

METHOD FOR DETECTION AND QUANTIFICATION OF METHYLATED DNA**FIELD OF THE INVENTION**

[0001] The present invention generally relates to the detection and quantification of nucleic acid. In particular, the present invention relates to the detection and quantification of methylated DNA.

BACKGROUND

[0002] DNA methylation is the covalent transfer of a methyl group to the 5-carbon position of the DNA base cytosine. In vertebrates, DNA methylation occurs at cytosines within a CpG site, i.e. a cytosine that immediately precedes a guanine base. This epigenetic modification is regulated by DNA methyltransferases and is widely known as a repressive mark that plays a key role in transcriptional silencing.

[0003] In the context of cancer, DNA methylation in promoter regions leads to a decrease in gene expression and is a common mechanism of silencing of tumor suppressor genes. DNA methylation can also result in the induction of mutations and decreased genomic stability. Spontaneous deamination of cytosine forms thymine, thus generating a point mutation. Cancer cells have distinct and aberrant patterns of DNA methylation compared to normal cells, and often display large regions of global hypo-methylation across the genome and localized areas of hyper-methylation, which are usually located at islands or clusters of CpG sites in gene promoter regions. Differential patterns of methylation in cancer cells can be used to detect the presence of cancer, such as for cancer screening purposes or for monitoring disease progression and treatment response.

[0004] Conventional methods for cancer screening, early cancer detection and disease monitoring have various drawbacks. For example, existing cancer screening methods, such as blood tumor marker tests or CT scans, are often limited by their sensitivity or specificity. These methods subject the patient to unnecessary follow-up that can be invasive, expensive and stressful. Further, conventional cancer screening methods such as colonoscopy and pap smear are often time-consuming, invasive and only detect one type of cancer per test. In addition, late cancer diagnosis when the cancer has already metastasized leaves the patient ineligible for curative surgery and limits the patient's effective therapeutic window and treatment options. Moreover, for cancer patients, disease monitoring by repeat tissue biopsy

is infeasible and repeat imaging scans are usually only recommended every 3 months to minimize radiation exposure. Also, disease monitoring through the detection of mutations in tissue or liquid biopsies is limited in sensitivity because mutations can occur anywhere along the length of a gene, rendering the comprehensive identification of mutations technically
5 challenging. Finally, cancer mutations are often not specific to a particular cancer type, making it difficult to identify the tissue of origin of the tumor.

[0005] In addition, conventional methods for multiplex detection of methylated DNA, for example, in plasma cell-free DNA (cfDNA), also face various challenges. Conventional treatment with sodium bisulfite to convert un-methylated cytosines to uracils is harsh and
10 often leads to DNA fragmentation and poor yield. This sodium bisulfite conversion method requires high starting amount of DNA, which can be challenging especially in the case of plasma cfDNA from individuals with no or low tumor load. Further, sequencing errors limit the sensitivity of detection as signal is indistinguishable from technical noise. In addition, target capture of CpG-rich hyper-methylated regions in cancer often requires two sets of
15 primers for the separate identification of un-methylated and methylated DNA in methyl-specific PCR reactions. These PCR reactions are also limited in the number of CpG sites that can be assessed in a single reaction, which is typically about one to three per primer pair. Moreover, the conditions for primer design in these PCR reactions are rather stringent, as the primer should contain the target CpG site(s), as well as at least three to five thymines
20 converted from unmethylated non-CpG cytosines, in order to ensure that only properly converted DNA will be amplified. These requirements of methyl-specific PCR reactions exclude the selection of targetable regions that do not fulfil the selection criteria.

[0006] Thus, there is a need for a method to address the disadvantages of the conventional methods as described above. The present disclosure describes a methodology
25 for the identification and quantification of methylated DNA for cancer screening and detection of early-stage (stage I/II) cancer which is often undetectable by conventional screening methods, minimal residual disease following cancer surgery or therapy, and cancer relapse. The method of the present disclosure seeks to achieve high sensitivity and specificity for the detection of methylated DNA, high efficiency of DNA conversion with
30 minimum fragmentation and loss in DNA yield, suppression of low-level errors due to sequencing, and minimal invasiveness.

SUMMARY

[0007] In a first aspect, the present disclosure refers to a method of detecting methylated DNA pattern in DNA in a biological sample, comprising:

- (a) converting un-methylated cytosine of the DNA to uracil by deamination to thereby
5 generate converted DNA;
- (b) purifying the converted DNA from step (a);
- (c) tagging a barcode sequence on the converted DNA, by performing a first PCR amplification using a plurality of forward and reverse primer pairs specific to the converted sequence of target regions, wherein each forward primer of the plurality of forward and
10 reverse primer pairs comprises a barcode sequence on its 5' end, wherein the barcode sequence of each forward primer is different from each other, wherein the target regions comprise one or more CpG sites;
- (d) subjecting the tagged converted DNA from step (c) to a second PCR amplification with universal indexed primers to thereby create a sequencing library with components required
15 for multiplex sequencing;
- (e) subjecting the sequencing library to multiplex sequencing on a next-generation sequencing platform;
- (f) detecting the presence of a barcode sequence using Bioinformatics methods to count and assign each DNA sequence from the next-generation sequencing to an original parental DNA
20 molecule carrying the same barcode sequence, comprising:
 - (i) performing cluster reassignment of sequencing reads with the same barcode sequence to thereby generate barcode clusters wherein each barcode cluster contains reads from the same amplicon and with the same barcode sequence; and
 - (ii) performing consensus calling for each barcode cluster to thereby obtain
25 consensus reads;
- (g) reconstructing the methylated DNA pattern of the DNA by
 - (I) comparing the DNA sequence to a reference genome using a sequence alignment tool; and
 - (II) conducting variant analysis of the DNA sequence by comparing the consensus
30 reads to the reference genome to detect the variations;to thereby assess 1) the conversion efficiency of non-CpG cytosines to thymine as quality control, and 2) the frequency of methylation at each CpG cytosine across all

consensus sequencing reads with distinct barcode sequences (i.e. reads that are derived from unique DNA molecules corresponding to a specific amplicon).

[0008] In a second aspect, the present disclosure refers to a kit for detecting methylated DNA pattern in DNA in a biological sample according to the method of the first aspect, comprising:

- (a) a first enzyme capable of oxidizing 5-methylcytosine and 5-hydroxymethylcytosine of the DNA;
- (b) a second enzyme capable of converting un-methylated cytosine of the DNA to uracil by deamination;
- 10 (c) a plurality of forward and reverse primer pairs specific to the converted sequence of target regions, wherein each forward primer of the plurality of forward and reverse primer pairs comprises a barcode sequence on its 5' end, wherein the barcode sequence of each forward primer is different from each other, wherein the target regions comprise one or more CpG sites;
- 15 (d) a plurality of universal indexed primers for creating the sequencing library;
- (e) a first DNA polymerase capable of amplifying DNA with uracil bases, for amplification of converted DNA;
- (f) a reagent capable of removing excess primers;
- (g) a second DNA polymerase capable of amplifying DNA, for creating the sequencing
- 20 library; and
- (h) sodium bisulfite.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] The invention will be better understood with reference to the detailed description when considered in conjunction with the non-limiting examples and the accompanying drawings, in which:

[0010] **Fig. 1** illustrates the overall experimental workflow, from the conversion of DNA to sequencing.

[0011] **Fig. 2** illustrates example of primer design for capturing converted DNA.

30 Top: For CLIP4_methyl_2F and CLIP4_methyl_2R, the italicised sequences represent the adaptor sequences required for the second amplification with the universal indexed Illumina P5 and P7 primers, respectively. The underlined sequence represents the target-specific

sequence. Y and R represent a degenerate base (C or T and A or G, respectively) following the IUB code. For CLIP4_methyl_2F, NNNNNNNNNN represents a random barcode sequence.

Bottom: For the indexed Illumina P5 and P7 primers, the underlined bases indicate an 8 bp index barcode. For multiplex sequencing, each sample will be assigned a unique combination of forward and reverse indexes.

[0012] **Fig. 3** illustrates expected sequencing library profile on TapeStation.

[0013] **Fig. 4**, comprising **Figs. 4(a)** and **4(b)**, illustrates examples of sequence alignment to Human hg19 genome for a single sample visualized using Integrated Genome Viewer (IGV), wherein **Fig. 4(a)** shows amplicon designed to the plus strand of the genome, and **Fig. 4(b)** shows amplicon designed to the minus strand of the genome.

[0014] **Fig. 5** illustrates the Conversion efficiency of non-CpG cytosines to thymines. Samples with conversion <0.97 will be repeated.

[0015] **Fig. 6** illustrates the examples of correlation of CpG methylation within each amplicon.

Top: Amplicon that contains highly correlated CpG methylation (Pearson Correlation Coefficient>0.9 at each site).

Bottom: Amplicon with low correlation of CpG methylation. The axes indicate chromosomal position.

[0016] **Fig. 7** illustrates examples of median methylation beta-values across normal (n=57) and cancer (n=152) samples. For amplicons with low CpG correlations (<0.8 correlation value), individual CpG position data is considered.

[0017] **Fig. 8** shows examples of average amplicon methylation values across normal, breast, colorectal, lung and ovarian cancer samples.

[0018] **Fig. 9** shows sample distribution used for training set and best 3-fold cross validation scores.

[0019] **Fig. 10** illustrates the N-gram method of detecting cfDNA methylation patterns.

[0020] **Fig. 11** illustrates the Skip-gram method of detecting cfDNA methylation patterns. Examples of 1-Skip and 2-Skip analyses are shown.

[0021] **Fig. 12** shows the sensitivity performance of different prediction models, set at 95% specificity threshold of the training set.

DETAILED DESCRIPTION

[0022] The present disclosure describes a methodology for detecting methylated DNA pattern in DNA with high sensitivity and specificity, for the purpose of cancer screening and detection of early-stage (stage I/II) cancer, minimal residual disease following cancer surgery or therapy, and cancer relapse.

[0023] In a first aspect, the present disclosure refers to a method of detecting methylated DNA pattern in DNA in a biological sample, comprising:

(a) converting un-methylated cytosine of the DNA to uracil by deamination to thereby generate converted DNA;

10 (b) purifying the converted DNA from step (a);

(c) tagging a barcode sequence on the converted DNA, by performing a first PCR amplification using a plurality of forward and reverse primer pairs specific to the converted sequence of target regions, wherein each forward primer of the plurality of forward and reverse primer pairs comprises a barcode sequence on its 5' end, wherein the barcode sequence of each forward primer is different from each other, wherein the target regions
15 comprise one or more CpG sites;

(d) subjecting the tagged converted DNA from step (c) to a second PCR amplification with universal indexed primers to thereby create a sequencing library with components required for multiplex sequencing;

20 (e) subjecting the sequencing library to multiplex sequencing on a next-generation sequencing platform;

(f) detecting the presence of a barcode sequence using Bioinformatics methods to count and assign each DNA sequence from the next-generation sequencing to an original parental DNA molecule carrying the same barcode sequence, comprising:

25 (i) performing cluster reassignment of sequencing reads with the same barcode sequence to thereby generate barcode clusters wherein each barcode cluster contains reads from the same amplicon and with the same barcode sequence; and

(ii) performing consensus calling for each barcode cluster to thereby obtain consensus reads;

30 (g) reconstructing the methylated DNA pattern of the DNA by

(I) comparing the DNA sequence to a reference genome using a sequence alignment tool; and

(II) conducting variant analysis of the DNA sequence by comparing the consensus reads to the reference genome to detect the variations;

to thereby assess 1) the conversion efficiency of non-CpG cytosines to thymine as quality control, and 2) the frequency of methylation at each CpG cytosine across all
5 consensus sequencing reads with distinct barcode sequences (i.e. reads that are derived from unique DNA molecules corresponding to a specific amplicon).

[0024] Firstly, the un-methylated cytosine of the DNA is converted to uracil by deamination to thereby generate converted DNA, as disclosed in step (a) of the method of the first aspect.

10 [0025] In one example, the DNA is extracted from the biological sample before step (a). The DNA may be extracted using any method or kit known in the art. In one example, the DNA is extracted from the biological sample before step (a) using organic extraction methods, such as phenol/chloroform extraction. In another example, the DNA is extracted from the biological sample before step (a) using kits such as, but not limited to, QIAamp
15 Circulating Nucleic Acid Kit (Qiagen), MagMAX Cell-Free DNA Isolation Kit (Applied Biosystems), Cell/Blood DNA Kit (CatchGene), Tissue DNA Kit (CatchGene) and DNeasy Blood and Tissue Kits (Qiagen).

