Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA

Shicheng Guo[1,5], Dinh Diep[1,5], Nongluk Plongthongkum[1], Ho-Lim Fung[1], Kang Zhang[2–4] & Kun Zhang[1,2]

Adjacent CpG sites in mammalian genomes can be co-methylated owing to the processivity of methyltransferases or demethylases, yet discordant methylation patterns have also been observed, which are related to stochastic or uncoordinated molecular processes. We focused on a systematic search and investigation of regions in the full human genome that show highly coordinated methylation. We defined 147,888 blocks of tightly coupled CpG sites, called methylation haplotype blocks, after analysis of 61 whole-genome bisulfite sequencing data sets and validation with 101 reduced-representation bisulfite sequencing data sets and 637 methylation array data sets. Using a metric called methylation haplotype load, we performed tissue-specific methylation analysis at the block level. Subsets of informative blocks were further identified for deconvolution of heterogeneous samples. Finally, using methylation haplotypes we demonstrated quantitative estimation of tumor load and tissue-of-origin mapping in the circulating cell-free DNA of 59 patients with lung or colorectal cancer.

Mammalian CpG methylation is a relatively stable epigenetic modification, which can be transmitted across cell division[1] through the DNA methyltransferase DNMT1 and dynamically either established or removed by DNMT3A, DNMT3B and the tet methylcytosine dioxygenase proteins (TET1, TET2 and TET3). Due to the locally coordinated activities of these enzymes, adjacent CpG sites on the same DNA molecules can share similar methylation status, although discordant CpG methylation has been observed, particularly in cancer[2]. The theoretical framework of linkage disequilibrium[3], which was developed to model the co-segregation of adjacent genetic variants on human chromosomes in 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[4], epi-alleles[5] or epi-haplotypes[6] have been reported, although at small numbers of genomic regions or limited numbers of cell and tissue types. Recent data production efforts, especially by large consortia[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 whole-genome bisulfite sequencing (WGBS) data generated in this study, allowed us to perform full-genome characterization of locally coupled CpG methylation across the largest set of human tissue types available to date and to 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 analyzing the relative cell composition of heterogeneous samples, such as different white blood cells in whole blood[8], fetal components in maternal circulating cell-free DNA (cfDNA)[9] or circulating tumor DNA (ctDNA) in plasma[9]. Most of these recent efforts rely on the methylation level of individual CpG sites, and they are fundamentally limited by the technical noise and sensitivity in measuring single-CpG methylation. Recently, Lehmann-Werman *et al.* demonstrated superior sensitivity with multi-CpG haplotypes in detecting tissue-specific signatures in cfDNA[10], although this was based on the sparse genome coverage of Illumina 450k methylation arrays (HM450K). Here we performed an exhaustive search of tissue-specific methylation haplotype blocks (MHBs) across the full genome and proposed a block-level metric, termed methylated haplotype load (MHL), for a systematic discovery of informative markers. By applying our analytical framework and identified markers, we demonstrate accurate determination of tissue origin and prediction of cancer status in clinical plasma samples from patients with lung cancer (LC) or colorectal cancer (CRC) (**Fig. 1a**).

RESULTS
Identification 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[3,4] and the r^2 metric to quantify the degree of coupled CpG methylation among different DNA molecules. Methylation status of multiple CpG sites in single- or paired-end Illumina sequencing reads were extracted to form methylation haplotypes,

[1]Department of Bioengineering, University of California at San Diego, La Jolla, California, USA. [2]Institute for Genomic Medicine, University of California at San Diego, La Jolla, California, USA. [3]Shiley Eye Institute, University of California at San Diego, La Jolla, California, USA. [4]Veterans Administration Healthcare System, San Diego, California, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to Kun Zhang (kzhang@bioeng.ucsd.edu).

and pairwise 'linkage disequilibrium' of CpG methylation r^2 was calculated from the fractions of different methylation haplotypes (Online Methods).

We started with 51 sets of published WGBS data from human primary tissues[11,12], the H1 human embryonic stem cells, *in-vitro*-derived progenitors[13] and human cancer cell lines[14,15]. We also included an in-house-generated WGBS data set from ten adult tissues of one human donor. Across these 61 samples (>2,000× combined genome coverage) we identified a total of 771 million methylation haplotype informative reads that covered 58.2% of autosomal CpGs. The uncovered CpG sites were either in regions with low mappability or in CpG-sparse regions in which there were too few CpG sites within the Illumina read pairs to derive informative haplotypes. We partitioned the human genome into blocks of tightly coupled CpG methylation sites (which we refer to as MHBs; **Fig. 1b**), using a r^2 cutoff of 0.5 (**Supplementary Note**). We identified 147,888 MHBs at an average size of 95 bp and a minimum of three 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 Fig. 1a,b** and **Supplementary Table 1**). The majority of CpG sites within the same MHBs were nearly perfectly coupled ($r^2 \sim 1.0$) regardless of the sample type. We found that the fraction of tightly coupled CpG pairs ($r^2 > 0.9$; **Fig. 1c**) slightly decreased over CpG spacing from stem and progenitor cells (94.8%; mostly cultured cells) to somatic cells (91.2%; mixture of primary adult tissues) to cancer cells (87.8%; mixture of CRC tissues and LC cell lines).

The loss of methylation linkage disequilibrium (LD) in cancer cells was validated by another independent WGBS data set from primary kidney cancer tissues[16] (**Supplementary Fig. 2**). Although the WGBS data came from different laboratories, which might have technical differences from batch to batch, we found that that methylation LD extended further over CpG distance in stem and progenitor cells, which is consistent with our previous observations on 2,020 CpG islands[4] for culture cell lines and with another report[17]. Notably, in cancer samples, we observed a reduction of perfectly coupled CpG pairs, which could be related to the pattern of discordant methylation that was recently reported in variable-methylation regions (VMRs)[2,18]. The cancer-specific decayed MHBs were enriched for cancer-related pathways and functions (**Supplementary Table 2**). Nonetheless, the majority of MHBs in cancers still contains tightly coupled CpGs (87.8%), allowing us to harness the pattern for detecting tumors in plasma. We further validated the co-methylation of these MHBs in 101 reduced-representation bisulfite sequencing (RRBS) data sets from the Encyclopedia of DNA Elements (ENCODE) and in 637 HM450K data sets from The Cancer Genome Atlas (TCGA) (**Supplementary Fig. 3** and **Supplementary Note**).

Co-localization of MHBs with known regulatory elements

The MHBs established by the WGBS data represent a distinct type of genomic feature that partially overlaps with multiple known genomic elements (**Fig. 1d**). Among all of the MHBs, 60,828 (41.1%) were located in intergenic regions, whereas 87,060 (58.9%) regions were located in transcribed regions. These MHBs were significantly enriched ($P < 1.0 \times 10^{-6}$) in enhancers, super-enhancers, promoters, CpG islands and imprinted genes. In addition, we observed a modest depletion in the lamina-associated domains (LADs)[19] and the large organized chromatin Lys9 modifications (LOCK) regions[20], as well as a modest enrichment in defined topologically associated domains (TADs)[21]. Notably, we observed a strong (26-fold) enrichment in VMRs (**Fig. 1e**), suggesting that increased epigenetic variability in a cell population or tissue can be coordinated locally among hundreds

of thousands of genomic regions[22]. We further examined a subset of MHBs that did not overlap with CpG islands and observed a consistent enrichment pattern (**Fig. 1e** and **Supplementary Fig. 1c**), suggesting that local CpG density alone does not account for the enrichment. Previous studies on mice and humans[23,24] demonstrated that dynamically methylated regions are associated with regulatory regions, such as enhancer-like regions marked by acetylation on Lys27 of histone H3 (H3K27ac) and transcription-factor-binding sites. In publicly available histone-mapping data for human adult tissues, we found co-localization of MHBs with marks for active promoters (trimethylated Lys4 on histone H3 (H3K4me3) with H3K27ac) but not for active enhancers[25] (no peak for H3K4me1) (**Supplementary Fig. 4**). We found that enhancers tended to overlap with CpG-sparse MHBs, whereas the co-localization with super-enhancers was independent of CpG density (**Supplementary Fig. 1c**). Therefore, MHBs probably capture the local coherent epigenetic signatures that are directly or indirectly coupled to transcriptional regulation.

