There is a need for improved methods for increasing sensitivity and reducing cost of using circulating DNA for cancer detection and screening assays for liquid biopsy.

SUMMARY OF THE INVENTION

The present invention is directed to methods for enriching CpG-rich regions of circulating cell-free DNA, that are known to harbor significant methylation changes in cancer irrespective of their methylation status, to provide methylation information that can be used for cancer-specific and other screening.

According to a first aspect, the present invention provides a method of enrichment of CpG islands comprising cancer-specific methylation information in isolated circulating cell-free DNA from a subject, comprising the steps:

i) provide a cell-free DNA sample;

ii) repair double-stranded DNA ends and add dA tail;

iii) heat-denature the cell-free DNA, wherein denaturation to single strands increases and remaining double stranded decreases with a decrease in GC content and denaturation to single strands decreases and remaining double stranded increases with an increase in GC content;

iv) ligate methylated adapters to both ends of the double-stranded DNA;

v) perform bisulfite conversion of the adapter-ligated DNA;

vi) amplify the bisulfite-converted adapter-ligated DNA from v);

vii) size select for 190-400 bp fragments of the amplified DNA of vi).

In some embodiments, low GC content fragments having lower than about 60% GC content are denatured and high GC content fragments having about 60% or higher GC content remain double stranded.

In some embodiments, the heat denaturing in step iii) is performed at a temperature in the range of about 87-92 ºC.

In some embodiments, the cell-free DNA sample is from a subject that has a medical condition.

In some embodiments, the method further comprises determining at least part of the sequence of one or more of the amplified molecules.

In some embodiments, the determining of at least part of the sequence comprises paired-end sequencing.

In some embodiments, the determining step provides diagnostic information for the subject.

In some embodiments, the diagnostic information comprises cancer diagnosis information for the subject.

According to a second aspect, the present invention provides a method to determine the tumor fraction in a circulating cell-free DNA sample from a subject, the method comprising:

A) identify differentially methylated clusters from a comparison of Whole Genome Bisulfite Sequencing datasets from normal plasma datasets with reference cancer methylation datasets;

B) compare each CpG site in each bisulfite-sequenced heat-enriched CpG-rich DNA fragment from said subject sample against the reference dataset and calculate, using a bimodal distribution, a class-specific probability for each bisulfite-sequenced heat-enriched CpG-rich DNA fragment; and

C) approximate the tumor fraction in said subject sample using maximum likelihood estimation.

In some embodiments of the second aspect, the method comprises:

A) identifying differentially methylated clusters in circulating cell-free DNA between normal subjects and subjects with cancer by;

(i) obtaining a normal plasma whole genome bisulfite sequencing methylation dataset;

(ii) obtaining a reference cancer methylation dataset and extrapolating, to ± 100 bp of each of a plurality of probe sites, methylation values with a standard deviation less than 0.4 between the various samples in the dataset;

(iii) identifying differentially methylated clusters from datasets (i) and (ii);

B) determining the class-specific probability of each site using a bimodal distribution by;

i) assigning to each sequenced fragment from the subject sample a normal and tumor class-specific proportional methylation status using the generated reference;

ii) for every site in the reference, the contribution from the unmethylated and methylated mode (0 and 1) is calculated, wherein the relative contributions of each mode in the two classes is used to assign normal or tumor class-specific probabilities for the methylation values in the assayed fragment; and

C) estimating the tumor fraction, denoted as θ, where $0 \leq \theta < 1$, of the sample;

wherein, each read is assumed to be independent, and normal and tumor class-specific probability is assigned to each read;

a global tumor fraction (θ_g) is calculated by applying a grid search to determine the highest probability event from a range of tumor fractions.

In some embodiments, in A)(iii) all CpG sites within clusters that have 0.5 difference in methylation are selected.

In some embodiments, each dataset from normal plasma is from circulating cell-free DNA from a normal subjects.

In some embodiments, each reference cancer methylation dataset is a colorectal adenocarcinoma (COAD) dataset.

According to a third aspect, the present invention provides a method of obtaining information in relation to a medical condition of a subject, the method comprising:

i) provide an isolated cell-free DNA sample from the subject;

ii) repair double-stranded DNA ends and add dA tail;

iii) heat-denature the cell-free DNA, wherein denaturation to single strands increases and remaining double stranded decreases with a decrease in GC content and denaturation to single strands decreases and remaining double stranded increases with an increase in GC content;

iv) ligate methylated adapters to both ends of the double-stranded DNA;

v) perform bisulfite conversion of the adapter-ligated DNA;

vi) amplify the bisulfite-converted adapter-ligated DNA from v);

vii) Size select for 190-400 bp fragments of the amplified DNA of vi);

viii) determine at least part of the sequence of one or more of the amplified molecules;

determine the tumor fraction in the sample from the subject, comprising:

A) Identify differentially methylated clusters from a comparison of Whole Genome Bisulfite Sequencing datasets from normal plasma datasets with reference cancer methylation datasets;

B) compare each CpG site in each bisulfite-sequenced heat-enriched CpG-rich DNA fragment from said subject sample against the reference dataset and calculate, using a bimodal distribution, a class-specific probability for each bisulfite-sequenced heat-enriched CpG-rich DNA fragment; and

C) approximate the tumor fraction in said subject sample using maximum likelihood estimation.

In some embodiments, low GC content fragments having lower than about 60% GC content are denatured and high GC content fragments having about 60% or higher GC content remain double stranded.

