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

Shicheng Guo[1,3], Dinh Diep[1,3], Nongluk Plongthongkum[1], Ho-Lim Fung[1], Kang Zhang[2], Kun Zhang[1,2*]

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

[3]Equally 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 RRBS and 637 sets of methylation array data. Using a metric called Methylation Haplotype Load (MHL), we performed tissue-specific methylation analysis at the block level. Subsets of informative blocks were further identified for deconvolution of heterogeneous samples. 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] 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], 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, 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 the NIH RoadMap Epigenomics project[3] and the EU Blueprint Epigenome project[4] 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 full-genome 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[5], fetal components in maternal cell-free DNA[6], or circulating tumor DNA in plasma[6]. 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[7]. 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 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[2,8] and the r^2 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 r^2 was calculated from the abundance of different methylation haplotypes (see Methods). 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 r^2 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[9,10], as well as the H1 human embryonic stem cells and *in vitro* derived progenitors[11]. We also included an in-house generated WGBS data set from 10 adult tissues of one human donor. Across this set of 61 samples (>2000x combined genome coverage) we identified a total of ~ 55 billion methylation haplotype informative reads that cover 58.2% of autosomal CpGs. 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**). The regions not covered by such blocks have low CpG density and hence too few CpG sites within Illumina read pairs for deriving informative haplotypes. The majority of CpG sites within the same MHBs are near perfectly coupled (r^2 ~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 (**Figure 1c**). This is consistent to our previous observations on a smaller BSPP data set on 2,020 CpG islands[8] for culture cell lines and another previous report[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 recently reported in VMR[13,14].

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 culture cells, we examined 101 published reduced representation bisulfite sequencing (RRBS) datasets from ENCODE cell lines and tissue samples, as well as 637 sets of Infinium HumanMethylation450 BeadChip (HM450K) data including 11 human normal tissues from TCGA project. 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 sample sets for block partitioning. Note that the presence of such correlated methylation blocks is a necessary but not sufficient condition for MHBs (**Supplementary Fig. 1a**). 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 ENCODE RRBS and TCGA HM450K array data respectively, among which 8,920 and 1,258 have significant overlaps with WGBS-defined MHBs. Additionally, we observed significantly higher correlation among the CpGs within the MHB regions compared CpG loci outside MHBs in HM450K and RRBS dataset, further supporting the block-like organization of local CpG co-methylation across a wide variety of cells and tissues (**Supplementary Fig. 1b**). 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 coupled with the local density of CpG dinucleotides.

Co-localization of methylation haplotype blocks with known regulatory elements. MHBs 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-value<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 LAD[15] and LOCK regions[16] (46% and 37% of the expected values), modest enrichment in TAD[17]. Importantly, we observed a very strong (26-fold) enrichment in variable methylation regions (VMR)[14] (**Figure 1e**), suggesting that increased epigenetic variability in a cell population or tissue can be coordinated locally among hundreds of thousands of genomic regions[18]. We further examined a subset of MHBs that do not overlap with CpG islands, and observed a consistent enrichment pattern (**Figure 1e**), suggesting that local CpG density alone does not account for the enrichment.

Previous studies on mouse and human[19,20] 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]. 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[21] (no peak for H3K4me1) (**Supplementary Fig. 2**). 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). 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 elements 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 cancer WGBS sets[22]) sets from human solid tissues based on the MHL. Unsupervised 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 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 1,360 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 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 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[23], 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 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 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[24]. 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. 4**). 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[25].

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 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 4-122ng (average 20ng) of cell-free DNA from an average of 866µL 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[26] 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 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 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 csHMH in all cancer patient plasma samples (Average=36, IQR=17,

Supplementary Table 5a). These HMHs were 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 that have no matched tumor samples, using 65 normal plasmas as the background. On average 60 (IQR=31) csHMH were identified for each cancer plasma sample (**Supplementary Table 5c**). Interestingly, a significant fraction (35%) of csHMH called on matched tumor-plasma pairs were also detected the expanded set of cancer patient plasma samples.

Next we quantified the tumor load in cancer plasma samples, using non-negative decomposition with quadratic programming, on the RRBS data from primary cancer biopsies (LC & CRC) and WGBS data from 10 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 recently based on shallow whole genome bisulfite sequencing (69.4%)[6]. Primary tumor and normal tissue-of-origin contributed at the similar level of 2.3% (95% CI: 0.4%-4.2%) and 3.0% (95% CI:1.2%-4.4%). In contrast, we applied the similar analysis to normal plasma, and found only residual tumor contributions (0.17% for CRC and 1.0% LC) to normal plasma, 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 certain (low) tumor contribution (**Supplementary Fig. 5**). Therefore, circulating cell-free DNA contains a relatively stable fraction of molecules released from various normal tissues, whereas in cancer patients tumor cells released DNA molecules that can be more abundant than normal tissues (**Supplementary Table 6**).