Block-level analysis using methylation haplotype load

To enable quantitative analysis of the methylation patterns within individual MHBs across many samples, we needed a single metric to define the methylated pattern of multiple CpG sites within each block. Ideally this metric should not only be a function of the average methylation level for all of the CpG sites in the block, but it should also be able to capture the pattern of co-methylation on single DNA molecules. Therefore, we defined MHL as the weighted mean of the fraction of fully methylated haplotypes and substrings at different lengths (i.e., all possible substrings; Online Methods). As compared to the other metrics used in the literature (methylation level, methylation entropy, epi-polymorphism and haplotype counts), the MHL is capable of distinguishing blocks that have the same average levels of methylation but various degrees of coordinated methylation (**Fig. 2**). In addition, the MHL is bounded between 0 and 1, which allows for direct comparison of different regions across many data sets.

We next asked whether treating MHBs as individual genomic features and performing quantitative analysis based on the MHL would provide an advantage over previous approaches that used individual CpG sites or weighted (or unweighted) averaging of multiple CpG sites in certain genomic windows. Therefore, we clustered 65 WGBS data sets (including four additional CRC and LC WGBS data sets[15]) from human solid tissues on the basis of the MHL. Unsupervised clustering with the 15% most-variable MHBs showed that, regardless of the data sources, samples of the same tissue origin clustered together (**Fig. 3a**), whereas cancer samples and stem cell samples showed patterns distinct from those of human adult tissues. Principal component analysis (PCA) on all MHBs yielded a similar pattern (**Supplementary Fig. 5**). To identify a subset of MHBs for effective clustering of human somatic tissues, we calculated a tissue specific index (TSI) for each MHB. Feature selection using random forests analysis[26] identified a set of 1,365 tissue-specific MHBs (**Supplementary Table 3**) that could predict tissue type at an accuracy of 0.89 (95% confidence interval (CI): 0.84–0.93), although several tissue types shared rather similar cell compositions (i.e., muscle versus heart). Using these MHBs, we compared the performance between MHL, average methylation fraction (AMF) in the MHBs and individual CpG methylation fractions (IMFs). MHL and AMF provided similar levels of tissue specificity, whereas the MHL had lower noise (background: 0.29; 95% CI: 0.23–0.35) than that in the AMF (background: 0.4; 95% CI: 0.32–0.48). Clustering based on individual CpGs in the blocks had the worst performance, which might be due to the higher biological or technical variability of individual CpG sites (**Fig. 3c**). Thus, block-level
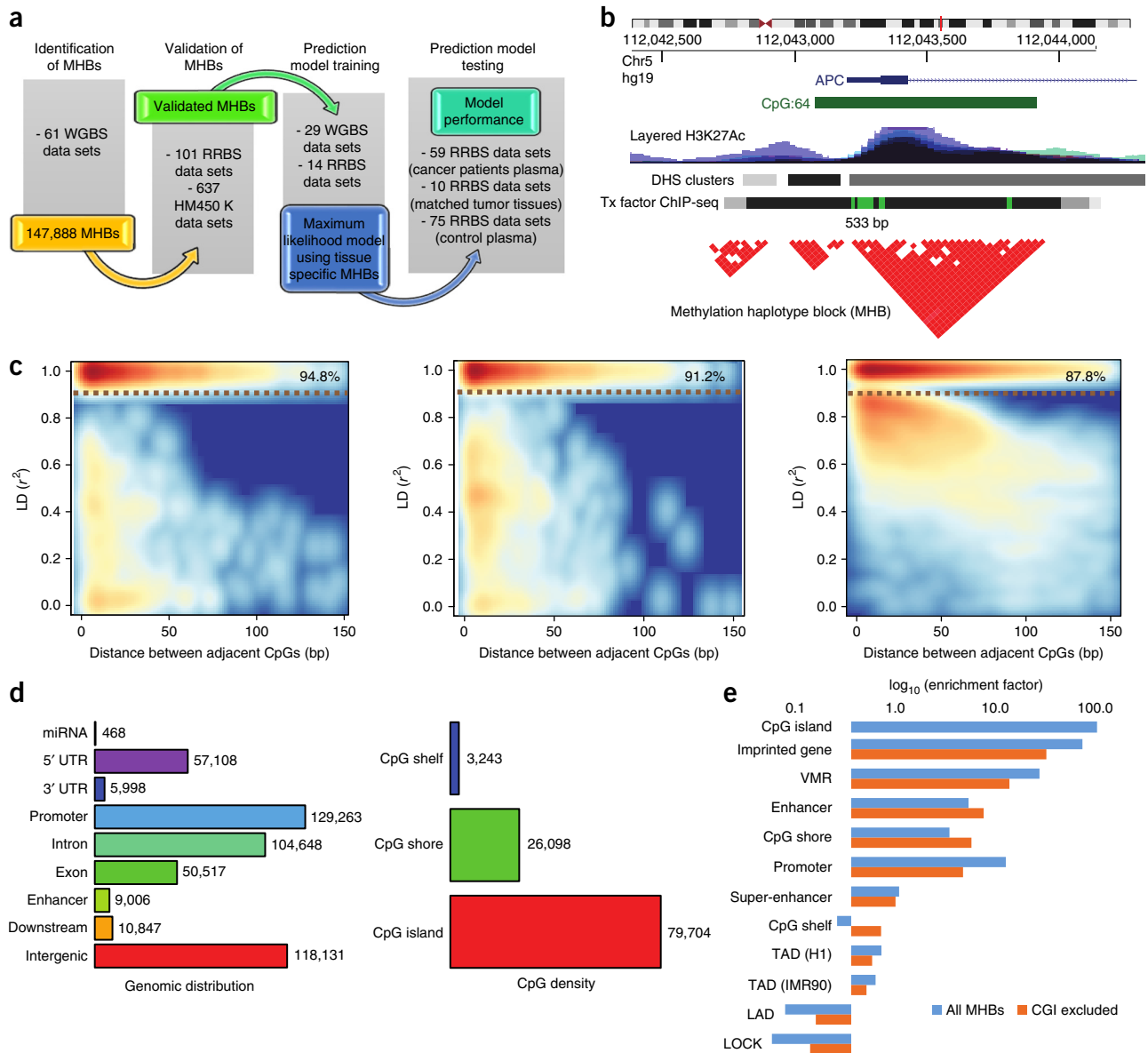
Figure 1 Identification and characterization of human methylation haplotype blocks (MHBs). (**a**) Schematic overview of data generation and analysis. (**b**) An example of an MHB at the promoter of the gene *APC*. Tx, transcription; DHS, DNase-I-hypersensitive sites. (**c**) Smooth scatter plots of methylation linkage disequilibrium within MHBs in stem and progenitor cells (left), somatic cells (middle) and cancer cells (right). Red indicates relative higher density, and blue indicates relative lower density. The yellow dashed lines and percentages highlight the reduction of high LD ($r^2 > 0.9$) with $n = 500,000$ sampling. (**d**) Co-localization of MHBs ($n = 147,888$) with known genomic features. (**e**) Enrichment of MHBs ($n = 147,888$) in known genomic features.

analysis based on the MHL is more advantageous than analysis using single-CpG sites or local averaging of multiple CpG sites in distinguishing tissue types.

