

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

3
4 Shicheng Guo^{1,3}, Dinh Diep^{1,3}, Nongluk Plongthongkum¹, Ho-Lim Fung¹, Kang Zhang², Kun
5 Zhang^{1,2*}

6
7
8 ¹Department of Bioengineering, ²Institute for Genomic Medicine, University of California at San
9 Diego, La Jolla, California, USA.

10
11 ³Equally contributed authors.

12
13 *Corresponding authors:

14 Kun Zhang, Email: kzheng@bioeng.ucsd.edu

15
16 Keywords: Methylation haplotype, epigenetic heterogeneity, circulating cell-free DNA

17
18 **Abstract**

19 Adjacent CpG sites in mammalian genomes can be co-methylated due to the processivity of
20 methyltransferases or demethylases. Yet discordant methylation patterns have also been observed,
21 and found to be related to stochastic or uncoordinated molecular processes. Here we focused on a
22 systematic search and investigation of regions in the full human genome that exhibit highly
23 coordinated methylation. We defined 147,888 blocks of tightly coupled CpG sites, called Methylation
24 Haplotype Blocks (MHBs), in the human genome with 61 sets of whole genome bisulfite sequencing
25 (WGBS) data, and further validated with 101 sets of RRBS and 637 sets of methylation array data.
26 Using a metric called Methylation Haplotype Load (MHL), we performed tissue-specific methylation
27 analysis at the block level. Subsets of informative blocks were further identified for deconvolution of
28 heterogeneous samples. Finally, we demonstrated quantitative estimation of tumor load and tissue-
29 of-origin mapping in the circulating cell-free DNA of 59 cancer patients using methylation haplotypes.
30

31 **Introduction**

32 CpG methylation in mammalian genomes is a relatively stable epigenetic modification, which can be
33 transmitted across cell division¹ through DNMT1, and dynamically established, or removed by
34 DNMT3 A/B and TET proteins. Due to the processivity of some of these enzymes, physically
35 adjacent CpG sites on the same DNA molecules can share similar methylation status, although
36 discordant CpG methylation has also been observed, especially in cancer cells. The theoretical
37 framework of linkage disequilibrium², which was developed to model the coordinated segregation of
38 adjacent genetic variants on human chromosomes among human populations, can be applied to the
39 analysis of CpG co-methylation in cell populations. A number of studies related to the concepts of
40 methylation haplotypes, epi-alleles, or epi-haplotypes have been reported, albeit at small numbers
41 of genomic regions or limited numbers of cell/tissue types. Recent data production efforts, especially
42 by large consortia such as the NIH RoadMap Epigenomics project³ and the EU Blueprint
43 Epigenome project⁴ have produced a large number of whole-genome, base-resolution bisulfite
44 sequencing data sets for many tissue and cell types. These public data sets, in combination with

45 additional WGBS data generated in this study, allowed us to perform full-genome characterization of
46 local coupled CpG methylation across the largest set of human tissue types available to date, and
47 annotate these blocks of co-methylated CpGs as a distinct set of genomic features.

48
49 DNA methylation is cell-type specific, and the pattern can be harnessed for deconvoluting the
50 relative cell composition of heterogeneous samples, such as different white blood cells in whole
51 blood⁵, fetal components in maternal cell-free DNA⁶, or circulating tumor DNA in plasma⁶. Most of
52 these recent efforts relies on the methylation level of individual CpG sites, and are fundamentally
53 limited by the technical noise and sensitivity in measuring single CpG methylation. Very recently,
54 Lehmann-Werman et al demonstrated a superior sensitivity with multi-CpG haplotypes in detecting
55 tissue-specific signatures in circulating DNA⁷. The markers in that study were discovered from
56 Infinium 450k methylation array data, which represent only a very limited fraction of the human
57 genome. Here we performed an exhaustive search of tissue-specific methylation haplotype blocks
58 across the full genome, and proposed a block-level metric, termed methylated haplotype load (MHL),
59 for a systematic discovery of informative markers. Applying our analytic framework and identified
60 markers, we demonstrated accurate determination of tissue origin as well as estimation of tumor
61 load in clinical plasma samples from patients of lung cancer (LC) and colorectal cancer (CRC)
62 (**Figure 1a**).

63 Results

64 **Identification and characterization of methylation haplotype blocks.** To investigate the co-
65 methylation status of adjacent CpG sites along single DNA molecules, we extended the concept of
66 genetic linkage disequilibrium^{2,8} and the r^2 metric to quantify the degree of coupled CpG methylation
67 among different DNA molecules of the same samples. CpG methylation status of multiple CpG sites
68 in single- or paired-end Illumina sequencing reads were extracted to form methylation haplotypes,
69 and pairwise “linkage disequilibrium” of CpG methylation r^2 was calculated from the abundance of
70 different methylation haplotypes (see Methods). We then partitioned the full human genome into
71 blocks of tightly coupled CpG methylation sites, which we called Methylation Haplotype Blocks
72 (MHBs, **Figure 1b**), using a r^2 cutoff of 0.5. Similar to the partitioning of genetic haplotype blocks,
73 slightly different cutoff values, such as 0.3 or 0.7, resulted in only minor quantitative differences in
74 the block size and number without affecting the global pattern (data not shown).

75
76 To characterize the global pattern and distribution of MHBs, we started with 51 sets of published
77 Whole Genome Bisulfite Sequencing (WGBS) data from human primary tissues^{9,10}, as well as the
78 H1 human embryonic stem cells, *in vitro* derived progenitors¹¹ and human cancer cell line^{12,13}. We
79 also included an in-house generated WGBS data set from 10 adult tissues of one human donor.
80 Across this set of 61 samples (>2000x combined genome coverage) we identified a total of ~ 55
81 billion methylation haplotype informative reads that cover 58.2% of autosomal CpGs. The
82 uncovered CpG sites were either in regions with low mappability, or CpG sparse regions where
83 there are too few CpG sites within Illumina read pairs for deriving informative haplotypes. We
84 identified 147,888 MHBs at the average size of 95bp and minimum 3 CpGs per block, which
85 represents ~0.5% of the human genome that tends to be tightly co-regulated on the epigenetic
86 status at the level of single DNA molecules (**Supplementary Table 1a**, **Supplementary Fig. 1ab**).
87 The majority of CpG sites within the same MHBs are near perfectly coupled ($r^2 \sim 1.0$) regardless of
88 the sample type. We found that methylation LD extends further along the DNA in stem cells and
89 progenitors, compared with normal adult tissue, both in the fraction of tightly coupled CpG pairs
90 (94.8% versus 91.2%, P-value < 2.6x10⁻¹⁶), and the over-representation of partially coupled CpG

91 pairs that are over 100 bp apart while the linkage was slightly decayed in primary cancer dataset
92 (87.8%, mixture of CRC and LC) and was validated by another independent WGBS data from
93 kidney cancer¹⁴ (**Figure 1c, Supplementary Fig. 2**). Gene Ontology analysis shown cancer loss
94 linkage regions were significantly associated with number of cancer related pathway and functions
95 (**Supplementary Table 1b**). This is consistent to our previous observations on a smaller BSPP data
96 set on 2,020 CpG islands⁸ for culture cell lines and another previous report¹⁵. Interestingly, in tumor
97 samples, we observed a reduction of perfectly coupled CpG pairs, which could be related to the
98 pattern of discordant methylation recently reported in VMR^{16,17}.