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 7a-b**). 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 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, 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 signal 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 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 4**). We identified a subset (Group II) of 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 CRC and LC, respectively). By computationally mixing the sequencing reads from cancer tissues and whole blood samples (WBC), 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 8a-b**). 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 (**Supplementary Fig. 6b**). 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 Fig. 7 and Supplementary Table 9**).

Recent studies[6,7,27] 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 5 and Supplementary Table 6**). 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 10**). We then used these tissue-specific MHBs or subsets to predict the tissue-of-origin for the cancer plasma sample. 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 very limited power, most likely due to the high diversity of the tissue classes (N=10). We then adopted an alternative approach by counting the total number of tissue-specific MHBs in the plasma samples and comparing with all other tissues, in order to infer the most probable tissue-of-origin. At the cutoff of minimal 10 tissue-specific methylated haplotypes 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 cancer, lung cancer, and control plasma samples respectively with 5-fold cross-validation (**Figure 5c, Supplementary Fig. 8, Supplementary Table 11**). 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. 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.

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[28]. 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[29]. Nonetheless, these short and punctated blocks capture discrete entities of epigenetic regulation in individual cells widespread in the human genome. Such a 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, 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 very 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 to hundreds of nanogram. We used 1 to 10 ng per patient for the sc-RRBS 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 very accurate prediction by direct counting of methylated haplotypes with the appropriate tissue-specific features. Further technical improvements on sample preparation and library construction, combined with larger sets of patient and normal plasma, will undoubtedly increase the coverage and further improve the specificity/sensitivity to the level required for clinical diagnosis.

Methods

Normal and cancer samples

Ten human primary tissues were purchased from BioChain. Cancer tissue and plasma samples were collected from UCSD Moores Cancer Center and normal plasma samples were obtained from UCSD Shirley Eye center under IRB protocols approved by UCSD Human Research Protections Program (HRPP). All data sets generated in this study or obtained from public databases were listed in **Supplementary Table 12.**

Generation of DNA libraries for sequencing

Extracted genomic DNA were prepared for bisulfite sequencing using published protocols. For whole genome bisulfite (WGBS) and reduced representation bisulfite sequencing (RRBS), the DNA fragments were adapted to barcoded methylated adaptors (Illumina). For WGBS, the adapted DNA were converted using the EZ DNA Methylation Lightning kit (Zymo Research) and then amplified for 10 cycles using iQ SYBR Green Supermix (BioRad). For RRBS, the adapted DNA were converted using the MethylCode™ Bisulfite Conversion kit (Thermo Fisher Scientific) and amplified using the PfuTurboCx polymerase (Agilent) for 12-14 cycles. Libraries were pooled and size selected using

6% TBE polyacrylamide gels. Libraries were sequencing using the Illumina HiSeq platform for paired-end 100 cycles, the Illumina MiSeq platform for paired-end 75 cycles, and the GAIIx (WGBS only) for single-end 36 cycles.

Methylation haplotype blocks (MHB)

Human genome was separated into non-overlapping “sequencible and mappable” segments using a set of in-house generated WGBS data from 10 tissues from a 25-year adult male individual. Mapped reads from WGBS data sets were converted into methylation haplotypes in each segment. Methylation linkage disequilibrium was calculated on the combined methylation haplotypes. We then partitioned each segment into methylation haplotype blocks (MHBs). MHBs were defined as the genomic region in which the r^2 value of two adjacent CpG sites is no less than 0.5. MHB regions inferred by GWBS dataset was also validated by bulk data of methylation level. Takai and Jones's sliding-window algorithm[30] was applied for methylation high linkage regions in HM450K (TCGA) and RRBS (Encode) dataset. Finally, simulation analysis to investigate the relationship between LD and correlation of average 5mC of two CpG loci were conducted based on random sampling different methylation haplotype with 1000 individual and each individual sampling 10 methylation haplotype.

Methylation haplotype load (MHL)

We define a methylated haplotype load (MHL) for each candidate region, which is the normalized fraction of methylated haplotypes at different length:

$$\text{MHL} = \frac{\sum_{i=1}^{l} w_i \times P(MH_i)}{\sum_{i=1}^{l} w_i}$$

$$w_i = i$$

Where l is the length of haplotypes, $P(MH_i)$is the fraction of fully methylated and un-methylated haplotype with i loci. For a haplotype of length L, we considered all the sub-strings with length from 1 to L in this calculation. w_i is the weight for i-locus haplotype. We typically used $w_i = i$ or $w_i = i^2$ to favor the contribution of longer haplotyes. In the present study, $w_i = i$ was applied. Quantile normalization, standardization (scale) as well as the batch effect elimination[31] were applied and the top quantile 15% MHL regions were selected in heatmap analysis to investigate the tissue relationship. The Euclidean distance and Ward.D aggregation were applied in the heatmap plot (R, gplots package).