The human adult tissues that we used have various degrees of similarity among each other. We hypothesized that this is primarily defined by their developmental lineage and that the related MHBs might reveal epigenetic insights relevant to germ-layer specification. We searched for MHBs that had differential MHLs among the data sets from the three germ layers. In total, we identified 114 ectoderm-specific MHBs (99 hypermethylated and 15 hypomethylated), 75 endoderm-specific MHBs (58 hypermethylated and 17 hypomethylated) and 31 mesoderm-specific MHBs (9 hypermethylated and 22 hypomethylated) (**Supplementary Table 4**). Cluster analysis based on layer-specific

MHBs showed the expected clustering among tissues of the same lineage (**Fig. 3b**). We speculated that some of these MHBs might capture binding events of transcription factors (TFs) specific to the developmental germ layers. We observed patterns of TF binding to layer-specific MHBs that overlapped with ENCODE-reported TF-binding events[27] (**Supplementary Fig. 6**). For layer-specific MHBs with low MHLs, we identified 53 TFs in mesoderm-specific MHBs, 71 TFs in endoderm-specific MHBs and 2 TFs in ectoderm-specific MHBs. Gene ontology analysis showed that mesoderm-specific TF-binding events had negative regulatory activity, whereas endoderm-specific TF-binding events had positive regulatory activity (**Supplementary Table 5**). For layer-specific MHBs with a high MHL, we identified 38 TFs in mesoderm-specific MHBs, 102 TFs in

endoderm-specific MHB and 145 TFs in ectoderm-specific MHBs. Notably, tissues in the ectoderm and endoderm lineage shared few bounded TFs, whereas mesoderm tissue shared multiple groups of TFs with the ectoderm and endoderm tissues. We identified two endoderm-specific high-MHL regions, which are associated with the transcription factors ESRRA (also known as ERR1) and NANOG. This is consistent with a previous finding that mouse embryonic stem cells differentiated spontaneously into visceral and parietal endoderm after knocking out *Nanog*[28]. The low-MHL regions shared by the mesoderm and endoderm might have regulatory functions in the fate commitment toward multiple tissues, whereas the ectoderm-specific high-MHL regions might induce ectoderm development by suppressing the path toward cells of 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 overexpressing lineage specifiers[29].

Methylation-haplotype-based analysis of circulating cfDNA

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, such as incomplete bisulfite conversion or sequencing errors. To explore potential clinical applications, we next focused on the methylation-haplotype analysis of cfDNA from healthy donors and patients with cancer, in which the low amounts of DNA molecules released from tumor cells (which we refer to as cancer DNA fraction) potentially carry epigenetic signatures distinct from those of white blood cells. We isolated cfDNA from the plasma of 75 healthy individuals (NCP), 29 patients with LC (LCP) and 30 patients with CRC (CCP). Owing to the limited amounts of available DNA, we performed single-cell RRBS (scRRBS)[30] and obtained an average of 13 million paired-end 150-bp reads per sample. On average, 57.7% of WGBS-defined MHBs were covered in our RRBS data set from the clinical samples.

We queried the presence of tumor-specific signatures in the plasma samples, using methylation haplotypes identified from tumor tissues as the reference and methylation haplotypes from the NCP samples as negative controls. For five LCP and five CCP samples, we obtained matched primary tumor tissues and generated RRBS data from 100 ng of genomic DNA from the tumor. By focusing on the MHBs with a low MHL in the blood, we identified cancer-associated highly methylated haplotypes (caHMHs). Such haplotypes were present in only the tumor tissues and the matched plasma from the same patient, but not in whole blood or any other non-cancer samples. We found caHMHs in all of the plasma samples from the patients with cancer (average = 36; interquantile range (IQR) = 17; **Supplementary Table 6**). These caHMHs were associated with 183 genes, some of which are known to be aberrantly methylated in human cancers, such as *WDR37*, *VAX1* and *SMPD1* (**Supplementary Table 6**). Next we extended the analysis to 49 additional plasma samples from patients with cancer that had no matched tumor samples, using 75 NCP samples as the background. On average 60 (IQR = 31) caHMHs were identified for each cancer plasma sample (**Supplementary Table 6**). Of note, a substantial fraction (35%) of caHMHs that were called on matched tumor–plasma pairs were also detected in the expanded set of plasma samples from the patients with cancer. Most of the caHMHs were individual specific, whereas several caHMHs were present in at least 53% of CCP and 62% of LCP samples (**Supplementary Fig. 7**). The improvement of sampling depth, by using more input cfDNA or reducing sample loss during the experiments, will likely
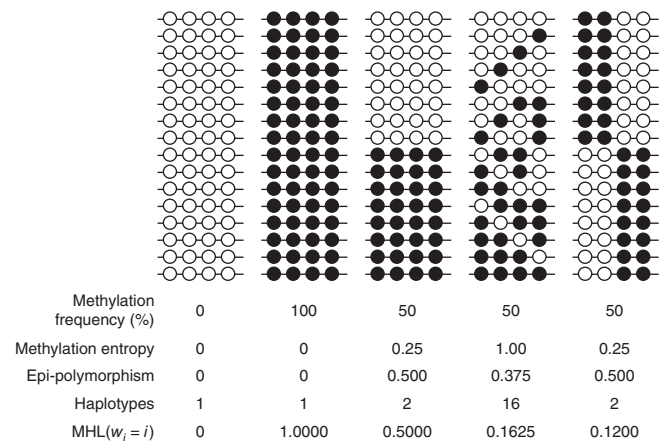

Figure 2 Comparison of methylation haplotype load with four other metrics used in the literature. Five patterns of methylation haplotype combinations (schematic) are used to illustrate the differences between methylation frequency, methylation entropy, epi-polymorphism and MHL. MHL is the only metric that can discriminate all five patterns.

Methylation frequency (%)	0	100	50	50	50
Methylation entropy	0	0	0.25	1.00	0.25
Epi-polymorphism	0	0	0.500	0.375	0.500
Haplotypes	1	1	2	16	2
MHL ($w_i = i$)	0	1.0000	0.5000	0.1625	0.1200

increase the number of caHMHs that are commonly observed in multiple patients.

