**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 fetal 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.

**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 dataset 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([1](#_ENREF_1)) 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 and Methylation 450K microarray data TCGA project as well as public database GEO. 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.

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

Random sampling were 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 from . Enhancer regions were collected from the study of Andersson ([2](#_ENREF_2)), Promoter regions were collected from the study of Thurman ([3](#_ENREF_3)). 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 ([4](#_ENREF_4)).

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, **Development layer 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.

**6, 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. Tumor-of-origin prediction were applied with random forest model. In the first stage, the prediction model were 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. 2111 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 5% variables were selected to be the input variable for the validation stage. In the second stage, the built prediction model were validated with our own RRBS dataset which including 30 colon cancer plasma, 29 lung cancer plasma, 10 pancreatic cancer plasma and 20 normal plasma samples.

1. Johnson, W.E., Li, C. and Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118-127.

2. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455-461.

3. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75-82.

4. Xie, H., Wang, M., de Andrade, A., Bonaldo Mde, F., Galat, V., Arndt, K., Rajaram, V., Goldman, S., Tomita, T. and Soares, M.B. (2011) Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic acids research*, **39**, 4099-4108.