(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 March 2016 (17.03.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/040602 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/US2015/049385

(22) International Filing Date:
10 September 2015 (10.09.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/049,284 11 September 2014 (11.09.2014) US

(71) Applicant: **EPICENTRE TECHNOLOGIES CORPORATION** [US/US]; 5602 Research Park Blvd., Suite 200, Madison, Wisconsin 53719 (US).

(72) Inventors: **BURGESS, Joshua**; 5602 Research Park Blvd., Suite 200, Madison, Wisconsin 53719 (US). **VAIDYANATHAN, Ramesh**; 5602 Research Park Blvd., Suite 200, Madison, Wisconsin 53719 (US). **BRUINSMA, Stephen Paul**; 5602 Research Park Blvd., Suite 200, Madison, Wisconsin 52719 (US). **GRUNENWALD, Haiying Li**; 5602 Research Park Blvd., Suite 200, Madison, Wisconsin 53719 (US).

(74) Agent: **MOORE, Brent C.**; 5200 Illumina Way, San Diego, California 92122 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU,

[Continued on next page]

(54) Title: REDUCED REPRESENTATION BISULFITE SEQUENCING USING URACIL N-GLYCOSYLASE (UNG) AND ENDONUCLEASE IV

(57) Abstract: In a first aspect, embodiments disclosed herein provide methods for preparing a sample for sequencing, comprising: treating nucleic acid molecules in the sample to convert at least a portion of unmethylated cytosine residues into uracil residues; and cleaving the nucleic acid molecules at at least a portion of the uracil residues to obtain nucleic acid fragments. Further provided are populations of nucleic acid fragments resulting from a sample treated with the methods disclosed herein.

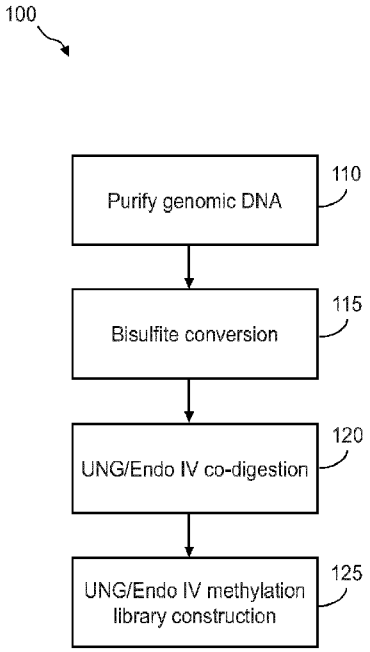

Figure 1

WO 2016/040602 A1

TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

— *with international search report (Art. 21(3))*

REDUCED REPRESENTATION BISULFITE SEQUENCING USING URACIL N-GLYCOSYLASE (UNG) AND ENDONUCLEASE IV

FIELD

[0001] Embodiments provided herein relate to methods and compositions for next generation sequencing. Some embodiments relate to methods of preparing a library for reduced representation bisulfite sequencing (RRBS), followed by subsequent sequencing of the library.

BACKGROUND

[0002] Reduced representation bisulfite sequencing (RRBS) is a sequencing-based methylation analysis technique that is used to analyze cytosine methylation patterns across regions of the genome. Current RRBS techniques combine restriction enzyme digestion, bisulfite conversion, and sequencing of the bisulfite-converted DNA in order to enrich and identify areas of the genome that have methylated CpG content.

[0003] There is a need for improved methods for constructing a genomic DNA library for RRBS from much lowered input DNA, with better data resolution and coverage.

SUMMARY OF EMBODIMENTS OF THE DISCLOSURE

[0004] In a first aspect, embodiments disclosed herein provide methods for preparing a sample for sequencing, comprising: treating nucleic acid molecules in the sample to convert at least a portion of unmethylated cytosine residues into uracil residues; and cleaving the nucleic acid molecules at at least a portion of the uracil residues to obtain nucleic acid fragments. Further provided are populations of nucleic acid fragments resulting from a sample treated with the methods disclosed herein.

[0005] In some embodiments, the sample is treated with bisulfite to convert at least a portion of unmethylated cytosine residues into uracil residues. In some embodiments, cleaving the nucleic acid molecules at the at least a portion of uracil residues comprises treating the nucleic acid molecules with a uracil DNA glycosylase resulting in a plurality of abasic residues in place of at least a portion of uracil residues. In some embodiments, the

methods disclosed herein comprise treating the nucleic acid molecules with an endonuclease after treating with said uracil DNA glycosylase, wherein said endonuclease cleaves said nucleic acid molecules at at least a portion of the plurality of abasic residues. In some embodiments, said uracil DNA glycosylase and said endonuclease are present in a single reaction mixture. In some embodiments, the methods disclosed herein comprise treating the nucleic acid molecules with heat after treating with said uracil DNA glycosylase, wherein said nucleic acid molecules are broken at at least a portion of the plurality of abasic residues. In some embodiments, the treatments with said uracil DNA glycosylase and heat are conducted simultaneously. In some embodiments, alternative endonucleases may be used to cleave the genome. In some embodiments, the treatment with said uracil DNA glycosylase is conducted for about 1 minute to about 240 minutes. In some embodiments, the treatment with said uracil DNA glycosylase is conducted for about 1 minute to about 10 minutes. In some embodiments, the treatment with said uracil DNA glycosylase is conducted for about 1 minute to about 5 minutes. In some embodiments, the treatment with heat is conducted at 70 °C. In some embodiments, said uracil DNA glycosylase is human uracil-DNA glycosylase (UNG). In some embodiments, said UNG is available from Epicentre (Cat#UG13100; Madison, WI). In some embodiments, said UNG is at a concentration of about 0.02 U/1 μg of DNA to about 1 U/1 μg of DNA. In some embodiments, said UNG is at a concentration of about 0.04 U/1 μg of DNA to about 0.2 U/1 μg of DNA. In some embodiments, said UNG is at a concentration of about 0.04 U/1 μg of DNA to about 0.1 U/1 μg of DNA. In some embodiments, said endonuclease is endonuclease IV (Endo IV). In some embodiments, said Endo IV is at a concentration of about 2 U/1 μg of DNA. In some embodiments, said Endo IV is at a concentration of about 0.2 U/1 μg of DNA. In some embodiments, said Endo IV is at a concentration of about 0.08 U/1 μg of DNA. In some embodiments, the methods disclosed herein comprise selecting the nucleic acid fragments based on their size. In some embodiments, said selecting the nucleic acid fragments based on their size comprises a bead-based method. In some embodiments, the portion of unmethylated cytosine residues is at least 50%, 60%, 70%, 80%, 90%, 95%, or 100%. In some embodiments, the portion of the uracil residues is at least 50%, 60%, 70%, 80%, 90%, 95%, or 100%. In some embodiments, the portion of the plurality of abasic residues is at least 50%, 60%, 70%, 80%, 90%, 95%, or 100%.

[0006] In a second aspect, embodiments disclosed herein provide methods for sequencing nucleic acid fragments, comprising: obtaining the nucleic acid fragments from a sample treated with a method disclosed herein; and sequencing the nucleic acid fragments.

[0007] In some embodiments, the methods disclosed herein comprise amplifying said nucleic acid fragments by PCR. In some embodiments, said PCR comprises about 10 cycles to about 15 cycles. In some embodiments, said sequencing is sequencing by synthesis (SBS). In some embodiments, the sequences of said nucleic acid fragments are compared to a reference sequence. In some embodiments, the methods disclosed herein comprise identifying methylated cytosines in said nucleic acid fragments. In some embodiments, the reference sequence comprises a methylome. In some embodiments, said methylated cytosines are identified using Bismark. In some embodiments, said methylated cytosines are analyzed using an integrated genome viewer. In some embodiments, said methylated cytosines comprise methylated CpG, methylated CHG and/or methylated CHH. In some embodiments, the methylation calls are significantly enriched in comparison to the methylation calls in a sample treated with bisulfite to convert unmethylated cytosine residues into uracil residues, but without cleaving the nucleic acid molecules at the uracil residues. In some embodiments, the enriched methylation calls comprise methylated CpG. In some embodiments, the enriched methylation calls comprise methylated CHG and/or methylated CHH. In some embodiments, the methylation calls comprise sites with a read depth of greater than about 10. In some embodiments, the cytosine sites are significantly enriched in comparison to the cytosine sites in a sample treated with bisulfite to convert unmethylated cytosine residues into uracil residues, but without cleaving the nucleic acid molecules at the uracil residues. In some embodiments, the cytosine sites comprise CpG. In some embodiments, the cytosine sites comprise CHG and/or CHH.

[0008] In a third aspect, embodiments disclosed herein provide nucleic acid samples comprising fragmented nucleic acid molecules wherein substantially all cytosine residues in said fragmented nucleic acid molecules are methylated.

[0009] In some embodiments, the first three residues of some of said fragmented nucleic acid molecules are not cytosine-guanine-guanine and/or the last four residues of some of said fragmented nucleic acid molecules are not thymine-cytosine-guanine-guanine. In some embodiments, the fragmented nucleic acid molecules comprise an

apurinic/apyrimidinic site at the 5' or 3' end. In some embodiments, the fragmented nucleic acid molecules have a size from about 10 bp to about 2,000 bp. In some embodiments, the fragmented nucleic acid molecules have a size from about 50 bp to about 500 bp. In some embodiments, the fragmented nucleic acid molecules have a size from about 100 bp to about 200 bp. In some embodiments, the fragmented nucleic acid molecules have a mean size of about 140 bp to about 170 bp. In some embodiments, a portion of the CpG sites in said fragmented nucleic acid molecules are methylated. In some embodiments, at least about 0.1% of the CHG sites in said fragmented nucleic acid molecules are methylated. In some embodiments, at least about 0.1% of the CHH sites in said fragmented nucleic acid molecules are methylated.

[0010] In a fourth aspect, embodiments disclosed herein provide kits comprising at least one container means, wherein the at least one container means comprises a reagent that cleaves a nucleic acid molecule at a uracil residue. In some embodiments, the reagent comprises UNG and endonuclease IV (Endo IV).

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] Figure 1 illustrates a flow diagram of an exemplary embodiment of a method for construction of a bisulfite-treated UNG/Endo IV methylation library for RRBS.

[0012] Figure 2 shows pictorially the steps of an exemplary embodiment of a method for construction of a bisulfite-treated UNG/Endo IV methylation library for RRBS.

[0013] Figure 3 shows a photograph of an agarose gel used to evaluate a UNG/Endo IV digestion time-course of bisulfite-treated genomic DNA in an exemplary embodiment.

[0014] Figure 4 shows a panel of BioAnalyzer traces of the fragment size distribution in RRBS libraries generated using Hela DNA in an exemplary embodiment.

[0015] Figure 5A, Figure 5B, and Figure 5C show a screenshot of the Genome Analyzer set-up parameters, a screenshot of the status pane with quality metrics, and a screenshot of the read summary, read 1, and read 2 metrics, respectively, for the analysis of RRBS libraries from Hela or Coriell gDNA in an exemplary embodiment.

