# Deconvolution of epigenetic heterogeneity in human tissues and plasma DNA by tightly coupled CpG methylation.

Shicheng Guo1,3, Dinh Diep1,3, Nongluk Plongthongkum1, Ho-Lim Fung1, Kang Zhang2, Kun Zhang1,2\*

1Department of Bioengineering, 2Institute for Genomic Medicine, University of California at San Diego, La Jolla, California, USA.

3Equally contributed authors.

\*Corresponding authors:

Kun Zhang, Email: kzhang@bioeng.ucsd.edu

Keywords: Methylation haplotype, epigenetic heterogeneity, circulating cell-free DNA

## Abstract

Adjacent CpG sites in mammalian genomes can be co-methylated due to the processivity of methyltransferases or demethylases. 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. We defined 147,888 blocks of tightly coupled CpG sites, called methylation haplotype blocks (MHBs), in the human genome with 61 sets of whole genome bisulfite sequencing (WGBS) data, and further validated with 101 sets of reduced representation bisulfite sequencing (RRBS) data and 637 sets of methylation array data. Using a block level metric called methylation haplotype load (MHL), we performed deconvolution of heterogeneous samples and tissue-of-origin methylation analysis. Finally, we demonstrated quantitative estimation of tumor load and tissue-of-origin mapping in the circulating cell-free DNA of 59 cancer patients using methylation haplotypes.

## Introduction

CpG methylation in mammalian genomes is a relatively stable epigenetic modification, which can be transmitted across cell division[1](#_ENREF_1) through DNMT1, and dynamically established, or removed by DNMT3 A/B and TET proteins. Due to the processivity of some of these enzymes, physically adjacent CpG sites on the same DNA molecules can share similar methylation status, although discordant CpG methylation has also been observed, especially in cancer cells. The theoretical framework of linkage disequilibrium[2](#_ENREF_2), which was developed to model the coordinated segregration of adjacent genetic variants on human chromosomes among human populations, can be applied to the analysis of CpG co-methylation in cell populations. A number of studies related to the concepts of methylation haplotypes[3](#_ENREF_3), epi-alleles[4](#_ENREF_4), or epi-haplotypes[5](#_ENREF_5) 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 the NIH RoadMap Epigenomics project[6](#_ENREF_6) and the EU Blueprint Epigenome project[7](#_ENREF_7) 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 WGBS data generated in this study, allowed us to perform genome-wide characterization of local coupled CpG 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 heterogeneous samples, such as different white blood cells in whole blood[8](#_ENREF_8), fetal components in maternal cell-free DNA[9](#_ENREF_9), or circulating tumor DNA in plasma[9](#_ENREF_9). 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. Recently, Lehmann-Werman et al demonstrated a superior sensitivity with multi-CpG haplotypes in detecting tissue-specific signatures in circulating DNA[10](#_ENREF_10). The markers in that study were discovered from Infinium 450k methylation array data, which represent only a limited fraction of the human genome. Here we performed an exhaustive search of tissue-specific methylation haplotype blocks across 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 identified markers, we demonstrated accurate determination of tissue origin as well as estimation of tumor load in clinical plasma samples from patients of lung cancer (LC) and colorectal cancer (CRC) (**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 the concept of genetic linkage disequilibrium[2](#_ENREF_2),[3](#_ENREF_3) and the r2 metric to quantify the degree of coupled CpG methylation among different DNA molecules of the same samples. 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 abundance of different methylation haplotypes (see Methods).

