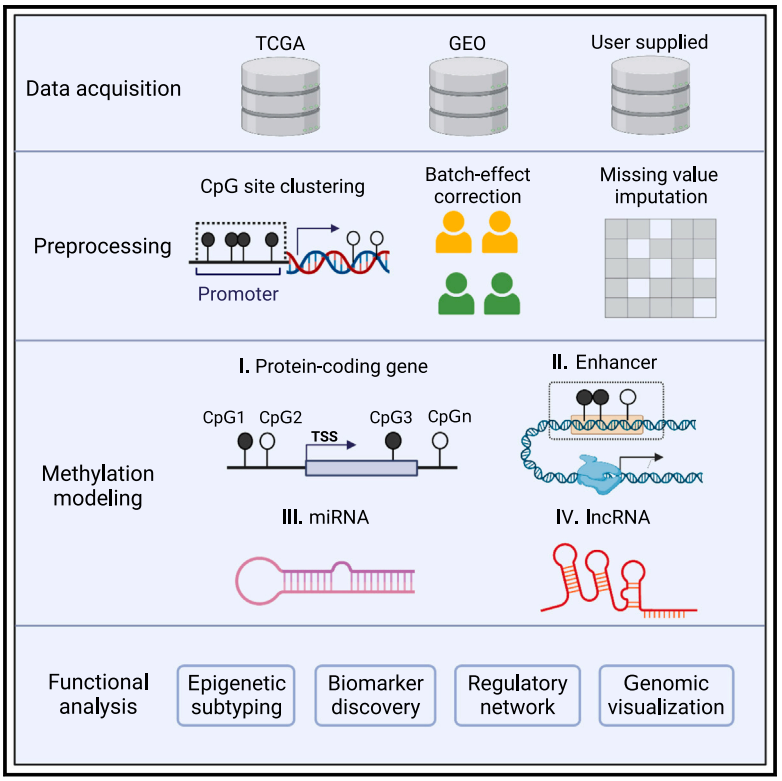
Cell Reports
Methods

Article

EpiMix is an integrative tool for epigenomic subtyping using DNA methylation

Graphical abstract



Highlights

- EpiMix detects abnormal DNA methylation (DNAme) present in small patient subsets
- EpiMix enables DNAme analysis of distal enhancers and non-coding RNAs
- Abnormal DNAme underlies suboptimal lipid metabolism in T cells
- Non-coding RNAs regulated by abnormal DNAme are biomarkers for lung cancer prognosis

Authors

Yuanning Zheng, John Jun, Kevin Brennan, Olivier Gevaert

Correspondence

ogevaert@stanford.edu

In brief

Zheng et al. introduce EpiMix, a computational tool for the integrated analysis of DNA methylation and gene expression at population levels. EpiMix employs a model-based approach that identifies epigenetic biomarkers for disease subtypes and personalized therapeutic targets.

Zheng et al., 2023, Cell Reports Methods *3*, 100515
July 24, 2023
https://doi.org/10.1016/j.crmeth.2023.100515

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

EpiMix is an integrative tool for epigenomic subtyping using DNA methylation

Yuanning Zheng,[1] John Jun,[1] Kevin Brennan,[1] and Olivier Gevaert[1,2,*]
[1]Stanford Center for Biomedical Informatics Research (BMIR), Department of Medicine & Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA
[2]Lead contact
*Correspondence: ogevaert@stanford.edu
https://doi.org/10.1016/j.crmeth.2023.100515

MOTIVATION Recent technological advancements have enabled genome-wide quantification of DNAme in large human populations. However, in diseases like cancer, abnormal DNA methylation patterns may only be present in specific subsets of a patient cohort. We aim to develop a statistical approach that models the distribution of DNAme in large patient cohorts and to characterize patient subsets with distinct DNAme profiles. This epigenetic subtyping can be essential in improving personalized diagnosis, treatment, and drug discovery.

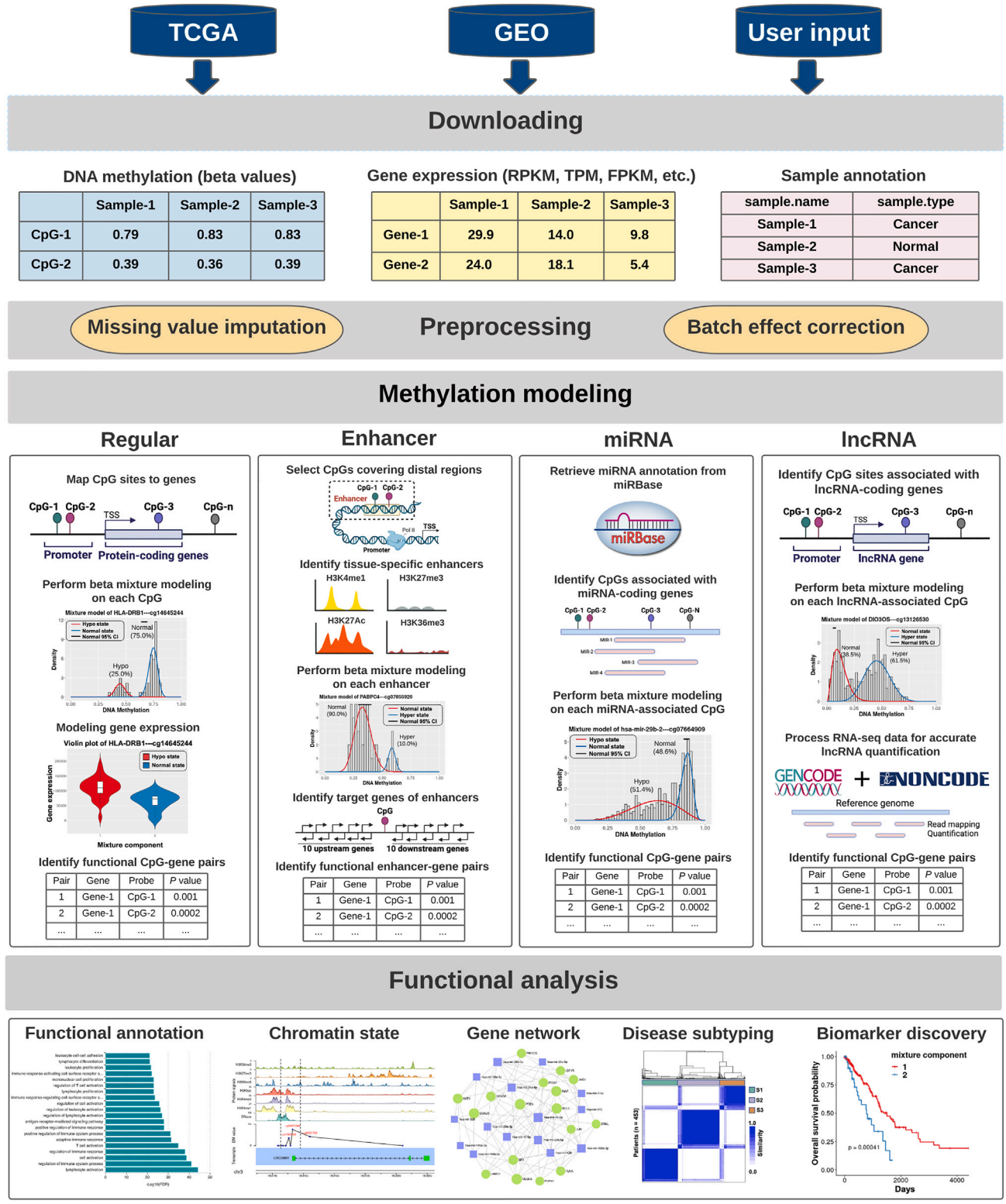
SUMMARY

DNA methylation (DNAme) is a major epigenetic factor influencing gene expression with alterations leading to cancer and immunological and cardiovascular diseases. Recent technological advances have enabled genome-wide profiling of DNAme in large human cohorts. There is a need for analytical methods that can more sensitively detect differential methylation profiles present in subsets of individuals from these heterogeneous, population-level datasets. We developed an end-to-end analytical framework named "EpiMix" for population-level analysis of DNAme and gene expression. Compared with existing methods, EpiMix showed higher sensitivity in detecting abnormal DNAme that was present in only small patient subsets. We extended the model-based analyses of EpiMix to *cis*-regulatory elements within protein-coding genes, distal enhancers, and genes encoding microRNAs and long non-coding RNAs (lncRNAs). Using cell-type-specific data from two separate studies, we discover epigenetic mechanisms underlying childhood food allergy and survival-associated, methylation-driven ncRNAs in non-small cell lung cancer.

INTRODUCTION

DNA methylation (DNAme) is one of the major epigenetic modifications that occurs primarily at the cytosine of cytosine-guanine dinucleotide (CpG) sequences in humans. This process involves the addition of a methyl (CH_3) group to DNA, and it plays a critical role in regulating gene expression and various biological processes. Aberrant DNAme has been associated with the development and progression of numerous human diseases.[1–3] Recent advances in microarray and next-generation sequencing technologies have enabled genome-wide quantification of DNAme at single-nucleotide resolution. Due to its quantitative and cost-effective nature, microarray-based technology has emerged as the method of choice for profiling DNAme in large human cohorts. Notably, The Cancer Genome Atlas (TCGA) project has utilized microarrays to generate DNAme profiles for over 10,000 specimens representing 33 cancer types. Several other public repositories, such as the Gene Expression Omnibus (GEO) database, also host DNAme data across a wide range of complex diseases.

Over the last decade, a number of computational approaches have been developed to identify genes that are abnormally methylated in human diseases. Some methods are tailored to the analysis of DNAme data from bisulfite sequencing,[4–7] while others are designed for array-based data or can be adapted to both platforms.[8–12] Many existing methods identify differentially methylated loci by comparing all samples from an experimental group with those from a control group. This type of comparison works well when the experimental population is assumed to be homogeneous. However, when the study population is large, abnormal DNAme may be present in only a subset of the patients, and this intra-population variation has been observed in cancers and many other diseases.[13–15] In addition, results from epigenetic screening of normal tissues showed that infrequent alterations in DNAme are associated with an increased risk of neoplastic transformation.[16–18] Detecting such infrequent DNAme changes may improve the identification of early carcinogenetic events and individuals at risk for developing cancers. In cases where abnormal DNAme occurred in only a small subset of



(legend on next page)

the patients, existing methods are not capable of capturing the signals of differential methylation. Therefore, there is a critical need to use a statistical approach to model the distribution of DNAme in large patient cohorts and to identify patient subsets with differential DNAme profiles. This epigenetic subtyping can be essential in improving personalized diagnosis, treatment, and drug discovery.

Furthermore, gene expression in mammalian cells is a result of a complex process coordinated by a broad range of genomic regulatory elements.[19,20] In many studies, CpG sites were mapped to genes based on linear genomic proximity. This mapping logic assumes that DNAme impacts transcriptional activity only when the genes are overlapped or close to the differentially methylated sites. However, emerging evidence has shown that distal enhancers, which may locate at a great linear genomic distance from their target genes, play a critical role in orchestrating spatiotemporal gene expression programs.[21] Abnormal DNAme at enhancers was frequently reported in cancers and many other diseases.[22,23] Therefore, the analysis of enhancer methylation can improve our understanding of how gene expression is regulated across physiological and pathological conditions.

Existing computational tools focus on the DNAme analysis of protein-coding genes. However, non-coding RNAs, such as microRNAs (miRNAs) and long non-coding RNAs (lncRNAs), also play an important role in regulating cellular processes.[24,25] Recent studies have shown that DNAme is a major epigenetic mechanism regulating ncRNA expression.[26,27] With existing methods, it is challenging to decipher how DNAme affects the expression of ncRNAs.

Here, we present EpiMix, a comprehensive analytical framework for the population-level analysis of DNAme and gene expression. Using a model-based computational approach, EpiMix is developed to identify abnormal DNAme at diverse genomic elements, including *cis*-regulatory elements within or surrounding protein-coding genes, distal enhancers, and genes encoding miRNAs and lncRNAs. In two separate studies, we showed that EpiMix identified methylation-driven pathways in T cells from childhood food allergy and methylation-driven ncRNAs in patients with non-small cell lung cancer. To improve usability, we disseminated EpiMix's algorithms in Bioconductor,[28] enabling end-to-end DNAme analysis. Furthermore, we developed a web tool for interactive exploration and visualization of EpiMix's results (https://epimix.stanford.edu). Overall, EpiMix offers an approach for discovering epigenetic biomarkers for disease subtypes and therapeutic targets for personalized medicine.

RESULTS

Overview of EpiMix workflow

EpiMix is an end-to-end analytical framework for modeling DNAme in heterogeneous patient cohorts and identifying differential DNAme associated with gene expression. The EpiMix framework consists of four functional modules: (1) data downloading, (2) preprocessing, (3) DNAme modeling, and (4) functional analysis (Figure 1). To analyze DNAme at functionally diverse genomic elements, we implemented four alternative analytic modes: "regular," "enhancer," "miRNA," and "lncRNA." The regular and enhancer modes aim to detect differential DNAme associated with the expression of protein-coding genes. Specifically, the regular mode analyzes DNAme sites within or immediately surrounding the genes, while the enhancer mode focuses on distal enhancers. The miRNA and lncRNA modes were built for the detection of DNAme affecting the expression of miRNAs and lncRNAs, respectively. With the functional analysis module, users can perform comprehensive exploratory analyses of the methylation-driven genes. This module integrates both in-house-developed and existing computational methods to enable diverse functional analyses and visualization of the differential DNAme.

Identifications of abnormal DNAme present in small sample subsets

To assess the sensitivity of EpiMix in detecting differential DNAme, we performed simulation experiments using a dataset containing DNAme measurement of quiescent $CD4^+$ T cells and antigen-activated T cells from 103 human subjects.[29] The dataset was generated using the Infinium MethylationEPIC array. We aimed to generate synthetic populations with differential DNAme that occurred only in specific subsets of the samples. First, we randomly selected a subset of CpGs (n = 300) from the quiescent group as baselines, such that the average beta values of the baseline CpGs ranged from 0.1 to 0.9. Next, we randomly selected a subset of samples from the activation group and combined them with the baseline group, while we controlled the number of samples and the mean difference in DNAme levels between the two groups. The final proportions of samples from the activation group in the combined dataset ranged from 3% to 50%, and the mean differences in beta values ranged from 0.1 to 0.7 (Figure 2A; STAR Methods). Finally, we compared the DNAme of the synthetic population with the baseline population (Figure 2A).