[0026] In one example, the extracted DNA is converted by the method disclosed herein, comprising:

- 20
- treating the DNA using a first enzyme that oxidizes 5-methylcytosine and 5-hydroxymethylcytosine of the DNA to thereby protect the 5-methylcytosine and 5-hydroxymethylcytosine from deamination;
 - purifying the DNA;
 - converting the un-methylated cytosine of the DNA to uracil by deamination using a
25 second enzyme to thereby generate converted DNA.

[0027] In one example, the first enzyme is a Ten-eleven translocation (TET) enzyme or an isoform thereof. In another example, the TET enzyme is selected from the group consisting of TET1 enzyme or an isoform thereof, TET2 enzyme or an isoform thereof, and TET3 enzyme or an isoform thereof.

30 [0028] In one example, the purification of DNA is performed using an agent such as paramagnetic beads. In one example, the paramagnetic beads are selected from the group consisting of AMPure XP beads, SPRI beads, and Dynabeads.

[0029] In one example, the second enzyme is a cytidine deaminase or an enzyme with cytidine deaminase properties. In another example, the cytidine deaminase is selected from the group consisting of APOBEC enzyme, CDA, and activation-induced cytidine deaminase. In another example, the enzyme with cytidine deaminase properties is selected from the group consisting of M. SssI and M.HpaII.

[0030] In another example, the extracted DNA is converted using sodium bisulfite.

[0031] In another example, the DNA is not extracted from the biological sample before step (a), and is converted using direct conversion methods in which no DNA extraction is required.

[0032] In another example, the un-methylated cytosine of the unextracted DNA is directly converted to uracil by deamination using bisulfite to thereby generate the converted DNA. In another example, the un-methylated cytosine of the unextracted DNA is directly converted to uracil using direct conversion kits selected from the group consisting of EpiTect Fast FFPE Bisulfite Kit, innuCONVERT Bisulfite All-In-One Kit, and Zymo EZ DNA Methylation-Direct Kit.

[0033] The DNA used in the method of the first aspect is present in a biological sample. In one example, the biological sample containing the DNA is selected from the group consisting of a liquid sample, a tissue sample, or a cell sample. In another example, the liquid sample is bodily fluids selected from the group consisting of blood, bone marrow, cerebral spinal fluid, peritoneal fluid, pleural fluid, lymph fluid, ascites, serous fluid, sputum, lacrimal fluid, stool, urine, saliva, ductal fluid from breast, gastric juice, and pancreatic juice. In one example, the bodily fluid is blood. In some examples, the tissue sample may include, but is not limited to frozen tissue sample, fixed tissue sample (such as formalin-fixed tissue sample). In another example, the tissue sample or the cell sample may be any type of tissue or cell in the body. For example, the tissue sample or cell sample may be a tissue or cell from bone, epithelial, cartilage, adipose tissue, nerves, muscle, connective tissue, esophagus, stomach, liver, gallbladder, pancreas, adrenal glands, bladder, gallbladder, large intestine, small intestine, kidneys, liver, pancreas, colon, stomach, thymus, spleen, brain, spinal cord, heart, lungs, eyes, corneal, skin, or islet tissue or organs. The cell sample may also be from blood, such as white blood cells and platelets. In another example, the cell sample may be cancer cells, stem cells, endothelial cells, or fat cells.

[0034] In another example, the biological sample is obtained from a subject having and/or suspected of having a disease. In another example, the disease is cancer. In yet another example, the cancer is selected from the group consisting of leukemia, lymphoma, ovarian cancer, lung cancer, colorectal cancer, breast cancer, pancreatic cancer, prostate cancer, nasopharyngeal cancer, liver cancer, cholangiocarcinoma, esophageal cancer, urothelial cancer, and gastrointestinal cancer. In another example, the cancer is an early stage cancer. In another example, the cancer is a Stage I cancer. In another example, the cancer is a Stage II cancer. In another example, the cancer is a Stage III cancer. In another example, the cancer is a late stage cancer. In another example, the cancer is an original cancer. In another example, the cancer is a relapsed cancer. In another example, the cancer is relapsed if cancer cells are detected at, in the region of, or distant from the primary site of the tumour, about 1 week, about 2 weeks, about 3 weeks, about 1 month, about 2 months, about 3 months, about 4 months, about 5 months, about 6 months, about 7 months, about 8 months, about 9 months, about 10 months, about 11 months, about 1 year, about 2 years, about 3 years, about 4 years, about 5 years, about 6 years, about 7 years, about 8 years, about 9 years, or about 10 years after complete remission of the primary cancer. In another example, the disease is minimal residual disease of the primary cancer following curative surgery or therapy. As used herein, minimal residual disease (MRD) is a term used to describe the presence of tumour cells disseminated from the primary lesion to distant organs in patients who lack any clinical or radiological signs of metastasis, or residual tumour cells left behind after therapy, that eventually lead to cancer relapse.

[0035] In one example, the DNA is cell-free DNA (cfDNA). As used herein, cfDNA refers to non-encapsulated DNA which is circulating in a liquid sample disclosed herein and not contained within cells. In one example, plasma cfDNA is derived from both normal (healthy, non-diseased) cells and tumor cells. In one example, the DNA is circulating tumor DNA (ctDNA). In one example, the cfDNA fragments from tumor cells are shorter than cfDNA fragments from normal cells. In one example, the differences in plasma cfDNA concentrations and cfDNA fragment lengths between individuals with and without cancer can be assayed as cancer-specific signals. In one example, the liquid sample is bodily fluids selected from the group consisting of blood, bone marrow, cerebral spinal fluid, peritoneal fluid, pleural fluid, lymph fluid, ascites, serous fluid, sputum, lacrimal fluid, stool, urine, saliva, ductal fluid from breast, gastric juice, and pancreatic juice.

[0036] In another example, the DNA is encapsulated within tissues and/or cells. In another example, the tissue or cell may be any type of tissue or cell in the body. In some examples, the tissue sample may include, but is not limited to frozen tissue sample, fixed tissue sample (such as formalin-fixed tissue sample). In another example, the tissue is from bone, epithelial, cartilage, adipose tissue, nerves, muscle, connective tissue, esophagus, stomach, liver, gallbladder, pancreas, adrenal glands, bladder, gallbladder, large intestine, small intestine, kidneys, liver, pancreas, colon, stomach, thymus, spleen, brain, spinal cord, heart, lungs, eyes, corneal, skin, or islet tissue or organs. In one example, the cell is from bone, epithelial, cartilage, adipose tissue, nerves, muscle, connective tissue, esophagus, stomach, liver, gallbladder, pancreas, adrenal glands, bladder, gallbladder, large intestine, small intestine, kidneys, liver, pancreas, colon, stomach, thymus, spleen, brain, spinal cord, heart, lungs, eyes, corneal, skin, or islet tissue or organs. In another example, the cell may be a cancer cell, a stem cell, an endothelial cell, or a fat cell. In yet another example, the cell is a blood cell. The blood cell may be a white blood cell, or a platelet.

[0037] As used herein, when cfDNA or DNA encapsulated in blood cells in the peripheral blood is used, the method as disclosed herein is carried out on a non-invasive basis.

[0038] In one example, the amount of DNA used in the method disclosed herein is at least 5 ng. In another example, the amount of DNA used in the method disclosed herein is about 5 ng, or about 10 ng, or about 15 ng, or about 20 ng, or about 30 ng, or about 40 ng, or about 50 ng, or about 60 ng, or about 70 ng, or about 80 ng, or about 90 ng, or about 100 ng, or about 110 ng, or about 120 ng, or about 130 ng, or about 140 ng, or about 150 ng, or about 160 ng, or about 170 ng, or about 180 ng, or about 190 ng, or about 200 ng, or about 300 ng, or about 400 ng, or about 500 ng, or about 600 ng, or about 700 ng, or about 800 ng, or about 900 ng, or about 1000 ng, or at least 1000 ng.

[0039] After conversion of un-methylated cytosine of the DNA to uracil by deamination, the converted DNA is then purified as disclosed in step (b) of the method of the first aspect, using an agent such as DNA purification beads. The DNA purification beads may be paramagnetic beads, such as AMPure XP beads, and SPRI beads.

[0040] The converted and purified DNA is then tagged with a barcode sequence by performing a first PCR amplification using a plurality of forward and reverse primer pairs specific to the converted sequence of target regions, wherein each forward primer of the

plurality of forward and reverse primer pairs comprises a barcode sequence on its 5' end, wherein the barcode sequence of each forward primer is different from each other, wherein the target regions comprise CpG sites, as disclosed in step (c) of the method of the first aspect. As used herein, the term "barcode sequence" is a commonly used term in the art of nucleic acid sequencing and used within the definition as known in the art. Thus, the term "barcode sequence" refers to the encoded molecules or barcodes that include variable amount of information within the nucleic acid sequence. For example, the barcode sequence is a tag that can be read out using any of a variety of sequence identification techniques, for example, nucleic acid sequencing, probe hybridization based assay, and the like. In some examples, the barcode sequence is used in the method as described herein to tag different converted DNA sequences of target regions of a sample, such that when the barcode sequence tags to the converted DNA sequences of target regions, each different converted DNA sequence of target region would then have a unique barcode sequence that is attached to it and read out with the converted DNA sequence of target region from the sample.

[0041] In one example, the barcode sequence is an oligonucleotide comprising 10 to 16 random nucleotides, or 10 to 15 random nucleotides, or 10 to 13 random nucleotides, or 10 random nucleotides, or 11 random nucleotides, or 12 random nucleotides, or 13 random nucleotides, or 14 random nucleotides, or 15 random nucleotides, or 16 random nucleotides. In another example, the barcode sequence is an oligonucleotide comprising 10 random nucleotides. As exemplified in the Experimental Section (Fig. 2), the barcode sequence may be defined as NNNNNNNNNN, which may have the sequences such as, but is not limited to, TAGCTAACGT, GCAAGGTCAA, ACCTGTGTAT and the like.

[0042] In one example, the number of the forward and reverse primer pairs is at least 5. In another example, the number of the forward and reverse primer pairs is at least 10. In another example, the number of the forward and reverse primer pairs is at least 15. In another example, the number of the forward and reverse primer pairs is at least 20. In another example, the number of the forward and reverse primer pairs is at least 30. In another example, the number of the forward and reverse primer pairs is at least 40. In another example, the number of the forward and reverse primer pairs is at least 50. In another example, the number of the forward and reverse primer pairs is at least 60. In another example, the number of the forward and reverse primer pairs is at least 70. In another example, the number of the forward and reverse primer pairs is at least 80. In

another example, the number of the forward and reverse primer pairs is at least 90. In another example, the number of the forward and reverse primer pairs is at least 100. In another example, the number of the forward and reverse primer pairs is at least 110. In another example, the number of the forward and reverse primer pairs is at least 120. In another example, the number of the forward and reverse primer pairs is at least 130. In another example, the number of the forward and reverse primer pairs is at least 140. In another example, the number of the forward and reverse primer pairs is at least 150. In another example, the number of the forward and reverse primer pairs is at least 160. In another example, the number of the forward and reverse primer pairs is at least 170. In another example, the number of the forward and reverse primer pairs is at least 180. In another example, the number of the forward and reverse primer pairs is at least 190. In another example, the number of the forward and reverse primer pairs is at least 200. In another example, the number of the forward and reverse primer pairs is 5. In another example, the number of the forward and reverse primer pairs is 22. In another example, the number of the forward and reverse primer pairs is 95. In another example, the number of the forward and reverse primer pairs is 159. In another example, there is no upper limit on the number of the forward and reverse primer pairs.

[0043] In another example, the forward and reverse primer pairs comprise sequences as disclosed in Table 1.

