# Deconvolution of epigenetic heterogeneity by tightly coupled CpG methylation.

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

Figure counts: 4

Abstract counts: 250

Main body counts: 3586

Running title: Methylation haplotype in human cancer

Keywords: Cell-free circulating DNA methylation, Methylation haplotype loading, Non-invasive diagnosis

Corresponding authors:

Dr. Kun Zhang, Institute for Genomic Medicine and Institute of Engineering in Medicine, Department of Bioengineering, University of California at San Diego, 9500 Gilman Drive, MC0412, La Jolla, CA 92093-0412, Tel: 858-822-7876, Fax: 858-534-5722, E-mail: kzhang@eng.ucsd.edu

## Abstract

Adjacent CpG sites in mammalian genomes tend to be co-methylated due to the processivity of enzymes responsible for adding or removing the methyl group. Yet discordant methylation patterns have also be observed, and found to be related to stochastic or uncoordinated molecular processes. Here we focused on a systematic search and investigation of regions in the human genome that exhibit highly coordinated methylation.

By examining the co-methylation patterns of multiple adjacent CpG sites, termed methylation haplotypes, in single bisulfite sequencing reads, we applied a greedy-searching strategy to defined blocks of tightly coupled CpG sites, called Methylation Haplotype Blocks (MHBs), based on 53 sets of whole genome bisulfite sequencing (WGBS) data, including 43 published sets from human adult tissues, ESC and in vitro differentiated cell lines, as well as 10 sets from human adult tissues generated in this study. The MHBs were then further validated with 101 sets of RRBS ENCODE data, and 1,274 sets of Illumina 450k methylation array data from TCGA tumor and normal samples. Globally, MHBs are enriched in but only partially overlap with several well-known genomic features, including CpG islands, promoters, enhancers and VMRs.

To perform quantitative analysis of the MHBs, we defined a metric called Methylation Haplotype Load (MHL), which is covered both average methylation level and methylation complexity and therefore more informative than average methylation level or Shannon entropy. Using a feature selection strategy, we identified a set of tissue-specific MHBs that cluster by developmental germ-layers. Interestingly, examination of these MHBs revealed two distinct mechanisms for fate commitment during development: epigenetic silencing of pluripotent genes, such as NANOG, for mesoderm induction; and epigenetic induction (or de-suppression) of lineage-specific factors for ectoderm commitment.

## Introduction

DNA methylation is one of most important inherited epigenetic modification in the human genome that occurs primarily on the cytosines in CpG dinucleotides. the dynamic change of the methylation profile are involved in most human biological and cellular, physiological, pathological changes. While previous studies have defined methylation heterogeneity based mainly on the methylation levels of individual CpG sites (Beta-value or M-value), the underlying methylation pattern within each allele and haplotype have not yet been examined. What’s more, the assumption of the independent between the adjacent of CpG sites is irrational from previous evidences. Therefore, it is important to determine the methylation pattern of consecutive CpG sites within a specific genomic region (methylation haplotype). Next generation based sequencing technique provide the opportunity to investigate the co-methylation between adjacent CpG sites. In our previous study, we have investigated the linkage disequilibrium (LD) between adjacent CpG sites on single DNA molecules analysis to characterize the correlation of methylation genomic variations such as SNPs (4). However, genome-wide investigation of the linkage disequilibrium (LD) between adjacent CpG sites on single DNA molecules in a large number samples were not conducted yet.

In addition, a range of molecular alterations found in tumor cells, such as DNA mutations and DNA methylation, is reflected in cell-free circulating DNA (cfc-DNA) released from the tumor into the blood, thereby making cfc-DNA an ideal candidate for the basis of a blood-based cancer diagnosis test. Comparing with mutation or copy number variation, DNA methylation own different aberrant characteristics in cancers, as it mainly happened in CpG island and shore with a relatively high frequency, therefore, it was considered as the most potential biomarker for cancer non-invasive diagnosis. Methylation based on single CpG locus and average of methylation in a genomic region have been applied in biomarker identification, However, such measurement ignored the correlation between adjacent CpG. Finally, how to apply the linkage disequilibrium to cancer diagnosis and other clinical application were not considered yet. In the present study, we explored methylation haplotype in normal and cancer tissues and proposed a novel metric termed as methylation haplotype load which is better than previous methylation entropy (5) and epi-polymorphism (6) to depict the methylation heterogeneity for the genomic regions. We carried out a comprehensive genome-wide DNA methylation analysis across 10 human normal tissues with WGBS and 45 plasma from cancer patients, 30 normal plasma and 45 solid cancer tissues with RRBS. We demonstrate that the methylation haplotype based metrics could be applied the cancer diagnosis and tissue-of-origin prediction to plasma from cancer patients.

