**Online Method**

**Experiments (Noi/Dinh).**

**Question:** Do we need to provide the sources of samples processed in our lab in this section?

*Processing of fetal tissues (N37).*

(Note: We received genomic DNA of N37 sample from Jin Billy Li’s lab, so I did not include DNA extraction step here)

Approximately 200 ng of genomic DNA of human primary tissues (N37) in the volume of 50 µL was sheared into an average size of 400 bp in a Covaris micro TUBE with Covaris E210 ultrasonicator. Fragmented genomic DNA was constructed 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 (Zymoresearch) 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 both GAIIx for SE 36 cycles and HiSeq Run for PE 100 cycles.

*Processing of patient tumor tissues.*

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 resuspend 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 HiSeq RapidRun for PE 100 cycles.

*Processing of patient plasma*

Plasma 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, 1-1.5 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 MiSeq and HiSeq RapidRun for PE 100 cycles.

**Read mapping (Dinh).**

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 (Dinh)**

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 **BSmooth**[1](#_ENREF_1)

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.

1. Next, a matrix is generated from the smoothed methylation values for each chromosome across all the WGBS data.
2. 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 ,** Wood, S.N. (2006) Generalized Additive Models: An Introduction with R. CRC). (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).
3. 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** (<http://www.sciencedirect.com/science/article/pii/S016794730800426X>) performs expectation-maximization to find the model’s parameters and performs a global decoding to determine the hidden state sequence using the Viterbi algorithm.
4. 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.
5. The average methylation frequency matrix is analyzed using the ROKU function from the R package **TCC**[**2**](#_ENREF_2)**.** 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 (Kun).**

1. We first divided 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

1. Mapped reads from five WGBS data sets were converted into methylation haplotypes in each segment. Calculation of methylation linkage disequilibrium (the r2 statistics) was performed on the combined methylation haplotypes from all the five data sets. A greedy algorithm was used to partition each segment into methylation haplotype blocks (MHBs). We define a methylation haplotype block as a genomic region in which the r2 value of two adjacent CpG sites is no less than a threshold (r2 >= 0.5). At this threshold, 50% of the variance of a CpG site can be predicted by the methylation 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

1. 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

**Statistical analysis of MHB and MHL**

**1, Genome-wide methylation haplotype load matrix integration**

Methylation haplotype load was calculated as the formula for each BS-seq samples, however, these datasets were derived from different laboratory with different study design. The data were merge and then quantile normalization, standardization (scale) as well as the batch effect elimination[3](#_ENREF_3) were conducted to decrease the random noise. The top quantile 15% MHL regions were selected in heatmap analysis to investigate the tissue relationship. The Euclidean distance and Ward.D aggregation were applied in the heatmap plot (R, gplots package).

**2, Methylation high linkage regions estimated by RRBS and Meth450K**

We hypothesis that the MHB regions inferred by GWBS dataset should be highly occurred or over-represented in bulk data of methylation level. Therefore, we collected RRBS data from Encode project (Download from UCSC Browser) and Methylation 450K microarray data TCGA project. The traditional Pearson correlation coefficient were calculated between adjacent CpG site. An intuitive method adopted from Takai and Jones's sliding-window algorithm were applied to identify the Methylation high linkage regions within methylation 450K microarray and RRBS dataset as the following steps. Set a 100-base window in the beginning of genomic position of the array and shift the window to the right when there are least 2 probes in the windows. Calculate the total probes in extent regions until the last window does not meet the criteria. All the regions in which at least 4 probes were collected and be defined as high CpG density regions and calculate the average Pearson correlation among all the probes in cancer and normal samples, respectively. Simulation analysis to investigate the relationship between LD and correlation of average 5mC of two CpG loci were conducted based on random sampling different methylation haplotype with 1000 individual and each individual sampling 10 methylation haplotype (**Supp. Figure**).

**3, Enrichment analysis for methylation haplotype blocks in known functional elements**

Random sampling was applied in enrichment analysis as previous paper. Same number and same length distributed genomic regions 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. Empirical significance was estimated with native definition of P-value. Fold change (enrichment coefficient) were estimated with the ratio of observation to expectation. The average percentage of the overlap regions to the known functional elements and MHB regions could be used to indicate the dependence degree between MHB and known functional element. exon, intron, 5-UTR, 3-UTR were collected UCSC databse. Enhancer regions were collected from the study of Andersson [4](#_ENREF_4), Promoter regions were collected from the study of Thurman[5](#_ENREF_5). All the genomic coordinates were based on GRCh37/hg19.

**4, Relationship between linkage disequilibrium (D’) and distance of all CpGs in a MHB regions.**

The linkage disequilibrium (D’) between two CpGs in the MHB regions were calculated and sampling 500,000 D’-distance. The distance between the adjacent CpG loci and the D’-distance were recorded and selected to show the expected negative correlation between R2 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.

**5, Definition of methylation haplotype, Methylation entropy and epi-polymorphism**

We define 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 methylated and un-methylated haplotype 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. We typically used or to favor the contribution of longer haplotye. In the present study, was applied.