Our simulation experiments showed that the sensitivity of EpiMix was determined by the magnitude of differences in DNAme between the quiescent and the activated subjects. With a delta beta of 0.1, EpiMix was able to detect differential DNAme that was present in 3%–25% of the synthetic population, with a mean minimum detection threshold of 11%, corresponding to absolute sample count of 13 (Figures 2B and 2C). When the delta beta was 0.2 or higher, the minimum detection threshold ranged from 3% to 10%, with a mean threshold of 3.4% (absolute sample count = 4) (Figures 2B and 2D). These results indicated that EpiMix was able to detect abnormal DNAme

Figure 1. The EpiMix workflow

EpiMix framework includes four functional modules: downloading, preprocessing, methylation modeling, and functional analysis. EpiMix accepts user-provided custom datasets or can automatically download and preprocess data from public repositories. The methylation modeling module provides four alternative analytic modes: regular, enhancer, miRNA, and lncRNA. Each mode uses a custom algorithm to analyze DNAme at a specific type of genomic element. One major output from the DNAme modeling is a matrix of functional CpG-gene pairs, illustrating the differentially methylated CpGs whose DNAme states were associated with gene expression. With the functional analysis module, users can perform diverse analytical tasks for the differentially methylated genes.

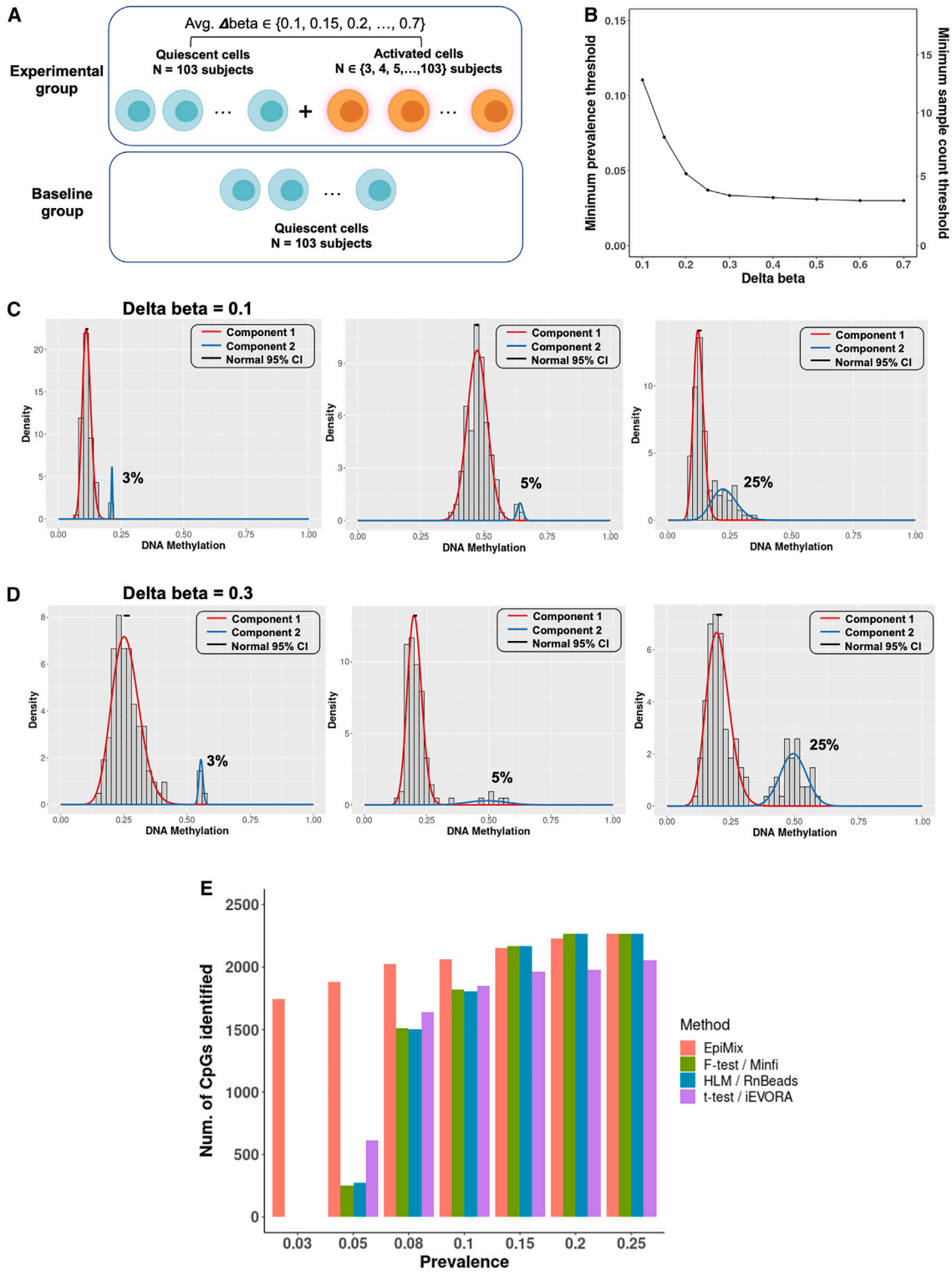


Figure 2. Evaluation of EpiMix performance using simulated data

(A) Design of the simulation study.

(B) Correlation between the delta beta values and the minimum detection threshold for the prevalence (left axis) and actual count (right axis) of the activated samples in the experimental group. The simulation was repeated 300 times using a different CpG site at each time, and the mean detection threshold is shown.

(legend continued on next page)

present in only small subsets of the tested population, and its sensitivity increased with larger differences in DNAme levels.

Next, we compared the performance of EpiMix with other statistical methods for identifications of differential DNAme, including the F-test used in Minfi,[10] the hierarchical linear model (HLM) used in RnBeads,[12,30] and the t test used in iEVORA.[17] Notably, the primary purpose of iEVORA was to identify CpGs with differential variances in preneoplastic tissues, and the differential methylation test was used only as an optional step for ranking the differentially variable features.[16–18] When the differential DNAme was present in 3% of the population, EpiMix detected the differential methylation signals at 1,747 CpG sites, whereas the other methods did not capture any differential DNAme (Figure 2E). When the differential DNAme was present in 5% of the population, EpiMix identified 3.1 times more differentially methylated CpGs than iEVORA and 3.6 times more CpGs than Minfi and RnBeads. Minfi and RnBeads only detected CpGs with high magnitude differences in DNAme (with an average delta beta of 0.6). In contrast, EpiMix detected CpGs with delta beta ranging from 0.1 to 0.7, with an average threshold of 0.3. When the prevalence of differential DNAme was 15% or higher, EpiMix detected a similar number of CpGs as the other three methods. These results indicated that EpiMix had higher sensitivity to detect differential DNAme, especially when the differential methylation was present in small sample subsets.

Modeling of DNAme at *cis*-regulatory elements within protein-coding genes

To test the regular mode of EpiMix, we used the complete dataset of antigen-activated T cells and quiescent T cells (n = 103 subjects per group).[29] In the activated T cells, 1,090 CpGs were differentially methylated compared with the quiescent cells (Figure 3A). By integrating sample-matched RNA sequencing (RNA-seq) data, we identified 748 protein-coding genes transcriptionally associated with these CpGs (Table S1). Among the differentially methylated CpGs, 746 (68.4%) CpGs associated with 504 genes were hypomethylated and 327 (30%) CpGs associated with 238 genes were hypermethylated (Figure 3A). These findings indicated a widespread loss of DNAme induced by antigens, consistent with the results from previous reports.[29] Gene Ontology (GO) analysis showed that the hypomethylated genes were associated with lymphocyte proliferation (e.g., *CCND2*, *CCND3*, *CDK6*, *CDK14*), T cell activation (e.g., *BCL2*, *CCL5*, *HLA-DPA1*, *HLA-DRB1* [*human leukocyte antigen DRB1*]), glycoprotein biosynthesis (e.g., *AGO2*, *ALG9*, *B3GNT5*, *B4GALT5*), and cytokine receptor activity (*IL1R1*, *IL1R2*, *IL21R*, *IL23R*) (Table S2). This result confirmed that EpiMix identified differential DNAme associated with T cell activation.

The differential methylation of many CpGs was observed only in a subset of patients. For instance, the *HLA-DRB1* gene was found to be hypomethylated in the antigen-activated T cells of only 25% of the subjects, while the majority (75%) had a normal methylation state similar to the quiescent T cells (Figure 3B). As expected, the gene expression levels of *HLA-DRB1* were significantly increased in the hypomethylated subjects compared with those with normal methylation levels (Figure 3C). Overall, the prevalence of hypomethylation ranged from 5.9%–100%, with a mean prevalence of 69.6% (Figure 3D). The prevalence of hypermethylation ranged from 5.8%–100%, with a mean prevalence of 47.3% (Figure 3E). These results indicated that the response to antigen stimulation in T cells varied between individuals.

We next investigated the genomic distribution of the differentially methylated CpGs. 39% (39.5% of the CpGs were located at the promoters, and 56.4% were located at introns (Figure S1A). To determine the enrichment of abnormal DNAme at various genomic regions, we analyzed publicly available chromatin immunoprecipitation sequencing (ChIP-seq) data of human naive $CD4^+$ T cells. To correct the background genomic distribution of CpG probes within the DNAme array, we used Monte Carlo-based permutation tests to calculate a ratio of observed to expected overlap between the differentially methylated sites and histone modification-enriched regions.[31] Our analysis revealed that the differentially methylated CpGs were significantly enriched at active promoters marked by H3K4me3 and H3K27ac, active enhancers marked by H3K4me1, and, to a lesser extent, actively transcribed gene bodies marked by H3K36me3 (Figure S1B). These results demonstrated that EpiMix was able to identify aberrant DNAme at lineage-defining *cis*-regulatory elements.

To allow users to explore the genomic locations and chromatin states associated with differentially methylated sites, EpiMix features genome browser-style visualization. We demonstrated this functionality by showcasing two regions of the interleukin-receptor gene *IL21R* where hypomethylation occurred (Figure 3F). The first region was located at the promoter, which overlapped with DNase I hypersensitivity sites and activating histone modifications (i.e., H3K4me1, H3K4me3, and H3K27ac). The second region was located at the 3′ untranslated region, enriched with histone modifications marking for active enhancers (i.e., H3K4me1 and H3K27ac). Consistent with the observed DNA hypomethylation, *IL21R* expression levels were significantly increased (Table S1; Wilcoxon rank-sum test, $p < 3.19E{-}08$).

Identification of functional DNAme at distal enhancers in food allergy

To demonstrate the enhancer mode of EpiMix, we investigated the impact of food allergy on DNAme in $CD4^+$ T cells from 82 allergic patients and 21 non-allergic controls from a publicly available dataset.[29] We hypothesized that the differential response of T cells to antigen-induced activation may be linked to allergic status. To identify allergy-associated changes in

(C) Density plots showing the mixture models when delta beta was 0.1 and the differential methylation was present in 3%, 5%, and 25% of the experimental group.

(D) Density plots showing the mixture models when delta beta was 0.3 and the differential methylation was present in 3%, 5%, and 25% of the experimental group.

(E) Number of differentially methylated CpGs detected by different methods when the differential methylation was present in from 3% to 25% of the population. For all methods, the same set of CpGs (n = 2, 700) was used.

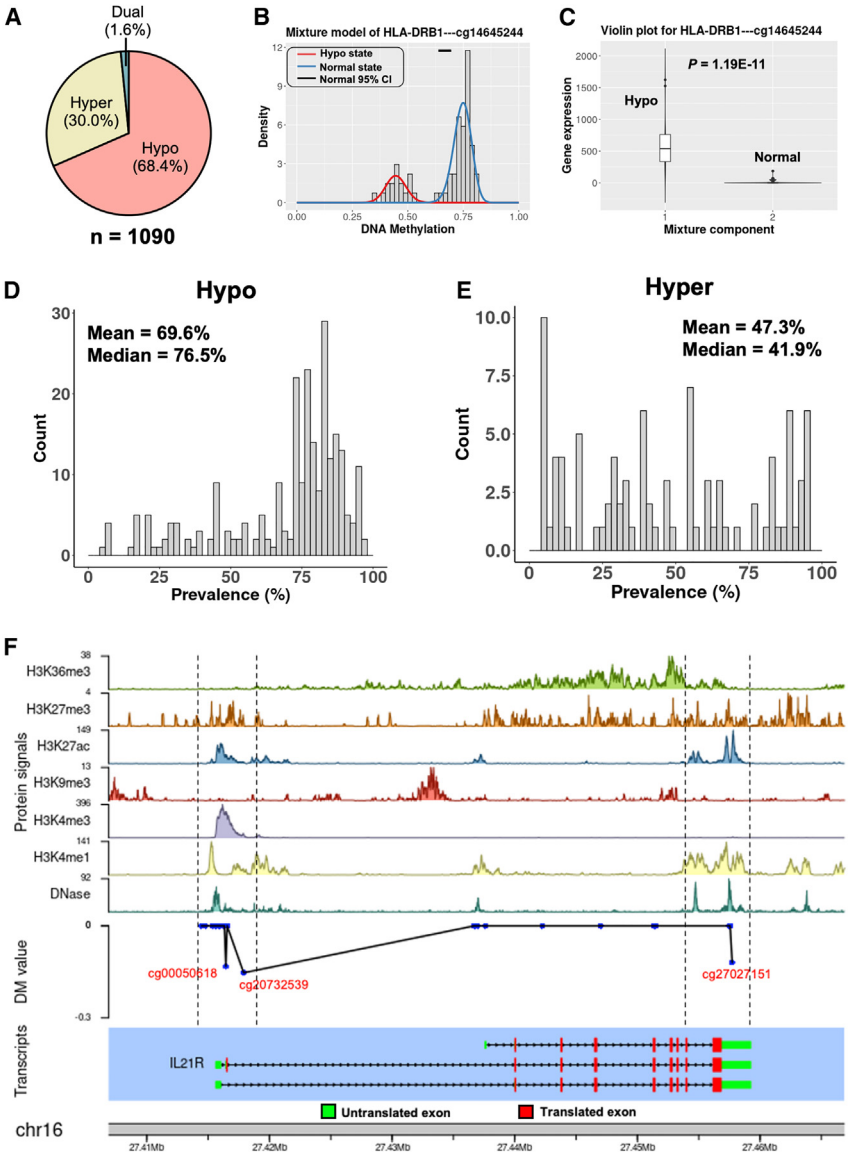


Figure 3. Identifications of differential DNAme resulting from antigen-induced T cell activation

(A) Proportions of the hypo-, hyper-, and dual-methylated CpGs in antigen-activated T cells. The dual-methylated CpGs refer to the CpGs that were hypomethylated in some individuals and hypermethylated in some other individuals.

