# 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

Table counts: 4

Abstract counts: 250

Mainbody counts: 4000

Running title: Methylation haplotype in human cancer

Key words: 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. It is involved in most human biological and cellular, physiological, pathological changes. Current evidence shows DNA methylation could be powerful tag for cell differentiation, aging estimation, forensic identification and disease status, especially in cancer. DNA methylation can be used as the sensitive biomarker for diagnosis, prognosis surveillance and chemo-response tracking. More and more genome-wide DNA methylation profile for cancers has been completed with different strategies ([1](#_ENREF_1)). Most of the current method is preferred to microarray or capture-based sequencing technology. However, the coverage is limited and the signal is a fuzzy evaluation of DNA methylation for the population of a specific tissue. Although capture-based DNA methylation sequencing technology, such as MBD-seq and MeDIP-seq have low cost, they cannot obtain single-based methylation status, to decrease the cost of the design. In addition, majority of these profiles are based on solid tissues rather than circulating free-cell DNA, which is considered as the most powerful media for non-invasive diagnosis.

As the evidence shown, the survival time of the cancer patients is highly dependent on the stage of the cancer in which the patient were diagnosed, especially for NSCLC, pancreatic cancer and colon cancer([2](#_ENREF_2)). For example, while the overall 5-year survival rates for late stage III and IV of NSCLC patients were just 5%-14% and 1% respectively, the rate could come up to 50% for the early stage of the NSCLC patients who are typically treated with surgery ([3](#_ENREF_3)). However, early diagnosis of cancer also require another 3 prerequisites, including early biomarkers, non-invasive detection and high specificity. The methylation detection in circulating cell-free DNA do provide such platform for cancer non-invasive early diagnosis.

The methylation pattern observed from single-molecule could provide more ultra-information than the present average based methylation level, such as Beta-value (methylation signal vs methylation signal plus un-methylation signals) or M-value (log transformed methylation signals vs un-methylation signals).

While previous studies have defined methylation heterogeneity based mainly on the methylation levels of individual CpG sites, the underlying methylation pattern within each allele and haplotype have not yet been examined. Accordingly, it is important to determine the methylation pattern of consecutive CpG sites within a specific genomic region (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 and have found the prevalence of the methylation interacted with genomic element such as SNPs ([4](#_ENREF_4)). methylation entropy ([5](#_ENREF_5)) and epi-polymorphism ([6](#_ENREF_6)) were also applied to measure the correlation of the adjacent CpGs within a genomic regions. The dynamics of the correlation between adjacent CpGs on single DNA molecules have shown to be widely proved to be associated with cancer evolutionary and heterogeneity ([7](#_ENREF_7)). 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, how to employ such linkage disequilibrium to cancer diagnosis and other clinical application were not considered yet.

The survival time is highly dependent on the stage of the cancer in which the patient were diagnosed, especially for NSCLC, pancreatic cancer and colon cancer([2](#_ENREF_2)). For example, while the overall 5-year survival rates for late stage III and IV of NSCLC patients were just 5%-14% and 1% respectively, the rate could come up to 50% for the early stage of the NSCLC patients who are typically treated with surgery ([3](#_ENREF_3)). However, early diagnosis of cancer also require another 3 prerequisites, including early biomarkers, non-invasive detection and high specificity. The methylation detection in circulating cell-free DNA do provide such platform for cancer non-invasive early diagnosis.

Human peripheral blood contains low levels of DNA molecules from other tissues or cell types, such as circulating cancer stem cells or cell-free DNA (cf-DNA) from apoptotic cancer cells in cancer patients. Analysis of DNA epigenetic mutations in the circulating cell-free DNA is becoming to a prospective trend for creation of noninvasive methods for the diagnosis and treatment efficiency monitoring in cancer. In the past few year, the basic characteristics of cfDNA has been depicted. Plasma rather than serum was considered to be the perfect media to collect the cfDNA within 8 hours of the storage ([8](#_ENREF_8)). Two main fragment components of cfDNA, 180bp and 350bp, could be found in non-white-cell contaminated plasma. The yield of cf-DNA in the plasma of cancer patients is a very low concentration which ranged from 1.0-100 ng/ml while it ranges from 1.0-10 ng/ml in healthy individuals. To detect and quantify such low abundance DNA molecules, some significant regions, hyper-methylated in the circulating DNAs derived from cancer cell while non-methylated in [white blood cell](http://labtestsonline.org/understanding/analytes/wbc/tab/test) (WBC), should be identified.