Developmental germ layers and tissue specific MHB regions.

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

$$\text{GSI} = \frac{\sum_{j=1}^{n} 1 - \frac{log_2\big(MHL(j)\big)}{log_2(MHL_{max})}}{n-1}$$

n indicates the number of the groups. $MHL(j)$ denotes the average of MHL of j^{th} group. $MHL(max)$ denotes the average of MHL of highest methylated group.

Simulation and real-data deconvolution analysis

Deconvolution analysis were conducted by simulation and real-data ways. The deconvolution references were constructed by human normal solid tissues, WBC, colorectal cancer tissues (CCT) and lung cancer tissues (LCT). For the simulation analysis, methylation haplotypes were mixture by CCT and WBC with specific gradients (CCT contents ranging from 0.1% to 50%) and then expected and observed CCT contents were compared. Since our MHL is a non-linear metric when mixing CCT and WBC, we found the deconvolution result is perfect, median root-mean-square-error < 5%, which is within the acceptable region of the deconvolution method[32] when the contribution of colorectal fraction is less than 20% (**Supplementary Fig. 6a**). Tissue specific MHB regions were applied to be the candidate features for deconvolution based on non-negative decomposition with quadratic programming[6,32,33]. Raw MHL signals were applied of logit transform before deconvolution analysis. The contribution of the WBC to cancer plasma, normal plasma samples were estimated. Meanwhile, the contribution of the cancer plasma from CCT and LCT were estimated respectively. Finally, the contribution of CCT and LCT for cancer plasma and normal plasma were compared.

Diagnosis biomarker identification and tissue mapping algorithm for plasma cancer DNA. Tumor specific methylation haplotype blocks based on were identified by 2-tailed t-test with False Discovery Rate (FDR) correction. Other statistical analysis to MHL were also conducted by 2-tailed t-test without explicitly notification. Tumor-of-origin prediction were applied with tissue-specific MHBs counting (MHC) strategy in which the tissue-of-origin of the plasma were assigned to the group for which have maximum tissue-specific MHB fragments (assignment by maximum likelihood). For the detail, In the first stage, the tissue-specific MHBs was identified with WGBS and RRBS dataset from solid tissues in the training samples. Tissue specific MHB regions (each tissue ~ 300 MHBs) were obtained by filtered with the moderate GSI> 0.1 so that we could select the most powerful biomarkers which can be detected in RRBS and GWBS. In the second stage, the built prediction model was validated with our own RRBS dataset which including 30 colorectal cancer plasma, 29 lung cancer plasma and 75 normal plasma samples. In the test dataset, we separated the samples into 5 parts so that 5-fold cross-validation could be applied to measure the stability of the prediction, number of tissue-specific MHB features were iterating from 50 to 300 and the minimum feature number was selected when accuracy for cancer plasma higher than 0.8 and normal plasma higher than 0.9 since we require high specificity in the realistic application in 4-fold samples. The selected number of features and then were used in the remaining samples to measure the accuracy of tissue-mapping. The variations of sensitivity, specificity and accuracy in different subsets of 5-fold cross-variation were quite slight (training dataset standard deviation<0.04 while testing dataset standard deviation<0.14, see supplementary Table 11), indicating the current sample size could provide enough prediction power.

Further method details are available in **Online Supplementary Method section**.

Data Availability

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

Code Availability

All Relevant codes and scripts were attached in the Supplementary with thorough usage while parts of analysis pipeline also shared in on-line method section.

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: Circulating cell-free DNA; RRBS: Reduced representation bisulfite sequencing; 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; csHMH: cancer specific 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. Patterns and distribution of methylation haplotype blocks(MHBs) in the human genome. (a) Schematic overview of data generation and analysis. (b) An example of MHB at the promoter of the gene APC. (c) Distribution of methylation linkage disequilibrium between adjacent CpG sites in stem cells and progenitors (mixture of 10 samples), normal adult tissues (mixture of 49 samples), and primary tumors (mixture of 2 samples). (d) Co-localization of MHBs with known genomic features. (e). Enrichment of MHBs in known genomic features.

Figure 2. Comparison of methylation haplotype load with four metrics used in the literature.

Figure 3. Tissue clustering based on methylation haplotype load. (a) Unsupervised clustering based on MHL grouped human tissues according to the expected similarity. (b) Supervised classification identified germ-layer specific MHBs. (c) MHL exhibit better signal-to-noise ratio than average methylation frequency (AMF) and methylation for all CpG site (MAS) for sample clustering. Note: Tissue specificity value (TSV) was the average MHL for the corresponding tissue specific MHL in the correctly assigned samples, while the background value (BV) were the average MHL in mis-assigned samples. Contrast was defined as the ratio TSV/BV.