[0016] Figure 6A and Figure 6B show a series of insert size plots in RRBS libraries from Coriell gDNA treated for different amounts of time and with different enzyme

dilutions, respectively, generated from the output sequencing data in an exemplary embodiment.

[0017] Figure 7 shows a data table of the Bismark methylation call results for the untreated and treated RRBS libraries from Hela or Coriell gDNA in an exemplary embodiment.

[0018] Figure 8 shows a screenshot of the IGV interface showing methylation patterns across a region of chromosome 8 in the untreated and 120-minute-treated libraries from Coriell gDNA in an exemplary embodiment.

[0019] Figure 9 shows a screenshot of the IGV interface showing methylation patterns across a region of chromosome 5 in RRBS libraries from Coriell gDNA treated with different enzyme dilutions in an exemplary embodiment.

[0020] Figure 10A and Figure 10B show a bar graph of Coriell CpG sites and a bar graph of Hela CpG sites in the RRBS libraries, respectively, in an exemplary embodiment.

[0021] Figure 11A and Figure 11B show bar graphs of total number of CpG sites and CHH/CHG sites, respectively, at a read depth of >10 in RRBS libraries from Coriell gDNA treated with different enzyme dilutions in an exemplary embodiment.

[0022] Figure 12A and Figure 12B show correlation between Coriell CpG methylation calls in the RRBS libraries using the HiSeq method versus the UNG/Endo IV method with different enzyme dilutions, at a read depth of >10 and >50, respectively, in an exemplary embodiment.

[0023] Figure 13A and Figure 13B show a graph of total number of CpG sites and methylated CpG sites, respectively, containing *Msp*I recognition sequence in the RRBS libraries treated with different enzyme dilutions, compared directly to the CpG sites and methylated CpG sites called by the UNG/EndoIV method, in an exemplary embodiment.

DETAILED DESCRIPTION OF EMBODIMENTS OF THE DISCLOSURE

Definitions

[0024] All patents, applications, published applications and other publications referred to herein are incorporated by reference to the referenced material and in their entireties. If a term or phrase is used herein in a way that is contrary to or otherwise

inconsistent with a definition set forth in the patents, applications, published applications and other publications that are herein incorporated by reference, the use herein prevails over the definition that is incorporated herein by reference.

[0025] As used herein, the singular forms "a", "an", and "the" include plural references unless indicated otherwise, expressly or by context. For example, "a" dimer includes one or more dimers, unless indicated otherwise, expressly or by context.

[0026] The terms "polynucleotide," "oligonucleotide," "nucleic acid" and "nucleic acid molecule" are used interchangeably herein to refer to a polymeric form of nucleotides of any length, and may comprise ribonucleotides, deoxyribonucleotides, analogs thereof, or mixtures thereof. This term refers only to the primary structure of the molecule. Thus, the term includes triple-, double- and single-stranded deoxyribonucleic acid ("DNA"), as well as triple-, double- and single-stranded ribonucleic acid ("RNA"). It also includes modified, for example by alkylation, and/or by capping, and unmodified forms of the polynucleotide. More particularly, the terms "polynucleotide," "oligonucleotide," "nucleic acid" and "nucleic acid molecule" include polydeoxyribonucleotides (containing 2-deoxy-D-ribose), polyribonucleotides (containing D-ribose), including tRNA, rRNA, hRNA, and mRNA, whether spliced or unspliced, any other type of polynucleotide which is an N- or C-glycoside of a purine or pyrimidine base, and other polymers containing normucleotidic backbones, for example, polyamide (*e.g.*, peptide nucleic acids ("PNAs")) and polymorpholino (commercially available from the Anti-Virals, Inc., Corvallis, OR., as NeuGene®) polymers, and other synthetic sequence-specific nucleic acid polymers providing that the polymers contain nucleobases in a configuration which allows for base pairing and base stacking, such as is found in DNA and RNA. Thus, these terms include, for example, 3'-deoxy-2',5'-DNA, oligodeoxyribonucleotide N3' to P5' phosphoramidates, 2'-O-alkyl-substituted RNA, hybrids between DNA and RNA or between PNAs and DNA or RNA, and also include known types of modifications, for example, labels, alkylation, "caps," substitution of one or more of the nucleotides with an analog, internucleotide modifications such as, for example, those with uncharged linkages (*e.g.*, methyl phosphonates, phosphotriesters, phosphoramidates, carbamates, etc.), with negatively charged linkages (*e.g.*, phosphorothioates, phosphorodithioates, etc.), and with positively charged linkages (*e.g.*, aminoalkylphosphoramidates, aminoalkylphosphotriesters), those containing pendant

moieties, such as, for example, proteins (including enzymes (*e.g.,* nucleases), toxins, antibodies, signal peptides, poly-L-lysine, etc.), those with intercalators (*e.g.*, acridine, psoralen, etc.), those containing chelates (of, *e.g.*, metals, radioactive metals, boron, oxidative metals, etc.), those containing alkylators, those with modified linkages (*e.g.*, alpha anomeric nucleic acids, etc.), as well as unmodified forms of the polynucleotide or oligonucleotide.

[0027] As used herein, "sequence identity" or "identity" or "homology" in the context of two protein sequences (or nucleotide sequences) includes reference to the residues in the two sequences which are the same when aligned for maximum correspondence over a specified comparison window. The portion of the amino acid sequence or nucleotide sequence in the comparison window may comprise additions or deletions (i.e., gaps) as compared to the reference sequence for optimal alignment of the two sequences. When percentage of sequence identity is used in reference to proteins it is recognized that residue positions which are not identical often differ by conservative amino acid substitutions, where amino acids are substituted for other amino acid residues with similar chemical properties (e.g. charge or hydrophobicity) and therefore do not change the functional properties of the molecule. Where sequences differ in conservative substitutions, the percentage sequence identity may be adjusted upwards to correct for the conservative nature of the substitutions. Sequences, which differ by such conservative substitutions are said to have "sequence similarity" or "similarity". Means for making these adjustments are well known to persons skilled in the art. The percentage is calculated by determining the number of positions at which the identical amino acid or nucleic acid base residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity. Typically this involves scoring a conservative substitution as a partial rather than a full mismatch, thereby increasing the percentage sequence identity. Thus, for example, where an identical amino acid is given a score of 1 and a non-conservative substitution is give a score of zero, a conservative substitution is given a score between 0 and 1. The scoring of conservative substitutions is calculated, e.g. according to the algorithm of Meyers and Miller (*Computer Applic. Biol. Sci.*, 1998, 4, 11-17).

[0028] As used herein, "substantially complementary or substantially matched" means that two nucleic acid sequences have at least 90% sequence identity. Preferably, the two nucleic acid sequences have at least 95%, 96%, 97%, 98%, 99% or 100% of sequence identity. Alternatively, "substantially complementary or substantially matched" means that two nucleic acid sequences can hybridize under high stringency condition(s).

[0029] In general, the stability of a hybrid is a function of the ion concentration and temperature. Typically, a hybridization reaction is performed under conditions of lower stringency, followed by washes of varying, but higher, stringency. Moderately stringent hybridization refers to conditions that permit a nucleic acid molecule such as a probe to bind a complementary nucleic acid molecule. The hybridized nucleic acid molecules generally have at least 60% identity, including for example at least any of 70%, 75%, 80%, 85%, 90%, or 95% identity. Moderately stringent conditions are conditions equivalent to hybridization in 50% formamide, 5x Denhardt's solution, 5x SSPE, 0.2% SDS at 42°C, followed by washing in 0.2x SSPE, 0.2% SDS, at 42°C. High stringency conditions can be provided, for example, by hybridization in 50% formamide, 5x Denhardt's solution, 5x SSPE, 0.2% SDS at 42°C, followed by washing in 0.1x SSPE, and 0.1% SDS at 65°C.

[0030] Low stringency hybridization refers to conditions equivalent to hybridization in 10% formamide, 5x Denhardt's solution, 6x SSPE, 0.2% SDS at 22°C, followed by washing in 1x SSPE, 0.2% SDS, at 37°C. Denhardt's solution contains 1% Ficoll, 1% polyvinylpyrolidone, and 1% bovine serum albumin (BSA). 20x SSPE (sodium chloride, sodium phosphate, ethylene diamide tetraacetic acid (EDTA)) contains 3M sodium chloride, 0.2M sodium phosphate, and 0.025 M (EDTA). Other suitable moderate stringency and high stringency hybridization buffers and conditions are well known to those of skill in the art.

[0031] It is understood that aspects and embodiments of the invention described herein include "consisting" and/or "consisting essentially of" aspects and embodiments.

[0032] Other objects, advantages and features of the present invention will become apparent from the following specification taken in conjunction with the accompanying drawings.

[0033] ***Msp*I-based method**

- Digests at C'CGG

- Recognition sequence cleaves at CpG sites, but misses all CpG sites not in proximity to C'CGG.
- High inputs of DNA required (1-5μg)
- No CHH or CHG methylation context coverage
- CpG coverage extremely limited due to required recognition sequence
- Resolution of 1st base required to make appropriate methylation call

[0034] Proposed novel method: UNG/EndoIV RRBS

- Method does not rely upon Restriction Endonuclease
- Method detects CpG, CHG, and CHH methylation contexts
- Input required: 1μg or less
- Method enriches for methylated regions of the genome on a context-neutral basis
- Method detects greater number of methylated sites than current RRBS methods
- Method detects wider scope of current CpG sites located by *Msp*I method

Methods for Preparing Nucleic Acid Molecules

[0035] Current RRBS techniques combine restriction enzyme digestion, bisulfite conversion, and sequencing of the bisulfite-converted DNA in order to enrich and identify areas of the genome that have methylated CpG content. Specifically, *Msp*I-based methods typically employ the restriction enzyme *Msp*I to digest genomic DNA at C′CGG sites to enrich for CpG sites. However, because the recognition sequence of MspI cleaves at CpG sites, any CpG sites not in proximity to the recognition sequence C'CGG are not recognized. Because the current technique relies on recognition and restriction enzyme digestion of the 4 bp C′CGG sequence, the coverage of all CpG sites in the genome is limited and restricted to only CpG sites with flanking 5′-C and 3′-G nucleotides. Further, relatively high input levels of DNA (e.g., about 1 μg to about 5 μg) are required. MspI-based methods also provide no CHH or CHG methylation context coverage. Also, a relatively high level of resolution and confidence in the first nucleotide of a read in a sequencing reaction is required for a methylation call. There is a need for improved methods for constructing a genomic DNA library for RRBS from much lowered input DNA, with better data resolution and coverage.

[0036] Thus, there exists_a need for improved methods for constructing a genomic DNA library for RRBS from much lowered input DNA, with better data resolution and coverage.

[0037] Presented herein are methods which overcome one or more of the above-described disadvantages. For example, embodiments of the methods presented herein do not rely on restriction endonucleases such as MspI, and as such are capable of detecting_CpG, CHG, and CHH methylation contexts. Embodiments of the methods presented herein require much lowered input DNA with better data resolution and coverage, enriching for methylated regions of the genome on a context-neutral basis, and detecting a greater number of methylated sites and a wider scope of CpG sites than current RRBS methods.