We started with 51 sets of published whole genome bisulfite sequencing (WGBS) data from human primary tissues[11](#_ENREF_11),[12](#_ENREF_12), as well as the H1 human embryonic stem cells, *in vitro* derived progenitors[13](#_ENREF_13) and human cancer cell line[14](#_ENREF_14),[15](#_ENREF_15). We also included an in-house generated WGBS dataset from 10 adult tissues of one human donor. Across these 61 samples (>2000x combined genome coverage) we identified a total of ~ 55 billion methylation haplotype informative reads that cover 58.2% of autosomal CpGs. The uncovered CpG sites were either in regions with low mappability, or CpG sparse regions where there are too few CpG sites within Illumina read pairs for deriving informative haplotypes. *We then partitioned the full human genome into blocks of tightly coupled CpG methylation 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).* We identified 147,888 MHBs at the average size of 95bp and minimum 3 CpGs per block, which represents ~0.5% of the human genome that tends to be tightly co-regulated on the epigenetic status at the level of single DNA molecules (**Supplementary Table 1, Supplementary Figure 1ab).** The majority of CpG sites within the same MHBs are near perfectly coupled (r2 ~1.0) regardless of the sample type. 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%, P-value<2.6x10-16), and the over-representation of partially coupled CpG pairs that are over 100 bp apart while the linkage was slightly decayed in primary cancer dataset (87.8%, mixture of CRC and LC) and the result was validated by another independent WGBS data from kidney cancer[16](#_ENREF_16) (**Figure 1c, Supplementary Figure 2**). Gene Ontology enrichment analysis to MHB regions whose r2 is decayed compared with normal found significant association with cancer related pathways and functions (**Supplementary Table 2**). This is consistent with our previous observations on a smaller dataset with 2,020 CpG islands[3](#_ENREF_3) for culture cell lines and another previous report[17](#_ENREF_17). Interestingly, in tumor samples, we observed a reduction of perfectly coupled CpG pairs, which could be related to the pattern of discordant methylation recently reported in variable methylation regions (VMR)[18](#_ENREF_18),[19](#_ENREF_19).

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 presence of MHBs in a wider range of human tissues and cultured cells, we examined 101 published reduced representation bisulfite sequencing (RRBS) datasets from the ENCODE project that included cell line and normal tissue samples, as well as 637 published Infinium HumanMethylation450K BeadChip (HM450K) datasets from the TCGA project that included 11 human cancer tissues. The RRBS datasets 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 Pearson’s correlation coefficient from methylation levels of adjacent CpGs across different sample sets for block partitioning. Note that the presence of such correlated methylation blocks is a necessary but not sufficient condition for MHBs (**Supplementary Figure 3a**). Nonetheless, the absence of correlated methylation blocks in these data would invalidate the pattern of MHBs. We identified 23,517 and 2,212 correlated methylation blocks from RRBS and HM450K data respectively, among which 8,920 and 1,258 have significant overlaps with WGBS-defined MHBs. Additionally, we observed significantly higher correlation (r2) among the CpGs within the MHB regions compared CpG loci outside MHBs in HM450K and RRBS dataset (**Supplementary Figure 3b**), further supporting the block-like organization of local CpG co-methylation across a wide variety of cells and tissues. Taken together, the MHBs that we have 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 coupled with the local density of CpG dinucleotides.

**Co-localization of methylation haplotype blocks with known regulatory elements.** The MHBs established by 61 sets of WGBS data appear to represent a distinct type of genomic feature that partially overlaps with multiple well-documented genomic elements (**Figure 1d**). Among all the methylation blocks, 60,828 (41.1%) were located in intergenic regions while 87,060 (58.9%) regions in transcribed regions. These MHBs were significantly (P<10-6) enriched in enhancers (enrichment factor=7.6), super enhancers (enrichment factor=2.3), promoter regions (enrichment factor=14.5), CpG islands (enrichment factor=70.4) and imprinted genes (enrichment factor=54.6). In addition, we observed modest depletion in the lamina-associated domains (LAD) [*20*](#_ENREF_20) and the large organized chromatin K9 modifications (LOCK) regions[*21*](#_ENREF_21) (46% and 37% of the expected values), modest enrichment in TAD[*22*](#_ENREF_22). Importantly, we observed a strong (26-fold) enrichment in VMR (**Figure 1e**), suggesting that increased epigenetic variability in a cell population or tissue can be coordinated locally among hundreds of thousands of genomic regions[*23*](#_ENREF_23). We further examined a subset of MHBs that do not overlap with CpG islands, and observed a consistent enrichment pattern (**Figure 1e, Supplementary Figure 1c**), suggesting that local CpG density alone does not account for the enrichment.