A range of molecular alterations found in tumor cells, such as DNA mutations and DNA methylation, is reflected in cell-free circulating DNA (cfcDNA) released from the tumor into the blood, thereby making circDNA an ideal candidate for the basis of a blood-based cancer diagnosis test.  DNA mutations driving tumor development and progression are present in a wide range of oncogenes and tumor suppressor genes. However, even when a gene is consistently mutated in a particular cancer, the mutations can be spread over very large regions of its sequence, making evaluation difficult. For example, in lung cancer, the mutation ratio for significant genes are very high, ranging from 2.43% to 47.57%. However, the mutation rate for specific site is very low, as Eric’s previous report ([9](#_ENREF_9)), there are up to 139 mutations in TP53 occurred in 196 NSCLC patients, which means each patients averagely only have 1.41 mutations. However, the DNA methylation own different aberrant characteristics in cancers, as it mainly happened in CpG island, shore, shelf, as well as some other specific genomic regions which can be detected in a convenient way.

In the present study, we carried out a comprehensive genome-wide DNA methylation analysis across 45 plasma from cancer patients, 30 normal plasma and 45 solid cancer tissues with RRBS. 68 cancer tissues and corresponding plasma and 25 normal samples were enrolled in RRBS assay. Methylation haplotype was constructed as our previous method and Methylation haplotype loading (MHL) was proposed to assess the level or the proportion of the DNA methylation. Diagnostic biomarker based on MHL were identified and validated by RRBS dataset.

## Results

#### **Identification of methylation haplotype blocks.**

Methylation linkage disequilibrium (LD) were defined similarly with genetic linkage disequilibrium, which indicate the methylation status of the CpG loci have high pair-wise correlation within a specific region in the human haploid genome. Under different assumption, methylation LD can be inferred with various type of data from different platforms, such as GWBS, RRBS, and even widely prevalent high density methylation microarray data. In terms of GWBS and RRBS data, the correlation between the adjacent CpG loci could be calculated within the reads in a specific individual and thus the methylation LD could be collected when the average correlation is higher than the pre-set threshold, which can be considered as individual level methylation haplotype. On the other hand, methylation LD could also be inferred by methylation array data based on the correlation between adjacent CpG sites in a large human population, which can be termed as population level methylation blocks.

In the present study, methylation blocks were inferred from three kinds of different data including GWBS, RRBS and methylation 450K microarray. GWBS dataset included H1 ESCs, NIH Roadmap Epigenomics Project, and three whole blood WGBS data from Heyn’s lab as well as 10 normal tissues from our own lab. RRBS dataset included 101 samples from ENCODE project. Methylation 450K microarray dataset included 1274 samples (637 cancer and 637 normal samples) from 11 type of cancers and adjacent normal tissues in TCGA project. In the beginning stage, WGBS data were chose to provide completely unbiased methylation block. The cutoff threshold of the average R2 was arbitrarily set as 0.5 after the multiple pre-test from 0.3 to 0.7. It is interesting that we found the average length of the methylation block were almost similar when the R2>0.1 and the total number of the methylation blocks were reduced as the increment of the R2 (Figure 1B). The methylation blocks in some typical cancer genes, such as SEPT9 and SDC2, were manually inspected and to guarantee the genomic length of the methylation blocks were not too long so that the comparison of the methylation haplotype within the methylation block would be reasonable (Supplementary Figure).

In the next stage, similar methylation block identification were conducted in 101 RRBS and 637 methylation 450 microarray data. 23,517 and 2,421 merged methylation blocks were identified in RRBS and MH450K (only normal samples from 12 tissues included) dataset. The methylation of RRBS were measured in Encode cell lines and methylation of HM450K were measured in adjacent normal tissues in TCGA project. We found the methylation blocks were significantly conservative (P-value<10-16). The overlapped methylation block between GWBS and RRBS as well as HM450K are enriched with fold-change of 19.8 and 23.4, respectively. Furthermore, the overlap methylation block between GWBS and RRBS as well as HM450 have significantly strong linked disequilibrium (correlation) compared not shared methylation blocks (P-value<10-22, Figure 1). What’s more, the conservation of the methylation between the different tissues were also assessed in 11 human normal tissues with HM450K microarray data. The number of the shared methylation blocks were significantly deviate from the theoretical distribution (exponential decay) as the increasing of the tissues (P<2.2 × 10-16, chi-square test, Figure 1), indicating the methylation blocks the methylation block have strong conservation in different tissues. The conservation of the methylation block were also found in the comparison between TCGA dataset (solid tissue) with another three peripheral blood mononuclear cell (PBMC) HM450 microarray dataset and we can exclude the probability of the influence from the CpG Island (Figure 1).