99

100 While WGBS data allowed us to unbiasedly identify MHBs across the entire genome, the 61 sets of
101 data did not represent the full diversity of human cell/tissue types. To validate the presence of
102 MHBs in a wider range of human tissues and culture cells, we examined 101 published reduced
103 representation bisulfite sequencing (RRBS) datasets from ENCODE cell lines and tissue samples,
104 as well as 637 sets of Infinium HumanMethylation450 BeadChip (HM450K) data including 11
105 human normal tissues from TCGA project. The ENCODE RRBS data sets were generated with
106 short (36bp) Illumina sequencing reads, greatly limiting the length of methylation haplotypes that
107 can be called. Similarly, Illumina methylation arrays only report average CpG methylation of all DNA
108 molecules in a sample, preventing a methylation linkage disequilibrium analysis. Therefore, we
109 calculated the pairwise correlation coefficient of adjacent CpG methylation levels across different
110 sample sets for block partitioning. Note that the presence of such correlated methylation blocks is a
111 necessary but not sufficient condition for MHBs (**Supplementary Fig. 3a**). Nonetheless, the
112 absence of correlated methylation blocks in these data would invalidate the pattern of MHBs. We
113 identified 23,517 and 2,212 correlated methylation blocks from ENCODE RRBS and TCGA
114 HM450K array data respectively, among which 8,920 and 1,258 have significant overlaps with
115 WGBS-defined MHBs. Additionally, we observed significantly higher correlation among the CpGs
116 within the MHB regions compared CpG loci outside MHBs in HM450K and RRBS dataset, further
117 supporting the block-like organization of local CpG co-methylation across a wide variety of cells and
118 tissues (**Supplementary Fig. 3b**). Taken together, the MHBs that we identified represent a distinct
119 class of genomic feature where local CpG methylation is established or removed in a highly
120 coordinated manner at the level of single DNA molecules, presumably due to the processive
121 activities of the related enzymes coupled with the local density of CpG dinucleotides.

122

123 **Co-localization of methylation haplotype blocks with known regulatory elements.** The MHBs
124 established by 61 sets of WGBS data appear to represent a distinct type of genomic feature that
125 partially overlaps with multiple well-documented genomic elements (**Figure 1d, Supplementary**
126 **Fig.1**). Among all the methylation blocks, 60,828 (41.1%) were located in intergenic regions while
127 87,060 (58.9%) regions in transcribed regions. These MHBs were significantly (p-value<10⁻⁶)
128 enriched in enhancers (enrichment factor=7.6), super enhancers (enrichment factor=2.3), promoter
129 regions (enrichment factor=14.5), CpG islands (enrichment factor=70.4) and imprinted genes
130 (enrichment factor=54.6). In addition, we observed modest depletion in LAD¹⁸ and LOCK regions¹⁹
131 (46% and 37% of the expected values), modest enrichment in TAD²⁰. Importantly, we observed a
132 very strong (26-fold) enrichment in variable methylation regions (VMR)¹⁷ (**Figure 1e**), suggesting
133 that increased epigenetic variability in a cell population or tissue can be coordinated locally among
134 hundreds of thousands of genomic regions²¹. We further examined a subset of MHBs that do not
135 overlap with CpG islands, and observed a consistent enrichment pattern (**Figure 1e**), suggesting
136 that local CpG density alone does not account for the enrichment.

137

138 Previous studies on mouse and human^{22,23} demonstrated that dynamically methylated regions were
139 associated with regulatory regions such as enhancer-like regions marked by H3K27ac and
140 transcription factor binding sites. In human, 21.8% of autosomal CpGs were found to be
141 differentially methylated across 30 human cell and tissue types¹⁷. These CpGs were enriched at low
142 to intermediate CpG density promoters. Using publicly available histone mapping data for human
143 adult tissues, we found co-localization of methylation haplotype blocks with marks for active
144 promoters (H3K4me3 with H3K27ac), but not for active enhancers²⁴ (no peak for H3K4me1)
145 (**Supplementary Fig. 4**). Meanwhile, we found that enhancers tend to overlap with CpG sparse
146 MHBs, whereas the overlap with superenhancers were independent of CpG density
147 (**Supplementary Fig. 1c**). Therefore, MHBs likely capture the local coherent epigenetic signatures
148 that are directly or indirectly coupled with transcriptional regulation.
149

150 **Block-level analysis of human normal tissues and stem cell lines with methylation haplotype**
151 **load.** To enable quantitative analysis of the methylation patterns within individual MHBs across
152 many samples, we need a single metric to define the methylated pattern of multiple CpG sites within
153 each block. Ideally this metric is not only a function of average methylation level for all the CpG sites
154 in the block, but also can capture the pattern of co-methylation on single DNA molecules. For this
155 purpose, we defined Methylation Haplotype Load (MHL), which is a weighted mean of the fraction of
156 fully methylated haplotypes and substrings at different lengths (i.e. all possible substrings).
157 Compared with other metrics used in the literature (methylation level, methylation entropy, epi-
158 polymorphism and haplotypes counts), MHL is capable of distinguishing blocks that have the same
159 average methylation but various degrees of coordinated methylation (**Figure 2**). In addition, MHL is
160 bounded between 0 and 1, which allows for direct comparison of different regions across many data
161 sets without normalization.
162

163 We next asked whether treating MHBs as individual genomic elements and performing quantitative
164 analysis based on MHL would provide an advantage over previous approaches using individual
165 CpG sites or weighted (or unweighted) averaging of multiple CpG sites in certain genomic windows.
166 To this end, we sought to cluster 65 WGBS data (including 4 additional **colon and lung** cancer
167 WGBS sets²⁵, **Supplementary Table S12**) sets from human solid tissues based on the MHL.
168 Unsupervised clustering with the top 15% most variable MHBs showed that, regardless of the data
169 sources, samples of the same tissue origin clustered together (**Figure 3a**), while cancer samples
170 and stem cell samples exhibit distinct patterns from adult human somatic tissues. PCA analysis on
171 all MHBs genome-wide yielded a similar pattern (**Supplementary Fig. 5**). To identify a subset of
172 MHBs for effective clustering of human somatic tissues, we constructed a tissue specific index (TSI)
173 for each MHB (see Methods). Random Forest based feature selection identified a set of 1,360
174 tissue-specific MHBs (**Supplementary Table 2**) that can predict tissue type at an accuracy of 0.89
175 (95%CI: 0.84-0.93), despite the fact that several tissue types share rather similar cell compositions
176 (i.e. muscle vs. heart). Using this set of MHBs, we compared the performance between MHL,
177 average methylation fraction in the MHL regions (AMF) and all individual CpG methylation fraction
178 (IMF). MHL and the average methylation provided similar tissue specificity, while MHL has a lower
179 noise (background noise: 0.29, 95%CI: 0.23-0.35) compared with average methylation (background
180 noise: 0.4, 95%CI: 0.32-0.48). Clustering based on individual CpGs in the blocks has the worst
181 performance, which might be due to higher biological or technical viability of individual CpG sites
182 (**Figure 3c**). Thus block-level analysis based on MHL is advantageous over single CpG or local
183 averaging of multiple CpG sites in distinguishing tissue types from regions of coupled CpG
184 methylation and heterogeneity.

185 The human adult tissues that we used in this study have various degrees of similarity amongst each
186 other. We hypothesize that this is primarily defined by their developmental lineage, and that the
187 related MHBs might reveal epigenetic insights related to germ layer speciation. We grouped all the
188 data sets based on the three germ layers, and searched for MHBs that have differential MHL. In
189 total we identified 114 ectoderm-specific MHBs (99 hyper- and 15 hypo-methylated), 75 endoderm
190 specific MHBs (58 hyper and 17 hypo-methylated) and 31 mesoderm specific MHBs (9 hyper and
191 22 hypo-methylated) (see Methods, **Supplementary Table 3**). We speculated that some of these
192 MHBs might capture binding events of transcription factors (TF) specific to developmental germ-
193 layers. Compared with ENCODE TFBS data²⁶, we observed distinctive patterns of TFs binding to
194 layer specific MHBs. (**Supplementary Fig. 6**). For layer specific MHBs with hypo-methylation MHL,
195 which tends to represent activation signals, we identified 53 TF binding events in mesoderm specific
196 MHBs, 71 in endoderm specific MHB and 2 in ectoderm specific MHBs. Gene ontology analysis
197 showed TFs binding to mesoderm exhibit negative regulator activity, while TFs binding to endoderm
198 exhibited positive regulator activity (**Supplementary Table 4**). For layer specific MHBs with hyper-
199 methylation MHL, which tend to represent repressive signals, we identified 38 TF binding events in
200 mesoderm specific MHBs, 102 in endoderm specific MHB and 145 in ectoderm specific MHBs.
201 Interestingly, ectoderm and endoderm shared few bounded TFs, while mesoderm tissues share
202 multiple groups of TFs with ectoderm and endoderm. We identified two endoderm specific hyper-
203 MHL regions, which are related to *ESRRA* and *NANOG*. This is consistent with a previous finding
204 that mouse ES cells differentiated spontaneously into visceral/parietal endoderm upon *NANOG*
205 knock-out²⁷. Gene ontology analysis showed that mesoderm and endoderm shared hypo-MHL
206 regions might have regulatory functions in the fate commitment towards multiple tissues, whereas
207 ectoderm specific hyper-MHL regions might induce the ectoderm development by suppressing the
208 path towards the immune lineage (**Supplementary Fig. 6**). These observations are indicative of two
209 distinctive “push” and “pull” mechanisms in the transition of cell states that have been harnessed for
210 the induction of pluripotency by over-expressing lineage specifiers²⁸.