Figure 4. Quantitative estimation of tumor load in cell-free DNA based on MHL of informative MHBs. (a) Colorectal cancer (b) Lung cancer. Informative MHBs were selected based on the presence of high-MHL in cancer solid tissues and the absence of MHL in WB.

Figure 5. Methylation Haplotype Load in Cancer Diagnosis and Tumor-of-Origin Deconvolution. (a) Detection of tumor-specific or tissue-specific MHL in the plasmas of cancer patients, but not normal plasma or whole blood. (b) Identification of informative MHBs for tissue prediction. (c) Application of the predictive model to plasma samples from cancer patients and normal individuals.

Supplementary Figure Legends:

Supplementary Figure 1. Validation of MHB with Illumina 450k methylation array and RRBS data. (a) Absolute Pearson's r versus absolute LD r^2 (b) The Pearson's r^2 in RRBS and HM450K were significantly higher in overlapped MHBs with WGBS compared with the MHBs without overlapping with WGBS MHBs

Supplementary Figure 2. Profiles of H3K27ac, H3K4me3 and H3K4me1 over methylation haplotype blocks for 12 adult tissue types. X-axis are distances from the center of methylation haplotype blocks (+/- 1000) and y-axis are the average reads density in RPKM (input normalized reads per kilobase per million).

Supplementary Figure 3. PCA analysis of human tissues and cells based on MHL.

Supplementary Figure 4. Distinctive patterns of functional enrichment for TF associated with MHBs of hypo- or hyper MHL.

Supplementary Figure 5. Estimated tumor fraction for all cancer plasma and normal plasma. CCP denotes colorectal cancer plasma, LCP denotes lung cancer plasma and NP denotes normal plasma.

Supplementary Figure 6. Deconvolution into cancer and normal plasma using non-negative decomposition with quadratic programming. (a) accurate deconvolution when cancer fraction was lower. Red line indicates diagonal line while black line indicates deconvolution result. (b) Cancer fraction estimated by deconvolution analysis to cancer and normal plasma samples.

Supplementary Figure 7. Estimated tumor fraction in plasma is generally correlated with the normalized yield of DNA extraction. CCP denotes colorectal cancer plasma, LCP denotes lung cancer plasma and NP denotes normal plasma.

Supplementary Figure 8. Tissue-specific MHBs counting approach mapping the plasma to its tissue-of-origin. The cancer plasma would carry more tissue-of-origin specific MHBs. CCP denotes colorectal cancer plasma, LCP denotes lung cancer plasma and NP denotes normal plasma.

Supplementary Tables:

Supplementary Table 1. Genome-wide MHBs identified from 65 sets of WGBS data.

Supplementary Table 2. Tissue specific MHBs identified based on tissue specificity index.

Supplementary Table 3. Layer specific MHBs identified based on layer specificity index.

Supplementary Table 4. Complete list for highly methylated haplotype shared by primary cancer tissue and matched plasma for CRC and lung cancer patients.

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 of CRC, LC and normal plasma samples by 10 normal tissues and LCT, CCT

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

Supplementary Table 8. The signal of MHL is higher than average 5mC based on cancer DNA and WB DNA mixture simulation analysis.

Supplementary Table 9. Significant correlation between estimated cancer DNA fraction with cell-free DNA yield from the patients.

Supplementary Table 10. Predictors applied in prediction model from CRC, LC and normal plasma.

Supplementary Table 11. Prediction performance of tissue-specific MHBs counting approach with 5-fold cross-validation.

Supplementary Table 12. Tissue-specific MHBs counting approach with 5-fold cross-validation.

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. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-8 (2010).
4. Jones, P.A. & Martienssen, R. A blueprint for a Human Epigenome Project: the AACR Human Epigenome Workshop. *Cancer Res* **65**, 11241-6 (2005).
5. Houseman, E.A. *et al.* Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics* **17**, 259 (2016).
6. 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).
7. 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).
8. 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).
9. Schultz, M.D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**, 212-6 (2015).
10. Heyn, H. *et al.* Distinct DNA methylomes of newborns and centenarians. *Proc Natl Acad Sci U S A* **109**, 10522-7 (2012).
11. Dixon, J.R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-6 (2015).
12. 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).
13. 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).
14. Hansen, K.D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43**, 768-75 (2011).
15. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-51 (2008).
16. 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).
17. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-80 (2012).
18. Pujadas, E. & Feinberg, A.P. Regulated noise in the epigenetic landscape of development and disease. *Cell* **148**, 1123-31 (2012).
19. 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).
20. Ziller, M.J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477-81

(2013).

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

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

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

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

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

26. 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).

27. 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).

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

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

30. Takai, D. & Jones, P.A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* **99**, 3740-5 (2002).

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

32. Houseman, E.A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).

33. Gong, T. & Szustakowski, J.D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29**, 1083-5 (2013).