# Deconvolution of epigenetic heterogeneity by tightly coupled CpG methylation.

1Department of Bioengineering, Institute for Genomic Medicine and Institute of Engineering in Medicine, University of California at San Diego, La Jolla, California, USA.

Figure counts: 5

Abstract counts: 250

Main body counts: 3586

Running title: **Methylation haplotypes in human tissues and cell-free DNA**

Keywords: Methylation haplotype, Circulating cell-free DNA, Tumor-of-origin mapping

Corresponding authors:

Dr. Kun Zhang, Institute for Genomic Medicine and Institute of Engineering in Medicine, Department of Bioengineering, University of California at San Diego, 9500 Gilman Drive, MC0412, La Jolla, CA 92093-0412, Tel: 858-822-7876, Fax: 858-534-5722, E-mail: kzhang@eng.ucsd.edu

## Abstract (NG:100, GR:250)

Adjacent CpG sites in mammalian genomes tend to be co-methylated due to the processivity of enzymes responsible for adding or removing the methyl group. Yet discordant methylation patterns have also been observed, and found to be related to stochastic or uncoordinated molecular processes. Here we focused on a systematic search and investigation of regions in the full human genome that exhibit highly coordinated methylation. By examining the co-methylation patterns of multiple adjacent CpG sites, termed methylation haplotypes, in single bisulfite sequencing reads, we defined blocks of tightly coupled CpG sites, called Methylation Haplotype Blocks (MHBs), with 61 sets of whole genome bisulfite sequencing (WGBS) data. Subsets of these MHBs were further validated with 101 sets of RRBS ENCODE data, and 637 sets of Illumina 450k methylation array data. Globally, MHBs are enriched in but only partially overlap with several well-known genomic features, including CpG islands, promoters, enhancers and VMRs. We uncovered a strong enrichment of MHBs in hypermethylated over hypomethylated DMRs, suggesting a processive spreading mechanism in establishing methylation and a less processive and targeted mechanism for demethylation.

To perform quantitative analysis of the MHBs, we defined a metric called Methylation Haplotype Load (MHL), which could cover both average methylation level and methylation complexity and therefore more informative than average methylation level or Shannon entropy. Integrated with feature selection strategy, we demonstrate MHL of tissue-specific MHBs could be useful in developmental germ-layers and tumor-of-origin prediction.

Interestingly, examination of these MHBs revealed two distinct mechanisms for fate commitment during development: epigenetic silencing of pluripotent genes, such as NANOG, for mesoderm induction; and epigenetic induction (or de-suppression) of lineage-specific factors for ectoderm commitment. Finally, MHL of tissue specific MHB regions were demonstrated to be powerful in tumor-of-origin prediction.

## Introduction

CpG methylation in mammalian genome is a relatively stable epigenetic modification, which can be either transmitted across cell division[1](#_ENREF_1" \o "Wigler, 1981 #1) through DNMT1, or dynamically established or removed by DNMT3 A/B and TET proteins. Due to the processivity of these enzymes, physically adjacent CpG sites on the same DNA molecules tends to share similar methylation status, although discordant CpG methylation has also been observed, especially in cancer cells. The theoretical framework of linkage disequilibrium, which was developed to model the co-segregration of adjacent genetic variants on human chromosomes among human populations, can be applied to the analysis of CpG co-methylation in human cells. A number of studies related to the concepts of methylation haplotypes, epi-alleles, or epi-haplotypes have been reported, albeit at small numbers of genomic regions or limited numbers of cell/tissue types. Recent data production efforts, especially by large consortia such as NIH RoadMap Epigenomics project and EU Blueprint Epigenome project have produced a large number of whole-genome, base-resolution bisulfite sequencing data sets for many tissue and cell types. These public data sets, in combination with additional internally generated data, allowed us to perform full genome characterization of local CpG co-methylation across the largest set of human tissue types available to date, and annotate these blocks of co-methylated CpGs as a distinct set of genomic features.

DNA methylation is cell-type specific, and the pattern can be harnessed for deconvoluting the relative cell composition of mixed samples, such as different white blood cells in whole blood, fetal components in maternal cell-free DNA, or circulating tumor DNA in plasma. Most of these recent efforts relies on the methylation level of individual CpG sites, and are fundamentally limited by the technical noise and sensitivity in measuring single CpG methylation. Very recently, Lehmann-Werman et al demonstrated a superior sensitivity with multi-CpG haplotypes in detecting tissue-specific signatures in circulating DNA[2](#_ENREF_2" \o "Lehmann-Werman, 2016 #1017). The markers in that study were discovered from Infinium 450k methylation array data, which represent only a very limited fraction of the human genome. Here we performed an exhaustive search of tissue-specific methylation haplotype blocks in the full genome, and proposed a block-level metric, termed methylated haplotype load (MHL), for a systematic discovery of informative markers. Applying our analytic framework and markers identified, we demonstrated accurate determination of cancer tissue of origin as well as estimation of tumor load in clinical plasma samples from patients of lung and colon cancers (Figure 1A).