211

212 **Methylation-haplotype based analysis of circulating cell-free DNA in cancer patients and**
213 **healthy donors.** A unique aspect of methylation haplotype analysis is that the pattern of co-
214 methylation, especially within MHBs, is robust in capturing low-frequency alleles among a
215 heterogeneous population of molecules or cells, in the presence of biological noise or technical
216 variability (ie. incomplete bisulfite conversion or sequencing errors). To explore the clinical potential,
217 we next focused on the methylation haplotype analysis of cell-free DNA from healthy donors and
218 cancer patients, of which various low fractions of DNA molecules were released from tumor cells
219 and potentially carry epigenetic signatures different from blood. We isolated 4-122ng (average 20ng)
220 of cell-free DNA from an average of 866µL human plasma from 75 normal individuals and 59 cancer
221 patients, except for four with unusually high yield due to cell lysis. Due to the limited DNA availability,
222 we performed scRRBS²⁹ on 1 to 10 ng of cfDNA from 134 plasma samples and obtained an
223 average of 13 million paired-end 150bp reads per sample. On average, 57.7% WGBS-defined
224 MHBs were covered in our RRBS data set on clinical samples.

225

226 We sought to detect the presence of tumor specific signatures in the plasma samples, using
227 methylation haplotypes identified from tumor tissues as the reference and normal samples as the
228 negative controls. For five lung cancer plasma samples and five colorectal cancer plasma samples,
229 we also obtained matched primary tumor tissues, and generated RRBS data (30 million reads per
230 sample) from 100ng of tumor genomic DNA. We focused on MHBs with low MHL (i.e. genomic
231 regions that have low or no methylation) in the blood, and asked whether we can detect **cancer-**

232 associated highly methylated haplotypes (caHMH). We required that such haplotypes were present
233 only in the tumor tissues and the matched plasma from the same patient, but not in whole blood or
234 any other non-cancer samples. We considered these highly confident tumor signature in circulating
235 DNA. We detected caHMH in all cancer patient plasma samples (Average=36, interquartile range
236 (IQR)=17, **Supplementary Table 5a**). These HMHS were associated with 183 genes, some of which
237 are known to be aberrantly methylated in human cancers such as *WDR37*, *VAX1*, *SMPD1*
238 (**Supplementary Table 5b**). Next, we extended the analysis to 49 additional cancer plasma
239 samples that have no matched tumor samples, using 65 normal plasmas as the background. On
240 average 60 (IQR=31) caHMH were identified for each cancer plasma sample (**Supplementary**
241 **Table 5c**). Interestingly, a significant fraction (35%) of caHMH called on matched tumor-plasma
242 pairs were also detected the expanded set of cancer patient plasma samples. We noticed majority
243 of caHMHS were individual specific while few caHMHS were present in at least 53% (16/30) and 62%
244 (18/29) cancer plasma samples for CRC and LC (**Supplementary Fig. 7**). Improving the sampling
245 depth, by either using more input cfDNA or reducing sample loss during the experiments, will likely
246 increase the number of caHMHS commonly observed in multiple patients.

247
248 Next we quantified the tumor load in cancer plasma samples, using non-negative decomposition
249 with quadratic programming, on the RRBS data from primary cancer biopsies (LC & CRC) and
250 WGBS data from 10 normal tissues. We estimated that a predominant fraction, 72.0% (IQR=40%) in
251 the cancer and normal plasma were contributed by white blood cells, which is consistent with the
252 levels reported recently based on shallow whole genome bisulfite sequencing (69.4%)⁶. Primary
253 tumor and normal tissue-of-origin contributed at the similar level of 2.3% (IQR=3.7%) and 3.0%
254 (IQR=4.4%). In contrast, we applied the similar analysis to normal plasma, and found only residual
255 tumor contributions (0.17%, IQR=2.9% for CRC and 1.0%, IQR=3.1% LC) to normal plasma, which
256 were significantly lower ($P=3.4\times 10^{-5}$ and 5.2×10^{-10} for CRC and LC, respectively) than cancer
257 plasma. We also found that 76.7% plasma samples from CRC patients and 89.6% from LC patients
258 had detectable contribution from tumor tissues while only 13% and 26% normal plasmas have
259 certain (low) tumor contribution (**Supplementary Fig. 8**). Therefore, circulating cell-free DNA
260 contains a relatively stable fraction of molecules released from various normal tissues, whereas in
261 cancer patient tumor cells released DNA molecules that can be more abundant than normal tissues
262 (**Supplementary Table 6**). The fractions of white blood cells observed are lower than what was
263 reported previously⁶, mostly likely due to the inclusion of 10 normal tissue types in our
264 deconvolution analysis.

265
266 We next asked whether we can identify a small subset of MHBs among all the RRBS targets that
267 have significantly higher levels of MHL in cancer plasma than in normal plasma. We found 81 and
268 94 MHBs with significantly higher MHL for colorectal and lung cancer (**Supplementary Table 7a-b**).
269 The majority (71/81 for CRC and 83/94 for LC) were also present in at least one of the matched
270 primary tumor and plasma pairs. Some of these regions (such as *HOXA3*) have been reported to be
271 aberrantly methylated in lung cancer and colorectal cancer Using these MHBs as markers, the
272 diagnostic sensitivity is 96.7% and 93.1% for colorectal cancer and lung cancer at the specificity
273 94.6% and 90.6%. As a comparison, we also performed a prediction based on average 5mC
274 methylation level within these MHB regions, or based on genome-wide single CpG sites. MHL was
275 found to be superior to average 5mC methylation level (sensitivity of 90.0% and 86.2%; specificity of
276 89.3% and 90.6% for CRC and lung cancer) and methylation signal of individual CpG site
277 (sensitivity of 89.6% and 80.6%; specificity of 89.3% and 92.0%).

278
279 We then sought to use the information from normal human tissues, primary tumor biopsies and

280 cancer cell lines to improve the detection of ctDNA. We started by selecting a subset of MHBs that
281 show high MHL (>0.5) in primary cancer biopsies and low MHL (<0.1) in whole blood, then clustered
282 these MHBs into three groups based on the MHL in all normal and cancer plasma, as well as
283 cancer and normal tissues (Figure 4). We identified a subset (Group II) of MHBs that have high
284 MHL in cancer tissues and low MHLs in normal tissues (Supplementary Table 8a-b). Cancer
285 plasma showed significantly higher MHL in these regions than normal plasma ($P=1.4\times10^{-12}$ and
286 6.2×10^{-8} for CRC and LC, respectively). By computationally mixing the sequencing reads from
287 cancer tissues and whole blood samples (WBC), we created synthetic admixtures at various levels
288 of tumor fraction. We found that MHL is 2-5 fold higher than the methylation level of individual
289 CpG sites across the full range of tumor fractions (Supplementary Table 8c-d). Remarkably, MHL
290 provides additional gain of signal-to-noise ratio (mean divided by standard deviation) compared with
291 AMF as the fraction of tumor DNA decreases below 10%, which is typical for clinical samples
292 (Figure 4c). We then took the individual plasma data sets, and predicted the tumor fraction based
293 on the MHL distribution established by computational mixing (Figure 4a-b). Except for a small
294 number ($N<5$) of outliers, we observed significantly higher average MHL in cancer plasma than in
295 normal plasma (Figure 4d). Note that all Group II MHBs were selected without using any
296 information from the plasma samples, and hence they should be generally applicable to other
297 plasma samples. Interestingly, we also found that the estimated tumor DNA fraction were positive
298 correlated with normalized cfDNA yield from the cancer patients ($P<0.000023$, Supplementary Fig.
299 9 and Supplementary Table 9).