Next we sought to quantify the cancer DNA fraction in plasma samples from patients with cancer using deconvolution analysis (Online Methods). We used the reference data from the biopsies of primary cancers (LC tissue and CRC tissue) and ten healthy tissues and estimated that a predominant fraction of DNA (72.0%; IQR = 40%) in the plasma from healthy individuals and patients with cancer was contributed by white blood cells, which is consistent with the levels reported recently based on shallow WGBS analysis (69.4%)[9]. DNA from the primary tumor and the healthy tissue of origin was present at similar levels (2.3% (IQR = 3.7%) and 3.0% (IQR = 4.4%), respectively). In contrast, when we applied the same deconvolution analysis to plasma samples from healthy individuals, we found only residual amounts of cfDNA fragments with a tumor signature (0.17% (IQR = 2.9%) for CRC tissue and 1.0% (IQR = 3.1%) for LC tissue), which were significantly lower ($P = 3.4 \times 10^{-5}$ for CRC tissue and $P = 5.2 \times 10^{-10}$ for LC tissue, by two-sample *t*-test) than that of plasma from patients with cancer. We also found that 23/30 CCP samples and only 10/75 NCP samples had detectable contribution from CRC tissue, whereas 26/29 LCP samples and 20/75 NCP samples had detectable contribution from LC tissue (**Supplementary Fig. 8**). Therefore, cfDNA contains a relatively stable fraction of molecules that are released from various normal tissues, whereas tumor cells from patients with cancer released DNA molecules at higher levels than normal tissues (**Supplementary Table 7**). The fractions of white blood cells observed were lower than those reported previously[9] and is likely due to the inclusion of more healthy tissue types as the references (ten instead of four) in the deconvolution analysis.

Next we searched for a small subset of MHBs among all of the RRBS targets that have significantly higher levels of MHL in cancer plasma than in normal plasma. We found 81 and 94 MHBs with markedly higher MHL for CCP and LCP samples, respectively (**Supplementary Table 8**). The majority (71/81 for CCP samples and 83/94 for LCP samples) were also present in at least one of the matched primary tumor–plasma pairs. Some of these regions (such as *HOXA3*) have been shown to be aberrantly methylated in LC and CRC. By using these MHBs as markers, we found that the diagnostic sensitivity was 96.7% and 93.1% for CRC and LC, respectively, at the specificities of 94.6% and 90.6%, respectively. As a comparison, we also performed
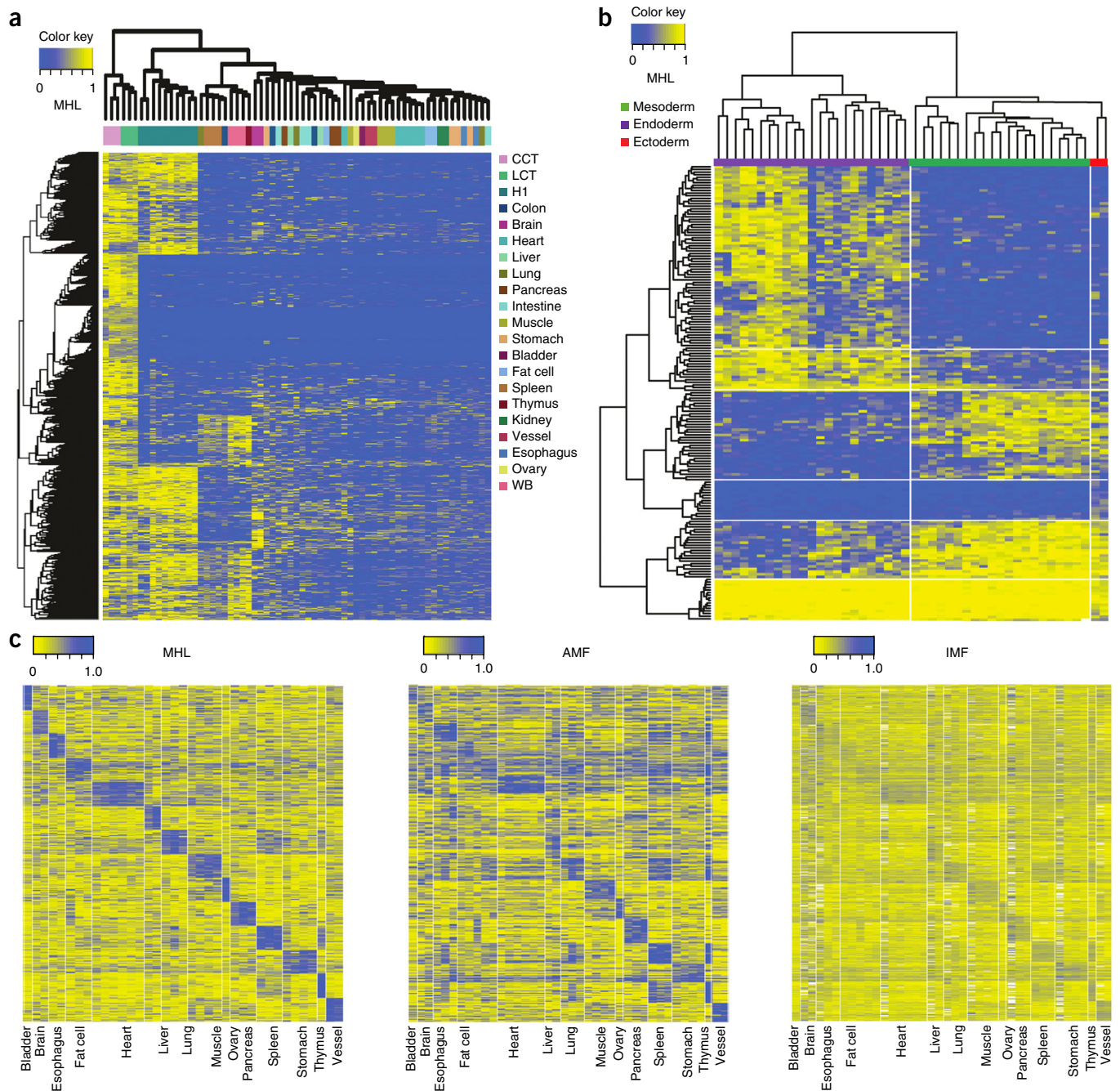
Figure 3 Tissue clustering based on methylation haplotype load. (**a**) MHL-based unsupervised clustering of human tissues using the top 15% most-variable regions. Colored bar indicates the MHL value. (**b**) Supervised clustering of germ-layer-specific MHBs. (**c**) Comparison of cluster performance to different samples using the MHL, AMF and IMF metrics. MHL exhibits a better signal-to-noise ratio than the AMF and IMF for sample clustering.

a prediction based either on the average methylation level within these MHB regions or on single CpG sites. The MHL was found to be superior to the average methylation level (sensitivity of 90.0% and 86.2%, respectively; specificity of 89.3% and 90.6% for CCP and LCP samples, respectively) and to the methylation of individual CpG sites (sensitivity of 89.6% and 80.6%, respectively; specificity of 89.3% and 92.0%, respectively).

