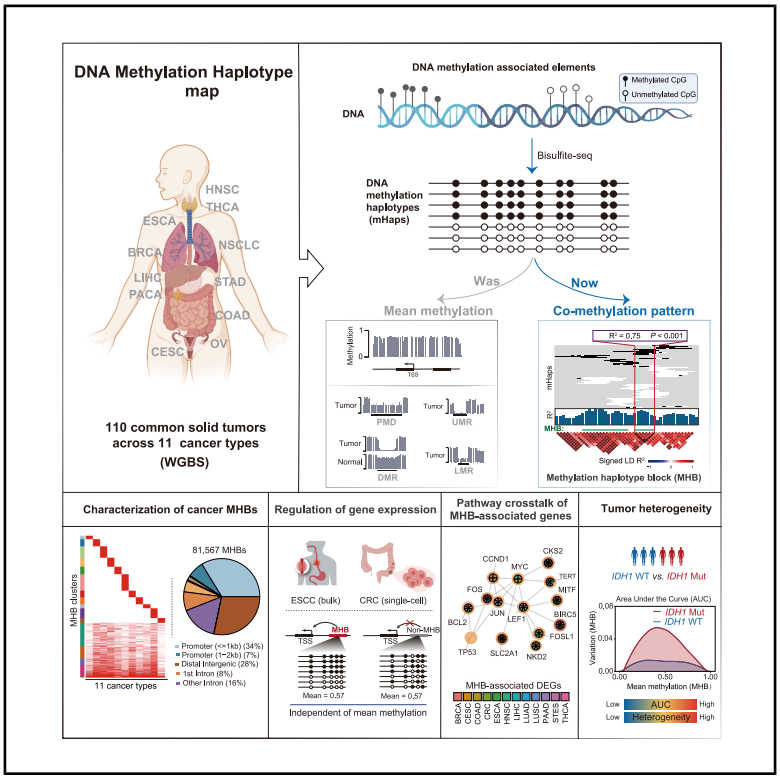
Cell Reports

Resource

Toward the DNA methylation haplotype map of 11 common solid cancers

Graphical abstract



Highlights

- Sequencing of 110 solid tumors reveals 81,567 methylation haplotype blocks (MHBs)
- MHBs are associated with gene expression-independent changes of methylation levels
- MHB-linked genes are dysregulated in cancers and enriched in oncogenic pathways
- Inter-tumor heterogeneity analysis links methylation discordance to driver mutations

Authors

Zhiqiang Zhang, Yuyang Hong, Shirong Zhang, ..., Hongcang Gu, Hai Fang, Jiantao Shi

Correspondence

fh12355@rjh.com.cn (H.F.), jtshi@sibcb.ac.cn (J.S.)

In brief

Zhang et al. present a resource profiling 110 primary tumors across 11 cancer types, identifying coordinated DNA methylation patterns that function as potential regulatory elements for gene expression. These cancer-specific methylation signatures connect to oncogenic pathways and provide new biomarkers for cancer detection.

Zhang et al., 2025, Cell Reports *44*, 116197
September 23, 2025
https://doi.org/10.1016/j.celrep.2025.116197

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

Toward the DNA methylation haplotype map of 11 common solid cancers

Zhiqiang Zhang,[1,2,7] Yuyang Hong,[3,7] Shirong Zhang,[4] Xin Zhu,[5] Leiqin Liu,[3] Xiqi Liao,[3] Hongcang Gu,[6] Hai Fang,[1,*] and Jiantao Shi[3,8,*]

[1]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
[2]School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200030, China
[3]Key Laboratory of RNA Innovation, Science and Engineering, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China
[4]Translational Medicine Research Center, Key Laboratory of Clinical Cancer Pharmacology and Toxicology Research of Zhejiang Province, Affiliated Hangzhou First People's Hospital, Cancer Center, Zhejiang University School of Medicine, Hangzhou 310022, China
[5]Key Laboratory of Head & Neck Cancer Translational Research of Zhejiang Province, Zhejiang Cancer Hospital, Hangzhou, Zhejiang Province 310022, China
[6]Anhui Province Key Laboratory of Medical Physics and Technology, Institute of Health and Medical Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, Anhui Province 230031, China
[7]These authors contributed equally
[8]Lead contact
*Correspondence: fh12355@rjh.com.cn (H.F.), jtshi@sibcb.ac.cn (J.S.)
https://doi.org/10.1016/j.celrep.2025.116197

SUMMARY

In heterogeneous tumors, adjacent CpG sites form methylation haplotype blocks (MHBs), genomic regions where methylation status reflects local epigenetic concordance. While MHBs have been implicated in gene dysregulation, their pan-cancer dynamics and clinical relevance remain unclear. We profiled 110 primary tumors across 11 common solid cancer types, identifying 81,567 MHBs. These MHBs exhibit high cancer-type specificity, with enrichment in regulatory elements. Integrative bulk and single-cell analyses reveal that MHBs associate with gene expression independently of mean methylation changes. Moreover, pan-cancer prioritization of MHB-associated differentially expressed genes highlights their roles in oncogenic pathways such as the G2/M checkpoint, MYC targets, and E2F signaling. Inter-tumor heterogeneity links MHB discordance to driver mutations and inflammatory pathways. Finally, we demonstrate that MHBs serve as effective biomarkers for cancer detection, performing competitively to existing methods. This resource positions MHBs as multimodal epigenetic regulators, bridging tumor heterogeneity, transcriptional control, and liquid biopsy diagnostics.

INTRODUCTION

Abnormal DNA methylation is a characteristic epigenetic feature of cancer, playing an essential role in cancer initiation, progression, and drug resistance.[1] DNA methylation regulators, such as DNMT3A and TET2, are recurrently mutated in acute myeloid leukemia with a population frequency of 26% and 10%, respectively.[2] Loss of TET function was shown to promote lung cancer development through impaired demethylation.[3] Targeting DNA methylation alterations, using inhibitors of DNA methyltransferases[4] or CRISPR-based epigenetic editing,[5] represents a promising cancer therapeutic strategy.[6] Characterizing aberrant DNA methylation in cancer is of great importance to develop novel therapeutic targets.

Cancer DNA methylation profiles demonstrate a unique pattern characterized by global hypomethylation across the genome, accompanied by focal hypermethylation at specific CpG islands (CGIs). This epigenetic signature contrasts markedly with that of normal cells, which maintain broad genomic methylation while exhibiting hypomethylation at CGIs.[7] Genome-wide DNA hypomethylation in cancer predominantly occurs within partially methylated domains (PMDs), which are genomic regions characterized by their association with the nuclear lamina and late replication timing.[8] Hypermethylation of promoter-associated CGIs was demonstrated to repress tumor suppressor gene expression, thereby facilitating tumorigenesis.[9] According to the distribution of mean methylation, the genome has been segmented into potential functional elements, including unmethylated regions (UMRs), low-methylation regions (LMRs), PMDs, and high-methylation regions (HMRs).[10] Notably, LMRs serve as regulatory elements that recruit cell type-specific transcription factors.[11] The characterization of these DNA methylation-associated regulatory elements, such as LMRs and PMDs, has significantly advanced our understanding of the roles of DNA methylation in maintaining the identities of cancer cells.[12]

Sequencing-based techniques such as whole-genome bisulfite sequencing (WGBS) have emerged as widely adopted methods for measuring DNA methylation at single-nucleotide resolution on individual sequencing reads.[13] Building on this, proportion of discordant reads (PDR) has been introduced to measure the degree of epigenetic heterogeneity within tumor samples,[14] demonstrating superior capability in explaining gene expression variation in chronic lymphocytic leukemia compared to average methylation. In contrast to traditional analyses that focus on mean methylation levels across cell populations,[15] WGBS enables the identification of DNA methylation haplotypes (mHaps), defined as the patterns of CpG methylation on individual DNA fragments.[16] Importantly, this approach allows for the characterization of intra-sample heterogeneity, a key feature of cancer that manifests at both genetic and epigenetic levels.[17] Single-cell methylation studies have recently provided unprecedented opportunities to evaluate cell type-specific methylomes, although these analyses predominantly focus on mean methylation levels at individual CpG sites.[18–20] Building upon the concept of mHaps, several metrics have been developed to quantify epigenetic heterogeneity. In addition to PDR, methylation concurrence ratio[21] further quantifies the fraction of unmethylated CpGs within partially methylated reads, providing a complementary perspective on methylation patterns and an insight into the relationship between DNA methylation and gene regulation. The characterization of mHap patterns has thus emerged as a powerful approach for investigating the regulatory roles of DNA methylation, particularly in cancer development and progression.[22]

Within genomic regions exhibiting heterogeneous mHaps, adjacent CpG sites may display co-methylation patterns, which can be identified through linkage disequilibrium analysis.[23] These regions containing tightly coupled CpG sites are termed methylation haplotype blocks (MHBs), analogous to the haplotype blocks observed in genomic data. MHBs have demonstrated significant potential as biomarkers for cancer detection and tissue-of-origin determination in plasma-derived DNA methylation data.[24,25] Previous research has indicated that CpG sites within individual MHBs typically exhibit highly synchronized methylation states, largely independent of tissue type or disease condition.[26] However, subsequent studies have revealed the existence of tissue-specific MHBs, which are predominantly enriched in regulatory elements crucial for maintaining tissue identity.[27,28] Despite these advances, the dynamic nature and potential functional roles of MHBs in human cancers remain largely unexplored.

Our previous study,[27] focusing on normal tissue MHBs, relied on curated datasets from diverse public sources, potentially introducing variability due to batch effects and biological heterogeneity. Currently, primary tumor DNA methylation data are mostly derived from European and American populations and are profiled using Infinium DNA methylation arrays. For example, the large-scale projects, The Cancer Genome Atlas (TCGA),[29] covered tens of thousands of samples. The demand for WGBS profiles to identify MHBs of tumors is particularly high for underrepresented populations, particularly for Chinese patients. To address these gaps, we performed WGBS on 110 solid tumors spanning 11 cancer types from Chinese individuals. Our study establishes a pan-cancer MHB atlas, revealing their regulatory associations, pathway linkages, and clinical potential as liquid biopsy biomarkers. This resource advances understanding of methylation-mediated oncogenesis while providing tools for early cancer detection development.

RESULTS

Identification of MHBs in 11 common solid cancer types

We analyzed 11 common solid cancer types (Figure 1A), which collectively represent 48% of new cancer cases worldwide[30] and 79% of new cases in China.[31] These types included head and neck squamous cell carcinoma (HNSC), thyroid carcinoma (THCA), non-small cell lung cancer (NSCLC), breast carcinoma (BRCA), esophageal carcinoma (ESCA), liver hepatocellular carcinoma (LIHC), stomach adenocarcinoma (STAD), pancreatic carcinoma (PACA), colon adenocarcinoma (COAD), ovarian serous adenocarcinoma (OV), and cervical and endocervical cancer (CESC). WGBS was performed on fresh-frozen tumor samples (n = 9–11 per cancer type) across 11 cancer types, yielding a total of 110 samples. After read deduplication, the samples achieved a median sequencing coverage of approximately 15-fold (Table S1).

To enhance MHB detection sensitivity, sequencing reads from the same cancer type were pooled to increase coverage depth. MHBs were defined as genomic regions where adjacent CpGs exhibit correlated methylation patterns. By modeling CpGs as ''epialleles,'' we quantified co-methylation using linkage disequilibrium (LD) R^2 to assess methylation concordance across individual DNA strands rather than population-averaged levels.[26] For example, when two CpG sites are covered by 52 reads, with predominantly methylated (n = 13) or unmethylated (n = 36) states, they demonstrate high correlation (LD R^2 = 0.75, $p < 0.001$; Figure 1B). We implemented a seed-and-extend algorithm in the mHapSuite tool[32]: starting with five consecutive CpGs exhibiting significant high pairwise correlation (LD $R^2 > 0.5$), we iteratively expanded the window by adding adjacent CpGs meeting the same threshold until no further extensions were possible.

Across 11 cancer types, we identified 191,395 MHBs with at least 5 CpGs required per block (Figure 1C; Table S2). Among the analyzed cancer types, gastrointestinal cancers such as ESCA, STAD, and COAD demonstrate a higher frequency of MHBs, and such a pattern is independent of the sequencing depth (Figure S1A). Considering that some MHBs are shared by different cancer types, we unionized them to obtain a total of 81,567 non-redundant MHBs at an average length of 107 bp (Figure S1B), 40.65% of which were found at promoter regions (Figure 1D). These regions were significantly enriched in known regulatory regions (permutation test, $p < 0.001$), even after controlling for confounding factors such as promoters and CGIs (Figure S1C). These non-redundant MHBs were further assigned into 16 non-overlapping clusters, including 11 cancer type-specific clusters and 5 common clusters shared by two or more cancer types (Figure 1E; Table S3). The cluster assignment was clearly driven by co-methylation rather than average methylation (Figures S1D and S1E), with consistent results at individual tumor resolution (Figure S2). Notably, ~50% of MHBs identified in

CellPress
OPEN ACCESS

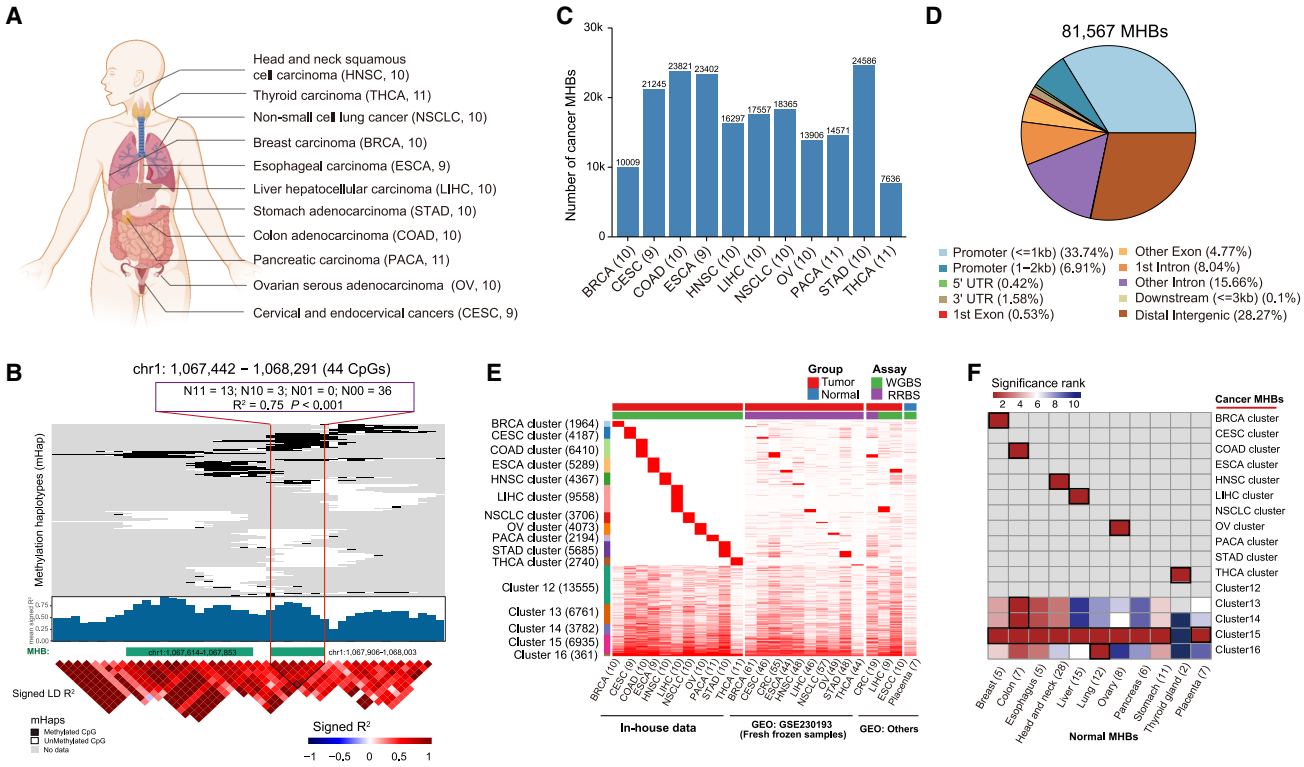


Figure 1. Identification of MHBs in cancers

(A) Cancer types included in this study. Abbreviations and numbers of sequenced WGBS samples were indicated.

(B) Representative MHBs in colon cancer (chr1: 1,067,442–1,068,291). Top: DNA methylation status of individual fragments (black: methylated CpGs; white: unmethylated CpGs). Middle: mean signed LD R^2 between adjacent CpGs. Bottom: pairwise signed LD R^2 between CpG sites. The *p* value was calculated using a binomial test.

(C) Number of MHBs identified in each cancer type.

(D) A pie chart illustrates the proportion of MHBs annotated to promoter, exonic, intronic, and intergenic regions.

(E) Assignment of MHBs into 16 non-overlapping clusters, including data from in-house samples, fresh-frozen tissues (GSE230193), and public tumor datasets.

(F) Cancer-specific MHB cluster enrichment in corresponding normal tissues. Enrichment analysis performed using LOLA package with the union of MHBs as the background. Black rectangles indicate top-ranked significant enrichment ($p < 0.05$); gray indicates non-significant results.

single tumors lacked overlap with cancer-type MHBs, highlighting inter-tumoral heterogeneity (Figure S3).

To validate the identified cancer MHBs, we analyzed independent public WGBS and reduced representation bisulfite sequencing (RRBS) datasets of colorectal cancer (CRC, n = 19),[26,33] esophageal squamous cell carcinoma (ESCC, n = 10),[34] and LIHC (n = 9)[35–37] samples (Table S4). As expected, the MHBs identified in this study showed substantial overlap with those from external datasets for all three cancer types (Figures S4A–S4C; Table S5). For example, 52.56% (n = 12,300) of ESCA MHBs identified in this study were also found in a public dataset of ESCC (Figure S4B). Similarly, we compared MHBs called in this study to those from our pan-cancer study[38] that profiled 10 cancer types using RRBS and found cancer type-specific enrichment (Figure 1E; Table S6). Previously, we have shown that the placenta shares a DNA methylation landscape with human cancers,[7] and consistent with our previous findings, here we showed that placental MHBs tended to overlap within clusters shared by multiple cancer types (Figures 1E and S4D; Tables S4 and S7).

