Methylation haplOtyping for Non-invasive Diagnosis (MONOD)

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 from apoptotic cancer cells in cancer patients, or fetal DNA in pregnant women. 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 other cells of interest (such as cancer or fetal cells). These are the marker regions used for assay and detection. For example, a region containing 6 CpG sites might be completely unmethylated in whole blood, and fully methylated in placenta. If a maternal whole blood sample contain 3% of fetal 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 distriminating 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.

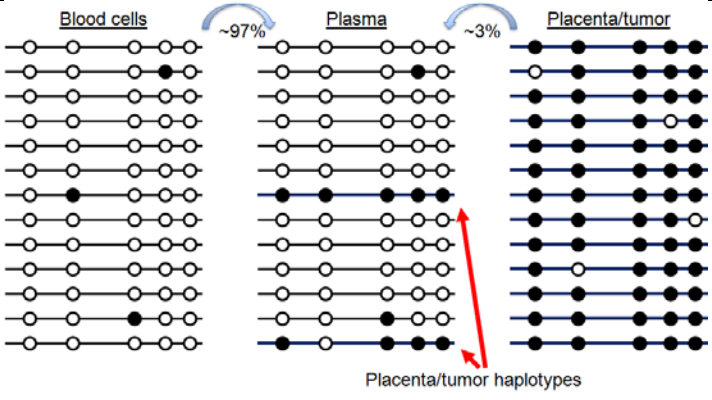


Figure 1. Sensitive detection and quantification of circulating placenta or tumor DNA in blood plasma using DNA methylation haplotypes. Each line represents a single DNA molecule, and each circle is a CpG site. Open circle is unmethylated CpG and solid circle is methylated CpG.

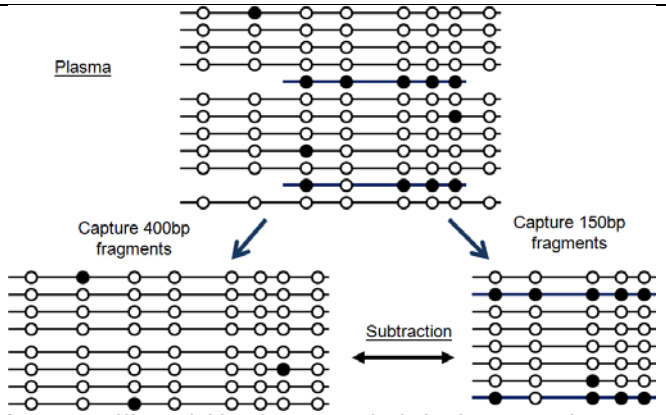


Figure 2. Differential haplotype analysis by fragment size. Targeted capture and sequencing of long and short fragments improve sensitivity by subtraction.

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:

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

Where i is the length of haplotypes, $P(MH_i)$ is the fraction of fully methylated haplotype with i loci. For a haplotype (a string) of length L, we considered all the sub-strings with length from 1 to L in this calculation. w_i is the weight for i-locus haplotype. We typically used $w_i=i$ or $w_i=i^2$ to favor the contribution of longer sub-strings. 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, random forest or MRMR.

We performed multiple rounds of screening for a subset of probes that are sensitivity, robust and informative. A set of 2,489 cancer probes were selected for detecting various levels of cancer and whole blood DNA mixtures (0%, 0.2%, 0.5%, 1%, 2%, 5%, 10%, 20%) (Figure 5). This first proof-of-concept experiment clearly indicated that multi-CpG methylation haplotypes are very effective in separating cancer DNA from whole blood for sensitive and quantative detection.

We next asked whether we could identify fetal specific haplotypes and use the sequencing reads mapped to these haplotypes for quantifying fetal chromosome dosage for non-invasive prenatal diagnosis of aneuploidy such as chromosome 21 trisomy in blood. By subsampling the captured sequencing reads that mapped to individual chromosomes and generating synthetic mixture of cancer trisomy 18 and 21 with diploid whole blood DNA, we were able to detect significant gains of a single copy at 2% or 5% levels (Figure 6). These results demonstrate that quantatitive detection of aneuploid in DNA/cells that carry methylation signatures distinct from whole blood is possible with methylation haplotypes. The detection sensitivity can be further improved by additional optimization of the marker selection and refining the statistical model when larger data sets are available.