[0044] **Table 1. Sample primer sequences (159 pairs).**

Primer name	Sequence (Read 1)	Primer name	Sequence (Read 2)
ADCY4_methyl_1F	GGGAGGGTTTGAAAA AGGAG	ADCY4_methyl_1R	ACCATAATCTCCCCAACC C
ADHFE1_methyl_1F	TGGTTTGAGGTTTAG ATAGGTG	ADHFE1_methyl_1R	CTTAAAATCCTCTTCCCT CCT
ALX4_methyl_1_1F	GATAAGAAGGAGATA GAATAGATTTTT	ALX4_methyl_1R	CATCCCTTAAACTAATA ACTTAC
ALX4_methyl_1_2F	GGGTTTGAGGGAGTA TAGATTT	ALX4_methyl_2R	AACACCTTTTCCTAAATT TCCAAC
ALX4_methyl_1_3F	TAAGTTTGTAGTATTTA TGTTTGGTTTG	ALX4_methyl_3R	CCCAAACCTCCAACCAA AA
ALX4_methyl_1_4F	GGTAAGGAGTGTATA GTTATAGTTTT	ALX4_methyl_4R	CAAATCTCAACATTCATA CCTAACTT
AMPD3_methyl_1F	GTTTAGGAAGGATGG GTTAGTTG	AMPD3_methyl_1R	CAACCCTATAATACTACT CTATACACAC
ANKRD54_methyl_1F	GATAGGTTTTTGGGA TTAGGATATGG	ANKRD54_methyl_1R	TCCTCCCAATCATTAATA CACC
APC_methyl_1F	GAGGGTAAGTAGTAA	APC_methyl_1R	CTAAATACTTCACCTTCC

1F	GGGGG	1R	TCACC
APC_methyl_2F	TATTGGAGATAGAATGGAGGTGTT	APC_methyl_2R	TTAACATACTACAATAAACCCCTATACCAAA
APC_methyl_3F	TAGTGGGGTATAGGGTTT	APC_methyl_3R	CATCAACTTCCTTACTTACTAAAAACT
APC_methyl_4F	ATATGTGGTTGTATTGTGTAGTT	APC_methyl_4R	CCATAATAACTCCAACACCTACC
APC_methyl_5F	GAGAGAGAAGTAGTTGTGTAATT	APC_methyl_5R	CATTCTATCTCCAATAACACCCTA
ATG9B_methyl_1F	GTTTTGAGGGATTTGGTTTTAGG	ATG9B_methyl_1R	CAACAACATTACTAAATAACTCAACC
ATOH8_methyl_1F	GGAGTTGGGATTTTTGTTGG	ATOH8_methyl_1R	ACTCAATTTCCCTATCTTAAACCC
AUTS2_methyl_1F	ATAAGGAAGATAATGGGAAGTTTT	AUTS2_methyl_1R	CTAATCATAACAAATCCATCAATAATATCCTC
BAIAP2_methyl_1F	GTTGAGTTAGGTAGGGGAGG	BAIAP2_methyl_1R	CCACAACAAAACCTCAAAACC
BCAT1_methyl_1F	GGGTTTGTTTATATYGTGTTTT	BCAT1_methyl_1R	CAATCCTCCCCCCTTC
BCAT1_methyl_2F	GTTTGGGGGAGTAGTTTAGT	BCAT1_methyl_2R	ACTACAACAAAACCTAAACAATACC
BCAT1_methyl_3F	GTTGTAGTAAGATTTGGGGTAGTG	BCAT1_methyl_3R	CAACCTTAATATCTTCCTACTAATACAATC
BCKDK_methyl_1F	GGTTTAGTGAATTTGAATGGGG	BCKDK_methyl_1R	CACCCTAAAACCTCCAAAC
C9orf50_methyl_1F	TTAGGTAGTTAGGGTTTGAGAG	C9orf50_methyl_1R	CACCTCTCCTAAAACTAATACAAAACTC
CALCA_methyl_1F	TTTTTAGGGGGAAGAGATGTAG	CALCA_methyl_1R	TTAAAACAAAACAAAAATTAACCCCTA
CALCA_methyl_2F	TATAGAGTTTGAGGTAAAGAGAAAT	CALCA_methyl_2R	AACCTATAACTAAAACCTTACAATTTAAAC
CALCA_methyl_3F	GGGGATTGAGGATAAAGGTG	CALCA_methyl_3R	CAAAACAAAATTAACCCCTAAATATAC
CALCA_methyl_4F	TGGGTTTTAGATAGGTAGGGAT	CALCA_methyl_4R	CCCTCCTCACACATATAATAAATCTAAA
CCDC102A_methyl_1F	GTGGTYGGATTGTATGTTAATTG	CCDC102A_methyl_1R	CCCRCCAACCTCCTTAATAAA
CCL26_methyl_1F	GGAAGAAGGAAAGTAGTAAGAAAGA	CCL26_methyl_1R	CTAATAAAACCAAATTACTTTTATATTCACC
CCNA1_methyl_1F	GTTYGGTTTTGGGAGGGA	CCNA1_methyl_1R	TCTCCCRAACCAAAATTC
CDH13_methyl_1F	TTGAGGTGTAGAGGTGATTGTAGG	CDH13_methyl_1R	AAAAAATTCAAACCTTACCCACC
CDH13_methyl_2F	GTTTTAATAGTTTAAAGAAGTAAATGGGA	CDH13_methyl_2R	AACCAAAACCAATAACTTTACAAAA
CDH13_methyl_3F	AAGAAGTAAATGGGATGTTATTTT	CDH13_methyl_3R	CCAAAACCAATAACTTTACAAAA

CDH13_methyl_4F	TTTTTTTATTTGGGATAGGAGAA	CDH13_methyl_4R	CCTCAAAACCTAACTCCCA
CDH13_methyl_5F	TGGGAAGTTGGTTGGTTG	CDH13_methyl_5R	CCCTCTTCCTACCTAAAC
CDKN2A_methyl_1F	GGGGAGATTTAATTTGGGG	CDKN2A_methyl_1R	CCTATCCCTCAAATCCTCTAAAAA
CDKN2A_methyl_2F	TTGGTTGAGGAGTTGGTTA	CDKN2A_methyl_2R	CTCTAAACTTTAAAAACTCTCAAAATAC
CDKN2A_methyl_3F	GGGGTAGTTGTGGTTTGTG	CDKN2A_methyl_3R	ATACCACACATCTTTAACCTCAAAAT
CDKN2A_methyl_4F	GTAGTTGTGGTTTTGTAGGATTTT	CDKN2A_methyl_4R	CTACCCTCTACAAACATCC
CFTR_methyl_1F	GGGAAGTTGGTGGTAG	CFTR_methyl_1R	CCTCTCTTTAAATCCAATTAACAA
CLIP4_methyl_1_1F	AGGTTGAGGGTTGTGAAGG	CLIP4_methyl_1_1R	CACTCCCCAACAATCAAATAAT
CLIP4_methyl_1_2F	GGGTTYGGAGAGGTGTG	CLIP4_methyl_1_2R	CCTACCTCCTCCRAATCCTC
CUX1_methyl_1_1F	GGTTAAGTGTAGGGGATAGAAGA	CUX1_methyl_1_1R	TAACCCCCTACCACCCTCT
DLK1_methyl_1_2F	GGTAGTAGATGATGGGTAGAAAATTG	DLK1_methyl_1_1R	CAAACCTTCTATTATCTCCAAACATATC
DOK7_methyl_1_1F	GGTAGATTAGTATTGGAGATGGGTA	DOK7_methyl_1_1R	CAAAACCTAAACACAAACCCAC
DQ580846_methyl_1F	GAAGTTTTTAGGATGGGGAGG	DQ580846_methyl_1R	CACTAATACAAAAACTATCAAACCTCTC
DYNLL2_methyl_1F	GGTTTAAGAGATTATTTTTGGGAGAG	DYNLL2_methyl_1R	ACCAAAATACTTAATTCCCAATAAAAAAC
EP300_methyl_1_1F	TATTTAGGTGGTTAAATGGGAG	EP300_methyl_1_1R	CAATAAATACACCCTACCTAATAAAT
EPHA10_methyl_1F	TGTTATTGGTTTTAGGGTTT	EPHA10_methyl_1R	TAAAATAAATCAAAATAAAAAATCCCA
EPHA10_methyl_2F	TTTTGGGAGGAGGATGAGATT	EPHA10_methyl_2R	CTCCAAATCCCTCCCAAAAC
EPHA10_methyl_3F	GGTGGGTAAAGGTGGAGAAT	EPHA10_methyl_3R	TATCATTAACCCCAAAAACCTC
FAIM2_methyl_1F	GAGAGGGTYGGGTTTGG	FAIM2_methyl_1R	AACCCATCCACTAAAATTACTC
FOXA1_methyl_1F	GGATAGAGTAGGGTAGAGGTTT	FOXA1_methyl_1R	CCAAACAAAACAAACAACA
FOXA1_methyl_2F	TAGGGTTGGTAGATAGTAGGTT	FOXA1_methyl_2R	TCCACACCAAACATAACTAAATTACTAAA
FOXA1_methyl_3F	GTTTAGAAGTGAGGAAGGTTTG	FOXA1_methyl_3R	TAACCCCTTCCCCAACT
FOXD3_methyl_1F	GGAAGGTTTTTGTGTGTTTTG	FOXD3_methyl_1R	CCTATAAACTTTTAACTACAACTACAAC
FRMD4B_methyl_1F	GGGTATAGTTGTAAG	FRMD4B_methyl_1R	CTATCCAAACTCCAACCTC

thyl_1F	TGTGGAGT	thyl_1R	AACA
GALR1_methyl_1F	GAGTTTGGGAAGAGG GATT	GALR1_methyl_1R	ACAACAAATACAACACR CAACC
GLI2-far_methyl_2F	GAGGGTTYGTGAGAG TTAATGAG	GLI2-far_methyl_2R	CCAATAACTAACCTAAA ACRAAACC
GPC6_methyl_1F	GAGGGGTAAGGTGAA GAG	GPC6_methyl_1R	CCAAACTAAATCCCTTA ACAC
GRASP_methyl_1F	TAGGAAGTTGTAGTA GAAGGAGG	GRASP_methyl_1R	ACAACCTCRACAAAATAA TAATCCTCC
GRIA4_methyl_1F	GTATTTAGGGGGTGA AGTTAGT	GRIA4_methyl_1R	CTCACATACACACAAAC ACTAA
GRK5_methyl_1F	GTAGATGAGTTATYG GTTTTTGTGG	GRK5_methyl_1R	CCCTAAAAACCRAAAAA ACTACTCTC
GSTP1_methyl_1F	GAGGGAATAGGGAAA TAGGTTTT	GSTP1_methyl_1R	ACTACAACCCACAACCC CT
GSTP1_methyl_2F	GTTGGGAGTTTTGAG TTTTATTTT	GSTP1_methyl_2R	AAACCACCCTTATAAAA CTC
HOXA9_methyl_1F	GGAGTTGGTTAGGAG GGAG	HOXA9_methyl_1R	CCTAAAAATCCTAATTAC CAACTAATAAAA
HOXA9_methyl_2F	GGGAGGTTGGTTTAG GG	HOXA9_methyl_2R	CCACCACTAAAACCCTA AACAAC
HOXB4_methyl_1F	GGTTTAGGGTTYGG TAGGT	HOXB4_methyl_1R	TTCCCTCCATACRAAAAA TATTCAC
HOXC13-far_methyl_1F	AGGGTTYGAAGATT TGGGG	HOXC13-far_methyl_1R	CAAACCCCTCTCCAAACT TCTC
HSP90AA1_methyl_1F	TAATTTGTTTAGGAGT AGGGGT	HSP90AA1_methyl_1R	CTACTAATTCCTTCCAC TTCCC
HSP90AA1_methyl_2F	GTAGTTGTAGGAAAG TTTGGGT	HSP90AA1_methyl_2R	CCCTTAAAAACRTATA AAATCAACA
IKZF1_methyl_1F	GTTYGGAGTTGYGGT TGAGA	IKZF1_methyl_1R	ATCCCTACTCRACCTACC C
intergenic_methyl_1F	TAGTTATAGGAGTAT TATTATGTAGTTGGT	intergenic_methyl_1R	CCAACCTAAAATATTCA ATACAAACTCAAC
intergenic_methyl_2F	GGGGAAGTTTTTGTT GTTTGT	intergenic_methyl_2R	ATCACATTTAAACTAACT TTCTCACAC
intergenic_methyl_3F	GGTTTTAAGGTAGAG ATTTTAAAGAAGG	intergenic_methyl_3R	CCTATTTTCAAAAAATA TAACACACCC
IRF4_methyl_1F	GGTTTATAGGGTYGG GTTTG	IRF4_methyl_1R	ACTCTACCCRAACCTCCC
KCNQ5_methyl_1F	GGGTAGGGTGTTGTT GAATT	KCNQ5_methyl_1R	ACTTCCCCAACAACTC ATC
KIF1A_methyl_1F	TTAGGGTGAGGTTAG TGTTAGTT	KIF1A_methyl_1R	CTTCCCCCAAAAATAA AAAACC
KLF16_methyl_1F	GGTTTAGTTTGAGTTA GAGGGTTAAG	KLF16_methyl_1R	CAAATACACAACCACCC AACTAC