## Results

#### **Identification and characterization of methylation haplotype blocks.**

To investigate the co-methylation status of adjacent CpG sites along single DNA molecules, we extended an approach that we previously established[3](#_ENREF_3" \o "Shoemaker, 2010 #633), in which we applied the concept of genetic linkage disequilibrium and the r2 metric to quantify the degree of co-methylation among different DNA molecules. CpG methylation status of multiple CpG sites in single- or paired-end Illumina sequencing reads were extracted to form methylation haplotypes, and pairwise “linkage disequilibrium” of CpG methylation r2 was calculated from the distribution of methylation haplotypes (see Methods). We then partitioned the genome into blocks of tightly co-methylated CpG sites, which we called Methylation Haplotype Blocks (MHBs, Figure 1B), using a r2 cutoff of 0.5. Similar to the partitioning of genetic haplotype blocks, slightly different cutoff values, such as 0.3 or 0.7, resulted in only minor quantitative differences in the block size and number without affecting the global pattern (data not shown).

To characterize the global pattern and distribution of MHBs, we started with 51 sets of published Whole Genome Bisulfite Sequencing (WGBS) data from human primary tissues, as well as the H1 human embryonic stem cells and *in vitro* derived progenitors. We also included an in-house generated WGBS data set from 10 human adult tissues. Across this set of 61 data (>2000x combined genome coverage) we identified ~ 55 billion methylation haplotype informative reads that cover 58.2% of autosomal CpGs. We identified a total of 147,888 MHBs at the average size of 95bp and minimum **3** CpGs per block, which represents ~0.5% of the human genome (**Supplementary Table 1**). The uncovered regions have low CpG density and hence too few CpG sites within Illumina read pairs for determining informative haplotypes. The majority of CpG sites within the same MHBs are near perfectly coupled (r2 ~1.0) regardless of the sample types. We found that methylation LD extends further along the DNA in stem cells and progenitors, compared with normal adult tissue, both in the fraction of tightly coupled CpG pairs (94.8% versus 91.2%), and the over-representation of partially coupled CpG pairs that are over 100 bp apart (**Figure 1c**). This is consistent to our previous observations on a smaller BSPP data set on 2,020 CpG islands[3](#_ENREF_3" \o "Shoemaker, 2010 #633) and another previous report[4](#_ENREF_4" \o "Shao, 2014 #634). Interestingly, in primary tumor tissues, we observed a reduction of perfectly coupled CpG pairs, which could be related to the pattern of discordant methylation recently reported in VMR[5](#_ENREF_5" \o "Landau, 2014 #631).

While WGBS data allowed us to unbiasedly identify MHBs across the entire genome, the 61 sets of data did not represent the full diversity of human cell/tissue types. To validate the MHBs in a wider range of human tissues and culture cells, we examined 101 published reduced representation bisulfite sequencing (RRBS) data from ENCODE cell lines and tissue samples, as well as 637 sets of Illumina 450k methylation array data from 11 human normal tissues. The ENCODE RRBS data sets were generated with short (36bp) Illumina sequencing reads, greatly limiting the length of methylation haplotypes that can be called. Similarly, Illumina methylation arrays only report average CpG methylation of all DNA molecules in a sample, preventing a methylation linkage disequilibrium analysis. Therefore, we calculated the pairwise correlation coefficient of adjacent CpG methylation levels across different data sets for block partitioning. Note that the presence of such correlated methylation blocks is a necessary but not sufficient condition for MHBs (**Supplementary Fig. 1**). Nonetheless, the absence of correlated methylation blocks in these data would indicate potential false discovery of MHBs. We identified 23,517 and 2,212 correlated methylation blocks from ENCODE RRBS and HM450 array data respectively, among which 8,920 and 1,258 have significant overlaps WGBS-defined MHBs. At the same time, as our expectation, we observed significant higher correlation among the CpGs within the MHB regions compared CpG loci beyond of MHBs in HM450K and RRBS dataset indicating the MHBs we identified is highly credible (**Supplementary Fig. 1**). Taken together, the MHBs that we identified represent a distinct class of genomic feature where local CpG methylation is established or removed in a highly coordinated manner at the level of single DNA molecules, presumably due to the processive activities of the related enzymes and the local density of CpG dinucleotides.

#### **Role of methylation haplotype blocks in formation of specifically methylated regions**

multiple**d** (p-value<10-6)[6](#_ENREF_6" \o "Guelen, 2008 #699)[7](#_ENREF_7" \o "Wen, 2009 #700)[8](#_ENREF_8" \o "Dixon, 2012 #722)[9](#_ENREF_9" \o "Hansen, 2011 #9)**e** **e**.

Previous studies on mouse and human [10](#_ENREF_10" \o "Irizarry, 2009 #721),[11](#_ENREF_11" \o "Ziller, 2013 #1016) demonstrated that dynamically methylated regions were associated with regulatory regions such as enhancer-like regions marked by H3K27ac and transcription factor binding sites. In the human study, 21.8% of autosomal CpGs were found to be differentially methylate d across 30 human cell and tissue types. These CpGs were enriched at low to intermediate CpG density promoters. We have found that genes promoters tend to overlap with methylation haplotype blocks and that many blocks also have strong H3K4me3 marks as well as H3K27ac marks (Supp. Figure 1). These findings suggest that the differentially methylated regulatory regions can form through a spreading mechanism by processive enzymes.