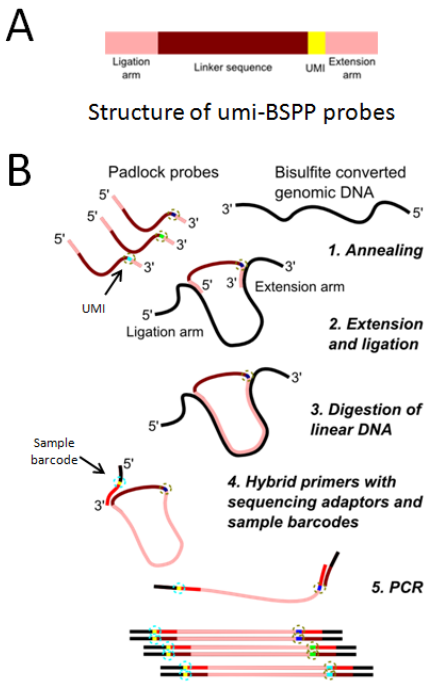


Figure 3. Multiplex capture of MONOD marker panel with umi-BSPP. (A) Each umi-BSPP probe contains a randomized unique molecule identifier (UMI) for uniquely label the capture product from each template DNA molecule. (B) Bisulfite converted DNA from clinical specimens is converted into sequencing libraries with 5 steps in single-tube reactions.

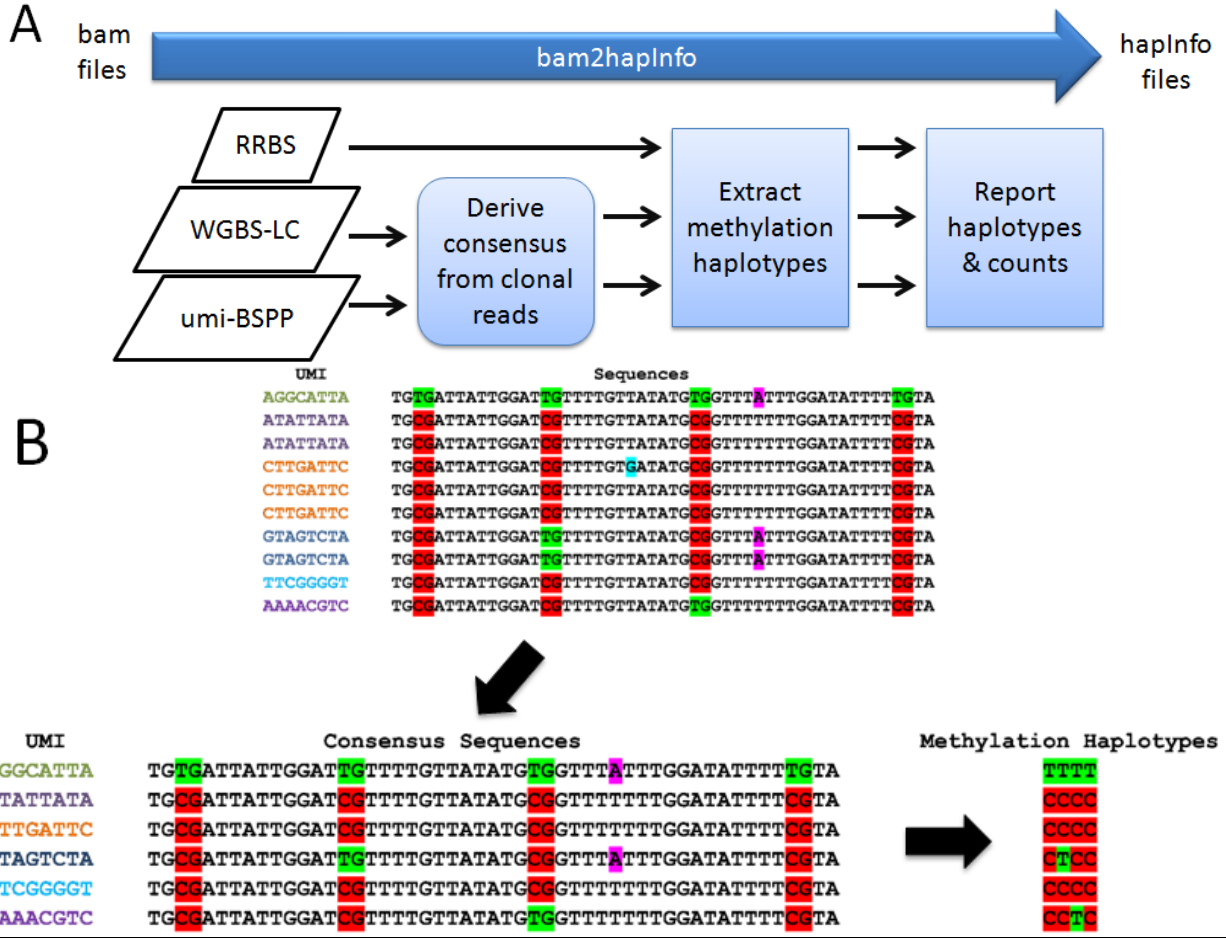
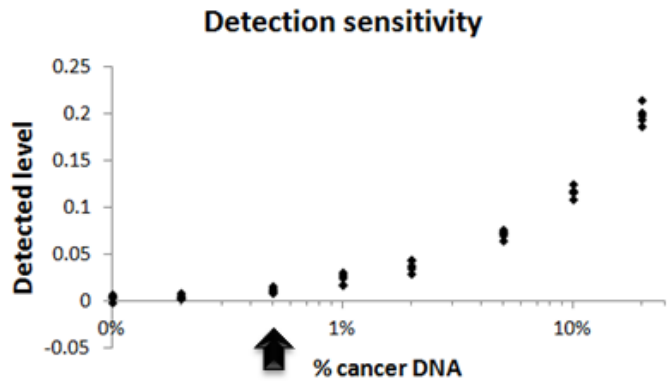