(B and C) Mixture model of a CpG associated with the *HLA-DRB1* gene (B) and *HLA-DRB1* gene expression levels in different mixtures (C). Red indicates hypomethylation (n = 26), while blue indicates normal methylation (n = 77). Gene expression levels were compared with Wilcoxon rank-sum test.

(D and E) Density plots showing the prevalence distribution of the (D) hypo- and (E) hypermethylated CpGs.

(F) Genome-browser style visualization of the chromatin state, DM values, and transcript structure of the *IL21R* gene. The hypomethylated CpGs are labeled in red. The DM value represents the mean difference in beta values between the hypomethylated subjects vs. the normally methylated subjects. DM = 0: normal methylation; DM < 0: hypomethylation.

DNAme, we compared antigen-activated T cells from the allergic patients with those from the non-allergic controls. Using a permutation approach (Figure S2; STAR Methods), we found 107 differentially methylated enhancers that were functionally linked to the expression of 119 genes (Figure 4A). The number of target genes of each enhancer ranged from 1 to 3, resulting in 131 significant enhancer-gene pairs (Table S3). This result is consistent with previous studies showing that enhancers typically loop to and are associated with the activation of 1–3 promoters.[32,33] Of the functional enhancers, 21 out of 107 (19.6%) enhancers associated with 24 genes were hypomethylated, while 82 out of 107 (76.7%) enhancers associated with 92 genes were hypermethylated (Figure 4A). This result indicated that there was a global gain of DNAme at enhancers in food allergy.

The genomic distance between enhancers and their target genes ranged from 4.5 kb to 1.7 Mb, with a median distance of 148 kb (Figure 4B). In a previous study, Jin et al. used the high-throughput chromosome conformation capture (Hi-C) assay to investigate promoter-enhancer interactions and showed that approximately 25% of the enhancer-promoter pairs are within a 50 kb range and approximately 57% span 100 kb or greater genomic distance, with a median distance of 124 kb.[34] Another study by Rao et al. showed that the distance between enhancers and promoters spans from 40 kb to 3 MB, with a median distance of 185 kb.[35] Our data agree with these experimentally generated results.

To further characterize the enhancer-gene linkage, we investigated how often the functional enhancers associated with the nearest gene promoter. We ranked the 20 adjacent genes of each enhancer by their genomic distance to the enhancer. As shown in Figure 4C, only 6.1% of the times did the enhancer associate with the nearest promoter, whereas the majority of the enhancers skipped one or more intervening genes to associate with promoters farther away. In line with his result, a previous study using the chromosome 5C assay showed that only ~7% of the time did the distal elements loop to the promoter of the nearest gene, whereas the majority of enhancers bypass the nearest promoter and loop to promoters farther away.[36] These results confirmed that EpiMix identified true distal *cis*-regulatory events.

Genes linked to the differentially methylated enhancers were related to the lipid metabolism (*LDLR*, *CAT*, *LPIN2*, *SREBF1, PIK3C2B*) and T cell activation (*CASP3*, *MALT*, *PRKCZ*, *SMAD3*). As shown in Figure 4D, the enhancer linked to the

CellPress
OPEN ACCESS

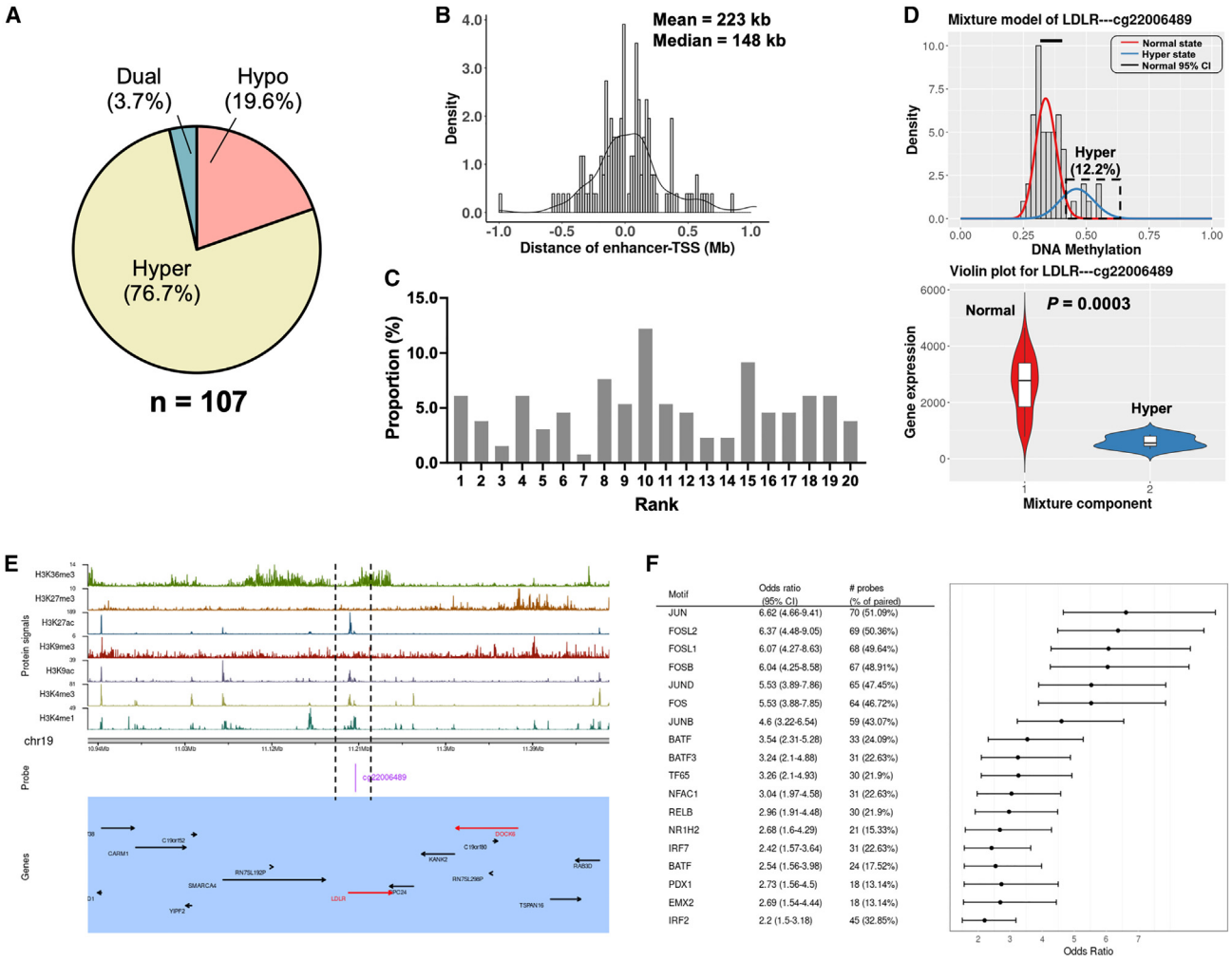


Figure 4. Identifications of differentially methylated enhancers associated with food allergy

(A) Proportions of the hypo-, hyper-, and dual-methylated enhancers in children with food allergy.

(B) Distribution of the linear genomic distance between enhancers and their gene targets.

(C) For each functional enhancer, the 20 adjacent genes were ranked by genomic distance. Bars show the proportions of the functionally linked genes in each rank.

(D) Mixture model of the *LDLR* gene (top panel) and *LDLR* gene expression levels in different mixtures (bottom panel). Red indicates normal methylation (n = 72), while blue indicates hypermethylation (n = 10). Gene expression levels were compared by Wilcoxon rank-sum test.

(E) Integrative visualization of the chromatin states and the adjacent genes of the hypermethylated enhancer shown in (D). The genes in the functional CpG-gene pairs are shown in red, while the others are shown in black.

(F) Enriched TF motifs and odds ratios for the differentially methylated enhancers. To find significantly enriched motifs, we used all the distal CpGs as the background and the functional enhancers as the targets.

LDLR gene was hypermethylated in 12.2% of the allergic patients, and the gene expression of *LDLR* was significantly decreased in the hypermethylated patients. Integrative visualization (Figure 4E) showed that the hypermethylated enhancer overlapped with the DNase I hypersensitivity site and was enriched with histone modifications marking for active enhancers, including H3K4me1 and H3K27ac and, to a lesser extent, H3K4me3 and H3K9ac. The *LDLR* gene encodes a low-density lipoprotein receptor that transports cholesterol from the blood into the cell, which plays a critical role in regulating T cell lipid metabolism.[37] Our results suggested that T cells from a small subset of the allergic patients may have an abnormal lipid metabolic profile due to enhancer hypermethylation.

Enhancers are enriched for sequences bound by site-specific transcription factors (TFs). Hypermethylation of enhancers can suppress gene transcription by decreasing the binding affinity of TFs.[38,39] To investigate potential regulatory mechanisms through which abnormal enhancer methylation impacted gene expression and susceptibility to food allergy, we carried out motif enrichment analysis of the differentially methylated enhancers identified by EpiMix. Our analysis revealed significant enrichment of binding sites for several key TFs that play critical roles in

regulating immune gene activation in T cells, including Jun-related factors (JUN, JUND), Fos-related factors (FOS, FOSL1, FOSL2, FOSB), BATF-related factors (BATF, BATF3), and interferon-regulatory factors (IRF2, IRF5, IRF7) (Figure 4F; Table S4). These findings are consistent with previous studies showing that dysregulation of these TFs can lead to aberrant immune responses and contribute to the development of allergies.[40,41] Our results demonstrated that the abnormal DNAme at enhancers affected the target gene response of these TFs and increased the subsequent risk for developing food allergy.

Identification of methylation-driven miRNAs in human lung cancer

To demonstrate the miRNA mode of EpiMix, we used a lung adenocarcinoma dataset containing DNAme and miRNA expression profiles of 457 tumors and 32 adjacent normal tissues.[42] Since both tumors and normal lung tissues are composed of multiple cell types, including epithelial cells, fibroblasts, hematopoietic cells, and endothelial cells, DNAme and gene expression profiles collected at the tissue level ("bulk") may not accurately reflect alterations present in specific cell types.[43,44] Since the majority of DNAme alterations in lung cancers have been found to occur in epithelial cells,[45,46] we aimed at identifying DNAme alterations specific to the epithelial cell population. To resolve the confounding effects of other cell types, we leveraged previously validated computational methods to estimate the proportions of epithelial cells and infer epithelial-specific methylomes and transcriptomes (Figure S3; STAR Methods). We then validated our results using an independent dataset of lung cancer epithelial cells.[46] Our results showed that the majority (86.6%–88.4%) of the differentially methylated CpGs identified from the independent dataset were also detected using our deconvoluted data for epithelial cells (Figure S4). To further demonstrate the effectiveness of the deconvolution procedures, we compared the hypo- and hypermethylated CpGs identified using the bulk tissue data with those identified from the deconvoluted data. While the bulk data led to the discovery of a greater number of differentially methylated CpGs, the overlap with the independent validation dataset was found to be 32%–45% lower than that of the deconvoluted data (Figures S4B and S4D). These results underscored the effectiveness of the deconvolution procedure in removing noise from the DNAme data, enabling us to obtain a more accurate assessment of epithelial-specific DNAme alterations.

Using the deconvoluted data of epithelial cells, we identified 272 differentially methylated CpGs that were associated with the expression of 92 miRNA genes (Figure 5A; Table S5). Among these CpGs, 138 (50.8%) CpGs associated with 66 genes were hypomethylated, and 55 (20.2%) CpGs associated with 37 genes were hypermethylated. 65% (63.6%) of the functional CpGs were located at the miRNA promoters, and this proportion was significantly higher than randomly selected CpGs (Figure S1C; Fisher's exact test, $p = 0.003$). To further investigate these findings, we analyzed publicly available ChIP-seq data of lung and found that the differentially methylated regions were enriched with histone modifications (i.e., H3K27ac, H3K4me1, and H3K4me3) marking for actively transcribed promoters and enhancers (Figure S1D). The prevalence of hypomethylation ranged from 1.1% to 66.7%, with a mean prevalence of 18% (Figure 5B). Similarly, the prevalence of hypermethylation ranged from 2.6% to 83.7%, with a mean prevalence of 24.9% (Figure 5C). These results indicated that the majority of differential DNAme associated with miRNA genes occurred in less than 25% of the patient population.

miRNAs are essential regulators of gene expression, mediating the destabilization and translational suppression of target messenger RNAs.[47] To gain systematic insight into the biological functions of the methylation-driven miRNAs, we curated experimentally validated miRNA targets from the literature,[48] resulting in a preliminary set of 10,374 protein-coding genes associated with 78 miRNAs. To further refine this list in the context of lung cancers, we compared messenger RNA expression levels of each target gene between patients with abnormal states of the miRNA regulator with those with normal methylation states (STAR Methods). This led to the discovery of 4,430 protein-coding genes whose messenger RNA expression levels were significantly altered in the expected direction with their miRNA regulators. Functional enrichment analysis of these miRNA targets revealed their associations with the regulation of cell cycle (e.g., *CCNE1*, *CDK4*), focal adhesion (e.g., *COL4A2*, *VEGFA*), and PD-L1 expression and PD-1 checkpoint pathway (e.g., *CD274*, *CD28*) (Figures 5D–5F; Table S6). These results provided mechanistic insights into how abnormal DNAme of miRNAs contributes to the development and progression of lung cancer.

We next investigated whether the DNAme of miRNAs was associated with patient survival.[49] Among the 92 methylation-driven miRNAs, we identified 22 miRNAs whose methylation states were significantly correlated with patient survival (Table S7; log rank test, $p < 0.05$). Half (11/22, 50%) of the miRNAs were hypomethylated, and the others (11/22, 50%) were hypermethylated. Some of the miRNAs were already known to be associated with lung cancer survival, such as *MIR29C*,[50] *MIR34A*,[51] and *MIR148A*.[52] However, we also identified many new survival-associated miRNAs. We found that *MIR30A*, which was related to the PD-L1 checkpoint pathway, was hypermethylated in 8.6% of the patients, and the hypermethylated patients showed significantly worse survival than the normally methylated patients (Figures 5F and 5G; hazard ratio = 1.50, $p = 0.001$). Next, *MIR1292* was hypomethylated in 8.6% of the patients, and the hypomethylated patients had significantly worse survival (Figure 5H; hazard ratio = 1.39, $p = 0.0008$). These results demonstrated that EpiMix enabled us to identify miRNAs that were differentially methylated in only small patient subsets yet had a significant impact on patient prognosis. The data suggested that targeting miRNA expression can be a therapeutic strategy to inhibit tumor progression and improve patient survival.

