Online Method

Processing of human normal tissues

Ten human primary normal tissues were purchased from BioChain. Approximately 200 ng of genomic DNA from ten human primary tissues in the volume of 50 µL was fragmented into an average size of 400 bp in a Covaris micro TUBE with Covaris E210 ultrasonicator. Fragmented genomic DNA was converted into Illumina paired-end sequencing libraries using KAPA Library Preparation kit (KAPA Biosystems) following manufacturer's instruction with modifications. After end-repair and dA-tailing, ligation with methylated adapters was performed at 20 ˚C for 15 min in the presence of 10-fold molar excess of Illumina methylated adapters (Illumina). The ligation mixture was purified with an equal volume of Agencourt AMPure XP beads (Beckman Coulter) and eluted with 23 µL of 10mM Tris-HCl, pH8.5. Next, 20 µL of adaptor ligated DNA was bisulfite converted using EZ DNA Methylation-Lightning kit (Zymo Research) following manufacturer's protocol and eluted with 30 µL of 10mM Tris-HCl, pH8.5. Bisulfite converted DNAs were amplified using iQ SYBR Green Supermix (Bio-Rad) with 200 nM each of PCR primer PE1.0 and multiplexing PCR primer for 10 cycles in 100 µL total volume. PCR products were purified with 0.8X volume of Agencourt AMPure XP beads (Beckman Coulter) and eluted with 50 µL of 10mM Tris-HCl, pH8.5, pooled in equimolar ratios, and size selected using 6% TBE gels for 400-600 bp. The concentration of sequencing libraries was quantified by qPCR using KAPA Library Quantification kit (KAPA Biosystems). Libraries were sequenced on HiSeq2500 for PE 100 cycles.

Processing of patient tumor tissues.

Cancer tissue and plasma samples were collected from UCSD Moores Cancer Center. Clinical information, gender, age and TNM staging, on the patients was limited because the samples were de-identified. Informed consent was obtained from all subjects. All the samples are diagnosis to corresponding cancers according to the World Health Organization classification criteria[1]. 88.4% samples were derived from Caucasian population while 6.8% and 3.3% samples were from Asian and African population (detail see Supplementary Table 12). Genomic DNAs were extracted from 20-50 mg of primary tumor tissues from lung, colon and pancreatic cancer patients using DNeasy Blood and Tissue kit (QIAGEN) following the manufacturer's instruction and eluted in 400 µL of AE buffer (QIAGEN). The concentration and quality of genomic DNA were assessed by Qubit dsDNA HS Assay kit (Life Technologies) and NanoDrop (Thermo Scientific), respectively. To generate RRBS sequencing libraries, 100 ng of gDNA were digested with 20 U of *Msp*I (Thermoscientific) in 1X Tango buffer (Thermoscientific) and 1 ng of unmethylated lambda DNA (Promega) in order to assess for bisulfite conversion rate in 30µL total volume for 3 h at 37 ˚C and heat inactivated at 65 ˚C for 20 min. Next, 5U of Klenow fragment, exo- (Thermoscientifc) and a mixture of dATP, dGTP, and dCTP (New England Biolabs) were added to *Msp*I-digested DNAs for a final concentration of 1 mM, 0.1 mM, and 0.1 mM for dATP, dGTP, and dCTP, respectively in 32 µL for end-repair and dA-tailing. The mixture was mixed and incubated at 30 ˚C for 20 min, 37 ˚C for 20 min, and heat inactivated at 75 ˚C for 10 min. dA-tailed DNA was purified with 2X volume of Agencourt AMPure XP beads (Beckman Coulter) and resuspended dA-tailed DNA with 20 µL nuclease-free water without discarding the magnetic beads. dA-tailed DNAs were then ligated to methylated adaptors in 30 µL total volume containing 30 U of T4 DNA ligase, HC (Thermoscientific), 1X Ligation buffer (Thermoscientific), and 500 nM individual TruSeq multiplexing methylated adaptors (Illumina). The ligation mixture was mixed well and incubated at 16 ˚C for 20 h, heat inactivated at 65 ˚C for 20 min, purified by adding 60 µL of PEG 8000/5M NaCl buffer (Teknova) to adaptor ligated DNA and bead mixture, and eluted in 20 µL of nuclease-free water. Next, the adaptor ligated DNA were bisulfite converted using the MethylCode Bisulfite Conversion kit (Life Technologies) following manufacturer’s protocol and eluted in 35 µL of Elution buffer (Life Technologies). Bisulfite treated DNAs were amplified using 5 U of PfuTurboCX (Agilent Technologies) and 300 nM each of TruS_F and TruS_R primers for 14 cycles in 100 µL total volume. PCR products were purified with an equal volume of Agencourt AMPure XP beads (Beckman Coulter) and eluted with 50 µL of 10mM Tris-HCl, pH8.5, pooled in equimolar ratios, and size selected using 6% TBE gels for 150-400 bp. The concentration of sequencing libraries was quantified by qPCR using KAPA Library Quantification kit (KAPA Biosystems). Libraries were sequenced on Illumina HiSeq2500 for PE 100 cycles.