Cancer type-specific MHBs can result from either cancer or tissue specificity. To clarify these two possibilities, we compared cancer type-specific MHBs to those identified in normal tissues by LOLA enrichment[39] (Figures 1F and S4E; Table S8). It shows that six cancer types have highest overlap percentages with corresponding normal tissues, including BRCA (7.99%), COAD (15.49%), HNSC (8.17%), LIHC (12.96%), OV (11.71%), and THCA (12.04%). Nonetheless, the majority of cancer type-specific MHBs in all cancers represent characteristics of cancer (Figure S5).

Cancer MHBs are enriched in accessible chromatin regions

Given the significant enrichment of MHBs in regulatory regions, we examined their overlap with assay for transposase-accessible chromatin using sequencing (ATAC-seq)-defined open chromatin regions from TCGA.[40] In most cancer types analyzed, more than 30% of MHBs overlapped with accessible chromatin regions (Figure 2A), which are statistically significant compared to random regions (permutation test, $p < 0.001$) (Figure S6A), indicative of

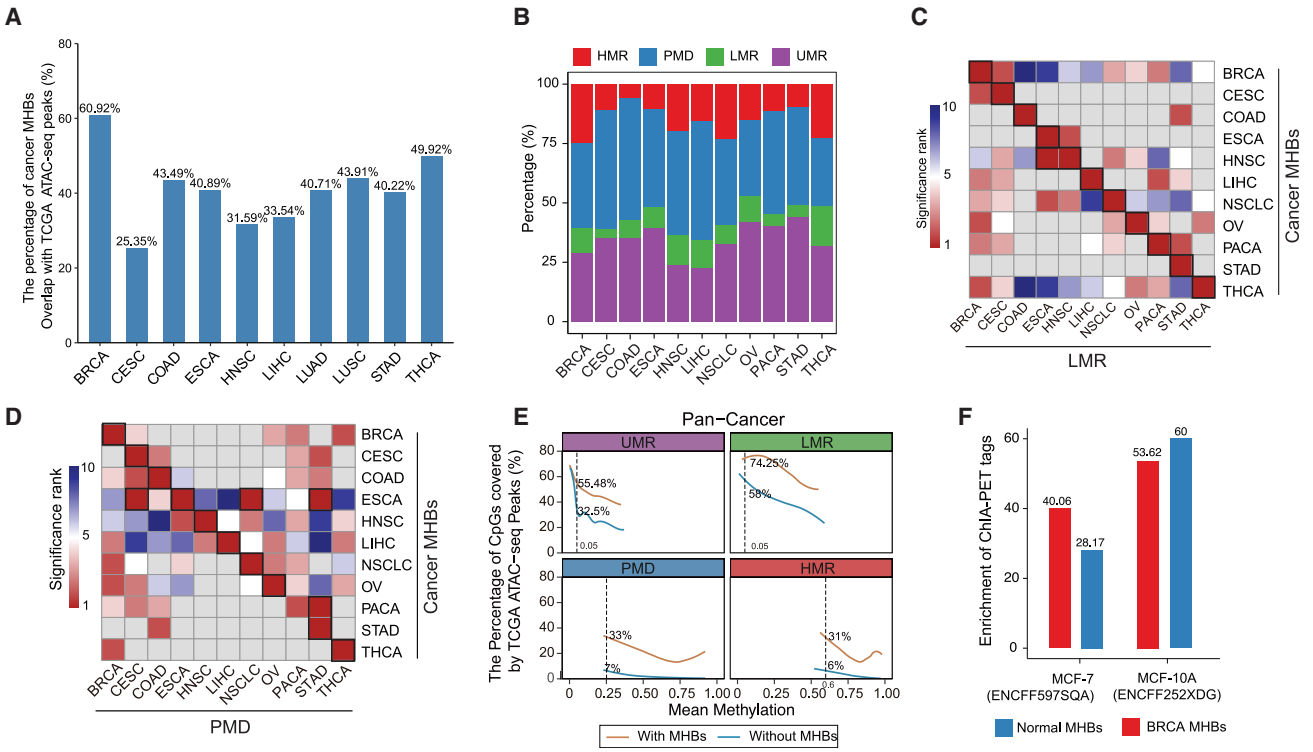


Figure 2. Cancer MHBs are enriched in accessible chromatin regions

(A) Overlapping of cancer MHBs with regions of open chromatin in matched cancer types. ATAC-seq peaks were defined previously and downloaded from NCI's Genomic Data Commons.

(B) Annotation of MHBs to regions of DNA methylation states, including UMRs, LMRs, PDR, and HMRs.

(C and D) The enrichment of MHBs in LMRs and PMDs, respectively. Enrichment test was performed by R package LOLA, with the union of MHBs serving as the background set. The resulting adjusted *p* values for significant enrichment were ranked across all cancer types (FDR < 0.01). Non-significant results are shown in gray.

(E) Genomic regions with MHBs are more enriched in regions of open chromatin when controlling for mean methylation levels. The enrichment score was calculated as the percentage of CpGs covered by open chromatin regions.

(F) Enrichment of MHBs in ChIA-PET in a disease status-specific manner. The enrichment score was calculated as the ratio of observed to expected overlaps between MHBs and ChIA-PET regions with permutation test.

regulatory roles of MHBs in cancer. It is known that segmented DNA methylation statuses such as UMRs and LMRs are enriched in active regulatory elements.[11] We thus annotated cancer MHBs with segmented DNA methylation states, including UMRs, LMRs, PMDs, and HMRs. It shows that MHBs are mainly located in UMRs and PMDs, together representing around 60% of MHBs in each cancer type (Figure 2B). While methylation states were defined solely by mean methylation levels, with PMDs comprising a substantial portion of the cancer genome (Figure S6B), region-set enrichment analysis revealed cancer type-specific enrichment of MHBs in LMRs and PMDs (Figures 2C and 2D) but not in UMRs or HMRs (Figures S6C and S6D).

We then performed enrichment analysis of open chromatin regions in MHBs across different DNA methylation states, controlling for region size and methylation levels. Each methylation state was stratified by MHB presence. MHB-containing regions showed substantially higher open chromatin enrichment across all methylation states, including UMRs and LMRs (Figure 2E). For example, at mean methylation of 0.05, ATAC-seq peaks covered 55.48% and 74.25% of CpG sites in MHB-containing UMRs and LMRs, respectively, compared to 32.5% and 58% in regions without MHBs. This pattern was consistent across all cancer types examined (Figure S7), suggesting that MHBs function as regulatory elements associated with DNA methylation patterns rather than mean methylation in human cancers.

To investigate the role of MHBs in long-range regulation, we analyzed chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) data, focusing on breast cancer as a model.[41] In MCF-7 breast cancer cells, Pol II ChIA-PET tags showed higher enrichment in breast cancer MHBs (40.06-fold) compared to normal MHBs (28.17-fold). Conversely, in non-malignant MCF-10A breast epithelial cells, Pol II ChIA-PET tags were more enriched in normal breast tissue MHBs (60-fold) than in cancer MHBs (53.62-fold) (Figures 2F and S8A). We then classified long-range chromatin contacts into 10 groups based on forward and reverse tag overlap with UMRs, LMRs, PMDs, and HMRs. Region-set enrichment analysis revealed that LMR-mediated long-range contacts showed the highest context-specific enrichment in MHBs (permutation test, $p < 0.001$; Figure S8B).

CellPress
OPEN ACCESS

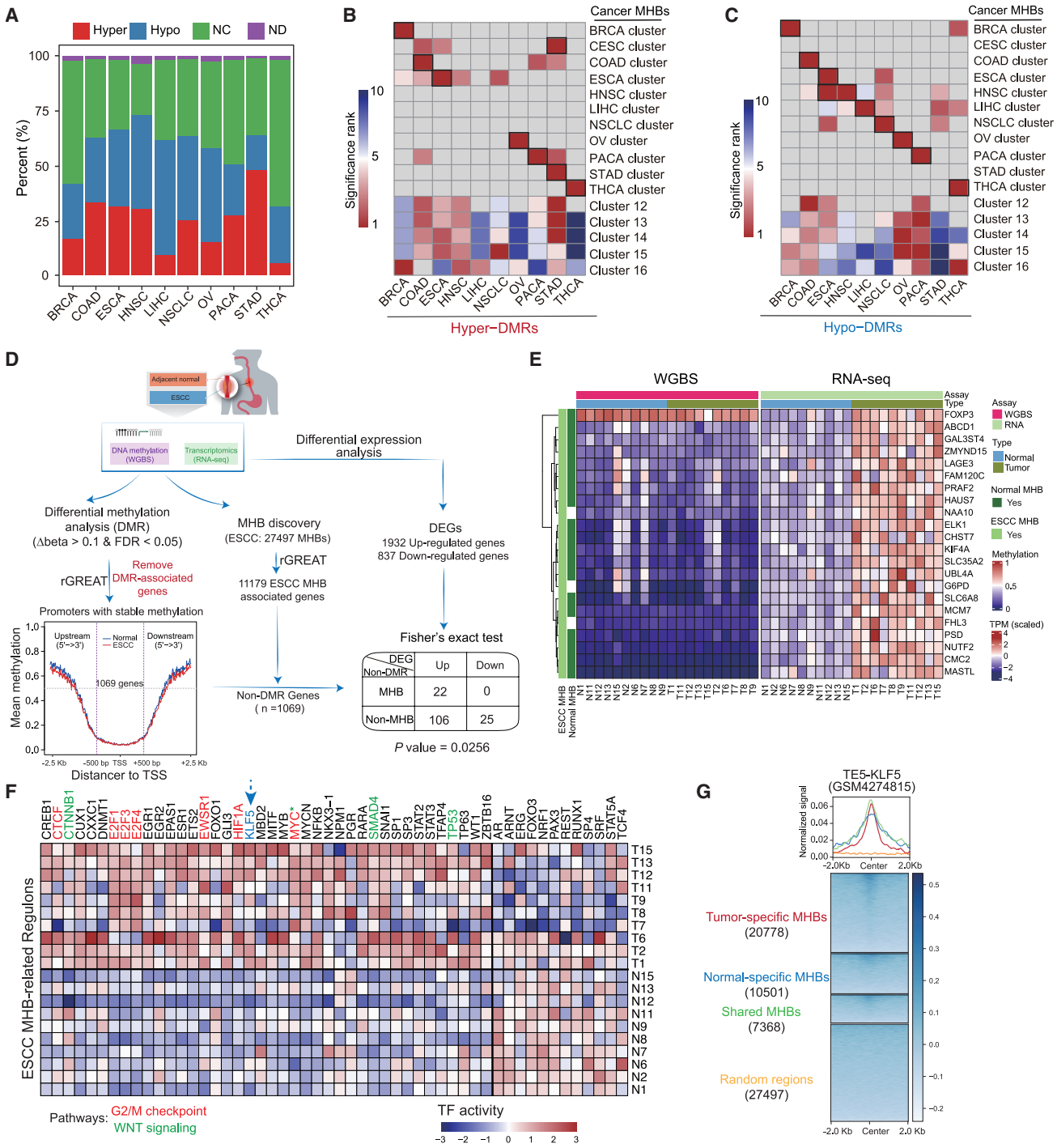


Figure 3. Cancer MHBs are associated with dysregulation of gene expression

(A) Annotation of MHBs to DMRs. Hyper, hyper-DMR; Hypo, hypo-DMR; NC, no significant change; ND, not determined.

(B and C) The enrichment of MHBs in hyper- and hypo-DMRs, respectively. Enrichment test was performed by R package LOLA, using the union of MHBs as the background. The resulting adjusted *p* values for significant enrichment were ranked across all cancer types (FDR < 0.01). The black rectangle highlights the top one. Non-significant results are shown in gray.

(D) Association of MHBs and dysregulation of gene expression in ESCC. For a dataset of ESCC (GSE149612) with 10 tumor and paired 10 normal WGBS and RNA-seq samples, DMRs were identified using Metilene (FDR < 0.05 and Δbeta > 0.1). ESCC MHBs (n = 27,497) were identified using mHapSuite. The identification of potential genes regulated by MHBs and DMRs using rGREAT with default parameters. Meanwhile, the DEGs (n = 2,769) between ESCC and adjacent normal tissue were identified by Wilcoxon rank-sum test (fold change > 2 and FDR < 0.05). Genes without DMRs were used to construct a 2 × 2 contingency table

(legend continued on next page)

Cancer MHBs are associated with dysregulation of gene expression

Differentially methylated regions (DMRs) are routinely utilized to characterize dynamic changes in DNA methylation, for example, by comparing cancer and normal samples. We therefore explored whether DMRs were covered by cancer MHBs, which are not necessarily differentially methylated in cancer compared to normal tissues. Over 50% of MHBs overlapped with hyper- or hypomethylated DMRs in most cancer types, except BRCA and THCA (Figures 3A and S9). Cancer-specific MHBs showed significant enrichment in hypermethylated DMRs in 7 cancer types (Figure 3B) and stronger cancer specificity in hypomethylated DMRs across 9 cancer types (Figure 3C).

To assess MHB-associated gene regulation independent of DMRs, we analyzed matched RNA sequencing (RNA-seq) and WGBS data from ESCC (GSE149612).[34] We identified 1,069 genes without DMRs (Δbeta > 0.1 and false discovery rate [FDR] < 0.05), out of which 153 genes were differentially expressed. Fisher's exact test demonstrated that genes with ESCC MHBs are more likely to be upregulated ($p = 2.56 \times 10^{-2}$; Figure 3D). Among 22 differentially expressed genes (DEGs) with MHBs but not DMRs, all were significantly upregulated (Figure 3E).

Given the regulatory roles of MHBs, we predicted regulators that potentially bind to MHBs and regulate gene expression, by utilizing transcriptional regulatory networks curated in the CollecTRI database.[42] Of the 1,186 transcriptional regulons in this database, 140 were significantly enriched in ESCC MHBs (FDR < 0.05), as assessed by testing for enrichment of their target genes within MHBs using the rGREAT tool.[43] Regulatory activity analysis further identified 52 regulators with the most significant activity differences between ESCC and normal esophagus samples ($p < 0.05$; Figure 3F). Regulators with increased activity were enriched in two key tumorigenic pathways: the G2/M checkpoint (MYC, HIF1A, E2F1, E2F3, E2F4, EWSR1, and CTCF) and WNT signaling (CTNNB1, MYC, SMAD4, and TP53). Chromatin immunoprecipitation sequencing (ChIP-seq) data confirmed the increased activity of the core transcription factor KLF5 in ESCC (Figure 3G).[44]

Validation of colon cancer MHBs with single-cell data

Single-cell bisulfite sequencing (scBS-seq) enables the characterization of epigenetic heterogeneity at the individual cell level. By filtering out CpG sites with allele-specific methylation, each chromosome from a single cell can be modeled as two DNA mHaps, one from the plus strand and the other from the minus strand. With this assumption, CpG site-level methylation can be converted to single-cell mHaps (sc-mHaps), typically for more than 95% of covered CpG sites (Figure 4A, upper). A key advantage of scBS-seq is the ability to measure LD between long-range CpG sites, even in the presence of missing values in between (Figure 4A, bottom).

To validate bulk-derived cancer MHBs and their potential regulatory roles, we analyzed a CRC dataset[20] that profiled DNA methylation and gene expression simultaneously at single-cell resolution. After data quality control, the single-cell methylation dataset comprised cells from five groups: primary tumor (PT, $n = 581$), lymph node metastasis (LN, $n = 346$), liver metastasis ($n = 144$), post-treatment liver metastasis ($n = 115$), and normal colon cell ($n = 93$) (Table S9). With this dataset, we reconstructed chromosome-level DNA mHaps, which were used for the identification of MHBs. For instance, two MHBs were identified in single cells at one locus (chr1: 3,229,375–3,230,473), and they largely overlapped with those identified from COAD bulk samples (Figure 4B). Additionally, MHBs identified from different types of malignant cells shared a substantial proportion, suggesting the presence of robust colon cancer-associated MHBs (Figure 4C; Table S10). Using all malignant cells, we have identified 24,087 MHBs (Table S10), covering 37.02% ($n = 8,818$) of regions identified in bulk WGBS (Figure S10). We then assessed the enrichment specificity by conducting a comparative analysis between MHBs identified from single-cell data of primary tumors and those derived from bulk sequencing across 11 different primary cancer types. LOLA enrichment analysis demonstrated that these regions are specifically enriched in COAD bulk MHBs with the highest odds ratio (odds ratio = 2.99, FDR < 1.00×10^{-300}; Figure 4D).

Extensive research has established a well-documented inverse correlation between promoter methylation and gene expression levels. To further investigate the regulatory function of MHBs, we categorized promoters by their mean methylation levels in single cells: low (<0.2), high (>0.8), and intermediate (0.2–0.8). Genes with MHBs showed significantly higher expression than those without MHBs across all methylation categories ($p < 2.22 \times 10^{-16}$; Figure 4E). Among genes containing promoter MHBs, increased MHB methylation correlated with reduced expression in both intermediate- ($p = 5.3 \times 10^{-3}$) and low- ($p < 2.22 \times 10^{-16}$) methylation groups (Figure 4F).

We continued to investigate the association between MHBs and DEGs, while controlling for changes in mean methylation levels. Comparing cells from primary tumors and normal tissues, we identified 354 DEGs that were not associated with any DMRs. These genes harboring MHBs tended to be upregulated (odds ratio = 1.67, $p = 0.03$; Figure 4G, left). Similar results were observed when comparing cells from lymph node metastasis and normal tissue (odds ratio = 1.78, $p = 3 \times 10^{-3}$; Figure 4G, right). We next focused on upregulated genes harboring MHBs in promoters, finding that 42 genes were shared between cells from primary sites and lymph nodes (Figure 4H). These genes

that separates each gene into one of four categories based on two factors, i.e., status of MHB and differential expression. Statistical significance was evaluated by Fisher's exact test.

(E) A heatmap shows the mean methylation and expression of upregulated genes that also contain MHBs.

(F) A heatmap shows the transcription factor (TF) activities between ESCC and normal esophageal tissues. TF activities were estimated by decoupleR using RNA-seq data and compared using Student's t test. Significant regulators ($p < 0.05$, $n = 52$) are shown, with G2/M checkpoint and WNT signaling genes highlighted in red and green, respectively. MYC, involved in both pathways, is highlighted in red with a green asterisk (*).