Previous studies on mouse and human[24](#_ENREF_24),[25](#_ENREF_25) demonstrated that dynamically methylated regions were associated with regulatory regions such as enhancer-like regions marked by H3K27ac and transcription factor binding sites. In human, 21.8% of autosomal CpGs were found to be differentially methylated across 30 human cell and tissue types[17](#_ENREF_17). These CpGs were enriched at low to intermediate CpG density promoters. Using publicly available histone mapping data for human adult tissues, we found co-localization of methylation haplotype blocks with marks for active promoters (H3K4me3 with H3K27ac), but not for active enhancers[26](#_ENREF_26) (no peak for H3K4me1) (**Supplementary Figure 4).** Meanwhile, we found that enhancers tend to overlap with CpG sparse MHBs, whereas the overlap with super enhancers were independent of CpG density **(Supplementary Figure 1c)**. Therefore, MHBs likely capture the local coherent epigenetic signatures that are directly or indirectly coupled with transcriptional regulation.

#### **Block-level analysis 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 lengths (i.e. all possible substrings, see Methods). 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 genomic features and performing quantitative analysis based on MHL would provide an advantage over previous approaches using individual CpG sites or weighted (or unweighted) averaging of multiple CpG sites in certain genomic windows. To this end, we sought to cluster 65 WGBS data (including 4 additional colon and lung cancer WGBS sets[27](#_ENREF_27), **Supplementary Table 13d**) 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 Figure 5**). To identify a subset of MHBs for effective 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 1,365 tissue-specific MHBs (**Supplementary Table 3**) 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 fraction in the MHL regions (AMF) and all individual CpG methylation fraction (IMF). MHL and the average methylation provided similar tissue specificity, while MHL has 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 3c**). Thus block-level analysis based on MHL is advantageous over single CpG or local averaging of multiple CpG sites in distinguishing tissue types from regions of coupled CpG methylation and heterogeneity.

The human adult tissues that we used in this study have various degrees of similarity amongst 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 114 ectoderm-specific MHBs (99 hyper- and 15 hypo-methylated), 75 endoderm specific MHBs (58 hyper and 17 hypo-methylated) and 31 mesoderm specific MHBs (9 hyper and 22 hypo-methylated) (see Methods, **Supplementary Table 4**). Supervised cluster analysis based on layer specific MHBs shows aggregation among tissues of same the lineage (**Figure 3b**). We speculated that some of these MHBs might capture binding events of transcription factors (TF) specific to developmental germ-layers. Overlapped with TF binding events identified from ENCODE transcription factor binding sites data[28](#_ENREF_28), we observed patterns of TFs binding to layer specific MHBs. (**Supplementary Figure 6**). 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 MHBs and 2 in ectoderm specific MHBs. Gene ontology analysis showed TFs binding to mesoderm exhibit negative regulator activity, while TFs binding to endoderm exhibited positive regulator activity (**Supplementary Table 5**). For layer specific MHBs with hyper-methylation MHL, which tend to 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 shared few bounded TFs, while mesoderm tissues share multiple 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[29](#_ENREF_29). Gene ontology analysis showed that mesoderm and endoderm shared hypo-MHL regions might have regulatory functions in the fate commitment towards multiple tissues, whereas ectoderm specific hyper-MHL regions might induce the ectoderm development by suppressing the path towards the immune lineage (**Supplementary Fig. 6**). These observations are indicative of two distinctive “push” and “pull” mechanisms in the transition of cell states that have been harnessed for the induction of pluripotency by over-expressing lineage specifiers[30](#_ENREF_30).

#### **Methylation-haplotype based analysis of circulating cell-free DNA in cancer patients and healthy donors.** 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 heterogeneous 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 potential, we next focused on the methylation haplotype analysis of cell-free DNA (cfDNA) from healthy donors and cancer patients, of which various low fractions of DNA molecules were released from tumor cells and potentially carry epigenetic signatures different from blood. We isolated 4-122 ng (average 20 ng) of cell-free DNA from an average of 866 µL human plasma from 75 normal individuals and 59 cancer patients, except for four with unusually high yield due to cell lysis. Due to the limited DNA availability, we performed scRRBS[31](#_ENREF_31) on 1 to 10 ng of cfDNA from 134 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 on clinical samples.

