimmed reads were aligned to reference genomes with BSMAP (v2.90) (19) using the following parameters: '-q 20 -f 5 -r 0 -v 0.05 -s 16 -S 1'. For paired-end WGBS and targeted BS-seq, duplicates were marked by sambamba (20).

Processing of mHaps

DNA mHaps were extracted from BAM files using mHapTools (v1.1) (21) and stored in mHap format. These files were then processed with mHapSuite (v2.0), available at https://github.com/yoyoong/mHapSuite, using the following parameters: '-minK 1 -maxK 10 -K 4 -strand both -cpgCov 5 -r2Cov 10'. This generated nine genome-wide tracks for each sample: coverage (Cov), mean methylation (MM), PDR, CHALM, MCR, MBS, MHL, entropy and linkage disequilibrium (LD) R-squared (R^2).

DNA methylation metrics

To calculate mHap metrics for a single CpG site, all reads covering the site were utilized. For calculating PDR, CHALM, MBS and entropy, only reads covering at least four CpG sites were considered.

MM: MM is defined as the proportion of methylated CpG sites over all covered CpG sites, calculated as

$$\mathrm{MM}\,(c) = \frac{\text{no. of methylated CpGs}}{\text{total no. of CpGs}}.$$

PDR: PDR was calculated as the number of discordant reads over the total number of reads. A read was classified as concordant if it exhibited consistent DNA methylation states at all covered CpG positions, or as discordant otherwise. PDR for CpG c is defined as

$$\mathrm{PDR}\,(c) = \frac{\text{no. of discordant reads covering } c}{\text{total no. of reads covering } c}.$$

CHALM: CHALM was developed based on the assumption that methylation on a single CpG site can recruit repressors to chromatin. CHALM for CpG c is defined as

$$\mathrm{CHALM}\,(c) = \frac{\text{no. of methylated reads covering } c}{\text{total no. of reads covering } c}.$$

A read containing at least one methylated CpG site is classified as a methylated read.

MCR: MCR is defined as the ratio of number of unmethylated CpG sites in partially methylated reads to the total number of covered CpG sites, calculated as

$$\mathrm{MCR}\,(c) = \frac{\sum_{i=1}^{P} p_i}{\sum_{i=1}^{T} t_i},$$

where P and T represent number of partially methylated reads and total number of reads covering CpG c, respectively. p_i is the number of unmethylated CpG sites in partially methylated reads. t_i is the total number of CpG sites in read i.

MHL: MHL calculates weighted mean of the fraction of fully methylated substrings at different lengths. MHL for CpG c is calculated as

$$\mathrm{MHL}\,(c) = \frac{\sum_{i=1}^{10} i \times P\,(\mathrm{MH}_i)}{\sum_{i=1}^{10} i},$$

where $P(\mathrm{MH}_i)$ is the fraction of fully methylated substrings of length i. For a haplotype of length l ($\geq$10), all substrings from length 1 to 10 were considered in this calculation.

MBS: MBS is another metric used to quantify the level of successive methylation patterns. MBS is calculated as

$$\mathrm{MBS}\,(c) = \frac{1}{n} \times \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{m} l_{ij}^2}{L_i^2} \right),$$

where n is the total number of reads that cover CpG c site. L_i denotes the number of CpG sites covered on ith read, while l_{ij} represents the length of jth successive methylated CpG block split by unmethylated CpG sites on ith read, and m is the total count of the blocks.

Entropy: Following the concept of Shannon entropy $H(x)$, methylation entropy is calculated as

$$\mathrm{entropy}\,(c) = -\frac{1}{4} \sum_{i=1}^{n} P\,(H_i) \times \log_2 P\,(H_i)\,.$$

To account for the variability of fragment length, substrings of mHaps were counted with a length of 4, representing 16 types of haplotypes. $P(H_i)$ represents the probability of observing mHap H_i.

R^2. R^2 metric quantifies the co-methylation level of pairwise CpGs. The R^2 value for CpG c is the mean R^2 value of CpG c between two CpG sites before and after CpG c:

$$R^2\,(a, b) = \frac{p_{ab} - p_a p_b}{p_a\,(1 - p_a)\,p_b\,(1 - p_b)},$$

$$R^2\,(c) = \frac{R^2\,(c-2, c) + R^2\,(c-1, c) + R^2\,(c+1, c) + R^2\,(c+2, c)}{4},$$

where $R^2(a, b)$ represents the R^2 value of CpGs a and b, p_a and p_b represent the methylated proportions at CpGs a and b, respectively, and p_{ab} represents the proportion of co-methylation at CpGs a and b.

Identification of genes with differential promoter dispersion

For a given sample, promoter PDR exhibits a nonlinear relationship with mean methylation (8). To identify genes with higher or lower promoter dispersion than expected, we utilized a nonlinear method previously applied for constructing cell-specific networks in single-cell RNA-seq analysis (22). Using default parameters (box size = 0.1, P-value = 0.01), genes fitting a nonlinear model were identified. Remaining genes were classified as having higher or lower dispersion based on their relative position in the PDR versus mean methylation scatter plot.

Database implementation

The mHapBrowser was developed using Node.js (https://nodejs.org) and deployed on a Red Hat Linux Server. The frontend was constructed using JavaScript and TypeScript, rendered by React (https://react.dev) and styled using Material-UI (https://mui.com). The backend interface was built using the hapi framework (https://hapi.dev/). Dataset information was stored and managed in MySQL (https://www.mysql.com/), while the raw data were stored on the Linux Server. The track data, with the exception of annotation tracks, were accessed through backend API interfaces and stored on the Linux Server. The annotation tracks were loaded into MongoDB (https://www.mongodb.com/). All online analysis modules were implemented by executing the mHapSuite Jar package (https://github.com/yoyoong/mHapSuite).