300 Recent studies^{6,7,30} have demonstrated that epigenetic information imbedded in cfDNA has the
301 potential for predicting tumor's tissue-of-origin. Consistently, we found that tissue-of-origin derived
302 methylation haplotypes were the most abundant fraction in cancer plasma (Supplementary Table 5
303 and Supplementary Table 6). Here we asked whether a MHL-based framework and a set of
304 targets derived from whole genome data would allow us to predict tissue-of-origin with quantifiable
305 sensitivity and specificity, which is crucial for future clinical applications. We compiled 43 WGBS and
306 RRBS data sets for 10 human normal tissues that have high cancer incident rate, and identified a
307 set of 2,880 tissue-specific MHBs as the candidates (Supplementary Table 10). We then used
308 these tissue-specific MHBs or subsets to predict the tissue-of-origin for the cancer plasma sample.
309 Although we found a large number of tissue-of-origin specific MHBs that have low MHL in normal
310 plasma (Figure 5a), the multiclass prediction based on random forest yielded very limited power,
311 most likely due to the high diversity of the tissue classes ($N=10$). We then adopted an alternative
312 approach by counting the total number of tissue-specific MHBs in the plasma samples and
313 comparing with all other tissues, in order to infer the most probable tissue-of-origin. At the cutoff of
314 minimal 10 tissue-specific methylated haplotypes per tissue type, we observed an average 90%
315 accuracy for mapping a data set from the primary tissue to its tissue type (Figure 5b). We then
316 applied this method to the full set of plasma data from 59 cancer patients and 75 normal individuals,
317 and achieved an average prediction accuracy of 82.8%, 88.5%, 91.2% for the plasma from
318 colorectal cancer, lung cancer, and control plasma samples respectively with 5-fold cross-validation
319 (Figure 5c, Supplementary Fig. 10, Supplementary Table 11). For the incorrectly classified
320 samples, we noticed that 4 out of 5 colorectal cancer plasma were from metastatic colorectal cancer
321 patients while the fifth was in fact tubular adenoma. In the case of lung cancer, one misclassified
322 sample came from a patient with benign fibrous tissue. Taken together, we demonstrated for the
323 first time that both tumor load and tissue of origin can be quantitatively characterized by methylation
324 haplotype analysis of cell free DNA in plasma.

325

327 **Discussions**

328 In this study we extended a well-established concept in population genetics, linkage disequilibrium,
329 to the analysis of co-methylated CpG patterns. While the mathematical representations are identical,
330 there are two key differences. First, traditional linkage disequilibrium was defined on human
331 individuals in a population, whereas in this study the analysis was performed on the diploid genome
332 of individual cells in a heterogeneous cell population. Second, linkage disequilibrium in human
333 populations depends on the mutation rate, frequency of meiotic recombination, effective population
334 size and demographic history. The LD level decays typically over the range of hundreds of
335 kilobases to megabases. In contrast, CpG co-methylation depends on DNA methyltransferases and
336 demethylases, which tend to have lower processivity, and, in the case of hemi-methyltransferases,
337 much lower fidelity compared with DNA polymerases³¹. Therefore, methylation LD decays over
338 much shorter distance in tens to hundreds of bases, with the exception of imprinting regions. Even if
339 longer-read sequencing methods were used, we do not expect a radical change of the block-like
340 pattern presented in this work, which is supported by another recent study³². Nonetheless, these
341 short and punctuated blocks capture discrete entities of epigenetic regulation in individual cells
342 widespread in the human genome. Such a phenomenon can be harnessed to improve the
343 robustness and sensitivity of DNA methylation analysis, such as the deconvolution of data from
344 heterogeneous samples including circulating cell-free DNA.

345
346 While we demonstrated a superior power of MHL over single-CpG methylation level or average
347 methylation level in classification and deconvolution, the accuracy is slightly less than what has
348 been reported on the deconvolution of blood cell types. One major difference is that each reference
349 tissue type itself is a mixture of multiple cell types that might share various degrees of similarity with
350 another reference tissue type. Furthermore, most solid tissues also contain blood vessels and blood
351 cells. Given such background signals, the accuracy that we achieved is very promising, and will be
352 further improved once reference methylomes of pure adult cell types are available.

353
354 Practically, the amount of cell-free DNA per patient is rather limited, typically in the range of tens to
355 hundreds of nanogram. We used 1 to 10 ng per patient for the sc-RRBS experiment. Considering
356 the material losses during bisulfite conversion and library preparation, as well as the sequencing
357 depth, there were most likely no more than 30 genome equivalents in each data set. Our data set is
358 rather sparse, especially when the fraction of tumor DNA is low. Hence the chance of finding
359 cancer-specific methylation haplotypes in a specific region consistently across many samples is low.
360 This is likely the reason that marker sets selected based on random forest has limited sensitivity and
361 specificity. However, epigenetic abnormalities tend to be more widespread across the genome
362 (compared with somatic mutations), and hence we were able to integrate the sparse coverage
363 across many loci to achieve very accurate prediction by direct counting of methylated haplotypes
364 within the appropriate tissue-specific features. Further technical improvements on sample
365 preparation and library construction, combined with larger sets of patient and normal plasma, will
366 undoubtedly increase the coverage and further improve the specificity/sensitivity to the level
367 adequate for clinical diagnosis.

368 **Methods**

369 **Normal and cancer samples**

370 Ten human primary tissues were purchased from BioChain. Cancer tissue and plasma samples
371 were collected from UCSD Moores Cancer Center and normal plasma samples were obtained from

372 UCSD Shirley Eye center under IRB protocols approved by UCSD Human Research Protections
373 Program (HRPP). All data sets generated in this study or obtained from public databases were listed
374 in **Supplementary Table 12**.

375

376 **Generation of DNA libraries for sequencing**

377 Extracted genomic DNA were prepared for bisulfite sequencing using published protocols. For
378 whole genome bisulfite (WGBS) and reduced representation bisulfite sequencing (RRBS), the DNA
379 fragments were adapted to barcoded methylated adaptors (Illumina). For WGBS, the adapted DNA
380 were converted using the EZ DNA Methylation Lightning kit (Zymo Research) and then amplified for
381 10 cycles using iQ SYBR Green Supermix (BioRad). For RRBS, the adapted DNA were converted
382 using the MethylCode™ Bisulfite Conversion kit (Thermo Fisher Scientific) and amplified using the
383 PfuTurboCx polymerase (Agilent) for 12-14 cycles. Libraries were pooled and size selected using 6%
384 TBE polyacrylamide gels. Libraries were sequencing using the Illumina HiSeq platform for paired-
385 end 100-111 cycles, the Illumina MiSeq platform for paired-end 75 cycles, and the GAIIx (WGBS
386 only) for single-end 36 cycles.

387

388 **Methylation haplotype blocks (MHB)**

389 Human genome was separated into non-overlapping “sequenceable and mappable” segments
390 using a set of in-house generated WGBS data from 10 tissues from a 25-year adult male individual.
391 Mapped reads from WGBS data sets were converted into methylation haplotypes in each segment.
392 Methylation linkage disequilibrium was calculated on the combined methylation haplotypes. We then
393 partitioned each segment into methylation haplotype blocks (MHBs). MHBs were defined as the
394 genomic region in which the r^2 value of two adjacent CpG sites is no less than 0.5. MHB regions
395 inferred by GWBS dataset was also validated by bulk data of methylation level. Takai and Jones's
396 sliding-window algorithm³³ was applied for methylation high linkage regions in HM450K (TCGA) and
397 RRBS (Encode) dataset. Finally, simulation analysis to investigate the relationship between LD and
398 correlation of average 5mC of two CpG loci were conducted based on random sampling different
399 methylation haplotype with 1000 individual and each individual sampling 10 methylation haplotype.