LINC00114_methyl_1F	TGTTATTGTGATGAA GGGATTAATG	LINC00114_methyl_1R	CATAACTATATATACAA CACTCAACTTTTAC
LMX1B_methyl_1F	GTTTAAGGGGTGTGG ATTTAGTAG	LMX1B_methyl_1R	CTCCTACCAATCATATCC CA
LOC153684_methyl_1F	GAATTAAGTAGATTT TGTTGGTTTTAGG	LOC153684_methyl_1R	CCCCTCCTATAAAATCCT ATAAACTC
LRIG1_methyl_1F	GTTATTGTTATTGGTT GGTGGTTAG	LRIG1_methyl_1R	CACAAACCTTCCTTCCAA CAC
MARCH11_methyl_1F	AGGTTGYGGGGGAGG TT	MARCH11_methyl_1R	CTCTCTTCCCTACAACC AAAACC
MARCH11_methyl_2F	AAGGGGGGGTYGGAG	MARCH11_methyl_2R	AAATCCCCCACTTCCAAC C
MARCH11_methyl_3F	TTAGTGGGAYGGAGT TYG	MARCH11_methyl_3R	CCCCTACTTCCCAACTCA
MYC_methyl_1F	GGATTGTGTTTGTGTG GTTGTTG	MYC_methyl_1R	ACCCCCTTCTTAACATAAT CAAAAC
MYO1G_methyl_1F	GTTTTATAAGGAGGT TGTGTT	MYO1G_methyl_1R	CCAAAATTAAAAACCAA CCCAC
NALCN_methyl_1F	GTTTTGTTATAGGGAT TGTAAGTAGT	NALCN_methyl_1R	AAAACCTAACTATTACT ATAAACTAACTA
NDRG4_methyl_1F	GGTTTGGGGTTAGTTT YGGA	NDRG4_methyl_1R	CAACRCACCCAACACAA TC
NXPH2_methyl_1F	GGTATTTATTAGTTGT AGTAAGTTAGGG	NXPH2_methyl_1R	TCCTCAATCCTACCCC
PCDH17_methyl_1F	GATGAAATTGATGGG GATGAAGAG	PCDH17_methyl_1aR	AAATCTCTAAATACAAC AAAACACTACC
PCDHGA12_methyl_1F	GATTATAAGTAGTTG YGTGTTTAAGG	PCDHGA12_methyl_1R	CTACTCACCACCTAATCC TT
PRIMA1_methyl_1F	GTTYGGGGTATTGGG GTG	PRIMA1_methyl_1R	CTCCACCCACRACTTTAC T
PTGER4_methyl_1F	AAGTGGTTATAATTT AGAAAGTAGG	PTGER4_methyl_1R	ACCTTAAATACAAAAC TAAATCTAAAAAT
PTGER4_methyl_2F	TTAGATTTAGTTTTGT ATTTTAAGGTTG	PTGER4_methyl_2R	TAACCACCAAATTACCC ACC
PUF60_methyl_1F	GAATTATGATTTGAG GGATAGGTTTG	PUF60_methyl_1R	CCTTCAAACCAAACATTC CCTA
RASSF1A_methyl_1F	TTTTTAAGGAGGGAG TAGGAGTAT	RASSF1A_methyl_1R	AACCCCTTAATTTCTAT CACAAAC
RASSF1A_methyl_2F	GAAGGGGTAGTTAAG GGGTAG	RASSF1A_methyl_2R	CCACCTCTACTCATCTAT AACCCAA
RASSF1A_methyl_3F	TTTGGGTGTAGGGAT TGTG	RASSF1A_methyl_3R	AATAAACTAACCTCCA AAAACAC
RASSF1A_methyl_5F	GTTGGTAGAAGGGGG TGA	RASSF1A_methyl_5R	CCAAAAAACATAATAC AACATACAAATCC
RASSF1A_methyl_6F	GAAATTTGGGTGTAG GGATTGTG	RASSF1A_methyl_6R	CACAATAAACTAACCT CCAAAAAC
RPL13A_methyl_1F	GGGGTTTATGATAGA	RPL13A_methyl_1R	CCCTTATCTCCACAAAT

hyl_1F	GGGGT	hyl_1R	CC
RPL23A_met hyl_1F	TGTATTTGGTAGTTAT GGTTTGTTG	RPL23A_met hyl_1R	AACTCTACTCAAATCTAT ACTCCC
RPL3_methyl _1F	TGGGTTTTTTTAGGGT AGAAAGG	RPL3_methyl _1R	CTTCTAAAAAACCTTTTC CACCAAC
RPL7A_meth yl_1F	TGTTTTTATTGGGAGT TGTTGAG	RPL7A_meth yl_1R	TACAACAAAATTACAAT CCTAATACCT
RPS18_methy l_1F	GTTAGTGGTTTTGGTG TGTTG	RPS18_methy l_1R	CTCATCTCCCTACCTACC T
RPS18_methy l_2F	AAGTGGGGAGGAGGT ATG	RPS18_methy l_2R	TACCCAATAATCCTCCAA TTCTC
SBNO2_meth yl_1F	GGGTTGATTTTTAGG AAATGATGT	SBNO2_meth yl_1R	CATACCCACCTAAACATT CTACA
SCT_methyl_ 1F	GGAGTAGTAGTAGTA GTAGGAGGG	SCT_methyl_ 1R	ACAACTAAAAAACRAC CCTAACC
SDC2_methyl _1F	GATGGGGGTTAGATT TAAGAGATT	SDC2_methyl _1R	CCCCAATATCCCCTTAAA AA
SDC2_methyl _2F	GGAAGGTTGTTTTTG AAGTTT	SDC2_methyl _2R	ATTTATTCCTTAAAACCA AAAACCTAACC
SDC2_methyl _3F	GAGTGTAGAAATTAA TAAGTGAGAGGG	SDC2_methyl _3R	CAAAACRAAACCTCCTA CCCAA
SDC2_methyl _4F	GAGAAAGGAAGAAA AGGATTGAG	SDC2_methyl _4R	CRACAATATAACTCCCA AATAAACC
SEPT9_meth yl_1F	AGAGTTAAAGGGAGG GGAT	SEPT9_methy l_1R	CCCTCACAAAACCCTAT ACT
SEPT9_meth yl_2F	GTGGAGTGAGGATGT TAGT	SEPT9_methy l_2R	TTCTCCTCTACTTAATAA AACA
SEPT9_meth yl_5F	AATAGGGTGATAAGA GGGATGATTT	SEPT9_methy l_5R	AAAAAACTTTTATTACAC AAAAACCC
SEPT9_meth yl_6F	AGGGGGTTATGAGTA GATTAGGAGT	SEPT9_methy l_6R	CCCAAATAAAACTTTTC CTCAACTTTA
SEPT9_meth yl_10F	GGTGGTTTTTYGTTGG TGTTATG	SEPT9_methy l_10R	AACAACCAAACAAAATT CTCTATCAC
SEPT9_meth yl_11F	GYGGTTAGTTTTGTAT TGTAGGAG	SEPT9_methy l_11R	ATCCRAAATAATCCCATC CAAC
SFRP1_meth yl_1F	GGAATTTGTTATATTT AAGTATTTYGGGT	SFRP1_methy l_1R	ACACCCAAATCTTCCTCT ACT
SFRP2_meth yl_1F	GTAGAAATTTYGGAT TGGGGTA	SFRP2_methy l_1R	CTTTAAACAACAAACAA AAAAACCTAAC
SHOX2_meth yl_1F	AGATTATTTGGTTGTT TAGGGGA	SHOX2_meth yl_1R	CAAACTTACTAATCTAAT TTAAAAACAAT
SHOX2_meth yl_2F	TAGGTTAGTATGGGT GTTAAAAA	SHOX2_meth yl_2R	CCATAAATCCACAACCTA CAAACC
SHOX2_meth yl_5F	GAGAGAATAGAAAGA GYGGTTGT	SHOX2_meth yl_5R	ATCCCCTAAACAACCAA ATAATCTC
SHOX2_meth yl_6F	GGGATTTTATGTAGG GTTTGGT	SHOX2_meth yl_6R	TCTCAAACCAACTCTCAC CAC

SHOX2_methyl_7F	GAATTTGTTGATTTTT GTGGGTAAAG	SHOX2_methyl_7R	CCCACRCCATATTAAC TAACTA
SKAP2_methyl_1F	GTAAGTATAGTAAAT TAGTAAGAGYGTTAG	SKAP2_methyl_1R	AAACTAACCTTTTCTTAT TATCAAATTACT
SLC12A7_methyl_1F	GTTATGGTATATAGG TGTAATGAGTTAG	SLC12A7_methyl_1R	CATAACCAAACCATACC CATTTTC
SLC25A25_methyl_1F	GTTGTTAATGGTTGGT AATTGGGATA	SLC25A25_methyl_1R	AACATTTCTACCTCCTA AAACT
SNHG1_methyl_1F	GAAATGATTTTTTGG GGAGGG	SNHG1_methyl_1R	CTCTAACATACTATTAAC TTTTATACCTCAC
SNHG1_methyl_2F	ATTTGTTTGGTTTAAA GGTGAGT	SNHG1_methyl_2R	CAAATTTCAAAATCTCA ACAACAAATCATC
SNHG32_methyl_1F	GGGAATGATGATTTT ATAGATTAGAGTT	SNHG32_methyl_1R	CAAAATTATAACCTCCC ATTAACC
SOX17_methyl_1F	GGGGTAGGTGTAGTT TTGGG	SOX17_methyl_1R	CCTTTTCTACACAAATAT AACCAATAAAA
SOX17_methyl_2F	GGGGAAATAATTTTT ATAAGTAAGAG	SOX17_methyl_2R	CTTCCTTACCCAACCTTT AAAAC
SOX17_methyl_4F	TTTATGGTGTGGGTTA AGGA	SOX17_methyl_4R	CTCACCCAACATCTTACT CAAC
SOX17_methyl_6F	GAAAAGTAAGGTTTG TAGGAAGTTAT	SOX17_methyl_6R	ATAAAACCTTCTATTTCA ATCTCTTCAC
ST6GALNAC5_methyl_1F	GTGGTTATTTTAGTGG GGGAGA	ST6GALNAC5_methyl_1R	TACACCACCTACCCTCTA C
TAC1_methyl_2F	GGTTTAAGGAGGTTG GGATAAATAT	TAC1_methyl_2R	AAATACTAAATTTCTCTA ATTCCTCC
TAP2_methyl_1F	GGTATGAAAAATAAT ATAAGAATGTGTTGG	TAP2_methyl_1R	CTTAATATAAAAACTACT AAAACCTCTCCAAAC
TRH_methyl_1F	GTGGAGTTTATTTTTG GTAGGTAAT	TRH_methyl_1R	CCCTATTCTCCAAAAAAC TCTC
TRIM15_methyl_1F	GGAGAGTAGAAGAGG GGATT	TRIM15_methyl_1R	AACAAATAAAAAACCATA AACTTCTAACTC
TRIM58_methyl_1F	TTGYGGTTATAGTTTT TGTTTTAGGT	TRIM58_methyl_2F	ACCAAACCCRCCAAC TAACTA
TSPAN10_methyl_1F	GTATGGGAATGGGGG AG	TSPAN10_methyl_1R	CATACCTACCTAAATCAC ACCTA
TSPAN4_methyl_1F	GTATTAAGTTAGTTG GGGTGGG	TSPAN4_methyl_1R	CTAACAACCCCAAAACA AACC
TSPAN4_methyl_2F	GGGGTTGYGATTGGT T	TSPAN4_methyl_2R	CCCAAACATCTACCCRA C
TSPAN4_methyl_3F	GAGYGTTGGGTTTG T	TSPAN4_methyl_3R	CCCAAACATCTACCCRA CC
TWIST1_methyl_1F	TTAAGTGAGGTGGGA AGGTTG	TWIST1_methyl_1R	CCACCCCTCAACAAAA C
TWIST1_methyl_2F	TTTTTTTATGGATTGG GTTTTT	TWIST1_methyl_2R	TCTAAACCCTCCTTTCCC CT

TWIST1_met hyl_3F	GAGTGGTTGTGATAG TAGTAATGG	TWIST1_met hyl_3R	TAACACTTTTCTTAACAT ACCCCC
TWIST1_met hyl_5F	AATTTTGTTTTAGTTT AGAAAGGGG	TWIST1_met hyl_5R	AACTAAAATCAATACAC AATATTAAAATAAAAAC
TWIST1_met hyl_7F	AGGTGTTTGGGAGTT GGG	TWIST1_met hyl_7R	AAAAATCCCTCCTCCTCA
TWIST1_met hyl_8F	GTTGGAGTTAGTATA TAGTGTTGGG	TWIST1_met hyl_8R	CTACAACRCRAACCTTA CTCC
VWC2_meth yl_1F	TAGTAAAGGGAGYGG GTG	VWC2_meth yl_1R	CTCCCTCCTCTACTACCC
WIPI2_meth yl_1F	TAGGTTGAGATAGTG GGGAT	WIPI2_meth yl_1R	CCCTCTAACCTTACACCA C
YTHDC1_me thyl_1F	TTAGGTATTAGTGGG TTGAGGG	YTHDC1_me thyl_1R	TTCCTCTCTAAAAACRTC ACAATA
ZNF577_met hyl_1F	GTAGTAGGGTTGAAA ATGGTTG	ZNF577_met hyl_1R	AACCCCACACACTACAA A
ZNF678_met hyl_1F	GTGTAGGAGATAGAG GTTTTTGG	ZNF678_met hyl_1R	ACTAATCTACTCTCCCC TCTAC
ZNF737_met hyl_1F	GTTTTAGGTTTTGTTT TTATTGTTGTGTG	ZNF737_met hyl_1R	ACACTCTCACCATTCTA AACTTCC
ZNRF4- far_methyl_1 F	TTAGGGAATTGTTTA AGTAGYGTT	ZNRF4- far_methyl_1 R	ACTAATCCCAACCAAAC CC

[0045] The exemplified sequences disclosed in Table 1 show only the target-specific sequences of each primer. These sequences do not show the barcode sequence (for forward primers only) and the adaptor sequence required for the second amplification with universal indexed primers.

