It could be expected that the methylation blocks would be different considering the context of the epigenomes changes in different physiologies and tissues. However, the conservation of the methylation blocks were evaluated between different dataset and different tissues. We found the methylation blocks were significantly conservative between different dataset compared with the distribution of the random sampling (Figure)

Furthermore, to examine MHBs in cancers and explore its clinical utility, we generated 162 sets of RRBS data on primary tumor tissues and matched plasma from patients of lung cancer, colon cancer and pancreatic cancer, as well as control plasma from healthy individuals. Compared with normal tissues and stem cells, primary tumor tissues exhibit a distinct methylation pattern within MHBs, related to locally disordered methylation recently discovered in CLL. Importantly, among the set of MHBs that we identified from normal adult tissues, we derived a subset of blocks that can predict the tumor tissue of origin from circulating DNA in the plasma. Finally, we demonstrated the detection of tumor-specific methylation haplotypes in the plasma at various low levels, which are otherwise difficult to be separated from biological or technical noise when analyzing individual CpG sites.

[Background] Circulating cell-free DNA methylation in plasma have been demonstrated to be powerful potential in non-invasive cancer early diagnosis. However, the genome-wide profile of DNA methylation for the free circulating DNA methylation has not been depicted in a single-base resolution. [Method] In the present study, we carried out a genome-wide survey of single-base resolution methylome across 45 plasma from cancer patients, 30 normal plasma and solid cancer tissues with GWBS and RRBS assays. [Result] In the discovery stage, both RRBS and SeqCap dataset identified large number classification potential biomarkers (N=1592 and 516, respectively) and high level prediction ability with Random Forest model (accuracy=100% and 98.18%, respectively). The above biomarkers were validated in the BSPP dataset and we found the methylation status of 30 derived regions in plasma could explain 93.75% cancers incidence and the specificity is 100%. What’s more, the pattern of the DNA methylation fragment transferring were depicted with paired tissue-plasma methylation dataset. We identified 1590 fragments significantly prefer-selected in the releasing process of the DNA methylation fragment from solid tissues to plasma (P<1.98\*10-6, binomial test, [Bonferroni correction](http://en.wikipedia.org/wiki/Bonferroni_correction)). These fragments were located in the regions of 1190 genes. Function enrichment analysis showed these genes were significantly associated with cancer relevant biological functions, including embryonic morphogenesis, regulation of transcription, neuron differentiation, regionalization, tissue morphogenesis, transcription factor activity, sequence-specific DNA binding, transcription regulator activity. What’s more, our analysis showed MHL could provide certain evaluation to the degree of the leukocytes contamination to the plasma cell-free circulating DNA which would provide useful application in the process of the clinical operation of cancer diagnosis [Conclusion] Methylation haplotype loading based DNA methylation biomarker would be potential cancer diagnosis biomarker and DNA methylation fragment releasing process were somehow regulated by certain uncovered mechanism rather than a stochastic event.

we found the top 289 most importance regions could take account of 80.0% contribution to the accurate prediction.

Random forest algorithm showed 1585 regions could provide positive ability to distinguish cancer samples from normal samples while the top 286 most importance regions could take account of about 80.0% contribution, with sensitivity of 97.06%, specificity of 100% and accuracy of 97.37%.

The second round prediction process of random forest model based on top 206 regions with mtry of 14 and mtrees of 500 showed 100% sensitivity, 100 specificity and 100% accuracy. One the other side, 248 regions were hypermethylated in at least 50% cancer samples. The average length of the regions were 103bp (IQR=95bp, SD=126) and were located in the promoter region of 182 genes (2000bp up-stream of TSS). With the help of text mining, 21 genes of them were validated to be methylation relevant cancer related genes (Table).

Next-generation methylation sequencing and quality control

Unique mappable reads

In the first step,

In the next step, genome-wide DNA methylation profiles of 30 samples including 15 solid cancer tissues (5 colon cancer, 5 lung cancer and 5 pancreatic cancer) and corresponding plasmas detected with [reduced representation bisulfite sequencing](http://en.wikipedia.org/wiki/Reduced_representation_bisulfite_sequencing) (RRBS) were collected to discover the pattern of the shedding for the methylated DNA fragments from tissues to blood.