Identification of methylation-driven lncRNAs in human lung cancer

To demonstrate the lncRNA mode, we analyzed the same lung adenocarcinoma dataset.[42] Compared with protein-coding genes, lncRNAs are shorter, lower expressed, less evolutionarily conserved, and expressed in a more tissue-specific manner.[53] To precisely quantify lncRNA expression from RNA-seq data, we utilized our previously developed pipeline.[54] We combined the transcriptome annotations from GENCODE and NONCODE.[55] Raw sequencing reads were aligned to the combined

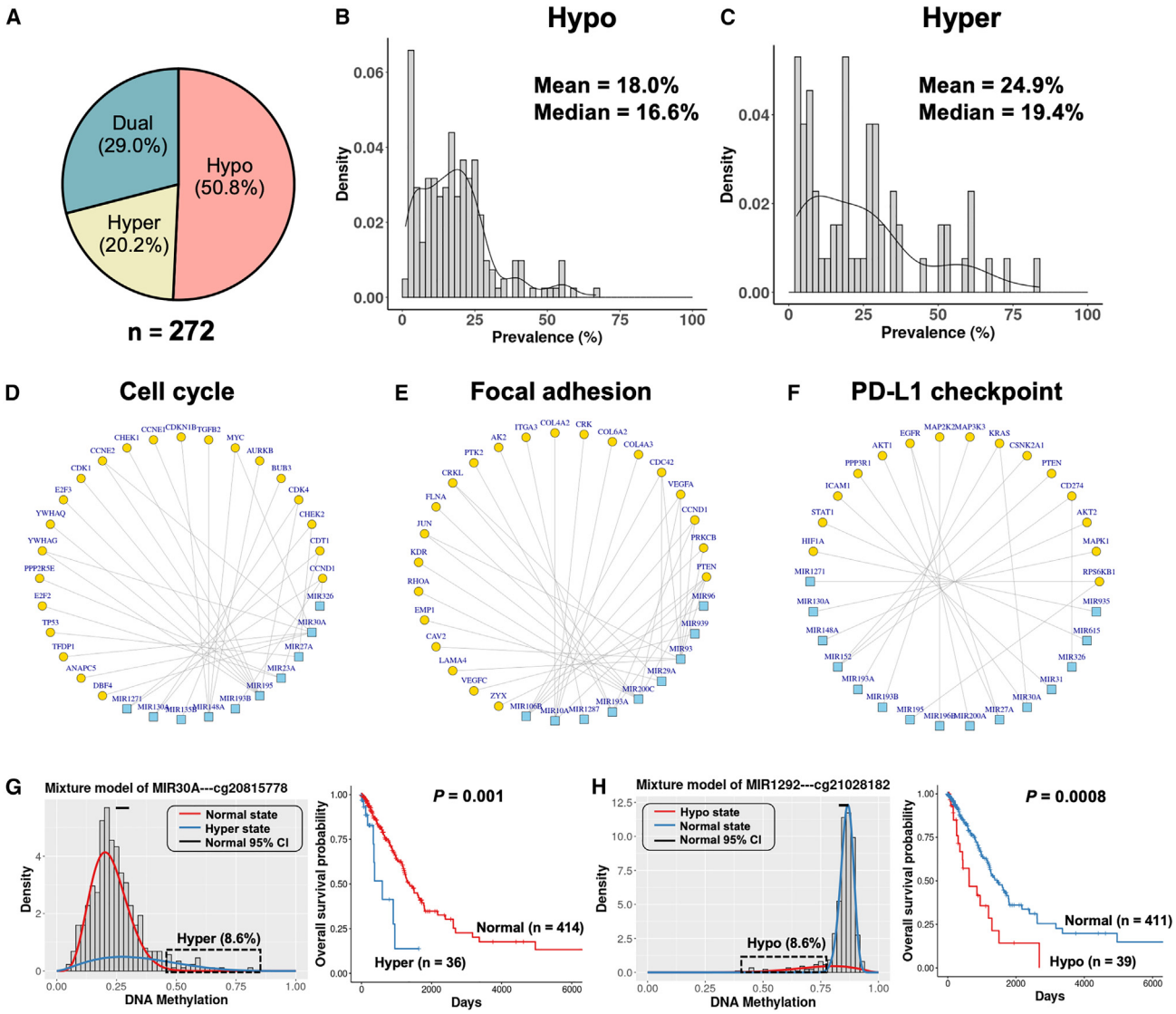


Figure 5. Identifications of differentially methylated miRNA-coding genes in human lung cancers

(A) Proportions of the hypo-, hyper-, and dual-methylated CpGs of miRNAs in lung cancer.

(B–C) Density plots showing the (B) prevalence distribution of the differentially methylated miRNAs in lung cancers (n = 457), (B) prevalence of hypomethylation, and (C) prevalence of hypermethylation.

(D–F) Network visualization of the methylation-driven miRNAs and their target genes related to the (D) cell cycle pathway, (E) focal adhesion, and (F) PD-L1 expression and PD-L1 checkpoint pathway. Blue squares: miRNAs, yellow circles: miRNA targets.

(G) Mixture model of the *MIR30A* gene (left panel) and Kaplan-Meier survival curves of patients in different mixtures (right panel). Red indicates normal methylation, and blue indicates hypermethylation. Gene expression levels were compared by Wilcoxon rank-sum test.

(H) Mixture model of the *MIR1292* gene (left panel) and Kaplan-Meier survival curves of patients in different mixtures (right panel). Red indicates hypomethylation, and blue indicates normal methylation.

transcriptome reference and quantified using the Kallisto-Sleuth algorithm.[56,57] This pipeline allowed us to detect the expression of 2,475 lncRNAs, which was three times higher compared with the lncRNAs detected by the traditional STAR-HTSeq pipeline. We then computationally deconvoluted bulk DNAme and lncRNA expression data of each sample to epithelial-specific DNAme and lncRNA expression values, and we aimed at identifying epithelial-specific DNAme alterations in cancers compared with normal tissues (Figure S3; STAR Methods).

We found 397 CpGs functionally associated with the expression of 132 lncRNAs in epithelial cells (Figure 6A; Table S8). Of these CpGs, 146 (36.8%) CpGs associated with 69 genes were hypomethylated, and 187 (47.1%) CpGs associated with 73 genes were hypermethylated. 72% of the functional CpGs were located at the promoters, and this proportion was significantly higher than randomly selected CpGs (Figure S1E; Fisher's exact test, $p < 0.0001$). The differentially methylated regions were enriched with histone modifications marking for actively

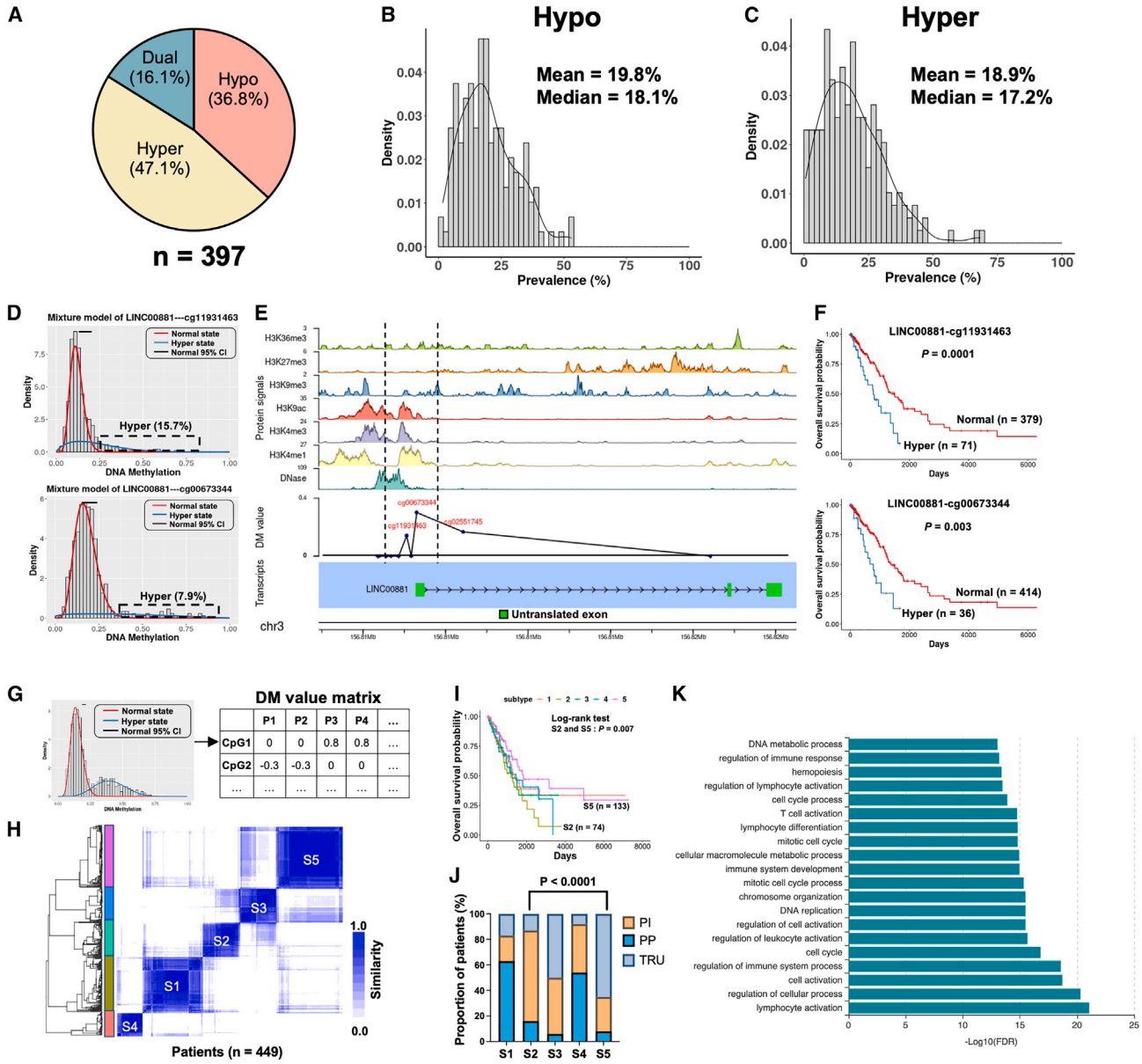


Figure 6. Identifications of differentially methylated lncRNA-coding genes in human lung cancers

(A) Proportions of the hypo-, hyper-, and dual-methylated CpGs of lncRNA genes in epithelial cells from lung cancers.

(B and C) Density plot showing the prevalence distribution of the (B) hypo- and (C) hypermethylated lncRNAs in the lung cancer cohort (n = 457).

(D) Mixture models of the *LINC00881* gene at two different CpG sites. Red indicates normal methylation, and blue indicates hypermethylation.

(E) Integrative visualization of the transcript structure, DM values, and chromatin state associated with the *LINC00881* gene. DM = 0: normal methylation; DM > 0: hypermethylation.

(F) Kaplan-Meier survival curves of patients in the normally methylated and the hypermethylated mixtures. Red indicates normal methylation, and blue indicates hypermethylation.

(G) Schematic representation of the DM value matrix. The rows correspond to CpG sites, and the columns correspond to patients. DM values represent the mean differences in DNAme levels between patients in each mixture component identified in the experimental group compared with the control group. At each CpG site, patients in the same mixture component have the same DM values.

(H) Consensus matrix showing patient clusters based on the DM values of lncRNAs (n1 = 120, n2 = 74, n3 = 72, n4 = 50, n5 = 133).

(I) Kaplan-Meier survival curves of patients in different clusters.

(J) Proportions of PP, PI, and TRU subtypes in different patient clusters. Fisher's exact test was used to compare subtype distributions (S2 vs. S5: $p < 0.0001$).

(K) Top 20 enriched GO terms of the methylation-driven lncRNAs in lung cancer. DM, differential methylation; PP, proximal proliferative; TRU, terminal respiratory unit; PI, proximal inflammatory.

transcribed promoters and enhancers, including H3K27ac, H3K4me1, and H3K4me3 (Figure S1F).

The majority of differential methylation events were detected in less than half of the patient cohort. The prevalence for hypomethylation ranged from 1.8% to 53%, with a mean value of 19.8% (Figure 6B), and the prevalence for hypermethylation ranged from 0.6% to 68.2%, with a mean value of 18.9% (Figure 6C). For instance, *LINC00881* was hypermethylated at CG11931463 in 15.7% of the patients and CG00673344 in 7.9% of the patients (Figure 6D). Both CpGs were located at the promoter (Figure 6E). Further analysis revealed that the hypermethylation of *LINC00881* was associated with significantly worse patient survival (Figure 6F; log rank test, $p < 0.01$). These data demonstrated that many lncRNAs exhibit differential methylation in only a small subset of patients with lung cancer, and these events were associated with clinical outcomes.

One of the major outputs from EpiMix is a differential methylation (DM) value matrix, which reflects the homogeneous subpopulations of samples with a particular methylation state (Figure 6G). An application of the DM value matrix is to identify DNAme-associated subtypes, where patients are clustered into robust and homogeneous groups based on their differential DNAme profiles. Using unsupervised consensus clustering, we discovered five DNAme subtypes (S1–S5) (Figure 6H). S5 had a significantly higher proportion of females (89/133 = 66.9%) compared with S1 (54/120 = 45.0%), S2 (36/74 = 48.6%), and S4 (16/50 = 32%) (Figure S5A; Fisher's exact test, $p < 0.01$). Importantly, patients from S5 had significantly better survival than those in S2 (Figure 6I; log rank test, $p = 0.007$). To assess how our classifications of lung adenocarcinoma were associated with published classifications, we compared our patient annotations with those based on molecular and histopathological features: (1) terminal respiratory unit (TRU), (2) proximal inflammatory (PI), and (3) proximal proliferative (PP).[42] We found that the TRU was significantly enriched in cluster S5 compared with S2, whereas S2 was enriched with the PI subtype (Figure 6J). This result was consistent with previous studies showing that the TRU has a favorable prognosis compared with the other subtypes, while the PI subtype has the worst prognosis.[42] In addition, S5 had a significantly lower number of patients with *TP53* mutations compared with S2 (Figure S5A). However, the proportions of patients with *KRAS* and *BRAF* mutations were not significantly different between patient groups.