[0038] Some embodiments disclosed herein provide methods for preparing a sample for sequencing, comprising: treating nucleic acid molecules in the sample to convert at least a portion of unmethylated cytosine residues into uracil residues. Typically, at least about 98% of unmethylated cytosine residues are converted into uracil residues. In some embodiments, the sample is treated with bisulfite to convert at least a portion of unmethylated cytosine residues into uracil residues.

[0039] In some embodiments, the methods disclosed herein produce a population of nucleic acid molecules suitable for reduced representation bisulfite sequencing (RRBS). For example, the methods provided herein may be used for construction of an RRBS library. In some embodiments, the nucleic acid molecules are genomic DNA.

[0040] In some embodiments, the nucleic acid molecules may be treated with a uracil DNA glycosylase resulting in a plurality of abasic residues in place of uracil residues. Persons skilled in the art would appreciate that any suitable DNA glycosylase, including but not limited to uracil DNA glycosylases, may be used to convert the uracil residues into abasic residues. For example, UNG (human Uracil-DNA glycosylase), or its orthologs in organisms other than human, may be used.

[0041] In some embodiments, the methods disclosed herein could practically be applied to other sequencing methods, including but not limited to: 5mC detection (oxBS-Seq; Song et al., *Nat. Biotech.* 30: 1107-16 (2012)); 5hmC detection (TAB-Seq; Song et al., *supra*); 5caC detection (CAB-Seq; Liu et al., *J. Am. Chem. Soc.* 135: 9315-7 (2013)). Alternatively, some embodiments disclosed herein provide methods that investigate modified

residues not converted by bisulfite from C to U. Therefore, other DNA glycosylases, such as a guanine DNA glycosylase, an adenine DNA glycosylase, a thymine DNA glycosylase, etc., may be used to convert a target nucleotide residue to an abasic residue.

[0042] In some embodiments, the methods further comprise cleaving the nucleic acid molecules at at least a portion of the uracil residues to obtain nucleic acid fragments. In some embodiments, the nucleic acid molecules having a plurality of abasic residues resulting from treatment with a uracil DNA glycosylase, e.g., UNG, may be further treated to cleave the nucleic acid molecules at the abasic residues. Any suitable treatment that is effective in cleaving the nucleic acid molecules at the abasic residues can be used for the methods disclosed herein. In some embodiments, an endonuclease, such as Endonuclease IV (Endo IV), may be used to cleave the nucleic acid molecules at the abasic residues. In other embodiments, heat treatment of the nucleic acid molecules can be used to cleave the molecules at the abasic residues. In some embodiments, the nucleic acid molecules having a plurality of abasic residues resulting from treatment with a uracil DNA glycosylase, e.g., UNG, may be used in the absence of a cleaving step because most polymerases will not amplify at the abasic sites, generating various sized fragments during PCR amplification.

[0043] **Figure 1** illustrates a flow diagram of an exemplary embodiment of a method 100 for construction of a bisulfite-treated nucleic acid methylation library for RRBS. **Figure 2** shows pictorially the steps of method 100 of Figure 1. UNG and EndoIV are used in the diagrams as examples of a uracil DNA glycosylase and a treatment to cleave the nucleic acid molecules at abasic residues, and should not be interpreted to limit the presently disclosed methods. It would be readily appreciated by those skilled in the art that other suitable enzymes or treatments may be used in substitute but would generally achieve similar functions. Method 100 may include, but is not limited to, the following steps.

[0044] At a step 110, a nucleic acid sample, such as genomic DNA, is purified through an accepted genomic DNA extraction protocol (e.g., MasterPure™ Complete DNA and RNA Purification Kit from Epicentre, Madison, WI).

[0045] In some embodiments, a nucleic acid sample includes a double-stranded nucleic acid. In some embodiments, a nucleic acid sample includes genomic DNA, or cDNA. In some embodiments, mitochondrial or chloroplast DNA is used. In some embodiments, a nucleic acid sample includes RNA or derivatives thereof such as mRNA or

cDNA. Some embodiments described herein can utilize a plurality of different nucleic acid species (e.g., nucleic acid molecules having different nucleotide sequences being present in the plurality). The nucleic acid sample may be prepared from nucleic acid molecules obtained from a single organism or from populations of nucleic acid molecules obtained from sources that include more than one organism. A nucleic acid sample can be from a single cell; from multiple cells, tissue(s) or bodily fluids of a single organism; from cells, tissues or bodily fluids of several organisms of the same species; or from multiple species, as with metagenomic samples, such as from environmental samples. Sources of nucleic acid molecules include, but are not limited to, organelles, cells, tissues, organs, or organisms.

[0046] The methods disclosed herein enable the construction of RRBS library using less nucleic acid from the sample, e.g., genomic DNA than other methods known in the art. For example, about 50 ng of nucleic acid (e.g., genomic DNA) may be used for construction of an RRBS library using the methods disclosed herein. Therefore, in some embodiments, about 1 μg of input nucleic acid (e.g., genomic DNA) may be used for bisulfite conversion. In some embodiments, less than 1 μg of input nucleic acid (e.g., genomic DNA) may be used for bisulfite conversion. For example, in some embodiments, the amount of input nucleic acid (e.g., genomic DNA) used for the construction of an RRBS library is, is about, or is less than,10 ng, 20 ng, 30 ng, 40 ng, 50 ng, 60 ng, 70 ng, 80 ng, 90 ng, 100 ng, 200 ng, 500 ng, 1 μg, 10 μg, 100 μg of nucleic acid (e.g., genomic DNA), or is an amount of nucleic acid (e.g., genomic DNA) in a range defined by any two of these values, for example, 10 μg to 100 μg, 1 μg to 10 μg, 10 ng to 1 μg, 10 ng to 100 ng, 10 ng to 50 ng, 30 ng to 100 ng, etc.

[0047] At a step 115, the purified sample nucleic acid, e.g., genomic DNA, is bisulfite converted. In some embodiments, sodium bisulfite treatment may be used to convert unmethylated cytosine into uracil, which is replaced by thymine after amplification (e.g., PCR), while 5-methylcytosine remains unchanged. This step is also shown pictorially in Figure 2.

[0048] As outlined herein, there are a variety of bisulfite conversion kits available. For example, use of the Zymo EZ DNA Methylation-Lightning Kit (Zymo Research, Irvine, CA) bisulfite conversion protocol is sufficient for conversion of >98% of the non-methylated cytosines.

[0049] At a step 120, the bisulfite-converted nucleic acid, e.g., genomic DNA, is co-digested with UNG and Endo IV. UNG catalyzes the hydrolysis of the N-glycosydic bond between uracil and sugar in nucleic acid, leaving an apyrimidinic (abasic) site in uracil-containing nucleic acids (e.g., DNA). The abasic sites formed in the nucleic acid by UNG may be subsequently cleaved by Endonuclease IV. For efficient conduction of this protocol add 1U UNG/2U Endonuclease IV and incubate for 5 minutes at 37 °C. Other embodiments of this procedure can use a ½, 1/5, 1/10, 1/25, and 1/50 dilution of this original concentration. The exemplary embodiments herein are encompassed between 1/5 and 1/25 dilution, or otherwise understood to be 0.2U UNG/0.4U Endonuclease IV to 0.04U UNG/0.08U Endonuclease IV incubated for the time dictated above. The digested DNA is purified over a Zymo Clean and Concentrator genomic DNA column (Zymo Research, Irvine, CA), after which library preparation using the EpiGnome Library Preparation Kit (Epicentre, Madison, WI) begins according to published protocols. Other embodiments of this procedure only require UNG for fragmentation of the genome. This step is also shown pictorially in Figure 2.

[0050] As will be appreciated by persons skilled in the art, any suitable procedures may be used to cleave the abasic sites in the DNA treated with UNG or another DNA glycosylase. For example, treatment of the nucleic acid molecules with a physical condition, such as heat, sonication, etc., may be sufficient to cleave the nucleic acid molecules at the abasic residues.

[0051] In some embodiments, the treatment to cleave the nucleic acid molecules may be conducted simultaneously or after the treatment of the nucleic acid molecules with a uracil DNA glycosylase. For example, UNG and Endo IV may be added to a reaction mixture at the same time in order to obtain nucleic acid fragments that are enriched in methylated cytosines.

[0052] As would be appreciated by persons skilled in the art, the treatment conditions, such as the treatment time, enzyme conditions, temperature, etc., may be varied in order to control the size distribution of the nucleic acid fragments and/or to enrich methylated cytosines in the nucleic acid fragments. Accordingly, the bisulfite-treated nucleic acid molecules may be treated at a temperature that is, is about, is lower than, or is higher than, 40 °C, 50 °C, 60 °C, 70 °C, 80 °C, 90 °C, 100 °C, or at a temperature that is in a range

defined by any two of these values, for example, 40 °C to 100 °C, 50 °C to 90 °C, 60 °C to 80 °C, 70 °C to 80 °C, etc, to cleave the nucleic acid molecules at the abasic residues. Similarly, the bisulfite-treated nucleic acid molecules may be treated with a uracil DNA glycosylase for an amount of time that is, is about, is less than, or is more than, 1 minute, 2 minutes, 3 minutes, 4 minutes, 5 minutes, 10 minutes, 20 minutes, 30 minutes, 60 minutes, 120 minutes, 240 minutes, or an amount of time that is in a range defined by any two of these values, for example, 1 minute to 240 minutes, 5 minutes to 120 minutes, 10 minutes to 60 minutes, 20 minutes to 30 minutes, etc. In some embodiments, the bisulfite-treated nucleic acid molecules may be treated with a uracil DNA glycosylase for 1 minute to 5 minutes.

[0053] Similarly, the nucleic acid molecules treated with a uracil DNA glycosylase, such as UNG, may be treated with Endo IV for an amount of time that is, is about, is less than, or is more than, 1 minute, 2 minutes, 3 minutes, 4 minutes, 5 minutes, 10 minutes, 20 minutes, 30 minutes, 60 minutes, 120 minutes, 240 minutes, or an amount of time that is in a range defined by any two of these values, for example, 1 minute to 240 minutes, 5 minutes to 120 minutes, 10 minutes to 60 minutes, 20 minutes to 30 minutes, etc. In some embodiments, the nucleic acid molecules treated with a uracil DNA glycosylase, such as UNG, may be treated with Endo IV for 1 minute to 5 minutes.