We combined ideas from previous approaches to systematically assess methylation specificity across the genome. Briefly, segmentation was performed on the genome using a five states Hidden Markov Model. We discovered distinct segments of the genome that can represent varying levels of methylation specificity [Sug Fig XXa]. MHBs contain overlapping CpGs with 36% of the most specific regions and with 10% of the least specific regions. Corroborating with previous works, high specificity regions are often hypo-methylated in the outlying samples [Sug Fig XXb]. High specificity regions also have transcription factor binding site frequencies different from other segmented regions (Pearson’s correlation = 0.69, -0.82, -0.80, and -0.48) but they were more similar to the transcription factor binding site frequencies in MHBs (Pearson’s correlation = 0.88). The cellular machinery for forming methylation blocks could play an important role in maintaining a tightly regulated hypermethylated state in most cell types and could be turned off when a cell needs to differentiate itself. Interestingly, majority of the high specificity regions do not form methylation haplotype blocks, probably due to stochastic rather than coordinated loss of methylation.

#### **Characterizations of human normal tissues and stem cell lines with methylation haplotype load**

To enable quantitative analysis of the methylation patterns within individual MHBs across many samples, we need a single metric to define the methylated pattern of multiple CpG sites within each block. Ideally this metric is not only a function of average methylation level for all the CpG sites in the block, but also can capture the pattern of co-methylation on single DNA molecules. For this purpose, we defined Methylation Haplotype Load (MHL), which is a weighted mean of the fraction of fully methylated haplotypes and substrings at different length (i.e. all possible substrings). Compared with other metrics used in the literature (methylation level, methylation entropy, epi-polymorphism and haplotypes counts), MHL is capable of distinguishing blocks that have the same average methylation but various degrees of coordinated methylation (**Figure 2**). In addition, MHL is bounded between 0 and 1, which allows for direct comparison of different regions across many data sets without normalization.

We next asked whether treating MHBs as individual units and performing quantitative analysis based on MHL would provide an advantage over existing approaches based on individual CpG sites or averaging multiple CpG sites in certain genomic windows. To this end, we sought to cluster 65 WGBS data (including 4 additional cancer WGBS sets[12](#_ENREF_12" \o "Heyn, 2016 #726)) sets from human solid tissues based on the MHL. U[nsupervised](http:///h) clustering with the top 15% most variable MHBs showed that, regardless of the data sources, samples of the same tissue origin clustered together (**Figure 3a**), while cancer samples and stem cell samples exhibit distinct patterns from adult human somatic tissues. PCA analysis on all MHBs genome-wide yielded a similar pattern (**Supplementary Fig. 3**). To identify a subset of MHBs for effectively clustering of human somatic tissues, we constructed a tissue specific index (TSI) for each MHB (see Methods). Random Forest based feature selection identified a set of 1,360 tissue-specific MHBs (**Supplementary Table 2**) that can predict tissue type at an accuracy of 0.89 (95%CI: 0.84-0.93), despite the fact that several tissue types share rather similar cell compositions (i.e. muscle vs. heart). Using this set of MHBs, we compared the performance between MHL, average methylation level in the MHL regions (AMF) and 5mC level of individual CpG site (MAS). MHL and the average methylation provided similar tissues specificity, while MHL have a lower noise (background noise: 0.29, 95%CI: 0.23-0.35) compared with average methylation (background noise: 0.4, 95%CI: 0.32-0.48). Clustering based on individual CpGs in the blocks has the worst performance, which might be due to higher biological or technical viability of individual CpG sites (**Figure 3b**). Thus block-level analysis based on MHL is advantageous over single CpG or local averaging of multiple CpG sites in distinguishing tissue types.