#### **Characterizations of MHBs in human adult tissues and stem cells.**

Eventually, 147,888 blocks were identified which is covering about 0.5% of the genome (Supplementary Table). We found methylation blocks widely disperse in human genome, especially in gene regulation and body region (Figure 1). Among all the methylation blocks, 60828 (41.1%) were located in intergenic regions while 87060 (58.9%) regions in transcript regions. 15,712 gene body regions, 14,761 gene promoter regions and 9,006 enhancer regions were covered respectively. Compared with random sampling distribution, we found methylation blocks were significantly enriched in super enhancer regions (enrichment factor=2.3, P<10-6, Permutation test), enhancer regions (enrichment factor=7.6, P<10-6, Permutation test), promoter region (enrichment factor=14.53, P<10-6, Permutation test), CpG island regions (enrichment factor=70.4, P<10-6, Permutation test) and imprint gene regions(Enrichment factor=54.61, P<10-6, Permutation test). In addition, we also investigated the correlation between methylation blocks with previously described important epigenetic modules, such as large organized chromatin K9 modification (LOCK) domains, lamina-associated domains (LADs)([10](#_ENREF_10)), topological domains identified by Hi-C([11](#_ENREF_11)), variable methylation region (VMR)([12](#_ENREF_12)), MHB were significantly deceased in LAD and LOCK regions with 46% and 37% compared with random expectation. We also observed MHB highly enriched in variable methylation regions (VMR) with 25.9 times than the random distribution, indicating the status of MHB were playing important role in cancer initiation and development. In all cases, we can see that MHB are related to, but independent from, each of these previously described domain-like structures (Table). We noticed that MHB were significantly enriched in CpG island regions, therefore the enrichment analysis to the function regions were analysis again to the MHBs beyond CpGI. The enrichment were strengthen in enhancer (7.6 to 9.9), CpG shore (5.4 to 8.0) while the enrichment degree were decrease in promoter (14.5 to 6.9), imprinted regions (54.6 to 29.2). No significant influence to the enrichment in super enhancer (2.29 to 2.15), LAD (0.46 to 0.54), LOCK (0.37 to 0.49).

#### **Characterizations of methylation haplotype load**

Average methylation level and methylation entropy as well as epi-polymorphism could measure either methylation level or methylation dynamics between adjacent CpGs within a genomic regions. However, none platform could measure both of methylation level and methylation dynamic correlation simultaneously. We proposed methylation haplotype load as an average methylation weighted by the CpG position within a genomic region (see method). Compared with correlation coefficient, methylation frequency，methylation entropy, epi-polymorphism, haplotypes counts, only our proposed MHL could separate the 5 scenarios (Figure 2). As we expected, in the second and third scenario, continuous methylated CpGs might derived from cancer genome, therefore, we want to identify these fragment or regions to detect the cancer with circulating DNAs.

#### **Genome-wide methylation haplotype load represent tissue and development information.**

In next stage, we want to investigate whether methylation haplotype load could provide the ability to reflect the tissue and development layer information for the given samples. First, [unsupervised](https://en.wikipedia.org/wiki/Unsupervised) cluster analysis based heatmap plot on top quantile 15% MHL regions shown same origin tissues would cluster together and the samples of H1 and cancer cells were significantly different with other somatic tissues (Figure 3A). Similar result could be found with genome-wide MHL regions (Supplementary cluster tree figure) which indicate the MHL would be used as the biomarker for the tissue identification. Furthermore, tissue specific index (TSI) based on MHL were calculated (see method) to provide the tissue specificity for each MHL regions in certain tissues. We expected the MHL regions with high TSI would distinguish the different tissues with high efficiency. Heatmap plot based on these specific regions shown the tissue the tissues could be assigned with high accuracy with the specificity of 0.89 (95%CI: 0.84-0.93). We also 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 significant lower noise compared with average methylation (background value: 0.4, 95%CI: 0.32-0.48). The totally CpGs have worst performance which might be cause by high variants noisy without the average operation (Figure 3A and 3B).

In the next step, we tried to identify development layer specific MHL regions. 38, 102 and 145 layer specific MHL regions were identified for mesoderm, endoderm and ectoderm. We found these regions would be distinguish three layers clearly, especially, 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 (Figure 4). Gene ontology analysis shown mesoderm and endoderm shared hypo-MHL regions play the significant role to induce the development of widely tissues (Figure). On the other side, Ectoderm specific hyper-MHL regions impress the immune system to induce the ectoderm development.