We then sought to use the information from healthy human tissues, primary tumor biopsies and cancer cell lines to improve the detection of ctDNA. We started by selecting a subset of MHBs that showed high a MHL (>0.5) in primary cancer biopsies and a low MHL (<0.1) in

whole blood; we then clustered these MHBs into three groups based on the MHLs in the plasma samples from healthy individuals and patients with cancer, as well as the MHLs from cancer and healthy tissues (**Fig. 4a,b**). We identified a subset (group II) of MHBs that had high MHLs in cancer tissues and low MHLs in healthy tissues (**Supplementary Table 9**). Plasma from patients with cancer showed a significantly higher MHL in these regions than plasma from healthy individuals ($P = 1.4 \times 10^{-12}$ for CCP and $P = 6.2 \times 10^{-8}$ for LCP). By computationally mixing the sequencing reads from cancer tissues and whole-blood samples, we created synthetic admixtures at various levels of tumor fraction. We found that the MHL was 2- to 5-fold higher
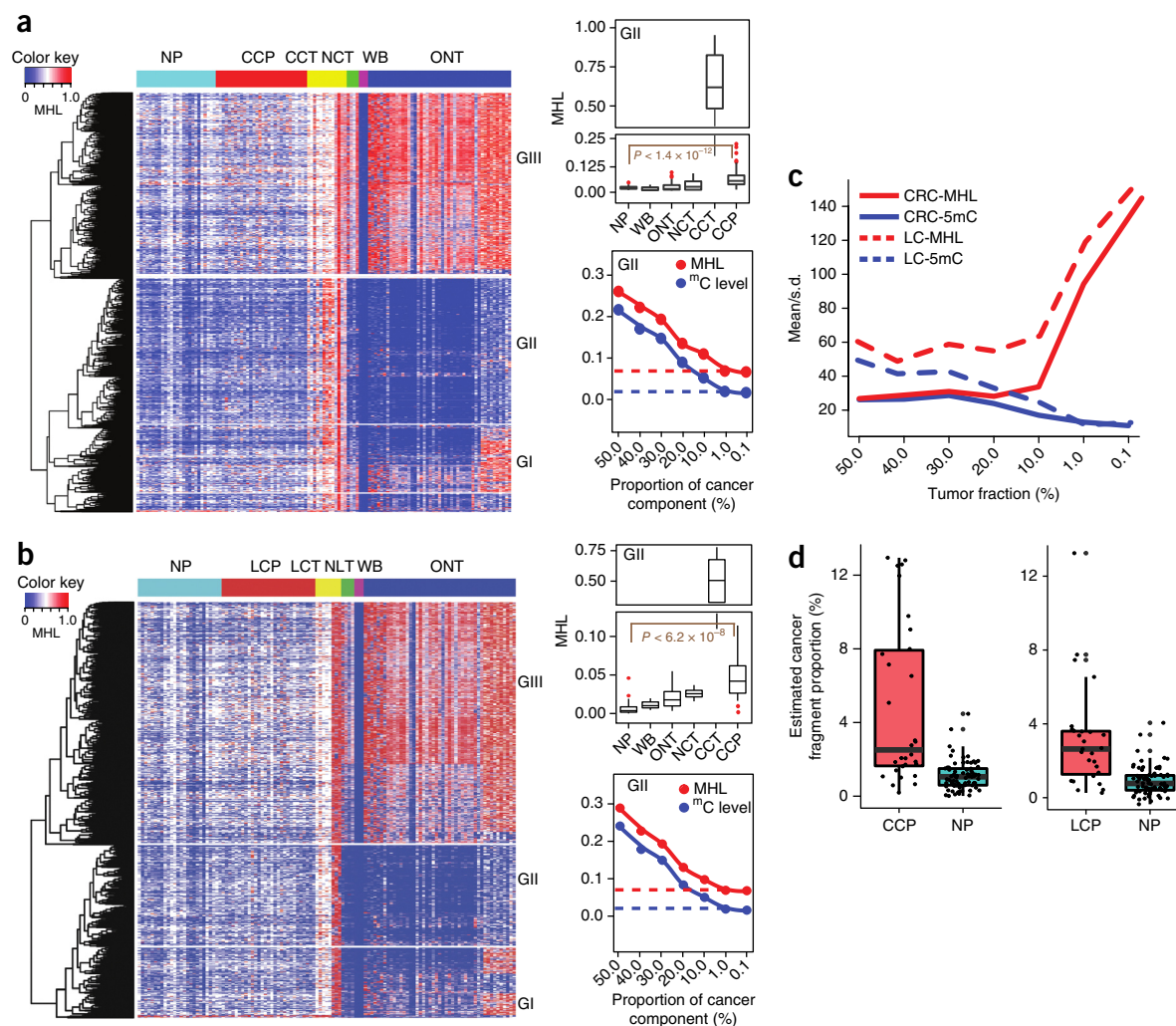
Figure 4 Quantitative estimation of the proportion of DNA derived from cancer cells in cell-free DNA, using the MHL of informative MHBs. (**a**,**b**) Left, heat maps showing the different patterns of MHL in patients with CRC (CCP, $n = 30$; CCT, $n = 13$) (**a**) or LC (LCP, $n = 29$; LCT, $n = 8$) (**b**), as compared to that in healthy individuals (NP, $n = 26$; NLT, $n = 4$; NCT, $n = 4$). Three whole-blood (WB) and 47 'other healthy tissue' (ONT) samples were used in the analysis; a similar number of plasma samples from healthy individuals ($n = 30$) were selected, and some of them ($n = 4$) were excluded because they had >30% missing values. Informative MHBs were selected on the basis of the presence of high MHLs in solid tissues from patients with cancer (CT) and the absence of MHLs in whole blood (WB). Group II (GII) regions have high MHL values in tissues (MHL > 0.5) and plasma from patients with cancer and low MHL values in WB and healthy tissues (MHL < 0.1), and hence, these were selected for further analysis. Bar plots show average MHL values in different groups of samples. MHLs in the plasma of patients with CRC (CCP) or LC (LCP) and in the plasma of healthy individuals (NP) were compared with a two-tailed Student's *t*-test. NCT denotes healthy colon tissues, NLT denotes healthy lung tissues, and ONT denotes other healthy tissues. mC, methylcytosine. (**c**) Comparison between signal-to-noise ratio of MHL and changes in the levels of 5-methylcytosine (5mC) as the reduction of tumor DNA fraction. MHL has higher signal-to-noise ratio (mean/s.d. ratio) than individual 5mC levels as tumor fraction decreases. *x* axis shows the tumor fraction in synthetic mixtures. 30 CRC and 29 LC samples were used in the analysis. (**d**) Estimation of the cancer DNA proportions in CCP ($n = 30$), LC ($n = 29$) and NP ($n = 75$) samples.

than the methylation level of individual CpG sites across the full range of tumor fractions (**Supplementary Table 9**). Notably, the MHL provides an additional gain of signal-to-noise ratio (mean divided by s.d.) than the AMF value as the fraction of tumor DNA decreased to below 10% (**Fig. 4c**), which is typical for clinical samples. We then took the data sets obtained from the individual plasma samples and predicted the tumor fraction on the basis of the MHL distribution established by computational mixing (**Fig. 4a,b**). Except for a small number ($n < 5$) of outliers, we observed significantly higher average MHL values in plasma from patients with cancer than in plasma from healthy individuals (**Fig. 4d**). Note that all group II MHBs were selected without using any information from the plasma samples, and hence these markers should be generally applicable to other

plasma samples. Notably, we also found that the estimated tumor DNA fractions positively correlated with normalized cfDNA yields from the patients with cancer ($P = 0.000023$; **Supplementary Fig. 9** and **Supplementary Table 10**).