We sought to detect the presence of tumor specific signatures in the plasma samples, using methylation haplotypes identified from tumor tissues as the reference and normal samples as the negative controls. For five lung cancer plasma samples and five colorectal cancer plasma samples, we also obtained matched primary tumor tissues, and generated RRBS data (30 million reads per sample) from 100 ng 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-associated highly methylated haplotypes (caHMH). We required that such haplotypes were 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. We considered these highly confident tumor signature in circulating DNA. We detected caHMH in all cancer patient plasma samples (Average=36, interquartile range (IQR)=17, Supplementary **Table 6a**). These caHMHs were associated with 183 genes, some of which are known to be aberrantly methylated in human cancers such as *WDR37*, *VAX1*, *SMPD1* (**Supplementary Table 6b**). Next, we extended the analysis to 49 additional cancer plasma samples that have no matched tumor samples, using 65 normal plasmas as the background. On average 60 (IQR=31) caHMH were identified for each cancer plasma sample (**Supplementary Table 6c**). Interestingly, a significant fraction (35%) of caHMH called on matched tumor-plasma pairs were also detected the expanded set of cancer patient plasma samples. We noticed majority of caHMHs were individual specific while few caHMHs were present in at least 53% (16/30) and 62% (18/29) cancer plasma samples for CRC and LC **(Supplementary Figure 7)**. Improving the sampling depth, by either using more input cfDNA or reducing sample loss during the experiments, will likely increase the number of caHMHs commonly observed in multiple patients.

Next, we sought to quantify the cancer DNA fraction in cancer plasma samples using non-negative decomposition with quadratic programming deconvolution. We used the reference data from primary cancer biopsies (LC and CRC) and from 10 normal tissues and estimated that a predominant fraction, 72.0% (IQR=40%) in the cancer and normal plasma were contributed by white blood cells, which is consistent with the levels reported recently based on shallow whole genome bisulfite sequencing (69.4%)[9](#_ENREF_9). Primary tumor and normal tissue-of-origin contributed at the similar levels of 2.3% (IQR=3.7%) and 3.0% (IQR=4.4%). In contrast, when we applied the same deconvolution analysis to normal plasma, we found only residual plasma fragments with a tumor signature (0.17%, IQR=2.9% for CRC and 1.0%, IQR=3.1% LC), which were significantly lower (*P*=3.4x10-5 and 5.2x10-10 for CRC and LC, respectively) than cancer plasma. We also found that 76.7% plasma samples from CRC patients and 89.6% from LC patients had detectible contribution from tumor tissues while only 13% and 26% normal plasmas have residual tumor MHB signature (**Supplementary Figure 8**). Therefore, circulating cfDNA contains a relatively stable fraction of molecules released from various normal tissues, whereas in cancer patient tumor cells released DNA molecules that can be more abundant than normal tissues (**Supplementary Table 7**). The fractions of white blood cells observed are lower than what was reported previously[9](#_ENREF_9), mostly likely due to the inclusion of 10 normal tissue types in our deconvolution analysis.

We next 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 colorectal and lung cancer (**Supplementary Table 8**). The majority (71/81 for CRC and 83/94 for LC) were also present in at least one of the matched primary tumor and plasma pairs. Some of these regions (such as *HOXA3*) have been reported to be aberrantly methylated in lung cancer and colorectal cancer. Using these MHBs as markers, the diagnostic sensitivity is 96.7% and 93.1% for colorectal cancer and lung cancer respectively at the specificity 94.6% and 90.6% respectively. As a comparison, we also performed a prediction based on average 5mC methylation level within these MHB regions, or based on genome-wide single CpG sites. MHL was found to be superior to average 5mC methylation level (sensitivity of 90.0% and 86.2%; specificity of 89.3% and 90.6% for CRC and lung cancer) and methylation of individual CpG site (sensitivity of 89.6% and 80.6%; 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 circulating tumor DNA (ctDNA). We started by selecting a subset of MHBs that show high MHL (>0.5) in primary cancer biopsies and low MHL (<0.1) in whole blood, 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 4ab**). We identified a subset (Group II) of MHBs that have high MHL in cancer tissues and low MHLs in normal tissues (**Supplementary Table 9a-b**). Cancer plasma showed significantly higher MHL in these regions than normal plasma (*P*=1.4×10-12 and 6.2×10-8 for CRC and LC, respectively). By computationally mixing the sequencing reads from cancer tissues and whole blood samples (WB), we created synthetic admixtures at various levels of tumor fraction. We found that MHL is 2-5 folder higher than the methylation level of individual CpG sites across the full range of tumor fractions (**Supplementary Table 9c-d**). Remarkably, MHL provides additional gain of signal-to-noise ratio (mean divided by standard deviation) compared with AMF as the fraction of tumor DNA decreases below 10%, which is typical for clinical samples (**Figure 4c**). We then took the individual plasma data sets, and predicted the tumor fraction based on the MHL distribution established by computational mixing (**Figure 4a-b)**. Except for a small number (N<5) of outliers, we observed significantly higher average MHL in cancer plasma than in normal plasma (**Figure 4d**). 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<0.000023, **Supplementary Figure 9 and Supplementary Table 10**).

