# 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: 4

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, Non-invasive detection

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), based on 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 then investigated the role of MHBs in the formation of high specificity methylated regions by first defining segments of genomic regions with high to low methylation variability across diverse samples. We found that the high specificity methylated regions are more likely to be stable and tightly regulated than low specificity regions; perhaps by their interactions with difference groups of transcription factors. 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)] 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)]. 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, called methylated haplotype load (MHL), for a systematic discovery of 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 patient plasma of three cancer types.

Methylation is one of most important inherited epigenetic modification in the human genome that occurs primarily on the cytosines in CpG dinucleotides. The dynamic change of the methylation profile are involved in most human biological and cellular, physiological, pathological changes [[3](#_ENREF_3)]. Similar with the DNA replication, the methylation status of symmetric CpG methylations could be faithfully inherited in a semi-conservative way with the assistant of DNMT1 [[4](#_ENREF_4)], therefore, the DNA methylation information conform to the classic theory in the population genetics, such as linkage disequilibrium [[5](#_ENREF_5)]. Consequently, the recombination of methylation status from the successive CpG loci provided basic profile for methylation haplotype. While previous studies have defined methylation heterogeneity based mainly on the methylation levels of individual CpG sites (Beta-value or M-value), the underlying methylation pattern within each allele and haplotype have not yet been examined. What’s more, methylation haplotype is a powerful matric to connect DNA methylation level, DNA methylation variability, DNA methylation complexity and cancer evolution [[6](#_ENREF_6), [7](#_ENREF_7)]. Therefore, it is important to determine the methylation pattern of consecutive CpG sites within a specific genomic region (methylation haplotype). Next generation based sequencing technique provide the opportunity to investigate the co-methylation between adjacent CpG sites. In our previous study, we have investigated the linkage disequilibrium (LD) between adjacent CpG sites on single DNA molecules analysis to characterize the correlation of methylation genomic variations such as SNPs [[8](#_ENREF_8)]. However, genome-wide investigation of the linkage disequilibrium (LD) between adjacent CpG sites on single DNA molecules in a large number samples were not conducted yet.

In addition, a range of molecular alterations found in tumor cells, such as DNA mutations and DNA methylation, is reflected in cell-free circulating DNA (cfc-DNA) released from the tumor into the blood, thereby making cfc-DNA an ideal candidate for the basis of a blood-based cancer diagnosis test. Comparing with mutation or copy number variation, DNA methylation own different aberrant characteristics in cancers, as it mainly happened in CpG island and shore [[9](#_ENREF_9)] with a relatively high frequency, therefore, it was considered as the most potential biomarker for cancer non-invasive diagnosis. Methylation based on single CpG locus and average of methylation in a genomic region have been applied in biomarker identification, However, such measurement ignored the correlation between adjacent CpG and the prediction accuracy would be discounted caused by high error rate for single CpG site. Finally, how to apply the linkage disequilibrium to cancer diagnosis and other clinical application were not considered yet. In the present study, we explored methylation haplotype in normal and cancer tissues and proposed a novel metric termed as methylation haplotype load (MHL) which we proved to be better than previous methylation entropy [[10](#_ENREF_10)] and epi-polymorphism [[11](#_ENREF_11)] to depict the methylation heterogeneity for the genomic regions. We carried out a comprehensive genome-wide DNA methylation analysis across 10 human normal tissues with WGBS and 45 plasma from cancer patients, 30 normal plasma and 15 solid cancer tissues with RRBS/scRRBS. We demonstrate that the methylation haplotype based metrics could be powerful to be applied in cancer diagnosis and tissue-of-origin prediction for plasma from cancer patients (Figure 1A).

## Results

#### **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 [[8](#_ENREF_8)], 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-end or paired-end Illumina sequencing reads were extracted to form methylation haplotypes, and pairwise “linkage disequilibrium” of DNA methylation r2 was calculated from the distribution of methylation haplotypes (Materials and 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, using 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 61 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 we identified ~ 55 billion methylation haplotype informative reads with total coverage of 58.2% of autosomal CpGs. A total of 147,888 MHBs at the average size of 95bp and minimum **2** CpGs within each block, which represents ~0.5% of the human genome (Supplementary Table 1).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 [[8](#_ENREF_8)] and previous report [[12](#_ENREF_12)]. Interestingly, in primary tumor tissues, we observed a reduction of perfectly coupled CpG pairs, which could be related to the pattern of discordant methylation that has been reported in VMR [[13](#_ENREF_13)].