## Results

#### **Characterization of methylation haplotype blocks.**

To investigate the co-methylation status of adjacent CpG sites along single DNA molecules, we extended an approach that we previously established (Shoemaker et a. GR, 2011), in which we applied the concept of genetic linkage disequilibrium and the r2 metric to quantify the degree of co-methylation among different DNA molecules. CpG methylation status of multiple CpG sites in single-end or paired-end Illumina sequencing reads were extracted to form methylation haplotypes, and pairwise “linkage disequilibrium” of DNA methylation r2 was calculated from the distribution of methylation haplotypes (Materials and Methods). We then partitioned the genome into blocks of tightly co-methylated CpG sites, which we called Methylation Haplotype Blocks (MHBs), using a r2 cutoff of 0.5. Similar to the partitioning of genetic haplotype blocks, using different cutoff values, such as 0.3 or 0.7, resulted in only minor quantitative differences in the block size and number without affecting the global pattern (data not shown).

To characterize the global pattern and distribution of MHBs, we started with 51 sets of published Whole Genome Bisulfite Sequencing (WGBS) data from human primary tissues, as well as the H1 human embryonic stem cells and in vitro derived progenitors. We also included an in-house generated WGBS data set from 10 human adult tissues. [What’s the total genome coverage of these data sets combined?] Across this 61 data set we identified a total of 147,888 MHBs at the average size of 95bp, which represents ~0.5% of the human genome(Supplementary Table 1). Note that our analysis was limited by the length of Illumina sequencing reads, and hence our results might represent the lower bound of the true MHBs in the human genome. The majority of CpG sites within the same MHBs are near perfectly coupled (r2 ~1.0) regardless of the sample types. We found that methylation LD extends further along the DNA in stem cells and progenitors, compared with normal adult tissue, both in the fraction of tightly coupled CpG pairs (94.8% versus 91.2%), and the over-representation of partially coupled CpG pairs that are over 100 bp apart (Figure 1B). This is consistent to our previous observations on a smaller BSPP data set on 2,020 CpG islands (Shoemaker et al. 2010). Interestingly, in primary tumor tissues, we observed a reduction of perfectly coupled CpG pairs, which could be related to the pattern of discordant methylation that has been reported in AML (Landau et al., Cancer Cell, 2014).

These MHBs appear to represent a distinct type of genomic feature that partially overlap with well-documented genomic elements (Figure 1C). Among all the methylation blocks, 60,828 (41.1%) were located in intergenic regions while 87,060 (58.9%) regions in transcribed regions. These MHBs were significantly enriched in enhancers (enrichment factor=7.6, p-value<10-6), super enhancer (enrichment factor=2.3, p-value<10-6), promoter regions (enrichment factor=14.5, p-value<10-6), CpG islands (enrichment factor=70.4, P<10-6) and imprinted genes (enrichment factor=54.6, p-value<10-6). In addition, we observed modest depletion in LAD and LOCK regions (46% and 37% of the expected values), modest enrichment in TAD, and a strong (26x) enrichment in variable methylation regions (VMR). To decouple the influence of CpG islands, we further examined the subset of MHBs that do not overlap with CpG islands, and observed a consistent enrichment pattern (Figure 1D).

In addition, enrichment of MHB in histone H3K27ac, H3K4me1 and H3K4me3 were also found which indicating the MHB was preferred in genomic regulatory regions (Supplementary Figure 1S3).