Figure 4. Deriving methylation haplotypes from sequencing data. (A) The bam2hapInfo program takes mapped raw sequencing data in the bam format, reports haplotypes and their counts in hapInfo files. (B) A example on deriving 4-locus methylation haplotypes from raw sequencing reads.



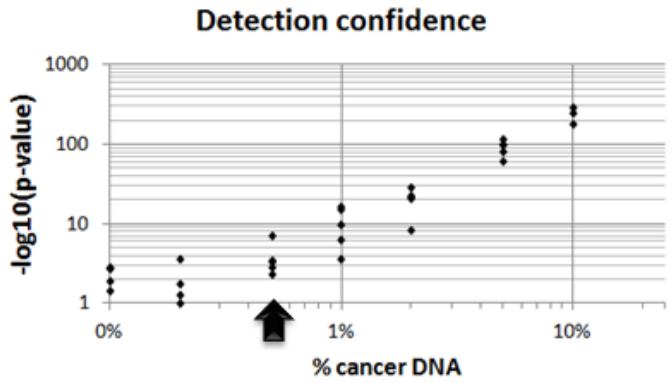
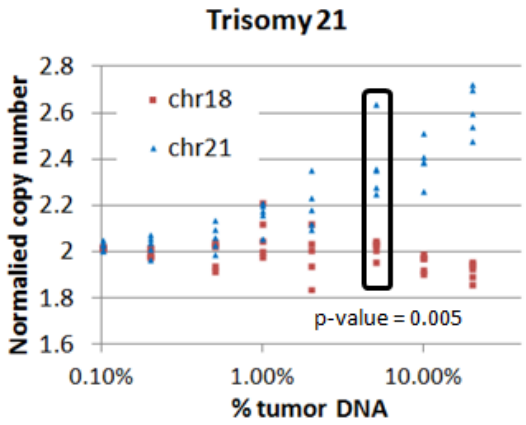


Figure 5. Quantitative detection of cancer DNA in whole blood. Top panel: level of cancer DNA detected in mixture of cancer and blood DNA at various ratios. Bottom panel: statistical significance of detection.



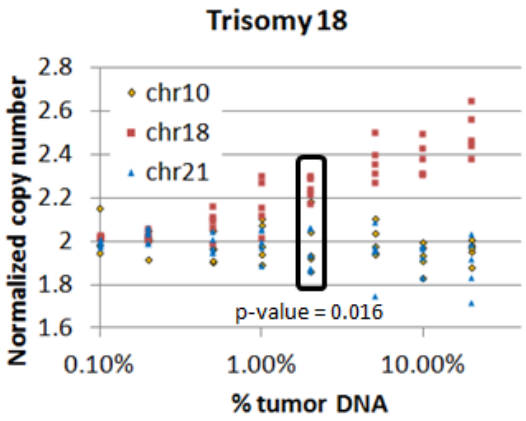


Figure 6: Detection of cancer specific aneuploidy on chromosomes 21 and 18.

We next sought to further demonstrate the feasibility of detecting tumor signatures in clinical samples based on the methylation haplotypes. We obtained 30 patient plasma samples, 10 from each of the three cancer types: colon cancer, lung cancer and pancreatic cancer. We also collected 15 fresh frozen primary cancer tissues, 5 per cancer type from the same patients that plasma samples were collected. In addition, we included 30 plasma samples from normal individuals with no detectible cancer as the negative controls.