In some embodiments, the method of the third aspect comprises:

A) identifying differentially methylated clusters in circulating cell-free DNA between normal subjects and subjects with cancer by;

(i) obtaining one or more normal plasma whole genome bisulfite sequencing methylation dataset(s);

(ii) obtaining one or more reference cancer methylation dataset(s) and extrapolating, to ± 100 bp of each of a plurality of probe sites, methylation values with a standard deviation less than 0.4 between the various samples in the dataset;

(iii) identifying differentially methylated clusters from datasets (i) and (ii);

B) determining the class-specific probability of each site using a bimodal distribution by;

i) assigning to each sequenced fragment from the subject sample a normal and tumor class-specific proportional methylation status using the generated reference;

ii) for every site in the reference, the contribution from the unmethylated and methylated mode (0 and 1) is calculated, wherein the relative contributions of each mode in the two classes is used to assign normal or tumor class-specific probabilities for the methylation values in the assayed fragment; and

C) estimating the tumor fraction, denoted as θ, where $0 \leq \theta < 1$, of the sample;

wherein, each read is assumed to be independent, and normal and tumor class-specific probability is assigned to each read;

a global tumor fraction (θ_g) is calculated by applying a grid search to determine the highest probability event from a range of tumor fractions.

In some embodiments, in A)(iii) all CpG sites within clusters that have 0.5 difference in methylation are selected.

In some embodiments, the heat denaturing in step iii) is performed at a temperature in the range of about 87-92 ºC.

In some embodiments, the determining of at least part of the sequence comprises paired-end sequencing.

In some embodiments, the determining step provides diagnostic information for the subject.

In some embodiments, the diagnostic information comprises cancer diagnosis information for the subject.

In some embodiments, each dataset from normal plasma is from circulating cell-free DNA from a normal subject.

In some embodiments, each reference cancer methylation dataset is a colorectal adenocarcinoma (COAD) dataset.

In some embodiments, the method further comprises treating the subject based on whether or not the subject is identified as having a tumor fraction indicative of a cancer.

In some embodiments, the method is capable of >98% predictive accuracy from as low as 3 million sequencing reads.

In some embodiments, the method is capable of detecting a 0.2% tumor fraction at a probability of 0.82.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1a-d shows enrichment of CpG-dense regions using the relationship with GC content. **(a)** Percentage of differentially methylated sites in different genomic regions. **(b)** Proportion of DMRs and Illumina 450k methylation array probes in CGI with respect to the genomic distribution. **(c)** Relationship between GC content and number of CpGs in randomly generated 200 bp fragments of the human genome. **(d)** Number of DMRs of different cancers detected per 1000 fragments using different GC-content thresholds. COAD: colorectal adenocarcinoma, BRCA: breast invasive carcinoma, LUAD: lung adenocarcinoma, KIRC: kidney renal clear cell carcinoma, UCEC: uterine corpus endometrial carcinoma. Fragments above 0.6 GC content contain nearly 8-fold more DMRs across different cancers.

Figure 2a-f shows selection of CpG-rich fragments using heat denaturation. **(a)** Workflow of Heatrich-BS to select for GC-rich fragments. **(b)** Trend of GC content (upper curve)

and read enrichment at CGI (lower curve) over a range of temperatures. **(c)** Average GC content of sequenced fragments with and without heat denaturation. **(d)** Distribution of Heatrich and RRBS reads in CpG islands, shores and other regions. **(e)** Localization of Heatrich-BS and RRBS reads to CGI regions. CGIs are marked with solid bars. **(f)** Distribution of RRBS reads in different genomic regions (left). Distribution of Heatrich reads in different genomic regions (right).

Figure 3a-d shows Heatrich-BS effectively enriches for reads in CpG-rich regions. **(a)** Piling up of reads and localization to CGI using WGBS, RRBS and Heatrich-BS methods on circulating DNA. CGI regions are marked by solid bars. **(b)** Percentage of reads in CGI with and without heat denaturation. **(c)** Percentage of reads in CGI using Heatrich-BS, RRBS and WGBS on circulating DNA. **(d)** Number of reads in DMR for different total reads with and without heat denaturation.

Figure 4 shows the number of DMRs detected for different total reads with and without heat denaturation.

Figure 5a-i shows the development and validation of a tumor fraction prediction algorithm using Heatrich-BS data. **(a)** Workflow of tumor fraction prediction algorithm. **(b)** Workflow of the tumor fraction prediction algorithm. DMRs were identified by comparing healthy volunteer plasma with TCGA CRC methylation array data. Class-specific probabilities were assigned to each sequencing fragment and maximum likelihood estimation was used to infer global tumor fraction. Tumor purity correction was applied to account for normal cell infiltration in TCGA data. **(c)** ROC analysis of 0.5% tumor fraction WGBS simulated dataset at different sequencing depths. **(d)** ROC analysis of 0.5% tumor fraction Heatrich-BS simulated dataset at different sequencing depths. **(e)** Probability of detection of different tumor fractions using WGBS simulated datasets at different sequencing depths. **(f)** Probability of detection of different tumor fractions using Heatrich-BS simulated datasets at different sequencing depths. **(g)** True and algorithm predicted values of simulated plasma WGBS cfDNA samples at different sequencing depths. Confident tumor fraction prediction is achieved beyond 150 million reads (5X sequencing depth). **(h)** True and algorithm predicted value of simulated Heatrich-BS samples using different total sequencing reads. Confident tumor fraction prediction is achieved with as few as 3 million reads. **(i)** ROC analysis of 0.5% tumor fraction using simulated Heatrich-BS and WGBS cfDNA samples at 3 million total reads. AUC of Heatrich-BS at 3 million reads much higher (0.988) compared to AUC of WGBS (0.547).

Figure 6 shows the baseline tumor fraction of normal samples at different sequencing depths.

Figure 7 shows spike-in percentage and algorithm predicted value of WGBS *in silico* datasets at different sequencing depths.

Figure 8 shows spike-in percentage and algorithm predicted value of Heatrich-BS *in silico* datasets at different sequencing depths.

Figure 9 shows the number of CpGs per fragment with (left bar) and without (right bar) heat denaturation.

Figure 10 shows the tumor probability of circulating DNA fragments with (left bar) and without (right bar) heat denaturation.