(G) KLF5 binding profiles from ESCC TE5 cell line ChIP-seq data. MHB regions were classified as ESCC tumor-specific ($n = 20,778$), normal-specific ($n = 10,501$), or shared ($n = 7,368$). Random genomic background ($n = 27,497$) was generated by "bedtools shuffle" function based on ESCC MHBs.

CellPress
OPEN ACCESS

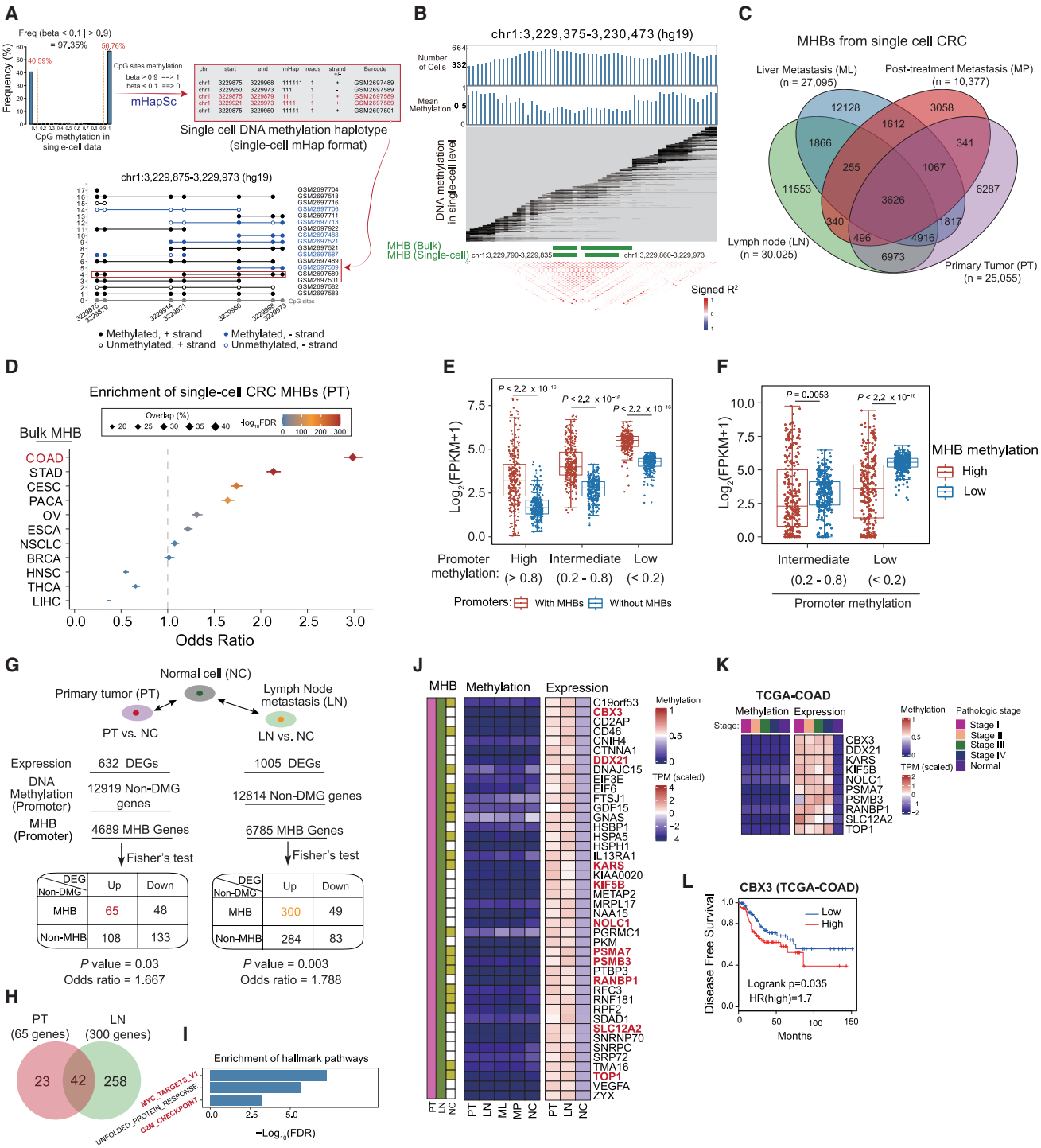


Figure 4. Validation of cancer MHBs and their associated regulatory role using single-cell data

(A) Schematic of single-cell DNA methylation haplotype (sc-mHap). Top: binary classification of CpG mean methylation (methylated: >0.9; unmethylated: <0.1) to generate sc-mHaps. Bottom: representative mHap patterns across 15 single cells.

(B) An example of COAD MHB that was also identified in region (chr1: 3,229,375–3,230,473) using CRC single-cell data.

(C) Overlap of MHBs identified across different single-cell sample types.

(D) Enrichment analysis of primary tumor single-cell MHBs against bulk sample MHBs from 11 cancer types using LOLA, with pan-cancer MHBs as the background.

(legend continued on next page)

were mainly enriched in cancer-associated pathways, including MYC targets and the G2/M checkpoint (Figure 4I). These genes were activated during cancer initiation, progression, and metastasis while exhibiting stable DNA methylation (Figure 4J). This conclusion was further supported by data from TCGA profiling COAD at different stages (Figure 4K), which identified *CBX3*, associated with disease-free survival (hazard ratio = 1.7, $p = 3.5 \times 10^{-2}$) and potentially regulated by MHBs (Figure 4L).

Cancer MHBs are associated with DEGs in pan-cancer

As demonstrated in ESCC and COAD using bulk and single-cell data, their MHBs are associated with DEGs, even for genes without annotated DMRs. To validate this observation across cancer types, we performed an integrative analysis of TCGA data for samples available with matched DNA methylation and gene expression profiles. As a sanity check, we tested the enrichment of DEGs in DMRs using the rGREAT tool. Consistent with previous studies,[7] genes associated with hypermethylated DMRs were significantly enriched in downregulated genes in 11 out of 12 cancer types (Figure S11, left; FDR < 0.05), while genes associated with hypomethylated DMRs were significantly enriched in upregulated genes in 8 out of 12 cancer types (Figure S11, right; FDR < 0.05). After excluding DMR-associated genes, MHB-containing genes showed significant enrichment among both upregulated (10 cancer types) and downregulated genes (11 cancer types), independent of mean methylation changes (Figure 5A; $p < 0.05$).

Next, we utilized our previously established priority index (Pi) prioritization approach[45,46] to elucidate candidate genes likely targeted by pan-cancer MHBs. This prioritization leveraged the information on TCGA expression profiles of pan-cancer MHB-associated genes as well as network evidence (see STAR Methods for more details). As a result, we prioritized 8,852 pan-cancer MHB-associated genes that were ranked according to their Pi rating (Table S11), with the leading prioritized genes more likely to be shared among multiple tumors (Figure S12); this finding indicated common dysfunctions across pan-cancer. For further illustration, we focused on pan-cancer MHB target genes shared by more than 6 tumors, including the well-known oncogene gene *MYC*. Among these pan-cancer MHB targets, those upregulated were significantly enriched in hallmark pathways closely related to E2F targets, the G2/M checkpoint, and MYC-activated targets (FDR < 0.01) (Figure 5B, upper). Notably, MHB-associated genes in these pathways tended to be activated in various cancer types, despite minor changes in mean methylation in most genes (Figure 5B, bottom). For instance, *MYC* activation in COAD coincided with a co-methylated region near its transcription start site (TSS) (chr8: 128,750,038–128,750,058; Figure S13A), despite a single repressive gene-body hypo-DMR (Figure S13B). This apparent contradiction suggests that the activation of *MYC* is more likely attributed to the MHB presence than conventional methylation patterns. Similarly, the gene *SLC2A1*, an MYC target activated in 10 cancer types, harbored a TSS-proximal MHB while maintaining partial methylation in NSCLC/normal lung tissues (Figure S13C), with complex DMR patterns (i.e., three hypo-DMRs in the gene body along with a hyper-DMR located upstream close to the TSS and a hypo-DMR in the distal upstream region) suggesting regulation involving mean methylation and co-methylation patterns (Figure S13D).

Furthermore, the Pi approach allowed us to identify which pan-cancer MHB-associated and highly prioritized genes mediated pathway crosstalk. The resulting 53-gene network for pathway crosstalk (Figure 5C; Table S12) was unlikely to be observed by chance based on the degree-preserving node permutation test ($p = 1.5 \times 10^{-5}$), and these crosstalk genes were enriched in pathways related to MYC, the G2/M checkpoint, and E2F targets (Figure S14). These findings indicated that the potential genes controlled by pan-cancer MHBs mainly contribute to core cancer-related pathways, leading to tumorigenesis with shared biological characteristics in pan-cancer. Furthermore, we found that 44 of these 53 MHB-associated crosstalk genes were associated with overall survival in one or more tumor types in TCGA dataset (Table S13), as exemplified by two genes (i.e., *RRM2* and *SLC2A1*) in CESC, LIHC, LUAD, and PAAD (Figure 5D).

Characterizing DNA methylation inter-tumor heterogeneity in cancers

MHBs exhibit coordinated methylation patterns, with adjacent CpGs showing high correlations in methylation levels, as

(E) Association between gene expression and presence of MHBs in promoter regions. For each single cell, promoters were categorized into three groups based on mean methylation: high (>0.8), intermediate (0.2–0.8), and low (<0.2). Within each group, the expression of genes with MHBs in their promoters was compared to those without MHBs. Statistical significance was evaluated using the Wilcoxon rank-sum test. Promoters were defined as ± 1 kb region around the TSS.

(F) Impact of MHB methylation levels on gene expression in intermediate- and low-methylation promoter groups. Analysis excluded high-methylation promoters due to limited MHB-containing genes. Significance was assessed by Wilcoxon rank-sum test.

(G) Association of MHBs and dysregulation of gene expression in single-cell CRC. The DEGs (fold change > 2 and FDR < 0.05) and differentially methylated promoters (Δbeta > 0.1 and FDR < 0.05) were identified using the Wilcoxon rank-sum test. Genes without differentially methylated promoters were used to construct a 2 × 2 contingency table that separates each gene into one of four categories based on two factors, i.e., status of the MHB and differential expression. Statistical significance was evaluated by Fisher's exact test.

(H) Venn diagram showing the number of upregulated genes in PT and LN tumors compared to normal colon tissues that harbor MHBs but lack DMRs in the promoter regions.

(I) Pathway enrichment of 42 shared MHB-related genes in PT and LN. The online tool in MSigDB was used and top 3 hallmark pathways were shown.

(J) A heatmap shows the mean methylation and expression of 42 shared genes that also contains MHB across single cell. The genes annotated in MYC targets and G2/M checkpoint pathways were colored in red.

(K) The DNA methylation and gene expression profiles of the 10 genes highlighted in (J) were validated using the TCGA-COAD dataset at different pathological stages.

(L) The Kaplan-Meier survival curves comparing disease-free survival (DFS) between patient subgroups stratified by the high/low expression (median cutoff) of *CBX3* gene in TCGA-COAD dataset. *p* value was calculated using the log rank test, and the hazard ratio (HR) was calculated using univariable Cox regression analysis.

CellPress
OPEN ACCESS

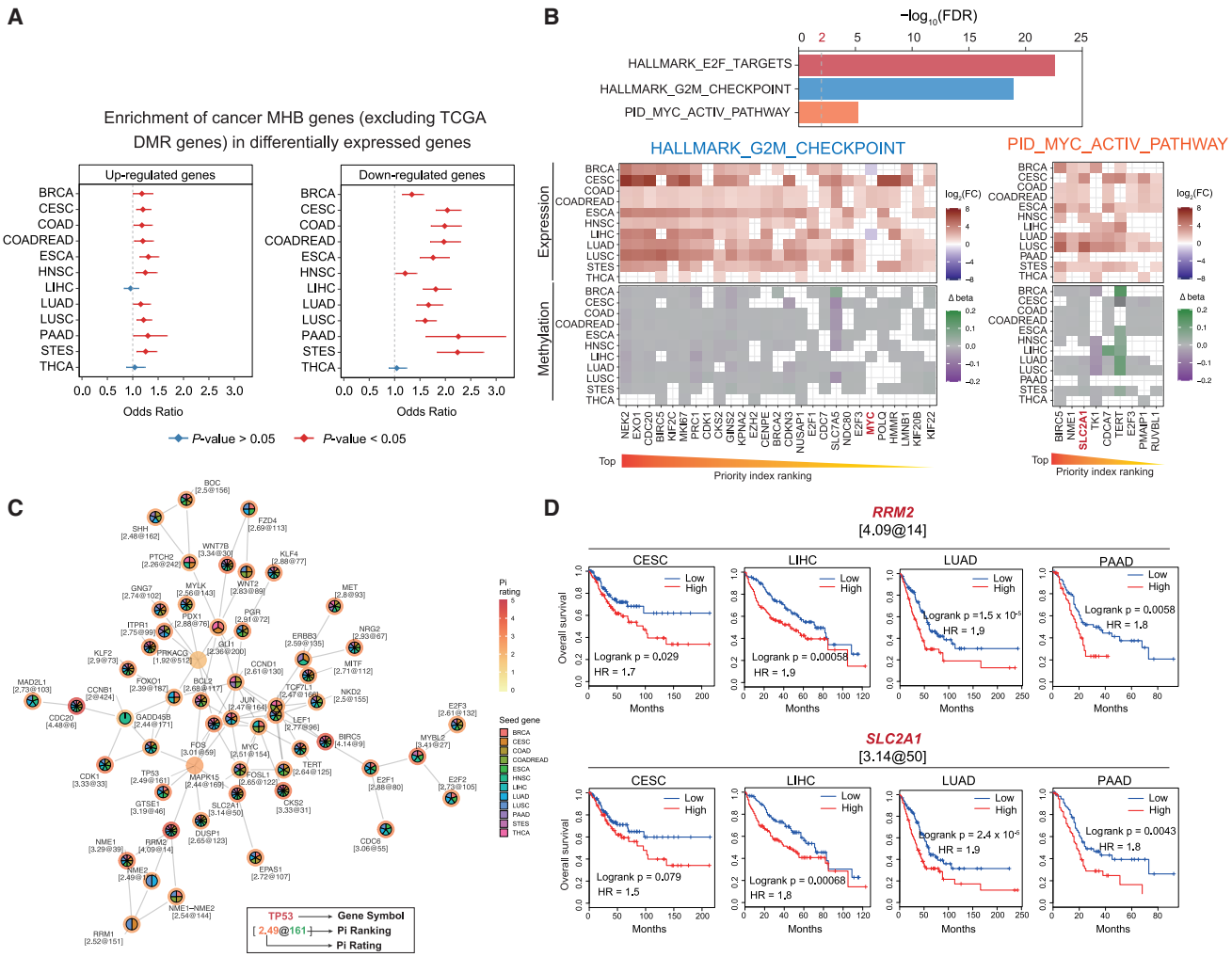


Figure 5. Cancer MHBs are associated with differentially expressed genes in pan-cancer

(A) Enrichment of cancer MHB-associated genes in differentially expressed genes. The forest plot shows the enrichment of cancer MHB-associated genes, excluding DMR-related genes, in DEGs. Statistical significance was evaluated using the Fisher's exact test, with p value < 0.05 labeled in red and p value > 0.05 labeled in blue, respectively.

(B) Pathway enrichment of MHB-associated genes in pan-cancer. An integrative prioritization approach utilizing network information (see STAR Methods) was employed to prioritize MHB-related genes shared by more than six cancer types. Upper: the gene set enrichment of shared upregulated genes. Bottom: the differences in methylation and expression profiles of selected genes, annotated in the G2/M checkpoint or MYC activity pathway across cancer types.

(C) Pathway crosstalk analysis for pan-cancer. The optimal subnetwork was identified by integrating the target priority rating information with pathway-derived gene interactions, which were merged from KEGG pathways.

(D) The Kaplan-Meier survival curves compare overall survival between patient subgroups stratified by high/low expression (median cutoff) of *RRM2* and *SLC2A1* genes in CESC, LIHC, LUAD, and PAAD of TCGA dataset. p values were calculated using the log rank test, and the HR was calculated through univariable Cox regression analysis.

observed in DNA methylation array data. Analysis of TCGA data revealed median pairwise CpG correlations >0.85 within cancer MHBs across all cancer types, significantly exceeding correlations in CGIs or random genomic regions (most < 0.6; Figure 6A). To quantify inter-tumor heterogeneity, we calculated the area under the curve (AUC) of methylation concordance per tumor, where lower AUC values reflect higher intra-block coordination. For example, COAD tumors exhibited stronger concordance in MHBs (AUC = 0.009) versus random regions (AUC = 0.03; Figure 6B).

We next linked methylation heterogeneity to transcriptional changes by comparing tumors with extreme AUC values (low vs. high: median cutoff). Tumors with increased concordance (lower AUC) showed significant upregulation of genes (fold change > 2 and FDR < 0.05; Figure 6C), enriched in immune-related pathways (such as allograft rejection, inflammatory response, and IL6-JAK-STAT3 signaling) and epithelial-mesenchymal transition across multiple cancers (Figure 6D). Downregulated genes lacked shared pathway enrichment (Figure S15).