Recent studies[9](#_ENREF_9),[10](#_ENREF_10),[32](#_ENREF_32) 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 most abundant fraction in cancer plasma (**Supplementary Table 6 and Supplementary Table 7**). 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. We compiled 43 WGBS and RRBS data sets for 10 human normal tissues that have high cancer incident rate, and identified a set of 2,880 tissue-specific MHBs as the candidates (**Supplementary Table 11**). We then used these tissue-specific MHBs or subsets to predict the tissue-of-origin for the cancer plasma samples. Although we found a large number of tissue-of-origin specific MHBs that have low MHL in normal plasma (**Figure 5a**), the multiclass prediction based on random forest yielded limited power, most likely due to the high diversity of the tissue classes (N=10). We then adopted an alternative approach by counting the number of methylated (or high MHL) tissue-specific MHBs in the plasma samples and comparing with all other tissues, to infer the most probable tissue-of-origin. At the cutoff of minimal 10 tissue-specific MHL signals per tissue type, we observed an average 90% accuracy for mapping a data set from the primary tissue to its tissue type (**Figure 5b**). We then applied this method to the full set of plasma data from 59 cancer patients and 75 normal individuals, and achieved an average prediction accuracy of 82.8%, 88.5%, 91.2% for the plasma from [colorectal](https://en.wikipedia.org/wiki/Colorectal_cancer) cancer, lung cancer, and control plasma samples respectively with 5-fold cross-validation (**Figure** **5c, Supplementary Figure 10, Supplementary Table 12**). For the incorrectly classified samples, we noticed that 4 out of 5 colorectal cancer plasma were from metastatic colorectal cancer patients while the fifth was in fact tubular adenoma. In the case of lung cancer, one misclassified sample came from a patient with benign fibrous tissue.

Finally, we asked whether we can combine cancer detection and tissue-of-origin mapping in a unified analytical framework. To this end, we pooled the RRBS and WGBS datasets of three cancer types (8 CRC, 8 LC, 2 KC) as a “pan-cancer tissue”, and included it together with the data sets from 10 normal tissues for tissue-specific MHB identification. Note that 7 CRC, 5 LC, and 2 KC references were from primary tumor tissues while 1 CRC and 3 LC were cancer cell lines. With 200 MHBs specific to each of the 11 reference tissues, we counted the number of MHBs that has high MHL based on an empirically cutoff, then established a background distribution of the counts for each of the 11 tissues based on the 75 normal plasma samples. The number of MHBs with high MHL in cancer plasma clearly has a different distribution (**Figure 5c**, **Supplementary Figure 11a**). For each cancer patient’s plasma, we determined whether there is any enrichment (Z-scores assuming Gaussian distributions, then converted them to negative log P-values) in each of the 11 reference tissues based on the background distribution. We observed consistent enrichments in both the “pan-cancer tissue” and the tissue-of-origin for the cancer plasma (Figure 6c; **Supplementary Figure 11b**), suggesting that tumor growth might lead to the release of cell-free DNA by adjacent normal cells, and that including such signals can boost the accuracy for prediction (**Supplementary Figure 12**). In fact, by integrating both types of signals, we achieved a 99% specificity and 80% sensitivity in predicting cancer, and a 90% accuracy in predicting the tissue-of-origin for the 59 cancer patients. For non-invasive detection of cancer in plasma, the area under the ROC curves was improved from 0.81 for CRC (0.75 for LC) when examining cancer-specific MHBs alone to 0.90 for CRC (0.85 for LC) if normal colon specific MHBs are included (Figure 5d; **Supplementary Figure 11c**). Taken together, we demonstrated for the first time that both tumor load and tissue-of-origin can be quantitatively characterized by methylation haplotype analysis of cell free DNA in plasma, and there is gain of detection accuracy by integrating the two aspects in a unified analytical framework.