Task 1.

1, You need to filter all the hypermethylated fragement in cancer solid tissues and corresponding plasma circulating DNA while no methylation signals in health plasma.

2, and then validation these signals in TCGA database.

Task 2. Different method comparison.

Can it be used on RRBS data? We can downsample high coverage data to lower the effect of clonal reads on the analysis.

Noi to check the input requirement for performing RRBS, can we perform RRBS in parallel on the test samples.

The optimal number of variables tried at each split is 16 (mtry) and number of trees (mtrees) makes no difference to the prediction accuracy from 500 to 7000. RF prediction model showed 1585 regions could provide positive ability to distinguish cancer samples from normal samples with high specificity of 100%, however, the sensitivity was only 25%, which indicated large number low predictive biomarkers were enrolled into the prediction model. After removed the most 60% lower informative regions in the random forest model the sensitivity was only 78%, (Informative regions from BSPP see supplementary Table \*)

Therefore, we merged with the informative biomarkers identified in RRBS and Cap-seq dataset, eventual, there are 2 biomarkers were remained, including:

In the prediction section, top 62 regions could provide the distinguish accuracy of 86.87% which did not indicated Capseq method were worse than RRBS in the prediction ability. First capture, On the another side, when you will top 10 predictive regions, the separate ability of prediction model based on Capseq dataset could come up to the sensitivity of 90.32%, specificity of 95.83% and accuracy of 92.49%. However, small number of predictors would greatly decrease the robust or reproducibility of the prediction model and would bring biased inference in the process of clinical application of Capseq assay.

63.4% MHL distance of BSPP were less than 200bp, indicating they are almost located in the same region/CpG island in human genome, therefore, we need merge these MHL region together to increase the sensitivity of the prediction.

We need to look into each cancer individually on the list, maybe rank the cancer samples from most preferable to least preferable. Once we get optimistic results, we can further ask for samples from other centers. We cannot realistically ask specifically for each stage of cancer, although having samples from multiple stages might be better.

prevalence

treatment options for early detection (stage 1)

current diagnostic methods

**We need to ask for buffy coat (1 tube) and matched serum samples (2 tubes) per patient**

**We need to ask for primary tumor samples**, they might be Formalin-fixed paraffin-embedded (FFPE) samples

check Blueprint data to see how successful we were at capturing DNA purified from FFPE samples

look into the kit for fixing degraded DNA

Noi said we successfully captured with **50 ng converted DNA** before, and it is possible to obtain this much from 2 tubes of serum samples, however, we can also look into the amplified DNA from Illumina (can we perform capture on these samples?)

RRBS option

Dinh need to check the bayes classification algorithm

Can it be used on RRBS data? We can downsample high coverage data to lower the effect of clonal reads on the analysis.

Noi to check the input requirement for performing RRBS, can we perform RRBS in parallel on the test samples.

Can it be used on RRBS data? We can downsample high coverage data to lower the effect of clonal reads on the analysis.

Noi to check the input requirement for performing RRBS, can we perform RRBS in parallel on the test samples.

#### **The stochastically mechanism of the releasing process of tumor DNA to plasma**

## Method and Materials

#### **Clinical sample and DNA collection**

#### **Methylation** **haplotype**

 methylation LD, the pair-wise LD for all sites in a bin should be above a certain threshold. Formally, we can partition all CpG sites in the entire genome into methylation LD blocks. Each block would be a bin

 derive the consensus haplotypes from multiple clonal reads based on read quality

 To be completely unbiased, we want to use WGBS data in the ideal situation. We have the N37 WGBS data from ten human tissues, plus the Heyn2013 whole blood WGBS data. All these data have been mapped. The Epigenomics Roadmap project just release a large number of WGBS data sets, but Dinh hasn't completed the mapping. These are all from non-cancerous tissues.

