## Methods

#### **Normal and cancer samples**

Ten human primary tissues were purchased from BioChain Institute Inc. Cancer patient tissues and plasma samples were purchased from UCSD Moores Cancer Center and normal plasma samples were obtained from UCSD Shirley Eye center under IRB protocols approved by [UCSD Human Research Protections Program](https://irb.ucsd.edu/) (HRPP). All data sets generated in this study or obtained from public databases were listed in **Supplementary Table 13.**

#### **Generation of DNA libraries for sequencing**

Extracted genomic DNA were prepared for bisulfite sequencing using published protocols. For whole genome bisulfite (WGBS) and reduced representation bisulfite sequencing (RRBS), the DNA fragments were adapted to barcoded methylated adaptors (Illumina). For WGBS, the adapted DNA were converted using the EZ DNA Methylation Lightning kit (Zymo Research) and then amplified for 10 cycles using iQ SYBR Green Supermix (BioRad). For RRBS, the adapted DNA were converted using the MethylCode™ Bisulfite Conversion kit (Thermo Fisher Scientific) and amplified using the PfuTurboCx polymerase (Agilent) for 12-14 cycles. Libraries were pooled and size selected using 6% TBE polyacrylamide gels. Libraries were sequenced using the Illumina HiSeq platform for paired-end 100-111 cycles, the Illumina MiSeq platform for paired-end 75 cycles, and the GAIIx (WGBS only) for single-end 36 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 GATCGGAAGAGCACACG-TCTGAACTCCAGTCAC -a2 AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGT-GGTCGCCGTATCATT” for RRBS data. Next, the reads were encoded to map to a three-letter gen-ome 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.

#### **Methylation haplotype blocks (MHB)**

Human genome was separated into non-overlapping “sequenceable and mappable” segments using a set of in-house generated WGBS data from 10 tissues from a 25-year adult male individual. Mapped reads from WGBS data sets were converted into methylation haplotypes within each segment. Methylation linkage disequilibrium was calculated on the combined methylation haplotypes. We then partitioned each segment into methylation haplotype blocks (MHBs). MHBs were defined as the genomic region in which the r2 value of two adjacent CpG sites is no less than 0.5. MHB regions inferred by WGBS dataset was also validated by bulk data of methylation level. Takai and Jones's sliding-window algorithm[1](#_ENREF_1) was applied for methylation high linkage regions in HM450K (TCGA) and RRBS (Encode) dataset. Finally, simulation analysis to investigate the relationship between LD and correlation of average 5mC of two CpG loci were conducted based on random samplings of different methylation haplotypes with 1000 individuals and each individual sampling 10 methylation haplotype.

#### **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[1](#_ENREF_1) 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[2](#_ENREF_2). 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[3](#_ENREF_3), super enhancer was derived from Hnisz’s study[4](#_ENREF_4) and promoter regions were based on the definition by Thurman et al[5](#_ENREF_5). All the genomic coordinates were based on GRCh37/hg19.

#### **Methylation haplotype load (MHL)**

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

Where s the length of haplotypes, is the fraction of fully successive methylated CpGs with i loci. For a haplotype of length L, we considered all the sub-strings with length from 1 to L in this calculation. is the weight for i-locus haplotype. Options for weights are or to favor the contribution of longer haplotyes. In the present study, was applied.

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

For a genome region with CpG loci and methylation haplotype, represents the probability of observing methylation haplotype , 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[6](#_ENREF_6).

Epipolymorphism[7](#_ENREF_7" \o "Landan, 2012 #715) was calculated as

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

#### **Developmental germ layers and tissue specific MHB regions.**

In order to investigate the germ layer and tissue specific MHB regions, group specific index (see below) was applied. An empirical threshold 0.6 was selected to filter out layer and tissue specific MHB regions. Layer specific MHB regions were selected again to show the ability to distinguish different development layers. Tissue specific MHB regions were further used for tissue mapping and cancer diagnosis.

indicates the number of the groups. denotes the average of MHL of group. denotes the average of MHL of highest methylated group.

#### **Genome-wide methylation haplotype load matrix (MHL) analysis**

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. The Euclidean distance and Ward.D aggregation were used in the heatmap plot (R, gplots package[8](#_ENREF_8)). PCA (R package prcomp[9](#_ENREF_9" \o "Team, 2016 #669)) was conducted with default setting of the corresponding R packages[9](#_ENREF_9) (Supplementary Fig. 5). Before the PCA analysis, raw data quantile normalization within same tissue/cell groups, standardization (scale) as well as the batch effect elimination (Combat algorithm[10](#_ENREF_10)) 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.