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 next searched for MHBs in 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. On the other hand, Illumina methylation arrays only report average CpG methylation of all DNA molecules in a sample, which cannot be used for methylation linkage disequilibrium analysis. Therefore, we calculated the pairwise correlation coefficient of adjacent CpG methylation levels across the 101 RRBS and 637 450k methylation array data sets for block partitioning. Note that the presence of such correlated methylation blocks is a necessary but not sufficient condition for MHBs. Nonetheless, the absence of correlated methylation blocks in these data would indicate potential false discovery of MHBs.

Of the 101 ENCODE RRBS data sets, with Pearson's r to high density CpG regions, we identified a total of 23,517 MHBs, which have significant overlaps with pre-defined MHBs in WGBS data sets (8,920, P-value<2.2×10-16 ). Similarly, from the 637 human solid tissue samples (HM450 array) we identified 2,212 correlated methylation blocks, of which 1258 (56.8%, P-value<1.0×10-9) were overlapping with the MHBs. In addition, 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 (P-value<10-22, Supplementary Figure 2). However, we also notice that 83% (1045/1258) shared MHBs by WGBS and MH450 are located in CpG Island and 89.32% (7968/8920) shared MHBs by WGBS and RRBS are located in CpG Island, indicating the CpG island might play as the proxy to connect these two MHBs. Taken together, the MHBs that we identified from the 799 data sets 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 related enzymes.

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

[[14](#_ENREF_14" \o "Guelen, 2008 #699)][[15](#_ENREF_15" \o "Wen, 2009 #700)][[16](#_ENREF_16" \o "Dixon, 2012 #722)]

Previous studies on mouse and human [9][[17](#_ENREF_17" \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 dynamically methylated across 30 human cell and tissue types. These human dynamic CpGs were also 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 (Supplementary Figure 1). These findings suggest that MHBs may play a role in the formation of differentially methylated regulatory regions.

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 methylation haplotype load**

To enable quantitative analysis of the methylation patterns within MHBs across many samples, we need a single metric to define the methylated pattern of many 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 at different length. Compared with other metrics used in the literature, such as methylation level, methylation entropy, epi-polymorphism and haplotypes counts, MHL can represent the overall level of methylation, and 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 (4 extra cancer WGBS set from [Holger Heyn](http://www.ncbi.nlm.nih.gov/pubmed/?term=Heyn%20H%5Bauth%5D)’s study [[18](#_ENREF_18)]) sets from human solid tissues based on the MHL. U[nsupervised](http:///h) clustering analysis 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 were quite different from adult human somatic tissues. PCA analysis to all MHBs genome-wide yielded a similar pattern (Supplementary Figure 3). To identify a subset of MHBs for clustering human somatic tissues, we calculated a tissue specific index (TSI) for each MHB (see Methods). Random Forest based feature selection identified a set of 1360 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 similar cell compositions (ie. muscle vs. heart). Using this set of MHBs, we compared the performance between MHL, average methylation frequency in the MHL regions (AMF) and methylation level of all the CpG site (MAS), we found MHL and the average methylation provide 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 had the worst performance which might be due to higher biological or technical viability of individual CpG sites (Figure 3B).