* Generate hapInfo file from each bam file (use the latest Jan2015 scripts, which derive the consensus haplotypes from multiple clonal reads based on read quality).
* Identify informative regions based on the primary tumor and normal plasma data. Do this for RRBS and WGBS/LC separately, then merge the regions.
* Generate another batch of hapInfo files for these informative regions only.
* Run mixMethHapAnalysis and also determine tumor HMH load in patient plasma samples
* 101 RRBS data were downloaded. 2,646,999 CpG loci were covered by 101 RRBS data while 866,979 CpG loci (32.8%) were detected in at least 80% samples.

In the methylation anlaysis method which is based on next generation sequencing, such as BS-seq, RRBS, BSPP, a genomic region would be covered with lots of sequencing reads and in each reads might cover several CpG sites whose methylation status would be different for different CpG sites or different reads in same patients (reflecting the heterogeneity of cells). Methylation haplotype is the analogous concept in human genetics which indicate the combination composed by the methylation status of multiple CpG loci.

Fastq and bed files can be downloaded from Encode Project. 101 RRBS data (bed files) were downloaded. 2,646,999 CpG loci were covered by 101 RRBS data while 866,979 CpG loci (32.8%) were detected in at least 80% samples. 46541 HDRs and 23517 methylation blocks were identified. Significant decreasing of average correlation within the methylation block were observed when the CpG number (P<2.0×10-16, Beta=-0.0045) and the length of the methylation block region (P<2.0×10-16, Beta=-0.0054) was increasing while there is no any correlation with the density of the CpG sites (P=0.473).

The advent of massively parallel sequencing provided large number of resource to make deep analysis of methylation haplotype between difference tissues, status (cancer, damage and immune-disease).

#### **Tracing of the origin of plasma with specific hyper-methylation haplotype regions**

GSI (Group Specificity Index) were applied to measure the tissue specificity for each hyper-methylation blocks.

indicates the number of the groups. denotes the average of MHL of group. denotes the average of MHL of highest methylated group. The statistic has been ever applied to identify tissues specific expression genes ([10](#_ENREF_10)). Here, the regions whose GSI>0.6 were assigned as hypermethylated methylation blocks regions of corresponding cancers/tissues. The performance of distinguish to the tissue origin which is based on methylation haplotype load was pretty well (Figure) while it is very limited with raw methylation signal or average methylation level (AML).

#### **Methylation sequencing assays.**

#### **The definition and statistical assessment of methylation haplotype to identify rare heterogeneous fragments**

WGBS data of ten normal tissues (colon, frontal lobe, heart, small intestine, liver, Lung, skeletal muscle, pancreas, stomach),

two public dataset: three whole blood ([11](#_ENREF_11)) and 36 WGBS from Epigenomics Roadmap project.