The human adult tissues that we used in this study have various degree of similarity among each other. We hypothesize that this is primarily defined by their developmental lineage, and that the related MHBs might reveal epigenetic insights related to germ layer speciation. We grouped all the data sets based on the three germ layers, and searched for MHBs that have differential MHL. In total we identified 31 ectoderm specific MHBs (16 hyper- and 15 hypo-methylated), 49 endoderm specific MHBs (34 hyper and 15 hypo-methylated) and 124 mesoderm specific MHBs (109 hyper and 15 hypo-methylated) (see Methods, **Supplementary Table 3**). We speculated that some of these MHBs might capture binding events of transcription factors (TF) specific to developmental germ-layers. Compared with ENCODE TFBS data[13](#_ENREF_13" \o ", 2012 #34), we observed distinctive patterns of TFs binding to layer specific MHBs. (**Supplementary Fig. 4**). For layer specific MHBs with hypo-methylation MHL, which tends to represent activation signals, we identified 53 TF binding events in mesoderm specific MHBs, 71 in endoderm specific MHB and 2 in ectoderm specific MHBs. Gene ontology analysis showed TFs binding events to mesoderm exhibit negative regulator activity, while TFs binding to endoderm exhibited positive regulator activity (**Supplementary Table 4**). For layer specific MHBs with hyper-methylation MHL which represent repressive signals, we identified 38 TF binding events in mesoderm specific MHBs, 102 in endoderm specific MHB and 145 in ectoderm specific MHBs. Interestingly, ectoderm and endoderm share few bounded TFs, while mesoderm tissues share different groups of TFs with ectoderm and endoderm. We identified two endoderm specific hyper-MHL regions, which are related to *ESRRA* and *NANOG*. This is consistent with a previous finding that mouse ES cells differentiated spontaneously into visceral/parietal endoderm upon Nanog knock-out[14](#_ENREF_14" \o "Mitsui, 2003 #698). Gene ontology analysis shown mesoderm and endoderm shared hypo-MHL regions play the significant role to induce the development of widely tissues while ectoderm specific hyper-MHL regions might induce the ectoderm development by suppressing path towards the immune lineage (**Supplementary Fig. 4**). In summary, MHB, a coordinated methylation region, might be involved in the tissue development independently or interacted with other known genomic regulation elements.

#### **Methylation-haplotype based deconvolution of circulating cell-free DNA in cancer patients.**

A unique aspect of methylation haplotype analysis is that the pattern of co-methylation, especially within MHBs, is robust in capturing low-frequency alleles among a mixed population of molecules or cells, in the presence of biological noise or technical variability (ie. incomplete bisulfite conversion or sequencing errors). To explore the clinical utility, we next focused on the methylation haplotype analysis of primary tumor tissues and plasma from cancer patients, of which various low fractions (down to ~0.1%) of DNA molecules were released from tumor cells and potentially carry epigenetic signatures different from blood. We isolated 4-122ng (average 20ng) of cell-free DNA on an average of 866µL human plasma from 59 cancer patients and 75 normal individuals, except for four outliers due to cell lysis. Due to the limited DNA availability, we performed scRRBS[15](#_ENREF_15" \o "Guo, 2013 #723) on 1ng of cfDNA from 95 plasma samples and obtained an average of 13 million paired-end 150bp reads per sample. On average, 57.7% WGBS-defined MHBs were covered in our RRBS data set.

We sought to detect tumor specific signatures in the plasma samples, by using tumor tissues as the reference and normal samples as the background. For 5 lung cancer plasma and 5 colon cancer plasma, we also obtained 10 matched primary tumor tissues, and generated RRBS data (30 million reads per sample) from 100ng of tumor genomic DNA. We focused on MHBs with low MHL (i.e. genomic regions that have low or no methylation) in the blood, and asked whether we can detect cancer-specific highly methylated haplotypes (csHMH). We required that these csHMH are present only in the tumor tissues and the matched plasma from the same patient, but not in whole blood or any other non-cancer samples in our data set. Hence these are highly confident signature in the circulating DNA. We detected csHMH in all the cancer patient plasma samples (Average=36, IQR=17, **Supplementary Table 5a**). Totally, these HMHs are associated with 183 genes, some of which are known to be aberrantly methylated in human cancers such as *WDR37*, *VAX1*, *SMPD1* (**Supplementary Table 5b**). Next, we extended the analysis to 49 additional cancer plasma samples without matched tumor samples, using 65 normal plasmas as the background for subtraction. On averagely 60 (IQR=31) csHMH were identified for each cancer plasma sample (**Supplementary Table 5c**). Interesting, 35% of csHMH called on matched tumor-plasma pairs were validated with the expanded set of non-matched tumor plasma and normal plasma data.

Next we sought to quantify the tumor load based on the deconvolution analysis to the cancer and normal plasma by non-negative decomposition with quadratic programming with the reference of LC, CRC and 10 type of normal tissues. We estimated that a predominant fraction, 72.0% (95% CI:0.659-0.782) in the cancer and normal plasma were contributed by white blood cells, which is consistent with the levels reported in a recent study based on lower-coverage whole genome bisulfite sequencing (69.4%)[16](#_ENREF_16" \o "Sun, 2015 #724). Primary tumor and normal tissue-of-origin contributes at the similar level of 2.3% (95% CI: 0.4%-4.2%) and 3.0% (95% CI:1.2%-4.4%). We also applied the similar analysis to normal plasma, and we found the tumor contribution to normal plasma were significantly lower (P=3.4x10-5 and 5.2x10-10 for CRC and LC, respectively) than that to cancer plasma, with the contribution of 0.17% and 1.0% for colon and lung cancer respectively. We also found 76.7% CRC and 89.6% LC samples were detected the contribution from tumor tissues while only 13% and 26% normal plasmas have certain (low) tumor contribution. Therefore, circulating cell-free DNA contains a relatively stable fraction of molecules released from various normal tissues, whereas in cancer patient tumor cells and apoptosis cells released DNA molecules that can be more abundant than normal tissues (**Supplementary Table 6**).