Figure 11a-h shows application of Heatrich-BS on patient cfDNA samples. **(a)** Tumor fractions predicted by genomic methods and Heatrich-BS for patient cfDNA samples. High degree of concordance (Pearson r = 0.92) was achieved between the tumor percentages obtained through the two methods. **(b)** Longitudinal tumor monitoring of Patient 1. XELOX (left arrow) and FOLFIRI-Cetuximab (right arrow) treatment introduction points are marked with arrows. **(c)** Longitudinal tumor monitoring of Patient 2. 5 FU/Oxaliplatin (left arrow) and Irinotecan (right arrow) treatment introduction points are marked with arrows. **(d)** PCA analysis of Patients 1 and 2. PC1 separates patients, with patient 1 on the left of the vertical broken line, while PC2 separates along tumor burden. **(e)** Longitudinal monitoring of 14 CRC patients with Heatrich-BS tumor fraction, SLD (sum of longest diameter) and CEA measurements. CEA positive and negative is indicated by vertical black and grey lines, respectively. Heatrich-BS positive and negative is indicated by filled circles and open circles, respectively. Tumor fraction is represented by the size of the dots. Bottom panel: Cancer detected from both CEA and Heatrich-BS. Middle panel: CEA failed to detect cancer at multiple timepoints. Top panel: Earlier detection of cancer recurrence by Heatrich-BS compared to CEA. **(f)** Heatrich-BS tumor fraction, CEA and SLD values for Patient 357, which shows Heatrich-BS tumor fractions increasing before CEA values. CEA, Heatrich-BS tumor fractions and SLD values are represented by triangles, circles and squares respectively. Open symbols indicate values below threshold. **(g)** Heatrich-BS tumor fraction, CEA and SLD values for Patient 507, which shows Heatrich-BS tumor fractions being informative even when CEA values are uninformative. CEA, Heatrich-BS tumor fractions and SLD values are represented by triangles, circles and squares respectively. Open symbols indicate values below threshold. **(h)** Distribution of Heatrich-BS tumor fractions and CEA values across different tumors sizes. Heatrich-BS tumor fractions produce a more linear relationship with SLD compared to CEA values. Dashed line indicates cancer detection threshold of CEA (5.3 ng/mL).

Figure 12 shows the trend of Heatrich-BS tumor fractions, CEA levels, quantitative tumor measurements (SLD) and disease status across different timepoints of 14 CRC patients.

CEA, Heatrich-BS tumor fractions and SLD values are represented by triangles, circles and squares respectively. Open symbols indicate values below threshold.

Figure 13a-b show the comparison of quantitative SLD values with orthogonal tumor measurements. **(a)** Relationship between Heatrich-BS tumor fractions and SLD measurements at individual timepoints. Pearson r = 0.62. **(b)** Relationship between SLD measurements and CEA level at individual timepoints. Pearson r = 0.26.

Figures 14a-f shows the characterization of patient cfDNA using Heatrich-BS. **(a)** Methylation status of TCGA CRC samples and normal plasma at 635 CpG markers identified for CIMP subtype classification. **(b)** Methylation score of 233 CRC tissues with CIMP annotation in TCGA. Cut-off thresholds to distinguish different CIMP subtypes was determined by a decision tree classifier. **(c)** Raw and corrected methylation score of simulated cfDNA with different tumor fractions and methylation subtypes. Deconvolution removes the effect of tumor fraction on raw methylation score calculated from cfDNA. **(d)** Corrected methylation scores of different CIMP subtypes and tumor fraction simulated samples with decision tree classifier thresholds for CIMP classification. Corrected methylation scores are concordant across tumor fractions more than 10% for different CIMP clusters. **(e)** Corrected methylation score for patient cfDNA samples with tumor fractions above 10%. **(f)** Methylation status of TCGA CRC samples and normal plasma at DMRs identified for tumor fraction determination. These DMRs are incapable of distinguishing the CIMP subtypes.

DETAILED DESCRIPTION OF THE INVENTION

Bibliographic references mentioned in the present specification are for convenience listed at the end of the examples. The whole content of such bibliographic references is herein incorporated by reference for the material contained in them that is discussed in the sentence in which the reference is relied upon.

Definitions

Certain terms employed in the specification, examples and appended claims are collected here for convenience.

It must be noted that as used herein and in the appended claims, the singular forms "a", "an", and "the" include plural reference unless the context clearly dictates otherwise.

As used herein, the term "comprising" or "including" is to be interpreted as specifying the presence of the stated features, integers, steps or components as referred to, but does not preclude the presence or addition of one or more features, integers, steps or components, or groups thereof. However, in context with the present disclosure, the term "comprising" or

"including" also includes "consisting of". The variations of the word "comprising", such as "comprise" and "comprises", and "including", such as "include" and "includes", have correspondingly varied meanings.

The terms “Heatrich” and “Heatrich-BS” are herein used interchangeably as shorthand to describe the method of the invention, being **heat** en**rich**ment for **b**isulfite **s**equencing.

The term "subject" is herein defined as vertebrate, particularly mammal, more particularly human. For purposes of research, the subject may particularly be at least one animal model, e.g., a mouse, rat and the like. In particular, for treatment of cancer, the subject may be a human.

The term "treatment", as used in the context of the invention refers to ameliorating, therapeutic or curative treatment.

A person skilled in the art will appreciate that the present invention may be practiced without undue experimentation according to the methods given herein. The methods, techniques and chemicals are as described in the references given or from protocols in standard biotechnology and molecular biology textbooks. Standard molecular biology techniques known in the art and not specifically described were generally followed as described in Sambrook and Russel, Molecular Cloning: A Laboratory Manual, Cold Springs Harbor Laboratory, New York (2001).

EXAMPLES

Methods

Generating sheared DNA

K562 cells (ATCC® CCL-243™) were cultured in high glucose Dulbecco's modified Eagle's medium (DMEM) (Gibco) supplemented with 10% Fetal Bovine Serum (FBS) (Gibco) and 1% penicillin-streptomycin (Gibco). Genomic DNA was extracted from cultured K562 cells using the DNeasy Blood and Tissue Kit (Qiagen). The extracted gDNA was fragmented using the LE220 Focused ultrasonicator (Covaris) at the following settings: 450W peak incidence power, 30% duty factor, 200 cycles per burst for 420 sec. The fragmented DNA was size selected for 100-200 bp fragments using BluePippin 2% agarose cassette (Sage Sciences).