Processing of RNA-seq data

For normal esophagus tissue (GSE149608), transcript abundance was estimated using Kallisto (v0.46.1) (23), an ultrafast RNA-seq quantification program utilizing a pseudoalignment approach based on a preconstructed transcriptome index. Abundance was reported in transcripts per million (TPM). For the Cancer Cell Line Encyclopedia (CCLE) data, the gene expression matrix was downloaded from the CCLE website (https://sites.broadinstitute.org/ccle/). Differential gene expression and methylation were analyzed by two-sided rank sum test. Resulting P-values were adjusted for multiple testing by applying the Benjamini–Hochberg (BH) procedure. Statistical significance was reported as false discovery rate (FDR) (24).

Results

Overview of mHapBrowser

The mHapBrowser database is a comprehensive atlas of DNA mHaps for human, mouse and rat (Figure 1). It contains 1815 WGBS, 4251 RRBS and 300 targeted BS-seq samples across three species: human (n = 5807), mouse (n = 490) and rat (n = 69). Blood has the most abundant data, with 54 datasets and 978 samples. Rich quality control metrics are provided for each sample, including alignment rate, mean coverage inside and outside CpG islands (CGIs), and mean cytosine methylation levels in CpG, CHG and CHH contexts. Samples were processed through a unified pipeline to generate mHap files, serving as standard inputs for online analysis and visualization. In mHapBrowser, the mHap files were directly utilized to visualize phased DNA methylation patterns in specified genomic regions. Additionally, the mHap files were used to construct genome-wide tracks depicting DNA methylation patterns across the genome, characterized by eight DNA methylation metrics. Briefly, mean methylation is the classical metric used to measure methylation levels. Entropy and PDR were developed to quantify epigenetic heterogeneity. CHALM and MCR were proposed to better explain gene expression variation. MHL and MBS were designed specifically for noninvasive early cancer detection. Finally, co-methylation between pairs of CpGs was measured by LD R^2.

User interface

The mHapBrowser database serves as a centralized resource for the analysis of DNA mHaps, consisting of three core modules: a dataset browser for finding samples of interest; an epigenome browser for visualizing DNA methylation metrics at the haplotype level across the genome; and an online server for analyzing and comparing haplotype-level DNA methylation patterns across samples. Additionally, the mHap files for all samples in dataset browser are freely downloadable to fa-

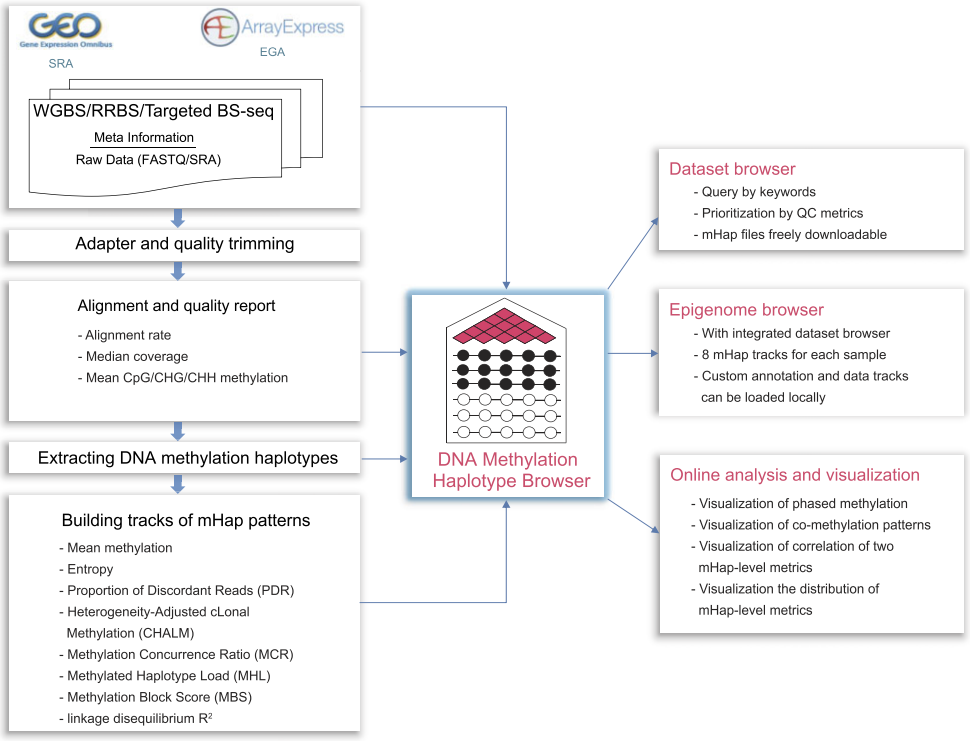


Figure 1. Overview of mHapBrowser. mHapBrowser processes raw data from public databases using a unified pipeline and presents the results as a centralized resource for the visualization and analysis of DNA mHaps.

cilitate local processing using mHapSuite (17) and publicly available tools such as deepTools (18).