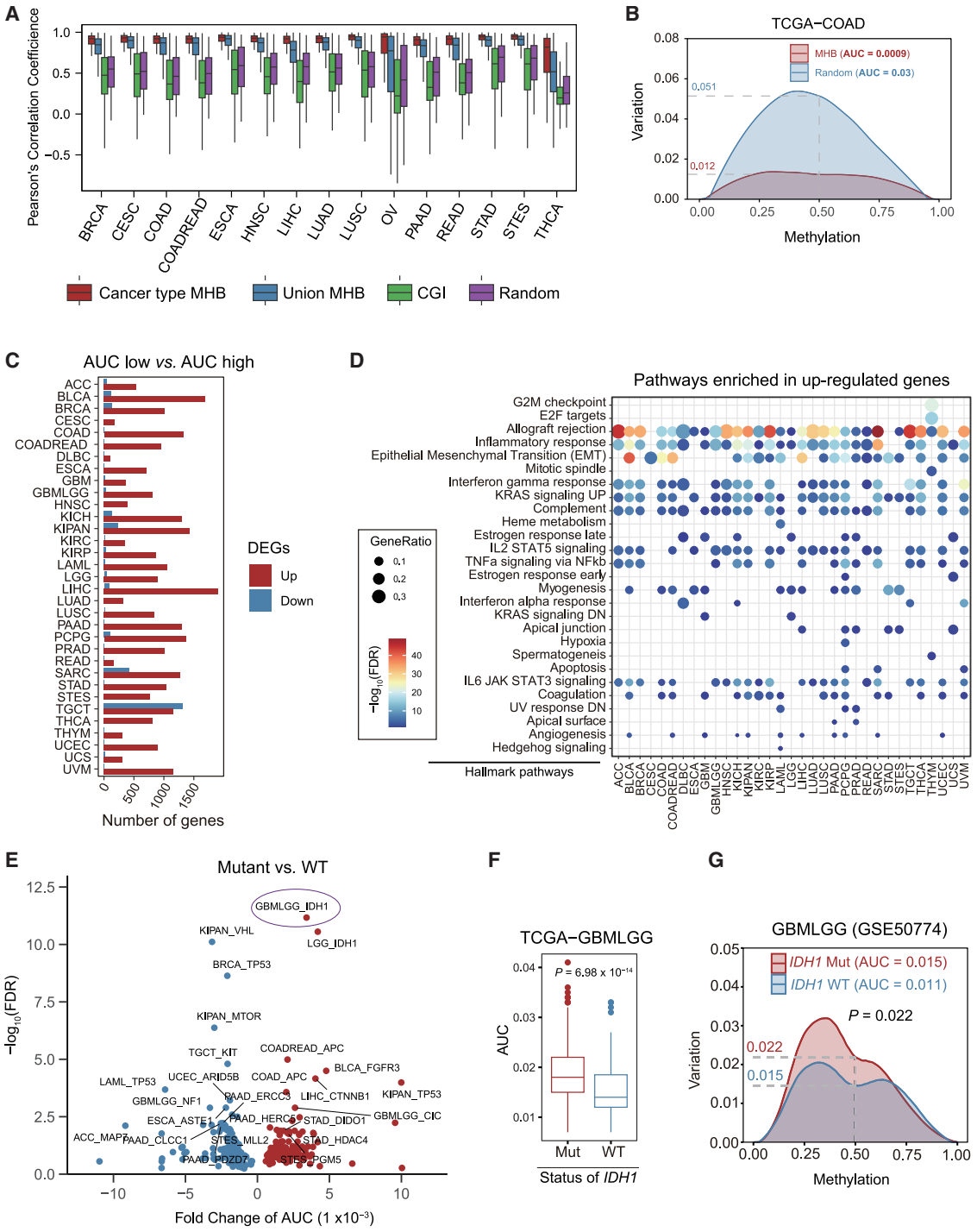


Figure 6. Characterization of DNA methylation inter-tumoral heterogeneity in cancers

(A) Boxplots showing Pearson correlation coefficients of average methylation between pairwise CpGs in TCGA 450K array data. Regions analyzed: pan-cancer MHBs (all cancer types, union MHBs), cancer type-specific MHBs, CGIs, and random genomic regions.

(B) Quantification of concordant methylation levels. Using COAD as an example, locally estimated scatterplot smoothing (LOESS) curves were fitted between mean methylation and variance for MHBs versus random regions. The AUC quantifies concordance (lower AUC = higher concordance).

(C) Number of DEGs between tumors with low/high AUCs. Red: upregulated; blue: downregulated.

(legend continued on next page)

We also analyzed somatic mutations associated with AUC changes to identify the potential drivers of methylation heterogeneity (FDR < 0.05; Figure 6E; Table S14). *IDH1*-mutant glioblastoma and low-grade glioma tumors exhibited reduced concordance (higher AUC; $p = 6.98 \times 10^{-14}$; Figure 6F), validated in an independent cohort (mutant: AUC = 0.015 vs. wild type: AUC = 0.011; $p = 2.2 \times 10^{-2}$; Figure 6G). These findings implicate genetic alterations in shaping methylation heterogeneity across tumors.

Cancer MHBs as biomarkers for non-invasive cancer detection from plasma DNA

MHBs were enriched with fully methylated fragments and demonstrated to preserve higher signal-to-noise ratio for non-invasive cancer detection.[26] We calculated methylation haplotype load (MHL)[26] and methylation block score (MBS)[47] to quantify fully methylated *k*-mers within MHBs. For example, a universal cancer marker identified in this study (chr1: 119,527,091–119,527,476) was hypermethylated in all 14 cancer types (Figure S16). When *k*-mer width was set to 4, the frequency of 16 different *k*-mers in tumor and normal samples was compared. As expected, the relative abundance of fully methylated *k*-mer achieved the highest odds ratio, i.e., 10.28 times of that in normal tissue (Figure 7A). Using cancer MHBs coupled with DNA methylation metric MHL/MBS, we tested a published dataset[48] with cell-free DNA methylation profiles from six gastrointestinal cancer types, including CRC (n = 40), ESCA (n = 60), esophageal adenocarcinoma (EAC, n = 12), gastric cancer (GC, n = 37), hepatocellular carcinoma (HCC, n = 43), and pancreatic ductal adenocarcinoma (PDAC, n = 74), as well as healthy individuals (n = 46) (Table S15). The original design targeted 67,832 DMRs, which were covered by 20,946 cancer MHBs. Our cancer prediction models achieved the best performance for CRC (MHL AUC = 0.97; MBS AUC = 0.95), EAC (MHL AUC = 0.80; MBS AUC = 0.80), ESCC (MHL AUC = 0.95; MBS AUC = 0.96), GC (MHL AUC = 0.88; MBS AUC = 0.85), HCC (MHL AUC = 0.97; MBS AUC = 0.97), and PDAC (MHL AUC = 0.79; MBS AUC = 0.78) (Figure 7B). Even when specificity was set to 98%, sensitivity still reached 56.07%–91%, in contrast to 36.89%–72.22% in CpG-level mean methylation-based prediction (Figure 7B), excluding PDAC. We also benchmarked CancerDetector,[49] a method that estimates cancer fraction using read-level statistical inference. Our analysis revealed that MHBs utilizing methylation pattern-based MHL/MBS outperform CancerDetector in detecting cancer from plasma DNA across all cancer types (Figure S17). These results highlight the superiority of co-methylation patterns over bulk methylation metrics, establishing MHB-derived biomarkers as powerful tools for non-invasive early cancer detection.

DISCUSSION

DNA methylation patterns in cancer have been extensively characterized in TCGA project, primarily using the HM450K BeadChip array. This method predominantly measures average methylation in CGIs and promoter regions. Cancer DMRs were defined by comparing cancer methylomes to those of normal tissues. Most attention has focused on hyper-DMRs that repress tumor suppressor genes.[50] With the advancement of sequencing technologies, epigenetic heterogeneity of cancer has been depicted by the analysis of disordered methylation at the fragment level.[14,26] The methylation of a region may evolve overtime and achieve fixation, i.e., fully unmethylated or methylated states, or a mixture of these two states, forming MHBs. While MHBs have been primarily used as potential biomarkers, their functions remain to be fully elucidated.[26] In our previous work, we demonstrated that MHBs were enriched in enhancers and tissue-specific genes in normal tissues.[50] Building upon this foundation, the present study profiles 11 common cancer types using WGBS to construct an atlas of DNA mHaps and a landscape of cancer MHB in common solid cancers.

Our results demonstrated that MHBs are features of cell identity rather than static genomic structures shared across tissue types and disease statuses. When compared to normal tissues, most cancer MHBs were not observed in non-malignant counterparts, suggesting that MHBs are specific to disease status. We validated our findings using external datasets for CRC, ESCC, and LIHC. Most convincingly, our discoveries were corroborated by single-cell data, which represent the ultimate method for resolving cancer heterogeneity. The CRC MHBs identified from scBS-seq were significantly and specifically enriched in colon cancer among the 11 cancer types studied. These results collectively demonstrate that MHBs are features of cell identity that are both tissue- and disease status specific.

We investigated the regulatory functions of MHBs from multiple angles. Initially, we found that a significant number of MHBs overlap with open chromatin regions, demonstrating a higher enrichment compared to conventional DNA methylation-associated regulatory regions like UMRs and LMRs. Second, MHBs are involved in long-range chromatin interactions, as evidenced by their enrichment in ChIA-PET loops in a disease status-specific manner. Specifically, cancer MHBs were more enriched in cancer-associated ChIA-PET loops, while normal MHBs were more enriched in normal tissue ChIA-PET loops. Third, cancer MHBs were enriched in DMRs in a cancer type-specific manner. Fourth, MHBs were significantly associated with gene expression independent of their DMR status. Most importantly, this observation is validated by single-cell data. We utilized a unique single-cell dataset, which profiled DNA methylation and gene expression simultaneously at single-cell resolution. Analysis of this dataset

(D) Enriched pathways in upregulated DEGs, hallmark pathways from MSigDB.
(E) Volcano plot associating driver mutations with concordant methylation levels (measured by AUC). Statistical significance was calculated by Student's two-sided t test (FDR < 0.05). Red: AUC increased; blue: AUC decreased.
(F) Boxplots showing *IDH1* mutation significantly associated with methylation concordance in TCGA dataset. p value was calculated using two-sided Student's t test.
(G) Independent validation of *IDH1* mutation effect (GSE50774). Mean methylation-variance plots compare *IDH1*-mutant (Mut, n = 13) and wild-type (WT, n = 13) samples. p value was calcuated using two-sided Student's t test.

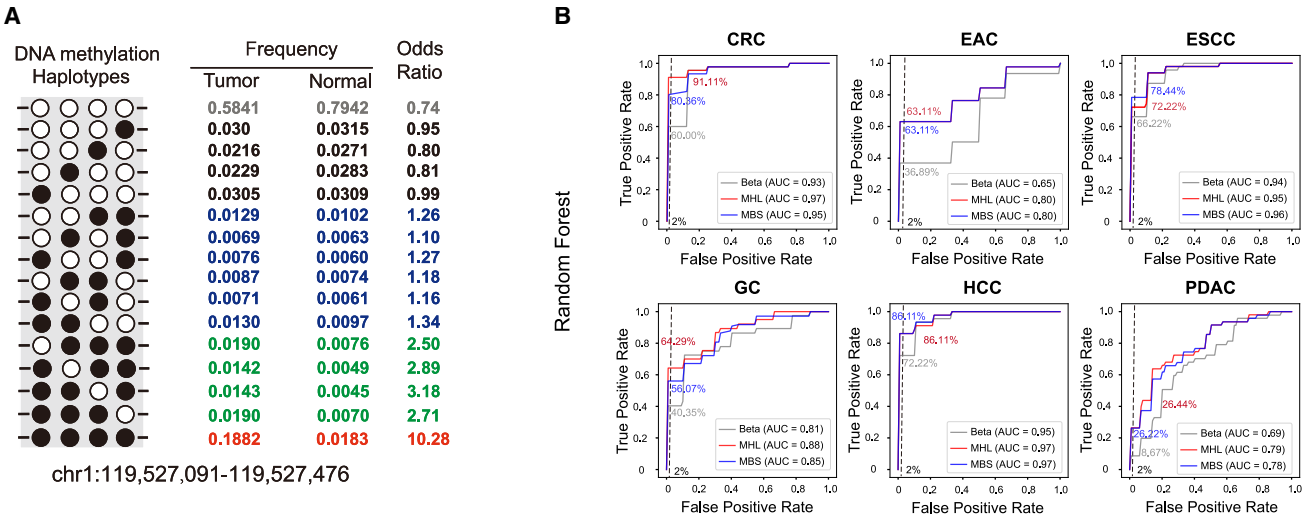


Figure 7. Cancer MHB-based non-invasive cancer detection with plasma DNA

(A) MHB regions as biomarkers for cancer detection. A universal cancer marker (chr1: 119,527,091–119,527,476) identified from cancer MHBs, the *k*-mers width was set to 4, and the frequency of 16 different *k*-mers in tumor and normal samples was displayed.

(B) Cancer detection performance using random forest models based on MHL, MBS, and CpG-level mean methylation in plasma DNA. Receiver operating characteristic (ROC) curves and AUC values from 5-fold cross-validation comparing CRC, ESCC, HCC, GC, EAC, and PDAC against normal samples.

revealed that among genes without annotated DMRs, those containing MHBs were significantly upregulated in CRC, suggesting the regulatory role of MHBs independent of mean methylation changes. Collectively, results from bulk bisulfite sequencing (BS-seq), scBS-seq, and HM450K array demonstrated that these cancer MHB-containing genes were enriched in oncogenic pathways such as the G2/M checkpoint and MYC pathway.

Clinically, we demonstrated that MHBs achieve superior diagnostic performance in liquid biopsies versus conventional metrics. Using DNA mHap-level metrics (MHL and MBS) outperformed mean methylation-based approaches (i.e., using beta values). Comparative analysis against established methods like CancerDetector highlighted the advantage of co-methylation patterns, underscoring MHBs' translational potential.

Limitations of the study

We acknowledge several limitations of our work. First, we only profiled tumor samples, lacking corresponding data from adjacent normal tissues. Consequently, the identification of DMRs and MHBs in normal tissues relied on curated datasets from diverse sources, potentially introducing variability due to batch effects and biological heterogeneity. Additionally, the analytical tool used does not distinguish between paternal and maternal fragments, which might result in the incorrect classification of imprinted regions as MHBs. In a single cell, each chromosome potentially has four DNA mHaps—two from the paternal chromosome (forward and reverse strands) and two from the maternal chromosome (forward and reverse strands). Fully resolving these haplotypes would require phased SNP calling, which remains challenging for both bulk BS-seq and scBS-seq data. In bulk BS-seq analysis, we treat each sequencing read as one fragment of an mHap without explicitly addressing allele-specific information. For scBS-seq, we simplified the approach by converting mean methylation values to binary states. This allows us to identify regions with coordinated methylation patterns, though this strategy masks CpG sites with allele-specific methylation in each single cell. Our previous research suggests that approximately 1% of MHBs in normal tissues are in imprinted regions.[27]

To explore the potential roles of MHBs in pan-cancer analysis, we utilized gene expression profiles from TCGA. These profiles were not entirely matched to the samples sequenced in this study, potentially introducing bias due to cancer heterogeneity. In addition, we observed that MHBs were associated with gene expression changes when controlling for changes in mean methylation, possibly through the regulation of transcription factors enriched in MHBs. However, this observation lacks experimental validation. Lastly, we leveraged cancer MHBs as biomarkers for non-invasive cancer detection from plasma DNA, but using targeted BS-seq data did not cover all the regions of cancer MHBs, leading to bias in cancer prediction. Despite these constraints, our work establishes MHBs as key players in maintaining cancer identity through coordinated epigenetic regulation, providing a foundation for future multi-omics exploration of mHaps in oncology.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Jiantao Shi (jtshi@sibcb.ac.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The mHap data for 110 human tumor samples profiled in this study have been deposited at GEO (GSE212391). Processed mHap data from

publicly available BS-seq datasets for tumor tissue, normal tissue, and cfDNA have been deposited at Zenodo (DOI: https://doi.org/10.5281/zenodo.16496803 and https://doi.org/10.5281/zenodo.16485955, respectively).
- All original code is available at Zenodo (DOI: https://doi.org/10.5281/zenodo.16496342).
- Additional information required to reanalyze the data is available from the lead contact upon request.

ACKNOWLEDGMENTS

This study was supported in part by the National Natural Science Foundation of China (grant ID: 32270691 and 32170663), Noncommunicable Chronic Diseases-National Science and Technology Major Project (grant ID: 2024ZD0519600), and Innovative Research Team of High-level Local Universities in Shanghai. The results presented in this study are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. Schematic representations of human tissues (Figure 1A) were created with BioRender: https://biorender.io. We are grateful for the high-performance computing center of the Center for Excellence in Molecular Cell Science (CEMCS), CAS, for its support in data processing.

AUTHOR CONTRIBUTIONS

Conceptualization, J.S. and H.F.; methodology, J.S., Y.H., H.F., and Z.Z.; computational, multi-omics, and statistical analyses, Z.Z., Y.H., and L.L.; software development, Y.H. and Z.Z.; writing – original draft, J.S., Z.Z., H.G., and H.F.; writing – review and editing, J.S., H.F., Z.Z., Y.H., H.G., X.L., and S.Z.; funding acquisition, J.S. and H.F.; resources, S.Z. and X.Z.; supervision, J.S. and H.F.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT/Claude to enhance its readability. After using this tool or service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - WGBS library construction and sequencing
  - External data collection
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Preprocessing of BS-seq data
  - Identification of DNA methylation-associated regions
  - Differential gene expression analysis
  - Identification of MHBs from bulk and single-cell BS-seq data
  - Genomic region enrichment analysis
  - Transcription factor activity inference
  - Priority index (Pi) prioritization for pan-cancer MHB target identification
  - Pathway enrichment analysis
  - Survival analysis
  - Quantifying the level of concordant methylation
  - Machine learning-based cancer detection

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2025.116197.