400

401 **Methylation haplotype load (MHL)**

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

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

$$w_i = i$$

404 Where l is the length of haplotypes, $P(MH_i)$ is the fraction of **fully successive methylated** with i loci.
405 For a haplotype of length L, we considered all the sub-strings with length from 1 to L in this
406 calculation. w_i is the weight for i-locus haplotype. We typically used $w_i = i$ or $w_i = i^2$ to favor the
407 contribution of longer haplotypes. In the present study, $w_i = i$ was applied. Quantile normalization,
408 standardization (scale) as well as the batch effect elimination³⁴ were applied and the top quantile 15%
409 MHL regions were selected in heatmap analysis to investigate the tissue relationship. The
410 Euclidean distance and Ward.D aggregation were applied in the heatmap plot (R, gplots package).

411

412 **Developmental germ layers and tissue specific MHB regions.**

413 In order to investigate the layer and tissue specific MHB regions, group specific index (see below)
414 were applied. An empirical threshold 0.6 were selected to filter out layer and tissue specific MHB
415 regions. Layer specific MHB regions were selected again to show the distinguish ability to different

416 development layers. Tissue specific MHB regions were further used to apply tissue mapping and
417 cancer diagnosis.

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

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

419 $MHL(max)$ denotes the average of MHL of highest methylated group.

420

421 **Simulation and real-data deconvolution analysis**

422 Deconvolution analysis were conducted by simulation and real-data ways. The deconvolution
423 references were constructed by human normal solid tissues, WBC, colorectal cancer tissues (CCT)
424 and lung cancer tissues (LCT). For the simulation analysis, methylation haplotypes were mixture by
425 CCT and WBC with specific gradients (CCT contents ranging from 0.1% to 50%) and then expected
426 and observed CCT contents were compared. Although our MHL is a non-linear metric when mixing
427 CCT and WBC, we found the deconvolution result is perfect with logit transform, median root-mean-
428 square-error < 5%, which is within the acceptable region of the deconvolution method³⁵ when the
429 contribution of colorectal fraction is less than 20% (**Figure 4d**). Tissue specific MHB regions were
430 applied to be the candidate features for deconvolution based on non-negative decomposition with
431 quadratic programming^{6,35,36}. Raw MHL signals were applied of logit transform before deconvolution
432 analysis. The contribution of the WBC to cancer plasma, normal plasma samples were estimated.
433 Meanwhile, the contribution of the cancer plasma from CCT and LCT were estimated respectively.
434 Finally, the contribution of CCT and LCT for cancer plasma and normal plasma were compared.

435

436 **Diagnosis biomarker identification and tissue mapping algorithm for plasma DNA.**

437 The flowchart of the analysis in current study was shown in **Supplementary Fig. 11**, especially for
438 the sample size in each section. Tumor specific methylation haplotype blocks based on were
439 identified by 2-tailed t-test with False Discovery Rate (FDR) correction. Other statistical analysis to
440 MHL were also conducted by 2-tailed t-test without explicitly notification. **CRC plasma and LC**
441 **plasma distinguish prediction evaluation were applied random forecast therefore the test and**
442 **validation sample were independent.** Tumor-of-origin prediction were applied with tissue-specific
443 MHBs counting (MHC) strategy in which the tissue-of-origin of the plasma were assigned to the
444 group for which have maximum tissue-specific MHB fragments (assignment by maximum likelihood).
445 For the detail, In the first stage, the tissue-specific MHBs was identified with WGBS and RRBS
446 dataset from solid tissues in the training samples. Tissue specific MHB regions (each tissue ~ 300
447 MHBs) were obtained by filtered with the moderate $GSI > 0.1$ so that we could select the most
448 powerful biomarkers which can be detected in RRBS and GWBS. In the second stage, the built
449 prediction model was validated with our own RRBS dataset which including 30 colorectal cancer
450 plasma, 29 lung cancer plasma and 75 normal plasma samples. In the test dataset, we separated
451 the samples into 5 parts so that 5-fold cross-validation could be applied to measure the stability of
452 the prediction, number of tissue-specific MHB features were iterating from 50 to 300 and the
453 minimum feature number was selected when accuracy for cancer plasma higher than 0.8 and
454 normal plasma higher than 0.9 since we require high specificity in the realistic application in 4-fold
455 samples. The selected number of features and then were used in the remaining samples to
456 measure the accuracy of tissue-mapping. The variations of sensitivity, specificity and accuracy in
457 different subsets of 5-fold cross-variation were quite slight (training dataset standard deviation<0.04
458 while testing dataset standard deviation<0.14, see supplementary Table 11), indicating the current
459 sample size could provide enough prediction power.

460

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

462 **Data Availability**

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

465 **Code Availability**

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

468
469 **Acknowledgements**
470
471 This study was supported by NIH grants R01GM097253 (to K.Z.) and P30CA23100. We thank S.
472 Kaushal for managing and handling patient samples in UCSD Moores Cancer Center BTTSR.

473 **Author's Contributions**

474 Ku.Z. conceived the initial concept and oversaw the study. S.G., D.D. and Ku.Z. performed
475 bioinformatics analyses. N.P., D.D., and H.F. performed experiments. Ka. Z. contributed normal
476 plasma samples. Ku. Z., S.G. and D.D. wrote the manuscript with inputs from all co-authors.
477

478 **Competing Financial interests**

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

481 **Abbreviation**

482 MHB: methylation haplotype load; MHL: Methylation Haplotype Load; cf-DNA: Circulating cell-free
483 DNA; RRBS: Reduced representation bisulfite sequencing; scRRBS: single-cell reduced-
484 representation bisulfite sequencing; WGBS: genome-wide bisulfite sequencing; TCGA: The Cancer
485 Genome Atlas project; ENCODE: the Encyclopedia of DNA Elements; GEO: Gene Expression
486 Omnibus; LC: Lung Cancer; CRC: Colorectal cancer; ACC: Accuracy; **caHMH: cancer associated**
487 **high methylation haplotype**; ts-MHB: tissue specific methylation haplotype block regions. CCT:
488 Colorectal cancer tissue; CCP: colorectal cancer plasma; LCT: lung cancer tissue; LCP: lung cancer
489 plasma; NP: normal plasma.
490

491 **Figure legends**

492 **Figure 1.** Identification and characterization of human methylation haplotype blocks(MHBs). (a)
493 Schematic overview of data generation and analysis. (b) An example of MHB at the promoter of the
494 gene APC. (c) **Smooth scatterplots of methylation linkage disequilibrium decay distance of adjacent**
495 **CpG sites.** 500,000 adjacent CpG loci in MHB regions were randomly sampling and the attenuation
496 of the the r^2 with the distance of the CpG loci in different scenario shown different characteristics.
497 Yellow dot line to show the decay of the linkage disequilibrium. 94.8%, 91.2% and 87.8% were

498 maintained high linkage ($r^2 > 0.9$). stem cells and progenitors (pooling of 10 samples), normal adult
499 tissues (pooling of 49 samples), and primary tumors (pooling of 6 samples from CRC and LC). (d)
500 Co-localization of MHBs with known genomic features. Genome distribution (left) and CpG island
501 nearby status show MHB are widely dispersed in human genome (e). Enrichment of MHBs in known
502 genomic features. Bootstrap random sampling regions with same size for 10,000 times to estimate
503 empirical statistical significance and enrichment factor (fold-change).