[0054] In some embodiments, the bisulfite-treated nucleic acid molecules may be treated with a uracil DNA glycosylase, such as UNG, at a concentration that is, is about, is less than, or is more than, 0.01 U/1 μg of DNA, 0.02 U/1 μg of DNA, 0.03 U/1 μg of DNA, 0.04 U/1 μg of DNA, 0.05 U/1 μg of DNA, 0.1 U/1 μg of DNA, 0.2 U/1 μg of DNA, 0.5 U/1 μg of DNA, 1 U/1 μg of DNA, 2 U/1 μg of DNA, 5 U/1 μg of DNA, or a concentration that is defined by any of these values, for example, 0.01 U/1 μg of DNA to 5 U/1 μg of DNA, 0.02 U/1 μg of DNA to 2 U/1 μg of DNA, 0.03 U/1 μg of DNA to 1 U/1 μg of DNA, 0.04 U/1 μg of DNA to 0.5 U/1 μg of DNA, etc. In some embodiments, the bisulfite-treated nucleic acid molecules may be treated with UNG at a concentration of 0.02 U/1 μg of DNA to 0.1 U/1 μg of DNA.

[0055] In some embodiments, the nucleic acid molecules treated with a uracil DNA glycosylase, such as UNG, may be treated with Endo IV at a concentration that is, is about, is less than, or is more than, 0.01 U/1 μg of DNA, 0.02 U/1 μg of DNA, 0.03 U/1 μg of DNA, 0.04 U/1 μg of DNA, 0.05 U/1 μg of DNA, 0.1 U/1 μg of DNA, 0.2 U/1 μg of

DNA, 0.5 U/1 μg of DNA, 1 U/1 μg of DNA, 2 U/1 μg of DNA, 5 U/1 μg of DNA, or a concentration that is defined by any of these values, for example, 0.01 U/1 μg of DNA to 5 U/1 μg of DNA, 0.02 U/1 μg of DNA to 2 U/1 μg of DNA, 0.03 U/1 μg of DNA to 1 U/1 μg of DNA, 0.04 U/1 μg of DNA to 0.5 U/1 μg of DNA, etc. In some embodiments, the nucleic acid molecules treated with a uracil DNA glycosylase, such as UNG, may be treated with Endo IV at a concentration of 0.04 U/1 μg of DNA to 0.2 U/1 μg of DNA.

[0056] At a step 125, the highly methylated regions are captured in a DNA methylation library for RRBS sequencing. In one example, the EpiGnome™ Methyl-Seq Kit (Epicentre, Madison, WI) may be used for the construction of a DNA methylation library. In another example, the TruSeq® DNA Sample Prep Kit (Illumina Inc., San Diego, CA) may be used for the construction of a DNA methylation library. The TruSeq® library preparation may be used to construct a library with fragment sizes of 500 bp or less. In yet another example, EpiNext Post-Bisulfite DNA Library Preparation Kit, Pico Methyl-Seq Library Prep Kit, Ovation Ultralow Mehtyl-Seq Library Kit, EpiTect Whole Bisulfitome Kit, NEXTflex Bisulfite-Seq Kit, or any other suitable kits may be used to construct a library. This step is also shown pictorially in Figure 2.

RRBS Library of Fragmented Nucleic Acid Molecules

[0057] Some embodiments disclosed herein provide nucleic acid samples having fragmented nucleic acid molecules wherein substantially all cytosine residues in the fragmented nucleic acid molecules are methylated. In some embodiments, the fragmented nucleic acid molecules result from the methods for preparing a sample for sequencing disclosed herein.

[0058] In some embodiments, cleaving the nucleic acid molecules at the previous location of a uracil residue, made abasic by UNG, and subsequently fragmented by, for example, using Endo IV or heat, may result in an apurinic/apyrimidinic site at a 5' or 3' end of the nucleic acid molecule. For example, the Endonuclease IV cleaves at the phosphodiester bond 5' to the AP DNA lesion.

[0059] Accordingly, from about 50% to about 100% of all post-bisulfite conversion cytosine residues in the fragmented nucleic acid molecules may be methylated. For example, a percentage that is, is about, is less than, or is more than, 50%, 60%, 70%,

80%, 90%, 95%, 98%, 99%, 100% of the cytosine residues, or a percentage that is a range between any two of these values, for example, 50% to 100%, 60% to 99%, 70% to 98%, 80% to 95%, etc., of the cytosine residues, in the fragmented nucleic acid molecules are methylated. In some embodiments, a portion of the CpG sites in the fragmented nucleic acid molecules are methylated. In some embodiments, at least about 0.1% of the CHG sites in the fragmented nucleic acid molecules are methylated. In some embodiments, at least about 0.1% of the CHH sites in the fragmented nucleic acid molecules are methylated.

[0060] In some embodiments, the methylated cytosines in the nucleic acid samples having fragmented nucleic acid molecules are enriched in comparison to a reference sample or samples. For example, the reference sample may be a sample treated with bisulfite to convert unmethylated cytosine residues into uracil residues, but without cleaving the nucleic acid molecules at the uracil residues. Alternatively, the reference sample may be a sample that is not treated with bisulfite to convert unmethylated cytosines residues into uracil residues, with or without being treated for cleaving at the uracil residues.

[0061] As outlined herein, methylated cytosines may be methylated CpG, CHH (H = A, T, or C) or CHG, or a combination thereof. In some embodiments, methylated cytosines may be located in a CpG island and/or a CpG shore (Pollard et al, *Cell Stem Cell.* 5(6): 571-2 (2009); Suzuki et al., *Mol. Oncol.* 6(6); 567-78 (2012)). In some embodiments, the treatment conditions, such as bisulfite treatment, conversion of unmethylated cytosine to uracil, and/or cleavage of AP sites, may be optimized to enrich the methylated cytosines in the nucleic acid samples having fragmented nucleic acid molecules, and/or to enrich CpG coverage in the nucleic acid samples having fragmented nucleic acid molecules. For example, the treatment time, enzyme conditions, temperature, etc., may be varied in order to enrich methylated cytosines in the nucleic acid fragments. It will be appreciated by those skilled in the art that by reducing enzyme concentration, fragmentation of the nucleic acid molecules may be reduced, which leads to increased coverage of CpG sites. On the other hand, by increasing enzyme concentration and fragmentation of the nucleic acid molecules, increased enrichment of methylated cytosines may be achieved, at the expense of decreased CpG coverage. It would also be appreciated that a balance may be achieved that results in both enrichment of methylated cytosines as well as acceptable CpG coverage. In some embodiments, one or more of the parameters of the treatment time, enzyme conditions,

temperature, etc., may be varied in order to enrich methylated CpG, CHH, or CHG, or a combination thereof, in the nucleic acid fragments.

[0062] Accordingly, the bisulfite-treated nucleic acid molecules may be treated at temperature conditions disclosed elsewhere herein to cleave the nucleic acid molecules at the abasic residues. Similarly, the bisulfite-treated nucleic acid molecules may be treated with a uracil DNA glycosylase for an amount of time that is, is about, is less than, or is more than, 1 minute, 2 minutes, 3 minutes, 4 minutes, 5 minutes, 10 minutes, 20 minutes, 30 minutes, 60 minutes, 120 minutes, 240 minutes, or an amount of time that is in a range defined by any two of these values, for example, 1 minute to 240 minutes, 5 minutes to 120 minutes, 10 minutes to 60 minutes, 20 minutes to 30 minutes, etc. In some embodiments, the bisulfite-treated nucleic acid molecules may be treated with a uracil DNA glycosylase for 1 minute to 5 minutes.

[0063] Similarly, the nucleic acid molecules treated with a uracil DNA glycosylase, such as UNG, may be treated with Endo IV for an amount of time that is, is about, is less than, or is more than, 1 minute, 2 minutes, 3 minutes, 4 minutes, 5 minutes, 10 minutes, 20 minutes, 30 minutes, 60 minutes, 120 minutes, 240 minutes, or an amount of time that is in a range defined by any two of these values, for example, 1 minute to 240 minutes, 5 minutes to 120 minutes, 10 minutes to 60 minutes, 20 minutes to 30 minutes, etc. In some embodiments, the nucleic acid molecules treated with a uracil DNA glycosylase, such as UNG, may be treated with Endo IV for 1 minute to 5 minutes.

[0064] In some embodiments, the bisulfite-treated nucleic acid molecules may be treated with a uracil DNA glycosylase, such as UNG, at a concentration that is, is about, is less than, or is more than, 0.01 U/1 μg of DNA, 0.02 U/1 μg of DNA, 0.03 U/1 μg of DNA, 0.04 U/1 μg of DNA, 0.05 U/1 μg of DNA, 0.1 U/1 μg of DNA, 0.2 U/1 μg of DNA, 0.5 U/1 μg of DNA, 1 U/1 μg of DNA, 2 U/1 μg of DNA, 5 U/1 μg of DNA, or a concentration that is defined by any of these values, for example, 0.01 U/1 μg of DNA to 5 U/1 μg of DNA, 0.02 U/1 μg of DNA to 2 U/1 μg of DNA, 0.03 U/1 μg of DNA to 1 U/1 μg of DNA, 0.04 U/1 μg of DNA to 0.5 U/1 μg of DNA, etc. In some embodiments, the bisulfite-treated nucleic acid molecules may be treated with UNG at a concentration of 0.02 U/1 μg of DNA to 0.1 U/1 μg of DNA.

[0065] In some embodiments, the nucleic acid molecules treated with a uracil DNA glycosylase, such as UNG, may be treated with Endo IV at a concentration that is, is about, is less than, or is more than, 0.01 U/1 μg of DNA, 0.02 U/1 μg of DNA, 0.03 U/1 μg of DNA, 0.04 U/1 μg of DNA, 0.05 U/1 μg of DNA, 0.1 U/1 μg of DNA, 0.2 U/1 μg of DNA, 0.5 U/1 μg of DNA, 1 U/1 μg of DNA, 2 U/1 μg of DNA, 5 U/1 μg of DNA, or a range of concentration that is defined by any of these values, for example, 0.01 U/1 μg of DNA to 5 U/1 μg of DNA, 0.02 U/1 μg of DNA to 2 U/1 μg of DNA, 0.03 U/1 μg of DNA to 1 U/1 μg of DNA, 0.04 U/1 μg of DNA to 0.5 U/1 μg of DNA, etc. In some embodiments, the nucleic acid molecules treated with a uracil DNA glycosylase, such as UNG, may be treated with Endo IV at a concentration of 0.04 U/1 μg of DNA to 0.2 U/1 μg of DNA.