Patient recruitment and Extraction of cfDNA from patient blood samples

Colorectal cancer patients were recruited at the National Cancer Centre Singapore under studies 2018/2795 and 2019/2401 approved by the SingHealth Centralised Institutional Review Board. From these patients, blood specimens and tumor specimens were collected

where possible and consented for. Blood samples from healthy individuals were collected under study 2012/733/B. Retrospective review of medical records was performed to collect clinicopathological details such as patient demographics, tumor staging, serum CEA and mutational status from clinical testing where available (Table 1).

Table 1: Clinicopathological features of 14 longitudinal CRC patients

Characteristic	**Number (n=14)**
Median age at diagnosis (years)	55 (40-72)
Gender	
Male	9
Female	5
Ethnicity	
Chinese	14
Sidedness	
Left	9
Right	5
Stage at Diagnosis	
IIIB	1
IVA	10
IVB	3
MSS/MSI	
MSS	14
Mutations (Clinical test)	
KRAS p.G12D	4
KRAS p.G13D	1
NRAS p.Q61R	1
TP53 p.R342*	1

To assess the sensitivity of Heatrich-BS for tumor monitoring in comparison to CEA measurements, patients included those whose CEA measurements were informative or uninformative of disease progression. All plasma was separated from whole blood collected in EDTA tubes within 2 h of venipuncture via centrifugation at 10 min x 300 g and 10 min x 9730 g, and subsequently frozen at −80 °C. Cell-free DNA was extracted using the QiaAmp Circulating Nucleic Acids kit (Qiagen) as per manufacturer's protocol.

Heatrich-BS protocol

5-10 ng of cfDNA was used as input for the Heatrich-BS protocol. Library preparation was done using KAPA Hyper Prep Kit (Kapa Biosystems). 1.4 µl of End Repair & A-tailing buffer (Kapa Biosystems) and 0.6 µl of End Repair & A-tailing enzyme mix (Kapa Biosystems) was added to 10 µl of input DNA, and incubated at 20°C for 30 min, 65°C for 30 min. Following this, the sample was heated at 88°C for 5 min, and immediately placed on ice. The sample

was then topped up with 6 µl of Ligation buffer (Kapa Biosystems), 2 µl of DNA Ligase (Kapa Biosystems), 1 µl of nuclease-free water and 1 µl of 750 nM adapter (Kapa Biosystems). For no-heat controls, 1 µl of 1.5 µM adapter (Kapa Biosystems) was used instead. After adding these reagents, the sample was incubated at 25°C for 1 hour and then cleaned up by performing two rounds of 1.2x SPRI Select (Beckman Coulter). The sample was then subject to bisulfite conversion following the recommended protocol of Zymo EZ DNA Methylation-Gold kit (Zymo Research). The bisulfite converted DNA was amplified for 15 cycles using Pfu Polymerase (Agilent), cleaned up using 1.2x SPRI Select (Beckman Coulter) and re-amplified (10 cycles) using KAPA Hyper Hot-Start Polymerase (Kapa Biosystems) until plateau was reached. The amplified sample was cleaned up using 1.2x SPRI Select (Beckman Coulter), size selected for 190-400 bp fragments using 2% agarose Bluepippin kits (Sage Sciences), quantified using Kapa Library quantification kits (Kapa Biosystems) and sequenced using MiSeqv3 150 cycle kit or Novaseq (Illumina). Pair end sequencing of 75 bp each was performed.

Heatrich-BS Analysis Pipeline

Fastqc [Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010)] was used to check the quality of the pair-end reads generated by a MiSeq sequencer (Illumina). After adapter trimming using Cutadapt Software [Martin., M. *EMBnet.journal* 17, 10–12 (2011)], the reads were aligned to the hg38 human genome using Bismark Software [Krueger, F. & Andrews, S. R. *Bioinformatics* 27, 1571–1572 (2011)]. The aligned reads were deduplicated using Picard tools [Broad Institute. Picard Tools. (2018)], following which the Bismark methylation extractor was used to obtain per-base methylation status of each fragment.

GC content calculation

To calculate the GC content of each fragment, the forward and reverse reads were aligned separately, and then combined to generate a single coordinate range encompassing the entire fragment. The coordinates of the fragment were then used to obtain its sequence from the reference genome. For each fragment, the GC content was defined as the number of Cs and Gs, divided by the total length of the fragment. Percentage of reads in CGI was defined as the proportion of sequenced reads that coincided with hg38 CpG island annotation from UCSC.

Tumor fraction determination algorithm

The tumor fraction determination algorithm has three major steps:

Step 1: Identifying the differentially methylated clusters

To identify differentially methylated clusters for tumor-specific cfDNA detection, normal plasma whole genome methylation data [Sun, K. *et al. Proc. Natl. Acad. Sci. U. S. A.* 112, E5503–E5512 (2015)] and colorectal adenocarcinoma (COAD) methylation array from TCGA was used. 23 WGBS datasets for normal plasma and 353 450k methylation array datasets from TCGA were used for cluster generation. The TCGA methylation values were extrapolated to ± 100 bp of each probe site. To ensure selection of only consistent sites, only methylation values with a standard deviation less than 0.4 between the various sample in that class were chosen, to ensure confidence for the reference. DMRfinder Software [Gaspar, J. M. & Hart, R. P. *BMC Bioinformatics* 18, 1–8 (2017)] was used to identify differentially methylated clusters. Within these clusters, sites with a 0.5 difference in methylation were selected.

Step 2: Using a bimodal distribution to calculate the class-specific probability of each site.

Using the generated reference, a normal and tumor class-specific must be assigned to each assayed fragment. Since methylation values are binary, the average methylation value observed in the reference is a proportional combination of the unmethylated and methylated reads. As such, a bimodal distribution can accurately represent the proportional methylation status of the reference. For every site in the reference, the contribution from the unmethylated and methylated mode (0 and 1) was calculated. The relative contributions of each mode in the two classes were used to assign class-specific probabilities for the methylation values in the assayed fragment. In this way, a bimodal reference was used to assign normal or tumor probability values to each site assayed.