Received: November 15, 2024
Revised: May 18, 2025
Accepted: August 4, 2025

REFERENCES

1. Feinberg, A.P. (2018). The Key Role of Epigenetics in Human Disease Prevention and Mitigation. N. Engl. J. Med. *378*, 1323–1334.
2. Cancer Genome Atlas Research Network; Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., Robertson, A.G., Hoadley, K., Triche, T.J., Jr., Laird, P.W., et al. (2013). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. N. Engl. J. Med. *368*, 2059–2074.
3. Xu, Q., Wang, C., Zhou, J.X., Xu, Z.M., Gao, J., Sui, P., Walsh, C.P., Ji, H., and Xu, G.L. (2022). Loss of TET reprograms Wnt signaling through impaired demethylation to promote lung cancer development. Proc. Natl. Acad. Sci. USA *119*, e2107599119.
4. Jang, H.J., Shah, N.M., Maeng, J.H., Liang, Y., Basri, N.L., Ge, J., Qu, X., Mahlokozera, T., Tzeng, S.C., Williams, R.B., et al. (2024). Epigenetic therapy potentiates transposable element transcription to create tumor-enriched antigens in glioblastoma cells. Nat. Genet. *56*, 1903–1913.
5. Wang, D., Zhang, F., and Gao, G. (2020). CRISPR-Based Therapeutic Genome Editing: Strategies and In Vivo Delivery by AAV Vectors. Cell *181*, 136–150.
6. Lee, A.V., Nestler, K.A., and Chiappinelli, K.B. (2024). Therapeutic targeting of DNA methylation alterations in cancer. Pharmacol. Ther. *258*, 108640.
7. 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.
8. Zhou, W., Dinh, H.Q., Ramjan, Z., Weisenberger, D.J., Nicolet, C.M., Shen, H., Laird, P.W., and Berman, B.P. (2018). DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat. Genet. *50*, 591–602.
9. Baylin, S.B. (2005). DNA methylation and gene silencing in cancer. Nat. Clin. Pract. Oncol. *2*, 4–11.
10. Burger, L., Gaidatzis, D., Schübeler, D., and Stadler, M.B. (2013). Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. *41*, e155.
11. Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.
12. Nishiyama, A., and Nakanishi, M. (2021). Navigating the DNA methylation landscape of cancer. Trends Genet. *37*, 1012–1027.
13. Harris, R.A., Wang, T., Coarfa, C., Nagarajan, R.P., Hong, C., Downey, S.L., Johnson, B.E., Fouse, S.D., Delaney, A., Zhao, Y., et al. (2010). Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat. Biotechnol. *28*, 1097–1105.
14. 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.
15. Zhou, W., Hinoue, T., Barnes, B., Mitchell, O., Iqbal, W., Lee, S.M., Foy, K.K., Lee, K.-H., Moyer, E.J., VanderArk, A., et al. (2022). DNA methylation dynamics and dysregulation delineated by high-throughput profiling in the mouse. Cell Genom. *2*, 100144.

16. 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.
17. Marusyk, A., Janiszewska, M., and Polyak, K. (2020). Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. Cancer Cell *37*, 471–484.
18. Luo, C., Liu, H., Xie, F., Armand, E.J., Siletti, K., Bakken, T.E., Fang, R., Doyle, W.I., Stuart, T., Hodge, R.D., et al. (2022). Single nucleus multiomics identifies human cortical cell regulatory genome diversity. Cell Genom. *2*, 100107.
19. Flint, J., Heffel, M.G., Chen, Z., Mefford, J., Marcus, E., Chen, P.B., Ernst, J., and Luo, C. (2023). Single-cell methylation analysis of brain tissue prioritizes mutations that alter transcription. Cell Genom. *3*, 100454.
20. Bian, S., Hou, Y., Zhou, X., Li, X., Yong, J., Wang, Y., Wang, W., Yan, J., Hu, B., Guo, H., et al. (2018). Single-cell multiomics sequencing and analyses of human colorectal cancer. Science *362*, 1060–1063.
21. 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.
22. 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*, 400–409.
23. Shoemaker, R., Deng, J., Wang, W., and Zhang, K. (2010). Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res. *20*, 883–889.
24. Acton, R.J., and Bell, C.G. (2017). Cancer detection and tissue of origin determination with novel annotation and scoring of cell-free methylated DNA. AME Med. J. *2*, 110.
25. Jia, Z., Wang, Y., Xue, J., Yang, X., Bing, Z., Guo, C., Gao, C., Tian, Z., Zhang, Z., Kong, H., et al. (2021). DNA methylation patterns at and beyond the histological margin of early-stage invasive lung adenocarcinoma radiologically manifested as pure ground-glass opacity. Clin. Epigenet. *13*, 153.
26. 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.
27. Feng, Y., Zhang, Z., Hong, Y., Ding, Y., Liu, L., Gao, S., Fang, H., and Shi, J. (2023). A DNA methylation haplotype block landscape in human tissues and preimplantation embryos reveals regulatory elements defined by comethylation patterns. Genome Res. *33*, 2041–2052.
28. Ma, Q., Xu, Z., Lu, H., Xu, Z., Zhou, Y., Yuan, B., and Ci, W. (2018). Distal regulatory elements identified by methylation and hydroxymethylation haplotype blocks from mouse brain. Epigenetics Chromatin *11*, 75.
29. Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell *173*, 291–304.e6.
30. Siegel, R.L., Miller, K.D., Fuchs, H.E., and Jemal, A. (2022). Cancer statistics, 2022. CA Cancer J. Clin. *72*, 7–33.
31. Xia, C., Dong, X., Li, H., Cao, M., Sun, D., He, S., Yang, F., Yan, X., Zhang, S., Li, N., and Chen, W. (2022). Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chin. Med. J. *135*, 584–590.
32. 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.
33. Lin, X., Stenvang, J., Rasmussen, M.H., Zhu, S., Jensen, N.F., Tarpgaard, L.S., Yang, G., Belling, K., Andersen, C.L., Li, J., et al. (2015). The potential role of Alu Y in the development of resistance to SN38 (Irinotecan) or oxaliplatin in colorectal cancer. BMC Genom. *16*, 404.
34. Cao, W., Lee, H., Wu, W., Zaman, A., McCorkle, S., Yan, M., Chen, J., Xing, Q., Sinnott-Armstrong, N., Xu, H., et al. (2020). Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma. Nat. Commun. *11*, 3675.
35. Li, X., Liu, Y., Salz, T., Hansen, K.D., and Feinberg, A. (2016). Whole-genome analysis of the methylome and hydroxymethylome in normal and malignant lung and liver. Genome Res. *26*, 1730–1741.
36. Jenkinson, G., Pujadas, E., Goutsias, J., and Feinberg, A.P. (2017). Potential energy landscapes identify the information-theoretic nature of the epigenome. Nat. Genet. *49*, 719–729.
37. Heyn, H., Vidal, E., Ferreira, H.J., Vizoso, M., Sayols, S., Gomez, A., Moran, S., Boque-Sastre, R., Guil, S., Martinez-Cardus, A., et al. (2016). Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. Genome Biol. *17*, 11.
38. Zhang, S., He, S., Zhu, X., Wang, Y., Xie, Q., Song, X., Xu, C., Wang, W., Xing, L., Xia, C., et al. (2023). DNA methylation profiling to determine the primary sites of metastatic cancers using formalin-fixed paraffin-embedded tissues. Nat. Commun. *14*, 5686.
39. Sheffield, N.C., and Bock, C. (2016). LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics *32*, 587–589.
40. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*, eaav1898.
41. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
42. Müller-Dott, S., Tsirvouli, E., Vazquez, M., Ramirez Flores, R.O., Badia-I-Mompel, P., Fallegger, R., Türei, D., Lægreid, A., and Saez-Rodriguez, J. (2023). Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. Nucleic Acids Res. *51*, 10934–10949.
43. Gu, Z., and Hübschmann, D. (2023). rGREAT: an R/bioconductor package for functional enrichment on genomic regions. Bioinformatics *39*, 17–19.
44. Li, L.-Y., Yang, Q., Jiang, Y.-Y., Yang, W., Jiang, Y., Li, X., Hazawa, M., Zhou, B., Huang, G.-W., Xu, X.-E., et al. (2021). Interplay and cooperation between SREBF1 and master transcription factors regulate lipid metabolism and tumor-promoting pathways in squamous cancer. Nat. Commun. *12*, 4362.
45. Fang, H., ULTRA-DD Consortium; De Wolf, H., Knezevic, B., Burnham, K. L., Osgood, J., Sanniti, A., Lledó Lara, A., Kasela, S., De Cesco, S., et al. (2019). A genetics-led approach defines the drug target landscape of 30 immune-related traits. Nat. Genet. *51*, 1082–1091.
46. Fang, H., and Knight, J.C. (2022). Priority index: Database of genetic targets in immune-mediated disease. Nucleic Acids Res. *50*, D1358–D1367.
47. 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.
48. Kandimalla, R., Xu, J., Link, A., Matsuyama, T., Yamamura, K., Parker, M. I., Uetake, H., Balaguer, F., Borazanci, E., Tsai, S., et al. (2021). EpiPanGI Dx: A Cell-free DNA Methylation fingerprint for the early detection of Gastrointestinal cancers. Clin. Cancer Res. *27*, 6135–6144.
49. Li, W., Li, Q., Kang, S., Same, M., Zhou, Y., Sun, C., Liu, C.-C., Matsuoka, L., Sher, L., Wong, W.H., et al. (2018). CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. Nucleic Acids Res. *46*, e89.
50. Herman, J.G., and Baylin, S.B. (2003). Gene Silencing in Cancer in Association with Promoter Hypermethylation. N. Engl. J. Med. *349*, 2042–2054.
51. Lin, I.-H., Chen, D.-T., Chang, Y.-F., Lee, Y.-L., Su, C.-H., Cheng, C., Tsai, Y.-C., Ng, S.-C., Chen, H.-T., Lee, M.-C., et al. (2015). Hierarchical clustering of breast cancer methylomes revealed differentially methylated and expressed breast cancer genes. PLoS One *10*, e0118453.
52. Hon, G.C., Hawkins, R.D., Caballero, O.L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L.E., et al. (2012). Global DNA

hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. Genome Res. *22*, 246–258.

53. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.
54. Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.-Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. Nature *500*, 477–481.
55. Lowe, R., Gemma, C., Beyan, H., Hawa, M.I., Bazeos, A., Leslie, R.D., Montpetit, A., Rakyan, V.K., and Ramagopalan, S.V. (2013). Buccals are likely to be a more informative surrogate tissue than blood for epigenome-wide association studies. Epigenetics *8*, 445–454.
56. Guo, H., Zhu, P., Yan, L., Li, R., Hu, B., Lian, Y., Yan, J., Ren, X., Lin, S., Li, J., et al. (2014). The DNA methylation landscape of human early embryos. Nature *511*, 606–610.
57. Cai, W., Mao, F., Teng, H., Cai, T., Zhao, F., Wu, J., and Sun, Z.S. (2015). MBRidge: an accurate and cost-effective method for profiling DNA methylome at single-base resolution. J. Mol. Cell Biol. *7*, 299–313.
58. McDonald, O.G., Li, X., Saunders, T., Tryggvadottir, R., Mentch, S.J., Warmoes, M.O., Word, A.E., Carrer, A., Salz, T.H., Natsume, S., et al. (2017). Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis. Nat. Genet. *49*, 367–376.
59. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol. *28*, 1045–1048.
60. Wang, Y., Qin, J., Wang, S., Zhang, W., Duan, J., Zhang, J., Wang, X., Yan, F., Chang, M., Liu, X., et al. (2016). Conversion of Human Gastric Epithelial Cells to Multipotent Endodermal Progenitors using Defined Small Molecules. Cell Stem Cell *19*, 449–461.
61. Schroeder, D.I., Blair, J.D., Lott, P., Yu, H.O.K., Hong, D., Crary, F., Ashwood, P., Walker, C., Korf, I., Robinson, W.P., and LaSalle, J.M. (2013). The human placenta methylome. Proc. Natl. Acad. Sci. USA *110*, 6037–6042.
62. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., and Ishiguro-Watanabe, M. (2023). KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. *51*, D587–D592.
63. Xi, Y., and Li, W. (2009). BSMAP: Whole genome bisulfite sequence MAPping program. BMC Bioinf. *10*, 232–239.
64. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.
65. Wang, Q., Li, M., Wu, T., Zhan, L., Li, L., Chen, M., Xie, W., Xie, Z., Hu, E., Xu, S., and Yu, G. (2022). Exploring Epigenomic Datasets by ChIPseeker. Curr. Protoc. *2*. e585–27.
66. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.
67. Badia-I-Mompel, P., Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., Taus, P., Dugourd, A., Holland, C.H., Ramirez Flores, R.O., et al. (2022). decoupleR: ensemble of computational methods to infer biological activities from omics data. Bioinformatics Advances *2*, 1–3.
68. Fang, H., and Gough, J. (2014). The "dnet" approach promotes emerging research on cancer patient survival. Genome Med. *6*, 64.
69. Jühling, F., Kretzmer, H., Bernhart, S.H., Otto, C., Stadler, P.F., and Hoffmann, S. (2016). metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome Res. *26*, 256–262.
70. 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.
71. Fang, H., Knezevic, B., Burnham, K.L., and Knight, J.C. (2016). XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. Genome Med. *8*, 129.
72. Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. *45*, W98–W102.
73. Court, F., Tayama, C., Romanelli, V., Martin-Trujillo, A., Iglesias-Platas, I., Okamura, K., Sugahara, N., Simón, C., Moore, H., Harness, J.V., et al. (2014). Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. Genome Res. *24*, 554–569.
74. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The Human Genome Browser at UCSC. Genome Res. *12*, 996–1006.
75. Zhang, Z., Wang, S., Jiang, L., Wei, J., Lu, C., Li, S., Diao, Y., Fang, Z., He, S., Tan, T., et al. (2024). Priority index for critical Covid-19 identifies clinically actionable targets and drugs. Commun. Biol. *7*, 189.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
110 fresh-frozen primary solid tumor samples	Hangzhou First People's Hospital	Listed in Table S1
Deposited data		
Processed WGBS data of 110 human primary solid tumors	This study	GEO: GSE212391
TCGA Methylation, RNA-seq and ATAC-seq data	TCGA et al.[29,40]	https://gdac.broadinstitute.org; https://gdc.cancer.gov/about-data/publications/ATACseq-AWG
Pan-cancer RRBS data	Zhang et al.[38]	GEO: GSE230193
Public WGBS data of human normal tissues	Guo et al.,[26] Cao et al.,[34] Li et al.,[35] Jenkinson et al.,[36] Hevn et al.,[37] Lin et al.,[51] Hon et al.,[52] ENCODE,[41] Lister et al.,[53] Ziller et al.,[54] Court et al.,[55] Guo et al.,[56] Cai et al.,[57] McDonald et al.,[58] Bernstein et al.,[59] Wang et al.,[60] Schroeder et al.[61]	Listed in Table S4, and the mHap files are available at https://doi.org/10.5281/zenodo.16474855
Paired WGBS and RNA-seq of ESCC	Cao et al.[34]	GEO: GSE149612
Colorectal cancer RRBS data	Guo et al.,[26] Lin et al.[32]	GEO: GSE79211, GSE56269
Colorectal cancer single-cell multi-omics data	Bian et al.[18]	GEO: GSE97693
Liver cancer WGBS data	Li et al.,[35] Jenkinson et al.,[36] Hevn et al.,[37]	GEO: GSE70090, GSE79799 SRA: SRX381661, SRX381666
DNA methylation of cfDNA in plasma	Kandimalla et al.[48]	GEO: GSE149438
Pol II ChIA-PET (MCF-7, MCF-10A)	ENCODE[41]	ENCODE: ENCFF597SQA, ENCFF252XDG
KLF5 ChIP-seq	Li et al.[44]	GEO: GSM4274815
CollecTRI database	Müller-Dott et al.[42]	OmniPath (https://omnipathdb.org/) and DoRothEA (https://saezlab.github.io/dorothea)
CpG island	UCSC	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/cpgIslandExt.txt.gz
FANTOM5 enhancer	FANTOM5	https://enhancer.binf.ku.dk/presets/permissive_enhancers.bed
TAD (A549, HMEC, IMR90)	ENCODE	ENCODE: ENCFF336WPU, ENCFF351HVI, ENCFF307RGV
Lamin B1 domain (LAD)	UCSC	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/laminB1Lads.txt.gz
Ultra-conserved elements (UCEs)	UCSC	http://www.cse.ucsc.edu/~jill/ultra.html
KEGG data	Kanehisa et al.[62]	https://www.genome.jp/kegg/
Software and algorithms		
Original code	This study	https://doi.org/10.5281/zenodo.16496342
mHapTools v0.10	Zhang et al.[16]	https://github.com/butyuhao/mHapTools
mHapTK (mHapSuite) v2.0	Ding et al.[32]	https://github.com/yoyoong/mHapSuite

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
mHapSC v1.0	This study	https://github.com/yoyoong/mHapSc
BSMAP v2.90	Xi et al.[63]	http://code.google.com/p/bsmap/
bedtools v2.25.0	Quinlan et al.[64]	http://code.google.com/p/bedtools
ChIPseeker v1.28.3	Wang et al.[65]	https://www.bioconductor.org/packages/release/bioc/html/ChIPseeker.html
clusterProfiler V4.0.5	Yu et al.[66]	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
decoupleR v2.9.11	Badia-I-Mompel et al.[67]	https://saezlab.github.io/decoupleR/
dnet v1.1.7	Fang et al.[68]	https://github.com/hfang-bristol/dnet
FastQC v0.11.8	N/A	https://github.com/s-andrews/FastQC
LOLA v1.22.0	Sheffield et al.[39]	http://code.databio.org/LOLA/
MethylDackel v0.5.0	N/A	https://github.com/dpryan79/MethylDackel
MethylSeekR v1.32.0	Burger et al.[10]	https://bioconductor.org/packages/release/bioc/html/MethylSeekR.html
Metilene v0.2-8	Jühling et al.[69]	http://legacy.bioinf.uni-leipzig.de/Software/metilene/
pracma v2.4.4	N/A	https://cran.r-project.org/web/packages/pracma/index.html
rGREAT v2.4.0	Gu et al.[43]	https://jokergoo.github.io/rGREAT/
Sambamba v0.7.1	Tarasov et al.[70]	https://lomereiter.github.io/sambamba/
SRA Toolkit v2.9.1	NCBI	https://github.com/ncbi/sra-tools
Trim galore v0.6.2	N/A	https://github.com/FelixKrueger/TrimGalore
TxDb.Hsapiens.UCSC.hg19.knownGene v3.2.2	N/A	https://bioconductor.org/packages/TxDb.Hsapiens.UCSC.hg19.knownGene/
XGR v1.1.9	Fang et al.[71]	https://github.com/hfang-bristol/XGR
liftOver	Kent et al.[61]	https://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/liftOver
GEPIA	Tang et al.[72]	http://gepia.cancer-pku.cn

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study included tumor samples from 110 patients across 11 different cancer types, with approximately 10 samples per cancer type. Sample collection was approved by the Ethics Committee of Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China. Participants ranged in age from 34 to 87 years and were stratified by cancer type for analysis. Comprehensive demographic and clinical characteristics, including age, sex, cancer types, and clinical stages, are detailed in Table S1. Due to the limited sample size per cancer type (approximately 10 samples each) and the pooled analysis approach employed in this study, we were unable to assess the influence of sex or gender on the results. This represents a limitation of our study design, and future studies with larger cohorts would be needed to evaluate potential sex-specific effects.