Recent studies[9,10,31] have demonstrated that epigenetic information imbedded in cfDNA has the potential for predicting a tumor's tissue of origin. Consistently, we found that tissue-of-origin-derived methylation haplotypes were the most abundant fraction in plasma from patients with cancer (**Supplementary Tables 6** and **7**). To predict the tissue of origin with quantifiable sensitivity and specificity using MHBs, we compiled 43 WGBS and RRBS data sets for ten human tissues types that have high cancer incidence rates and identified a set of 2,880 tissue-specific MHBs (**Supplementary Table 11**).
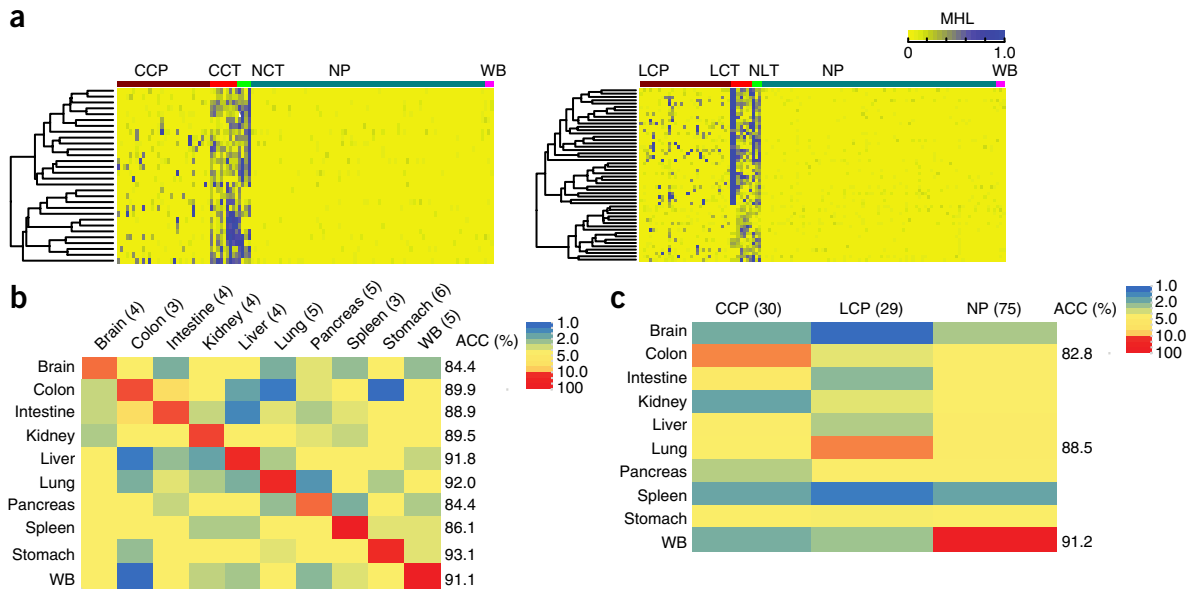
Figure 5 MHL-based prediction of cancer tissue of origin from plasma DNA. (**a**) Detection of tissue-specific MHL in the plasma of patients with cancer but not in plasma or whole blood from healthy individuals. Tissue-specific MHLs were visible in corresponding tissue and plasma samples from patients with cancer, indicating the feasibility for tissue-of-origin mapping. (**b**) Identification of informative MHBs for tissue prediction, using training data included in the WGBS and RRBS data sets from healthy tissues. The color key indicates the tissue-of-origin mapping accuracy (ACC). (**c**) Application of the prediction model to plasma samples from patients with CRC (*n* = 30) or LC (*n* = 29) and from healthy individuals (*n* = 75).

We then used these tissue-specific MHBs or subsets of these to predict the tissue of origin for the plasma samples from the patients with cancer. Although we found many tissue-of-origin-specific MHBs that have low MHLs in the plasma from healthy individuals (**Fig. 5a**), the multiclass prediction based on random forests analysis yielded limited power. This is likely due to the large number of 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 the numbers among all ten tissues to infer the most probable tissue of origin. At the cutoff of a minimum of ten features per tissue type, we observed an average of 90% accuracy for mapping a data set from the primary tissue to its tissue type (**Fig. 5b**). We then applied this method to the plasma data and achieved an average prediction accuracy of 82.8%, 88.5% and 91.2% for the CCP, LCP or NCP samples, respectively, with fivefold cross-validation (**Fig. 5c**, **Supplementary Fig. 10** and **Supplementary Table 12**). The misclassified samples were mainly due to the inclusion of samples with heterogeneous clinical status: four of five CCP samples were from patients with metastatic CRC, whereas the fifth was in fact tubular adenoma; one LCP sample came from a patient with cryptococcal pulmonary infection who later developed lung cancer.

DISCUSSION

Here we extended a well-established concept in population genetics, linkage disequilibrium, to the analysis of co-methylated CpG patterns. Although 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 methyltransferases and demethylases, which tend to have much lower processivity (if any), and, in the case of hemi-methyltransferases, much lower fidelity than DNA polymerases[32]. 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 did not expect a radical change of the block-like pattern presented in this work, which is supported by another recent study[33]. Nonetheless, these short and punctuated blocks capture discrete entities of epigenetic regulation in individual cells that are 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 cfDNA.

Although we demonstrated a superior power of MHL over single-CpG methylation levels or average methylation levels in the classification and deconvolution using MHBs as features, the accuracy was 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 cfDNA per patient is rather limited, typically in the range of tens of nanograms. We therefore used 1–10 ng per patient for the scRRBS experiments. After considering the material losses during bisulfite conversion and library preparation, as well as the sequencing depth, we calculated that there were on average five genome equivalents in each data set. Our data set is very sparse, especially when the fraction of tumor DNA was 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 using random forest analysis had limited sensitivity

and specificity. However, epigenetic abnormalities tend to be more widespread across the genome (as compared to somatic mutations), and hence, we managed to integrate the sparse coverage across many loci to achieve accurate prediction by direct counting of methylated haplotypes within the informative genomic regions. Notably, we showed that, in patients with cancer, plasma contains circulating DNA fragments from both normal and malignant cell types that are detectible with methylation haplotyping. This allowed us to detect the presence of cancer and map the tissue or organ of tumor growth. Of note, when we combined all of the data from primary tumors and cancer cell lines as a 'pan-cancer' tissue and included it as the eleventh reference for tissue-of-origin mapping, the detection sensitivity and specificity was further improved (**Supplementary Figs. 11** and **12**), suggesting that a joint analysis of the cancer signature and the tissue-of-origin signature is more sensitive than focusing on the cancer signature alone. In summary, methylation haplotyping in plasma is a promising strategy for the early detection of a tumor and its primary growth site, as well as for the continuous monitoring of tumor progression and metastasis to multiple organs. With more plasma samples from patients at multiple clearly defined cancer stages and from healthy controls, it is possible to further improve the prediction sensitivity and specificity to a level adequate for clinical testing.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS

Kun Zhang conceived the initial concept and oversaw the study; S.G., D.D. and Kun Zhang performed the bioinformatics analyses; N.P., D.D. and H.-L.F. performed the experiments; Kang Zhang contributed plasma samples from healthy individuals; and Kun Zhang, S.G. and D.D. wrote the manuscript with input from all co-authors.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Wigler, M., Levy, D. & Perucho, M. The somatic replication of DNA methylation. *Cell* **24**, 33–40 (1981).
2. Landau, D.A. *et al.* Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* **26**, 813–825 (2014).
3. Slatkin, M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).
4. 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**, 883–889 (2010).
5. Jones, B. DNA methylation: switching phenotypes with epialleles. *Nat. Rev. Genet.* **15**, 572 (2014).
6. Schwartzman, O. & Tanay, A. Single-cell epigenomics: techniques and emerging applications. *Nat. Rev. Genet.* **16**, 716–726 (2015).
7. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
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 non-invasive prenatal, cancer and transplantation assessments. *Proc. Natl. Acad. Sci. USA* **112**, E5503–E5512 (2015).
10. Lehmann-Werman, R. *et al.* Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc. Natl. Acad. Sci. USA* **113**, E1826–E1834 (2016).
11. Schultz, M.D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**, 212–216 (2015).
12. Heyn, H. *et al.* Distinct DNA methylomes of newborns and centenarians. *Proc. Natl. Acad. Sci. USA* **109**, 10522–10527 (2012).
13. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (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**, 103–118 (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. Hansen, K.D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43**, 768–775 (2011).
19. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
20. 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**, 246–250 (2009).
21. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
22. Pujadas, E. & Feinberg, A.P. Regulated noise in the epigenetic landscape of development and disease. *Cell* **148**, 1123–1131 (2012).
23. 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**, 178–186 (2009).
24. Ziller, M.J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).
25. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350–354 (2015).
26. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
27. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
28. Mitsui, K. *et al.* The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113**, 631–642 (2003).
29. Shu, J. *et al.* Induction of pluripotency in mouse somatic cells with lineage specifiers. *Cell* **153**, 963–975 (2013).
30. 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**, 2126–2135 (2013).
31. 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).
32. Williams, K. *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343–348 (2011).
33. Saito, D. & Suyama, M. Linkage-disequilibrium analysis of allelic heterogeneity in DNA methylation. *Epigenetics* **10**, 1093–1098 (2015).