We took an unbiased approach to systematically screen for candidate markers appropriate for MONOD. In the first strategy, we analyzed the published whole genome bisulfate sequencing data from human whole blood, and searched from blood unmethylation regions (UMRs). These UMRs represent the “windows” in the human genome in which tumor methylated haplotypes are not masked by blood methylation signatures if they are present in peripheral blood. A customized NimbleGen probe set was designed to capture a total of 29Mb blood UMRs for the entire panel of 75 samples for bisulfite sequencing. In addition, we used a second strategy to scan the CpG dense regions (including most CpG islands) for marker identification. We performed reduced representation bisulfite sequencing (RRBS) on the same panel of 75 samples. From the sequencing data generated by the two experimental approaches, we searched for regions where methylated haplotypes are detectible in primary tumor samples but present at a low level or non-detectible in plasma samples from normal individuals. This genome-wide screening identified a set of 5,301 candidate regions within blood UMRs, and 9,935 candidates from the RRBS data sets. Merging these two candidate lists yielded 11,901 candidate regions where we can screen for tumor specific haplotypes in plasma. From the 30 cancer patient plasma we have in this pilot, we have identified 5,871 regions within which various levels of tumor methylated haplotypes are detectible from patient’s plasma. Figure 7 shows an example at the promoter region of RhoB, which has been previously identified as a marker for cancer progression. Targeted methylation sequencing probe sets focusing on these 11,901 regions are being designed and optimized for examining larger sets of plasma samples from cancer patients and normal controls, in order to finalizing a specific set of markers with high sensitivity and specificity.

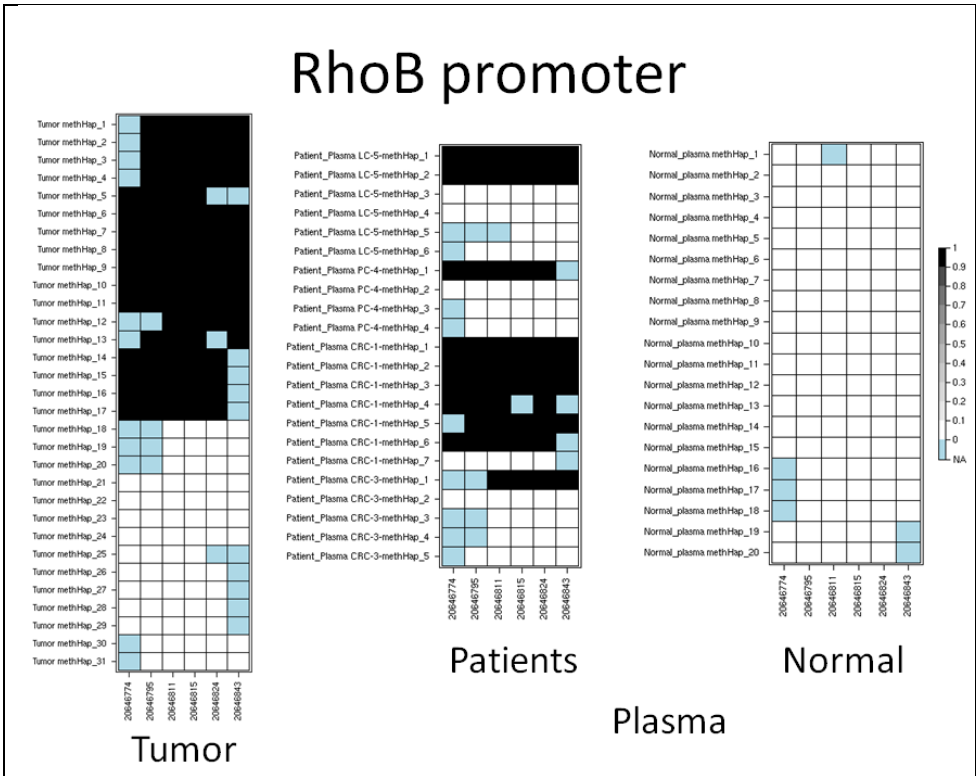


Figure 7. Tumor-specific haplotypes detected in the plasma of four cancer patients. Each row represent a methylation haplotype derived from a unique DNA molecules in the sample. A CpG site is represented as a square. Black is methylated, white is unmethylated and light blue is missing value.