Processing of plasma samples

Normal plasma samples were obtained from UCSD Shirley Eye center. Information such as gender and age was limited because the samples were de-identified. Informed consent was obtained from all subjects. Plasma samples from patients were processed using the QIAamp Circulating Nucleic Acid Kit (Qiagen) to extract circulating DNA. The DNA extracted from plasma were then concentrated using ethanol precipitation and eluted in 15 uL nuclease-free water. Next, 1-10 ng of DNA were digested with 10 U of *Msp*I (Thermoscientific), 1X Tango buffer (Thermoscientific), and 10 pg of unmethylated lambda DNA (New England Biolabs) as control for ~13 h at 37 ˚C, then heat inactivated at 65 ˚C for 20 min. Next, 5 U of Klenow fragment, exo- (Thermoscientifc) and a mixture of dATP, dGTP, and dCTP (New England Biolabs) were added for a final concentration of 1 mM, 0.1 mM, and 0.1 mM for dATP, dGTP, and dCTP respectively. The mixture was gently vortexed, and incubated at 30 ˚C for 20 min, 37 ˚C for 20 min, and finally 75 ˚C for 10 min. To perform adaptor ligation, the dA-tailed DNA were added to a 5 uL mixture of 1X Tango buffer, 30 U of T4 DNA Ligase, HC (Thermoscientific), 2.5 mM ATP, and 500 nM individual TruSeq multiplexing methylated adaptors. The combined mixture was gently vortexed, incubated at 16 ˚C for ~20 h, then heat inactivated at 65 ˚C for 20 min. The ligation mixture was purified using Agencourt AMPure XP beads (Beckman Coulter), and eluted in 20 uL of nuclease-free water. The ligated products were then bisulfite converted using the MethylCode Bisulfite Conversion kit (Life Technologies). Two rounds of amplification were performed after bisulfite conversion. The first round was using PfuTurboCX (Agilent Technologies) for 12 cycles in 50 uL total volume, then the second round was performed using Phusion HotStart Flex (New England Biolabs) master mix for 9 cycles in 50 uL total volume. Final PCR products were purified, pooled in equimolar ratios, and size selected using polyacrylamide gels for 150-400 bp. Libraries were sequenced on both Illumina MiSeq and HiSeq2500 for PE 100 cycles.

Read mapping

WGBS and RRBS data were processed in similar fashions. We first trimmed all PE or SE fastq files using trim-galore version 0.3.3 to remove low quality bases and biased read positions. We used the option “`--stringency 5 --clip_R1 5 --clip_R2 5 -a GATCGGAAGAGCACACGTCTGAACTCCAGTCAC -a2 AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT`” for WGBS data and the option “`--stringency 5 --rrbs --non-directional -a`

GATCGGAAGAGCACACGTCTGAACTCCAGTCAC -a2 AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT″ for RRBS data. Next, the reads were encoded to map to a three-letter genome via conversion of all C to T or G to A if the read appears to be from the reverse complement strand. Then the reads were mapped using BWA mem version 0.7.5a, with the options "-B2 -c1000″ to both the Watson and Crick converted genomes. The alignments with mapping quality scores of less than 5 were discarded and only reads with a higher best mapping quality score in either Watson or Crick were kept. Finally, the encoded read sequences were replaced by the original read sequences in the final BAM files. Overlapping pair end reads were also clipped with bamUtils clipOverlap function.

Differentially methylated regions analyses

We developed a software package calls BsmoothHMM to identify differential methylated regions (DMRs) from whole genome bisulfite sequencing data. The workflow for the program is described as follows:

1. First, each WGBS methylation frequency data is pre-processed for local linear smoothing using the R package[2] **BSmooth**[3]
   The smoothing model requires the smoothing parameters **h** (the minimum smoothing window size) and **ns** (the minimum number of site per smoothing window). The parameters for smoothing is determined for each chromosome by a cross validation test using the first 1 million CpG sites along the chromosome with 10% of sites randomly selected as the validation set and remaining 90% as training set. First the **h** parameter is kept constant at 500 bp while values for the parameter **ns** is first tested in increment of 2 from 14 to 50, and the lowest value which generates the highest correlation of methylation level with the validation set is chosen. Next, the **ns** value is kept constant at the chosen value, and the **h** parameter is tested in increment of 100 from 500 to 2000. The lowest **h** value with the highest correlation of methylation level with the validation set is chosen. The entire chromosome is then smoothed using the model generated with the chosen parameters.
2. Next, a matrix is generated from the smoothed methylation values for each chromosome across all the WGBS data.
3. Each matrix is then evaluated for differential methylation, in this case, over-dispersion analysis performed on the matrix. The dispersion value is first formulated as the natural log of the squared coefficient of variation. The over-dispersion value is estimated by the Pearson residual from the expected dispersion for a given mean methylation level across samples using a generalized additive model (R package **mgcv**[4]). (Pearson's residual is defined as *(y-m)/V(m)^0.5*, where *y* is data *m* is model fitted value and *V* is model mean-variance relationship).
4. Segmentation is performed using a five states Hidden Markov Model (HMM). The model is initialized with five Gaussian emission distributions for each state, each with equal starting probabilities, and each with transition probabilities which disfavors state changes and which allows only stepwise state changes. The R package **hsmm**[5] performs expectation-maximization to find the model's parameters and performs a global decoding to determine the hidden state sequence using the Viterbi algorithm.

5. CpG sites with the same hidden states and within 500 bp of each other are merged to form DMR windows. Regions with less than 2 CpGs are discarded. An average methylation frequency is calculated across each DMR window for each WGBS data.
6. The average methylation frequency matrix is analyzed using the ROKU function from the R package **TCC**[6]**.** This function first normalizes the methylation frequency by subtracting the one-step Tukey biweight and by taking the absolute value. Then a normalized Shannon entropy value is calculated across the normalized vector per DMR region. High entropy means more uniformity across the samples while low entropy means one or few samples are differently methylated. ROKU also tests all combinations of 30% outlier candidates starting from no-outlier, one hypermethylated outlier, one hypomethylated outliers, x hypermethylated outlier, x hypomethylated outliers, and so on. The minimum Akaike's information criterion (MAIC) is used to pick the best model. The outliers determined by ROKU also determines whether a region is hypermethylating or hypomethylating. Only regions passing a maximum 0.85 normalized entropy cutoff are considered to be a DMR. We estimated 0.7% for hypomethylating DMRs and 5.4% hypermethylating DMRs false discovery rates from the regions in the lowest dispersion state (S1) passing this cutoff.

Methylation haplotype analyses.

1. We first partitioned the human genome into non-overlapping "sequencible and mappable" segments using a set of in-house generated WGBS data from 10 tissues from a 25-yr adult male individual (5x mappable genome coverage per tissue, 50x for 10 tissues combined). A total of 1,072,789 autosomal segments (minimal size: 80bp; average size: 2.35Kb; total size: 2.52Gb) that have a minimal read depth of 10x were identified.

```
bedtools genomecov -bg -split -ibam N37_10_tissue_pool_chrXX.bam >
N37_10_tissue_pooled.chrXX.genomecov.bed

awk '$4>9 {print $1"\t"$2"\t"$3}
N37_10_tissue_pooled.chrXX.genomecov.bed | bedtools merge -d 10 -i - >
N37_10_tissue_pooled.chrXX.RD10.genomecov.bed

awk '$3-$2>80 {print $1"\t"$2"\t"$3"\t"$3-$2+1}'
N37_10_tissue_pooled.chrXX.RD10.genomecov.bed >
N37_10_tissue_pooled.chrXX.RD10_80up.genomecov.bed
```

2. Mapped reads from WGBS data sets were converted into methylation haplotypes in each segment. Calculation of methylation linkage disequilibrium (the r^2 statistics) was performed on the combined methylation haplotypes from all the five data sets. A binary partitioning strategy was used to split each segment into methylation haplotype blocks (MHBs). We define a methylation haplotype block as a genomic region in which the r^2 value of two adjacent CpG sites is no less than a threshold (r^2 >= 0.5). At this threshold,

50% of the variance of a CpG methylation status can be predicted by the status of an adjacent site.

```
mergedBam2hapInfo.pl
N37_10_tissue_pooled.chrXX.RD10_80up.genomecov.bed
N37_10_tissue_pool_chrXX.bam >
N37_10_tissue_pool_chr1.RD10_80up_bin.hapInfo.txt

cat *.chrXX.RD10_80up_bin.hapInfo.txt | /mergeHapInfo.pl >
WGBS_pooled_mappable_bins.chrXX.hapInfo.txt

hapInfo2mld_block.pl WGBS_pooled_mappable_bins.chrXX.hapInfo.txt 0.5 >
WGBS_pooled_mappable_bins.mld_blocks_r2-0.5.bed
```