Dataset browser

This module displays metadata for each sample, including data source, accession number, assay type, tissue type and disease status. Users can conveniently query samples of interest using one or multiple keywords. For instance, using the accession number 'GSE16256' and tissue type 'Lung' as keywords returned eight samples, comprising two primary samples and six cell lines (Figure 2A). Extensive quality control metrics are provided for each sample, such as alignment rate, median coverage inside and outside CGIs, and mean cytosine methylation levels in CpG, CHH and CHG contexts (Figure 2A). These metrics enable users to select high-quality samples for downstream analysis. Notably, cytosine methylation in the CHH context can be utilized to estimate bisulfite conversion rate, with a value below 1% indicating a conversion rate above 99%.

Epigenome browser

The mHapBrowser enables direct visualization of genome-wide DNA methylation patterns using the WashU Epigenome Browser (13–16). The dataset browser is seamlessly integrated into the genome browser, allowing users to easily select samples of interest via the 'mHap Data Hubs' function (Figure 2B). This integrated dataset browser offers identical functionality to the stand-alone version, including keyword-based querying and quality control metrics for all samples (Figure 2C). Once a sample is selected, nine corresponding tracks become available for visualization: Cov, MM, entropy, PDR, CHALM, MCR, MBS, MHL and R^2. These tracks enable users to explore comprehensive information related to the selected samples and their DNA methylation characteristics (Figure 2D). In addition to the built-in gene annotations, users can upload custom data and annotation tracks in formats such as BAM, bigWig and BED, provided they are supported by the genome browser.

Online analysis and visualization

The mHapBrowser offers a panel of functions for analyzing and visualizing DNA methylation patterns (Figure 2E). (i) The Tanghulu plot displays DNA mHaps in a specified genomic region. (ii) The mHapView module visualizes co-methylation patterns in particular genomic regions. (iii) Scatter plots display relationships between two continuous variables, including comparisons of two DNA methylation metrics from the same or different samples. (iv) Box plots illustrate the distribution of metric values across multiple samples. (v) Heatmaps show signals within defined regions, with a focus on the central area. (vi) Profile plots display the average signal profiles in predetermined genomic intervals and flanking regions. (vii) Enrichment plots display the percentage of genomic features overlapping user-specified intervals such as open chromatin regions.

Case study 1: mHap patterns are associated with gene expression

DNA methylation is widely recognized to be negatively correlated with gene expression. Utilizing mHap files in mHapBrowser facilitates exploration of the relationship between DNA methylation patterns and gene expression. In mHapBrowser, DNA methylation patterns can be quantified by various mHap-level metrics (Figure 3A). Notably, regions with