Step 3: Using maximum likelihood estimation to predict the tumor fraction of the sample.

After assigning class-specific probabilities to each fragment, the fraction of fragments that come from the tumor must be enumerated. The tumor-derived cfDNA in a sample, also known as tumor fraction, can be denoted as θ, where $0 \leq \theta < 1$. To estimate the tumor fraction θ, a maximum likelihood estimation approach and grid search, adapted from CancerDetector [Li, W. *et al. Nucleic Acids Res.* 46, e89 (2018)], was used to calculate the global tumor fraction (θ_g) for each sample. The determined tumor purity correction factor (γ) of 0.057 is then applied to the raw tumor fraction to generate the final tumor fraction.

GC content analysis in DMRs

The human genome was split into 200 bp tiling windows and the GC content of each window was calculated. Windows with GC content exceeding 60% were used as a theoretical representation of Heatrich output. To investigate the relationship between GC content and

CpG density, we used a random sequence generator to create half a million 200 bp fragments. The number of G+C bases (GC content) and number of CG dinucleotides (CpG content) was calculated as a fraction of the total length. To calculate the number of DMRs per 1000 fragments, we first generated DMRs for each cancer using the earlier mentioned approach (Step 1 of tumor fraction determination). We then utilized random 1000 fragments generated from a plasma cfDNA dataset subject to different GC content thresholds and counted the number of DMRs of each cancer that was detected. To generate the number of DMRs covered by different numbers of total sequencing reads, a similar approach was utilized where different datasets were sub-sampled to the required number of reads and the number of fragments that contained DMRs were counted. For cfRRBS, all fragments with MspI cut-sites between 20 and 160 bp were used as theoretical data.

Tumor burden estimation by whole genome or targeted sequencing

DNA libraries were prepared using the Kapa Hyper Prep Kit (Kapa Biosystems) and sent for whole genome sequencing or targeted sequencing. Hybridization capture was done for targeted sequencing using an IDT Xgen Custom Panel of 101 cancer genes and reagents as per manufacturer's instructions. Sequencing was performed on an Illumina Hiseq4000 (2 × 150 bp paired end reads). Tumor fraction estimation from whole genome sequencing data was carried out using the ichorCNA algorithm [Adalsteinsson V. A. et al. *Nat Commun* 8 (2017)]. Variant calling from targeted sequencing data was performed using MuTect Software [Sougnez, C et al. *Nat Biotechnol* 31:213-219 (2013)] with the tumor fraction estimation being the mean Variant Allele Frequency of 7 known colorectal cancer hotspots (KRAS, NRAS, BRAF, EGFR, APC, TP53, PIK3CA) present in a particular sample.

Tumor measurements and disease status classification

For each profiled timepoint, the nearest available CT scan image was retrieved from the patient's clinical records. Each lesion on the scan was measured in 2 dimensions (maximum width and maximum length). Indeterminate lesions were not measured. For each timepoint, the Sum of the Longest Diameter (SLD) was determined, providing a representation of the total tumor load present at the timepoint. To ensure consistency, all measurements were carried out by the same clinician. Disease classification for each timepoint was carried out according to the following criteria; Complete Response: disappearance of all lesions; Partial Response: ≥ 30% decrease in the SLD of the lesions compared with the SLD of the previous measured timepoint; Progressive Disease: ≥ 20% increase of at least 5 mm in the SLD of the lesion compared with the SLD of the previous measured timepoint OR the appearance of new lesions >10 mm in diameter; Stable Disease: Neither PR, PD nor CR.

Subtype classification

In order to identify marker sets for CIMP subtype classification, we utilized the data and annotations from The Cancer Genome Atlas Network's publication [Muzny, D. M. et al. *Nature* 487:330-337 (2012)]. We selected CpG sites with a minimum standard deviation of 0.25 and showed distinct methylation patterns in the different CIMP subtypes. Since Heatrich-BS disproportionately enriches for CpG-rich regions, we further restricted the marker list to CpGs that were covered by at least 50 Heatrich-BS samples. To determine the threshold values to distinguish CIMP clusters, we utilized the Decision Tree classifier software from KNIME [Berthold, M. R. et al. *Studies in Classification, Data Analysis, and Knowledge Organization, GfKL* (2007)]. The TCGA dataset with CIMP classification was split 70-30 for training and testing.

To perform CIMP classification of cfDNA samples, a raw methylation score across the marker sites was calculated for each sample. The Methylation Score, defined as the average of methylation values of the 635 loci, was used to summarize the degree of methylation in CIMP loci. To estimate the Methylation Score of the underlying tumor in cfDNA, we note that $M_{cfDNA} = \theta M_{tumor} + (1-\theta)M_{normal-plasma}$. Substituting $M_{normal-plasma}$ (Methylation Scores from cfDNA of healthy plasma) and θ (the tumor fraction calculated from Heatrich-BS), the Methylation Score of the underlying tumor (M_{tumor}) can be estimated. To calculate $M_{normal-plasma}$, samples with negative methylation scores or fewer than 100 reads falling into marker sites were excluded.

Data sources

In order to compare our assay with existing methods, we obtained the following data from NCBI GEO: RRBS (SRR222486), cell-free RRBS (GSM2090507). cfDNA WGBS data for algorithm development and validation was obtained by request from the EGA database (EGAS00001001219). Colorectal cancer tumor WGBS data was obtained from NCBI GEO (SRR1035745). Illumina 450k methylation array data for different cancers was obtained from The Cancer Genome Atlas (TCGA) data repository.