#### **Simulation and real-data deconvolution analysis**

#### Deconvolution analysis was performed on simulated and non-simulated datasets. The deconvolution references were constructed by human normal primary tissues, whole blood (WB), colorectal cancer tissues (CCT) and lung cancer tissues (LCT). For the simulation analysis, methylation haplotypes from CCT and WB were mixed to generate a gradient of CCT factions ranging from 0.1% to 50% and then the expected and observed CCT factions were compared. Although our MHL is a non-linear metric, when mixing CCT and WB, we found the deconvolution result is perfect with logit transform, median root-mean-square-error < 5%, which is within the acceptable region of the deconvolution method[11](#_ENREF_11)when the contribution of colorectal fraction is less than 20%. Tissue specific MHB regions were selected features for deconvolution based on non-negative decomposition with quadratic programming[11-13](#_ENREF_11). Raw MHL signals logit transformed before deconvolution analysis.

#### **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, 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. 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.

#### **Diagnosis biomarker identification and tissue mapping algorithm for plasma DNA.**

The flow of analyses in current study is described in **Supplementary Figure 13**. Tumor specific methylation haplotype blocks (tsMHBs) were identified by a 2-tailed t-test with [false discovery rate](http://brainder.org/2011/09/05/fdr-corrected-fdr-adjusted-p-values/) (FDR) correction. Other statistical analyses to MHL were also conducted by 2-tailed t-test without explicit notification. CRC plasma and LC plasma distinguish prediction evaluation were applied random forecast therefore the test and validation sample were independent. Tumor-of-origin prediction was performed using a tsMHBs counting strategy in which the tissue-of-origin of the plasma were assigned to the reference group with the maximum number of ts MHB fragments (assignment by maximum likelihood). In further details, in the first stage, the tissue-specific MHBs were identified with WGBS and RRBS dataset from solid tissues in the training samples. Tissue specific MHB regions (each tissue have ~ 300 MHBs) were obtained by a moderate GSI> 0.1 filter. In the second stage, the predictions were validated with our own RRBS dataset which included 30 colorectal cancer plasma, 29 lung cancer plasma and 75 normal plasma samples. In the test dataset, we separated the samples into 5 parts so that 5-fold cross-validation could be applied to measure the stability of the prediction, and the number of tissue-specific MHB features were iterating from 50 to 300. The minimum number of features was selected when the accuracy for cancer plasma is higher than 0.8 and the accuracy for normal plasma is higher than 0.9 since we require high specificity in clinical applications. The selected number of features were used in the remaining samples to measure the accuracy of tissue-mapping. The variations of sensitivity, specificity, and accuracy in different subsets of 5-fold cross-variation were low (training dataset standard deviation<0.04 while testing dataset standard deviation<0.14, see Supplementary Table 12).

#### **Joint analysis of tumor and normal tissue for non-invasive cancer detection in plasma.**

Pathology markers (GSI scores derived from 8 CRC, 8 LC and 2 KC) and tissue-specific marker panels were integrated and considered as a “pan-cancer tissue”, and then together with the data sets from 10 normal tissues were applied for the tissue/reference-specific MHB identification. The top 200 MHBs specific to each of the 11 reference tissues were selected as the prediction features. The distribution for the reference specific MHBs in 75 normal plasma samples, 30 CRC plasma and 29 LC plasma samples were constructed for 11 references. The p-value of each reference in the plasma could be inferred by comparison with background distribution of the reference in normal plasma. Meanwhile, tissue-of-origin was assigned by maximum Z-scores among different references. With leave-one out cross-validation on normal plasma, the Type-1 error (FDR) for the corresponding Z-score threshold and sensitivity were estimated. Finally, setting a predefined Z-score threshold could be also used for tissue-of-origin assignment, meanwhile, ROC curve was built to show the performance of the predictors.

1. 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).

2. 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).

3. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-61 (2014).

4. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).

5. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).

6. Xie, H. *et al.* Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Res* **39**, 4099-108 (2011).

7. 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).

8. 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).

9. Team, R.C. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://[www.R-project.org/](http://www.R-project.org/). (2016).

10. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-27 (2007).

11. Houseman, E.A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).

12. Gong, T. & Szustakowski, J.D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29**, 1083-5 (2013).

13. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci U S A* **112**, E5503-12 (2015).