We then asked whether we can identify a small subset of MHBs among all the RRBS targets that have significantly higher levels of MHL in cancer plasma than in normal plasma. We found 81 and 94 MHBs with significantly higher MHL for colon and lung cancer ( **Supplementary Table 7a-b**). Some of these regions (such as *HOXA3*) have been reported to be aberrantly methylated in lung cancer and colon cancer. Using these MHBs as markers, the diagnostic sensitivity is 96.7% and 93.1% for colorectal cancer and lung cancer at the specificity 94.6% and 90.6%. As a comparison, we also performed a prediction based on average 5mC methylation level within these MHB regions and based on genome-wide single CpG site. Our result showed MHL based cancer prediction were better than average 5mC methylation level (sensitivity of 90.0% and 86.2% while specificity of 89.3% and 90.6% for CRC and lung cancer) and methylation signal of individual CpG site (sensitivity of 89.6% and 80.6% while specificity of 89.3% and 92.0%).

We then sought to use the information from normal human tissues, primary tumor biopsies and cancer cell lines to improve the detection of cancer in cfDNA. We started by selected a subset of MHBs that show high MHL (>0.5) in cancer tissue samples and low MHL (<0.1) in whole blood cell. We then clustered these MHBs into three groups based on the MHL in all normal and cancer plasma, as well as cancer and normal tissues (**Figure 4**). We identified a Group II MHBs that have high MHL in cancer tissues and low MHLs in normal tissues. Cancer plasma showed significantly higher MHL in these regions than normal plasma (*P*=1.4×10-12 and 6.2×10-8 for colorectal cancer, lung cancer, respectively). By mixing the sequencing reads from cancer tissues and whole blood samples (WB), we computationally created synthetic admixture at different tumor fractions. We found that MHL is 2-5 folder higher (and hence easier to detect) than the methylation level of individual CpG sites across the full range of tumor fraction (**Supplementary Table 8a-b**). We then took the individual plasma data sets, and predicted the tumor fraction based on the MHL distribution established by computational mixing (**Figure 4)**. Except for a small number (N<5) of outliers, we observed significantly higher MHL in cancer plasma than in normal plasma (**Supplementary Fig. 7**). Note that all Group II MHBs were selected without using any information from the plasma samples, and hence they should be generally applicable to other plasma samples. Interestingly, we also found that the estimated tumor DNA fraction were positive correlated with normalized cfDNA yield from the cancer patients (P<0.002, **Supplementary Fig. 8 and Supplementary Table 9**).

Recent studies[2](#_ENREF_2" \o "Lehmann-Werman, 2016 #1017),[16](#_ENREF_16" \o "Sun, 2015 #724),[17](#_ENREF_17" \o "Snyder, 2016 #725) have demonstrated that epigenetic information imbedded in cfDNA has the potential for predicting tumor’s tissue-of-origin. Consistently, we found that tissue-of-origin derived methylation haplotypes were the third abundant fraction in cancer plasma. Here we asked whether a MHL-based framework and a set of targets derived from whole genome data would allow us to predict tissue-of-origin with quantifiable sensitivity and specificity, which is crucial for future clinical applications. W and RRBS normal 2,880 which could be used to predict the tissue types (Table S10) We then applied these tissue-specific MHBs or subsets to predict the tissue-of-origin for the cancer plasma sample.Although we found at least large number of tissue-of-origin specific MHBs that have low MHL in normal plasma (**Figure 5a**), the multiclass prediction based on traditional classification (random forest) were quite limited caused by high diversity of the tissue classes (N=10). Using the method which counting the total number of ts-MHBs in the plasma samples and comparing with all other potential tissues, the maximum likely tissue-of-origin could be conjectured. We demonstrate that the minimum 10 tsMHB fragment could provide at least average 90% accuracy for the tissue sample mapping (**Figure 5b**). And then applying the method and the predictors identified from human adult normal tissues to the full set of 59 plasma data from cancer patients and normal individual, we achieved an average prediction accuracy of 82.8%, 88.5% for the plasma from [colorectal](https://en.wikipedia.org/wiki/Colorectal_cancer) cancer, lung cancer and the accuracy of 91.2% for normal plasmas with 5-fold cross-validation setting (**Figure 5c**). In summary, we assemble a pipeline and demonstrate that tissue specific MHL can be used in the tissue prediction and meanwhile we identified a subset to be used in our cancer (CRC, LC) and normal plasma distinguish.

## Discussion

DNA methylation between the adjacent CpGs are usually highly correlated. However, the distribution of the linkage disequilibrium was not well profiled. In the present study, we first investigate the linkage disequilibrium of the adjacent CpGs in the whole human genome in the normal tissue samples and defined a specific genomic regions called methylation haplotype blocks (MHB). We demonstrated MHB regions were widely dispensed in human genome, however, they were significantly enriched in some regulatory regions, such as enhancer, promoter while were significantly decreased in LAD, LOCKS regions. In the next step were proposed a metric titled as methylation haplotype load (MHL) which was used to measure the average methylation level weighted by the length of the sequencing fragment. We demonstrate MHL have more powerful ability to indicate both the methylation diversity and methylation level, simultaneously, compared with currently average methylation level in specific genomic region, methylation entropy[18](#_ENREF_18" \o "Hon, 2013 #681) as well as epi-polymorphism [19](#_ENREF_19" \o "Landan, 2012 #675). Genome-wide analysis to the cancer tissue and plasmas from the cancer patients, we found MHL would be used as powerful predictor for tissue and development layer distinguish, cancer diagnosis as well as plasma origin tissue prediction.