EXAMPLE 1:

Enrichment of CpG-rich DMRs using GC content

The sequence content in the human genome is highly non-uniform. Long stretches of CpG poor regions are punctuated by short stretches of CpG-dense region that coincide with important gene regulatory elements such as promoters. These CpG-dense regions are often differentially methylated between tissues and disease such as cancer [Guo, S. *et al. Nat. Genet.* 49, 635–642 (2017)]. We use DMRfinder Software [Gaspar, J. M. & Hart, R. P. *BMC Bioinformatics* 18, 1–8 (2017)], a popular software, to identify DMRs between colorectal

cancer (CRC) tissue and healthy plasma and found that nearly 45% of the DMRs lie within CGIs (Fig. 1a), which originates from only 1% of the genome (Fig. 1b). As such, it is of value to focus on this small fraction of CpG-rich genome for epigenetic profiling. The traditional approach to enrich for CGIs is RRBS [Gu, H. *et al. Nat. Protoc.* 6, 468–481 (2011)], with a single-tube variant termed single-cell RRBS (scRRBS) [Guo, H. *et al. Nat. Protoc.* 10, 645–59 (2015)] for low input samples. However, this approach is still limited in utility for enrichment of fragmented DNA, as very few fragments can meet the requirement for a CCGG cut site on both ends. For example, one study that uses scRRBS to assay for methylation in cfDNA, henceforth referred to as cell-free RRBS, has only 6.4% of the reads in CGIs [Guo, S. *et al. Nat. Genet.* 49, 635–642 (2017)]. While there is no known means to physically enrich for CpG-dense DNA, it is long known that the G+C content of a double-stranded DNA fragment is closely related to its thermal stability. It has been shown that the GC bond in DNA has a binding energy of 25.4 kcal mol^{-1}, which is two times stronger than the AT bond, with a binding energy of 12.4 kcal mol^{-1}. As the presence of a CpG dinucleotide in a fragment adds 2 GC bonds to the duplex, whether effective selection of CpG-dense fragments can be achieved by selection of GC-rich fragments was tested. To check this hypothesis, the GC content and number of CpGs in each fragment in a sample of 0.5 million 200 bp fragments of the human genome were calculated (Fig. 1c). The excellent correlation between GC content and number of CpGs suggests that GC content can be used to enrich for CpG-rich fragments. Fragments with GC content greater than 0.6 constitutes only 2.5% of the genome, but disproportionately includes 85% of CGIs and 58% of our identified DMRs. We further verified the relationship between different cancer-specific DMRs and GC content (Fig. 1d). Selection of fragments above 0.6 GC content affords nearly 8-fold enrichment in proportion of reads in DMRs across different cancers (COAD: colorectal adenocarcinoma, BRCA: breast invasive carcinoma, LUAD: lung adenocarcinoma, KIRC: kidney renal clear cell carcinoma, UCEC: uterine corpus endometrial carcinoma). Therefore, we established that DMRs of various cancers, which are universally overrepresented in CpG-dense regions, can be effectively enriched with selection of high-GC DNA fragments.

Heatrich uses thermal denaturation to select for DNA fragments with high GC content (Fig. 2a). Fragmented DNA was first end-repaired and A-tailed. Following this, the sample was heated to denature the GC-poor fragments and adapter ligation was immediately performed. The process of adapter ligation allows selection of intact non-denatured GC rich double-stranded fragments, as T4 DNA ligase has a high selectivity for dsDNA [Doherty, A. J. & Wigley, D. B. *J. Mol. Biol.* 285, 63–71 (1999)]. The selected fragments were bisulfite-converted and subsequently sequenced. We found through empirical optimization experiments using sheared genomic DNA (Fig. 2b) that heating DNA to 88°C immediately prior to adapter ligation

yields the best enrichment of reads in CGI (28%) at high GC content (0.63±0.006) compared to the average GC content of the unheated samples (0.42±0.009) (Fig. 2c).

EXAMPLE 2

Using heat denaturation to select for GC-rich fragments

Having established that GC content correlates well with CpG density, we then explored the use of heating as a means of GC content selection. The workflow of the Heatrich-BS assay is shown in Figure 2a. Fragmented DNA was prepared for adapter ligation by performing end repair and A-tailing. Following this, the sample was heated to denature the GC-poor fragments and immediately underwent adapter ligation. The process of adapter ligation allows selection of intact double-stranded fragments, as T4 DNA ligase results has a high selectivity for dsDNA due to its high affinity for dsDNA and low affinity for ssDNA [Doherty, A. J. & Wigley, D. B. *J. Mol. Biol.* 285, 63–71 (1999)]. In this way, adapter ligation can be used to select for the non-denatured GC-rich DNA fragments. The selected fragments were bisulfite converted and subsequently sequenced.

In order to identify the temperature needed to achieve the desired enrichment, a range of temperatures from 75°C to 95°C on sheared genomic DNA were tested (Fig. 2b). The temperature range of 87°C to 90°C had the highest GC content and enrichment of reads in CGI. At higher temperatures, even the fragments with high GC content were denatured, therefore reducing the enrichment in CGIs. Further factoring in the alignment rate after bisulfite conversion, we chose 88°C as the denaturation temperature, as this condition had the highest alignment rate along with a sufficiently high GC content (~0.62) and enrichment of reads in CGIs (28%). From multiple experiments performed at the selected temperature, we can see that the GC content of the Heatrich samples was much higher than the average GC content of the genome and unheated samples (Fig. 2c). The consistent GC content and CpG enrichment obtained from multiple experiments also shows the high degree of reproducibility of our assay.

Analysis of the mapping of heat-treated sheared DNA showed that there is indeed a significant enrichment of reads that localized to CpG islands and shores, comparing favorably to the theoretical genomic distribution and even RRBS, the gold standard technique for CGI enrichment (Fig. 2d). The accumulation of reads around CGIs was also visualized (Fig. 2e), showing that Heatrich has significant piling up of reads around CGI. More detailed analysis of the genomic distribution of Heatrich reads showed a remarkable similarity to that of standard RRBS (Fig. 2f), suggesting that this non-enzymatic approach can be a viable alternative to RRBS for detailed methylation profiling in important genomic regulatory elements. Notably, Heatrich is independent of restriction enzyme sequence, and will thus robustly profile the same

regions even in the presence of restriction site polymorphisms, and could find applications when DNA is already fragmented prior to restriction digestion (e.g. FFPE, degraded DNA, cfDNA) where RRBS is not well-suited.