## Discussions

In this study we extended a well-established concept in population genetics, linkage disequilibrium, to the analysis of co-methylated CpG patterns. While the mathematical representations are identical, there are two key differences. First, traditional linkage disequilibrium was defined on human individuals in a population, whereas in this study the analysis was performed on the diploid genome of individual cells in a heterogeneous cell population. Second, linkage disequilibrium in human populations depends on the mutation rate, frequency of meiotic recombination, effective population size and demographic history. The LD level decays typically over the range of hundreds of kilobases to megabases. In contrast, CpG co-methylation depends on DNA methytransferases and demethylases, which tend to have lower processivity, and, in the case of hemi-methyltransferases, much lower fidelity compared with DNA polymerases[33](#_ENREF_33). Therefore, methylation LD decays over much shorter distance in tens to hundreds of bases, with the exception of imprinting regions. Even if longer-read sequencing methods were used, we do not expect a radical change of the block-like pattern presented in this work, which is supported by another recent study[34](#_ENREF_34). Nonetheless, these short and punctuated blocks capture discrete entities of epigenetic regulation in individual cells widespread in the human genome. This phenomenon can be harnessed to improve the robustness and sensitivity of DNA methylation analysis, such as the deconvolution of data from heterogeneous samples including circulating cell-free DNA.

While we demonstrated a superior power of MHL over single-CpG methylation level or average methylation level in classification and deconvolution using MHBs as features, the accuracy is slightly less than what has been reported on the deconvolution of blood cell types. One major difference is that each reference tissue type itself is a mixture of multiple cell types that might share various degrees of similarity with another reference tissue type. Furthermore, most solid tissues also contain blood vessels and blood cells. Given such background signals, the accuracy that we achieved is promising, and will be further improved once reference methylomes of pure adult cell types are available.

Practically, the amount of cell-free DNA per patient is rather limited, typically in the range of tens of nanogram. We therefore used 1 to 10 ng per patient for the scRRBS experiment. Considering the material losses during bisulfite conversation and library preparation, as well as the sequencing depth, there were most likely no more than 30 genome equivalents in each data set. Our data set is rather sparse, especially when the fraction of tumor DNA is low. Hence the chance of finding cancer-specific methylation haplotypes in a specific region consistently across many samples is low. This is likely the reason that marker sets selected based on random forest has limited sensitivity and specificity. However, epigenetic abnormalities tend to be more widespread across the genome (compared with somatic mutations), and hence we were able to integrate the sparse coverage across many loci to achieve accurate prediction by direct counting of methylated haplotypes within the appropriate tissue-specific features. Importantly, we showed that, in cancer patients, plasma contains circulating DNA fragments from both normal and malignant cell types, and an integrative analysis of such signatures improved the power of non-invasive detection. Further technical improvements on sample preparation and library construction, will undoubtedly increase the coverage and sensitivity. Finally, with larger sets of plasma samples from healthy controls and more primary tumor tissue samples at different clearly defined cancer stages, we will be able to build models that have more comprehensive coverage of the inter-individual variability, and further improve the specificity/sensitivity to the level adequate for clinical diagnosis.

## Data Availability

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

## Code Availability

All codes and scripts written for this study are released freely for non-commercial use and available as Supplementary Materials.

**Acknowledgements**

This study was supported by NIH grants R01GM097253 (to K.Z.) and P30CA23100. We thank S. Kaushal for managing and handling patient samples in UCSD Moores Cancer Center BTTSR.

## Author’s Contributions

Ku.Z. conceived the initial concept and oversaw the study. S.G., D.D. and Ku.Z. performed bioinformatics analyses. N.P., D.D., and H.F. performed experiments. Ka. Z. contributed normal plasma samples. Ku. Z., S.G. and D.D. wrote the manuscript with inputs from all co-authors.