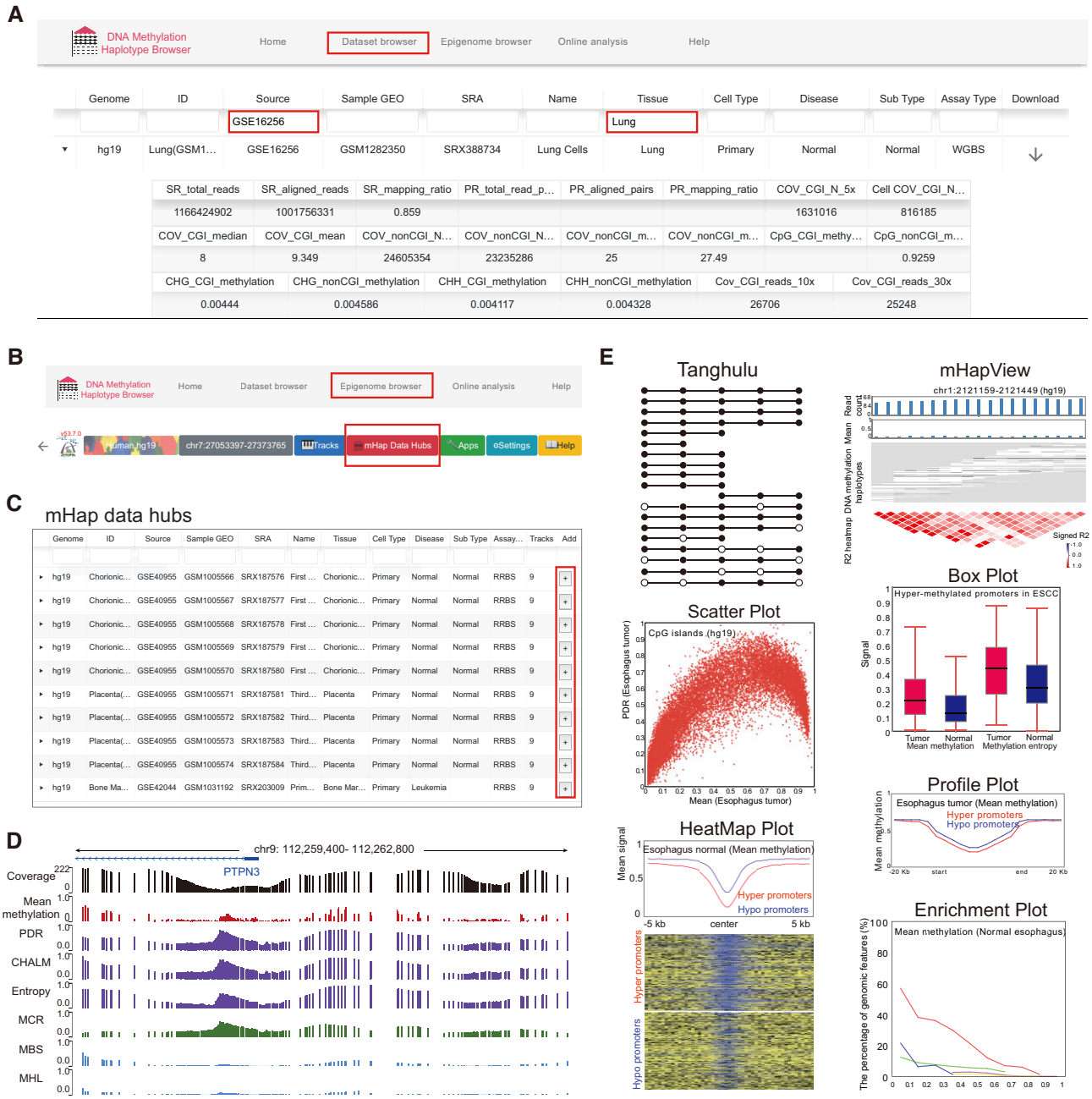


Figure 2. Core modules in mHapBrowser. (**A**) The dataset browser facilitates the search for samples of interest through keyword-based queries. (**B**–**D**) A re-engineered WashU Epigenome Browser enables the selection of samples and tracks of interest for visualization. (**E**) An online server, comprising seven analytical modules, enables the analysis and comparison of mHap patterns across samples.

similar mean methylation can exhibit distinct DNA methylation patterns, as measured by different mHap-level metrics. To demonstrate this, publicly available data from normal esophagus tissue were analyzed. The processed mHap files can be accessed under accession number GSE149608 (Supplementary Table S1). As expected, a negative correlation was observed between gene expression levels and mean methylation levels in promoter regions, with lower expressed genes tending to show higher DNA methylation (Figure 3B, upper panel). It is notable that even within normal tissues, the population consists of a heterogeneous mixture of cells. When quantifying epigenetic diversity using PDR, we found a similar negative correlation between PDR and gene expression in promoter regions (Figure 3B, middle panel). Additionally, MCR that quantifies methylation and demethylation was even more effective in distinguishing between genes with different expression levels (Figure 3B, lower panel). Previous studies have demonstrated a nonlinear correlation between PDR and mean methylation (8). Specifically, a positive correlation between PDR and mean methylation is observed at low methylation levels, whereas a negative correlation is observed at high methylation levels. We further identified pro-