To demonstrate the advantage of using the DM value matrix for identifying clinically relevant patient subsets, we compared the clustering results obtained from using the DM value matrix to those obtained from using raw beta values of the differentially methylated CpG sites (n = 397 sites). The patient subsets identified using raw beta values had low cluster consensus, and no significant association was found between patient subsets and survival outcome (Figures S5B and S5C). These results demonstrated that the DNAme subtypes discovered by EpiMix had prognostic values.

To investigate the biological functions of the differentially methylated lncRNAs, we curated 4,552 protein-coding genes transcriptionally associated with 76 lncRNAs.[58] GO analysis showed that the protein-coding genes were primarily associated with the regulation of lymphocyte activation, immune response, and DNA replication (Figure 6K; Table S9). These results indicated how DM of lncRNAs was involved in the regulation of lung cancer development and progression.

DISCUSSION

In this study, we present EpiMix, a comprehensive analytic framework for population-level analysis of DNAme and gene expression. We packaged the EpiMix algorithms in R, enabling end-to-end DNAme analysis. To enhance the user experience, we also implemented a web-based application (https://epimix.stanford.edu) for interactive exploration and visualization of EpiMix's results (Figure 7). EpiMix offers a range of diverse functionalities, including automated data downloading, preprocessing, methylation modeling, and functional analysis. The seamless connection of EpiMix to data from TCGA program and the GEO database enables DNAme analysis on a broad range of diseases. Here, we showed that EpiMix identified methylation-driven pathways in food allergy and lung cancer. However, EpiMix is not limited to these disease areas and can be readily applied to any other diseases.

EpiMix uses a beta mixture model to decompose the DNAme profiles in large patient populations. It enabled us to resolve the epigenetic subtypes within the patient population and pinpoint the individuals carrying differential DNAme profiles. We identified five DNAme subtypes in lung cancers using the DM values of lncRNAs. Patients of subtype 2 had significantly worse survival than patients of subtype 5, indicating that the DNAme subtypes discovered by EpiMix had prognostic values. The biological interpretation of DNAme subtypes requires the integration of data from other modalities, such as genetic mutations, lifestyle history, and other etiological features.

In addition, EpiMix was able to detect abnormal DNAme that was present in only small subsets of a patient cohort. In our simulation study, EpiMix detected more differentially methylated CpGs compared with existing methods when the DM occurred in only a small patient subset. Using the real lung cancer dataset, we identified miRNAs that were differentially methylated in only 1.1% of the patient population and lncRNAs differentially methylated in 0.6% of the patient population. We showed that over half of the miRNAs and lncRNAs were differentially methylated in only less than 20% of the patients. This unique feature of EpiMix to detect differential DNAme in small patient subsets enables us to identify epigenetic mechanisms underlying disease phenotypes. It can also be used to discover new epigenetic biomarkers and drug targets for improving personalized treatment.

Another feature of EpiMix is its ability to model DNAme at functionally diverse genomic elements. This includes *cis*-regulatory elements within or surrounding protein-coding genes, distal enhancers, and genes encoding miRNAs and lncRNAs. To model DNAme at distal enhancers, we selected the enhancers from the ENCODE and ROADMAP consortiums, in which enhancers of over a hundred human tissues and cell lines were identified using the chromatin-state discovery (ChromHMM).[59] Since enhancers are cell-type specific, EpiMix allows the users to select enhancers of specific cell types or tissues. In this study, we selected the enhancers of human blood and T cells, leading to the discovery of 40,311 CpGs of enhancers. In addition to enhancers, many other

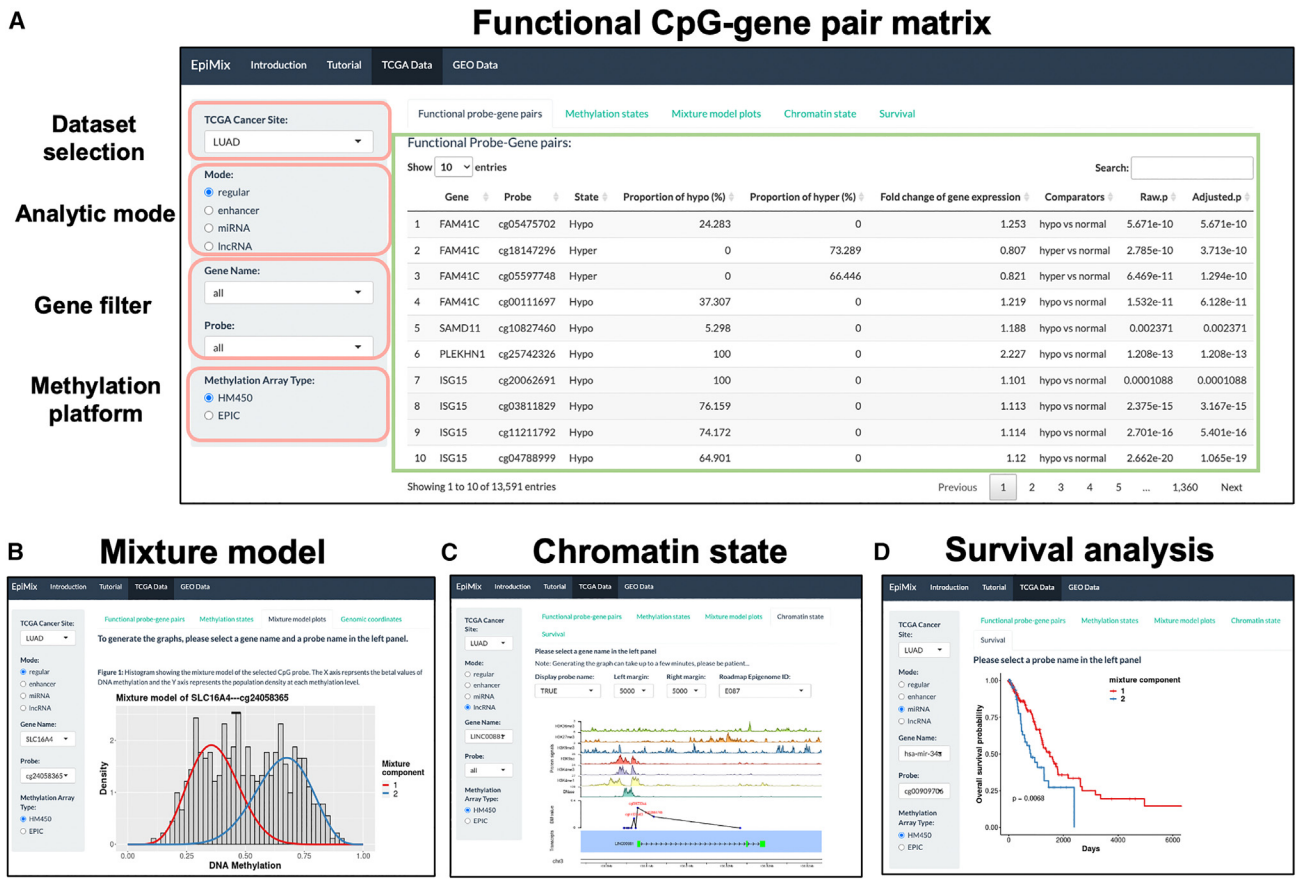


Figure 7. Screenshots of the EpiMix web application

(A) Interactive data filters and visualization of functional CpG-gene pair matrix.

(B) Visualization of the mixture model of the SLC16A4 gene in lung cancer.

(C) Genome-browser style visualization of the lncRNA gene LINC00881 in lung cancer.

(D) Kaplan-Meier survival curves of patients with different methylation states of the miRNA gene *MIR34A* in lung cancer.

regulatory elements were identified from the ROADMAP studies.[59] These include active transcription start site proximal promoters, zinc finger protein genes, bivalent regulatory elements, polycomb-repressed regions, and many others. By customizing the "chromatin state" parameter of EpiMix, users can target the DNAme analysis to any of these regulatory modules.

Despite the critical biological functions of ncRNAs, there are no existing tools that specifically analyze DNAme regulating their transcription. To analyze DNAme of miRNA genes, we utilized the miRNA annotation from miRBase, the largest and consistently updated knowledge base of miRNAs.[60] In addition, we selected CpGs at miRNA promoters by using a recent database that integrates the information of miRNA transcription start sites (TSSs) from 14 genome-wide studies across different human cell types and tissues.[61] This led to the discovery of 17,192 CpGs associated with 1,484 miRNAs in the HM450 array and 23,379 CpGs associated with 1,759 miRNAs in the EPIC array. With miRNA-seq data provided, EpiMix can select differential DNAme that was associated with miRNA expression. Different from profiling protein-coding gene expression, measuring miRNA expression requires special library preparation strategies that capture small RNAs from total RNAs.[62] Users are preferentially needed to supply miRNA expression data obtained from proper library preparation strategies.

Similarly, custom methods are needed to accurately quantify lncRNA expression from RNA-seq. We adopted the data processing pipeline developed from our previous study.[54] With this pipeline, we combined the transcriptome annotations from GENCODE and NONCODE. Raw sequencing reads were aligned to the combined transcriptome reference and quantified using the Kallisto-Sleuth algorithm.[56,57] Using this pipeline, we detected the expression of over 2,400 lncRNA genes. In this study, we have used our pipeline to generate lncRNA expression profiles for all the cancers in TCGA database, and users can retrieve these data with EpiMix. Note, if users plan to use EpiMix on non-TCGA datasets, they are encouraged to use this pipeline to profile lncRNA expression.

Future work will aim to extend the use of EpiMix to whole-genome bisulfite sequencing and to further improve the scalability. Furthermore, the rapid development of single-cell technologies enables coassay of DNAme and gene expression in thousands of cells. EpiMix can be used to identify differential DNAme that was present in only small subsets of a cell

population. Therefore, a joint analysis of single-cell methylome and transcriptome holds great promise for substantiating our goals, and the analytical framework presented here will be a valuable component for future research and applications.

Limitations of the study

Since EpiMix is designed for the analysis of data from large patient cohorts and the microarray-based DNAme assays are currently the most cost-effective approach for DNAme profiling, we tailored EpiMix to the analysis of microarray-based data. Future work is to extend the use of EpiMix to methylation sequencing data (e.g., bisulfite sequencing) and to further improve the scalability that would accommodate a broader range of applications.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data downloading
  - Preprocessing
  - Batch effect normalization
  - CpG annotation and filtering
  - CpG site aggregation and smoothing
  - Methylation modeling
  - Identifications of differentially methylated CpGs
  - Identifications of differential DNAme that was associated with transcription
  - Simulation study
  - Benchmark with existing methods
  - Imputation of cell-type-specific DNAme and gene expression values
  - Validation of deconvolution results
  - Genomic distribution of the differentially methylated CpGs
  - Motif enrichment analysis
  - Enrichment analysis of chromatin modifications
  - Functional enrichment analysis
  - Biomarker identification and survival analysis
  - Genome browser-style visualization
  - Identifications of DNAme subtypes
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2023.100515.

ACKNOWLEDGMENTS

We thank Dr. Sandra Steyaert, Alexander Henry Thieme (Department of Radiation Oncology, Charité - Universitätsmedizin Berlin), and Gautam Machiraju for helpful comments and suggestions. Research reported here was further supported by the National Cancer Institute (NCI) under awards R01 CA260271, U01 CA217851, and U01 CA199241. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.Z., K.B., and O.G.; methodology, Y.Z. and O.G.; investigation, Y.Z. and J.J.; writing – original draft, Y.Z.; writing – review & editing, Y.Z., J.J., K.B., and O.G.; funding acquisition, resources, and supervision, O.G.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 3, 2023
Revised: April 12, 2023
Accepted: June 1, 2023
Published: June 22, 2023

REFERENCES

1. Li, J., Li, L., Wang, Y., Huang, G., Li, X., Xie, Z., and Zhou, Z. (2021). Insights into the role of DNA methylation in immune cell development and autoimmune disease. Front. Cell Dev. Biol. *9*, 757318. https://doi.org/10.3389/fcell.2021.757318.
2. Si, J., Yang, S., Sun, D., Yu, C., Guo, Y., Lin, Y., Millwood, I.Y., Walters, R.G., Yang, L., Chen, Y., et al. (2021). Epigenome-wide analysis of DNA methylation and coronary heart disease: a nested case-control study. Elife *10*, e68671. https://doi.org/10.7554/elife.68671.
3. Zheng, Y., Luo, L., Lambertz, I.U., Conti, C.J., and Fuchs-Young, R. (2022). Early dietary exposures epigenetically program mammary cancer susceptibility through Igf1-mediated expansion of the mammary stem cell compartment. Cells *11*, 2558. https://doi.org/10.3390/cells11162558.
4. Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., and Mason, C.E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. *13*, R87. https://doi.org/10.1186/gb-2012-13-10-r87.
5. Park, Y., and Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics *32*, 1446–1453. https://doi.org/10.1093/bioinformatics/btw026.
6. Korthauer, K., Chakraborty, S., Benjamini, Y., and Irizarry, R.A. (2019). Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. Biostatistics *20*, 367–383. https://doi.org/10.1093/biostatistics/kxy007.
7. Wang, X., Hao, D., and Kadarmideen, H.N. (2021). GeneDMRs: an R package for gene-based differentially methylated regions analysis. J. Comput. Biol. *28*, 304–316. https://doi.org/10.1089/cmb.2020.0081.
8. Wang, D., Yan, L., Hu, Q., Sucheston, L.E., Higgins, M.J., Ambrosone, C.B., Johnson, C.S., Smiraglia, D.J., and Liu, S. (2012). IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics *28*, 729–730. https://doi.org/10.1093/bioinformatics/bts013.
9. Warden, C.D., Lee, H., Tompkins, J.D., Li, X., Wang, C., Riggs, A.D., Yu, H., Jove, R., and Yuan, Y.-C. (2013). COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. Nucleic Acids Res. *41*, e117. https://doi.org/10.1093/nar/gkt242.
10. Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363–1369. https://doi.org/10.1093/bioinformatics/btu049.