MHBs defined and identified by the present study represent a distinct set of genomic elements which would be provide important genomic structure or functional elements in human development and complex disease pathogenesis. These regions should be highly linked for the CpGs which might indicate the coordinate dynamic change was required by other cis or trans-regulatory elements in their biological function. MHBs were significantly enriched in some methylation classic functional genomic regions, such as CpG Island, imprint gene regions which were prior expected. It is very interesting, MHBs regions were also enriched in previous identified cancer associated DMR and VMR regions, which suggest MHBs would be taken as important predictor for cancer diagnosis and also indicate these DMR and VMR might have more power in the clinical cancer research since these CpGs would have high coordinate changes and its methylation status can be detected with high stability. Furthermore, we also found MHBs significantly enriched in enhancer, super-enhancer, promoter, UTR-5, which indicating MHBs would enriched in cis-regulatory genomic regions. Finally, MHBs were negative correlated with LAD and LOCKS regions which indicating the However, it should be investigated with more comprehensive study, since the genomic regions of LAD, LOCKS were highly large than MHBs, the enrichment analysis between these two kinds of regions might be not accurate.

We identified large number of significant different methylated MHB fragments in cancer plasmas and large number of them have been previously identified as the promising DNA methylation predictor for cancer non-invasive diagnosis were validated to be high frequently methylated in cancer plasmas. We also demonstrate that MHL have powerful ability to detect rare (<1.0%) long continuous methylated fragment compared with 5mC level and the power positively correlated with the number of continuous methylated CpGs. We observed that the MHL signal will be 10-20 time higher than 5mC level when the number of continuous CpG site come up to 6 CpGs. Our result provided a model to detect rare cancer specific long continuous methylated fragment. The prediction performances based on MHL and average 5mC within in genomic interval are better than the single CpG site analysis which mainly caused by high variation and high missing ration for single CpGs among the samples of same groups which indicating the methylation predictor would be more powerful when more continuous CpG site were measured and analysis together.

We conducted a comprehensive genome-wide DNA methylation haplotype based investigation to human genome in a large sample size including normal, cancer tissues and cancer plasmas. With the deconvolution analysis which is based on non-negative decomposition with quadratic programming, we found the majority of the contribution of cell-free DNA were derived from normal white cells and the tumor and tissue-of-origin fragment were in very low level which are consistent with previous report. However, we find a significant higher tumor fragment contribution in the plasma of cancer patients compared with normal plasma. In the deconvolution analysis, we also found the significant contribution from liver and Intestine, which might be caused by non-independent of reference (mixture) and single-cell based methylation reference to human tissues would provide important boost in the future. We validated the distinguish ability of the MHL between cancer plasma and normal plasma, the prediction performance was quite accurate for such binary prediction problem. However, for the realistic problem, we need to predict the cancer plasma to its tissue-of-origin (multiple mapping). Therefore, multi-class prediction to the plasma samples were also investigated. We found the traditional multi-class prediction strategy (random forest) could perform well when the tissue of mapping to no more than 3 tissues, however, when we are mapping plasma to 10 tissues, the perform was quite limited, the best average accuracy ranging from 30%-40% (data not shown), caused by multiple reasons such as the large number of the classes, low coverage of sequencing and stochastic releasing process of cancer fragment to the plasma. Our tsMHB fragment counting method was quite stable and suitable for genome-wide plasma based sequencing. We also demonstrate that the more tissue-specific MHBs enrolled in the prediction model, the more stable of the prediction performance would be obtained, however, in the clinical practice, more features indicate higher cost for the medical application. When we explore the mistakenly classified samples, we notice that 4/5 mis-classified colon plasma were metastatic colon cancer patients while the remaining is actually not colon cancer but tubular adenoma. For mis-classified lung cancer plasma, two lung cancer plasma (2/3) sample were failed to be mapped to lung without explicitly reasons while the remaining were actually not lung tissue but benign fibrous tissue.

We observed that the reads length in the BS-seq or RRBS would influence the length and number of the MHB and hence our results might represent the lower bound of the true MHBs in the human genome. On the other side, the profile of the methylation haplotype was also influenced by the sequencing depth. A high sequencing depth were required to obtain the high coverage for all the haplotypes especially when the genomic windows or sequencing length was long. However, our present study has shown the powerful advantage for the application of methylation haplotype based analysis in cancer diagnosis and tumor-of-origin prediction for the plasma from cancer patients with current prevalent GWBS and RRBS methodology. Since the MHB regions represent the specific human genome which exhibit highly coordinated methylation, the methylation status for MHB would be more useful to be predictor for the diagnosis or the phenotype identification. Together with previous evidence that epigenetic biomarker panel composed by large number of single CpG [16](#_ENREF_16" \o "Sun, 2015 #724), nucleosome footprint [17](#_ENREF_17" \o "Snyder, 2016 #725), methylation haplotype load in the methylation haplotype regions are also promising tool for non-invasive cancer diagnosis and tumor-of-origin mapping.