#### **Detection of tumor methylation haplotypes in circulating DNA.**

In our RRBS dataset, we have 30 colon cancer plasma, 29 lung cancer plasma, 10 pancreatic cancer plasma and 20 normal plasma. Cancer specific MHL biomarkers were then identified by differential analysis with multiple test correction (FDR P-value <0.5). We identified 4738, 4307 and 5428 significant different MHL regions between colon, lung as well as pancreatic cancer and normal (Supplementary Table), respectively. With the most significant differential 15 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%, respectively (See supplementary Tables).

On the other side, the circulating tissue derived free DNA would as carry on certain methylation characteristic with its origin-of-tissue. Therefore, we wonder whether there would be some methylation signals which could be separate the tissue derived materials (solid normal tissue, solid cancer tissue and cancer patient plasma) with non-tissue components (normal plasma and white blood cell). Considering methylation biomarkers were required to be high methylation in cancer plasma and low or non-methylated in normal plasma and WB cells. We identified a panel MHL biomarkers which would separate the tissue derived samples from non-tissue components with the sensitivity ranging from 76%, 86% and 72% for colon, lung and pancreas dataset (Figure 4B and supplementary Table). Furthermore, Although there is no requirement for the diagnosis or prediction biomarker to be have specific biological function, several biomarker were still found to be related with cancer initiation and development, such as NEURL1, FGF12, microRNA-10B. NEURL1 was demonstrated to be a tumor suppressor while FGF12 and microRNA-10B play important roles in cancer development. These evidence indicated that MHL signals from the solid tissue could be transferred to plasma provided a theoretical foundation for the tissue-of-origin mapping for the plasma from cancer patients.

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

We have 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 cancer, colon cancer and plasma 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 genome-wide DNA methylation dataset (Figure) was collected to be the training dataset to build the random forest prediction model.

1090 tissue specific MHL regions were introduced in the training model and 48 predictors with highest variable importance were selected in the validation stage (Supplementary Table). 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. Furthermore, 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. Finally, similar prediction analysis also successfully works in to distinguish cancer samples (plasma and solid tissues) from normal tissues and normal plasmas, indicating the powerful ability of MHL in the diagnosis and tissue-of-origin mapping.

## Discussion

DNA methylation between the adjacent CpGs are usually high 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 regulationary 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](#_ENREF_5)) as well as epi-polymorphism ([6](#_ENREF_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-regulationary 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. We also discovered the enrichment of MHBs in TF binding regions from ENCODE project and \*\*\*. 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.

## 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](#_ENREF_13)), therefore, the DNA methylation information conform to the classic theory in the population genetics, such as linkage disequilibrium ([14](#_ENREF_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://en.wikipedia.org/wiki/Reduced_representation_bisulfite_sequencing); 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.

Tables and Figure legends

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | Colon | Lung | Pancreas | Normal1 | SEN | SPE |
| Colon | 6.98 | 2.62 | 0.35 | 0.05 | 0.69(0.68-0.69) | 0.86(0.86-0.86) |
| Lung | 3.7 | 5.49 | 0.68 | 0.13 | 0.52(0.52-0.52) | 0.89(0.89-0.89) |
| Pancreas | 1.27 | 0.52 | 7.95 | 0.26 | 0.78(0.77-0.78) | 0.96(0.96-0.97) |
| Normal1 | 0.65 | 1.64 | 0.53 | 17.18 | 0.83(0.83-0.84) | 0.99(0.99-0.99) |

Random forest were conducted with 100 times to make sure the prediction model were with high reproducibility and the average prediction number were recorded in the tables.

In the second stage, RRBS data from 20 colon cancer, 19 lung cancer were collected. 58 MHL features were positively selected in the random forest prediction model (see supplementary Table)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Colon | Lung | SEN | SPE |
| Colon | 19.42 | 0.58 | 0.83(0.81-0.85) | 0.91(0.9-0.91) |
| Lung | 0.82 | 18.18 | 0.73(0.7-0.76) | 0.93(0.92-0.94) |

Random forest were conducted with 100 times to make sure the prediction model were with high reproducibility and the average prediction number were recorded in the tables.

When we merge the stage 1 and stage 2 samples together (pancreatic cancer plasma excluded since the sample size incomparable with other samples).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  | Colon | Lung | Normal | SEN | SPE |
| Colon | 21.72 | 1.7 | 6.58 | 0.79(0.78-0.81) | 0.91(0.9-0.91) |
| Lung | 4.26 | 18.04 | 6.7 | 0.69(0.67-0.72) | 0.94(0.93-0.94) |
| Normal | 0.12 | 0.54 | 19.34 | 0.92(0.92-0.93) | 0.90(0.89-0.92) |