While WGBS data allowed us to unbiasedly identify MHBs across the entire genome, the 61 sets of data did not cover the full diversity of human cell/tissue types. To validate the MHBs in a wider range of human tissues and culture cells, we next searched for MHBs in published Reduced Representation Bisulfite Sequencing (RRBS) data on ENCODE cell lines and tissue samples. From 101 RRBS data sets we identified a total of 23,517 MHBs, which have significant overlaps with the 147,888 MHBs discovered in 61 WGBS data sets ( , ). Due to the limited number of human primary tissue samples in the ENCODE RRBS data set, we further analyzed 637 sets of Illumina 450k methylation array data from 11 human normal tissues. While we used the same computational approach to analyze RRBS and WGBS data due to their high similarity, Illumina methylation arrays only report average CpG methylation of all DNA molecules in a sample, which cannot be used for methylation linkage disequilibrium analysis. Instead, we calculated the pairwise correlation coefficient of adjacent CpG methylation levels across 637 data sets for block partitioning. Note that the presence of such correlated methylation blocks is a necessary but not sufficient condition for MHBs. Nonetheless, the absence of correlated methylation blocks in the 450k methylation array data would indicate potential false discovery of MHBs. Instead, from the 637 human solid tissue samples we identified 1,635 correlated methylation blocks, of which 833 (51%) were overlapping with the MHBs. Taken together, the MHBs that we identified from the 799 data sets represent a distinct class of genomic feature where local CpG methylation is established or removed in a highly coordinated manner at the level of single DNA molecules, presumably due to the processive activities of related enzymes.

**Characterizations of methylation haplotype load**

To enable quantitative analysis of the methylation patterns within MHBs across many samples, we need a single metric to define each MHB. Ideally this metric is not only a function of average methylation level for all CpG sites within a block, but also can capture the pattern of co-methylation on single DNA molecules. For this purpose, we defined a metric, called Methylation Haplotype Load (MHL), which is essentially a weighted mean of the fraction of fully methylated haplotypes at different length. Compared with other existing metrices, such as correlation coefficient, methylation level，methylation entropy, epi-polymorphism and haplotypes counts, MHL can represent both the overall level of methylation, but also is capable of distinguishing blocks that have the same average methylation but various degrees of coordinated methylation (Figure 2). In addition, MHL is bounded between 0 and 1, which allows direct comparison of different regions across many data sets without normalization.

We next asked whether treating MHBs as individual units and performing quantitative analysis based on MHL would provide an advantage than existing approaches based on individual CpG sites or averaging multiple CpG sites in certain genomic windows. To this end, we sought to cluster XX data sets from human adult tissues based on the MHL metrix. U[nsupervised](http:///h) cluster analysis with the top 15% most variable MHBs showed that, regardless of the data source, samples of the same tissue origin clustered together (Figure 3A). Clustering using all MHBs genome-wide yielded a similar pattern(Supplementary Figure 3). To identify a subset of MHBs that have the best power for clustering tissues, we calculated a tissue specific index (TSI) for each MHB (see Methods). We expected the MHBs with high TSI would distinguish the different tissues with a high specificity. Random Forest based feature selection identified a set of tissue-specific MHBs(1360 MHL region, see Supplementary Table 2) that can predict tissue type at an accuracy of 0.89 (95%CI: 0.84-0.93), despite the fact that several tissue types have similar cell compositions (ie. stomach vs. intestine). With the set of MHBs, we compared the performance between MHL, average methylation in the MHL regions and all the CpGs in the MHL regions, we found MHL and the average methylation could provide similar tissues specificity, while MHL (background noise: 0.29, 95%CI: 0.23-0.35) have a much lower noise compared with average methylation (background value: 0.4, 95%CI: 0.32-0.48). Clustering based on individual CpGs in the blocks had the worst performance which might be due to higher biological noise or technical viability at individual CpG sites (Figure 3B).

The human adult tissues that we included in this study have various degree of similarity among each other, primarily determined by their developmental germ layers of origin. With supervised clustering and feature selection, we identified MHBs that exhibit systematic differences of MHL among the three developmental layers: 38 for mesoderm, 102 for endoderm and 145 for ectoderm (Figure 3D). Interestingly, ectoderm and endoderm were high different with each other while mesoderm tissues were distributed between ectoderm and endoderm. We mapped these layer specific regions to the TF binding regions from ENCODE project and we found 2 Endoderm specific MHL regions which included ESRRA and NANOG. As the previous evidence shown in the absence of Nanog, mouse embryonic stem cells differentiate into visceral/parietal endoderm (Mitsui, K. et al, cell, 2003). Gene ontology analysis shown mesoderm and endoderm shared hypo-MHL regions play the significant role to induce the development of widely tissues. On the other side, Ectoderm specific hyper-MHL regions impress the immune system to induce the ectoderm development (Figure 3E).