3. After MHBs were defined, methylation haplotypes for each MHB were extracted from the bam file, and the methylation haplotype load (MHL) for each MHB was calculated.

```
#Iterate through all the sample and chromosome combinations
mergedBam2hapInfo.pl WGBS_pooled_mappable_bins.mld_blocks_r2-0.5.bed
SampleID_chrXX.bam > SampleID_chrXX.hapInfo.txt

#Place all hapInfo.txt files for one data set in one folder, calculate
MHL and report the values for all samples at all MHBs in one matrix
get_methHapLoad_matrix.pl hapInfo_data_set_folder >
Data_set_name_mhl_matrix.txt

merge_mhl_matrix.pl Data_set_A_mhl_matrix.txt
Data_set_B_mhl_matrix.txt Data_set_C_mhl_matrix.txt >
All_data_sets_matrix.txt
```

Genome-wide methylation haplotype load matrix (MHL) and principle component analysis (PCA).

Methylation haplotype load was calculated as the formula for each BS-seq samples. The top quantile 15% MHL regions were selected in heatmap analysis to investigate the tissue relationship (**Figure 3**). The Euclidean distance and Ward.D aggregation were used in the heatmap plot (R, gplots package[7]). PCA (R package prcomp[2]) was conducted with default setting of the corresponding R packages[2] (**Supplementary Fig. 3**). Before the PCA analysis, raw data quantile normalization within same tissue/cell groups, standardization (scale) as well as the batch effect elimination (Combat algorithm[8]) were also applied to decrease the random noise. MAF and IMF were extracted from BAM files with customised PileOMeth (https://github.com/dpryan79/PileOMeth). Differential MHL analysis between cancer plasma and normal plasma were based on two-tailed Student's *t-test* or Wilcoxon rank sum test dependent on the normal distribution assumption or not while multiple test correction was conducted by false discovery rate (FDR) approach. Statistic variations were estimated among different groups and therefore one-way ANOVA analysis could be conducted.

Methylation high linkage regions estimated by RRBS and Meth450K

We collected RRBS data from ENCODE project (downloaded from UCSC Browser) and Methylation 450K microarray data from TCGA project. Pearson correlation coefficient were calculated between adjacent CpG sites across all samples. The Takai and Jones's sliding-window algorithm[9] was used to identify blocks of highly correlated methylation. (i) set a 100-base window in the beginning of genomic position and move the window to the downstream when there are least 2 probes in the window. Calculate the total probes in extended regions until the last window does not meet the criteria. The regions covering at least 4 probes were defined as CpG dense regions, and the average Pearson correlation coefficients among all the probes in cancer and normal samples were calculated respectively. Simulation analysis to investigate the relationship between LD at the single-read level and correlation coefficients of average 5mC between two CpG sites were performed based on random sampling of 10 different methylation haplotypes from each of the 1000 individuals.

Enrichment analysis of methylation haplotype blocks for known functional elements

Random sampling was performed in enrichment analysis as previous paper[10]. Genomic regions with same number (147,888), fragment length distribution and CpG ratios were sampling within sequencing accessible regions (genomic regions beyond CRG mappability blacklisted regions and non-cover regions in our WGBS dataset) by repeating 10,000 times. Statistical significance was estimated empirically based on empirical P-value. Fold changes (enrichment factors) were calculated as the ratios of observation over expectation. Exon, intron, 5-UTR, 3-UTR were collected UCSC database. Enhancer definition was based on the Andersson et al study[11], super enhancer was derived from Hnisz's study[12] and promoter regions were based on the definition by Thurman et al[13]. All the genomic coordinates were based on GRCh37/hg19.

The level of linkage disequilibrium (r^2) as a function of the distance between adjacent CpG sites.

The linkage disequilibrium (r^2) between two CpGs in the MHB regions were calculated and sampling 500,000 D'-distance. The distance between the adjacent CpG loci and the r^2 were recorded and selected to show the expected negative correlation between r^2 and distance of the CpGs. Density plot of the relationship were used to show the distribution of the correlation with the x-axis of distance of CpGs.

Definition of methylation haplotype, Methylation entropy and epi-polymorphism

We defined a methylated haplotype load (MHL) for each MHB, which is the normalized fraction of methylated haplotypes at different length:

$$\mathrm{MHL} = \frac{\sum_{i=1}^{l} w_i \times P(MH_i)}{\sum_{i=1}^{l} w_i}$$

$$w_i = i$$

Where l is the length of haplotypes, $P(MH_i)$is the fraction of fully methylated and un-methylated haplotype with i loci. For a haplotype of length L, we considered all the possible sub-strings with length from 1 to L in this calculation. w_i is the weight for i-locus haplotype. We typically used $w_i = i$ or $w_i = i^2$ to favor the contribution of longer haplotye. In this study, $w_i = i$ was used.