METHOD DETAILS

WGBS library construction and sequencing

Frozen tumor tissues were sliced to 3 mm^3, and genomic DNA was isolated using the DNeasy Blood & Tissue Kit (QIAGEN, Cat. No. 69504; Hilden, Germany) per the manufacturer's instructions. Then, 200 ng of purified genomic DNA were sheared to 100–500 bp using a Covaris S2 (Covaris; Woburn, MA, USA) sonicator. Next, we treated the fragmented DNA using the EpiTect Fast Bisulfite Conversion Kit (QIAGEN, Cat. No. 59824; Hilden, Germany) with a modified protocol–two cycles of 98°C for 5 min and 60°C for 25 min, followed by a hold at 25°C. The bisulfite-converted DNA was subjected to WGBS construction using the Accel-NGS Methyl-seq DNA Library Kit (Swift Biosciences, Cat. No. 51002025P; Ann Arbor, MI, USA) as recommended by the manufacturer. WGBS libraries were sequenced for 2 × 150 cycles in an Illumina NovaSeq 6000 sequencer.

External data collection

We curated a panel of public WGBS samples from normal tissues: breast (n = 5),[37,41,51,52] colon (n = 7),[26,37,41,53,54] head and neck (n = 28),[55] esophagus (n = 15),[34,41,53] liver (n = 15),[26,35–37,41,56,73] lung (n = 12),[26,35–37,41,53] ovary (n = 8),[41,53,57] ncreas (n = 6),[26,41,53,58]

stomach (n = 11),[26,41,53,59,60] thyroid gland (n = 2),[41] and placenta (n = 7).[37,59,61,73] Additionally, WGBS samples from liver cancer (n = 9)[35–37] and ESCC (n = 10),[34] as well as RRBS samples from CRC (n = 19)[26,33] and previous pan-cancer study (GSE230193, n = 498),[38] were included. Detailed sample annotations are described in Table S4.

We explored the potential application of MHBs in a targeted BS-seq dataset[48] of cell-free DNA from gastrointestinal cancers and normal individuals, including CRC (n = 40), ESCC (n = 48), EAC (n = 12), GC (n = 37), HCC (n = 43), PDAC (n = 74), and healthy donors (n = 46). Detailed annotation of these samples was described in Table S15.

For Pol II ChIA-PET data,[41] the processed results from MCF-7 (ENCFF597SQA) and MCF-10A (ENCFF252XDG) cell lines were retrieved from the ENCODE project. Regions of open chromatin (ATAC-seq peaks) across 11 cancer types from TCGA were downloaded from the NCI Genomic Data Commons (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG).[40] RNA-seq and DNA methylation HM450K array data for 11 cancer types from TCGA were obtained from the Genome Data Analysis Center at the Broad Institute (https://gdac.broadinstitute.org).[29]

The normalized gene expression profiles of ESCC dataset were obtained from GSE149612. The single-cell DNA Methylation data of colorectal cancer was obtained from GSE97693. When applicable, genomic coordinates of downloaded regions were converted to hg19 using the liftover tool.[74]

QUANTIFICATION AND STATISTICAL ANALYSIS

Preprocessing of BS-seq data

Raw SRA files were downloaded from NCBI GEO and converted to fastq files using the SRA Toolkit (version 2.9.1). Quality control was then performed with FastQC (version 0.11.8). Trim galore (version 0.6.2) was used to remove sequence adapters and low-quality bases, operating in paired-end or single-end mode with default settings. The trimmed sequences were subsequently aligned to the human genome version hg19 using BSMAP[63] with the parameters "-q 20 -f 5 -r 0 -v 0.05 -s 16 -S 1". For WGBS, duplicate reads were identified and marked using Sambamba (version 0.7.1).[70] Mean CpG methylation levels were extracted with MethylDackel (https://github.com/dpryan79/MethylDackel) (version 0.5.0).

Identification of DNA methylation-associated regions

Differentially methylated regions (DMRs) for WGBS were identified using Metilene (v.0.2–8)[69] with parameters "-t 10 -c 2 -m 5". Only DMRs with q value <0.05 and covered by at least five CpG sites were retained for downstream analysis. The genome of each cancer type was segmented into UMR, LMR, and PMDs using the MethylSeekR tool (version 1.32.0) with default parameters.[10] The rest of genome regions, excluding the gaps annotated by UCSC (https://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/-gap.txt.gz), were defined as HMRs. Biological replicates of the same cancer type or tissue type were pooled to increase coverage.

For methylation array data, the differentially methylated CGIs were defined as DMRs. Briefly, differentially methylated probes were identified by Wilcoxon rank-sum test with FDR <0.05 and Δbeta >0.1. Next, we retained the probes located in CGIs. Subsequently, differentially methylated CGIs were estimated by identifying CGIs harboring differentially methylated probes. A CGI was classified as a hypomethylated DMR if at least two-thirds of its probes were hypomethylated. The identification of hypermethylated DMRs followed the same criterion.

Differential gene expression analysis

DEGs between tumor and normal samples in both bulk and single-cell data were calculated using Wilcoxon rank-sum test with the thresholds of fold change >2 and FDR <0.05.

Identification of MHBs from bulk and single-cell BS-seq data

To identify MHBs from bulk and single-cell BS-seq data, we applied a seed-and-extend algorithm. The process starts by selecting a seeding window of five consecutive CpGs, requiring significant linkage disequilibrium correlation (LD $R^2 > 0.5$, binomial test $p < 0.05$) between all pairwise CpGs within the window. This window is iteratively extended by adding adjacent CpGs that maintain LD correlation (LD $R^2 > 0.5$, binomial test $p < 0.05$) with every CpG in the existing window. Extension continues until no additional CpGs meet the threshold, defining the contiguous region as an MHB.

For bulk BS-seq, BAM files were converted to mHap files with mHapTools (version 1.1).[16] MHBs were identified with mHapSuite (https://github.com/yoyoong/mHapSuite), a java implementation of mHapTk,[32] with default parameters "-window 5 -r2 0.5 -pvalue 0.05". Genome-wide methylation tracks were created with mHapSuite using default parameters "–minK 1 –maxK 10 –K 4".

For single-cell BS-seq, mean methylation values from forward and reverse strands were converted to binary methylation haplotypes by thresholding: values < 0.1 as 0 (unmethylated), >0.9 as 1 (methylated), excluding intermediate values. Single-cell MHBs were identified using mHapSc (https://github.com/yoyoong/mHapSc) with default parameters "-window 5 -r2 0.5 -pvalue 0.05". The genomic annotations of cancer MHBs were performed using ChIPseeker.[65]

Genomic region enrichment analysis

The R package LOLA (v1.22.0)[39] was used to test the enrichment of a query region set in a database of reference region sets, with FDR <0.05 considered statistically significant. To evaluate tissue or cancer specificity, the normal MHBs, methylation-associated

regions (PMD, LMR, UMR, HMR), and DMRs were compared with cancer MHBs by LOLA enrichment. The FDR were ranked, and the top *k* tissues with the highest significance were highlighted, where *k* was dependent on the database size. In case of tied ranks, more than *k* tissues may be highlighted. The specific *k* used in each dataset is indicated in the figure legends.

For ChIA-PET tag enrichment analysis within cancer MHBs, we employed a permutation-based approach to assess the enrichment of a set of genomic features (reference set) in a set of MHBs (query set). The "shuffle" function in bedtools (v2.25.0)[64] was used to generate 1000 random sets that match the size of the query set. The expected overlap was calculated as the average number of overlaps between the reference set and random sets. Enrichment score was defined as the ratio of observed and expected numbers of overlapping regions. The permutation *p*-value was determined as the fraction of random sets that showed more overlaps with the reference set than the query set.

Transcription factor activity inference

Transcription factor activities in ESCC were calculated using the decoupleR R package.[67] First, a high-confidence network comprising 1,186 TFs and their target genes was downloaded from the CollecTRI database[42] as prior knowledge. Next, the rGREAT R package[43] with default parameters was employed to assess the enrichment of TF regulons in ESCC MHB regions, retaining only the significant ones (FDR <0.05). Finally, a univariate linear model method was applied to infer TF enrichment scores using RNA-seq data from ESCC and paired adjacent normal esophageal samples.

Priority index (Pi) prioritization for pan-cancer MHB target identification

Following our well-established Pi prioritization approach,[45,46] we generated a ranked list of 8,852 pan-cancer MHB-associated genes (priority ratings: 0–5) and identified a 53-gene pathway crosstalk. Two inputs were considered: (i) TCGA expression profiles of pan-cancer MHB-associated genes identified in this study; and (ii) KEGG pathway-derived gene interactions. The prioritization comprised the following key components.

(1) *Core genes under expression evidence of pan-cancer MHB-regulated genes.* DEGs potentially regulated by cancer MHBs (excluding DMR-associated genes) were defined as core genes in each tumor type through four steps. At Step 1, DEGs were identified between tumors and normal controls using the Wilcoxon rank-sum test. At Step 2, the significance of genes was quantified simultaneously considering the combined information of $\log_2$(fold change) and FDR. At Step 3, MHB-related genes were identified using rGREAT[43] with default parameters, while DMR-associated genes identified by promoter methylation changes (1,000 bp around the TSS). At Step 4, DEGs potentially regulated by MHBs but excluding those DMR-associated genes were retained as core genes. These core genes were used as inputs for identifying peripheral genes (see the next).

(2) *Peripheral genes under network evidence of pathway interactions.* Using the random walk with restart (RWR) algorithm,[75] peripheral genes were prioritized based on connectivity to core genes in a KEGG-derived network (~6,300 nodes/genes and ~55,000 interactions/edges).[62] Seed nodes represented core genes; non-seed peripheral genes received affinity scores reflecting network proximity to seeds.

(3) *Gene-predictor matrix for target gene prioritization at the gene and pathway crosstalk levels.* For pan-cancer datasets (e.g., BRCA and ESCA), the process described above yielded a predictor containing both MHB-associated core and peripheral genes, along with affinity scores quantifying network connectivity to seed core genes. The resulting gene-predictor matrix contained rows for genes and columns for predictors. Within this matrix, Fisher's combined meta-analysis method was used to combine predictors, incorporating genetic evidence and network evidence. It involved affinity score conversion, that is, converting into *P*-like values using empirical cumulative density function (eCDF) that were subsequently combined across predictors via Fisher's combined method to produce priority rating (0–5). The identification of pan-cancer MHB-regulated and highly prioritized target genes at the pathway crosstalk involved a heuristic solution, with the statistical significance (*P* value) assessed using a degree-preserving node permutation test (see our previous publications).[68,71]

Pathway enrichment analysis

To investigate biological functions and pathways associated with MHB-linked genes, we performed enrichment analysis using the clusterProfiler[66] R package (v4.0.5). Gene sets were obtained from MSigDB: curated chemical and genetic perturbation sets (C2) identified pathways linked to optimal subnetwork genes from pathway crosstalk analysis, while Hallmark gene sets (H) representing 50 core biological processes assessed differences between high- and low-AUC tumors. Gene symbols were first converted to Entrez IDs using the "bitr" function from the org.Hs.e.g.,.db annotation package. All genes measured in the dataset served as the background. Results were visualized using dot plots and bar plots generated by the dotplot and barplot functions, with significant terms/pathways filtered at FDR <0.05.

Survival analysis

Survival analysis was performed using the online tool GEPIA (http://gepia.cancer-pku.cn/)[72] with TCGA RNA-seq data. Participants were stratified into two groups based on median expression across all tumor samples. Kaplan-Meier survival curves compared disease-free survival and overall survival between patient subgroups stratified by high/low gene expression (median cutoff). *p* values were calculated using log rank tests, and hazard ratios were determined through univariable Cox regression analysis.

Quantifying the level of concordant methylation

To quantify co-methylation patterns, we calculated the AUC derived from LOESS curves fitted between mean methylation and methylation variation within MHB regions using HM450K array data. First, MHB regions were mapped to HM450K probes, excluding regions with fewer than three probes. Mean methylation and variation (variance across probes) were computed for each MHB region. A LOESS curve (span = 0.3) was fitted to model the relationship between mean methylation and variation per sample, with AUC calculated using the "trapz" function from the pracma R packages. Samples were assigned an AUC score reflecting methylation concordance: lower AUC indicates higher methylation consistency (lower heterogeneity), while higher AUC reflects greater disorder (higher heterogeneity). Samples were categorized based on AUC scores for downstream comparisons. To assess mutation effects, samples were stratified by tumor driver mutation status, enabling analysis of mutation-associated methylation heterogeneity within MHB regions.

Machine learning-based cancer detection

We constructed a random forest model using the python "scikit-learn" package, incorporating CpG-level mean methylation, MHL,[26] and MBS[47] for 20,946 cancer MHBs overlapping 67,832 DMRs in the GSE149438 dataset across six cancer types. The model was configured with 1,000 decision trees, a maximum tree depth of 5, and randomized variable selection at each split. Performance was evaluated using receiver operating characteristic (ROC) curves, with 5-fold cross-validation and iterative parameter tuning to optimize ROC values. For benchmarking, we compared results against established methods such as CancerDetector[49] across the same six cancer types in MHB regions, using ROC curves for performance evaluation.

Cell Reports, Volume *44*

Supplemental information

Toward the DNA methylation haplotype map of 11 common solid cancers

Zhiqiang Zhang, Yuyang Hong, Shirong Zhang, Xin Zhu, Leiqin Liu, Xiqi Liao, Hongcang Gu, Hai Fang, and Jiantao Shi

Supplemental Figures

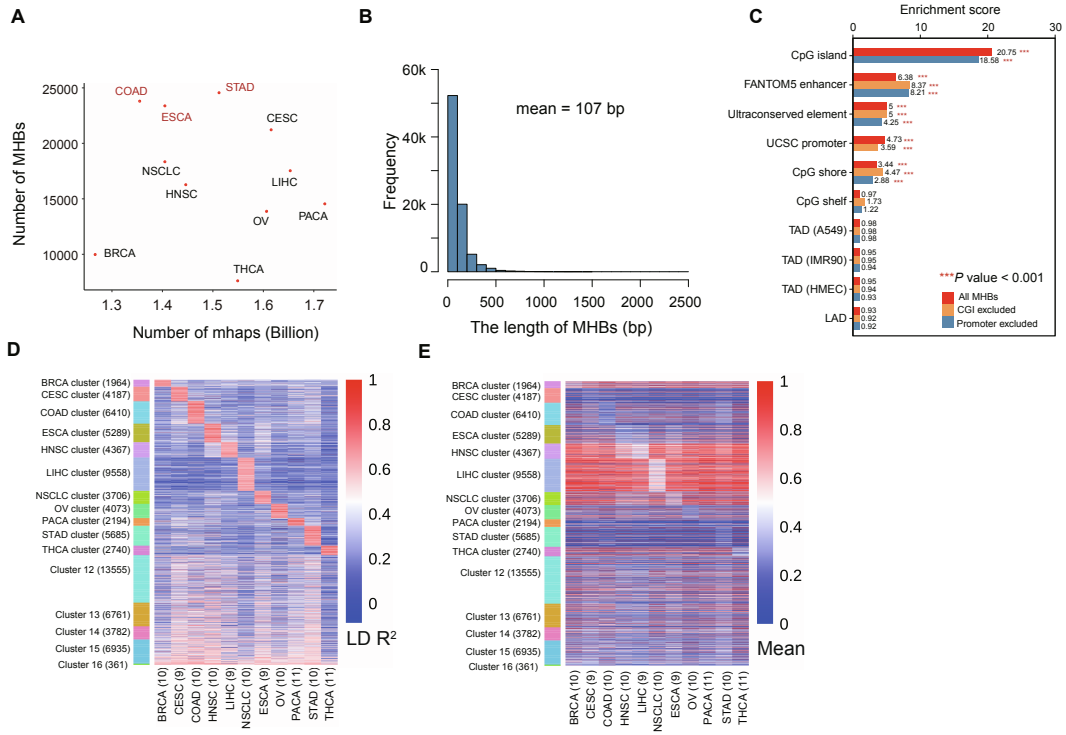


Figure S1. The MHB characteristics in 11 common solid cancers. (**A**) A scatter plot shows the relationship between number of MHBs and number of mHaps. (**B**) The length distribution of the union of MHBs from all cancer types. (**C**) Enrichment of MHBs in regions of genomic features. Enrichment scores were defined as the ratio of observed and random overlapping counts between MHBs and genomic regions. Significance was assessed by permutation test (1000 times). Enrichment analysis was also performed when confounding regions such as promoters and CGIs were removed. ***, $P < 0.001$. Regions of LADs, UCSC promoters, and CGIs were downloaded from UCSC Genome Browser (https://genome.ucsc.edu). CGI shores (< 2 kb flanking CGIs) and CGI shelves (<2 kb flanking outwards from the CGI shores) were defined based on genomic positions of CGIs. FANTOM5 enhancers were downloaded from FAMTOM5 data server: https://fantom.gsc.riken.jp/5/data/. TADs in IMR90 and HMEC cell lines were downloaded from ENCODE project (https://www.encodeproject.org). 481 UCEs (UltraConserved Elements) was obtained from http://www.cse.ucsc.edu/~jill/ultra.html. (**D-E**) A heatmap of signed LD R^2 and mean methylation of 16 MHB clusters in 11 cancer types.