## Competing Financial interests

A patent application (PCT/US2015/013562) has been filed related to the methods disclosed in this manuscript. Ku. Z. is a co-founder and scientific advisor of Singlera Genomics Inc.

## Abbreviation

MHB: methylation haplotype load; MHL: Methylation Haplotype Load; cf-DNA: cell-free DNA; RRBS: [Reduced representation bisulfite sequencing](http:///h); scRRBS: single-cell reduced-representation bisulfite sequencing; WGBS: genome-wide bisulfite sequencing; TCGA: The Cancer Genome Atlas project; ENCODE: the Encyclopedia of DNA Elements; GEO: Gene Expression Omnibus; LC: Lung Cancer; CRC: Colorectal cancer; ACC: Accuracy; caHMH: cancer associated high methylation haplotype; ts-MHB: tissue specific methylation haplotype block regions. CCT: Colorectal cancer tissue; CCP: colorectal cancer plasma; LCT: lung cancer tissue; LCP: lung cancer plasma; NP: normal plasma.

## Figure legends

**Figure 1**. Identification and characterization of human methylation haplotype blocks (MHBs). (a) Schematic overview of data generation and analysis. (b) An example of MHB at the promoter of the gene APC. (c) Smooth scatterplots of methylation linkage disequilibrium for adjacent CpG sites within MHBs. Red indicate relative higher density and blue indicates relative low density. 500,000 adjacent CpG loci in MHB regions were randomly sampled and the reduction of coupled adjacent CpGs based on r2 was observed from stem and progenitor cells to somatic cells to cancer cells. The yellow dotted lines and percentages highlight the reduction of high linkage disequilibrium (r2>0.9). The samples included in each group are as follows: stem and progenitor cells included 10 samples from 1 other study, somatic cells included 39 samples from 2 other studies and 10 from this study, and cancer cells included 6 samples from 2 other studies. (d) Co-localization of MHBs with known genomic features. Genome distribution (left) and CpG island nearby status show MHB are widely dispersed in the human genome (e). Enrichment of MHBs in known genomic features. Bootstrap random sampling of regions with same size 10,000 times to estimate empirical statistical significance and enrichment factor (fold-change).

**Figure 2**. Comparison of methylation haplotype load with four other metrics used in the literatures. Five patterns of methylation haplotype combinations are used to illustrate the difference between methylation frequency, methylation entropy, epi-polymorphism and methylation haplotype load. Methylation haplotype load can discriminate all the five patterns while other metrics cannot.

**Figure 3**. Tissue clustering based on methylation haplotype load. (a) MHL based unsupervised clustering of human tissues using the top 15% most variable regions. (b) Supervised clustering of germ-layer specific MHBs.(c) MHL exhibits better signal-to-noise ratio than average methylation frequency (AMF) and individual site methylation frequency (IMF) for sample clustering.

**Figure 4**. Quantitative estimation of cancer DNA proportion in cell-free DNA based on MHL of informative MHBs. (a) [Colorectal](https://en.wikipedia.org/wiki/Colorectal_cancer) cancer (b) Lung cancer. Informative MHBs were selected based on the presence of high-MHL in cancer solid tissues (CT) and the absence of MHL in whole blood (WB). Group II regions have high MHL in cancer tissues (MHL>0.5) and cancer plasma while low MHL in WB and normal tissues (MHL<0.1), and hence selected for further analysis. Barplots show average MHL in different groups of samples. MHL level in cancer plasma (CCP and LCP) and normal plasma (NP) were compared with a two-tail t-test. NCT denotes normal colon tissues, NLT denotes normal lung tissues, and ONT denotes other normal tissues. (c) Computational mixtures of cancer and whole blood DNA at different ratios (0.1% to 50%) were created by random sampling of haplotypes in the Group II regions, and repeated 1,000 times to empirically determined the mean and variance of MHL and 5mC levels at different fractions of cancer DNA. (d) After an empirical “standard curve” was constructed, it was used to estimate the cancer DNA proportions in plasma samples. CCP denotes colorectal cancer plasma, LCP denotes lung cancer plasma, and NP denotes normal plasma.