## Summary

We apply linkage disequilibrium to the methylation alleles on a single DNA molecule and defined methylation haplotype blocks in human genomes. We demonstrate MHB represent a distinct class of genomic feature with gene regulation and tissue differentiation. Furthermore, we proposed methylation haplotype load (MHL), an average methylation weighted by order of the CpGs within MHB, to quantitatively measure both DNA methylation level and methylation complexity simultaneously. Eventually, we demonstrate MHL based genome-wide DNA methylation analysis would provide a novel strategy in cancer biomarker identification and tumor-of-origin prediction to cancer plasma samples.

## Accession codes

WGBS and RRBS data are available at the Gene Expression Omnibus (GEO) under accession GSE79279.

## Acknowledgement

## Author’s Contribution

## Competing Financial interests

The authors declare no competing financial interests.

## Abbreviation:

MHB: methylation haplotype load; MHL: Methylation Haplotype Load; cf-DNA: Circulating cell-free DNA; RRBS: [Reduced representation bisulfite sequencing](http:///h); scRRBS: single-cell reduced-representation bisulfite sequencing; TCGA: The Cancer Genome Atlas project; GEO: Gene Expression Omnibus; LC: Lung Cancer; CRC: Colorectal cancer; ACC: Accuracy; csHMH: cancer specific high methylation haplotype; tsMHB: tissue specific methylation haplotype block regions.

## Figure legends

Figure 1. DNA methylation haplotype in human genome inferred by different tissues in different platforms. (a) Workflow diagram for methylation haplotype calling and its application in cancer diagnosis and tumor-of-origin prediction for cancer plasma. (b) DNA methylation blocks were inferred with genome-wide bisulfite-sequencing. The Pearson correlation was calculated and empirical threshold were applied for the methylation block inference genome-widely. (c) linkage disequilibrium between adjacent CpG loci in different cells including stem cells and progenitors, normal adult tissues as well as primary tumor. (d). the distribution characteristics of the methylation blocks in human genome including gene region and CpG high density regions. (e). Enrichment analysis to MHB in known regulatory elements.

Figure 2. The performance of methylation haplotype load in distinguishing the methylation complexity and methylation frequency compared with other metrics in different scenarios. Methylation entropy (ME) and epi-polymorphism were same with previous report and the formula were shown in method section.

Figure 3. Methylation haplotype load has ability in tissue, development layer and disease status distinguishing. (a) Unsupervised cluster analysis shown genome-wide DNA methylation haplotype load could represent the sample relationship. The sample with same origin were cluster together preferentially. (b) Layer specific methylation haplotype region showed the development layer relationship. (c) Tissues specific methylation haplotype loading (MHL) shown advantages in tissues distinguish compared with average methylation frequency (AMF) and methylation for all CpG site (MAS). Tissue specificity value (TSV) was the average MHL for the corresponding tissue specific MHL in the correct samples while the background value (BV) were the average MHL in mis-assigned samples. Contract value were the division of TSV by BV.

Figure 4. MHL based cancer non-invasive biomarker identification.

Genomic fragments were selected for which are of high-MHL in cancer solid tissues and non/low-MHL in WB and the MHL status for these fragments in normal plasma, cancer plasma were investigated. (a). [Colorectal](https://en.wikipedia.org/wiki/Colorectal_cancer) cancer (b). Lung cancer. Cancer DNA fragment would be released to cancer plasma and caused increased MHL.

Figure 5. Methylation Haplotype Load in Cancer Diagnosis and Tumor-of-Origin Deconvolution.

(a). Schematic diagram to show tissue specific MHL signature were existed in normal tissue, solid tissue and tumor plasma while they are clean in normal plasma and whole blood sample in colorectal cancer and lung cancer. (b, c) Prediction model for tumor-of-origin prediction based on tissue specific MHL regions. WGBS and RRBS dataset with balanced samples for each tissue were collected to train the prediction model and our lab RRBS dataset were used to validate the prediction performance.

## Supplementary Figure Legends:

**Supplementary Figure 1**. Significant Overlap between MHBs and Genomic Regulatory Regions (Histone Modifications). Profiles of H3K27ac, H3K4me3 and H3K4me1 (input normalized reads per kilobase per million mapped reads, RPKM) over genomic regions which are plus and minus 1 kbp from the middle of methylation haplotype blocks and in order of the highest to lowest total signal from H3K4me3. X-axis represent different samples. Histone modification data were downloaded from ENCODE project as the wig files.

**Supplementary Figure 2.** Validation of MHB with Beadchip and RRBS data. The Pearson's r in RRBS and HM450 were significantly higher in overlapped MHBs with WGBS compared with the MHBs without overlapping with WGBS MHBs

**Supplementary Figure 3.** Genome-wide MHL matrix were applied to infer the relationship between the different samples with PCA analysis. PC1-PC2 two-dimension plot were shown and the samples from the same group were together in the PCA plot indicating the genome-wide MHL could reflect the sample information and this result were consistent with cluster analysis based on most variable regions.