504

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

509

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

517

518 **Figure 4.** Quantitative estimation of tumor load in cell-free DNA based on MHL of informative MHBs.
519 (a) Colorectal cancer (b) Lung cancer. Informative MHBs were selected based on the presence of
520 high-MHL in cancer solid tissues and the absence of MHL in WB. Group II regions have high MHL in
521 cancer tissues (MHL>0.5) and cancer plasma while low MHL in WBC and normal tissues (MHL<0.1),
522 and hence selected for further analysis. Barplot showed MHL in different groups of samples. MHL
523 level in cancer plasma (CRC and LC) and normal plasma (NP) were compared with two-tail t-test.
524 ONT: other normal tissues. (c) Computational mixtures of cancer and whole blood DNA at different
525 ratios (0.1% to 50%) were created by random sampling of haplotypes in the Group II regions, and
526 repeated 1,000 times to empirically determined the mean and variance of MHL and 5mC levels at
527 different fractions of cancer DNA. (d) After an empirical “standard curve” was constructed, it was
528 used to estimate the fraction of cancer DNA in plasma samples. CCP denotes colorectal cancer
529 plasma, LCP denotes lung cancer plasma and NP denotes normal plasma.

530

531

532 **Figure 5.** Methylation haplotype load based cancer detection and tumor tissue-of-origin prediction
533 from plasma DNA. (a) Detection of tumor-specific or tissue-specific MHL in the plasma of cancer
534 patients, but not normal plasma or whole blood. Tissue specific MHL were observed in tissue and
535 corresponding cancer plasma, indicating the possibility for tissue-of-origin mapping. b) Identification
536 of informative MHBs for tissue prediction. A total of ~300 tissue-specific MHBs were selected with
537 the cutoff GSI>0.1. Training dataset including WGBS and RRBS dataset. (c) Application of the
538 predictive model to plasma samples from cancer patients and normal individuals. Plasma samples
539 (30 CRC, 29 LC and 75 NP) were separated into 5 parts (each group) so that 5-fold cross-validation
540 could be applied to measure the stability of the prediction; the number of tissue-specific MHB
541 features were iterating from 50 to 300 and the minimum feature number was selected when
542 accuracy for cancer plasma was above 0.8 and normal plasma above 0.9.

543

544 **Supplementary Figure Legends:**

545 **Supplementary Figure 1.** Characteristics of MHB in human genome. (a) Distribution of MHB sizes.
546 (b) Distribution of CpG density (CpGs/bp) in MHB regions. (c) Colocalization of MHB with known
547 genomics features breaking down based on CpG density. We split all MHBs into quartiles where the
548 CpGs/bp of each quantile is as follows: (0,0.046), (0.046,0.096), (0.096, 0.155), and (0.155,0.6).
549 The 1st quantile (MHBs with lowest CpG density) are mostly in CGI shelf or shore, and are enriched
550 for LAD, LOCK and enhancers.

551
552 **Supplementary Figure 2.** Loss of CpG linkage disequilibrium replicated in two additional kidney
553 cancer samples. Two sets of kidney cancer WGBS data were downloaded from NCBI
554 GEO(GSE63183), and processed with the same computational procedures.

555
556
557 **Supplementary Figure 3.** Validation of MHB with Illumina 450k methylation array and RRBS data.
558 (a) Pearson correlation coefficient (r^2) versus absolute LD r^2 (b) The Pearson correlation coefficient
559 (r^2) in RRBS and HM450K were significantly higher in overlapped MHBs with WGBS compared with
560 the MHBs without overlapping with WGBS MHBs. IN: denotes RRBS or HM450K regions within
561 MHB. OUT:denotes RRBS or HM450K regions beyond MHB regions.

562 **Supplementary Figure 4.** Profiles of H3K27ac, H3K4me3 and H3K4me1 over methylation
563 haplotype blocks for 12 adult tissue types. X-axis are distances from the center of methylation
564 haplotype blocks (+/- 1000) and y-axis are the average reads density in RPKM (input normalized
565 reads per kilobase per million).

566 **Supplementary Figure 5.** PCA analysis of human tissues and cells based on MHL.

567
568 **Supplementary Figure 6.** Distinctive patterns of functional enrichment for TF associated with
569 MHBs of hypo- or hyper MHL.

570
571 **Supplementary Figure 7.** Distribution of incidence of cancer-associated HMH in CRC and LC
572 plasma samples. y-axis denotes the frequency of caMHM and x-axis denotes the incidence (sample
573 number) of the caMHM in cancer plasmas. We found majority caMHM are patient specific while a
574 few of them will have high incidence among the cancer plasma samples.

575
576 **Supplementary Figure 8.** Deconvolution of cancer and normal plasma using non-negative
577 decomposition with quadratic programming. (a) deconvolution accuracy as a function of tumor
578 fraction. Red line indicates the diagonal line where prediction equals to the expected values; black
579 line indicates the deconvolution values. (b) Tumor fraction estimated by deconvolution analysis on
580 cancer and normal plasma samples.

581
582 **Supplementary Figure 9.** Estimated tumor fraction in plasma is generally correlated with the
583 normalized yield of DNA extraction. CCP denotes colorectal cancer plasma, LCP denotes lung
584 cancer plasma and NP denotes normal plasma.

585
586 **Supplementary Figure 10.** Tissue-of-origin mapping based on tissue-specific MHBs counting. CCP
587 denotes colorectal cancer plasma, LCP denotes lung cancer plasma and NP denotes normal
588 plasma.

589
590
591

Supplementary Figure 11. Flowchart of data analysis and samples used in each section.

592 **Supplementary Tables:**

- 593 **Supplementary Table 1.** Genome-wide MHBs identified from 65 sets of WGBS data.
594 **Supplementary Table 2.** Tissue specific MHBs identified based on tissue specificity index.
595 **Supplementary Table 3.** Germ-layer specific MHBs identified based on layer specificity index.
596 **Supplementary Table 4.** Complete list of highly methylated haplotype shared between primary
597 cancer tissue and matched plasma for CRC and lung cancer patients.
598 **Supplementary Table 5.** Component deconvolution of cancer plasma from WB, normal tissue and
599 primary cancer tissues based on high methylation haplotypes.
600 **Supplementary Table 6.** Deconvolution of CRC, LC and normal plasma samples by 10 normal
601 tissues and LCT, CCT
602 **Supplementary Table 7.** Significantly differential MHB regions between cancer and normal plasma.
603 **Supplementary Table 8.** Comparison of MHL and average 5mC based on computational mixtures
604 of cancer and blood DNA.
605 **Supplementary Table 9.** Correlation between estimated cancer DNA fraction and normalized cell-
606 free DNA yield.
607 **Supplementary Table 10.** Marker regions used in the prediction models for CRC, LC and normal
608 plasma.
609 **Supplementary Table 11.** Prediction accuracy based on tissue-specific MHBs counting with 5-fold
610 cross-validation.
611 **Supplementary Table 12.** Information of all samples used in this study.
612

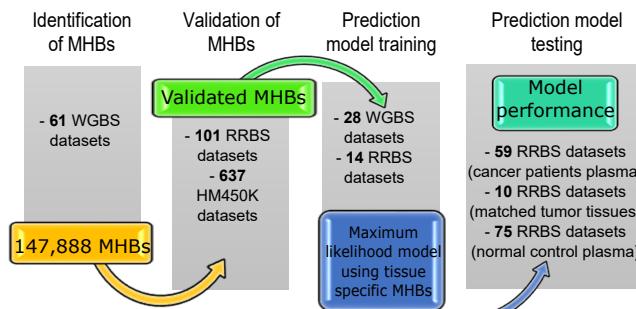
613 **Reference**

- 614 1. Wigler, M., Levy, D. & Perucco, M. The somatic replication of DNA methylation. *Cell* **24**, 33-40 (1981).
615 2. Slatkin, M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**, 477-85 (2008).
616 3. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-8 (2010).
617 4. Jones, P.A. & Martienssen, R. A blueprint for a Human Epigenome Project: the AACR Human Epigenome
618 Workshop. *Cancer Res* **65**, 11241-6 (2005).
619 5. Houseman, E.A. *et al.* Reference-free deconvolution of DNA methylation data and mediation by cell
620 composition effects. *BMC Bioinformatics* **17**, 259 (2016).
621 6. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal,
622 cancer, and transplantation assessments. *Proc Natl Acad Sci U S A* **112**, E5503-12 (2015).
623 7. Lehmann-Werman, R. *et al.* Identification of tissue-specific cell death using methylation patterns of circulating
624 DNA. *Proc Natl Acad Sci U S A* **113**, E1826-34 (2016).
625 8. Shoemaker, R., Deng, J., Wang, W. & Zhang, K. Allele-specific methylation is prevalent and is contributed by
626 CpG-SNPs in the human genome. *Genome Res* **20**, 883-9 (2010).
627 9. Schultz, M.D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**,
628 212-6 (2015).
629 10. Heyn, H. *et al.* Distinct DNA methylomes of newborns and centenarians. *Proc Natl Acad Sci U S A* **109**, 10522-7
630 (2012).
631 11. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**,
632 1134-48 (2013).
633 12. Blattler, A. *et al.* Global loss of DNA methylation uncovers intronic enhancers in genes showing expression
634 changes. *Genome Biol* **15**, 469 (2014).
635