[0066] These fragmented nucleic acid molecules may have a size distribution that is suitable for capture in the subsequent library prep and sequencing reaction. The size distribution of the fragmented nucleic acid molecules may be adjusted by a variety of reaction conditions, such as the time of the treatment, enzyme concentration, temperature, etc. In addition, a large portion of fragmented nucleotides may not be large enough for capture, as a result of the uracil site cleavage. This loss will eventually lead to the methylation call and CpG, CHG, and CHH site enrichment of the methylated residues containing cytosines. Persons of ordinary skill in the art would appreciate that the size distribution of the fragmented nucleic acid molecules may be optimized in order to achieve an enrichment of methylation calls, including CpG methylation, CHG methylation, and/or CHH methylation. For example, the fragmented nucleic acid molecules have a size that is, is about, is less than, or is more than, 10 bp, 20 bp, 30 bp, 40 bp, 50 bp, 60 bp, 70 bp, 80 bp, 90 bp, 100 bp, 110 bp, 120 bp, 150 bp, 200 bp, 300 bp, 400 bp, 500 bp, 100 bp, or 2,000 bp, or a size that is a range between any of these values, for example, 10 bp to 2000 bp, 20 bp to 1000 bp, 50 bp to 500 bp, 100 bp to 200 bp, etc. In some embodiments, the fragmented nucleic acid molecules have a size of 50 bp to 500 bp. In some embodiments, the fragmented nucleic acid molecules have a size of 100 bp to 200 bp. In some embodiments, the fragmented nucleic acid molecules have a mean size that is, is about, is less than, or is more than, 50 bp, 60 bp, 70 bp, 80 bp, 90 bp, 100 bp, 110 bp, 120 bp, 130 bp, 140 bp, 150 bp, 160 bp, 170 bp, 180 bp, 190 bp, 200 bp, or a size that is a range between any of these values, for example, 50

bp to 200 bp, 60 bp to 190 bp, 70 bp to 180 bp, 800 bp to 170 bp, etc. In some embodiments, the fragmented nucleic acid molecules have a mean size of 140 bp to 165 bp.

[0067] Persons skilled in the art would readily appreciate the advantages of the presently disclosed methods over those already known in the art, such as *Msp*I-based, *Taq*I-based, or a combination thereof, RRBS library construction methods. For example, because the methods disclosed herein are not dependent on a sequence-recognition enzyme, and/or coverage of methylated cytosine sites not in the proximity of the recognition sequence of a restriction enzyme, such as *Msp*I (which is limited to the sequence C'GCC), would be greatly increased. Further, coverage of methylated cytosines will not be limited to CpG sites, but also CHG and CHH sites.

Methods of RRBS Sequencing

[0068] Embodiments disclosed herein further provide methods for sequencing nucleic acid fragments, which comprises obtaining the nucleic acid fragments from a sample treated with a method disclosed herein, and sequencing the nucleic acid fragments. Without being bound by a theory, the methods disclosed herein may be optimized for constructing a library for RRBS sequencing.

[0069] Persons skilled in the art would appreciate that the methods disclosed herein may achieve enrichment of methylation calls by varying the treatment conditions of cleaving the nucleic acid molecules at the uracil residues to obtain nucleic acid fragments. It would also be appreciated that enrichment of methylation calls may be inversely related to unique alignment, because of the reduction in length of the fragmented nucleic acid molecules. Therefore, experiments may be conducted to identify optimal treatment time(s) and/or enzyme dilution(s) that result in optimal enrichment of methylation calls as well as desirable unique alignments.

[0070] The library construction protocol may also include a fragment size selection step. For example, the fragment size selection step may be a bead-based protocol that is used to select fragments within a desired base pair range. In some embodiments, SPRIselect beads (Beckman Coulter) may be used to select fragments in a certain base pair range. In some embodiments, Agencourt AmPure® XP system (Beckman Coulter) may be used to select fragments in a certain base pair range.

[0071] In some embodiments, the library of fragmented nucleic acid molecules is sequenced by sequencing by synthesis (SBS). In some embodiments, the sequences of the nucleic acid fragments are compared to a reference sequence. In some embodiments, methylated cytosines in the nucleic acid fragments are identified. In some embodiments, the reference sequence comprises a methylome. In some embodiments, the methylated cytosines are identified using Bismark. In some embodiments, the methylated cytosines are analyzed using an integrated genome viewer.

[0072] In some embodiments, the method of the invention provides for construction of an RRBS library with reduced bias for CpG methylation regions. Methylation of cytosine at CpG sites is considered to be the most important methylation site physiologically. However, methylation of cytosine also occurs in other sequence contexts such as CHG and CHH. In some embodiments, the methods disclosed herein provide for enrichment of CHG methylated regions. In some embodiments, the methods disclosed herein provide for enrichment of CHH methylated regions. Enrichment of CHG and CHH regions may provide a mechanism for analyzing the physiological role of cytosine methylation at alternative sites in the genome, e.g., cytosine methylation in the CHG and CHH context.

[0073] In some embodiments, the fragmented nucleic acid molecules may be constructed by the methods for preparing a sample for sequencing methods disclosed herein. In some embodiments, the fragmented nucleic acid molecules are subjected to amplification for preparing a sample for sequencing. Any suitable amplification methodology known in the art can be used. In some embodiments, nucleic acid fragments are amplified in or on a substrate. For example, in some embodiments, the nucleic acid fragments are amplified using bridge amplification methodologies as exemplified by the disclosures of U.S. Pat. No. 5,641,658; U.S. Patent Publ. No. 2002/0055100; U.S. Pat. No. 7,115,400; U.S. Patent Publ. No. 2004/0096853; 10 U.S. Patent Publ. No. 2004/0002090; U.S. Patent Publ. No. 2007/0128624; and U.S. Patent Publ. No. 2008/0009420, each of which is incorporated herein by reference in its entirety.

[0074] Bridge amplification methods allow amplification products to be immobilized in or on a substrate in order to form arrays comprised of clusters (or "colonies") of immobilized nucleic acid molecules. Each cluster or colony on such an array is formed from a plurality of identical immobilized polynucleotide strands and a plurality of identical

immobilized complementary polynucleotide strands. The arrays so-formed can be referred to herein as "clustered arrays". The products of solid-phase amplification reactions are so-called "bridged" structures when formed by annealed pairs of immobilized polynucleotide strands and immobilized complementary strands, both strands being immobilized on the solid support at the 5' end, preferably via a covalent attachment. Bridge amplification methodologies are examples of methods wherein an immobilized nucleic acid template is used to produce immobilized amplicons. Other suitable methodologies can also be used to produce immobilized amplicons from immobilized nucleic acid fragments produced according to the methods provided herein. For example one or more clusters or colonies can be formed via solid-phase PCR, solid-phase MDA, solid-phase RCA etc. whether one or both primers of each pair of amplification primers are immobilized.

[0075] It will be appreciated that any of the amplification methodologies described herein or generally known in the art can be utilized with universal or target-specific primers to amplify immobilized DNA fragments. Suitable methods for amplification include, but are not limited to, the polymerase chain reaction (PCR), strand displacement amplification (SDA), transcription mediated amplification (TMA) and nucleic acid sequence based amplification (NASBA), for example, as described in U.S. Patent No. 8,003,354, which is incorporated herein by reference in its entirety. The above amplification methods can be employed to amplify one or more nucleic acids of interest. For example, PCR, multiplex PCR, SDA, TMA, NASBA and the like can be utilized to amplify immobilized nucleic acid fragments. In some embodiments, primers directed specifically to the nucleic acid of interest are included in the amplification reaction.

[0076] Other suitable methods for amplification of nucleic acids can include oligonucleotide extension and ligation, rolling circle amplification (RCA) (Lizardi et al., *Nat. Genet.* 19:225-232 (1998), which is incorporated herein by reference in its entirety) and oligonucleotide ligation assay (OLA) (*See e.g.*, U.S. Pat. Nos. 7,582,420, 5,185,243, 5,679,524 and 5,573,907; EP 0320308; EP 0336731; EP 0439182; WO 90101069; WO 89/12696; and WO 89109835, each of which is incorporated herein by reference in its entirety). It will be appreciated that these amplification methodologies can be designed to amplify immobilized nucleic acid fragments. For example, in some embodiments, the amplification method can include ligation probe amplification or oligonucleotide ligation

assay (OLA) reactions that contain primers directed specifically to the nucleic acid of interest. In some embodiments, the amplification method can include a primer extension-ligation reaction that contains primers directed specifically to the nucleic acid of interest. As a non-limiting example of primer extension and ligation primers that can be specifically designed to amplify a nucleic acid of interest, the amplification can include primers used for the GoldenGate® assay (Illumina, Inc., San Diego, CA) or one or more assay set forth in U.S. Pat. No. 7,582,420 and 7,611,869, each of which is incorporated herein by reference in its entirety.

[0077] An isothermal amplification technique can be used in a method of the present disclosure. Exemplary isothermal amplification methods include, but are not limited to, Multiple Displacement Amplification (MDA) as exemplified by, for example, Dean et al., *Proc. Natl. Acad. Sci. USA* 99:5261-66 (2002) or isothermal strand displacement nucleic acid amplification as exemplified by, for example U.S. Pat. No. 6,214,587, each of which is incorporated herein by reference in its entirety. Other non-PCR-based methods that can be used in the present disclosure include, for example, strand displacement amplification (SDA) which is described in, for example Walker et al., *Molecular Methods for Virus Detection,* Academic Press, Inc., 1995; U.S. Pat. Nos. 5,455,166, and 5,130,238, and Walker et al., *Nucl. Acids Res.* 20:1691-96 (1992) or hyperbranched strand displacement amplification which is described in, for example Lage et al., *Genome Research* 13:294-307 (2003), each of which is incorporated herein by reference in its entirety.

[0078] Additional description of amplification reactions, conditions and components are set forth in U.S. Patent No. 7,670,810, which is incorporated herein by reference in its entirety. Other useful isothermal amplification techniques include recombinase-facilitated amplification techniques such as those sold commercially as TwistAmp™ kits by TwistDx (Cambridge, UK). Useful components of recombinase-facilitated amplification reagent and reaction conditions are set forth in US 5,223,414 and US 7,399,590, each of which is incorporated herein by reference in its entirety. Helicase dependent amplification can also be used, for example, as described in Xu et al. *EMBO Rep* 5:795-800 (2004), which is incorporated herein by reference in its entirety.

[0079] In some embodiments, it may be desirable to perform a re-seeding step. For example, modified nucleic acid fragments can be captured at locations within a region of

a surface, replicated on one or more cycles of an amplification process, the original fragments and/or amplicons thereof can be released from the locations, the released nucleic acids can be captured at other locations in the same region, and the newly captured nucleic acids can be amplified. In a specific example, a single cycle of bridge amplification can be carried out for a fragment that was seeded on a surface and instead of washing away the original template fragment upon release from the surface, the template fragment can re-seed the surface at a new location that is proximal to the location where it had originally seeded. Subsequent rounds of bridge amplification will allow cluster growth at both the original seed location and at the re-seed location. Using such methods replicate colonies can be created at a region of a surface to provide technical replicates. Analysis of the sequences for the technical replicates can provide the benefit of error checking. For example, observed sequence variants that occur in only a subset of proximal clusters (that are identified as technical replicates) can be identified as amplification errors, whereas sequence variants that occur in all clusters that are identified as technical replicates for a particular fragment are more likely to be true variants.

Sequencing Methods

[0080] The methods described herein can be used in conjunction with a variety of nucleic acid sequencing techniques. Particularly applicable techniques are those wherein nucleic acids are attached at fixed locations in an array such that their relative positions do not change and wherein the array is repeatedly imaged. Embodiments in which images are obtained in different color channels, for example, coinciding with different labels used to distinguish one nucleotide base type from another are particularly applicable. In some embodiments, the process to determine the nucleotide sequence of a target nucleic acid can be an automated process. Preferred embodiments include sequencing-by-synthesis ("SBS") techniques.