ONLINE METHODS

Samples from healthy individuals and patients with cancer. Ten human primary tissues were purchased from BioChain Institute, Inc. Tissues from cancer patients and plasma samples were purchased from the UCSD Moores Cancer Center, and plasma samples from healthy individuals were obtained from the UCSD Shirley Eye Center under Institutional Review Board (IRB) protocols approved by UCSD Human Research Protections Program (HRPP). All data sets generated in this study or obtained from public databases are listed in **Supplementary Table 13**.

Generation of DNA libraries for sequencing. Extracted genomic DNA was prepared for bisulfite sequencing, using published protocols. For whole-genome bisulfite sequencing (WGBS) and reduced-representation bisulfite sequencing (RRBS), the DNA fragments were adapted to barcoded methylated adaptors (Illumina). For WGBS, the adapted DNA was converted using the EZ DNA Methylation Lightning kit (Zymo Research) and then amplified for ten cycles using iQ SYBR Green Supermix (Bio-Rad). For RRBS, the adapted DNA was 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 sequenced using the Illumina HiSeq platform for paired ends for 100–111 cycles and the Illumina MiSeq platform for paired ends for 75 cycles.

Read mapping. WGBS and RRBS data were processed in similar fashions. We first trimmed all paired-end (PE) or single-end (SE) fastq files using trim-galore version 0.3.3 to remove low-quality bases and biased read positions. Next, the reads were encoded to map to a three-letter genome via conversion of all cytosines to thymidines or all guanines to adenines for the reads that seemed to be from the reverse-complement strand. Then the reads were mapped using BWA mem version 0.7.5a, with the options '-B2 -c1000' to both the Watson-strand- and Crick-strand-converted genomes. The alignments with mapping-quality scores of <5 were discarded, and only reads with a higher best-mapping-quality score in either the Watson or Crick strand were kept. Finally, the encoded read sequences were replaced by the original read sequences in the final BAM files. Overlapping paired-end reads were also clipped with the bamUtils clipOverlap function.

Methylation haplotype blocks (MHBs). The human genome was split into non-overlapping 'sequenceable and mappable' segments using a set of in-house-generated WGBS data from ten tissues of a 25-year adult male donor. Mapped reads from WGBS data sets were converted into methylation haplotypes within 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.

High-methylation linkage regions defined using ENCODE and TCGA data. We collected RRBS data from the ENCODE project (downloaded from the UCSC Browser) and HM450K data from the TCGA project. Pearson correlation coefficients were calculated between adjacent CpG sites across all samples. The Takai and Jones' sliding-window algorithm[34] was used to identify blocks of highly correlated methylation. We set a 100-bp sliding window starting at a CpG position and moved the window downstream when there were at least two probes in the window. We calculated the total number of probes in the extended regions until the last window that did not meet the criteria. Regions covering at least four probes were defined as CpG-dense regions, and the average Pearson correlation coefficients among all of the probes in the cancer and healthy samples were calculated. We then performed simulation analysis to generate 10,000 virtual individuals with 1,000 haplotypes to investigate the relationship between the LD at the single-read level and the correlation coefficients of the average 5mC levels between two CpG sites, which was based on random sampling of ten different methylation haplotypes from each of the 1,000 simulated individuals.

Enrichment analysis of methylation haplotype blocks for known functional elements. Enrichment analysis was performed by random sampling as previously described[35]. Genomic regions with same number of MHBs (147,888), fragment length distribution and CpG ratios were randomly sampled within the mappable regions (genomic regions meeting Center for Genomic Regulation (CRG) mappability criteria and minimum 10× depth of coverage in our WGBS data set) and repeated 1,000,000 times. Statistical significance was estimated based on the empirical p-value (P). Fold changes (enrichment factors) were calculated as the ratios of observed values to expected values. Exons, introns, 5′ untranslated regions (UTRs) and 3′ UTRs were collected from the UCSC database. The definition of enhancers was based on Andersson et al.[36], the definition of super-enhancers was derived from Hnisz et al.[37], and the definitions of promoter regions was based on the definition by Thurman et al.[38]. All of the genomic coordinates were based on the GRCh37/hg19 human genomic sequence.

Methylation haplotype load (MHL). We defined a methylated haplotype load (MHL) for each candidate region, which is the normalized fraction of methylated haplotypes at different lengths:

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

$$w_i = i$$

where l is the length of haplotypes, and $P(MH_i)$ is the fraction of fully successive methylated CpGs with i loci. For a haplotype of length L, we considered all substrings with length from 1 to L in this calculation. w_i is the weight for i-locus haplotype. Options for weights are $w_i = i$ or $w_i = i^2$ to favor the contribution of longer haplotyes. In the present study, $w_i = i$ was applied.

Following the concept of Shannon entropy $H(x)$, methylation entropy (ME) for the haplotype variable in specific genomic regions were calculated with the following formula:

$$H(x) = -\sum_{i=1}^{l} P(x) \times \log_2 P(x)$$

$$\mathrm{ME} = -\frac{1}{b}\sum_{i=1}^{n} P(H_i) \times \log_2 P(H_i)$$

$$P(H_i) = \frac{h_i}{N}$$

for a genome region with b CpG loci and n methylation haplotype; $P(H_i)$ represents the probability of observing methylation haplotype H_i, which can be calculated by dividing the number of reads carrying this haplotype by the total reads in this genomic region. ME is bounded between 0 and 1, and it can be directly compared across different regions genome wide and across multiple samples. ME values were widely used in the measurement of variability of DNA methylation in specific genomic regions[39].

Epipolymorphism[40] was calculated as

$$\mathrm{ppoly} = 1 - \sum_{i=1}^{n} P_i^2$$

where P_i is the frequency of epi-allele i in the population (with 16 potential epi-alleles representing all possible methylation states of the set of four CpGs).

Developmental germ layers and tissue-specific MHBs. To investigate the germ-layer and tissue-specific MHBs, the group-specific index (GSI) was defined. An empirical threshold GSI > 0.6 was used to define layer- and tissue-specific MHBs. Layer-specific MHBs were selected again to show the ability to distinguish different development layers. Tissue-specific MHBs were further used for tissue mapping and cancer diagnosis.

$$\mathrm{GSI} = \frac{\sum_{j=1}^{n} 1 - \dfrac{\log_2(\mathrm{MHL}(j))}{\log_2(\mathrm{MHL}_{max})}}{n - 1}$$

where n indicates the number of the groups, $\mathrm{MHL}(j)$ denotes the average of the MHL of jth group, and MHL_{max} denotes the average of MHL of the highest-methylated group.