RRBS

RRBS, SeqCap, Umi-BSPP

#### BSPP and Methylation Haplotype Loading (MHL)

To detect and quantify such low abundance DNA molecules, we focus on regions in the genome in which there are major differences in DNA methylation between whole blood and cancer cells which can be used for assay and detection. For example, a region containing 6 CpG sites might be completely unmethylated in whole blood, and fully methylated in cancer cells. If the whole blood sample contain 3% of cancer DNA, then we would detect a 3% methylation with an ideal assay. However, all methylation assays have technical errors, such as incomplete bisulfite conversion, incomplete enzyme digestion, sequencing error. Typically all the technical errors combined can contribute to ~1-2%. With the presence of these errors, a 3% methylation cannot be confidently detected. Such technical errors greatly compromised the sensitivity and confidence in detecting and quantifying fetal DNA molecules. Methylation haplotyping analysis can dramatically improve the discriminating power, as technical errors typically occur independently on all DNA molecules at random locations. In contrast, the 3% fetal DNA molecules are fully (or almost fully) methylated at all CpG sites, whereas all maternal DNA molecules are not methylated. In other words, the methylation status at multiple CpG sites on the same molecules are all “linked”. Using methylation haplotypes of four or more CpG sites, one can confidently identify rare DNA molecules at 0.01% or even lower, at least two orders of magnitude below the technical errors (1-2% per site). We were the first group that developed an analytical framework for methylation haplotype and linkage disequilibrium analysis (Shoemaker et al, 2009). In this invention, we developed an assay for digital quantification of methylation (see below), and greatly extended the previous framework to haplotype-based methylation marker analysis. In addition, combining information from multiple markers will further improve the sensitivity and robustness in the presence of biological variability. Finally, cell-free DNA typically come from apoptotic cells, and are in small fragments (Chan et al. 2004). In contrast, whole blood DNA typically have larger sizes (at least kilobases) even after DNA extraction. Using targeted methylation sequencing to analyze haplotypes from DNA molecules of different sizes adds another level of stringency in separating rare cancer or fetal DNA from whole blood DNA. In summary, this invention can achieve an ultra-high sensitivity for detecting rare DNA species from mixed DNA (such as whole blood DNA) using some combinations of the three concepts: (i) multi-locus methylation haplotype analysis (**Figure 1**); (ii)integrative analysis of multiple marker regions; and (iii) differential haplotype analysis of DNA fragments with different sizes (**Figure 2**). The invention can be implemented as genetic screening or diagnostic tests for non-invasive prenatal diagnosis, non-invasive monitoring of tumor loads in cancer patients after treatments, or early-stage cancer detection.

We have implemented this invention using a target methylation sequencing technology (Bisulfite Padlock Probes, or BSPP, Deng et al, 2008; Diep et al. 2012) developed by the Zhang lab. Note that alternative technologies, such as micro-droplet PCR (Komori et al. 2011) or Selector probes (Johansson et al. 2011), can also potentially be used with some differences in the requirement of input materials and/or cost. MeDiP is another alternative experimental method for data collection (Papageorgiou et al. 2010), with disadvantages including low efficiency, high cost, and low reliability (Tong et al. 2012). Therefore, this invention should cover the aforementioned three key concepts, not a specific implementation based on the BSPP technology.

Regardless of the specific sample preparation methods or sequencing platforms used, our method takes the bisulfite sequencing reads (single-ends or paired-ends) as the input. We derived methylation haplotypes and their abundance from the raw sequencing reads. Each haplotype represent the combination of binary methylation status (methylated or unmethylated) at multiple CpG sites of one sequencing read. For sample preparation methods (such as umi-RRBS, or hybridization-based target capture) that allow identifying multiple clonal sequencing reads originated from the same template DNA molecules, we also derive the consensus haplotypes from the clonal reads to improve the accuracy and avoid over-dispersion of haplotype counts. The probability that a methylation haplotype is present in a sample (or a pool of samples) is determined by the frequency if the exact haplotype is observed, or estimated from the methylation levels of individual CpG sites and the technical error rates (sequencing errors, bisulfite conversion errors) assuming no linkage between adjacent sites (or P(M1M2)=P(M1)\*P(M2) where M1M2 is the probability of two-locus methylated haplotype, P(M1) and P(M2) are the methylation level at the two loci). For each methylation haplotype from the patient plasma of interest, we determined the likelihoods of it originating from the pooled tumor primary biopsies data and from the pooled normal plasma data, and calculated the negative log likelihood ratio. A methylation haplotype is classified as the tumor haplotype when the negative log likelihood ratio is above 3. To improve the signal-to-noise ratios, we focus on methylation haplotypes that contain four or more CpG sites.

We have demonstrated the proof-of-concept for detecting low-abundant tumor DNA in blood DNA based on methylation haplotypes, both in synthetic DNA mixtures and clinical plasma samples of three cancer types.

For the first demonstration using synthetic DNA mixtures, we searched marker regions by extensive analyses of published and unpublished DNA methylation data on whole blood, cancers, and placenta. We have identified 12,446 candidate regions that exhibit large methylation difference between whole blood and three types of cancer cell lines (pancreatic cancer, glioblastoma multiforme, lung cancer), as well as 1,230 regions different between whole blood and placenta. We designed umi-BSPP (Figure 3), which is an improved version of Bisulfite Padlock Probes (BSPP, Deng et al, 2009; Diep et al. 2012) for targeted methylation sequencing of these candidate regions of whole blood and a panel of five cancer (pancreatic cancer and glioblastoma) cell lines. These probes also contain built-in Unique Molecule Identifier (Kivioja et al. 2011), so that we can perform deep sequencing and true single-molecule counting to avoid quantification artefacts due to DNA amplification. We have designed and synthesize 19,109 long oligonucleotides for the candidate targets.