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

Suppose, there is a genome region with CpG loci and methylation haplotype, then represent the probability for methylation haplotype and it can be calculated with the number of reads for such haplotype and the total reads in this genomic region. Since the proposed methylation entropy has been normalized by the number of CGs, the value of ME ranges from 0 to 1 and therefore can be applied in the genome-wide scales. Methylation entropy were widely used in the measurement of variability of DNA methylation in specific genome regions [6](#_ENREF_6).

Epipolymorphism 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).

**6, High Methylation haplotype in cancer plasma and normal tissues**

High methylation haplotype (HMH) was defined as the methylation haplotype which have at least 2 methylated Cs in the haplotype. Meanwhile, cancer-specific high-methylation haplotype (csHMH) was denoted as the high-methylation haplotype only occurred in cancer plasmas compared with normal plasma, normal tissues. In the present study, we have different comparisons for csHMH. In the matched tumor-plasma context, csHMHs were HMH occurred in cancer plasma and matched primary tissues while no occurring in all normal tissues and normal plasmas. In the non-matched tumor plasma vs normal plasma context, csHMHs were HMH occurred in cancer plasma while no occurring in all normal tissues and all normal plasmas. In the later analysis, matched tissue and plasma samples were excluded to guarantee the analysis could be independent. The contribution of the cancer plasma HMH were decomposed by predominant category and then remove all the HMH from the predominant category to infer the contribution from the such predominant category.

**7, Deconvolution of cancer and normal plasma**

Methylation haplotype were mixture by different tissues with specific gradients and then deconvolution was conducted to compare the result. . Tissue specific MHL were applied to be the candidate features for deconvolution based on non-negative decomposition with quadratic programming. Raw MHL signals were applied of logit transform before deconvolution analysis. Virtual samples of same counts were constructed by random sampling and averaging and then contribution of WB, primary tumor tissue, normal origin-of-tissue were estimated. The detail about non-negative decomposition with quadratic programming could be found in the study of Gong[7](#_ENREF_7) and Sun[8](#_ENREF_8).

8, **Developmental germ layers and tissue specific MHB regions.**

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

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

Cancer specific hyper-methylated MHBs measure by MHL could be considered as promising non-invasive biomarkers (Figure 4A, Group II). We could hypothesize that the cancer plasma was composed by these fragment and whole blood (WB) DNA fragments with different proportion (0.1% to 50%). These Group II regions were repeat sampling 1,000 times to investigate the MHL trends along with the concentration of cancer fragment in plasma and compare the difference between MHL and 5mC levels (Figure 4). After the standard curve was built, we could also estimate cancer cfDNA fraction with the average MHL of the plasma samples. Since we have the information of the cfDNA yield for each sample, then we could assess the relationship between estimated cfDNA fraction and normalized plasma cfDNA yield.

MHL in plasma could also be simulated with mixture of WB DNA fragment and cancer DNA fragment. Since we focus on hyper-methylation biomarker for cancer diagnosis, we hypothesis that the cancer plasma is composed with 99% hypo-methylated fragment from normal WB or other normal tissues while 1% fragments are from hyper-methylated cancer cells. In addition, in the NGS sequencing, usually, we will obtain some non-completed sequencing fragment and there also will be some stochastic methylated CpG site in non-methylated normal cell fragment. With above assumption, we simulated the situation with 90 completely 4-CpG non-methylation fragments and 9 partial sequenced and stochastic methylated fragment and 1 copy continuous 4-CpG methylated fragment. MHL and 5mC level were calculated as the definition of MHL. We can find MHL could be significantly increase the methylation signal compared with average 5mC level (Supp. Figure 6).

**10, Diagnosis biomarker identification, prediction model and tissue mapping algorithm for plasma cancer DNA.**

Tumor specific methylation haplotype blocks based on were identified by 2-tailed t-test with [False Discovery Rate](http://brainder.org/2011/09/05/fdr-corrected-fdr-adjusted-p-values/) correction. Other statistical analysis to MHL were also conducted by 2-tailed t-test without explicitly notification. The genomic regions and the samples who have more than 30% missing rations were filter out in the further analysis (6 of 26 normal plasma samples were filter out). Tumor-of-origin prediction were applied with random forest model. In the first stage, the prediction model was training with WGBS and RRBS dataset from 10 solid tissues, the sample size for each tissue was prepared as the balanced number to increase the prediction accuracy and stabilization of the prediction model. 1090 tissue specific MHL regions were obtained by filtered with the moderate GSI> 0.3 so that we could select the most powerful biomarkers which can be detected in RRBS and GWBS and the most important quantile 10% variables were selected to be the input variable for the validation stage. In the second stage, the built prediction model was validated with our own RRBS dataset which including 30 colon cancer plasma, 29 lung cancer plasma and 75 normal plasma samples. Finally, 52 biomarkers (predictors) were selected with the highest variable-importance in random forest model and the corresponding average prediction sensitivity, specificity or accuracy were reported for the training dataset prediction. Random Forest were repeat 500 times to guarantee the prediction result were robust.

1. Hansen, K.D., Langmead, B. & Irizarry, R.A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* **13**, R83 (2012).

2. 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).

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

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

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. 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).

8. 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).