EXAMPLE 3

Enrichment of CpG-rich regions using Heatrich-BS

cfDNA samples are ideally suited for Heatrich-BS due to their fragmented nature, and promising applications for noninvasive disease detection. To test the performance of Heatrich-BS on cfDNA, samples were obtained from cancer patients. Bisulfite treatment was performed on samples with and without heat denaturation. We first visualized the reads obtained from Heatrich-BS and compared it with previously reported cell-free RRBS and WGBS datasets [Guo, S. *et al. Nat. Genet.* 49, 635–642 (2017)] with similar read counts (Fig. 3a). The reads from cell-free RRBS and WGBS were distributed almost evenly across the genomic region. On the other hand, the vast majority of Heatrich-BS reads piled up at CpG islands and shores, with little reads in the remaining regions. This shows the specificity of Heatrich-BS in selecting only fragments of high GC content. Furthermore, with similar read counts, the average height of Heatrich-BS peaks is appreciably more than that of the other datasets. This shows that Heatrich-BS can obtain higher depth in informative regions using the same number of total reads.

The distribution of sequenced reads from Heatrich-BS and unheated cfDNA were then quantified, and it was seen that the Heatrich-BS samples displayed up to 30-fold enrichment of reads in CGIs (Fig. 3b). For 500k unique sequencing reads, almost 30% of Heatrich-BS reads fall within CGIs, while only 6% and 2% of cell-free RRBS and WGBS reads are in CGIs (Fig. 3c). This demonstrates that Heatrich-BS displays more effective CGI enrichment for fragmented DNA, compared to the performance of cell-free RRBS on similar samples. Furthermore, the number of reads in DMRs were measured for different sequencing reads (Fig. 3d). It is evident that Heatrich-BS has up to 10-fold more reads localizing to DMRs and Heatrich-BS is able to detect up to 10-fold more DMRs compared to the no heat controls (Fig. 4). This would provide higher sensitivity in detecting fragments of tumor origin for the same number of sequencing reads. These figures highlight the strength of Heatrich-BS, which is the effective enrichment of reads in CGI and DMRs, even with fewer sequencing reads.

EXAMPLE 4

Estimating the tumor fraction

In order to determine tumor fraction from methylation data, recent studies have utilized a fragment-wise approach as it offers high sensitivity [Guo, S. *et al. Nat. Genet.* 49, 635–642

(2017); Li, W. *et al. Nucleic Acids Res.* 46, e89 (2018)]. As such, we adopted a fragment-based approach for our tumor fraction determination algorithm. The workflow of our algorithm is shown in Figure 5a and 5b, and detailed in the methods section.

Using 450k array methylation data from TCGA to generate the colorectal cancer reference has a key pitfall: the resected tumor sample used for the array is often contaminated with normal cells. This contamination has been widely acknowledged [Leary, R. J. *et al. Proc. Natl. Acad. Sci. U. S. A.* 105, 16224–9 (2008)] and tools have been developed to account for this contamination in somatic variant calling [Sendorek, D. H. *et al. BMC Bioinformatics* 19, 28 (2018)]. In methylation-based analysis, this contamination could lead to overestimation of the tumor fraction. Therefore, we attempted to identify the contribution of non-neoplastic cells to the colorectal cancer reference, and eliminate its effect on our tumor fraction determination. To this end, we generated a colorectal cancer reference using 23 healthy volunteer cfDNA datasets [Sun, K. *et al. Proc. Natl. Acad. Sci. U. S. A.* 112, E5503–E5512 (2015)] for the normal and 353k TCGA array datasets for the tumor reference. This reference was then applied to 3 other healthy cfDNA datasets to estimate the tumor fraction. We observed a stable non-zero baseline value (Fig. 6), irrespective of the sequencing depth. Similar observations of non-zero baseline of healthy cfDNA has been reported when tissue samples rather than pure cell populations were used as references. This baseline value is attributed to the contribution of non-neoplastic cells from tumor tissue references and commonly leads to overestimation of the tumor fraction in methylation-based analysis. To identify the contribution of these non-neoplastic cells to the colorectal cancer reference, and eliminate its effect on our tumor fraction determination, we determined a tumor purity correction factor by performing Receiver Operating Characteristic (ROC) analysis of healthy and simulated plasma cfDNA WGBS samples at 0.5% tumor fraction. The specific correction factor (γ) that maximized sensitivity and specificity across multiple sequencing depths (Table 2) was determined for the generated CRC reference. This factor is dependent on the reference used and only needs to be determined once for each reference.

CLAIMS

1. A method of enrichment of CpG islands comprising cancer-specific methylation information in isolated circulating cell-free DNA from a subject, comprising the steps:

i) provide a cell-free DNA sample;

ii) repair double-stranded DNA ends and add dA tail;

iii) heat-denature the cell-free DNA, wherein denaturation to single strands increases and remaining double stranded decreases with a decrease in GC content and denaturation to single strands decreases and remaining double stranded increases with an increase in GC content;

iv) ligate methylated adapters to both ends of the double-stranded DNA;

v) perform bisulfite conversion of the adapter-ligated DNA;

vi) amplify the bisulfite-converted adapter-ligated DNA from v);

vii) size-select for 190-400 bp fragments of the amplified DNA of vi).

2. The method of claim 1, wherein low GC content fragments having lower than about 60% GC content are denatured and high GC content fragments having about 60% or higher GC content remain double stranded.

3. The method of claim 1 or 2, wherein the heat denaturing in step iii) is performed at a temperature in the range of about 87-92 ºC.

4. The method of any one of claims 1 to 3, wherein the cell-free DNA sample is from a subject that has a medical condition.

5. The method of any one of claims 1 to 4, further comprising determining at least part of the sequence of one or more of the amplified molecules.