5 [0046] The full sequence of each forward primer used in step (c) of the method of the first aspect contains the adaptor sequence, followed by the barcode sequence and then the target-specific sequence (the sequences disclosed in Table 1). **Fig. 2** shows the full sequence of CLIP4_methyl_2F, which is one exemplary forward primer among the 159 primer pairs comprising the target-specific sequences in Table 1.

10 [0047] The full sequence of each reverse primer used in step (c) of the method of the first aspect contains the adaptor sequence followed by the target-specific sequence (the sequences disclosed in Table 1). **Fig. 2** shows the full sequence of CLIP4_methyl_2R, which is one exemplary reverse primer among the 159 primer pairs comprising the target-specific sequences in Table 1.

15 [0048] In another example, the primer pair comprises degenerate bases. In one example, the forward primer in the primer pair comprises one or more degenerate bases, while the

reverse primer in the primer pair has no degenerate base. In another example, the reverse primer in the primer pair comprises one or more degenerate bases, while the forward primer in the primer pair has no degenerate base. In yet another example, both the forward and reverse primers in the primer pair comprise one or more degenerate bases. As used herein, 5 degenerate primers are used when the primer landing site overlaps with a CpG site. A CpG site bound by the forward primer has a sequence of either CG (methylated) or TG (un-methylated). The degenerate base Y is used in forward primers to specify either a cytosine or thymine, thus allowing the primer to cover both un-methylated and methylated DNA. In addition, a CpG site bound by reverse primers has a sequence of either CA (un-methylated) 10 or CG (methylated). The degenerate base R is used in reverse primers to specify either an adenine or guanine, thus allowing the primer to cover both un-methylated and methylated DNA. In another example, the degenerate base is selected from the group consisting of C, T, A and G. In another example, each primer of the primer pair comprises 1, 2, or 3 degenerate bases. In another example, each primer of the primer pair has one degenerate base. In 15 another example, the primer pair does not comprise a degenerate base, i.e. has no degenerate base.

[0049] In one example, the target regions comprise CpG sites. As used herein, CpG site refers to a cytosine that immediately precedes a guanine base. In vertebrates, DNA methylation occurs at cytosines within a CpG site.

20 [0050] In one example, each forward and reverse primer pair covers a target region which comprises at least 1 CpG site. In one example, each forward and reverse primer pair covers a target region which comprises at least 2 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at least 3 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at least 5 CpG sites. 25 In one example, each forward and reverse primer pair covers a target region which comprises at least 8 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at least 10 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at least 15 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at 30 least 20 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at least 25 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at least 30 CpG sites. In one example,

each forward and reverse primer pair covers a target region which comprises at least 35 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at least 40 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at least 50 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at least 60 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at least 70 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at least 80 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at least 90 CpG sites. In one example, each forward and reverse primer pair covers a target region which comprises at least 100 CpG sites. In another example, there is no upper limit on the number of CpG sites within the target region covered by each forward and reverse primer pair.

[0051] The first PCR amplification comprises a number of PCR cycles selected from the group consisting of 2, 3, 4 and 5 PCR cycles. In one example, the first PCR amplification comprises 2 PCR cycles. In one example, the first PCR amplification comprises 3 PCR cycles. In one example, the first PCR amplification comprises 4 PCR cycles. In one example, the first PCR amplification comprises 5 PCR cycles. As each forward primer carries on its 5' end a randomly assigned barcode sequence as disclosed herein, the first PCR amplification allows individual DNA molecules to be tagged uniquely in this first step of sequencing library formation.

[0052] After the first PCR amplification, a second PCR amplification is performed with universal indexed primers as disclosed in step (d) of the method of the first aspect, to create a sequencing library with components required for multiplex sequencing on a next-generation sequencing platform selected from the group consisting of Illumina platform, Ion Torrent sequencing technology, MGI sequencing platform, Oxford Nanopore sequencing, PacBio SMRT sequencing and 10X Genomics platform, as disclosed in step (e) of the method of the first aspect.

[0053] In one example, the universal indexed primers used in step (d) of the method of the first aspect are shown in **Fig. 2**, which comprise:

a forward primer comprising the sequence of
AATGATACGGCGACCACCGAGATCTACACCTAGCGCTACACTCTTTCCCTACAC
GACGCTCTTCCGATC*T; and

a reverse primer comprising the sequence of
CAAGCAGAAGACGGCATACGAGATAAACCGCGGGTGACTGGAGTTCAGACGTGT
GCTCTTCCGATC*T.

[0054] The above exemplary sequences of the universal indexed primers used in step (d)
5 of the method of the first aspect are the Indexed Illumina primers. The underlined index
barcodes are 8 bp barcode sequences that are specified by Illumina. The underlined part can
vary for different samples. Each sample within each sequencing run will have a unique
combination of forward and reverse indexes. In another example, the underlined index
barcodes has the sequences provided by Illumina for next-generation sequencing on the
10 Illumina platform, and may be any sequence listed in the “Illumina Adapter Sequences”
handbook, February 2019 version (<https://dnatech.genomecenter.ucdavis.edu/wp-content/uploads/2019/03/illumina-adapter-sequences-2019-1000000002694-10>). Exemplary
index barcodes for forward primers that may be used are listed in the column “i5 Bases for
Sample Sheet iSeq, MiniSeq, NextSeq, HiSeq 3000/4000”, for example CTAGCGCT and
15 TCGATATC. Exemplary index barcodes for reverse primers that may be used are listed in
the column “i7 Bases in Adapter”, for example AACCGCGG and GGTTATAA.

[0055] After next-generation sequencing, the presence of a barcode sequence is then
detected using Bioinformatics methods to count and assign each DNA sequence from the
next-generation sequencing to an original parental DNA molecule carrying the same barcode
20 sequence, as disclosed in step (f) of the method of the first aspect, comprising:

- (i) performing cluster reassignment of sequencing reads with the same barcode
sequence to thereby generate barcode clusters wherein each barcode cluster contains reads
from the same amplicon and with the same barcode sequence; and
- (ii) performing consensus calling for each barcode cluster to thereby obtain
25 consensus reads.

[0056] In one example, the assignment of DNA sequence to an original parental DNA
molecule refers to the cluster reassignment of sequencing reads with the same barcode
sequence. This generates barcode clusters wherein each cluster contains reads from the same
amplicon and with the same barcode sequence. Consensus calling is performed for each
30 barcode cluster to obtain the consensus reads. These consensus reads are the DNA sequence
that is subsequently compared to the reference genome for variations to be detected. The

initial step of cluster reassignment and generation of barcode clusters is important because it greatly reduces sequencing errors and improves confidence for accurate variant calling.

[0057] As used herein, the term “count” recited in step (f) of the method of the first aspect refers to the following process: Barcode sequences are extracted and clustered in two steps: 1. Initial grouping by exact match of the combination of amplicon_name + barcode sequence and 2. Cluster Reassignment, in each group of same amplicon_name, barcodes were further reassigned using global pairwise alignment with maximum 2 base differences between barcodes. Barcode clusters with number of associated reads less than 3 (after cluster reassignment) were considered unreliable clusters and removed from downstream analysis.

[0058] Next, the methylated DNA pattern of the DNA is reconstructed as disclosed in step (g) of the method of the first aspect. The consensus DNA sequence is compared to a reference genome using a sequence alignment tool and variant analysis of the DNA sequence is conducted by comparing the consensus reads to the reference genome to detect the variations. As used herein, the term “reference genome” refers to DNA sequences known in the art that may be obtainable from public databases. Exemplary Bioinformatics analysis method for reconstructing the methylated DNA pattern include bwa-meth, Bismark, MethylDackel, bisulfite-treated reads analysis tools (BRAT), methyQA, mrsFAST, BSMAP, VerJInxer, RMAP-bs, MethylCoder, BS-seeker2, and Bison.

[0059] Steps (a) to (g) of the method of the first aspect as described above thereby enable assessment of: 1) the conversion efficiency of non-CpG cytosines to thymine as quality control, and 2) the frequency of methylation at each CpG cytosine across all consensus sequencing reads with distinct barcode sequences (i.e. reads that are derived from unique DNA molecules corresponding to a specific amplicon).

[0060] In one example, the methods as disclosed herein may be used to detect mutations or polymorphisms at CpG sites. As used herein, relative to the reference genome, methylation of a CpG site is defined as concordance of the CG sequence for a CpG site, regardless of whether the site is on the plus or minus strand. As used herein, relative to the reference genome, non-methylation of a CpG site is defined as:

- A sequence of TG for a CpG on the plus strand. In this case, the unmethylated cytosine has been converted to thymine; or

- A sequence of CA for a CpG on the minus strand. In this case, the unmethylated cytosine on the minus strand was converted to thymine, which has a complementary adenine base on the plus strand (see **Fig. 4**).

[0061] During the reconstruction of the methylated DNA pattern of cfDNA, variations at CpG cytosines to a non-C/T base (i.e. mutation to A or G), will be flagged due to the unexpected occurrence of a non-C/T base that disrupts the CpG site. The allele frequency of this variation can be determined by its frequency across all consensus sequencing reads with distinct barcode sequences.

[0062] Variations at CpG guanines will also be flagged during this process due to the unexpected occurrence of a non-G base that disrupts the CpG site. The allele frequency of this variation can be determined by its frequency across all consensus sequencing reads with distinct barcode sequences.

[0063] In one example, the method as described in the first aspect further comprises the following steps:

(h) using a statistical modelling technique to thereby predict presence or absence of cancer; and

(i) using a statistical modelling technique to thereby identify specific cancer types when the presence of cancer is predicted in step (h). In one example, the method as described in the first aspect further comprises the step of analyzing methylated DNA pattern prior to performing step (h). In one example, Natural Language Processing, N-gram and Skip-gram are used for analyzing methylated DNA pattern. In one example, N-gram may be used to capture methylation pattern-specific information and generate new features that can be further analyzed. In one example, the generated new features can be used as data input for further statistical modelling techniques, such as those in step (h) and/or (i). In one example, the statistical modelling technique is logistic regression. In one example, Skip-gram may be used to determine patterns between initially unrelated or non-adjacent CpG sites by skipping N number of sites between 2 sites within an amplicon. In one example, the determined patterns can be used as data input for further statistical modelling techniques, such as those in step (h) and/or (i). In one example, the statistical modelling technique is logistic regression. In one example, the utilities of methylation frequency and methylation patterns derived from N-gram and Skip-gram may be used to detect cancer. In one example, the cancer is lung cancer.

[0064] In another example, the statistical modelling technique is selected from the group consisting of logistic regression, tree based classifiers and deep neural networks.

[0065] In a second aspect, the present disclosure refers to a kit for detecting methylated DNA pattern in DNA in a biological sample according to the method of the first aspect, comprising:

- (a) a first enzyme capable of oxidizing 5-methylcytosine and 5-hydroxymethylcytosine of the DNA;
- (b) a second enzyme capable of converting un-methylated cytosine of the DNA to uracil by deamination;
- 10 (c) a plurality of forward and reverse primer pairs specific to the converted sequence of target regions, wherein each forward primer of the plurality of forward and reverse primer pairs comprises a barcode sequence on its 5'end, wherein the barcode sequence of each forward primer is different from each other, wherein the target regions comprise one or more CpG sites;
- 15 (d) a plurality of universal indexed primers for creating the sequencing library;
- (e) a first DNA polymerase capable of amplifying DNA with uracil bases, for amplification of converted DNA;
- (f) a reagent capable of removing excess primers;
- (g) a second DNA polymerase capable of amplifying DNA, for creating the sequencing
- 20 library; and
- (h) sodium bisulfite.

[0066] The first enzyme, the second enzyme, the plurality of forward and reverse primer pairs, the barcode sequence, the CpG sites, and the plurality of universal indexed primers are disclosed herein.