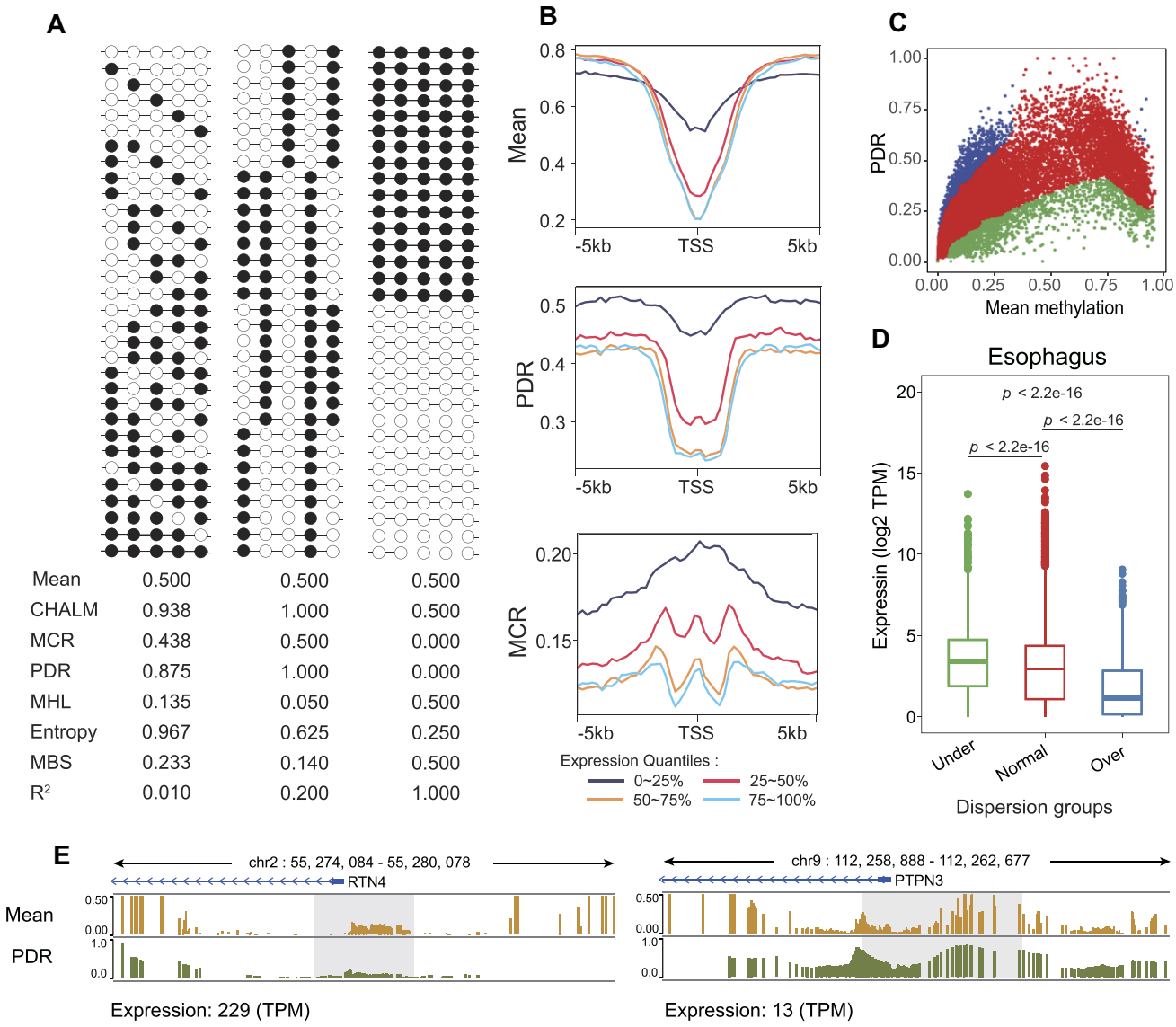


Figure 3. Association of DNA methylation patterns with gene expression. (**A**) Three hypothetical mHap patterns, each having an identical mean methylation level, are illustrated for a representative genomic region. (**B**) Correlation between gene expression and DNA methylation metrics. Using matched RNA-seq and WGBS data from normal esophageal tissue (GSE149608), genes were categorized into four equal expression quantile groups (0–25%, 25–50%, 50–75%, 75–100%). Mean methylation, PDR and MCR average profiles around transcription start site (TSS) are presented. (**C**) Identification of genes with differential promoter discordance. Genes that were explained by a nonlinear model are shown in the middle (see the 'Materials and Methods' section) ($P < 0.01$). Genes exhibiting higher or lower promoter discordance than expected are depicted in upper and lower parts, respectively. Gene expression distribution is shown in panel (**D**) as box plots. (**E**) Epigenome browser screenshots of sample genes with differential promoter discordance. Mean gene expression levels are shown for both genes.

moters with significantly higher or lower discordance than expected, considering their mean methylation as a conditional factor ($P < 0.01$) (Figure 3C, Supplementary Table S2). Interestingly, the distributions of gene expression among these three groups show statistically significant differences, indicating the association of DNA methylation patterns with gene expression (Figure 3D). Specifically, genes with underdispersion in promoters demonstrate activation, while overdispersed genes are repressed. For instance, both RTN4 and PTPN3 have similar mean methylation around 0.2; RTN4 is activated (TPM = 229) and PTPN3 is repressed (TPM = 13), a disparity that could be attributed to the higher promoter discordance in PTPN3 compared to the lower discordance in RTN4 (Figure 3E).

Case study 2: mHap patterns are associated with differential gene expression

To investigate the potential associations between DNA methylation patterns and differentially expressed genes (DEGs), we conducted an analysis using the CCLE dataset (PRJNA523380). The corresponding mHap files can be obtained from mHapBrowser by searching 'CCLE' (Supplementary Table S1). As an example, we compared the DNA methylation profiles of non-small cell lung cancer (NSCLC) ($n = 122$) and small cell lung cancer (SCLC) ($n = 49$). To control the influence of mean methylation changes, we first identified 13 205 genes that displayed no significant alterations in mean methylation levels within their promoter regions (|mean methylation change| < 0.1, FDR > 5%) (Figure 4A). In parallel, we com-