Following the concept of Shannon entropy $H(x)$, methylation entropy (ME) for haplotype variable in specific genome region were calculated with the following formula:

$$H(x) = -\sum_{i=1}^{l} P(x) \times \log_2 P(x)$$

$$ME = -\frac{1}{b}\sum_{i=1}^{n} P(H_i) \times \log_2 P(H_i)$$

$$P(H_i) = \frac{h_i}{N}$$

For a genome region with b CpG loci and n methylation haplotype, $P(H_i)$ represents the probability of observing methylation haplotype H_i, which can be calculated by dividing the number of reads carrying this haplotype by the total reads in this genomic region. ME is bounded between 0 and 1, and can be directly compared across different regions genome-wide and across multiple samples. Methylation entropy were widely used in the measurement of variability of DNA methylation in specific genome regions[14].

Epipolymorphism[15] was calculated as

$$\mathrm{ppoly} = 1 - \sum_{i=1}^{n} P_{\mathrm{i}}^{2}$$

where P_i is the frequency of epi-allele i the population (with 16 potential epialleles representing all possible methylation states of the set of four CpGs).

Highly methylated haplotype in cancer plasma and normal tissues

Highly methylated haplotype (HMH) was defined as the methylation haplotype which have at least 2 methylated CpGs in the haplotype. Cancer-specific highly methylated haplotypes (csHMH) were the ones only found in cancer plasma samples but absence in any of the normal plasma samples and normal tissues. For the analysis of matched tumor-plasma data from the same individuals, csHMHs were the HMHs present in both the cancer plasma and the matched primary cancer tissues, but absence in all normal samples. In the analysis of plasma samples with no matched primary tumor tissue, we identified csHMHs by subtracting HMHs found in cancer plasma with those present in all normal tissues and all normal plasma samples.

Simulation of MHL in plasma mixture and comparison between MHL and 5mC in the plasma mixture

In evaluating csHMHs as potential markers for non-invasive diagnosis (**Group II regions in Figure 4**), we hypothesized that cfDNA in plasma is a mixture of DNA fragments from cancer cells and white blood (WB) cells at different ratios (cancer DNA fragment from 0.1% to 50%). We created synthetic mixtures by random sampling of haplotypes in the Group II regions from cancer and WB data sets at different ratios, and repeated 1,000 times to empirically determined the mean and variance of MHL and 5mC levels at different fractions of cancer DNA (**Figure 4**). Once an empirical “standard curve” was constructed, we then used it to estimate the fraction cancer DNA in the plasma samples. In addition, we assessed the relationship between estimated cfDNA fraction and log-transformed normalized plasma cfDNA yield by linear regression. Signal-to-noise ratio to MHL and 5mC was conducted with the 1,000-time sampling procedures and then the average estimated tumor fraction as well as the variation (standard deviation) were recorded and the ratio was applied to measure the performance of the metric.

1. Gibbs, A.R. & Thunnissen, F.B. Histological typing of lung and pleural tumours: third edition. *J Clin Pathol* **54**, 498-9 (2001).
2. Team, R.C. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. (2016).
3. Hansen, K.D., Langmead, B. & Irizarry, R.A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* **13**, R83 (2012).
4. Wood, S.N. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC. (2006).
5. Bulla, J.B.a.I. hsmm: Hidden Semi Markov Models. R package version 0.4. https://CRAN.R-project.org/package=hsmm. (2013).
6. Kadota, K., Konishi, T. & Shimizu, K. Evaluation of two outlier-detection-based methods for detecting tissue-selective genes from microarray data. *Gene Regul Syst Bio* **1**, 9-15 (2007).
7. Gregory R. Warnes, B.B., Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz and Bill Venables. gplots: Various R Programming Tools for Plotting Data. R package version 3.0.1. https://CRAN.R-project.org/package=gplots. (2016).
8. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-27 (2007).
9. Takai, D. & Jones, P.A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* **99**, 3740-5 (2002).
10. Timmons, J.A., Szkop, K.J. & Gallagher, I.J. Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biol* **16**, 186 (2015).
11. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-61 (2014).
12. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).
13. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).
14. Xie, H. *et al.* Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Res* **39**, 4099-108 (2011).
15. Landan, G. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat Genet* **44**, 1207-14 (2012).