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

[WE NEED SOME DESCRIPTION OF OUR RRBS EXPERIMENTS, ESPECIALLY HOW WE HANDLE PLASMA, THE AMOUNT OF DATA WE GENERATED ETC. THAT PART REPRESENTS A DECENT FRACTION OF OUR EFFORTS INTO THIS PROJECT ]

A unique aspect of methylation haplotype analysis is that the pattern of co-methylation, especially within MHBs, is robust in capturing rare events in a mixed population of molecules or cells, in the presence of biological noise or technical variability (ie. incomplete bisulfite conversion, sequencing errors). To explore the clinical utility, we next focused on the methylation haplotype analysis of primary tumor tissues, as well as plasma from cancer patients of which various low fractions (typically <0.1%) of DNA molecules were released from tumor cells. We obtained freshly frozen primary tumor tissues and matched plasma from five patients of each of the three cancer types (lung cancer, colon cancer, pancreatic cancer). We also obtained additional cancer plasma samples (25 colon cancer patients, 24 lung cancer patients, 5 pancreatic cancer patients) without the matched primary tumor tissues, as well as 20 plasma samples from healthy (non-cancerous) donors. To generate bisulfite sequencing data from nanogram level of cell-free DNA from plasma, we applied the scRRBS protocol and obtained an average of XX PE150bp sequencing reads per sample.

We tried to identify some cancer specific MHL regions in the plasma compared with normal plasma samples to investigate the power for the cancer diagnosis with MHL. The genome-wide DNA methylation profiles for 30 colon cancer plasma, 29 lung cancer plasma, 10 pancreatic cancer plasma and 20 normal plasma were established with RRBS. Cancer specific MHL biomarkers were then identified by differential analysis with multiple test correction (FDR <0.05). We identified 4738, 4307 and 5428 significant different MHL regions between colon, lung as well as pancreatic cancer plasma and normal plasma (Supplementary Table 3, 4 and 5), respectively. Based on the significant differential MHL regions, the diagnostic sensitivity for colon cancer, lung cancer and pancreatic cancer could come up to 90%, 93.2% and 95% while the specificity were all as high as 90%. Through scanning positive MHL regions in cancer plasma, cancer tissues, normal tissues, and negative regions in normal plasma as well as WB cells, we could identify large number of tissue derived signals and therefore could be used for non-invasive diagnosis based on DNA methylation in plasma (Figure 4A). Furthermore, we hypothesis that cancer specific hyper-methylated haplotype would be released to plasma randomly and the MHL signals for these regions would be significant higher for cancer plasma than normal plasmas. Therefore, we collected all the MHL regions which contain cancer specific haplotype and we found the MHL signals were significantly higher in cancer plasma compared with normal plasma with P-value of 2.9×10-3, 8.5×10-5 and 2.8×10-2 for colon cancer, lung cancer and pancreatic cancer, respectively (Figure 4B).

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

We collected large number of genome-wide DNA methylation dataset covering about 15 different tissue. We tried to predict the tissues-of-origin for the plasma derived from 3 different kind of tumors (lung, colon and pancreatic cancer) based on tissues specific MHL regions. We collected the 10 tissues with high incidence of cancer as the fundamental components to build a prediction model so that we can capture the origin of the plasma for the cancer patients. 43 WGBS dataset (Figure 4C) was enrolled to be the training dataset to build the random forest prediction model. 1090 tissue specific MHL regions were introduced in the training model and 49 predictors with highest variable importance were selected in the validation stage (Supplementary Table 6). The prediction sensitivity to the plasma tissues-of-origin came up to 72.6%, 85.2%, 63.2% and 85.3% for the plasma from colon cancer, lung cancer, pancreatic cancer and normal individuals (Figure 4D). The methylation status of the informative predictor were found to the significantly different in at least one of the cancer types compared with normal plasma.