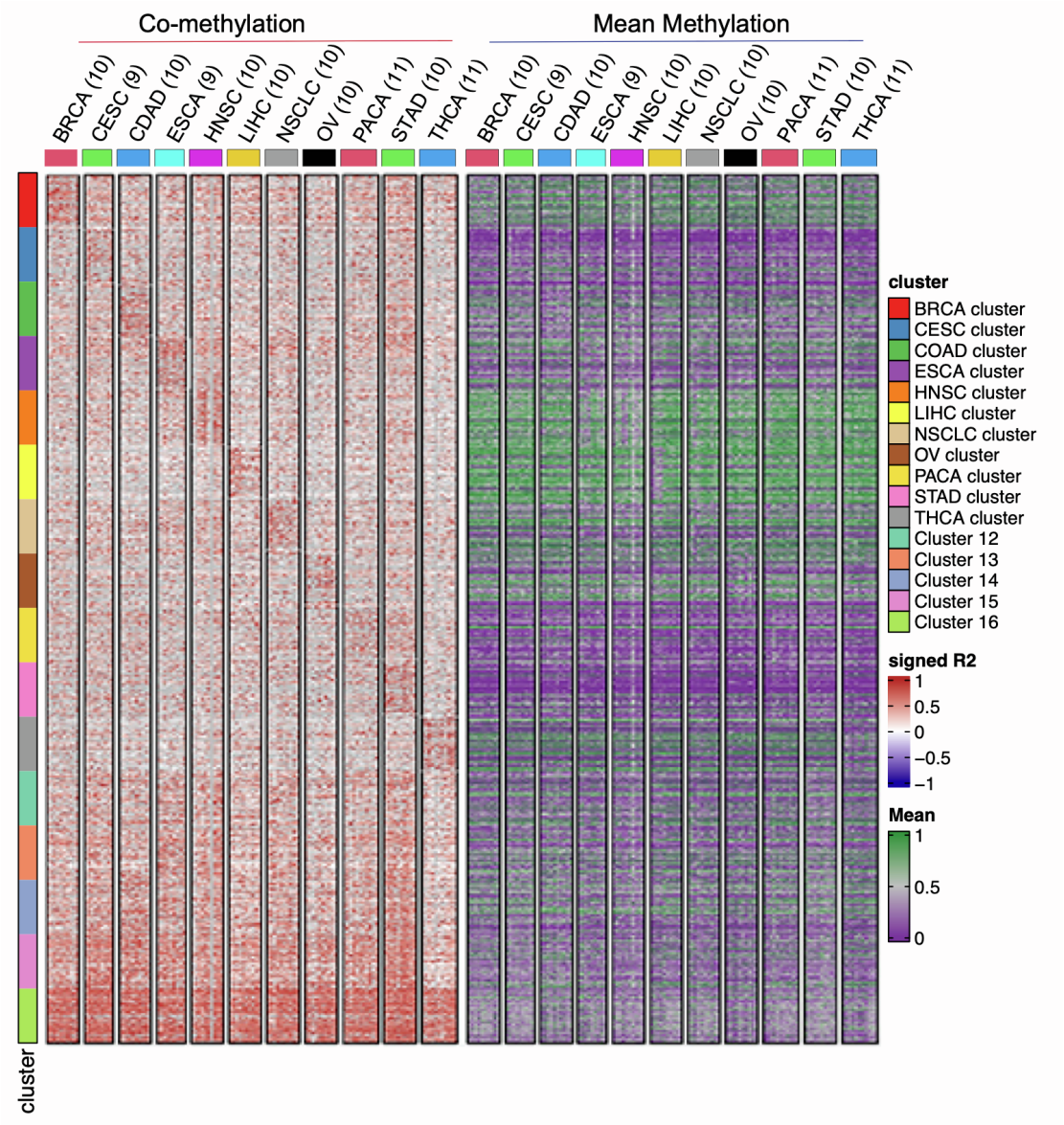


Figure S2. Heatmap of signed LD R^2 and mean methylation levels across 16 MHB clusters in 110 solid tumor samples. For visualization purposes, 100 regions were randomly selected from each of the 16 clusters.

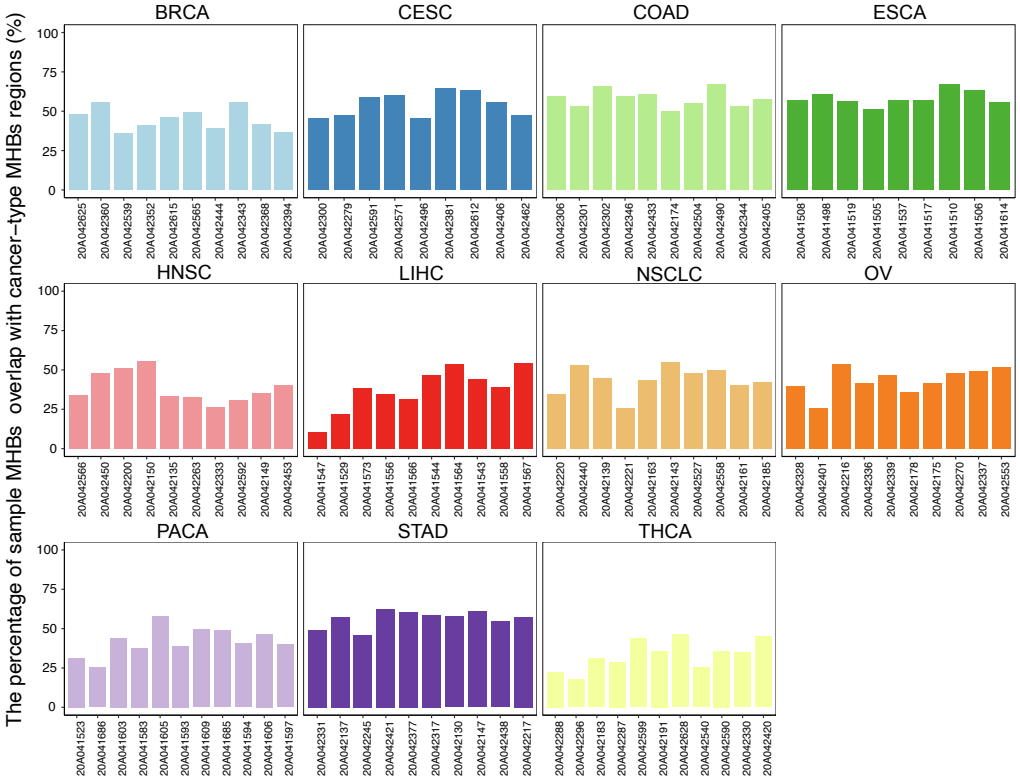


Figure S3. Comparison of individual tumor MHBs *versus* cancer type-specific MHBs. MHBs were identified from each individual tumor sample using the same parameters applied to the pooled samples. For each cancer type, the percentage of individual tumor MHBs that overlap with the corresponding cancer type-specific MHBs is shown.

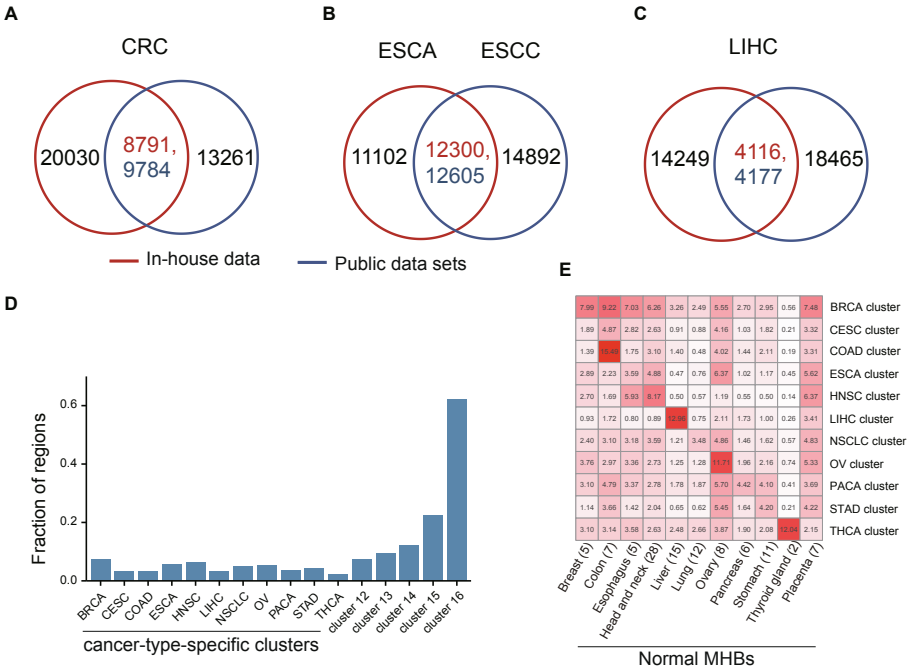


Figure S4. Comparison of MHBs identified in this study with those from public datasets. (**A-C**) Venn diagrams show number of MHBs shared between those identified from this study (in red) and external datasets (in blue). The comparison was performed for three cancer types, including CRC (**A**), ESCA (**B**), LIHC (**C**). (**D**) A bar plot shows fraction of regions in 16 MHB clusters covered by placenta MHBs. (**E**) A heatmap shows the percentages of cancer type-specific MHBs that are also present in normal tissues.

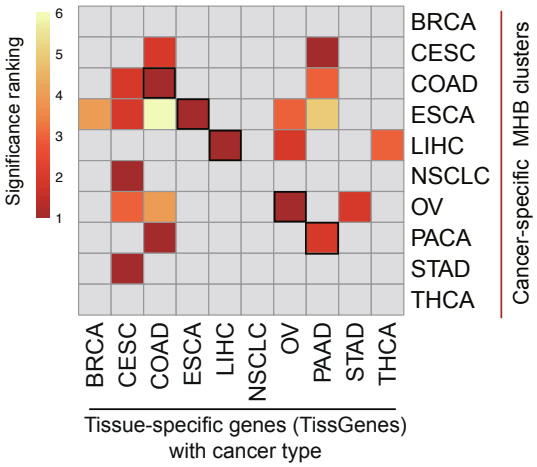


Figure S5. Enrichment of cancer type-specific MHBs in cancer-type-specific TissGenes. Tissue-specific genes in cancer were downloaded from TissGDB (https://bioinfo.uth.edu/TissGDB/)**.** The significance rank was defined by the values of FDR, which are calculated by R package rGREAT. Non-significant results were shown in gray (*P*-value > 0.05).

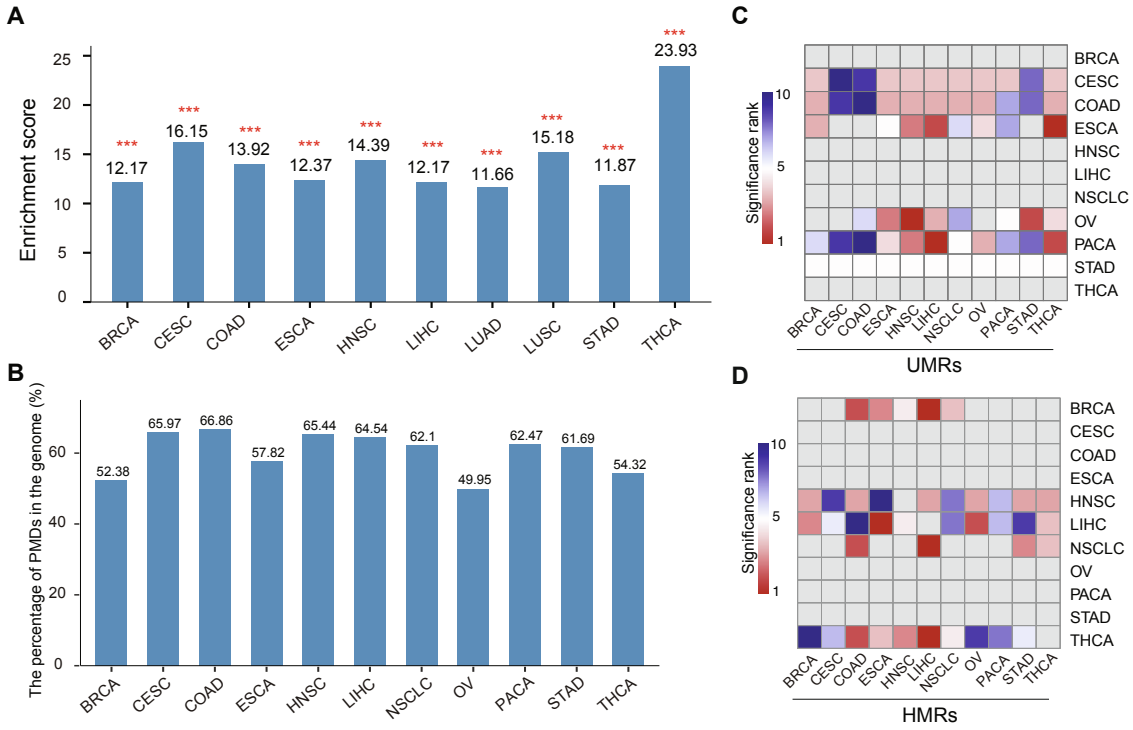


Figure S6. Cancer MHB enrichment in regulatory regions. (**A**) Enrichment of MHBs in regions of open chromatin. ***, p < 0.001. **, p < 0.01. *, p < 0.05. (**B**). The percentage of PMD in 11 cancer types. (**C-D**). The enrichment of MHBs in UMR and HMR, respectively. The significance rank was defined by the values of FDR, which are calculated by R package LOLA, all cancer types MHBs as background. The grey color shows the results with FDR > 0.01.

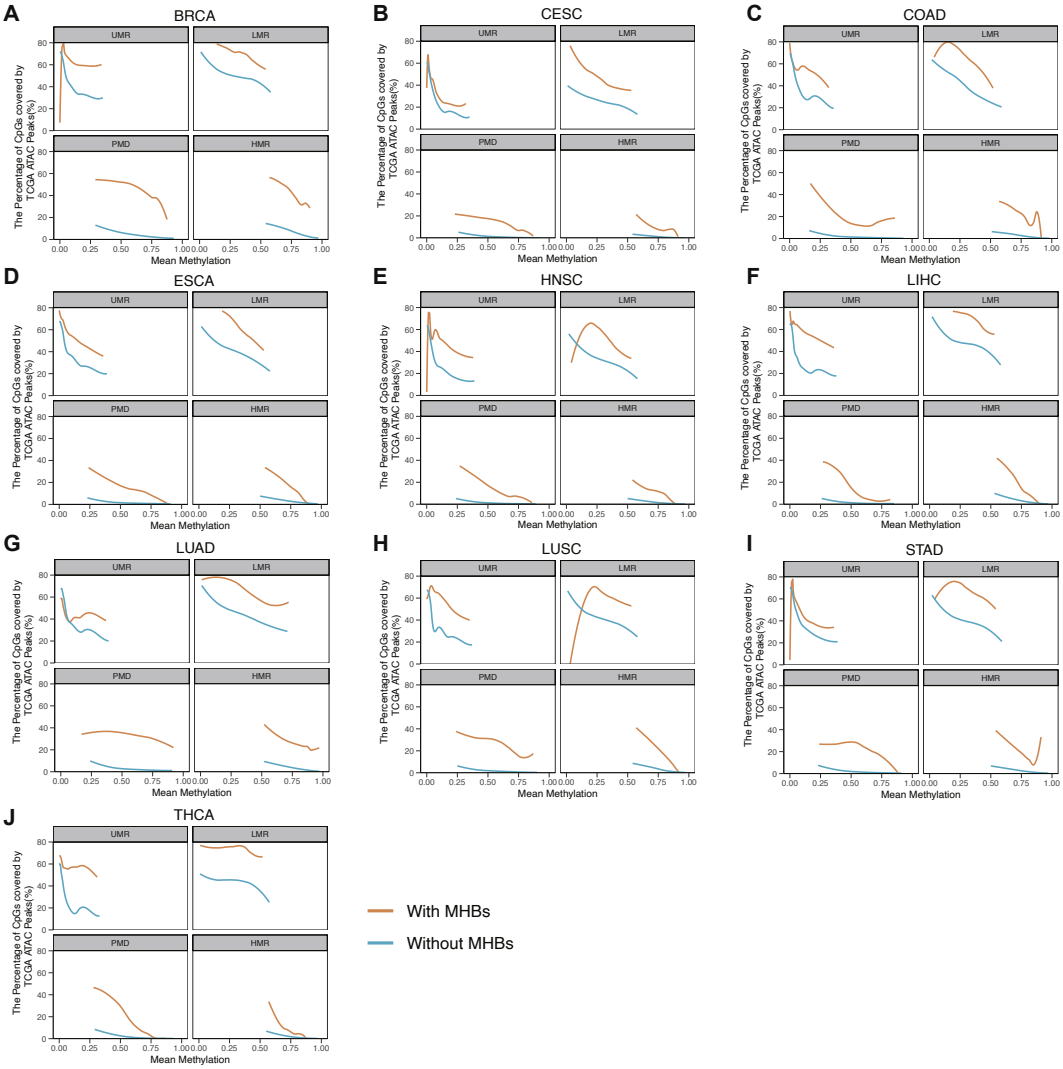


Figure S7. Enrichment of MHBs in regulatory regions is dependent of mean methylation. (**A-J**) Genomic regions with MHBs are more enriched in regions of open chromatin regardless of mean methylation levels. The enrichment score was calculated as the percentage of CpGs covered by regions of open chromatin. The open chromatin data (ATAC-seq peaks) from TCGA project.

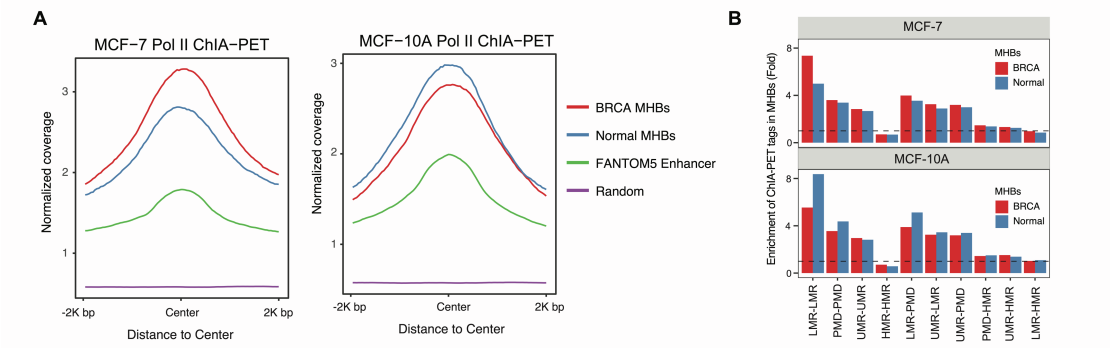


Figure S8. Genomic regions enrichment in MCF-7 or MCF-10A Pol II ChIA-PET. (**A**) The normalized coverage of breast cancer MHBs, normal breast MHBs, FANTOM5 enhancer and random regions in MCF-7 (ENCFF597SQA) and MCF-10A (ENCFF252XDG) Pol II ChIA-PET data, respectively. (**B**) Enrichment of ChIA-PET tags in MHBs. We divided the Pol II ChIA-PET tags into 10 groups, LMR, UMR, PMD, HMR and a combination of any two of them. Permutation test (100 times) was performed to evaluate the fold enrichment of MHBs from BRCA and normal tissues in MCF-7 and MCF-10A ChIA-PET tags.