[0081] "Sequencing-by-synthesis ("SBS") techniques" generally involve the enzymatic extension of a nascent nucleic acid strand through the iterative addition of nucleotides against a template strand. In traditional methods of SBS, a single nucleotide monomer may be provided to a target nucleotide in the presence of a polymerase in each delivery. However, in the methods described herein, more than one type of nucleotide

monomer can be provided to a target nucleic acid in the presence of a polymerase in a delivery.

[0082] SBS can utilize nucleotide monomers that have a terminator moiety or those that lack any terminator moieties. Methods utilizing nucleotide monomers lacking terminators include, for example, pyrosequencing and sequencing using γ-phosphate-labeled nucleotides, as set forth in further detail below. In methods using nucleotide monomers lacking terminators, the number of nucleotides added in each cycle is generally variable and dependent upon the template sequence and the mode of nucleotide delivery. For SBS techniques that utilize nucleotide monomers having a terminator moiety, the terminator can be effectively irreversible under the sequencing conditions used as is the case for traditional Sanger sequencing which utilizes dideoxynucleotides, or the terminator can be reversible as is the case for sequencing methods developed by Solexa (now Illumina, Inc.).

[0083] SBS techniques can utilize nucleotide monomers that have a label moiety or those that lack a label moiety. Accordingly, incorporation events can be detected based on a characteristic of the label, such as fluorescence of the label; a characteristic of the nucleotide monomer such as molecular weight or charge; a byproduct of incorporation of the nucleotide, such as release of pyrophosphate; or the like. In embodiments, where two or more different nucleotides are present in a sequencing reagent, the different nucleotides can be distinguishable from each other, or alternatively, the two or more different labels can be the indistinguishable under the detection techniques being used. For example, the different nucleotides present in a sequencing reagent can have different labels and they can be distinguished using appropriate optics as exemplified by the sequencing methods developed by Solexa (now Illumina, Inc.).

[0084] Preferred embodiments include pyrosequencing techniques. Pyrosequencing detects the release of inorganic pyrophosphate (PPi) as particular nucleotides are incorporated into the nascent strand (Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. (1996) "Real-time DNA sequencing using detection of pyrophosphate release." Analytical Biochemistry 242(1), 84-9; Ronaghi, M. (2001) "Pyrosequencing sheds light on DNA sequencing." Genome Res. 11(1), 3-11; Ronaghi, M., Uhlen, M. and Nyren, P. (1998) "A sequencing method based on real-time pyrophosphate." Science 281(5375), 363; U.S. Pat. No. 6,210,891; U.S. Pat. No. 6,258,568 and U.S. Pat. No.

6,274,320, the disclosures of which are incorporated herein by reference in their entireties). In pyrosequencing, released PPi can be detected by being immediately converted to adenosine triphosphate (ATP) by ATP sulfurylase, and the level of ATP generated is detected via luciferase-produced photons. The nucleic acids to be sequenced can be attached to features in an array and the array can be imaged to capture the chemiluminscent signals that are produced due to incorporation of a nucleotides at the features of the array. An image can be obtained after the array is treated with a particular nucleotide type (e.g., A, T, C or G). Images obtained after addition of each nucleotide type will differ with regard to which features in the array are detected. These differences in the image reflect the different sequence content of the features on the array. However, the relative locations of each feature will remain unchanged in the images. The images can be stored, processed and analyzed using the methods set forth herein. For example, images obtained after treatment of the array with each different nucleotide type can be handled in the same way as exemplified herein for images obtained from different detection channels for reversible terminator-based sequencing methods.

[0085] In another exemplary type of SBS, cycle sequencing is accomplished by stepwise addition of reversible terminator nucleotides containing, for example, a cleavable or photobleachable dye label as described, for example, in International Patent Pub. No. WO 04/018497 and U.S. Patent 7,057,026, the disclosures of which are incorporated herein by reference in their entireties. This approach is being commercialized by Solexa (now Illumina Inc.), and is also described in International Patent Pub. No. WO 91/06678 and International Patent Pub. No. WO 07/123,744, the disclosures of which are incorporated herein by reference in their entireties. The availability of fluorescently-labeled terminators in which both the termination can be reversed and the fluorescent label cleaved facilitates efficient cyclic reversible termination (CRT) sequencing. Polymerases can also be co-engineered to efficiently incorporate and extend from these modified nucleotides.

[0086] Preferably in reversible terminator-based sequencing embodiments, the labels do not substantially inhibit extension under SBS reaction conditions. However, the detection labels can be removable, for example, by cleavage or degradation. Images can be captured following incorporation of labels into arrayed nucleic acid features. In particular embodiments, each cycle involves simultaneous delivery of four different nucleotide types to

the array and each nucleotide type has a spectrally distinct label. Four images can then be obtained, each using a detection channel that is selective for one of the four different labels. Alternatively, different nucleotide types can be added sequentially and an image of the array can be obtained between each addition step. In such embodiments each image will show nucleic acid features that have incorporated nucleotides of a particular type. Different features will be present or absent in the different images due the different sequence content of each feature. However, the relative position of the features will remain unchanged in the images. Images obtained from such reversible terminator-SBS methods can be stored, processed and analyzed as set forth herein. Following the image capture step, labels can be removed and reversible terminator moieties can be removed for subsequent cycles of nucleotide addition and detection. Removal of the labels after they have been detected in a particular cycle and prior to a subsequent cycle can provide the advantage of reducing background signal and crosstalk between cycles. Examples of useful labels and removal methods are set forth below.

[0087] In particular embodiments some or all of the nucleotide monomers can include reversible terminators. In such embodiments, reversible terminators/cleavable fluors can include fluor linked to the ribose moiety via a 3' ester linkage (Metzker, *Genome Res.* 15:1767-1776 (2005), which is incorporated herein by reference in its entirety). Other approaches have separated the terminator chemistry from the cleavage of the fluorescence label (Ruparel et al., *Proc Natl Acad Sci USA* 102: 5932-7 (2005), which is incorporated herein by reference in its entirety). Ruparel et al described the development of reversible terminators that used a small 3' allyl group to block extension, but could easily be deblocked by a short treatment with a palladium catalyst. The fluorophore was attached to the base via a photocleavable linker that could easily be cleaved by a 30 second exposure to long wavelength UV light. Thus, either disulfide reduction or photocleavage can be used as a cleavable linker. Another approach to reversible termination is the use of natural termination that ensues after placement of a bulky dye on a dNTP. The presence of a charged bulky dye on the dNTP can act as an effective terminator through steric and/or electrostatic hindrance. The presence of one incorporation event prevents further incorporations unless the dye is removed. Cleavage of the dye removes the fluor and effectively reverses the termination. Examples of modified nucleotides are also described in U.S. Patent 7,427,673, and U.S.

Patent 7,057,026, the disclosures of which are incorporated herein by reference in their entireties.

[0088] Additional exemplary SBS systems and methods which can be utilized with the methods and systems described herein are described in U.S. Patent Pub. No. 2007/0166705, U.S. Patent Pub. No. 2006/0188901, U.S. Patent 7,057,026, U.S. Patent Pub. No. 2006/0240439, U.S. U.S. Patent Pub. No. 2006/0281109, International Patent Pub. No. WO 05/065814, U.S. Patent Pub. No. 2005/0100900, International Patent Pub. No. WO 06/064199, International Patent Pub. No. WO 07/010,251, U.S. U.S. Patent Pub. No. 2012/0270305 and U.S. Patent Pub. No. 2013/0260372, the disclosures of each of which are incorporated herein by reference in its entirety.

[0089] Some embodiments can utilize detection of four different nucleotides using fewer than four different labels. For example, SBS can be performed utilizing methods and systems described in U.S. Patent Pub. No. 2013/0079232, which is incorporated herein by reference in its entirety. As a first example, a pair of nucleotide types can be detected at the same wavelength, but distinguished based on a difference in intensity for one member of the pair compared to the other, or based on a change to one member of the pair (e.g., via chemical modification, photochemical modification or physical modification) that causes apparent signal to appear or disappear compared to the signal detected for the other member of the pair. As a second example, three of four different nucleotide types can be detected under particular conditions while a fourth nucleotide type lacks a label that is detectable under those conditions, or is minimally detected under those conditions (e.g., minimal detection due to background fluorescence, etc). Incorporation of the first three nucleotide types into a nucleic acid can be determined based on presence of their respective signals and incorporation of the fourth nucleotide type into the nucleic acid can be determined based on absence or minimal detection of any signal. As a third example, one nucleotide type can include label(s) that are detected in two different channels, whereas other nucleotide types are detected in no more than one of the channels. The aforementioned three exemplary configurations are not considered mutually exclusive and can be used in various combinations. An exemplary embodiment that combines all three examples, is a fluorescent-based SBS method that uses a first nucleotide type that is detected in a first channel (e.g., dATP having a label that is detected in the first channel when excited by a first excitation

wavelength), a second nucleotide type that is detected in a second channel (e.g., dCTP having a label that is detected in the second channel when excited by a second excitation wavelength), a third nucleotide type that is detected in both the first and the second channel (e.g., dTTP having at least one label that is detected in both channels when excited by the first and/or second excitation wavelength) and a fourth nucleotide type that lacks a label that is not, or minimally, detected in either channel (e.g., dGTP having no label).

[0090] Further, as described in the incorporated disclosure of U.S. Patent Pub. No. 2013/0079232, sequencing data can be obtained using a single channel. In such so-called one-dye sequencing approaches, the first nucleotide type is labeled but the label is removed after the first image is generated, and the second nucleotide type is labeled only after a first image is generated. The third nucleotide type retains its label in both the first and second images, and the fourth nucleotide type remains unlabeled in both images.

[0091] Some embodiments can utilize sequencing by ligation (SBL) techniques. Such techniques utilize DNA ligase to incorporate oligonucleotides and identify the incorporation of such oligonucleotides. The oligonucleotides typically have different labels that are correlated with the identity of a particular nucleotide in a sequence to which the oligonucleotides hybridize. As with other SBS methods, images can be obtained following treatment of an array of nucleic acid features with the labeled sequencing reagents. Each image will show nucleic acid features that have incorporated labels of a particular type. Different features will be present or absent in the different images due the different sequence content of each feature, but the relative position of the features will remain unchanged in the images. Images obtained from ligation-based sequencing methods can be stored, processed and analyzed as set forth herein. Exemplary SBL systems and methods which can be utilized with the methods and systems described herein are described in U.S. Patent 6,969,488, U.S. Patent 6,172,218, and U.S. Patent 6,306,597, the disclosures of which are incorporated herein by reference in their entireties.