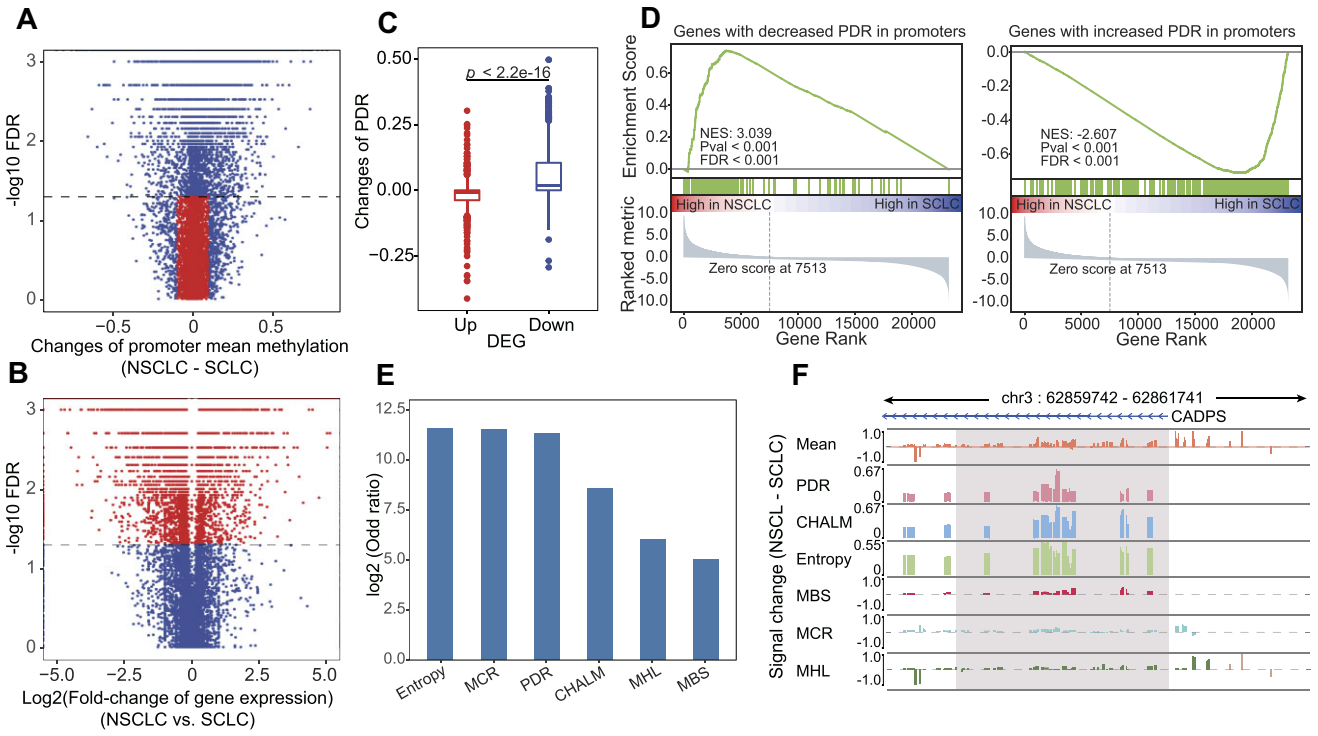


Figure 4. Association of DNA methylation patterns with differential gene expression. (**A**) Differential methylation analysis between NSCLC ($n = 122$) and SCLC ($n = 49$) samples. Statistical significance was assessed using a two-sided rank sum test with multiple testing correction via the BH procedure. Genes with unchanged promoter mean methylation (|change| < 0.1, FDR > 0.05) were identified. (**B**) DEGs between NSCLC and SCLC were identified using FDR < 5%. Statistical significance was assessed using a two-sided rank sum test with multiple testing correction via the BH method. (**C**) Genes exhibiting differential expression without changes in promoter mean methylation showed significant alterations in promoter PDR. (**D**) Gene set enrichment analysis demonstrated associations between PDR changes and expression changes. (**E**) Six DNA methylation metrics associated with expression changes as determined by Fisher's exact test (all $P < 2.22 \times 10^{-16}$); only odds ratios are shown. (**F**) An example gene CADPS exhibiting significant alterations in PDR, CHALM and entropy but not mean methylation in NSCLC compared to SCLC.

pared the gene expression profiles of NSCLC and SCLC and identified 13 798 DEGs (FDR < 5%) (Figure 4B). The shared genes represent DEGs without significant changes in promoter mean methylation. Remarkably, among these genes, upregulated genes and downregulated genes show significant difference in promoter PDR (FDR < 5%), indicating the presence of an association between DNA methylation patterns and DEGs that is independent of changes in mean methylation (Figure 4C). Alternatively, when we specifically select genes with a significant decrease in PDR but no significant change in mean methylation, they were found to be upregulated in NSCLC, as demonstrated by gene set enrichment analysis ($P < 0.001$) (Figure 4D, left panel). Similarly, genes with a significant increase in PDR but no significant change in mean methylation are repressed in NSCLC ($P < 0.001$) (Figure 4D, right panel). In this way, we found that differential gene expression is associated with DNA methylation patterns characterized by all six metrics, comprising entropy, MCR, PDR, CHALM, MHL and MBS (Fisher's exact test, P-value $< 10^{-16}$) (Figure 4E, Supplementary Table S3). As an example, CADPS is found to be significantly downregulated in NSCLC cell lines compared to SCLC cell lines, which can be potentially explained by changes of DNA methylation patterns in the promoter region, considering the minimal change in mean methylation (Figure 4F).

Case study 3: DNA methylation marker for cancer detection

DNA methylation has been extensively utilized as markers for cancer detection. Using mHapBrowser, we can easily assess predefined markers using cancer and normal tissue samples. For instance, Smith *et al.* identified a cluster of CGIs that displayed hypermethylation in mouse extraembryonic ectoderm and exhibited similar methylation patterns in most human cancer types, thus representing a universal cancer signature (25). The initial study showed distributions of mean methylation in The Cancer Genome Atlas tumors and their corresponding normal tissues, which were profiled using the 450K array. In this study, we demonstrate the presence of this signature in 11 tumor types that were profiled with RRBS, using DNA methylation metric MHL, which quantify co-methylation patterns (Supplementary Table S1). As expected, significant differences between tumor and normal samples were observed in all cancer types (rank sum test, $P < 10^{-16}$) (Figure 5A). The data in mHapBrowser also allow us to evaluate the performance of various DNA methylation metrics. Previous research has shown that MHL preserves a higher signal-to-noise ratio than mean methylation and, as a result, performs better in noninvasive cancer detection using DNA methylation of cell-free DNA (11). We utilized a publicly available dataset, GSE149438 (Supplementary Table S1), which profiled cell-free DNA from esophageal squamous cell carcinoma and normal individuals using targeted BS-seq, to compare the performance of mean methylation and MHL by a support vector machine. With different assignment of training and validation, the MHL-based model demonstrated superior performance compared to the mean methylation-based model (Figure 5B), as evidenced by a higher area under the receiver operating characteristic curve (AUC).