Computational analysis of methylation haplotypes starts with bisulfite sequencing reads mapped to the reference genome (using common bisulfite read mapping algorithms, such as bisReadMapper, bisMark) in the format of bam files. We derived consensus sequencing reads based on UMI (if available), then determined the methylation haplotypes on multiple CpG sites in single consensus sequencing reads (Figure 4). The haplotypes and their counts in all genomic regions assayed are reported. For selection of the marker set that can classify plasma samples, we define a methylated haplotype load (MHL) for each candidate region, which is the normalized fraction of methylated haplotypes at different length:

Where s the length of haplotypes, is the fraction of fully methylated and un-methylated haplotype with i loci. For a haplotype of length L, we considered all the sub-strings with length from 1 to L in this calculation. is the weight for i-locus haplotype. We typically used or to favor the contribution of longer haplotye. After calculating MHL for all candidate regions for all samples, we built a MHL matrix for feature selection, using standard machine learning approaches such as SVM and random forest.

As the principle of the traditional Shannon entropy, methylation entropy (ME) for haplotype variable in specific genome region could be calculated with the following formula:

Suppose, there is a genome region with CpG loci and methylation haplotype, then represent the probability for methylation haplotype and it can be calculated with the number of reads for such haplotype and the total reads in this genomic region. Since the proposed methylation entropy has been normalized by the number of CGs, the value of ME ranges from 0 to 1 and therefore can be applied in the genome-wide scales. Methylation entropy were widely used in the measurement of variability of DNA methylation in specific genome regions ([12](#_ENREF_12)).

The characteristics of the free circulating DNA derived from apoptotic or necrotic were quite different.

In RRBS, DNA fragments of 40-220 base pair are representative of the majority of promoter sequences and CpG islands

The methylation status of target genes in circulating DNA could be evaluated by two kinds of methods including bisulfite assistant methylation assay (RRBS, BSPP) and bisulfite-free methylation assay (MeDIP, MBD-seq).

The technique of DNA methylation has been widely applied to identify clinical associated biomarker or monitor biomarkers for disease prognosis. Genome-wide DNA methylation profiles has been in the solid tissues ([1](#_ENREF_1)).

The main contribution of our study includes as the following:

The methylation haplotype loading is significantly associated with cancer progress, such as TNM stages.

In present study, single-base genome-wide DNA methylation analysis were conducted in 75 solid cancer, adjacent tissues, cell-free circulating DNA from cancer and normal samples. Methylation haplotype based DNA methylation diagnosis to cancer was evaluated and most powerful biomarker were identified

#### **Bioinformatics and Statistics**

In the biomarker discover stage, the biomarker whose methylation were 100% un-methylated in normal plasmas were enrolled while any regions which were detected to be methylated in normal plasmas were filter out from the candidates. In the procedure of differential methylation test, regions whose variance in total samples were stage at lowest 30% quantile were removed to decrease the burden of multi-test correction. For the random forest prediction, optimal parameters were tuned before the tanning of the model the best number of tries and tress were determined by the grid search method with lowest out of beg prediction error. Intersection analysis of the genome position were conducted by Bedtools and the regions whose distance less than 25bp were taken as the same biomarker region.

Methylated haplotype load (MHL) was defined as

R packages of IRanges, Biostrings, stringr, randomForest, impute, rpart, e1071, biclust were used in the the statistic and bioinformatics analysis.

Potential biomarker based on published papers from NCBI were extracted for lung cancer (74 papers), colon cancer (8 papers) and pancreatic cancer (13 papers) with the strategy of title including “methylation” and corresponding cancer symbols as well as abstract including “Diagnosis”. \*\*