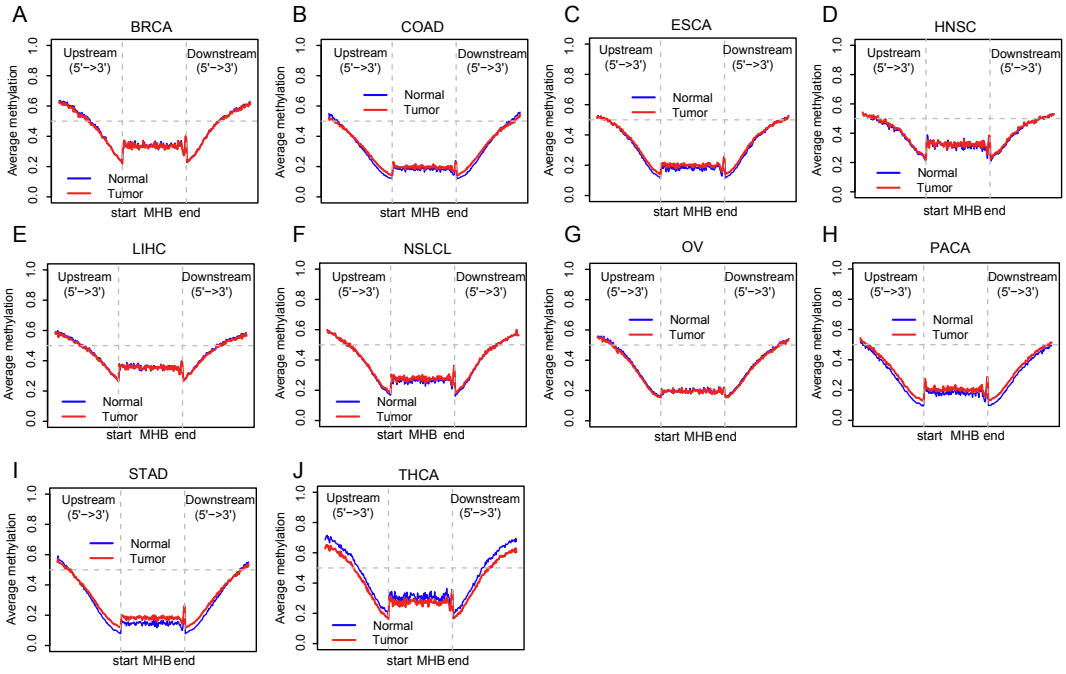


Figure S9. Mean methylation profiles of MHBs not overlapping DMRs. (A-J) In each cancer type, MHBs that don't overlap with DMRs were plotted. Mean methylation around center of MHBs (+/- 1000 bp) was shown.

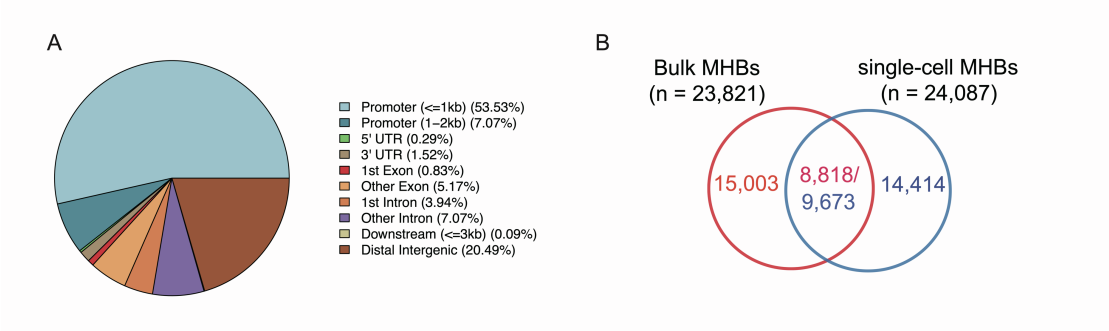


Figure S10. Validation of cancer MHBs and their regulatory roles using single-cell CRC data. (**A**). A Pie chart illustrates the proportion of single-cell MHBs of malignant tumor cells annotated to promoter, exonic, intronic, and intergenic regions. (**B**). Veen diagram of MHBs identified from single-cells data and bulk data.

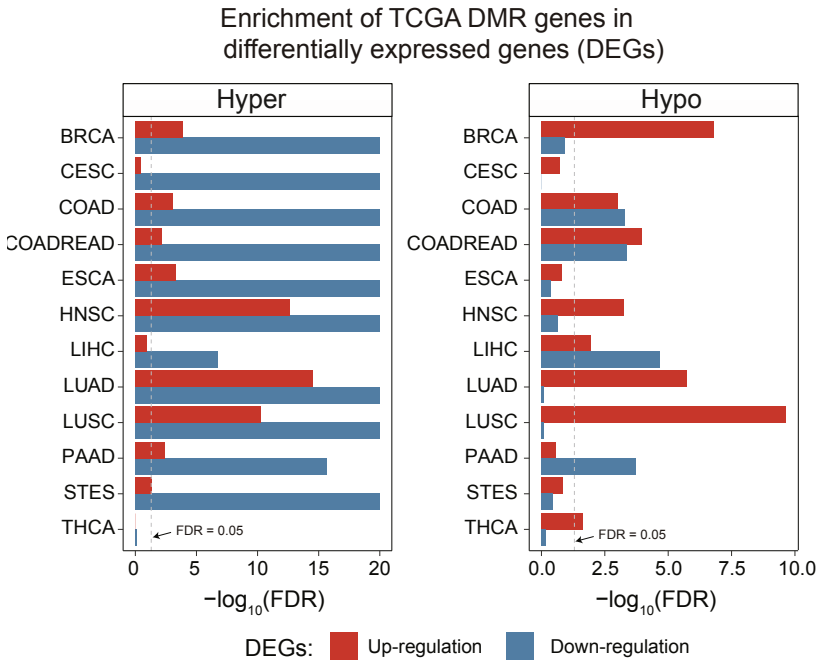


Figure S11. Association between DMRs and differentially expressed genes in pan-cancer. Enrichment of DMR-associated genes in differentially expressed genes. Statistical significance was evaluated using Fisher's exact test. The resulting *P*-values were adjusted for multiple testing and reported as FDR. Upregulated genes and downregulated genes are labelled in red and blue, respectively.

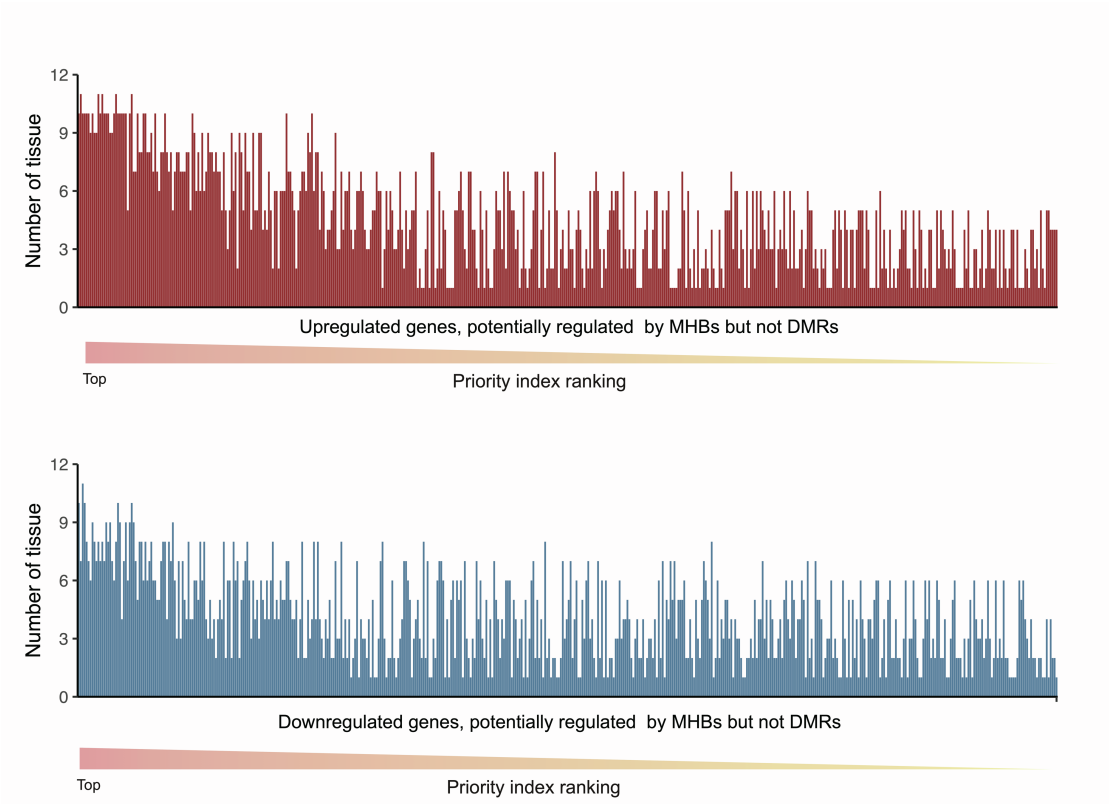


Figure S12. Distribution of MHB-related genes in pan-cancer ranked by Priority index. The bar plot shows the distribution of genes, which are potentially regulated by MHBs but not by DMRs, among upregulated and downregulated DEGs in pan-cancer. All of the genes ranked by priority index rating.

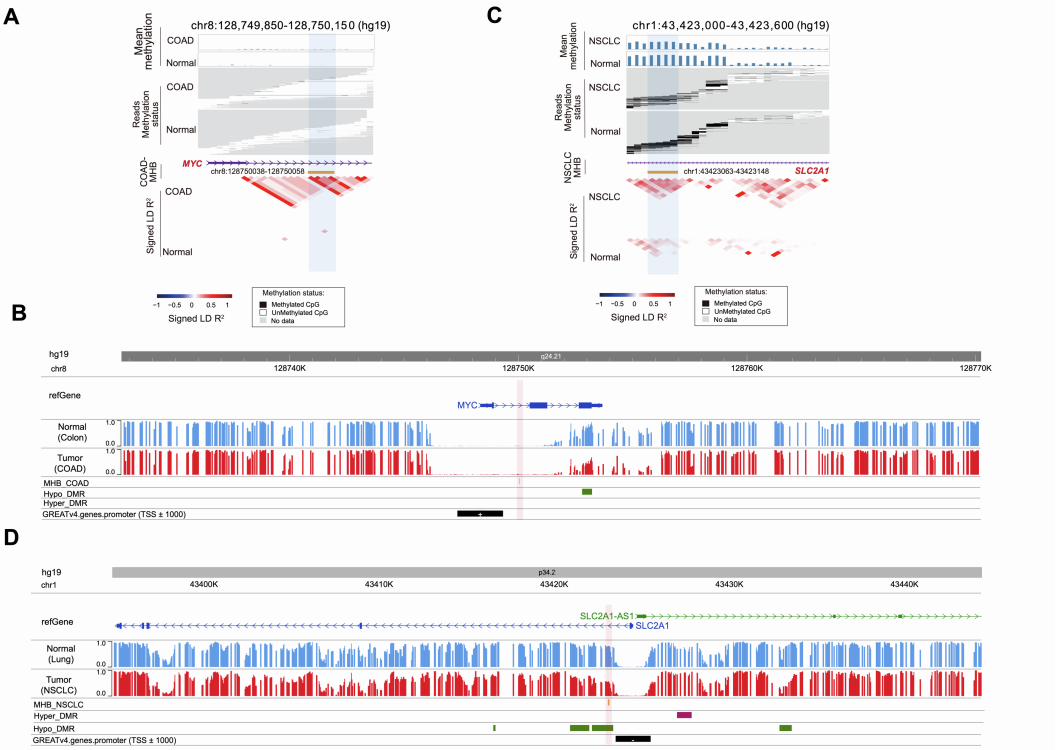


Figure S13. Methylation patterns of MHBs in *MYC* and *SLC2A1* gene loci in COAD and NSCLC. (**A**) MHB in the *MYC* gene locus (chr8: 128,749,850 - 128,750,150). The upper panel shows the mean methylation levels; the middle panel illustrates the DNA methylation status of individual fragments, with black and white representing methylated and unmethylated CpG sites, respectively; and the bottom panel displays the genomic position of the MHB relative to adjacent genes, alongside a heatmap of signed linkage disequilibrium (LD) R^2 values in the region. (**B**) IGV screenshot for *MYC* showing tracks of mean methylation in cancer and normal tissues, MHB locations, differentially methylated regions (DMRs), and promoter regions (defined as 1000 bp upstream to 1000 bp downstream of the TSS). (**C**) MHB in the *SLC2A1* gene locus (chr1: 43,423,000 - 43,423,600). Panel organization is the same as in (**A**). (**D**) Integrative IGV screenshot for *SLC2A1*.

Figure S14. Pathway enrichment of subnetwork genes. The curated gene set collection C2 from the MSigDB.

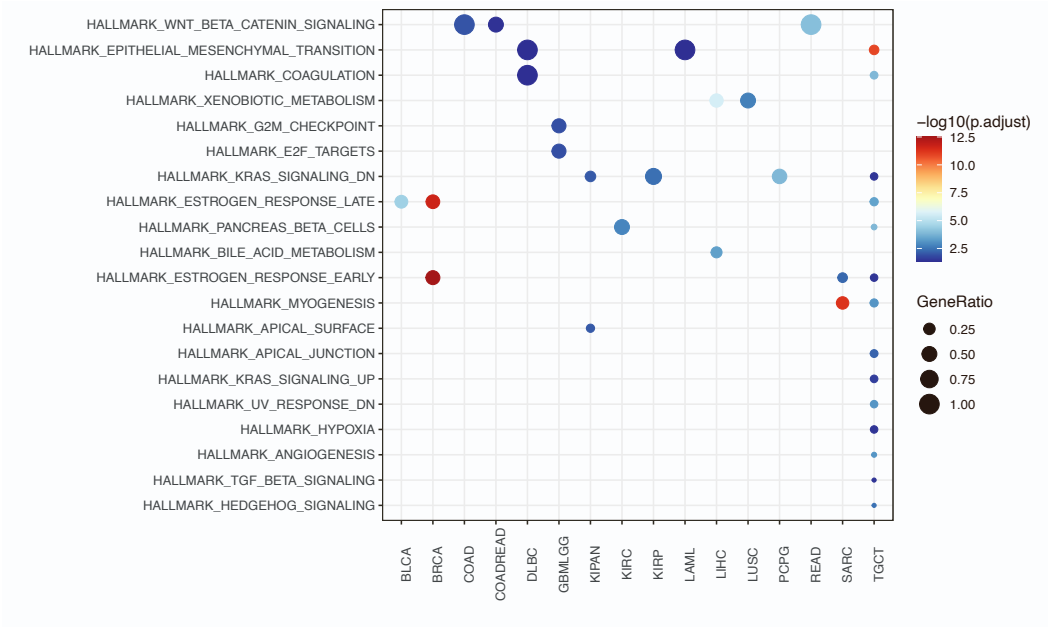


Figure S15. Dotplot of enriched pathways in downregulated genes associated with inter-tumoral heterogeneity across cancer types. Analysis was performed using clusterProfile R package with hallmark pathways from MSigDB.

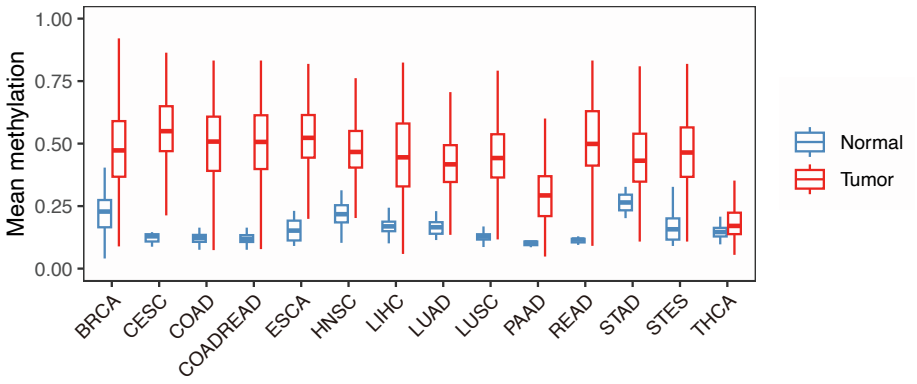


Figure S16. Mean methylation at chr1:119,527,091-119,527,476 across TCGA datasets.

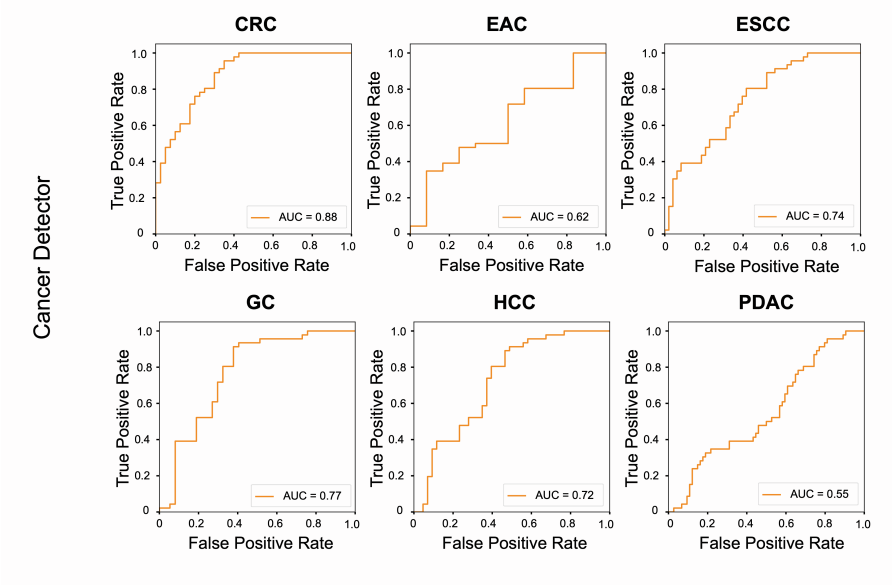


Figure S17. Evaluation the performance of cancer detection using the CancerDetector method. Six cancer types from GSE149438 dataset were tested.