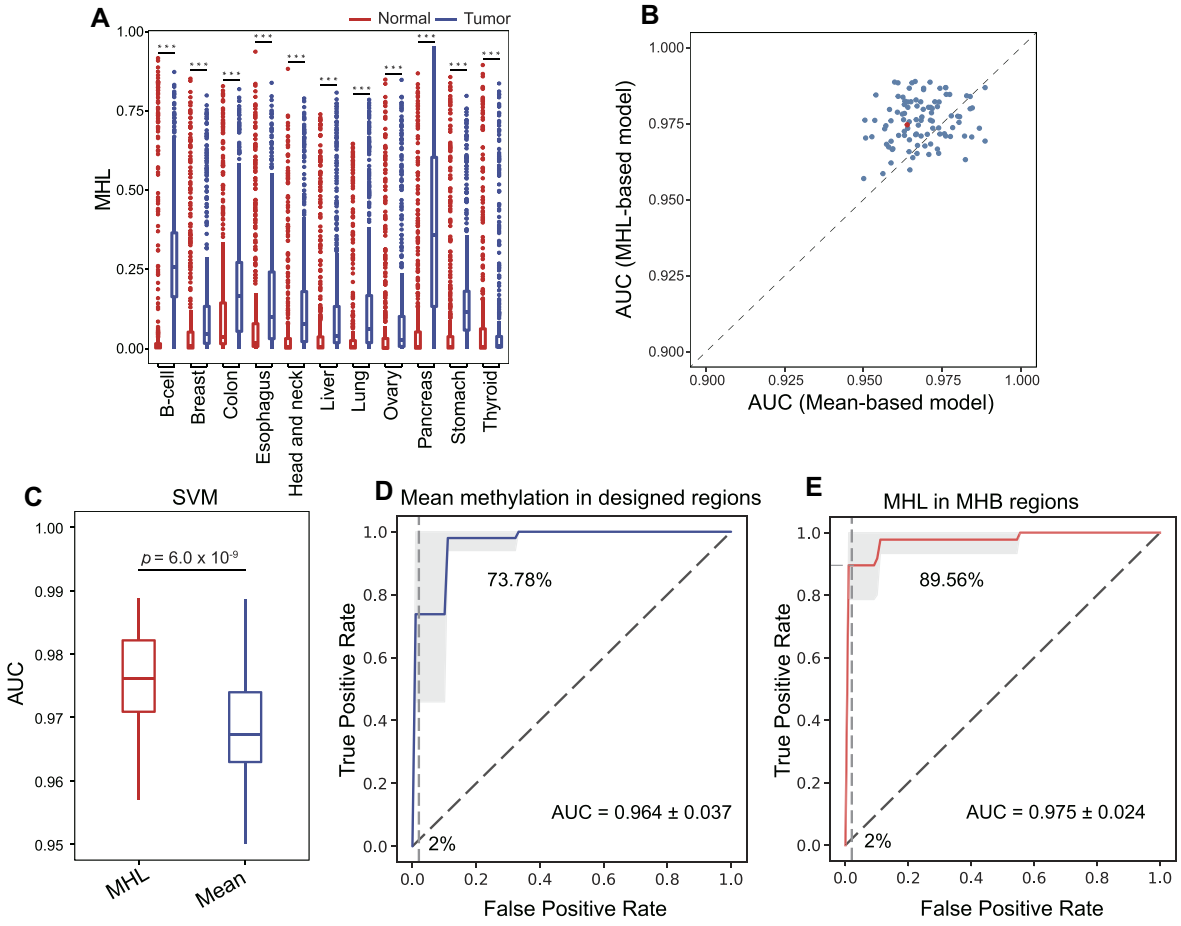


Figure 5. Utilizing mHap metrics for early cancer detection. (**A**) Evaluation of cancer signatures using mHapBrowser datasets. (**B**, **C**) Comparison of the performance of MHL and mean methylation for cancer detection using dataset GSE149438 with various training and validation set partitions. (**D**, **E**) The red point in panel (B) represents a typical result, highlighting that small differences in AUC can result in significant sensitivity differences at high specificity.

Although the improvement in overall performance was modest, the difference was found to be statistically significant (paired rank sum test P-value $< 6 \times 10^{-9}$) (Figure 5C). Importantly, even a minor enhancement in AUC can lead to substantial changes in sensitivity, especially in situations requiring high specificity. For instance, when specificity was set to 0.98, the MHL-based model achieved a sensitivity of 89.56% in contrast to 73.78% from the mean methylation-based model (Figure 5D and E).

Discussion and future developments

DNA methylation serves as an epigenetic regulatory mechanism involved in various critical biological processes. In BS-seq data, the DNA methylation status of CpG sites on the same sequencing read represents a discrete mHap, enabling decoding of the epigenetic code beyond mean methylation levels. However, there is a lack of well-curated resources offering tools and consistently processed mHap files for data integration. To address this need, we present the mHapBrowser database as a centralized resource for the analysis and visualization of DNA mHaps. Moving forward, our research will focus on the following aims: (i) expanding the datasets to include more species and cell types; (ii) continually updating the mHapSuite package to provide additional functionalities for offline processing of mHap files; and (iii) incorporating multi-omics data such as RNA-seq and ATAC-seq (assay for transposase-accessible chromatin using sequencing) for online integration and analysis. As demonstrated by the case studies, the continuous development of mHapBrowser will be devoted to both basic research and translational applications such as early cancer detection.

Data availability

The mHapBrowser database is freely accessible online, with all data available at http://mhap.sibcb.ac.cn. mHapSuite is available at https://github.com/yoyoong/mHapSuite. All source data and code have been uploaded to Figshare and are available at https://figshare.com/projects/mHapBrowser/175104.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

The authors sincerely thank the High-Performance Computing Center at the Shanghai Institute of Biochemistry and Cell

Biology for their invaluable support with data processing. Their generous assistance was greatly appreciated and played a pivotal role in the successful completion of this research.

Funding

National Natural Science Foundation of China [32270691 to J.S.]. Funding for open access charge: National Natural Science Foundation of China.

Conflict of interest statement

None declared.