## Discussion

**MHBs represent a distinct set of genomic elements.**

DNA methylation between the adjacent CpGs are usually highly correlated. However, the distribution of the linkage disequilibrium were not well profiled. In the present study, we first investigate the linkage disequilibrium of the adjacent CpGs in the whole human genome in the normal tissue samples and defined a specific genomic regions called methylation haplotype blocks (MHB). We demonstrated MHB regions were widely dispensed in human genome, however, they were significantly enriched in some regulatory regions, such as enhancer, promoter while were significantly decreased in LAD, LOCKS regions. In the next step were proposed a metric titled as methylation haplotype load (MHL) which was used to measure the average methylation level weighted by the length of the sequencing fragment. We demonstrate MHL have more powerful ability to indicate the methylation diversity compared with currently average methylation level in specific genomic region, methylation entropy (5) as well as epi-polymorphism (6). Genome-wide analysis to the cancer tissue and plasmas from the cancer patients, we found MHL would be used as powerful biomarker for tissue and development layer distinguish, cancer diagnosis as well as plasma origin tissue prediction.

MHBs defined and identified by the present study represent a distinct set of genomic elements which would be provide important genomic structure or functional elements in human development and complex disease pathogenesis. These regions should be highly linked for the CpGs which might indicate the coordinate dynamic change was required by other cis or trans-regulatory elements in their biological function. MHBs were significantly enriched in some methylation classic functional genomic regions, such as CpG Island, imprint gene regions which were prior expected. It is very interesting, MHBs regions were also enriched in previous identified cancer associated DMR and VMR regions, which suggest MHBs would be taken as important biomarkers for cancer diagnosis and also indicate these DMR and VMR might have more power in the clinical cancer research since these CpGs would have high coordinate changes and its methylation status can be detected with high stability. Furthermore, we also found MHBs significantly enriched in enhancer, super-enhancer, promoter, UTR-5, which indicating MHBs would enriched in cis-regulationary genomic regions. Finally, MHBs were negative correlated with LAD and LOCKS regions which indicating the However, it should be investigated with more comprehensive study, since the genomic regions of LAD, LOCKS were highly large than MHBs, the enrichment analysis between these two kinds of regions might be not accurate.

**Potential of cancer diagnosis based on MHBs.**

## Summary

Similar with the DNA replication, the methylation status of symmetric CpG methylations could be faithfully inherited in a semi-conservative way with the assistant of DNMT1 (13), therefore, the DNA methylation information conform to the classic theory in the population genetics, such as linkage disequilibrium (14). We apply linkage disequilibrium to the methylation alleles on a single DNA molecule and defined methylation haplotype blocks in human genomes. Then, a novel average methylation weighted by order of the CpGs within the genome interval were proposed to quantitatively measure both DNA methylation level and methylation complexity.

## Acknowledgement

## Author’s Contribution

## Competing Financial interests

The authors declare no competing financial interests.

## Abbreviation:

MHB: methylation haplotype load; cf-DNA: Circulating cell-free DNA; RRBS: [Reduced representation bisulfite sequencing](http:///h); MHL: Methylation Haplotype Loading; AUC: Area Under the Curve; TCGA: the Cancer Genome Atlas project; GEO: Gene Expression Omnibus; NSCLC: Non-Small Cell Lung Cancer; ROC: Receiver operating characteristic curve; ACC: Accuracy; SEN: Sensitivity; SPE: Specificity

## Reference

1. Zhao, Y., Sun, J., Zhang, H., Guo, S., Gu, J., Wang, W., Tang, N., Zhou, X. and Yu, J. (2014) High-frequency aberrantly methylated targets in pancreatic adenocarcinoma identified via global DNA methylation analysis using methylCap-seq. Clinical epigenetics, 6, 18.

2. Siegel, R.L., Miller, K.D. and Jemal, A. (2015) Cancer statistics, 2015. CA: a cancer journal for clinicians, 65, 5-29.

3. Hankey, B.F., Ries, L.A. and Edwards, B.K. (1999) The surveillance, epidemiology, and end results program: a national resource. Cancer Epidemiol Biomarkers Prev, 8, 1117-1121.