- 25 [0067] In one example, the first DNA polymerase is selected from the group consisting of Phusion U Hot Start DNA Polymerase (Thermo Scientific), ZymoTaq DNA Polymerase (Zymo Research) and Q5U Hot Start High-Fidelity DNA Polymerase (NEB). In another example, the reagent capable of removing excess primers is selected from the group consisting of paramagnetic beads and single-strand exonucleases. Exemplary paramagnetic
- 30 beads include AMPure XP beads, SPRI beads, and Dynabeads. In another example, the second DNA polymerase is selected from the group consisting of KAPA HiFi DNA

Polymerase (Roche), Platinum Taq DNA Polymerase or Platinum SuperFi DNA Polymerase (Invitrogen) and Q5 High-Fidelity DNA Polymerase (NEB).

[0068] As used in this application, the singular form “a,” “an,” and “the” include plural references unless the context clearly dictates otherwise. For example, the term “a primer”
5 includes a plurality of primers, including mixtures and combinations thereof.

[0069] As used herein, the terms “increase” and “decrease” refer to the relative alteration of a chosen trait or characteristic in a subset of a population in comparison to the same trait or characteristic as present in the whole population. An increase thus indicates a change on a positive scale, whereas a decrease indicates a change on a negative scale. The term
10 “change”, as used herein, also refers to the difference between a chosen trait or characteristic of an isolated population subset in comparison to the same trait or characteristic in the population as a whole. However, this term is without valuation of the difference seen.

[0070] As used herein, the term “about” in the context of concentration of a substance, size of a substance, length of time, or other stated values means +/- 5% of the stated value, or
15 +/- 4% of the stated value, or +/- 3% of the stated value, or +/- 2% of the stated value, or +/- 1% of the stated value, or +/- 0.5% of the stated value.

[0071] Throughout this disclosure, certain embodiments may be disclosed in a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an inflexible limitation on the scope
20 of the disclosed ranges. Accordingly, the description of a range should be considered to have specifically disclosed all the possible sub-ranges as well as individual numerical values within that range. For example, description of a range such as from 1 to 6 should be considered to have specifically disclosed sub-ranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6 etc., as well as individual numbers within that
25 range, for example, 1, 2, 3, 4, 5, and 6. This applies regardless of the breadth of the range.

[0072] The invention illustratively described herein may suitably be practiced in the absence of any element or elements, limitation or limitations, not specifically disclosed herein. Thus, for example, the terms "comprising", "including", "containing", etc. shall be read expansively and without limitation. Additionally, the terms and expressions employed
30 herein have been used as terms of description and not of limitation, and there is no intention in the use of such terms and expressions of excluding any equivalents of the features shown and described or portions thereof, but it is recognized that various modifications are possible

within the scope of the invention claimed. Thus, it should be understood that although the present invention has been specifically disclosed by preferred embodiments and optional features, modification and variation of the inventions embodied herein disclosed may be resorted to by those skilled in the art, and that such modifications and variations are considered to be within the scope of this invention.

[0073] The invention has been described broadly and generically herein. Each of the narrower species and subgeneric groupings falling within the generic disclosure also form part of the invention. This includes the generic description of the invention with a proviso or negative limitation removing any subject matter from the genus, regardless of whether or not the excised material is specifically recited herein.

[0074] Other embodiments are within the following claims and non-limiting examples.

EXAMPLES

[0075] **Methods**

[0076] ***Sample collection and Processing***

[0077] Blood from healthy individuals or patients with cancer was collected into Streck Cell-Free DNA tubes and plasma was isolated by centrifugation. Plasma cell-free DNA (cfDNA) was extracted using the QIAamp Circulating Nucleic Acid Kit (Qiagen). To convert all un-methylated cytosines in the genome to uracils while preserving methylated cytosines, the plasma cfDNA was subjected to enzymatic conversion using the NEBNext Enzymatic Methyl-Seq Conversion Module (New England BioLabs). Briefly, DNA was treated with the TET2 enzyme that oxidizes 5-methylcytosine and 5-hydroxymethylcytosine, protecting these bases from deamination by APOBEC in the next step. Next, the DNA was purified using AMPure XP beads (Beckman Coulter) prior to the addition of APOBEC enzyme which deaminates un-methylated cytosines to uracils. Lastly, purification using AMPure XP beads generated single-stranded DNA that is similar to that of sodium-bisulfite-converted DNA.

[0078] ***Design of Multiplex PCR panel for identification of DNA methylation in targeted regions***

[0079] A multiplex amplicon-based next generation sequencing (NGS) platform was developed to capture and sequence targeted regions of the converted genome. These regions were selected based on literature review of known methylated regions in specific cancers and

from analyses of methylation data from normal and tumor tissues in the Cancer Genome Atlas (TCGA) database. Each amplicon covers at least 1 CpG site. In initial validation experiments, primers for 22 amplicons were designed and the panel has since been increased to >100 amplicons (159 amplicons). The design of the assay is intended to be scalable to include multiple targets for the specific identification of multiple cancers.

[0080] Each forward primer additionally includes on the 5' end, a random 10 nucleotide sequence to serve as barcode sequence for the identification of unique DNA molecules. In CpG-rich regions in which it was not possible to design primers in between CpG sites, degeneracy was incorporated for the primer designs to enable the capture of both unmethylated and methylated CpGs.

[0081] A combinatorial amplicon-based NGS based assay targeting hotspot mutations in 32 genes that are commonly mutated in lung cancer was developed to complement the multiplex amplicon-based NGS platform described above, to improve the sensitivity of cancer detection. Said combinatorial amplicon panel incorporates molecular barcode sequences for error suppression and improved coverage, enabling 100% specificity and 100% detection sensitivity at 1% and 5% VAF for single nucleotide variants (SNVs) and insertions/deletions (indels) and 89% detection sensitivity at 0.1% VAF using HD780 (Horizon Discovery) reference standards. The design of the panel incorporates tiled amplicons that can generate longer or shorter amplicons, thus enabling the profiling of the size distribution of cfDNA fragments. In one example, the combinatorial amplicon panel can detect cfDNA methylation. In one example, the combinatorial amplicon panel can detect cfDNA concentration. In one example, the combinatorial amplicon panel can detect cfDNA fragmentation profile. In one example, the combinatorial amplicon panel can detect cfDNA methylation, cfDNA concentration and cfDNA fragmentation profile or any combinations thereof.

[0082] ***Preparation of whole-genome sequencing library***

[0083] The amount of converted cfDNA used for library preparation varied slightly depending on the amount used for enzymatic conversion, but typically represented 5-10 ng starting amount of cfDNA prior to conversion. For the target capture PCR, both forward and reverse primers were combined in a single reaction using Phusion U Hot Start DNA Polymerase (Thermo Fisher Scientific) under the following thermocycling conditions: Denaturation at 98°C for 30s, followed by 3 or 4 cycles of 98°C for 10s, 55-57°C for 6 min,

and 72⁰C for 5 min (3 cycles with 55⁰C for the 22-amplicon panel, 4 cycles with 56⁰C or 57⁰C for larger panels). At the end of the reaction, for the 22-amplicon panel, excess primers were removed by purification with 1.5x AMPure XP beads twice. For larger panels, excess primers were removed by purification with 1.2x AMPure XP beads, treatment with
5 Thermolabile Exonuclease I (New England BioLabs) for 10 min, and a second round of purification with 1.3x AMPure XP beads.

[0084] A final amplification was performed to amplify the targets and to complete the library with indexed sequencing adaptors for sequencing on the Illumina platform. Briefly, purified product was amplified with indexed P5 adapter sequence and indexed P7 adapter
10 sequence using KAPA HiFi HotStart ReadyMix (Roche) under the following thermocycling conditions: Denaturation at 98⁰C for 45 s, followed by 19 to 21 cycles of 98⁰C for 15 s, 60⁰C for 30 s, and 72⁰C for 30 s, with a final extension at 72⁰C for 1 min. The amplified library was purified twice with 0.8x then 0.7x AMPure XP beads to remove non-specific products. The quality and quantity of the sequencing library was assessed using the 4200 TapeStation
15 system (Agilent Technologies, USA) and KAPA Library Quantification Kit for Illumina® Platforms (Roche) respectively. Paired-end sequencing (2x151bp) of the final dual-indexed libraries was performed on the Illumina platform as per manufacturer's instructions.

[0085] ***Data Analysis***

[0086] FASTQ files were processed using a custom pipeline. First, expected amplicons
20 were identified and labelled in the FASTQ files based on the expected primer sequences in Read 1 and paired Read 2. For amplicons with degenerate primers, data formed from each pair of degenerate primers are aggregated and assigned to the same amplicon based on the expected primer sequences. Primer sequences and upstream barcode sequences were trimmed using cutadapt, primer trimmed sequences were mapped to the Homo sapiens
25 GRCh37 (hg19) reference genome using bwa-meth, which is specifically designed for the alignment of bisulfite-converted sequences. For "primer" trimmed fastq files, the name of the primer which has the best match to a read is concatenated to the name of the mapped output reads (for both read 1 and read 2). The primer name assigned to read 1 may not always match that of read 2, which can be due to non-specific binding. An "amplicon_name"
30 is assigned to each paired read by combining the matching primer name of read 1 and read 2 (concatenated by semicolon).

[0087] Molecular tag (or barcode) sequences were included in the trimmed “primer” sequences of read 1, and can be extracted given the unique structure of primer sequences in read 1. The extracted molecular tag sequences were clustered in two steps: 1. Initial grouping by exact match of the combination of amplicon_name + barcode sequence and 2. Cluster Reassignment, in each group of same amplicon_name, barcodes are further reassigned using global pairwise alignment with maximum 2 base differences between barcodes. Barcode clusters with number of associated reads less than 3 (after cluster reassignment) were considered unreliable clusters and removed from downstream analysis.

[0088] Consensus calling was done for each molecular tag (or barcode) cluster, by first performing global alignment among all associated reads using MAFFT. The consensus base in each aligned position is called by determining the majority representative base type, the percentage of which was no less than an automatically determined threshold. The threshold is a function of the total number of reads for that barcode sequence. If no representative base can be called, the position is assigned N (as opposed to one of A, C, T, G). A new quality score was assigned to each position, which is either 90th percentile of all the quality values from the representative base type in that position (if a consensus base is found), or 10th percentile of all quality values in that position (if no consensus bases is found). The consensus reads were then written to a new FASTQ file. With molecular barcoding, the sequencing is error-free and increases confidence of methylated/non-methylated calls due to the high quality of sequencing data.

[0089] ***Analysis of conversion efficiency and methylation frequency***

[0090] Adaptor-trimmed, barcode-clustered consensus FASTQ reads were mapped to the Homo sapiens GRCh37 (hg19) reference genome using bwa-meth. The reads were subjected to several filtering steps prior to the evaluation of conversion efficiency (non-CpG Cs) and methylation frequency (CpG Cs). First, each read was only considered if at least two-thirds (66%) of its CpG cytosines are properly covered and assigned to a base (A, C, T or G) instead of N. Reads with more than one-third of its CpG cytosines assigned as N were excluded. Subsequently, data from all the reads were aggregated at the amplicon level and cytosines that meet any of the following criteria are excluded:

- >40% N fraction at an expected cytosine position. This filters out positions with low quality sequencing.

- <80% C or T base fraction at position. This filters out potential SNPs or positions with low quality.
- <60% G base fraction of the adjacent G base of an expected CpG site. This allows for the identification of putative SNPs at the G coordinate of a CpG site that would disrupt the site.
- >40% G base fraction of the 3' adjacent base of a non-CpG cytosine. This allows for the identification of putative SNPs that result in the formation of an unexpected CpG.

[0091] Conversion efficiency is defined as the average conversion fraction of non-CpG cytosines to thymines. Samples with amplicons with conversion efficiency <0.97 were repeated.

[0092] A methylation fraction was calculated at each CpG position and mean methylation fraction of an amplicon is defined as the average methylation fraction of all the considered CpG cytosines.

[0093] In addition to evaluating the mean methylation frequency of an amplicon, the methylation pattern in DNA sequences can also contain information of their source. To supplement the methylation frequency data, cancer-specific methylation patterns were evaluated via alternative approaches namely N-gram and Skip-gram.

[0094] The N-gram approach, which is similar to Natural Language Processing technique, was adopted to capture pattern-specific information and create new features that can be further analyzed (**Fig. 9**). A N-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a $(n - 1)$ order Markov model. N-gram features such as bigram, trigram, quad-gram and pentagram combinations were constructed to capture methylation patterns in adjacent 2, 3, 4 or 5-CpG sites, respectively. As amplicons that cover more CpG sites would have higher numbers of N-gram combinations, the N-gram for each amplicon was normalized by taking the average of all the reads and then divided by the maximum number of N-grams that can be formed for the particular amplicon. Of all possible N-gram combinations derived herein, grid search was performed to reduce the number of features that were then used for the training of a logistic regression model for cancer/non-cancer prediction.