11. Silva, T.C., Coetzee, S.G., Gull, N., Yao, L., Hazelett, D.J., Noushmehr, H., Lin, D.-C., and Berman, B.P. (2019). ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. Bioinformatics *35*, 1974–1977. https://doi.org/10.1093/bioinformatics/bty902.

12. Müller, F., Scherer, M., Assenov, Y., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. (2019). RnBeads 2.0: comprehensive analysis of DNA methylation data. Genome Biol. *20*, 55. https://doi.org/10.1186/s13059-019-1664-9.

13. Shaknovich, R., Geng, H., Johnson, N.A., Tsikitas, L., Cerchietti, L., Greally, J.M., Gascoyne, R.D., Elemento, O., and Melnick, A. (2010). DNA methylation signatures define molecular subtypes of diffuse large B-cell lymphoma. Blood *116*, e81–e89. https://doi.org/10.1182/blood-2010-05-285320.

14. Chen, X., Zhang, J., and Dai, X. (2019). DNA methylation profiles capturing breast cancer heterogeneity. BMC Genom. *20*, 823. https://doi.org/10.1186/s12864-019-6142-y.

15. Schenkel, L.C., Aref-Eshghi, E., Rooney, K., Kerkhof, J., Levy, M.A., McConkey, H., Rogers, R.C., Phelan, K., Sarasua, S.M., Jain, L., et al. (2021). DNA methylation epi-signature is associated with two molecularly and phenotypically distinct clinical subtypes of Phelan-McDermid syndrome. Clin. Epigenet. *13*, 2. https://doi.org/10.1186/s13148-020-00990-7.

16. Teschendorff, A.E., Jones, A., Fiegl, H., Sargent, A., Zhuang, J.J., Kitchener, H.C., and Widschwendter, M. (2012). Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. Genome Med. *4*, 24. https://doi.org/10.1186/gm323.

17. Teschendorff, A.E., Gao, Y., Jones, A., Ruebner, M., Beckmann, M.W., Wachter, D.L., Fasching, P.A., and Widschwendter, M. (2016). DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. Nat. Commun. *7*, 10478. https://doi.org/10.1038/ncomms10478.

18. Teschendorff, A.E., and Relton, C.L. (2018). Statistical and integrative system-level analysis of DNA methylation data. Nat. Rev. Genet. *19*, 129–147. https://doi.org/10.1038/nrg.2017.86.

19. Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., 3rd, Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the cancer cell line encyclopedia. Nature *569*, 503–508. https://doi.org/10.1038/s41586-019-1186-3.

20. Partridge, E.C., Chhetri, S.B., Prokop, J.W., Ramaker, R.C., Jansen, C.S., Goh, S.-T., Mackiewicz, M., Newberry, K.M., Brandsmeier, L.A., Meadows, S.K., et al. (2020). Occupancy maps of 208 chromatin-associated proteins in one human cell type. Nature *583*, 720–728. https://doi.org/10.1038/s41586-020-2023-4.

21. Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. Nat. Rev. Genet. *20*, 437–455. https://doi.org/10.1038/s41576-019-0128-0.

22. Yao, L., Shen, H., Laird, P.W., Farnham, P.J., and Berman, B.P. (2015). Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. Genome Biol. *16*, 105. https://doi.org/10.1186/s13059-015-0668-3.

23. Cribbs, A.P., Kennedy, A., Penn, H., Amjadi, P., Green, P., Read, J.E., Brennan, F., Gregory, B., and Williams, R.O. (2015). Methotrexate restores regulatory T cell function through demethylation of the FoxP3 upstream enhancer in patients with rheumatoid arthritis. Arthritis Rheumatol. *67*, 1182–1192. https://doi.org/10.1002/art.39031.

24. Wang, L., Sinnott-Armstrong, N., Wagschal, A., Wark, A.R., Camporez, J.-P., Perry, R.J., Ji, F., Sohn, Y., Oh, J., Wu, S., et al. (2020). A MicroRNA linking human positive selection and metabolic disorders. Cell *183*, 684–701.e14. https://doi.org/10.1016/j.cell.2020.09.017.

25. Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. Nat. Rev. Mol. Cell Biol. *22*, 96–118. https://doi.org/10.1038/s41580-020-00315-9.

26. Watanabe, K., Emoto, N., Hamano, E., Sunohara, M., Kawakami, M., Kage, H., Kitano, K., Nakajima, J., Goto, A., Fukayama, M., et al. (2012). Genome structure-based screening identified epigenetically silenced microRNA associated with invasiveness in non-small-cell lung cancer. Int. J. Cancer *130*, 2580–2590. https://doi.org/10.1002/ijc.26254.

27. Zhang, M., Wu, J., Zhong, W., Zhao, Z., and He, W. (2021). DNA-methylation-induced silencing of DIO3OS drives non-small cell lung cancer progression via activating hnRNPK-MYC-CDC25A axis. Mol. Ther. Oncolytics *23*, 205–219. https://doi.org/10.1016/j.omto.2021.09.006.

28. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. *5*, R80. https://doi.org/10.1186/gb-2004-5-10-r80.

29. Martino, D., Neeland, M., Dang, T., Cobb, J., Ellis, J., Barnett, A., Tang, M., Vuillermin, P., Allen, K., and Saffery, R. (2018). Epigenetic dysregulation of naive CD4+ T-cell activation genes in childhood food allergy. Nat. Commun. *9*, 3308. https://doi.org/10.1038/s41467-018-05608-4.

30. Assenov, Y., Müller, F., Lutsik, P., Walter, J., Lengauer, T., and Bock, C. (2014). Comprehensive analysis of DNA methylation data with RnBeads. Nat. Methods *11*, 1138–1140. https://doi.org/10.1038/nmeth.3115.

31. Simovski, B., Vodák, D., Gundersen, S., Domanska, D., Azab, A., Holden, L., Holden, M., Grytten, I., Rand, K., Drabløs, F., et al. (2017). GSuite HyperBrowser: integrative analysis of dataset collections across the genome and epigenome. GigaScience *6*, 1–12. https://doi.org/10.1093/gigascience/gix032.

32. Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. *47*, 598–606. https://doi.org/10.1038/ng.3286.

33. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al. (2016). Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell *167*, 1369–1384.e19. https://doi.org/10.1016/j.cell.2016.09.037.

34. Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294. https://doi.org/10.1038/nature12644.

35. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021.

36. Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113. https://doi.org/10.1038/nature11279.

37. Bietz, A., Zhu, H., Xue, M., and Xu, C. (2017). Cholesterol metabolism in T cells. Front. Immunol. *8*, 1664. https://doi.org/10.3389/fimmu.2017.01664.

38. Rasmussen, K.D., Berest, I., Keβler, S., Nishimura, K., Simón-Carrasco, L., Vassiliou, G.S., Pedersen, M.T., Christensen, J., Zaugg, J.B., and Helin, K. (2019). TET2 binding to enhancers facilitates transcription factor recruitment in hematopoietic cells. Genome Res. *29*, 564–575. https://doi.org/10.1101/gr.239277.118.

39. Wang, L., Ozark, P.A., Smith, E.R., Zhao, Z., Marshall, S.A., Rendleman, E.J., Piunti, A., Ryan, C., Whelan, A.L., Helmin, K.A., et al. (2018). TET2 coactivates gene expression through demethylation of enhancers. Sci. Adv. *4*, eaau6986. https://doi.org/10.1126/sciadv.aau6986.

40. Li, P., Spolski, R., Liao, W., Wang, L., Murphy, T.L., Murphy, K.M., and Leonard, W.J. (2012). BATF–JUN is critical for IRF4-mediated transcription in T cells. Nature *490*, 543–546. https://doi.org/10.1038/nature11530.

41. Glasmacher, E., Agrawal, S., Chang, A.B., Murphy, T.L., Zeng, W., Vander Lugt, B., Khan, A.A., Ciofani, M., Spooner, C.J., Rutz, S., et al. (2012). A genomic regulatory element that directs assembly and function of immune-specific AP-1–IRF complexes. Science *338*, 975–980. https://doi.org/10.1126/science.1228309.

42. Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550. https://doi.org/10.1038/nature13385.

43. Montaño, C.M., Irizarry, R.A., Kaufmann, W.E., Talbot, K., Gur, R.E., Feinberg, A.P., and Taub, M.A. (2013). Measuring cell-type specific differential methylation in human brain tissue. Genome Biol. *14*, R94. https://doi.org/10.1186/gb-2013-14-8-r94.

44. Zheng, S.C., Breeze, C.E., Beck, S., and Teschendorff, A.E. (2018). Identification of differentially methylated cell types in epigenome-wide association studies. Nat. Methods *15*, 1059–1066. https://doi.org/10.1038/s41592-018-0213-x.

45. Stueve, T.R., Li, W.-Q., Shi, J., Marconett, C.N., Zhang, T., Yang, C., Mullen, D., Yan, C., Wheeler, W., Hua, X., et al. (2017). Epigenome-wide analysis of DNA methylation in lung tissue shows concordance with blood studies and identifies tobacco smoke-inducible enhancers. Hum. Mol. Genet. *26*, 3014–3027. https://doi.org/10.1093/hmg/ddx188.

46. Vaz, M., Hwang, S.Y., Kagiampakis, I., Phallen, J., Patil, A., O'Hagan, H.M., Murphy, L., Zahnow, C.A., Gabrielson, E., Velculescu, V.E., et al. (2017). Chronic cigarette smoke-induced epigenomic changes precede sensitization of bronchial epithelial cells to single-step transformation by KRAS mutations. Cancer Cell *32*, 360–376.e6. https://doi.org/10.1016/j.ccell.2017.08.006.

47. Hata, A., and Lieberman, J. (2015). Dysregulation of microRNA biogenesis and gene silencing in cancer. Sci. Signal. *8*, re3. https://doi.org/10.1126/scisignal.2005825.

48. Huang, H.-Y., Lin, Y.-C.-D., Cui, S., Huang, Y., Tang, Y., Xu, J., Bao, J., Li, Y., Wen, J., Zuo, H., et al. (2022). miRTarBase update 2022: an informative resource for experimentally validated miRNA-target interactions. Nucleic Acids Res. *50*, D222–D230. https://doi.org/10.1093/nar/gkab1079.

49. Center, B.I.T.G.D.A (2016). Analysis-ready Standardized TCGA Data from Broad GDAC Firehose 2016_01_28 Run (Broad Institute of MIT and Harvard).

50. Liu, L., Bi, N., Wu, L., Ding, X., Men, Y., Zhou, W., Li, L., Zhang, W., Shi, S., Song, Y., and Wang, L. (2017). MicroRNA-29c functions as a tumor suppressor by targeting VEGFA in lung adenocarcinoma. Mol. Cancer *16*, 50. https://doi.org/10.1186/s12943-017-0620-0.

51. Zhao, K., Cheng, J., Chen, B., Liu, Q., Xu, D., and Zhang, Y. (2017). Circulating microRNA-34 family low expression correlates with poor prognosis in patients with non-small cell lung cancer. J. Thorac. Dis. *9*, 3735–3746. https://doi.org/10.21037/jtd.2017.09.01.

52. Chen, Y., Min, L., Ren, C., Xu, X., Yang, J., Sun, X., Wang, T., Wang, F., Sun, C., and Zhang, X. (2017). miRNA-148a serves as a prognostic factor and suppresses migration and invasion through Wnt1 in non-small cell lung cancer. PLoS One *12*, e0171751. https://doi.org/10.1371/journal.pone.0171751.

53. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789. https://doi.org/10.1101/gr.132159.111.

54. Zheng, H., Brennan, K., Hernaez, M., and Gevaert, O. (2019). Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. GigaScience *8*, giz145. https://doi.org/10.1093/gigascience/giz145.

55. Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., Zhao, L., Li, X., Teng, X., Sun, X., et al. (2018). NONCODEV5: a comprehensive annotation database for long non-coding RNAs. Nucleic Acids Res. *46*, D308–D314. https://doi.org/10.1093/nar/gkx1107.

56. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527. https://doi.org/10.1038/nbt.3519.

57. Pimentel, H., Bray, N.L., Puente, S., Melsted, P., and Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. Nat. Methods *14*, 687–690. https://doi.org/10.1038/nmeth.4324.

58. Zhang, Y., Bu, D., Huo, P., Wang, Z., Rong, H., Li, Y., Liu, J., Ye, M., Wu, Y., Jiang, Z., et al. (2021). ncFANs v2.0: an integrative platform for functional annotation of non-coding RNAs. Nucleic Acids Res. *49*, W459–W468. https://doi.org/10.1093/nar/gkab435.

59. Roadmap Epigenomics Consortium; Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330. https://doi.org/10.1038/nature14248.

60. Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Res. *47*, D155–D162. https://doi.org/10.1093/nar/gky1141.

61. Wang, S., Talukder, A., Cha, M., Li, X., and Hu, H. (2021). Computational annotation of miRNA transcription start sites. Briefings Bioinf. *22*, 380–392. https://doi.org/10.1093/bib/bbz178.

62. Motameny, S., Wolters, S., Nürnberg, P., and Schumacher, B. (2010). Next generation sequencing of miRNAs–strategies, resources and methods. Genes *1*, 70–84.

63. Davis, S., and Meltzer, P.S. (2007). GEOquery: a bridge between the gene expression Omnibus (GEO) and BioConductor. Bioinformatics *23*, 1846–1847. https://doi.org/10.1093/bioinformatics/btm254.

64. Zhou, W., Laird, P.W., and Shen, H. (2017). Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. *45*, e22. https://doi.org/10.1093/nar/gkw967.

65. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

66. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127. https://doi.org/10.1093/biostatistics/kxj037.

67. Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191. https://doi.org/10.1038/nprot.2009.97.

68. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49. https://doi.org/10.1038/nature09906.

69. Grubert, F., Srivas, R., Spacek, D.V., Kasowski, M., Ruiz-Velasco, M., Sinnott-Armstrong, N., Greenside, P., Narasimha, A., Liu, Q., Geller, B., et al. (2020). Landscape of cohesin-mediated chromatin loops in the human genome. Nature *583*, 737–743. https://doi.org/10.1038/s41586-020-2151-x.

70. Aran, D., Sabato, S., and Hellman, A. (2013). DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biol. *14*, R21. https://doi.org/10.1186/gb-2013-14-3-r21.