The human adult tissues that we used in this study have various degree of similarity among each other. We hypothesize that this is primarily determined by their developmental germ layers of origin, and that the related MHBs might reveal epigenetic insights on germ layer speciation. Therefore 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. By matching against the ENCODE TFBS database, we found different profiles of TFs binding to the layer specific MHBs that exhibit systematic differences of MHL among the three developmental layers (Supplementary Figure 4). For layer specific MHBs with hypo-methylation MHL which represent activated cellular functions, we identified 53 TF binding in mesoderm specific MHBs, 71 TF in endoderm specific MHB and 2 TF in ectoderm specific MHBs. Gene ontology analysis shown TFs binding to mesoderm exhibit negative regulator activity while TFs binding to endoderm exhibited positive regulator activity. It is consisted to the functions of the different layers. Endoderm were mainly developed to functional organs which required abounded positive biological process while mesoderm were mainly developed to connective tissue including muscle, bone in which the biological process were limited. For layer specific MHBs with hyper-methylation MHL which represent de-activated cellular functions, we identified 38 TF binding in mesoderm specific MHBs, 102 TF in endoderm specific MHB and 145 TF in ectoderm specific MHBs. Interestingly, ectoderm and endoderm were high different with each other while mesoderm tissues were distributed between ectoderm and endoderm. We mapped these layer specific regions to the TF binding regions from ENCODE project and we found 2 Endoderm specific hyper-MHL regions which included ESRRA and NANOG. As the previous evidence shown in the absence of Nanog, mouse embryonic stem cells differentiate into visceral/parietal endoderm [[19](#_ENREF_19)]. Gene ontology analysis shown mesoderm and endoderm shared hypo-MHL regions play the significant role to induce the development of widely tissues. On the other side, Ectoderm specific hyper-MHL regions impress the immune system to induce the ectoderm development (Supplementary Figure 4).

#### **Detection of tumor methylation haplotype regions in circulating DNA.**

A unique aspect of methylation haplotype analysis is that the pattern of co-methylation, especially within MHBs, is robust in capturing rare events in 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, as well as plasma from cancer patients of which various low fractions (typically 0.1-1%) of DNA molecules were released from tumor cells.

We tried to identify cancer specific hyper-methylated regions to investigate the power of cancer diagnosis with MHL by comparing patient plasma with healthy plasma. The contents of cell-free DNA varies dramatically in cancer and normal plasma samples. The patient plasma samples (1 mL) yielded from 4 ng to 122 ng of DNA while the normal samples (1mL) yielded from 2 ng to 35 ng of DNA. Considering the low content of cfDNA in both patient and healthy plasma, we applied a scRRBS protocol to generate bisulfite sequencing library from as little as 1 ng of cfDNA [[20](#_ENREF_20)]. The genome-wide DNA methylation profiles for 30 colon cancer plasma, 29 lung cancer plasma, 10 pancreatic cancer plasma and 20 normal plasma were established by RRBS with an average of 3.4 million paired-end 150bp reads per sample. In average, 57.7% MHBs derived from WGBS dataset were covered in RRBS dataset with 52.4% MHBs were covered by fresh tumor tissue, 55.3% MHBs were covered by frozen tumor tissue as well as 58.1% MHBs were covered by plasma samples.

Cancer plasma specific MHL biomarkers were then identified by differential analysis with multiple test correction (FDR <0.05) compared with normal plasma samples. We identified 81, 94 and 37 MHBs with significantly different MHL between colon, lung as well as pancreatic cancer plasma and normal plasma (Supplementary Table 4, 5 and 6). Based on the significant differential MHL regions, the diagnostic sensitivity for colon cancer, lung cancer and pancreatic cancer could come up to 90%, 93.2% and 95% while the specificity were all as high as 90%.

We knew the process of the DNA fragment from the tumor releasing to the blood was truth, however, the process would be a stochastic event and it is hard to detect same genomic fragment is 100% patients. We hypothesis that cancer specific hyper-methylated haplotype would be released to plasma randomly, therefore, MHL signals for these regions would be significant higher for cancer plasma than normal plasmas. We collected all the MHB regions whose MHL were positive in cancer tissues and negative in whole blood (WB) samples and then we check the MHL distribution in different groups such as cancer plasma, normal plasma, cancer solid tissue, normal tissues. We found these MHBs would be cluster into three groups (Figure 4). Regardless of cancer dataset and MHB groups, significantly higher MHL in cancer plasma compared with normal plasma were found (Figure 4 and Supplementary Figure). We notice the MHB Group II (GII) were promising methylation biomarkers which could be applied to be cancer non-invasive diagnosis and were hypermethylated only in cancer tissues while hypomethylated in all the normal tissues and normal plasmas. For GII regions the MHL in cancer plasma were significantly higher than that in normal plasmas with P-value of 1.4×10-12, 6.2×10-8 and 2.7×10-9 for colon cancer, lung cancer and pancreatic cancer, respectively (Figure 4). We notice that these cancer specific hyper-MHL regions were distributed in cancer plasma were low frequency and the average MHL were quite low which is consistent with our previous assumption that cancer specific hypermethylated fragment only account for 0.1%-1.0% total circulating DNAs. We tried to compare the MHL signal and the average methylation level in this situation. Simulation analysis by the mixture of cancer fragment and WB fragment with different proportion, we could observe the MHL change as the concentration of cancer fragment, we found the signal of MHL in cancer plasma were quite closed to the theoretical strength when the DNA fragment from cancer account for 0.1-1.0% proportion. On the other side, we also simulated the situation in which the plasma have 99% non-methylation or low methylated fragement with stochastic methylated CpG, as well as one 4-CpG continuous methylated cancer fragment and then we calculate the MHL and 5mC level, we found these values were quite consistent with what we found in the real situation (Supplementary Figure). Meanwhile, through scanning positive MHL signals in cancer plasma, cancer tissues, normal tissues, and negative signals in normal plasma as well as WB cells, we found large number of tissue derived signals would be passed on from normal tissue, cancer tissue to the plasmas in the cancer patients(Figure 5A). Finally, MHL based DNA methylation measurement provided large number of biomarker which have high tissue specificity and cancer specificity and eventually would be useful in cancer non-invasive diagnosis and tumor-of-origin deconvolution.