[0092] Some embodiments can utilize nanopore sequencing (Deamer, D. W. & Akeson, M. "Nanopores and nucleic acids: prospects for ultrarapid sequencing." Trends Biotechnol. 18, 147-151 (2000); Deamer, D. and D. Branton, "Characterization of nucleic acids by nanopore analysis". Acc. Chem. Res. 35:817-825 (2002); Li, J., M. Gershow, D. Stein, E. Brandin, and J. A. Golovchenko, "DNA molecules and configurations in a solid-

state nanopore microscope" Nat. Mater. 2:611-615 (2003), the disclosures of which are incorporated herein by reference in their entireties). In such embodiments, the target nucleic acid passes through a nanopore. The nanopore can be a synthetic pore or biological membrane protein, such as α-hemolysin. As the target nucleic acid passes through the nanopore, each base-pair can be identified by measuring fluctuations in the electrical conductance of the pore. (U.S. Patent 7,001,792; Soni, G. V. & Meller, "A. Progress toward ultrafast DNA sequencing using solid-state nanopores." Clin. Chem. 53, 1996-2001 (2007); Healy, K. "Nanopore-based single-molecule DNA analysis." Nanomed. 2, 459-481 (2007); Cockroft, S. L., Chu, J., Amorin, M. & Ghadiri, M. R. "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution." J. Am. Chem. Soc. 130, 818-820 (2008), the disclosures of which are incorporated herein by reference in their entireties). Data obtained from nanopore sequencing can be stored, processed and analyzed as set forth herein. In particular, the data can be treated as an image in accordance with the exemplary treatment of optical images and other images that is set forth herein.

[0093] Some embodiments can utilize methods involving the real-time monitoring of DNA polymerase activity. Nucleotide incorporations can be detected through fluorescence resonance energy transfer (FRET) interactions between a fluorophore-bearing polymerase and γ-phosphate-labeled nucleotides as described, for example, in U.S. Patent 7,329,492 and U.S. Patent 7,211,414 (each of which is incorporated herein by reference in its entirety) or nucleotide incorporations can be detected with zero-mode waveguides as described, for example, in U.S. Patent 7,315,019 (which is incorporated herein by reference in its entirety) and using fluorescent nucleotide analogs and engineered polymerases as described, for example, in U.S. Patent 7,405,281 and U.S. Patent Pub. No. 2008/0108082 (each of which is incorporated herein by reference in its entirety). The illumination can be restricted to a zeptoliter-scale volume around a surface-tethered polymerase such that incorporation of fluorescently labeled nucleotides can be observed with low background (Levene, M. J. et al. "Zero-mode waveguides for single-molecule analysis at high concentrations." Science 299, 682-686 (2003); Lundquist, P. M. et al. "Parallel confocal detection of single molecules in real time." Opt. Lett. 33, 1026-1028 (2008); Korlach, J. et al. "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nano structures." Proc. Natl. Acad. Sci. USA 105, 1176-

1181 (2008), the disclosures of which are incorporated herein by reference in their entireties). Images obtained from such methods can be stored, processed and analyzed as set forth herein.

[0094] Some SBS embodiments include detection of a proton released upon incorporation of a nucleotide into an extension product. For example, sequencing based on detection of released protons can use an electrical detector and associated techniques that are commercially available from Ion Torrent (Guilford, CT, a Life Technologies subsidiary) or sequencing methods and systems described in U.S. Patent Pub. No. 2009/0026082; U.S. Patent Pub. No. 2009/0127589; U.S. Patent Pub. No. 2010/0137143; or U.S. Patent Pub. No. 2010/0282617, each of which is incorporated herein by reference in its entirety. Methods set forth herein for amplifying target nucleic acids using kinetic exclusion can be readily applied to substrates used for detecting protons. More specifically, methods set forth herein can be used to produce clonal populations of amplicons that are used to detect protons.

[0095] The above SBS methods can be advantageously carried out in multiplex formats such that multiple different target nucleic acids are manipulated simultaneously. In particular embodiments, different target nucleic acids can be treated in a common reaction vessel or on a surface of a particular substrate. This allows convenient delivery of sequencing reagents, removal of unreacted reagents and detection of incorporation events in a multiplex manner. In embodiments using surface-bound target nucleic acids, the target nucleic acids can be in an array format. In an array format, the target nucleic acids can be typically bound to a surface in a spatially distinguishable manner. The target nucleic acids can be bound by direct covalent attachment, attachment to a bead or other particle or binding to a polymerase or other molecule that is attached to the surface. The array can include a single copy of a target nucleic acid at each site (also referred to as a feature) or multiple copies having the same sequence can be present at each site or feature. Multiple copies can be produced by amplification methods such as, bridge amplification or emulsion PCR as described in further detail below.

[0096] The methods set forth herein can use arrays having features at a density that is, is about, is less than, or is more than, 10 features/cm^2, 100 features/cm^2, 500 features/cm^2, 1,000 features/cm^2, 5,000 features/cm^2, 10,000 features/cm^2, 50,000 features/cm^2, 100,000 features/cm^2, 1,000,000 features/cm^2, 5,000,000 features/cm^2, or a

density that is a range between any of these values, for example, 10 features/cm^2 to 5,000,000 features/cm^2, 100 features/cm^2 to 1,000,000 features/cm^2, 500 features/cm^2 to 100,000 features/cm^2, 1,000 features/cm^2 to 50,000 features/cm^2, 5,000 features/cm^2 to 10,000 features/cm^2, etc.

[0097] An advantage of the methods set forth herein is that they provide for rapid and efficient detection of a plurality of target nucleic acid in parallel. Accordingly the present disclosure provides integrated systems capable of preparing and detecting nucleic acids using techniques known in the art such as those exemplified above. Thus, an integrated system of the present disclosure can include fluidic components capable of delivering amplification reagents and/or sequencing reagents to one or more immobilized DNA fragments, the system comprising components such as pumps, valves, reservoirs, fluidic lines and the like. A flow cell can be configured and/or used in an integrated system for detection of target nucleic acids. Exemplary flow cells are described, for example, in U.S. Patent Pub. No. 2010/0111768 A1 and U.S. Patent App. No. 13/273,666, each of which is incorporated herein by reference in its entirety. As exemplified for flow cells, one or more of the fluidic components of an integrated system can be used for an amplification method and for a detection method. Taking a nucleic acid sequencing embodiment as an example, one or more of the fluidic components of an integrated system can be used for an amplification method set forth herein and for the delivery of sequencing reagents in a sequencing method such as those exemplified above. Alternatively, an integrated system can include separate fluidic systems to carry out amplification methods and to carry out detection methods. Examples of integrated sequencing systems that are capable of creating amplified nucleic acids and also determining the sequence of the nucleic acids include, without limitation, the MiSeq™ platform (Illumina, Inc., San Diego, CA) and devices described in U.S. Patent App. No. 13/273,666, which is incorporated herein by reference in its entirety.

Kits

[0098] Embodiments disclosed herein further provide kits comprising at least one container means, wherein the at least one container means comprises a reagent that cleaves a nucleic acid molecule at a uracil residue. In some embodiments, the reagent is UNG and/or

endonuclease IV (Endo IV). In some embodiments, the container means may be a tube, a well, a microtiter plate, etc.

EXAMPLES

[0099] The following examples are offered to illustrate but not to limit the invention.

[0100] In order to facilitate understanding, the specific embodiments are provided to help interpret the technical proposal, that is, these embodiments are only for illustrative purposes, but not in any way to limit the scope of the invention. Unless otherwise specified, embodiments do not indicate the specific conditions, are in accordance with the conventional conditions or the manufacturer's recommended conditions.

Example 1 Evaluation of libraries constructed using the UNG/Endo IV method.

[0101] The UNG/Endo IV library construction method 100 of Figure 1 was evaluated using two different genomic DNA sources, i.e., Coriell 18507 and Hela CpG hypermethylated DNA. The whole genome methylation profile (methylome) for Coriell 18507 is known and may be used as a reference for determining percent coverage in subsequent RRBS of the UNG/Endo IV generated RRBS libraries. The Hela genome is hypermethylated (i.e., every CpG site is methylated) and may be used as a reference for percent coverage of CpG sites in subsequent RRBS of the UNG/Endo IV generated libraries. For the purposes of this disclosure, all references to Hela, refer specifically to the Hela CpG hypermethylated genome.

[0102] **Figure 3** shows a photograph of an agarose gel used to evaluate a UNG/Endo IV digestion time-course of bisulfite-treated genomic DNAs. In this example, T2 DNA, which is fully hydroxymethylated, was used as a negative control for UNG/Endo IV digestion of bisulfite-treated DNA. The experimental details were as follows: 1 μg of bisulfite-treated DNA (Coriell 18507, Hela, or T2) was co-digested with UNG (1 unit) and Endo IV (2 units) in 50μl of a final concentration of reaction buffer containing 50mM Tris-HCl (pH 9.0), 20mM NH_2SO_4, and 10mM EDTA for 0 (no enzyme), 30, 60, 120, or 240 minutes at 37 °C. Non-bisulfite treated Coriell 18507 and Hela DNA were also co-digested

with UNG/Endo IV for 30 or 240 minutes. An aliquot (20 μL) of each reaction was loaded into individual lanes of an agarose gel. The enzyme treatment (UNG/Endo IV), bisulfite treatment, and lane designation on the agarose gel are shown in Table 1. A DNA size ladder (M) was loaded as shown in the Figure 3. The arrow indicates a DNA ladder fragment size of about 500 bp. The data show that UNG and Endo IV co-digest and fragment bisulfite-treated DNA appropriately when compared to T2 DNA (negative control). At the 30 minute time point, for both Coriell and Hela DNA there is fragmentation to approximately 300-700 bp. At the 240 minute time point, the DNA is digested to 100-200 bp. Excessive fragmentation of the genomic DNA may, for example, be reduced by reducing the digestion time period. In the absence of bisulfite treatment (i.e., Coriell lanes 6 and 7; Hela lanes 6 and 7) the DNA was not digested by UNG/Endo IV treatment.

Table 1. Genome source and UNG/Endo IV, BIS treatment

	Lane	**Enzyme Treatment**	**BIS**
Coriell	1	No enzyme	+
	2	30 min	+
	3	60 min	+
	4	120 min	+
	5	240 min	+
	6	30 min	-
	7	240 min	-
Hela	1	No enzyme	+
	2	30 min	+
	3	60 min	+
	4	120 min	+
	5	240 min	+
	6	30 min	-
	7	240 min	-
T2	1	Non enzyme	+
	2	30 min	+
	3	240 min	+

Example 2 Fragment size distribution of RRBS libraries.

[0103] The bisulfite-treated Coriell and Hela DNAs were used to construct RRBS libraries using the EpiGnome™ Methyl-Seq Kit (Epicentre). Briefly, the bisulfite-treated and UNG/Endo IV DNA is random primed using a polymerase able to read uracil nucleotides to synthesize DNA strands containing a specific sequence tag. The 3′-ends of the newly synthesized DNA strands are then selectively tagged with a second specific sequence tag resulting in di-tagged DNA molecules with known sequence tags at the 5′- and 3′-ends. The di-tagged DNA is PCR amplified and ready for sequencing.