**Supplementary Figure 4.** TFBS located in layer specific MHB regions shown different mechanism which MHB involved in layer development. Differential MHL regions were identified by comparison between the samples from different development layer. And then these differential MHL regions were mapped to TFBS and obtain the corresponding TFs. Genome ontology analysis to these TFs could reflect the roles of MHL regions in the layer development. Our result support that layer specific and share transcript factors could be used to infer the roles of MHB in layer development.

**Supplementary Figure 5.** Genomic fragments whose MHL were high in cancer samples and low in whole blood were collected as Figure 4. MHL of the regions were investigated in different sample groups (Left two figures). In our Figure 4, MHL change as the concentration of the mixture between WB and cancer solid tissue DNA ranging from 50% to 0.1% were shown, here the whole range from

99.99% to 0.001 were shown to descript the whole picture for the mixture simulation.

**Supplementary Figure 6.** MHL in plasma were 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 5-CpG non-methylation fragments and 9 partial sequenced and stochastic methylated fragment and 1 copy continuous 5-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.

**Supplementary Figure 7.** Estimated cancer fragment proportion by the derived standard curve proposed by mixture simulation with cancer and WB samples. The approximate cancer fragment proportion could be estimated for each plasma samples. We could expect that the cancer plasma would have significant higher cancer fragment component compared with normal plasma samples.

**Supplementary Figure 8.** Relationship between circulating tumor DNA fraction and circulating cell-free DNA yield from the patients and normal individuals. As we shown in the Figure 4, we could estimate potential tumor DNA fraction based on averaged MHL and therefore we could explore the relationship between circulating tumor DNA fraction and circulating cell-free DNA yield. We observed a significant positive correction between them within the suitable range of Estimated tumor fraction.

**Supplementary Figure 9.** Circulating tumor DNA components de-convoluted by DNA methylation haplotype. High-methylation haplotype of cancer plasma and normal plasma were compared with normal tissues, cancer primary tissue and WB. The most predominant contribution was removed one by one to estimate the contribution from different tissue or cells.

## Supplementary Tables:

**Supplementary Table 1.** Genome-wide identified MHBs based on 65 WGBS dataset.

**Supplementary Table 2.** Tissue specific MHB regions identified by tissue specificity index.

**Supplementary Table 3.** Layer specific MHB regions identified by layer specificity index.

**Supplementary Table 4.** Complete list for high-methylation haplotype shared by primary cancer tissue and matched plasma from CRC and lung cancer samples.

**Supplementary Table 5.** Component deconvolution of cancer plasma from WB, normal tissue and primary cancer tissues based on high-methylation haplotype.

**Supplementary Table 6.** Deconvolution to CRC, LC and normal plasma samples by 10 normal tissues and LCT, CCT

**Supplementary Table 7.** Significant differential MHB regions between CRC plasma and normal plasma.

**Supplementary Table 8.** Significant differential MHB regions between lung cancer plasma and normal plasma.

**Supplementary Table 9.** The signal of MHL is higher than average 5mC based on cancer DNA and WB DNA mixture simulation analysis.

**Supplementary Table 10.** Significant correlation between estimated cancer DNA fraction with cell-free DNA yield from the patients.

**Supplementary Table 11.** Predictors applied in prediction model from CRC, LC and normal plasma.

## Reference

1. Wigler, M., Levy, D. & Perucho, M. The somatic replication of DNA methylation. *Cell* **24**, 33-40 (1981).

2. Lehmann-Werman, R. *et al.* Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci U S A* **113**, E1826-34 (2016).

3. Shoemaker, R., Deng, J., Wang, W. & Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res* **20**, 883-9 (2010).

4. Shao, X., Zhang, C., Sun, M.A., Lu, X. & Xie, H. Deciphering the heterogeneity in DNA methylation patterns during stem cell differentiation and reprogramming. *BMC Genomics* **15**, 978 (2014).

5. Landau, D.A. *et al.* Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* **26**, 813-25 (2014).

6. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-51 (2008).

7. Wen, B., Wu, H., Shinkai, Y., Irizarry, R.A. & Feinberg, A.P. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat Genet* **41**, 246-50 (2009).

8. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-80 (2012).

9. Hansen, K.D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43**, 768-75 (2011).

10. Irizarry, R.A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* **41**, 178-86 (2009).

11. Ziller, M.J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477-81 (2013).

12. Heyn, H. *et al.* Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol* **17**, 11 (2016).

13. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

14. Mitsui, K. *et al.* The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113**, 631-42 (2003).

15. Guo, H. *et al.* Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res* **23**, 2126-35 (2013).

16. 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).

17. Snyder, M.W., Kircher, M., Hill, A.J., Daza, R.M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57-68 (2016).

18. Hon, G.C. *et al.* Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat Genet* **45**, 1198-206 (2013).

19. 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).