#### **Prediction of tumor-of-origin from circulating DNA based on tissue-specific MHBs.**

we tried to predict the tissues-of-origin for the plasma derived from 3 different kind of tumors (lung, colon and pancreatic cancer) based on tissues specific MHL regions. We collected the 10 tissues with high incidence of cancer as the fundamental components to build a prediction model so that we can capture the origin of the plasma for the cancer patients. 43 WGBS dataset (Figure 5B) was enrolled to be the training dataset to build the random forest prediction model. 1090 tissue specific MHL regions were introduced in the training model and 49 predictors with highest variable importance were selected in the validation stage (Supplementary Table 7). The prediction sensitivity to the plasma tissues-of-origin came up to 90.1%, 89.5%, 71.8% and 85.4% for the plasma from colon cancer, lung cancer, pancreatic cancer and normal individuals (Figure 5C). What’s more, the methylation status of the informative predictor were found to the significantly different in at least one of the cancer types compared with normal plasma, indicating these predictors would have high potential to be applied in the non-invasive cancer diagnosis.

## Discussion

DNA methylation between the adjacent CpGs are usually highly correlated. However, the distribution of the linkage disequilibrium were 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[[10](#_ENREF_10)] as well as epi-polymorphism [[11](#_ENREF_11)]. Genome-wide analysis to the cancer tissue and plasmas from the cancer patients, we found MHL would be used as powerful biomarker 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 biomarkers 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 conducted a comprehensive genome-wide DNA methylation investigation to human cancer genome in a large sample size including normal, cancer tissues and cancer plasmas. We identified large number of significant different methylated MHB fragments in cancer plasmas. In addition, large number of previous identified promising DNA methylation biomarker for cancer non-invasive diagnosis were validated to be high frequently methylated in cancer plasmas, such as SEPT9 [[21](#_ENREF_21)], RUNX3 [[22](#_ENREF_22)], TMEFF2 [[23](#_ENREF_23)], TCERG1L [[24](#_ENREF_24)] in colon cancer, RUNX1 [[25](#_ENREF_25)], SEPT9 [[26](#_ENREF_26)] in lung cancer and IKZF1 [[27](#_ENREF_27)], RB1 [[28](#_ENREF_28)] and SYK [[28](#_ENREF_28)]. 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 provide a model to detect rare cancer specific long continuous methylated fragment.

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 were 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 have 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 biomarker for the diagnosis or the phenotype identification. Together with previous evidence that DNA methylation biomarker panel composed by large number of single CpG [[29](#_ENREF_29)], nucleosome footprint [[30](#_ENREF_30)].

## Summary

We apply linkage disequilibrium to the methylation alleles on a single DNA molecule and defined methylation haplotype blocks in human genomes. Methylation haplotype load (MHL), an average methylation weighted by order of the CpGs within MHB, were proposed 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.

## Acession 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; NSCLC: Non-Small Cell Lung Cancer; ACC: Accuracy; SEN: Sensitivity; SPE: Specificity

## 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). Colon cancer (B). Lung cancer and (C) Pancreatic 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 colon cancer, lung cancer and pancreatic cancer. (B, C) Random forest 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)

Supplementary Figure 2.