6. The method of claim 5, wherein the determining of at least part of the sequence comprises paired-end sequencing.

7. The method of claim 6, wherein the determining step provides diagnostic information for the subject.

8. The method of claim 7, wherein the diagnostic information comprises cancer diagnosis information for the subject.

9. A method to determine the tumor fraction in a circulating cell-free DNA sample from a subject, the method comprising:

A) identify differentially methylated clusters from a comparison of Whole Genome Bisulfite Sequencing datasets from normal plasma with reference cancer methylation datasets;

B) compare each CpG site in each bisulfite-sequenced heat-enriched CpG-rich DNA fragment from said subject sample against the reference datasets and calculate, using a bimodal distribution, a class-specific probability for each bisulfite-sequenced heat-enriched CpG-rich DNA fragment; and

C) approximate the tumor fraction in said subject sample using maximum likelihood estimation.

10. The method of claim 9, comprising:

A) identifying differentially methylated clusters in circulating cell-free DNA between normal subjects and subjects with cancer by;

(i) obtaining a normal plasma whole genome bisulfite sequencing methylation dataset;

(ii) obtaining a reference cancer methylation dataset and extrapolating, to ± 100 bp of each of a plurality of probe sites, methylation values with a standard deviation less than 0.4 between the various samples in the dataset;

(iii) identifying differentially methylated clusters from datasets (i) and (ii);

B) determining the class-specific probability of each site using a bimodal distribution by;

i) assigning to each sequenced fragment from the subject sample a normal and tumor class-specific proportional methylation status using the generated reference;

ii) for every site in the reference, the contribution from the unmethylated and methylated mode (0 and 1) is calculated, wherein the relative contributions of each mode in the two classes is used to assign normal or tumor class-specific probabilities for the methylation values in the assayed fragment; and

C) estimating the tumor fraction, denoted as θ, where $0 \leq \theta < 1$, of the sample;

wherein, each read is assumed to be independent, and normal and tumor class-specific probability is assigned to each read;

a global tumor fraction (θ_g) is calculated by applying a grid search to determine the highest probability event from a range of tumor fractions.

11. The method of claim 10, wherein in A)(iii) all CpG sites within clusters that have 0.5 difference in methylation are selected.

12. The method of any one of claims 9 to 11, wherein each dataset from normal plasma is from circulating cell-free DNA from a normal subject.

13. The method of any one of claims 9 to 12, wherein each reference cancer methylation dataset is a colorectal adenocarcinoma (COAD) dataset.

14. A method of obtaining information in relation to a medical condition of a subject, the method comprising:

i) provide an isolated cell-free DNA sample from the subject;

ii) repair double-stranded DNA ends and add dA tail;

iii) heat-denature the cell-free DNA,wherein denaturation to single strands increases and remaining double stranded decreases with a decrease in GC content and denaturation to single strands decreases and remaining double stranded increases with an increase in GC content;

iv) ligate methylated adapters to both ends of the double-stranded DNA;

v) perform bisulfite conversion of the adapter-ligated DNA;

vi) amplify the bisulfite-converted adapter-ligated DNA from v);

vii) Size-select for 190-400 bp fragments of the amplified DNA of vi);

viii) determine at least part of the sequence of one or more of the amplified molecules;

determine the tumor fraction in the sample from the subject, comprising:

A) identify differentially methylated clusters from a comparison of Whole Genome Bisulfite Sequencing datasets from normal plasma datasets with reference cancer methylation datasets;

B) compare each CpG site in each bisulfite-sequenced heat-enriched CpG-rich DNA fragment from said subject sample against the reference dataset and calculate, using a bimodal distribution, a class-specific probability for each bisulfite-sequenced heat-enriched CpG-rich DNA fragment; and

C) approximate the tumor fraction in said subject sample using maximum likelihood estimation.

15. The method of claim 14, wherein low GC content fragments having lower than about 60% GC content are denatured and high GC content fragments having about 60% or higher GC content remain double stranded.

16. The method of claim 14 or 15, comprising:

A) identifying differentially methylated clusters in circulating cell-free DNA between normal subjects and subjects with cancer by;

(i) obtaining one or more normal plasma whole genome bisulfite sequencing methylation dataset(s);

(ii) obtaining one or more reference cancer methylation dataset(s) and extrapolating, to ± 100 bp of each of a plurality of probe sites, methylation values with a standard deviation less than 0.4 between the various samples in the dataset;

(iii) identifying differentially methylated clusters from datasets (i) and (ii);

B) determining the class-specific probability of each site using a bimodal distribution by;

i) assigning to each sequenced fragment from the subject sample a normal and tumor class-specific proportional methylation status using the generated reference;

ii) for every site in the reference, the contribution from the unmethylated and methylated mode (0 and 1) is calculated, wherein the relative contributions of each mode in the two classes is used to assign normal or tumor class-specific probabilities for the methylation values in the assayed fragment; and

C) estimating the tumor fraction, denoted as θ, where $0 \leq \theta < 1$, of the sample;

wherein, each read is assumed to be independent, and normal and tumor class-specific probability is assigned to each read;

a global tumor fraction (θ_g) is calculated by applying a grid search to determine the highest probability event from a range of tumor fractions.

17. The method of claim 16, wherein in A)(iii) all CpG sites within clusters that have 0.5 difference in methylation are selected.

18. The method of any one of claims 14 to 17, wherein the heat denaturing in step iii) is performed at a temperature in the range of about 87-92 ºC.

19. The method of any one of claims 14 to 18, wherein the determining of at least part of the sequence comprises paired-end sequencing.

20. The method of any one of claims 14 to 19, wherein the determining step provides diagnostic information for the subject.

21. The method of claim 20, wherein the diagnostic information comprises cancer diagnosis information for the subject.

22. The method of any one of claims 14 to 21, wherein each dataset from normal plasma is from circulating cell-free DNA a from normal subject.

23. The method of any one of claims 14 to 22, wherein each reference cancer methylation dataset is a colorectal adenocarcinoma (COAD) dataset.

24. The method of any one of claims 9-23, capable of >98% predictive accuracy from as low as 3 million sequencing reads.

25. The method of any one of claims 9-23, capable of detecting a 0.2% tumor fraction at a probability of 0.82.