**Figure 5**. Methylation haplotype load based tissue-of-origin prediction from plasma DNA. (a) Detection of tissue-specific MHL in the plasma of cancer patients, but not normal plasma or whole blood. Tissue specific MHL were observed in corresponding tissue and cancer plasma, indicating the possibility for tissue-of-origin mapping. (b) Identification of informative MHBs for tissue prediction. A total of ~300 tissue-specific MHBs were selected with the cutoff GSI>0.1. Training data included WGBS and RRBS datasets. (c) Application of the prediction model to plasma samples from cancer patients and normal individuals. Plasma samples (30 CRC, 29 LC, and 75 NP) were separated into 5 parts (each group) so that 5-fold cross-validation could be applied to measure the stability of the prediction; the number of tissue-specific MHB features were iterated from 50 to 300 and the minimum feature number was selected when accuracy for cancer plasma was above 0.8 and normal plasma above 0.9. ACC denotes accuracy. (d) A unified analytical framework for cancer status and tissue-of-origin joint prediction from plasma. Distribution of Z-score of each reference-specific MHBs for each colon cancer plasma samples (left) and lung cancer plasma samples (right). (e) Prediction performance based on colon cancer tissue and joint prediction by CT and colon tissue specific MHB markers for colon cancer (left) and lung cancer plasma samples (right). Adjusting the cutoff for the Z-scores changed the specificity and sensitivity, resulting in the ROC curves. AUC denotes area under the curve.

## Supplementary Tables:

## Supplementary Table 1: Methylation haplotype block regions (hg19)

## Supplementary Table 2: Gene Ontology of cancer loss linkage regions by GREAT

## Supplementary Table 3: Tissue specific MHBs regions for normal tissue of origin classification

## Supplementary Table 4: Layer specific MHBs regions with group specificity index

## Supplementary Table 5: Gene ontology analysis of TFs binding events

## Supplementary Table 6: Cancer-associated High-Methylation-Haplotype (caHMH) regions

## Supplementary Table 7: Deconvolution of plasma samples by 10 normal tissues, LCT, and CCT

## Supplementary Table 8: Differential MHLs between cancer plasma and normal plasma

## Supplementary Table 9: Estimation of cancer DNA proportion using Group II MHL regions for CRC and LC.

## Supplementary Table 10: Relationship between average MHL and cell-free DNA yield

## Supplementary Table 11: Predictors for colon cancer, lung cancer, and normal plasma tissue

## Supplementary Table 12: Prediction accuracy based on tsMHB counting with 5-fold cross-validation

## Supplementary Table 13: Information of all samples used in this study

## Reference

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

2. Slatkin, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**, 477-85 (2008).

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. Jones, B. DNA methylation: Switching phenotypes with epialleles. *Nat Rev Genet* **15**, 572 (2014).

5. Schwartzman, O. & Tanay, A. Single-cell epigenomics: techniques and emerging applications. *Nat Rev Genet* **16**, 716-26 (2015).

6. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-8 (2010).

7. Jones, P.A. & Martienssen, R. A blueprint for a Human Epigenome Project: the AACR Human Epigenome Workshop. *Cancer Res* **65**, 11241-6 (2005).

8. Houseman, E.A. *et al.* Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics* **17**, 259 (2016).

9. 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).

10. 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).

11. Schultz, M.D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**, 212-6 (2015).

12. Heyn, H. *et al.* Distinct DNA methylomes of newborns and centenarians. *Proc Natl Acad Sci U S A* **109**, 10522-7 (2012).

13. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134-48 (2013).

14. Blattler, A. *et al.* Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. *Genome Biol* **15**, 469 (2014).

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

16. Chen, K. *et al.* Loss of 5-hydroxymethylcytosine is linked to gene body hypermethylation in kidney cancer. *Cell Res* **26**, 103-18 (2016).

17. 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).

18. 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).

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

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

21. 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).

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

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

24. 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).

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

26. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350-4 (2015).

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

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

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

30. Shu, J. *et al.* Induction of pluripotency in mouse somatic cells with lineage specifiers. *Cell* **153**, 963-75 (2013).

31. 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).

32. 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).

33. Williams, K. *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343-8 (2011).

34. Saito, D. & Suyama, M. Linkage disequilibrium analysis of allelic heterogeneity in DNA methylation. *Epigenetics* **10**, 1093-8 (2015).