[0095] Skip-gram, another approach used in Natural Language Processing, was adopted to find patterns between initially unrelated or non-adjacent CpG sites by skipping N number of sites between 2 sites within an amplicon (**Fig. 10**). Similar to the N-gram approach, the

Skip-gram for each amplicon was normalized to account for different numbers of CpG sites in different amplicons.

[0096] ***Logistic regression model for cancer prediction***

[0097] To build a training set for a logistic regression model for cancer prediction, >200 samples from healthy individuals and cancer patients were processed and analyzed. Using the data from these samples, methylation across individual CpG sites within an amplicon was examined for concordance.

[0098] Concordance of CpG methylation across CpG sites for an amplicon was computed using the Pearson Correlation to identify highly correlated features. Absolute values of Pearson Correlation Coefficient (PCC) was calculated for methylation/non-methylation at each pair of CpG positions and a PCC threshold of >0.8 was used for filtering out highly concordant features. This was to ensure that there was no multicollinearity among the amplicons (independent variables) when building the Logistic Regression Model.

[0099] For amplicons with concordance of >0.8 for CpG methylation, the average percent methylation across all the sites was considered as one feature. For amplicons with poor CpG methylation concordance (<0.8), the methylation frequency of each CpG within the amplicon was considered as a separate feature.

[00100] Methylation frequency for each feature was log transformed using the formula $\text{np.log}(x+0.00001)$ where x is the methylation frequency also known as the methylation beta-value. Highly correlated features were removed from the model by calculating the variance inflation factor (VIF) score for each feature to detect multicollinearity. Features with the highest VIF scores were then dropped iteratively until a maximum VIF score of <10.

[00101] Recursive feature elimination (RFE) was performed on the remaining features to determine the set of features for the best performance of the model. If a sample was healthy, its corresponding target array value was set to 0, while if a sample was cancerous, its corresponding target array value was set to 1. For model-building set, the scikit-learn (<http://scikit-learn.org>) package's LogisticRegression module was used to machine-learn parameters for an LR classifier using the log-transformed methylation signatures as features and the target array as the target values. The liblinear solver implemented in scikit-learn (<http://www.csie.ntu.edu.tw/~cjlin/liblinear>) was utilized for this process. In order to avoid overfitting and build a robust model, a cross-validation approach with RIDGE penalty (L2) was utilized. 3-fold cross validation was performed for each iteration of the model and a

probability threshold of 0.5 was used to assign samples as normal (<0.5) or cancer (>0.5). Sensitivity and specificity values were calculated for each fold and finally overall sensitivity and specificity was reported by taking average of the fold scores.

[00102] The utilities of methylation frequency and methylation patterns by N-gram and Skip-gram for the detection of cancer were evaluated by training individual logistic regression models using a dataset derived from 60 healthy individuals and 39 early (stage I-III) and 56 late stage lung cancer patients, respectively. 3-fold cross validation, each with 50 repeats, of the logistic regression models with a threshold set at 95% specificity demonstrated 33.96-54.72% and 81.61-86.21% sensitivity of detection of early and late stage lung cancers, respectively (**Fig. 11**).

[00103] Plasma cfDNA methylation, cfDNA levels, cfDNA fragmentation profiles and ctDNA detection can each provide complementary information for enhanced accuracy in cancer detection. The combinatorial amplicon-based panel approach described herein combining the detection of cfDNA methylation, cfDNA concentration and ctDNA fragmentation profile can mitigate the limitations of individual approaches and improve the overall accuracy of cancer detection. Thus, a machine learning classifier prediction model that integrates these multiple classes of data generated from plasma cfDNA was used.

[00104] To establish a prediction classifier model of normal vs lung cancer status, individual logistic regression models of mAF, N-gram, Skip-gram or aggregate cfDNA 'Biomarker' features were first trained using the dataset of 60 healthy individuals and 39 early (stage I-III) and 56 late-stage lung cancer patients. Aggregate cfDNA Biomarker features encompass plasma cfDNA concentration, fragment size ratio and the ctDNA detection score, each of which are log normalized. The ctDNA detection score was determined by first classifying each variant to one of six classes based on evidence in public databases for the prevalence and pathogenesis of the variant in cancer. Each of the classes were assigned a score, with the highest score assigned to the highest class, and the ctDNA detection score of each sample was calculated by aggregating the score multiplied by allele frequency of each variant detected. The ctDNA detection score, plasma cfDNA concentration and cfDNA fragment size were incorporated into a single 'Biomarkers' logistic regression model. A Stacking Ensemble technique was adopted to merge the 3 (mAF + Biomarkers + N-gram or Skip-gram) models and generate a final prediction probability value for cancer.

[00105] At a specificity of 95%, 3-fold cross-validation analysis using an Ensemble mAF + Biomarkers + N-gram model yielded an average sensitivity of 79.49% and 91.07% for early- and late-stage lung cancer, respectively, with an overall sensitivity of 86.32% (**Fig. 12**). Considering both early and late-stage detection sensitivities, the combinatorial approach provided an additional diagnostic value of 24.8-45.5% for early-stage and 4.9-9.5% for late-stage lung cancer compared with individual models alone, supporting the clinical utility of the combinatorial approach.

[00106] Measurement of protein tumor marker levels are also commonly used in cancer screening and detection. Its utility is demonstrated in the combinatorial multi-omic approach by the assessment of plasma CEA levels. Plasma CEA levels, detected by the Beckman Access II immunoanalyzer, were higher in lung cancer samples compared to normal controls, giving a sensitivity of 46.15% and 73.21% for early and late-stage lung cancer detection, respectively, at a specificity of 95% (**Fig. 12**). When combined with a mAF, N-gram and Biomarkers Ensemble prediction model, the addition of CEA provided an additional diagnostic sensitivity of 5.2% and 3.6% in the detection of early and late-stage lung cancer, respectively.

[00107] *Random forest model for determination of cancer type*

[00108] For samples predicted to be cancer by the logistic regression model described above, a random forest classification algorithm was trained for identification of the specific type of cancer using data from several types of cancer samples, including breast, colorectal, lung and ovarian cancers.

[00109] Feature selection was done using ANOVA F-Test via the `f_classif()` function from scikit-learn (https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html). F_Scores for all methylation sites across 4 different cancer categories were computed and ranked.

[00110] Random Forest, as implemented in the scikit-learn (<http://scikit-learn.org>) package's `RandomForestClassifier` module was used, using the methylation signatures as features and cancer type as the target label. The default setting of the `RandomForestClassifier` were used. For robustness, five rounds of 3-fold CV were performed for each iteration of the model. The performance of the Random Forest Classifier seemed to plateau at around 23 features and these were selected the final features for the model. For prediction, probability scores were calculated for each cancer type and the cancer

type with the highest probability score was predicted as likely cancer type for that particular sample.

[00111] Finally, individual sensitivities and specificities for each cancer type across all 5 iterations of the models were combined and reported.

5 [00112] All analysis and modelling for both the modelling parts was conducted in Python Programming Language, version 3.7.3.

[00113] **Results**

[00114] The present disclosure describes the methodology for the identification of methylated DNA for the detection of early stage cancer, minimal residual disease following
10 cancer surgery or therapy, and cancer relapse, with high sensitivity and specificity, especially in situations that these disease are undetectable by conventional screening methods. In one example, blood-based test is used for the identification of methylated signatures in plasma cell-free DNA (cfDNA) that can indicate the presence of cancer and specify its tissue of origin (i.e. cancer type) before the development of overt symptoms. To
15 identify sites of DNA methylation, the present disclosure uses enzymatic conversion as an alternative to conventional sodium bisulfite treatment to convert un-methylated cytosines to uracils. First, cfDNA was treated with TET2 enzyme that oxidizes 5-methylcytosine and 5-hydroxymethylcytosine, protecting these bases from deamination by APOBEC in the next step. Next, the cfDNA was purified using AMPure XP beads prior to the addition of
20 APOBEC enzyme which deaminates un-methylated cytosines to uracils. Lastly, purification using AMPure XP beads generated single-stranded DNA that is similar to that of sodium-bisulfite-converted cfDNA, but typically obtained in higher recovery yields and with little fragmentation compared to bisulfite-converted DNA. As little as 5 ng starting amount of cfDNA has been successfully put through conversion, library preparation and sequencing in
25 the present workflow.

[00115] In the target capture and library amplification step, the converted cfDNA molecules were selectively enriched using a multiplicity of primers specific to the converted sequence of target regions in a single PCR reaction. The converted cfDNA was added to a PCR reaction containing more than 5 'forward' and 'reverse' primer pairs and subject to 2,
30 3, 4 or 5 cycles of PCR in a first limited amplification reaction. As each forward primer carries on its 5' end a randomly assigned barcode sequence, this PCR allows individual cfDNA molecules to be tagged uniquely in this first step of sequencing library formation.

Subsequently, the reactions were purified to remove excess primers. A final PCR amplification with universal indexed primers was done to create libraries with components required for multiplex sequencing on a next-generation sequencing platform such as Illumina.

- 5 [00116] Each 'forward' and 'reverse' primer pair forms an amplicon that covers at least 1 CpG site. In CpG-rich regions in which it was not possible to design primers not overlapping CpG sites, the primer designs incorporate degeneracy for the capture of both un-methylated and methylated CpGs and thus overcome methylation-related drop-off of coverage and capture.
- 10 [00117] Following sequencing, the presence of a barcode sequence was detected using specialized Bioinformatics methods to count and assign each DNA sequence from high-throughput sequencing to an original parental DNA molecule, carrying the same tag. In the method as disclosed herein, the parental DNA molecule is the original cfDNA molecule right after enzymatic conversion.
- 15 [00118] The cfDNA methylation pattern of the biological sample was then reconstructed. The number of unique cfDNA molecules corresponding to targeted regions of the genome were enumerated. The specific DNA methylation pattern of each molecule was reconstructed by comparing to a reference genome using a sequence alignment tool (for example, bwa-meth) designed for the alignment of bisulfite-converted sequences. Variations of the
- 20 samples' genome sequence compared to this reference genome were detected by variant analysis. This allows for the assessment of 1) the conversion efficiency of non-CpG cytosines to thymine as quality control and 2) the frequency of methylation at each CpG cytosine across all consensus sequencing reads with distinct barcode sequences (i.e. reads that are derived from unique cfDNA molecules corresponding to a specific amplicon). The
- 25 methylation information was therefore obtained and because of the incorporation of barcode sequences at the PCR step, low-level errors of sequencing were suppressed (<1%) which allowed for accurate determination of the methylation status at each CpG site.
- [00119] For training of a statistical model for cancer vs non-cancer prediction, methylation across individual CpG sites within an amplicon was examined for pairwise concordance
- 30 using a large number of samples. Concordance of CpG methylation across CpG sites for an amplicon was computed using the Pearson Correlation to identify highly correlated features. Absolute values of Pearson Correlation Coefficient (PCC) was calculated for

methylation/non-methylation at each pair of CpG positions and a PCC threshold of >0.8 was used to filter out highly concordant features. This was to ensure that there was no multicollinearity among the amplicons (independent variables) when building the Logistic Regression Model.

5 [00120] For amplicons with concordance of >0.8 for CpG methylation, the average percent methylation across all the sites was considered as one feature. For amplicons with poor CpG methylation concordance (<0.8), the methylation frequency of each CpG within the amplicon was considered as an individual feature. Methylation data obtained from 209 plasma cfDNA samples (57 normal, 152 cancer) were then used as a “training set” for a
10 logistic regression model to calculate probabilities and AUC/ROC curves for cancer prediction. 3-fold cross validation of this model reported 94.1% sensitivity and 87.7% specificity for cancer detection (**Fig. 9**). Data from specific cancer types was also used for a second model in which a random forest classification algorithm is used to predict the tissue of origin (or cancer type) in samples that cancer was detected in the first step (**Fig. 9**).

15 [00121] **Discussion**

[00122] The method of the present disclosure has the following advantages:

1. The target regions to be analysed were selected based on externally validated regions and from genome-wide analyses of methylation data in the TCGA database. Even when using a relatively small panel of 22 targets, the method has a high sensitivity
20 ($>90\%$) and specificity (90%) for the detection of cancer. This panel has been expanded to include more targets (159) and can be further expanded. Increased target number greatly improved the sensitivity and specificity of the test. The combination of target regions, and their associated CpG sites that are covered by each primer pair, renders the present method novel.
- 25 2. The method of the present disclosure may be used on a blood-based test (for example, to detect methylated DNA pattern in cfDNA in the blood) that is fast and non-invasive (only one draw of blood is needed). In addition, the method is scalable for the detection of multiple cancers in a single test and is suitable for cancer screening in an asymptomatic population.
- 30 3. Unlike somatic mutations that can occur anywhere along the length of a gene and are thus difficult to profile comprehensively, DNA methylation in cancer occurs predominantly in CpG islands within gene promoter regions and are thus more

accessible to comprehensive profiling. In addition, DNA methylation typically occurs in a tissue-specific manner which increases the specificity of identifying the tissue of origin of the cancer. The frequency of methylation can be calculated which gives an indication of tumor load and can be used for disease monitoring.