71. Cho, J.-W., Shim, H.S., Lee, C.Y., Park, S.Y., Hong, M.H., Lee, I., and Kim, H.R. (2022). The importance of enhancer methylation for epigenetic regulation of tumorigenesis in squamous lung cancer. Exp. Mol. Med. *54*, 12–22. https://doi.org/10.1038/s12276-021-00718-4.

72. Sham, P.C., and Purcell, S.M. (2014). Statistical power and significance testing in large-scale genetic studies. Nat. Rev. Genet. *15*, 335–346. https://doi.org/10.1038/nrg3706.

73. Ramalingam, P., Palanichamy, J.K., Singh, A., Das, P., Bhagat, M., Kassab, M.A., Sinha, S., and Chattopadhyay, P. (2014). Biogenesis of intronic miRNAs located in clusters by independent transcription and alternative splicing. RNA *20*, 76–87. https://doi.org/10.1261/rna.041814.113.

74. Loader, C. (1999). Local Regression and Likelihood (Springer).

75. Dempster, A.P., Laird, N.M. and Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B (Methodological).

76. Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E.W., Wu, B., Doucet, D., Thomas, N.J., Wang, Y., Vollmer, E., et al. (2006). High-throughput DNA

methylation profiling using universal bead arrays. Genome Res. *16*, 383–393. https://doi.org/10.1101/gr.4410706.

77. Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. *37*, 773–782. https://doi.org/10.1038/s41587-019-0114-2.
78. Rahmani, E., Schweiger, R., Rhead, B., Criswell, L.A., Barcellos, L.F., Eskin, E., Rosset, S., Sankararaman, S., and Halperin, E. (2019). Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. Nat. Commun. *10*, 3417. https://doi.org/10.1038/s41467-019-11052-9.
79. Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J. (2014). Epigenome-wide association studies without the need for cell-type composition. Nat. Methods *11*, 309–311. https://doi.org/10.1038/nmeth.2815.
80. Rahmani, E., Zaitlen, N., Baran, Y., Eng, C., Hu, D., Galanter, J., Oh, S., Burchard, E.G., Eskin, E., Zou, J., and Halperin, E. (2016). Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. Nat. Methods *13*, 443–445. https://doi.org/10.1038/nmeth.3809.
81. Houseman, E.A., Kile, M.L., Christiani, D.C., Ince, T.A., Kelsey, K.T., and Marsit, C.J. (2016). Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinf. *17*, 259. https://doi.org/10.1186/s12859-016-1140-4.
82. Lutsik, P., Slawski, M., Gasparoni, G., Vedeneev, N., Hein, M., and Walter, J. (2017). MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. Genome Biol. *18*, 55. https://doi.org/10.1186/s13059-017-1182-6.
83. Teschendorff, A.E., Breeze, C.E., Zheng, S.C., and Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinf. *18*, 105. https://doi.org/10.1186/s12859-017-1511-5.
84. Rahmani, E., Schweiger, R., Shenhav, L., Wingert, T., Hofer, I., Gabel, E., Eskin, E., and Halperin, E. (2018). BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference. Genome Biol. *19*, 141. https://doi.org/10.1186/s13059-018-1513-2.
85. Team, B.C., and Maintainer, B.P. (2019). TxDb. Hsapiens. UCSC. Hg38. knownGene: Annotation Package for TxDb Object (S). R Package.
86. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol. *9*, e1003118. https://doi.org/10.1371/journal.pcbi.1003118.
87. Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. *46*, D252–D259. https://doi.org/10.1093/nar/gkx1106.
88. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation *2*, 100141. https://doi.org/10.1016/j.xinn.2021.100141.
89. Carlson, M., and Maintainer, B. (2015). Txdb. Hsapiens. Ucsc. Hg19. Knowngene: Annotation Package for Txdb Object (S). R Package Version 3.
90. Gel, B., and Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics *33*, 3088–3090. https://doi.org/10.1093/bioinformatics/btx346.
91. Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573. https://doi.org/10.1093/bioinformatics/btq170.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
DNA methylation and RNA-seq data of human CD4$^+$ T cells	Gene Expression Omnibus - NCBI	GSE114135
DNA methylation and RNA-seq data of human lung cancer	Cancer Genome Atlas Research Network	https://www.nature.com/articles/nature13385
DNA methylation data of human bronchial epithelial cells	Vaz et al. Cancer Cell, 2017	https://www.cell.com/cancer-cell/fulltext/S1535-6108(17)30349-5
Survival data of human lung adenocarcinoma	Broad GDAC Firehose	http://firebrowse.org/?cohort=LUAD
Chromatin state discovery data	Roadmap Epigenomics Consortium	https://www.nature.com/articles/nature14248
Software and Algorithms		
EpiMix R package	Bioconductor	https://bioconductor.org/packages/release/bioc/html/EpiMix.html https://doi.org/10.5281/zenodo.7987093
CibersortX algorithm	Newman et al. Nat Biotechnol, 2019	https://cibersortx.stanford.edu
Tensor Composition Analysis	CRAN repository	https://cran.r-project.org/web/packages/TCA
miRTarBase 9.0	Huang et al., Nucleic Acids Res, 2022	https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2022/php/index.php
ncFANs V2.0 server	Zhang et al. Nucleic Acids Res, 2021	http://ncfans.gene.ac

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Olivier Gevaert (ogevaert@stanford.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. The accession numbers for the datasets are listed in the key resources table.
- EpiMix is available as an R package on Bioconductor (https://bioconductor.org/packages/release/bioc/html/EpiMix.html). We also deposited the code into a public GitHub repository (https://github.com/gevaertlab/EpiMix). Lastly, we developed an R-shiny-based web application (https://epimix.stanford.edu) for users to interactively visualize and explore the results from this study,
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data downloading

The downloading module enables automated data downloading from the GEO database and TCGA project. Users can also supply their own datasets. To retrieve data from GEO, we utilized the GEOquery R package (version 2.62).[63] In this study, we downloaded DNAme data and gene expression data using GEO accession number GSE114135. The DNAme data were in the form of beta values ranging from 0 to 1, indicating the proportion of the methylated signal to the total signal. The gene expression data were TMM values.

To retrieve data from TCGA, we used the Broad Institute Firehose tool (Firehose).[49] We downloaded level three DNAme data (beta values) and gene expression data. The protein-coding genes were represented as log-transformed RSEM values, and the pri-miRNA expression data were represented as RPKM values.

Preprocessing

EpiMix requires DNAme data in beta values as input. If users prefer to start from *.idat* files unloaded from the measuring platforms, they can use other existing libraries[10,12] to convert the data into beta values. Gene expression data can be represented in any formats, such as RPKM, TPM and FPKM. EpiMix's contribution to preprocessing includes imputing missing values, removing single-nucleotide polymorphism (SNP) probes, aggregating nearby CpGs with similar DNAme values, and correcting batch effects (see sections below). With default settings, EpiMix removes all SNP probes that were explicitly labeled in the Illumina arrays (i.e., 'rs' probes). However, users have an option to remove a larger set of probes that overlap with any SNPs with global minor allele frequency (MAF) greater than 1%.[64] In addition, we removed CpGs and samples with more than 20% missing values and imputed missing values on the remaining dataset using the k-nearest neighbor (KNN) algorithm with K = 15. CpGs on sex chromosomes were also removed in the current study.

Batch effect normalization

Data from large patient cohorts were typically collected in technical batches. Systematic variances between technical batches may affect downstream data analysis and interpretation. EpiMix provides two options to correct for batch effects: (1) an anchor-based data integration approach adapted from the Seurat package (version 4.0.1)[65] and (2) an empirical Bayes regression approach, Combat.[66] The anchor-based approach identifies shared subpopulations, termed "anchors," across different datasets using canonical correlation analysis and mutual nearest neighbors, and then integrates the data with a non-linear transformation. To identify anchors, we used the "vst" method to select the top 10% variable feature, and effective batch effect removal was confirmed using the PCA-based ANOVA analysis. Alternatively, the batch effect can be corrected using the Combat algorithm.[58] We found that the anchor-based approach was more time-efficient, completing batch correction in 2 h for the lung cancer dataset, while Combat took over 48 h. The evaluation was conducted using parallel computation with 10 CPU cores, each of which was an Intel(R) Xeon(R) CPU E5-2697 v3 @ 2.60GHz processor.

CpG annotation and filtering

Regular mode

In the Regular mode, DNAme is modeled at *cis*-regulatory elements within or immediately surrounding protein-coding genes. We paired each CpG site to the nearest genes based on the hg38 manifest generated from Zhou et al.[64] Unique CpG-gene pairs were identified, where a CpG was either within the gene body or in the surrounding region. Additionally, users have the option to restrict the analysis to the promoters, defined as 2 kb upstream and 500 bp downstream (-2000bp $\sim$ +500bp) of the transcription start sites (TSSs). TSS information was retrieved using the *biomaRt* R package (version 2.50.1)[67] from Ensembl.

Enhancer mode

The Enhancer mode is designed to model DNAme specifically at distal enhancers. We first selected the distal CpGs that were at least 2 kb away from any known TSSs. However, users can customize this distance based on their needs. To select the CpGs located within enhancers, we used the enhancer database established by the ENCODE and ROADMAP consortia, in which enhancers of over a hundred human tissues and cell lines were identified using the chromatin-state discovery (ChromHMM).[59] We looked for the DNA elements associated with the chromatin states of active enhancers ("EnhA1" and "EnhA2") and genic enhancers ("EnhG1" and "EnhG2"). Since enhancers are cell-type specific, EpiMix allows users to select enhancers of specific cell types or tissue groups. In this study, we selected the enhancers of human blood and T cells, leading to the discovery of 40,311 CpGs of enhancers. For each CpG, we retrieved 20 nearby genes as candidate genes targets. This gene number was determined by the previous studies showing that many of the enhancers can regulate a gene within a 10-gene distance.[32,68,69] Genes that are positively regulated by the enhancers should have a negative relationship between DNAme and gene expression.[39,70,71] Therefore, we performed a one-tailed Wilcoxon rank-sum test on each enhancer-gene pair to select the enhancers whose methylation states were inversely associated with the gene expression. The raw p value from the Wilcoxon rank-sum test was adjusted using a permutation approach,[72] where an empirical p value was determined by ranking the raw p value in a set of permutation p values from testing the expression of a set of randomly selected 1,000 genes (Figure S2).

miRNA mode

MicroRNAs are commonly classified into "intergenic" or "intronic" based on their genomic locations. Intergenic miRNAs are found at previously unannotated human genome and are transcribed from their own unique promoters as independent entities. In contrast, intronic miRNAs are believed to share promoters with their host genes and co-transcribed from respective hosts. Recent evidence shows that some intronic miRNAs can also be transcribed independently from their host genes, suggesting they have their own independent promoters.[73] To select CpGs associated with miRNAs, we used a combined strategy. First, we obtained the most recent annotation of miRNAs from miRBase (version 22.1).[60] For each miRNA gene, we selected CpGs that were located within 5 kb upstream and 5 kb downstream. Second, we selected CpGs at miRNA promoters by using a recent database that integrates miRNA TSS information from 14 genome-wide studies across different human cell types and tissues.[61] We included CpGs located with

miRNA promoters defined as 2000 bp upstream and 1000 bp downstream of the TSSs. This combined feature selection strategy resulted in the discovery of 17,192 CpGs associated with 1,484 miRNAs in the HM450 array and 23,379 CpGs associated with 1,759 miRNAs in the EPIC array.

lncRNA mode

The mechanisms for transcriptional regulation of lncRNAs are similar to protein-coding genes. We first selected lncRNA-coding genes using the GENCODE annotation (Version 36). We then selected CpGs associated with each lncRNA based on the hg38 manifest generated from Zhou et al.[64] Unique CpG-gene pairs were identified, where a CpG was either located within the gene body or at the immediately surrounding area. This resulted in the discovery of 98,320 CpGs associated with 11,280 lncRNAs in the HM450 array and 184,816 CpGs associated with 15,392 lncRNAs in the EPIC array. Alternatively, users can select to focus the analysis at lncRNA promoters, defined as 2 kb upstream and 500 bp downstream (-2000bp $\sim$ +500bp) of the TSSs. The TSS information was retrieved from Ensembl using the *biomaRt* R package (version 2.50.1).[67]

CpG site aggregation and smoothing

Aggregation

To avoid computational expenses and overfitting of DNAme data when identifying patient subsets, we developed a feature that groups correlated CpG sites into clusters. This feature takes advantage of the strong correlation of DNAme values between adjacent CpGs. We used the average linkage hierarchical clustering algorithm to group CpGs within a single gene based on their DNAme values. and then we set a Pearson correlation threshold of 0.4 to define CpG clusters and single CpG sites that do not correlate with other sites. For each CpG site cluster, we computed the mean DNAme levels of its CpGs to represent its DNAme value, resulting in potentially multiple CpG site clusters representing a single gene. Users can then perform DNAme modeling at each CpG site cluster or single CpG sites.

Smoothing

Smoothing is a technique to reduce noise and enhance the statistical power in analyzing whole-genome bisulfite sequencing data.[6] It estimates the DNAme levels at local regions by incorporating the data from neighboring CpGs within a user-defined genomic window. EpiMix allows users to smooth the DNAme data using local likelihood smoothing.[74] However, since the number of CpGs is typically lower in array-based data than in bisulfite sequencing data, caution should be exercised when applying smoothing on array-based data.

Methylation modeling

After preprocessing, the methylation data are beta values bounded between 0 and 1, representing the proportion of the methylated signal to the total signal. When the study population is large, the beta values can be assumed to come from multiple underlying probability distributions, in our case, beta distributions. To model the DNAme, we fit a beta mixture model to the methylation values at each CpG site. Let y_i denote the beta value from subject i at a CpG site, where $i \in \{1,\ldots,n\}$, and n represents the total number of subjects. Let k denote the class membership of subject i, where $k \in \{1,\ldots,K\}$, and K represents the total number of components in the mixture. Assume subject i belongs to component k with probability η_k, we will have $\sum_{k=1}^{K}\eta_k = 1$. Subsequently, the likelihood contribution from subject i is:

$$f(Y_i = y_i) = \sum_{k=1}^{K} \eta_k \frac{y_i^{\alpha_k - 1}(1 - y_i)^{\beta_k - 1}}{B(\alpha_k, \beta_k)}$$

where $B(\alpha_k, \beta_k) = \int_0^1 t^{\alpha_k - 1}(1 - t)^{\beta_k - 1}\, dt$ is the beta function. Since the population contains n subjects, the log likelihood for the complete dataset is

$$l(\alpha, \beta, \eta) = \sum_{i=1}^{n} \log\{f(Y_i = y_i)\}$$

The goal of our modeling is to estimate the α, β, η parameters of each component that best fit the methylation values. Let $\theta = \{\alpha_1, \beta_1, \eta_1 \ldots, \alpha_k, \beta_k, \eta_k\}$ be a vector of parameters that define the shape of each component in the mixture. We used the expectation–maximization (EM) algorithm[75] to iteratively maximize the log likelihood and update the conditional probability that y_i comes from the *k th* component.