4. Shoemaker, R., Deng, J., Wang, W. and Zhang, K. (2010) Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome research, 20, 883-889.

5. Hon, G.C., Rajagopal, N., Shen, Y., McCleary, D.F., Yue, F., Dang, M.D. and Ren, B. (2013) Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nature genetics, 45, 1198-1206.

6. Landan, G., Cohen, N.M., Mukamel, Z., Bar, A., Molchadsky, A., Brosh, R., Horn-Saban, S., Zalcenstein, D.A., Goldfinger, N., Zundelevich, A. et al. (2012) Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nature genetics, 44, 1207-1214.

7. Swanton, C. and Beck, S. (2014) Epigenetic noise fuels cancer evolution. Cancer cell, 26, 775-776.

8. Warton, K., Lin, V., Navin, T., Armstrong, N.J., Kaplan, W., Ying, K., Gloss, B., Mangs, H., Nair, S.S., Hacker, N.F. et al. (2014) Methylation-capture and Next-Generation Sequencing of free circulating DNA from human plasma. BMC genomics, 15, 476.

9. (2014) Comprehensive molecular profiling of lung adenocarcinoma. Nature, 511, 543-550.

10. Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W. et al. (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature, 453, 948-951.

11. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485, 376-380.

12. Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M. et al. (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nature genetics, 41, 178-186.

13. Leonhardt, H., Page, A.W., Weier, H.U. and Bestor, T.H. (1992) A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. Cell, 71, 865-873.

14. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. et al. (2001) Linkage disequilibrium in the human genome. Nature, 411, 199-204.

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

## Figure legends

Figure 1. DNA methylation haplotype in human genome inferred by different tissues in different platforms. (A) Workflow diagram for methylation haplotype calling and its application in cancer diagnosis and tumor-of-origin prediction for cancer plasma. (B) DNA methylation blocks were inferred with genome-wide bisulfite-sequencing.The pearson correlation was calculated and empirical threshold were applied for the methylation block inference genome-widely.(C) linkage disequilibrium between adjacent CpG loci in different cells including stem cells and progenitors, normal adult tissues as well as primary tumor. (D). The distribution characteristics of the methylation blocks in human genome including gene region and CpG high density regions. (E). Enrichment analysis to MHB in known regulatory elements. (F). MHB inferred by WGBS data were overrepresented in MHB inferred by RRBS and MH450K data. (G). MHB inferred by MH450K array were highly overlapped in solid tissue and circulating PBMC.

Figure 2. The performance of methylation haplotype load in distinguishing the methylation complexity and methylation frequency compared with other metrics in different scenarios. methylation entropy (ME) and epi-polymorphism were same with previous report and the formula were shown in method section.

Figure 3. Methylation haplotype load has ability in tissue origin, development layer and disease status distinguishing. (A) Unsupervised cluster analysis shown genome-wide DNA methylation haplotype load could represent the sample relationship. The sample with same origin were cluster together preferentially. (B) Layer specific methylation haplotype region showed the development layer relationship. (C) Tissues specific methylation haplotype loading (MHL) shown advantages in tissues distinguish compared with average methylation frequency (AMF) and methylation for all CpG site (MAS). Tissue specificity value (TSV) was the average MHL for the corresponding tissue specific MHL in the correct samples while the background value (BV) were the average MHL in mis-assigned samples. Contract value were the division of TSV by BV. (D). TFBS located in layer specific MHB regions shown different mechanism which MHB involved in layer development. (E). Genome ontology analysis to layer specific and share transcript factors to infer the roles of MHB in layer development.

Figure 4. Plasma signature by MHL with different analysis strategy. (A). schematic diagram to show tissue specific MHL signature were existed in normal tissue, solid tissue and tumor plasma while they are clean in normal plasma and whole blood sample in colon cancer, lung cancer and pancreatic cancer. (B) Tissue specific haplotype regions in plasma show increased MHL level in cancer plasma compared with normal plasma. (C, D) Random forest prediction model for tumor-of-origin prediction based on tissue specific MHL regions. WGBS and RRBS dataset with balanced samples for each tissue were collected to train the prediction model and our lab RRBS dataset were used to validate the prediction performance.