- 5 4. Enzymatic conversion of un-methylated cytosines to uracils enables high efficiency conversion with little fragmentation and loss in DNA yield.
5. Primers with barcode sequences in a multiplex amplicon-capture assay allow the suppression of low-level errors due to sequencing and improve sensitivity of identification of methylated sites with high confidence.
- 10 6. Degenerate primers are used to capture both methylated and un-methylated strands in regions that are CpG rich and would otherwise be inaccessible with regular primers for bisulfite sequencing.
7. The initial multiplex PCR reaction is scalable and allows the capture of multiple genomic regions for the identification of several cancer types in a single assay.
- 15 8. Use of dual index combinations reduces the possibility of index swapping during sequencing.
9. A statistical model trained using methylation sequencing data from hundreds of known normal and cancer cfDNA enables accurate detection of cancer in independent samples.
- 20 10. The technological significance lies in the generalizable use of primers for target capture, which allows working with smaller, limiting amounts of DNA, especially when enzymatic conversion is used instead of conventional sodium bisulfite treatment. In addition, the unique combination of targets is selected for the sensitivity and specific detection of multiple cancers.
- 25 11. The method of the present disclosure may be used in the following applications:
 - Identification of methylation signatures specific to cancers.
 - Identification of methylation signatures that are specific to particular cancers.
 - Cancer screening in healthy individuals and individuals at high risk for the tested
- 30 cancers. One of the intended uses of the method of the present disclosure is for cancer screening and early cancer detection. In a validation experiment, the method of the present disclosure showed 90.9% sensitivity for the detection of stage I colorectal cancer.

- Disease monitoring in cancer patients, including monitoring response to treatment and cancer relapse, and detecting minimal residual disease (MRD) following cancer surgery or therapy. The method of the present disclosure is suitable for regular disease monitoring as only a blood draw is required.

CLAIMS

1. A method of detecting methylated DNA pattern in DNA in a biological sample, comprising:
- 5 (a) converting un-methylated cytosine of the DNA to uracil by deamination to thereby generate converted DNA;
- (b) purifying the converted DNA from step (a);
- (c) tagging a barcode sequence on the converted DNA, by performing a first PCR amplification using a plurality of forward and reverse primer pairs specific to the converted
- 10 sequence of target regions, wherein each forward primer of the plurality of forward and reverse primer pairs comprises a barcode sequence on its 5' end, wherein the barcode sequence of each forward primer is different from each other, wherein the target regions comprise one or more CpG sites;
- (d) subjecting the tagged converted DNA from step (c) to a second PCR amplification with
- 15 universal indexed primers to thereby create a sequencing library with components required for multiplex sequencing;
- (e) subjecting the sequencing library to multiplex sequencing on a next-generation sequencing platform;
- (f) detecting the presence of a barcode sequence using Bioinformatics methods to count and
- 20 assign each DNA sequence from the next-generation sequencing to an original parental DNA molecule carrying the same barcode sequence, comprising:
- (i) performing cluster reassignment of sequencing reads with the same barcode sequence to thereby generate barcode clusters wherein each barcode cluster contains reads from the same amplicon and with the same barcode sequence; and
- 25 (ii) performing consensus calling for each barcode cluster to thereby obtain consensus reads;
- (g) reconstructing the methylated DNA pattern of the DNA by
- (I) comparing the DNA sequence to a reference genome using a sequence alignment tool; and
- 30 (II) conducting variant analysis of the DNA sequence by comparing the consensus reads to the reference genome to detect the variations;

to thereby assess 1) the conversion efficiency of non-CpG cytosines to thymine as quality control, and 2) the frequency of methylation at each CpG cytosine across all consensus sequencing reads with distinct barcode sequences (i.e. reads that are derived from unique DNA molecules corresponding to a specific amplicon).

5

2. The method of claim 1, further comprising the following steps before step (a):

- 1) extracting the DNA from the biological sample;
- 2) treating the DNA using a first enzyme that oxidizes 5-methylcytosine and 5-hydroxymethylcytosine of the DNA to thereby protect the 5-methylcytosine and 5-hydroxymethylcytosine from deamination in step (a); and
- 3) purifying the DNA from step (2).

10

3. The method of claim 2, wherein in step (a), the un-methylated cytosine of the DNA is converted to uracil by deamination using a second enzyme to thereby generate converted DNA.

15

4. The method of claim 1, wherein the DNA is unextracted from the biological sample, and wherein in step (a), the un-methylated cytosine of the DNA is converted to uracil by deamination using bisulfite to thereby generate converted DNA.

20

5. The method of any one of the preceding claims, further comprising:

- (h) performing a statistical modelling technique to thereby predict presence or absence of cancer; and
- (i) performing a statistical modelling technique to thereby identify specific cancer types when the presence of cancer is predicted in step (h).

25

6. The method of claim 5, wherein the statistical modelling technique is selected from the group consisting of logistic regression, tree based classifiers and deep neural networks.

30

7. The method of any one of the preceding claims, further comprising the following step prior to step (h):

analyzing methylated DNA pattern by capturing methylation pattern-specific information and generating new features and/or determining patterns within an amplicon as data input for statistical modelling techniques of step (h) and/or step (i).

5 8. The method of any one of the preceding claims, wherein the biological sample is selected from the group consisting of a liquid sample, a tissue sample, or a cell sample.

9. The method of claim 1, wherein the DNA is selected from the group consisting of cell-free DNA (cfDNA) and DNA encapsulated within tissues and/or cells.

10

10. The method of claim 2, wherein the first enzyme is a Ten-eleven translocation (TET) enzyme or an isoform thereof; and wherein optionally, the TET enzyme is selected from the group consisting of TET1 enzyme or an isoform thereof, TET2 enzyme or an isoform thereof, and TET3 enzyme or an isoform thereof.

15

11. The method of claim 3, wherein the second enzyme is a cytidine deaminase or an enzyme with cytidine deaminase properties;

wherein optionally, the cytidine deaminase is selected from the group consisting of APOBEC enzyme, CDA, and activation-induced cytidine deaminase; and

20

wherein optionally, the enzyme with cytidine deaminase properties is selected from the group consisting of M. SssI and M.HpaII.

12. The method of any one of the preceding claims, wherein the amount of DNA is at least 5 ng.

25

13. The method of any one of the preceding claims, wherein the barcode sequence is an oligonucleotide comprising 10 to 16 random nucleotides, or 10 to 15 random nucleotides, or 10 to 13 random nucleotides, or 10 random nucleotides, or 11 random nucleotides, or 12 random nucleotides, or 13 random nucleotides, or 14 random nucleotides, or 15 random nucleotides, or 16 random nucleotides; and

30

wherein optionally, the barcode sequence is an oligonucleotide having 10 random nucleotides.

14. The method of any one of the preceding claims, wherein the number of the forward and reverse primer pairs is at least 5; and wherein optionally, the number of the forward and reverse primer pairs is at least 159.
- 5 15. The method of any one of the preceding claims, wherein each forward and reverse primer pair covers a target region which comprises at least 1 CpG site.
16. The method of any one of the preceding claims, wherein the forward and reverse primer pairs comprise sequences as disclosed in Table 1.
- 10 17. The method of any one of the preceding claims, wherein the forward primer in the primer pair comprises one or more degenerate bases, and/or the reverse primer in the primer pair comprises one or more degenerate bases; wherein optionally, the degenerate base is selected from the group consisting of C, T, A and G.
- 15 18. The method of claim 17, wherein each primer of the primer pair comprises 1, 2, or 3 degenerate bases; and wherein optionally, each primer of the primer pair has one degenerate base.
- 20 19. The method of any one of claims 1-18, wherein the primer pair does not comprise a degenerate base.
20. The method of any one of the preceding claims, wherein the first PCR amplification comprises a number of PCR cycles selected from the group consisting of 2, 3, 4 and 5 PCR cycles.
- 25 21. The method of claim 1, wherein the universal indexed primers comprise:
a forward primer comprising the sequence of
AATGATACGGCGACCACCGAGATCTACACCTAGCGCTACACTCTTTCCCTACAC
30 GACGCTCTTCCGATC*T; and

a reverse primer comprising the sequence of

CAAGCAGAAAGACGGCATACGAGATAAACCGCGGGTGACTGGAGTTCAGACGTGT
GCTCTTCCGATC*T.

- 5 22. The method of claim 1, wherein the methylated DNA pattern is reconstructed using a Bioinformatics analysis method selected from the group consisting of bwa-meth, Bismark, MethylDackel, bisulfite-treated reads analysis tools (BRAT), methyQA, mrsFAST, BSMAP, VerJInxer, RMAP-bs, MethylCoder, BS-seeker2, and Bison.
- 10 23. A kit for detecting methylated DNA pattern in DNA in a biological sample according to the method of claim 1, comprising:
- (a) a first enzyme capable of oxidizing 5-methylcytosine and 5-hydroxymethylcytosine of the DNA;
- (b) a second enzyme capable of converting un-methylated cytosine of the DNA to uracil by
- 15 deamination;
- (c) a plurality of forward and reverse primer pairs specific to the converted sequence of target regions, wherein each forward primer of the plurality of forward and reverse primer pairs comprises a barcode sequence on its 5' end, wherein the barcode sequence of each forward primer is different from each other, wherein the target regions comprise one or more
- 20 CpG sites;
- (d) a plurality of universal indexed primers for creating the sequencing library;
- (e) a first DNA polymerase capable of amplifying DNA with uracil bases, for amplification of converted DNA;
- (f) a reagent capable of removing excess primers;
- 25 (g) a second DNA polymerase capable of amplifying DNA, for creating the sequencing library; and
- (h) sodium bisulfite.
24. The kit of claim 23, wherein the first enzyme is selected from the group consisting of a Ten-eleven translocation (TET) enzyme or an isoform thereof; and wherein optionally, the TET enzyme is selected from the group consisting of TET1 enzyme or an isoform thereof, TET2 enzyme or an isoform thereof, and TET3 enzyme or an isoform thereof.
- 30

25. The kit of claim 23, wherein the second enzyme is a cytidine deaminase or an enzyme with cytidine deaminase properties;
wherein optionally, the cytidine deaminase is selected from the group consisting of APOBEC enzyme, CDA, and activation-induced cytidine deaminase; and
5 wherein optionally, the enzyme with cytidine deaminase properties is selected from the group consisting of M. SssI and M.HpaII.
26. The kit of claim 23, wherein the reagent is selected from the group consisting of paramagnetic beads and single-strand exonucleases; wherein optionally the paramagnetic
10 beads are selected from the group consisting of AMPure XP beads, SPRI beads, and Dynabeads.
27. The kit of claim 23, wherein the barcode sequence is an oligonucleotide comprising
10 to 16 random nucleotides, or 10 to 15 random nucleotides, or 10 to 13 random
15 nucleotides, or 10 random nucleotides, or 11 random nucleotides, or 12 random nucleotides, or 13 random nucleotides, or 14 random nucleotides, or 15 random nucleotides, or 16 random nucleotides; and wherein optionally, the barcode sequence is an oligonucleotide having 10 random nucleotides.
- 20 28. The kit of claim 23, wherein the number of the forward and reverse primer pairs is at least 5; and wherein optionally, the number of the forward and reverse primer pairs is at least 159.
29. The kit of claim 23, wherein each forward and reverse primer pair covers at least 1
25 CpG site.
30. The kit of claim 23, wherein the forward and reverse primer pairs comprises sequences as disclosed in Table 1.
- 30 31. The kit of claim 23, wherein the forward primer in the primer pair comprises one or more degenerate bases, and/or the reverse primer in the primer pair comprises one or more

degenerate bases; and wherein optionally, the degenerate base is selected from the group consisting of C, T, A and G.

32. The kit of claim 31, wherein each primer of the primer pair comprises 1, 2, or 3 degenerate bases; and wherein optionally, each primer of the primer pair has one degenerate base.

33. The kit of claim 23, wherein the primer pair does not comprise a degenerate base.

34. The kit of claim 23, wherein the universal indexed primers comprise:
a forward primer comprising the sequence of
AATGATACGGCGACCACCGAGATCTACACCTAGCGCTACACTCTTCCCTACAC
GACGCTCTTCCGATC*T; and
a reverse primer comprising the sequence of
CAAGCAGAAGACGGCATACGAGATAACCGCGGGTGACTGGAGTTCAGACGTGT
GCTCTTCCGATC*T.