Genome-wide methylation haplotype load matrix analysis. The MHL was calculated for all MHBs on each sample. The MHBs with the top 15% MHL were included in the heat map to investigate the tissue relationship. The Euclidean distance and Ward.D aggregation were used in the heat map plot (R, gplots package; https://cran.r-project.org/package=gplots). PCA (R package prcomp; https://www.r-project.org/) was conducted with the default setting of the corresponding R packages. Before the PCA, raw data were quantile-normalized within same tissue and cell groups. Standardization (scale) and batch-effect elimination (the Combat algorithm[41]) were also applied to decrease the random noise. MAF and IMF were extracted from BAM files with customized PileOMeth software (https://github.com/dpryan79/PileOMeth). Differential MHL analysis between plasma from patients with cancer and healthy individuals was based on a two-tailed Student's t-test or a Wilcoxon rank-sum test. Correction for multiple testing was based on the false discovery rate (FDR). Statistical variations were estimated among different groups, and therefore, one-way analysis of variance (ANOVA) analysis could be conducted.

Deconvolution analysis of simulated and experimental data. Deconvolution analysis was performed on simulated and experimental data sets. The deconvolution references were constructed on data from healthy human primary tissues, whole blood, CRC tissues and LC tissues. For the simulation analysis, methylation haplotypes from CRC tissue and whole blood were randomly mixed to generate a series of synthetic data sets with CRC tissue factions ranging from 0.1% to 50%. We then plotted the expected and observed CRC tissue factions. Although MHL is a nonlinear metric, when mixing CRC tissue and whole blood, we found that the deconvolution result was accurate after log-transformation of the MHL values (median root-mean-square error < 5%), which was within the acceptable region of the deconvolution method[43] when the contribution of colorectal fraction was less than 20%. Tissue-specific MHBs were selected features for deconvolution based on non-negative decomposition with quadratic programming[9,42,43]. MHL values were log-transformed before deconvolution.

Highly methylated haplotype in plasma from individuals with cancer and in normal tissues. Highly methylated haplotype (HMH) was defined as the methylation haplotype that have at least two methylated CpGs in the haplotype. Cancer-associated highly methylated haplotypes (caHMHs) were the ones found in only plasma samples from patients with cancer but were absent in the plasma samples from healthy individuals and in healthy tissues. For the analysis of matched tumor–plasma data from the same individuals, caHMHs were the HMHs present in both the plasma and matched primary cancer tissues, but absent in all healthy samples. In the analysis of plasma samples with no matched primary tumor tissue, we identified caHMHs by subtracting HMHs found in plasma from patients with cancer from those present in all healthy tissues and plasma samples from healthy individuals.

Simulation of MHL in plasma mixture, and comparison between MHL and 5mC in the plasma mixture. In evaluating caHMHs as potential markers for non-invasive diagnosis, we hypothesized that cfDNA in plasma is a mixture of DNA fragments from cancer cells and whole-blood cells at different ratios (cancer DNA fragments from 0.1% to 50%). We created synthetic mixtures by random sampling of haplotypes in the group II regions from cancer and whole-blood data sets at different ratios, and we repeated the sampling analysis described above 1,000 times to empirically determine the mean and variance of MHL and 5mC levels at different fractions of cancer DNA. Once an empirical 'standard curve' was constructed, we then used it to estimate the fraction of cancer DNA in the plasma samples. In addition, we assessed the relationship between estimated cfDNA fraction and log-transformed normalized plasma cfDNA yield by linear regression. Signal-to-noise ratio to MHL and 5mC was conducted with the 1,000-time sampling procedure, and then the average estimated tumor fraction, as well as the variation (s.d.), were recorded, and the ratio was calculated to measure the performance of the metric.

Mapping cancer tissue of origin with plasma DNA. The workflow for data analysis is illustrated in **Supplementary Figure 13**. Tissue-specific MHBs (tsM-HBs) were identified by a two-tailed t-test with FDR correction. Additional statistical analyses with MHL were also conducted by two-tailed t-tests unless stated explicitly. Predictions on plasma from individuals with CRC or LC, or from plasma from healthy individuals, were performed with random forecast, so test and validation samples were independent. Tissue-of-origin prediction was performed using a tsMHB counting strategy, in which the tissue of origin of the plasma was assigned to the reference group with the maximum number of tsMHB fragments (assignment by maximum likelihood). Specifically, in the first stage, the tissue-specific MHBs were identified with the WGBS and RRBS data sets from solid tissues in the training samples. tsMHBs (each tissue has ~300 MHBs) were identified with the cutoff GSI > 0.1. In the second stage, the predictions were validated with our own RRBS data set, which included plasma samples from 30 patients with CRC, 29 patients with LC and 75 healthy individuals. In the test data set, we separated the samples into five parts so that fivefold cross-validation could be applied to estimate the stability of the prediction, and the number of tissue-specific MHB features were iterated from 50 to 300. The minimum number of features was selected when the accuracy for plasma from patients with cancer was >0.8 and the accuracy for plasma from healthy individuals was >0.9, as we required high specificity in clinical applications. The selected number of features were used in the remaining samples to measure the accuracy of tissue-mapping. The variations of sensitivity, specificity and accuracy in different subsets of fivefold cross-variation were low (training data set s.d. < 0.04, and testing data set s.d. < 0.14).

Joint analysis of tumor and normal tissue for non-invasive cancer detection in plasma. Cancer-specific markers (GSI scores derived from eight CRC, eight LC and two kidney cancer (KC) tumor samples) and tissue-specific markers were integrated and considered as a 'pan-cancer tissue'; then, together with the data sets from ten normal tissues, they were applied for tissue- and reference-specific MHB identification. The top 200 MHBs specific to each of the 11 reference tissues were selected as the prediction features. The distribution for the reference-specific MHBs in plasma samples from 75 healthy individuals, 30 patients with CRC and 29 patients with LC were constructed for 11 references. The P value of each reference in the plasma could be inferred by comparison with the background distribution of the reference in plasma from healthy individuals. Meanwhile, tissue of origin was assigned by maximum z-scores among different references. With leave-one-out cross-validation on plasma from healthy individuals, the type-1 error (FDR) for the corresponding z-score threshold and sensitivity were estimated. Finally, setting a predefined z-score threshold could be also used for tissue-of-origin assignment; meanwhile, a receiver-operating-characteristic (ROC) curve was built to show the performance of the predictors.

Code availability. All codes and scripts developed for this study are available for non-commercial use at http://genome-tech.ucsd.edu/public/MONOD_NG_TR44413/.

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

34. Takai, D. & Jones, P.A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA* **99**, 3740–3745 (2002).
35. Timmons, J.A., Szkop, K.J. & Gallagher, I.J. Multiple sources of bias confound functional enrichment analysis of global '-omics' data. *Genome Biol.* **16**, 186 (2015).
36. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
37. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
38. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
39. Xie, H. *et al.* Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Res.* **39**, 4099–4108 (2011).
40. Landan, G. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.* **44**, 1207–1214 (2012).
41. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
42. Houseman, E.A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
43. Gong, T. & Szustakowski, J.D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-seq data. *Bioinformatics* **29**, 1083–1085 (2013).