To determine the best number of components K, we used The Bayesian Information Criterion (BIC) for model selection and to avoid overfitting:

$$BIC = \log(n)(3K) - 2 \times \sum_{i=1}^{n} \log\{f(Y_i = y_i)\}$$

This process involves iteratively adding a new mixture component if the BIC improves. Each mixture component represents a subset of samples for whom a particular DNAme state is observed.

Identifications of differentially methylated CpGs

To determine whether a CpG site is hypo- or hyper-methylated, we can compare its methylation levels in the experimental group to its counterpart in the control group. We first performed beta mixture modeling to each CpG site using data from the experimental group to identify the mixture components. Next, we compare the methylation levels of each mixture component to the mean methylation levels of the control group. This methodology is based on the assumption that the DNAme profile is heterogeneous across patients in the experimental (i.e., disease) group but homogeneous in the control group. For instance, the DNAme profile can always be different across cancer patients with different driver mutations and subtypes, but in normal tissues the DNAme should be relatively homogeneous. In addition, the number of subjects in the experimental group is typically higher than the control group (e.g., TCGA projects). To determine the significant difference between the experimental and the control group, we used a Wilcoxon rank-sum to calculate the p-value, and multiple comparison was corrected with the false discovery rate (FDR). The Q-value threshold was set to 0.05. In addition, we required a minimum difference of 0.10 based on the platform sensitivity reported previously.[76]

Identifications of differential DNAme that was associated with transcription

When sample-matched gene expression data are provided, EpiMix can select CpGs whose methylation states were significantly associated with gene expression. In this study, we focused on the identification of DNAme that represses gene expression. However, users have the option to identify DNAme that is positively correlated with gene expression. For each CpG-gene pair, we used a one-tailed Wilcoxon rank-sum test to compare the mean levels of gene expression in patients showing an abnormal methylation state (hypo- or hyper-methylation state) to those with a normal methylation state. If a CpG was hypomethylated, we examined that the hypomethylated patients have higher gene expression levels compared to the normally methylated patients. Vice versa, if a CpG was hypermethylated, we tested that the hypermethylated patients have lower gene expression levels compared to the normally methylated patients. If a CpG was dual methylated (i.e., some samples were hypomethylated, while some others were hypermethylated), we tested that the hypomethylated patients have higher gene expression levels compared to the hypermethylated patients. Since a gene is typically paired with multiple CpGs, we adjusted the p-value using FDR to correct multiple comparisons. To select functionally significant CpG-gene pairs, we set the maximum threshold of the adjusted p-value to 0.01.

Simulation study

The goal of simulation study was to assess the sensitivity of EpiMix in detecting differential DNAme present in only specific subsets of the tested population. To achieve this goal, synthetic populations were created. First, CpGs exhibiting statistically similar DNAme levels that conformed to a unimodal beta distribution were selected from the activation and quiescent groups, respectively. Next, a subset of CpGs (n = 300) was randomly sampled from the quiescent group (n = 103 subjects) to serve as the baselines. The mean DNAme levels (beta values) of the CpGs in the baseline group ranged from 0.1 to 0.9, with a mean value of 0.6. To generate a synthetic population where the differential DNAme occurs only in specific subsets, we randomly selected a number of samples from the activation group and combined them with the baseline group, such that the final proportion (P) of samples from the activation group in the combined dataset ranged from 0.01 to 0.50, where $P \in \{0.01, 0.02, 0.05, 0.08, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$, and the mean difference in beta values ($\Delta beta$) between the two groups ranged from 0.1 to 0.7, where $\Delta beta \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 0.50, 0.60, 0.70\}$. This resulted in 2,700 synthetic CpGs. Finally, EpiMix was run on the synthetic population at each CpG, and its ability to detect differentially methylated signals was evaluated.

Benchmark with existing methods

We benchmarked the performance of EpiMix with other existing methods, including Minfi,[10] iEVORA[17] and RnBeads.[12,30]

Minfi includes a differential methylation step based on an F-test. We first transformed beta values to M values, and the differential methylation analysis was performed with the *dmpFinder* function. We set the significant p-value and Q-value thresholds to 0.05.

iEVORA is a two-step algorithm aimed at identifying infrequent alterations (outliers) in DNAme across a population.[16–18] It can be used to detect DNAm changes between normal cells at-risk versus those not-at-risk for neoplastic transformation. In the first step, the algorithm identifies differentially variable CpGs using a Bartlett's test. Once the differentially variable CpGs are identified, an optional second step can be added to rank them using differential methylation. The differential methylation was tested by t-statistics that compares the average levels of DNAme in the experimental group to the control group. In our study, we used the default parameters with a Q-value (FDR) threshold of 0.001 for testing differential variability and p-value threshold of 0.05 for testing differential methylation means. Our simulation studies showed that iEVORA was able to identify differentially variable CpGs even when the abnormal methylation was present in only a small sample subset. However, since the algorithm does not identify which subjects were abnormally methylated, and the average DNAme levels of the entire experimental group was compared to the control group, the differential methylation test could not generate statistically significant results.

RnBeads uses hierarchical linear models as implemented in the limma package to identify differential methylated CpGs. We set the differential methylation p-value threshold to 0.05.

Imputation of cell-type-specific DNAme and gene expression values

DNAme and gene expression are known to be cell-type specific. When the DNAme were measured at the tissue ("bulk") level, the differential DNAme profiles between patient subjects may result from the differences in tissue compositions.[43,44] From a clinical

perspective, tissue composition is meaningful in classifications of tumor subtypes and predictions of treatment response. However, from a biological perspective, users may be interested in identifying the differential DNAme present in only specific cell types. EpiMix focuses on the identification of differential DNAme across patient individuals. To resolve the confounding effect from tissue heterogeneity, we used previously validated algorithms to infer cell-type proportions and cell-type specific methylomes and transcriptomes (Figure S3). The first step was to estimate cell-type proportions in each sample using the CIBERSORTx algorithm.[77] This method leveraged the gene expression signatures established from experimentally purified cells from lung cancers and adjacent normal tissues, including epithelial cells, fibroblasts, hematopoietic cells and endothelial cells.[77] The estimated cell-type proportions were then used for sample-specific deconvolution of gene expression data, where the output was a three-dimensional tensor with shape *gene x cell type x sample* (Figure S3), indicating the gene expression levels in each cell type and in each sample. To deconvolute the DNAme data, we used Tensor Composition Analysis (TCA).[78] TCA requires knowledge of the estimated cell-type proportions for each sample. Since we restricted our analysis to samples with both DNAme data and gene expression data, we used the cell-type fractions estimated from CIBERSORTx in the previous step. The output from TCA was also a three-dimensional tensor with shape *methylome x cell type x sample*, indicating the DNAme levels in each cell type and in each sample. It is important to note that users can also leverage other existing tools[43,44,79–84] to adjust the effects from tissue compositions and then input the deconvoluted data to EpiMix.

Validation of deconvolution results

To validate the deconvolution results, we leveraged an independent dataset of human lung epithelial cells.[46] In this dataset, the DNAme data were collected from the HM450 array. We compared the DNAme profiles in *KRAS*-transformed lung epithelial cells (n = 4 replicates) to normal controls (n = 4 replicates). The hypomethylated CpGs were defined as those with a mean $\Delta\beta < -0.2$ and the hypermethylated CpGs were defined as those with a mean $\Delta\beta > 0.2$. For TCGA-LUAD data, we used the CpGs that were differentially methylated in at least 25% of the patients.

Genomic distribution of the differentially methylated CpGs

Genomic coordinates of the TSSs of the methylation-driven genes were retrieved from Ensembl using the *biomaRt* R package (version 2.50).[67] Exons and Introns of the protein-coding genes were retrieved from the TxDb object (*TxDb.Hsapiens.UCSC.hg38.knownGene*) (version 3.14).[85] The GenomicRanges R package (version 1.46)[86] was used to identify the differentially methylated CpGs located within promoters, exons and introns.

Motif enrichment analysis

TF binding motifs were retrieved from HOCOMOCO, a comprehensive database for TF binding sites.[87] HOMER (Hypergeometric Optimization of Motif EnRichment) was used to find motif occurrences in a ±250bp region around each differentially methylated regions (DMRs). We then combined all the DMRs to identify enriched motifs. Enrichments were quantified using Fisher's exact test and multiple comparisons were adjusted with the Benjamini-Hochberg procedure. To calculate the enrichment Odds Ratio, we used all the distal CpGs as the background probes and the functional CpGs of enhancers as the target probes. We set the significant *p* value cutoff to 0.05 and the smallest lower boundary of 95% confidence interval for Odds Ratio to 1.1. The enrichment analysis was performed using the *get.enriched.motif* function from the *ELMER* library (version 3.14) in R.[11]

Enrichment analysis of chromatin modifications

Enrichment analysis of histone modifications at the DMRs was performed using the Genomic Hyperbrowser GSUITE of tools.[31] A suite of tracks representing different chromatin features for human naive T cells (Epigenome ID: E038) and lung (Epigenome ID: E096) were retrieved from the ENCODE and ROADMAP consortiums.[59] To determine which tracks in the suite exhibit the strongest similarity by co-occurrence to the DMRs, the Forbes coefficient was used to obtain rankings of tracks, and Monte Carlo simulations were used to define a statistical assessment of the robustness of the rankings using randomization of genomic regions covered by the entire HM450 or EPIC array, and compute test statistics.

Functional enrichment analysis

Protein-coding genes

Gene sets were retrieved by curating the latest annotation databases of gene ontologies (GO) and KEGG pathways. Over-represented biological pathways in the methylation-driven genes were identified using hypergeometric testing.[88] We set the significant *p* value to 0.05 and Q value to 0.20. Highly similar GO terms were removed with a cutoff *p* value of 0.60 to retain the most representative terms. The enrichment analysis was implemented with the *clusterProfiler* R package (version 4.2.1),[88] and enrichment results can be visualized in both tabular and graphical formats.

miRNAs

MicroRNAs are known to mediate the destabilization and translational suppression of target messenger RNAs.[47] We queried experimentally validated miRNA targets from the miRTarBase.[48] Of the 92 methylation-drive miRNAs, we obtained 10,374 protein-coding genes targeted by 78 miRNAs. To further select miRNA targets, we compared the messenger RNA expression levels of each target between the patients with abnormal methylation states to those with a normal methylation state. If the miRNAs were

hypermethylated, we tested whether their target genes were upregulated. Conversely, when the miRNAs were hypomethylated, the target genes would be expected to be downregulated. We used a one-tailed Wilcoxon rank-sum test to compare the mean levels of target gene expression between patient groups. A significant *p* value was set to 0.05 and FDR-corrected *p* value was set to 0.2. KEGG pathway analysis was performed on the significant miRNA targets using hypergeometric testing.[88]

lncRNAs

To carry out functional annotation and pathway analysis of the differentially methylated lncRNAs, we used the ncFANs V2.0 server (http://ncfans.gene.ac/).[58] The genes in the significant CpG-gene pair matrix generated from EpiMix can be directly used as an input to ncFANs. NcFANs assigns the functions of protein-coding genes to lncRNAs based on pre-built co-expression networks in various normal tissues and cancers. We used the co-expression network built in the lung adenocarcinoma dataset from TCGA, and we set the correlation coefficient between lncRNAs and proteins-genes to 0.4 and the cutoff of the topological overlap measure similarity to 0.01.

Biomarker identification and survival analysis

Patient clinical data were retrieved from TCGA using the Firehose tool.[49] Alternatively, users can provide EpiMix with survival data if using their own datasets. We selected the CpGs with at least two methylation states. For each CpG, we fit a Cox proportional hazards regression model to assess the effect of methylation states on patient survival time. The log rank test was used to compare the survival curve and to calculate the significant p-value. $p < 0.05$ was considered as significant. The Kaplan-Meier survival plots were generated with the *survminer* R package (version 0.4.9).

Genome browser-style visualization

EpiMix enables genome browser-style visualization of the genomic coordinates and chromatin states of the differentially methylated genes and regions. We implemented two different forms of visualization. The gene-centric form shows the DM values of all the CpGs associated with a specific gene (e.g., Figure 3F). The CpG-centric form shows a differentially methylated CpG and its upstream and downstream genes (e.g., Figure 4E). Users can specify the number of nearby genes to display. Genes whose expression levels were significantly associated with the DNAme levels of the CpG are shown in red.

DNase I sensitivity and histone modification levels were retrieved from the ENCODE and ROADMAP consortiums.[59] By providing the Epigenome ID, users can retrieve data corresponding to the investigated tissue or cell type. In this study, we extracted the chromatin features for human naive T cells (Epigenome ID: E038) and fetal lung (Epigenome ID: E088). The genomic coordinates (X axis) were established on the hg19 genome built, and the enrichment signal (Y axis) represents negative log10 of the Poisson p-values. Human transcript annotation was retrieved from the TxDb object (*TxDb.Hsapiens.UCSC.hg19.knownGene*) (version 3.2.2).[89] The genomic coordinates of the adjacent genes of the differentially methylated CpGs were retrieved from Ensembl using the *biomaRt* R package (version 2.50.1).[67] The visualization was implemented with the *karyoploteR* package (version 1.20.0).[90]

Identifications of DNAme subtypes

DNAme subtypes can be discovered by applying consensus clustering to the DM-value matrix, where patients were clustered into robust and homogeneous groups (putative subtypes) based on their abnormal methylation profiles. Consensus clustering was performed with the ConsensusClusterPlus R package (version 1.58.0).[91] We used 1,000 rounds of k-means clustering and a maximum of K = 10 clusters. Selection of the best number of clusters was based on the visual inspection of ConsensusClusterPlus output plots.

QUANTIFICATION AND STATISTICAL ANALYSIS

All procedures involving statistical analysis were described in the "method details" section. The number of samples used in each experiment were described in both the figure legend and "Results" sections.