References

1. Greenberg,M.V.C. and Bourc'his,D. (2019) The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.*, **20**, 590–607.
2. Grunau,C., Renault,E., Rosenthal,A. and Roizes,G. (2001) MethDB—a public database for DNA methylation data. *Nucleic Acids Res.*, **29**, 270–274.
3. Song,Q., Decato,B., Hong,E.E., Zhou,M., Fang,F., Qu,J., Garvin,T., Kessler,M., Zhou,J. and Smith,A.D. (2013) A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One*, **8**, e81148.
4. Ding,W., Chen,J., Feng,G., Chen,G., Wu,J., Guo,Y., Ni,X. and Shi,T. (2020) DNMIVD: DNA methylation interactive visualization database. *Nucleic Acids Res.*, **48**, D856–D862.
5. Lebron,R., Gomez-Martin,C., Carpena,P., Bernaola-Galvan,P., Barturen,G., Hackenberg,M. and Oliver,J.L. (2017) NGSmethDB 2017: enhanced methylomes and differential methylation. *Nucleic Acids Res.*, **45**, D97–D103.
6. Zhang,M., Zong,W., Zou,D., Wang,G., Zhao,W., Yang,F., Wu,S., Zhang,X., Guo,X., Ma,Y., *et al.* (2023) MethBank 4.0: an updated database of DNA methylation across a variety of species. *Nucleic Acids Res.*, **51**, D208–D216.
7. Scherer,M., Nebel,A., Franke,A., Walter,J., Lengauer,T., Bock,C., Muller,F. and List,M. (2020) Quantitative comparison of within-sample heterogeneity scores for DNA methylation data. *Nucleic Acids Res.*, **48**, e46.
8. Landau,D.A., Clement,K., Ziller,M.J., Boyle,P., Fan,J., Gu,H., Stevenson,K., Sougnez,C., Wang,L., Li,S., *et al.* (2014) Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell*, **26**, 813–825.
9. Xu,J., Shi,J., Cui,X., Cui,Y., Li,J.J., Goel,A., Chen,X., Issa,J.P., Su,J. and Li,W. (2021) Cellular Heterogeneity-Adjusted cLonal Methylation (CHALM) improves prediction of gene expression. *Nat. Commun.*, **12**, 400.
10. Shi,J., Xu,J., Chen,Y.E., Li,J.S., Cui,Y., Shen,L., Li,J.J. and Li,W. (2021) The concurrence of DNA methylation and demethylation is associated with transcription regulation. *Nat. Commun.*, **12**, 5285.
11. Guo,S., Diep,D., Plongthongkum,N., Fung,H.L., Zhang,K. and 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**, 635–642.
12. Liang,N., Li,B., Jia,Z., Wang,C., Wu,P., Zheng,T., Wang,Y., Qiu,F., Wu,Y., Su,J., *et al.* (2021) Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning. *Nat. Biomed. Eng.*, **5**, 586–599.
13. Li,D., Harrison,J.K., Purushotham,D. and Wang,T. (2022) Exploring genomic data coupled with 3D chromatin structures using the WashU Epigenome Browser. *Nat. Methods*, **19**, 909–910.
14. Li,D., Hsu,S., Purushotham,D., Sears,R.L. and Wang,T. (2019) WashU Epigenome Browser update 2019. *Nucleic Acids Res.*, **47**, W158–W165.
15. Li,D., Purushotham,D., Harrison,J.K., Hsu,S., Zhuo,X., Fan,C., Liu,S., Xu,V., Chen,S., Xu,J., *et al.* (2022) WashU Epigenome Browser update 2022. *Nucleic Acids Res.*, **50**, W774–W781.
16. Zhuo,X., Hsu,S., Purushotham,D., Kuntala,P.K., Harrison,J.K., Du,A.Y., Chen,S., Li,D. and Wang,T. (2023) Comparing genomic and epigenomic features across species using the WashU Comparative Epigenome Browser. *Genome Res.*, **33**, 824–835.
17. Ding,Y., Cai,K., Liu,L., Zhang,Z., Zheng,X. and Shi,J. (2022) mHapTk: a comprehensive toolkit for the analysis of DNA methylation haplotypes. *Bioinformatics*, **38**, 5141–5143.
18. Ramirez,F., Dundar,F., Diehl,S., Gruning,B.A. and Manke,T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, **42**, W187–W191.
19. Xi,Y. and Li,W. (2009) BSMAP: whole genome bisulfite sequence mapping program. *BMC Bioinformatics*, **10**, 232.
20. Tarasov,A., Vilella,A.J., Cuppen,E., Nijman,I.J. and Prins,P. (2015) Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, **31**, 2032–2034.
21. Zhang,Z., Dan,Y., Xu,Y., Zhang,J., Zheng,X. and Shi,J. (2021) The DNA methylation haplotype (mHap) format and mHapTools. *Bioinformatics*, **37**, 4892–4894.
22. Dai,H., Li,L., Zeng,T. and Chen,L. (2019) Cell-specific network constructed by single-cell RNA sequencing data. *Nucleic Acids Res.*, **47**, e62.
23. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
24. Korthauer,K., Kimes,P.K., Duvallet,C., Reyes,A., Subramanian,A., Teng,M., Shukla,C., Alm,E.J. and Hicks,S.C. (2019) A practical guide to methods controlling false discoveries in computational biology. *Genome Biol.*, **20**, 118.
25. Smith,Z.D., Shi,J., Gu,H., Donaghey,J., Clement,K., Cacchiarelli,D., Gnirke,A., Michor,F. and Meissner,A. (2017) Epigenetic restriction of extraembryonic lineages mirrors the somatic transition to cancer. *Nature*, **549**, 543–547.

Received: August 11, 2023. **Revised:** September 13, 2023. **Editorial Decision:** September 29, 2023. **Accepted:** September 29, 2023