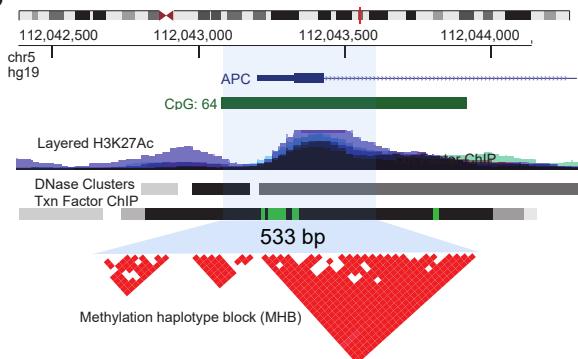
- 636 13. Heyn, H. *et al.* Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol* **17**, 11 (2016).
- 637 14. Chen, K. *et al.* Loss of 5-hydroxymethylcytosine is linked to gene body hypermethylation in kidney cancer. *Cell Res* **26**, 103-18 (2016).
- 638 15. 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).
- 639 16. 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).
- 640 17. Hansen, K.D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43**, 768-75 (2011).
- 641 18. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-51 (2008).
- 642 19. 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).
- 643 20. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-80 (2012).
- 644 21. Pujadas, E. & Feinberg, A.P. Regulated noise in the epigenetic landscape of development and disease. *Cell* **148**, 1123-31 (2012).
- 645 22. 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).
- 646 23. Ziller, M.J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477-81 (2013).
- 647 24. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350-4 (2015).
- 648 25. Heyn, H. *et al.* Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol* **17**, 11 (2016).
- 649 26. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
- 650 27. Mitsui, K. *et al.* The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113**, 631-42 (2003).
- 651 28. Shu, J. *et al.* Induction of pluripotency in mouse somatic cells with lineage specifiers. *Cell* **153**, 963-75 (2013).
- 652 29. 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).
- 653 30. 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).
- 654 31. Williams, K. *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343-8 (2011).
- 655 32. Saito, D. & Suyama, M. Linkage disequilibrium analysis of allelic heterogeneity in DNA methylation. *Epigenetics* **10**, 1093-8 (2015).
- 656 33. 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).
- 657 34. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-27 (2007).
- 658 35. Houseman, E.A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
- 659 36. 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).