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.

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.

Supplementary Figure 7.

Supplementary Figure 8.

**Reference**

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

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, 2016. **113**(13): p. E1826-34.

3. Pujadas, E. and A.P. Feinberg, *Regulated noise in the epigenetic landscape of development and disease.* Cell, 2012. **148**(6): p. 1123-31.

4. Leonhardt, H., et al., *A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei.* Cell, 1992. **71**(5): p. 865-73.

5. Reich, D.E., et al., *Linkage disequilibrium in the human genome.* Nature, 2001. **411**(6834): p. 199-204.

6. Feinberg, A.P. and R.A. Irizarry, *Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease.* Proc Natl Acad Sci U S A, 2010. **107 Suppl 1**: p. 1757-64.

7. Swanton, C. and S. Beck, *Epigenetic noise fuels cancer evolution.* Cancer Cell, 2014. **26**(6): p. 775-6.

8. Shoemaker, R., et al., *Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome.* Genome Res, 2010. **20**(7): p. 883-9.

9. Irizarry, R.A., et al., *The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.* Nat Genet, 2009. **41**(2): p. 178-86.

10. Hon, G.C., et al., *Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues.* Nat Genet, 2013. **45**(10): p. 1198-206.

11. Landan, G., et al., *Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues.* Nat Genet, 2012. **44**(11): p. 1207-14.

12. Shao, X., et al., *Deciphering the heterogeneity in DNA methylation patterns during stem cell differentiation and reprogramming.* BMC Genomics, 2014. **15**: p. 978.

13. Landau, D.A., et al., *Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia.* Cancer Cell, 2014. **26**(6): p. 813-25.

14. Guelen, L., et al., *Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions.* Nature, 2008. **453**(7197): p. 948-51.

15. Wen, B., et al., *Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells.* Nat Genet, 2009. **41**(2): p. 246-50.

16. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-80.

17. Ziller, M.J., et al., *Charting a dynamic DNA methylation landscape of the human genome.* Nature, 2013. **500**(7463): p. 477-81.

18. Heyn, H., et al., *Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer.* Genome Biol, 2016. **17**(1): p. 11.

19. Mitsui, K., et al., *The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells.* Cell, 2003. **113**(5): p. 631-42.

20. Guo, H., et al., *Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing.* Genome Res, 2013. **23**(12): p. 2126-35.

21. Payne, S.R., *From discovery to the clinic: the novel DNA methylation biomarker (m)SEPT9 for the detection of colorectal cancer in blood.* Epigenomics, 2010. **2**(4): p. 575-85.

22. Nishio, M., et al., *RUNX3 promoter methylation in colorectal cancer: its relationship with microsatellite instability and its suitability as a novel serum tumor marker.* Anticancer Res, 2010. **30**(7): p. 2673-82.

23. Lofton-Day, C., et al., *DNA methylation biomarkers for blood-based colorectal cancer screening.* Clin Chem, 2008. **54**(2): p. 414-23.

24. Yi, J.M., et al., *DNA methylation biomarker candidates for early detection of colon cancer.* Tumour Biol, 2012. **33**(2): p. 363-72.

25. Feng, Q., et al., *DNA methylation in tumor and matched normal tissues from non-small cell lung cancer patients.* Cancer Epidemiol Biomarkers Prev, 2008. **17**(3): p. 645-54.

26. Powrozek, T., et al., *Septin 9 promoter region methylation in free circulating DNA-potential role in noninvasive diagnosis of lung cancer: preliminary report.* Med Oncol, 2014. **31**(4): p. 917.

27. Kisiel, J.B., et al., *New DNA Methylation Markers for Pancreatic Cancer: Discovery, Tissue Validation, and Pilot Testing in Pancreatic Juice.* Clin Cancer Res, 2015. **21**(19): p. 4473-81.

28. Melson, J., et al., *Commonality and differences of methylation signatures in the plasma of patients with pancreatic cancer and colorectal cancer.* Int J Cancer, 2014. **134**(11): p. 2656-62.

29. 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, 2015. **112**(40): p. E5503-12.

30. Snyder, M.W., et al., *Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin.* Cell, 2016. **164**(1-2): p. 57-68.