Figure 1

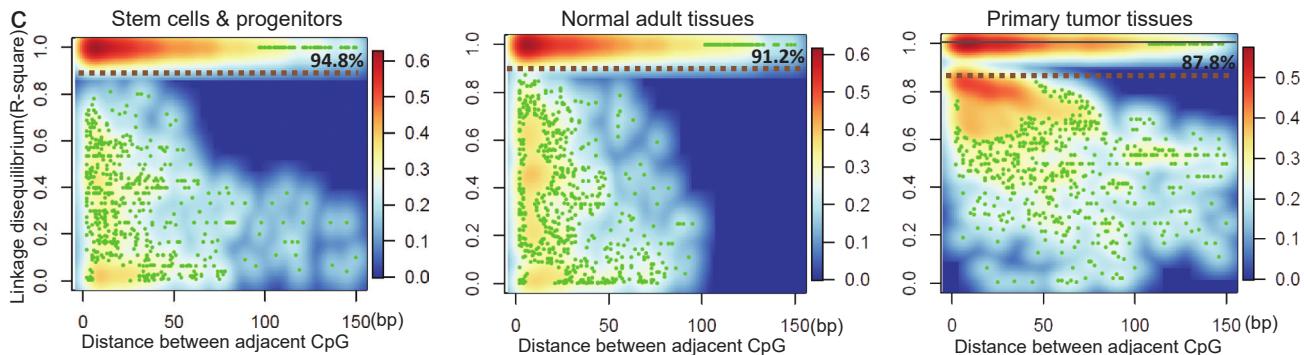
a



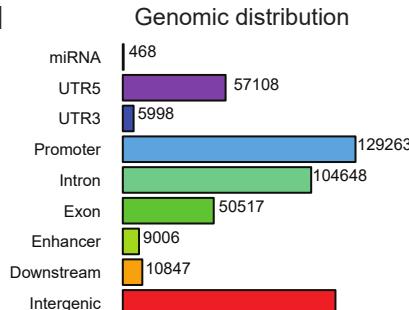
b



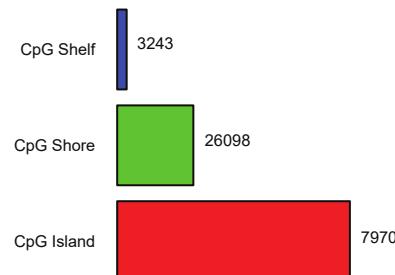
c



d



CpG distribution



e

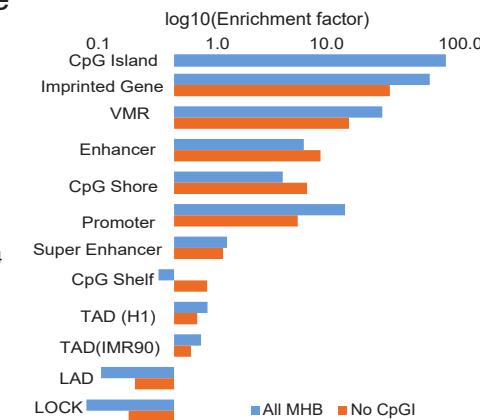


Figure 2

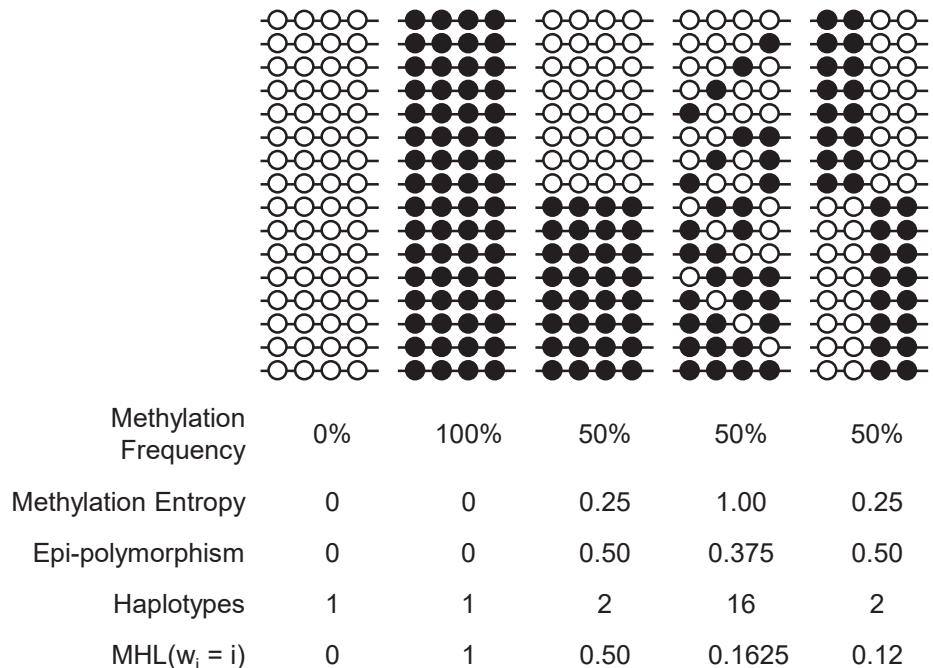


Figure 3

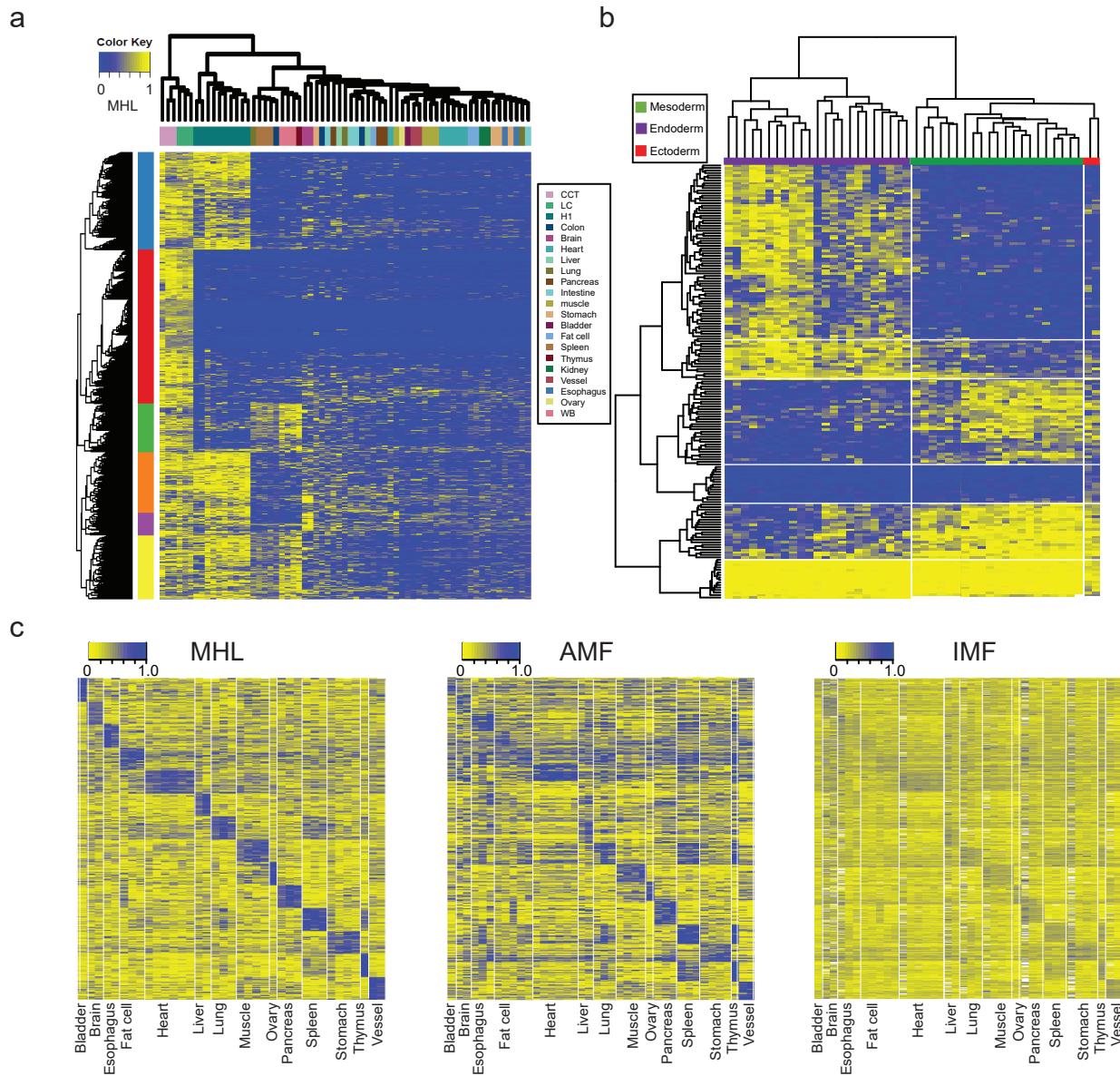


Figure 4

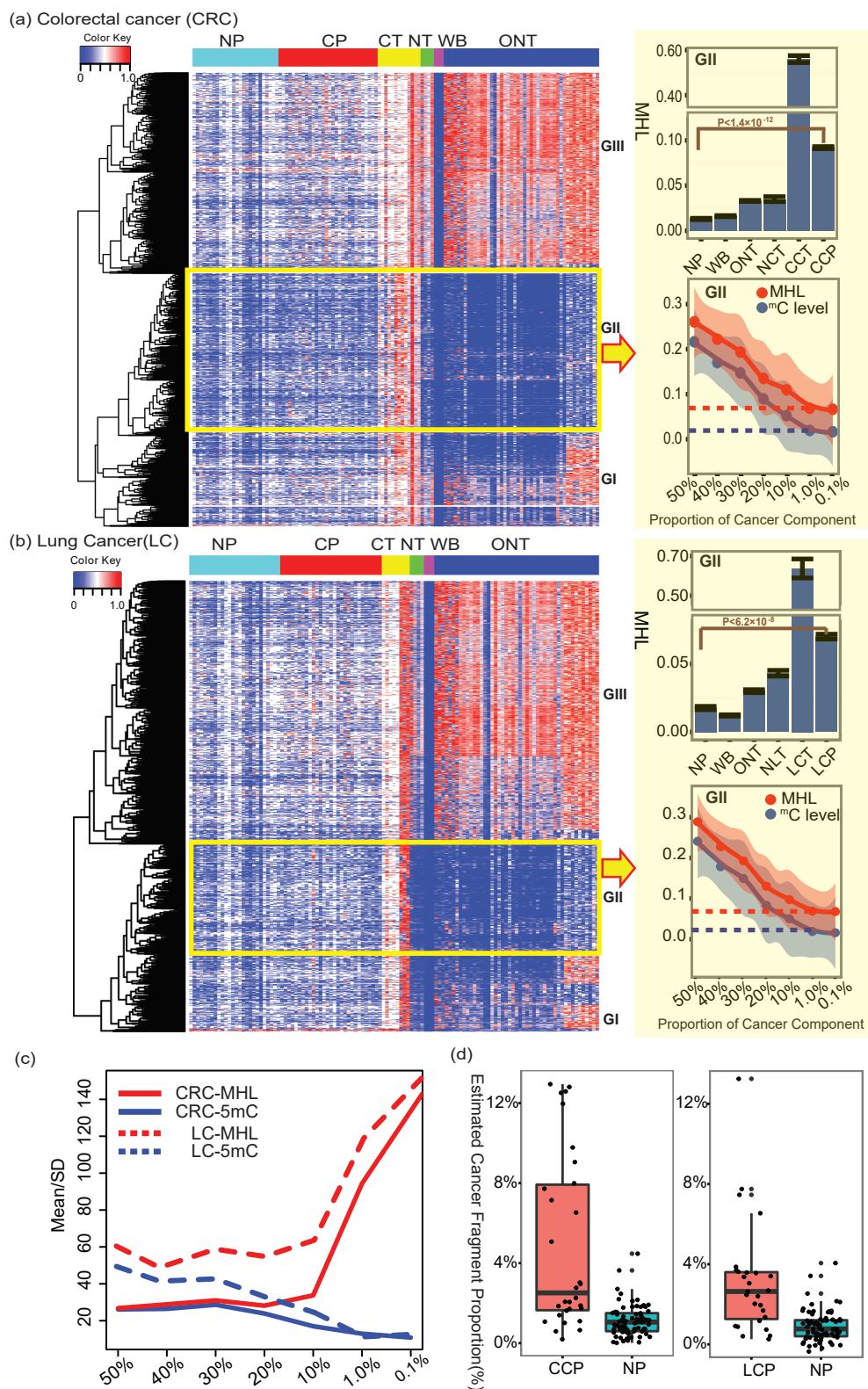


Figure 5.