[0104] **Figure 4** shows a panel of BioAnalyzer traces of the fragment size distribution in RRBS libraries from Hela gDNA. The libraries were prepared using 50 ng input of each untreated (no UNG/Endo IV digestion) and treated (UNG/Endo IV for 30, 120, and 240 minutes) bisulfite-converted DNA and 10 (untreated), 12 (UNG/Endo IV 30 minutes), or 15 (UNG/Endo IV 30, 120, and 240 minutes) cycles of PCR amplification. The untreated (no UNG/Endo IV) Hela DNA and the 30 minute treated Hela DNAs generated libraries with good size distribution and yield. The 120 and 240 minute treated Hela DNAs generated libraries with reduced yield which may be due to over fragmentation. These data indicate at this level of digestion, 15 cycles are necessary for good library amplification. But, fewer cycles may be possible if less fragmentation of the genome occurs. RRBS libraries generated with the untreated and treated Coriell DNA showed similar results (data not shown).

Example 3 Sequencing of RRBS libraries.

[0105] The Hela and Coriell RRBS libraries (untreated, 10 of cycles PCR and treated, 15 cycles of PCR) were subsequently clonally amplified (cluster generation) and sequenced on the Genome Analyzer platform (Illumina).

[0106] **Figure 5A, 5B, and 5C** show a screenshot of the Genome Analyzer Sequencer (Illumina Inc., San Diego, CA) set-up parameters, a screenshot of the status pane with quality metrics, and a screenshot of the read summary, read 1, and read 2 metrics, respectively, for the analysis of RRBS libraries from Hela or Coriell gDNA. Referring to Figure 5A, screenshot 500 shows a summary of the sequencing run parameters. For

example, the sequencing run was paired-end (no index read) with 76 cycles/read and 24 tiles per lane. The sample ID shows the lane designation, library type, insert size, and concentration for each sample analyzed.

[0107] Referring to Figure 5B, screenshot shows the status pane of the sequencing analysis viewer showing quality metrics for the run. The quality metrics demonstrate a high Q30 quality score, averaging 91.4% for Read 1 and Read 2. Typical acceptable Q30 is considered >80%. In addition, the data by cycle demonstrates the % Cytosine is depleted as one would expect after bisulfite conversion; however, the cytosine is slightly higher than the expected 1% remaining due to the effects of cytosine concentration in the UNG/Endo IV method.

[0108] Referring to Figure 5C, screenshot shows a run summary and the sequencing metrics for reads 1 and 2. The percent clusters that pass filter (Cluster PF (%)) is about 93% to about 95% for all libraries. Typical acceptable PF is >80%. The reads that pass filter (Reads PF (M)) were about 10 million for all libraries. For all libraries, the Q30 scores (% >=Q30) are in the mid-90s (about 93% to about 95%) for read 1 and the high-80s (about 88% to about 90%) for read 2.

Example 4 Effects of treatment time and enzyme dilution on fragment size.

[0109] **Figures 6A and 6B** show a series of plots of the insert size in the Coriell RRBS libraries treated with enzyme for different times and a series of plots 650 of the insert size in the Coriell RRBS libraries treated with enzyme dilutions for 5 minutes, respectively, generated from the output sequencing data.

[0110] Referring to Figure 6A, the mean insert size in the untreated Coriell library is about 163 bp, which is within the expected range. For the treated libraries, there is a strong shift to the left with a cut off at 75 bp. The mean insert size reads as a mean output of 141 bp for the 30 minute library, 138 bp for the 120 and 240 minute libraries, but the actual mean insert size for the treated libraries is potentially about 50 bp to about 60 bp, if extrapolated. Similar results were obtained for the Hela RRBS libraries.

[0111] The data show that UNG and Endo IV co-digestion fragment the DNA. After 30 minutes co-digestion, no further change in fragment size is observed.

[0112] Referring to Figure 6B, the mean insert size in the untreated Coriell library is about 164 bp, which is within the expected range. The mean insert size reads as a mean output of 143 bp for the 1x enzyme library, 155 bp for the 1/10x enzyme library and 153 bp for the 1/25x enzyme library (1x enzyme being 1U UNG and 2U Endonuclease IV per 1μg DNA).

[0113] The data show that enzyme dilution provides longer, more suitable mean fragment sizes at an incubation time of 5 minutes.

Example 5 Methylation call results using Bismark.

[0114] **Figure 7** shows a data table of the Bismark methylation call results for the untreated and treated Hela and Coriell RRBS libraries. Bismark is an open source program that may be used to analyze methylated DNA. In this example, for the untreated Hela library, CpG methylation is about 95%, CHG methylation is 0.60%, and CHH methylation is 0.60%. For the 120 and 240 minute treated Hela libraries, the CpG methylation is about 93.2% and 95%, respectively, CHG methylation is 70% and 87%, respectively, and CHH methylation is 90.2% and 95.8%, respectively. FASTQC analysis of the GC content of the 30 minute treated Hela library indicated that the GC content (data not shown) suggested that the library may be contaminated with a second genome.

[0115] For the enzymatically-untreated, bisulfite-converted, Coriell library, CpG methylation is 50%, CHG methylation is 0.50%, and CHH methylation is 0.60%%, as would be typically expected in Whole Genome Bisfulfite-converted samples. For the 30 minute treated Coriell library, CpG methylation is 78.3% CHG methylation is 71.2%, and CHH methylation is 91.3%. For the 120 and 240 minute treated Coriell libraries, the CpG methylation is about 77.4% and 81.1%, respectively, CHG methylation is 64.8% and 74.7%, respectively, and CHH methylation is 88.3% and 93.4%, respectively.

[0116] The data show that UNG/Endo IV treatment of bisulfite-treated DNA significantly increases enrichment of CpG, CHG, and CHH methylated regions. For example, CpG methylation in the untreated (no UNG/Endo IV digestion) Coriell library is 52% compared to about 80% in treated Coriell libraries. CHG methylation in the untreated (no UNG/Endo IV digestion) Coriell library is 0.5% compared to about 71% to about 75% in treated Coriell libraries. Similarly, CHH methylation in the untreated (no UNG/Endo IV

digestion) Coriell library is 0.6% compared to about 88% to about 93% in treated Coriell libraries.

[0117] The method of the invention provides a mechanism to enrich for CHG and CHH methylated regions of a genome. The method of the invention also provides a mechanism to further enrich (compared to *Msp*I digestion) CpG methylated regions of a genome.

[0118] For the untreated (no UNG/Endo IV) Hela and Coriell libraries, the unique alignment is about 67% and about 55%, respectively, which is within the expected range. However, in the UNG/Endo IV treated libraries, the unique alignment is decreased.

[0119] The enrichment of methylation content of the RRBS libraries may also be assessed using an antibody specific for cytosine methylation (data not shown).

[0120] **Figure 8** shows a screenshot 800 of the IGV interface showing methylation patterns across a region of chromosome 8 in the untreated and 120 minute treated Coriell libraries. This region of chromosome 8 includes CpG islands. A track 810 shows the captured reads across the region of chromosome 8 for the 120 minute treated Coriell library. A track 815 shows the captured reads across the region of chromosome 8 for the untreated (no UNG/Endo IV digestion) Coriell library. In this example, one read was captured in this region of chromosome 8 for the treated Coriell library (track 810). The failure to detect methylation in this region may be due to over-fragmentation of the DNA by UNG/Endo IV digestion. For the untreated Coriell library (track 815), reads were detected, but they did not necessarily correlate with known CpG islands.

[0121] **Figure 9** shows a screenshot of the IGV interface showing reads captured across a region of chromosome 5 in the Coriell libraries treated with different enzyme dilutions at 5 minute incubation times. The IGV data demonstrate that increasing enzyme concentration directly correlates with lower coverage of the genome, as expected. Enzyme digestion should enrich only methylated portions of the genome. Concentration of the enzyme may be optimized for maximizing methylation call detection and CpG, CHH, and CHG site enrichment.

Example 6 Enrichment in CpG sites in UNG/Endo IV treated libraries.

[0122] **Figures 10A and 10B** show a bar graph of Coriell CpG sites and a bar graph of Hela CpG sites in the RRBS libraries, respectively. Bar graphs and show the number of CpG sites with greater than 10 reads within each library (untreated (no UNG/Endo IV digestion), 30, 120, and 240 minute treated). For both the Coriell and the Hela untreated libraries the number of CpG sites (> 10 reads) is about 2000 to about 3000. In the 30 minute treated Coriell library, the number of CpG sites (> 10 reads) is increased about 8 to 10 fold. The number of CpG sites (> 10 reads) is further increased in the 120 minute (about 24,000) and 240 minute (about 32,000) treated Coriell libraries. Similarly, the number of CpG sites (> 10 reads) is increased in the 120 minute (about 32,000) and 240 minute (45,000) treated Hela libraries. The 30 minute treated Hela library is potentially contaminated and the data invalid.

[0123] The data show that CpG sites with greater than 10 reads is significantly enriched in UNG/Endo IV treated samples. A similar enrichment is also noted in CHG and CHH sites with greater than 10 reads (data not shown).

Example 7 CpG and CHH/CHG enrichment related to enzyme dilution.

[0124] **Figure 11** shows results on CpG and CHH (white bar)/CHG (black bar) enrichment related to enzyme dilution. The results show that optimizing for dilution demonstrates up to 20-fold enrichment of CpG sites and CHH/CHG sites (not based on methylation) due to enzymatic cleavage and improved unique alignment. Therefore, for best data, the genome may be partially digested for optimal unique alignment, and enrichment of non-cleaved regions that contain methylated residues. These results demonstrate that the strength of this method is that it does not discriminate between CpG/CHH/CHG, because the method only requires a concentration of methylated residues, regardless of surrounding sequence context.

Example 8 Methylation Call Correlation of Coriell 18507 HiSeq versus UNG/Endo IV.

[0125] **Figure 12** shows correlation of methylation calls between libraries prepped with UNG/Endo IV and EpiGnome vs. EpiGnome-only. High enzyme

concentration is excellent for correlating 100% methylated reads. For more refined analysis, lower concentration enzyme provides more nuanced, better correlation output to previous EpiGnome runs. The 10x read depth provides adequate correlation in the original analysis. With higher read depth (50x), the methylation call correlation becomes outstanding. Selecting the correct dilution relies upon using a balanced approach. After examining all of the crucial indicators from sequencing, the best data becomes available at either the 1/10x or 1/25x dilution of UNG and Endonuclease IV with a 5 minute incubation. This dilution provides:

a) Optimal CpG enrichment;

b) Enhancement of Methylation Calls;

c) Best correlation with previous methylation call data; and

d) Enrichment also present for CHG and CHH context.

Example 9 *Msp*I RRBS versus UNG/Endo IV RRBS.

[0126] The advantage of the UNG/Endo IV method is best demonstrated by comparing the CpG site enrichment of this method to other examples that used matching conditions and read depth (e.g. 1 lane of GA). In Lee et al., *Biol. Proced. Online* 16(1):1 (2014), they covered 1.8M CpG sites at 10x read depth with the *Msp*I, RRBS method. In Varley et al., *Genome Res.* 23(3): 555-67 (2013), they covered 1.2M CpG sites. Jin et al., *PLoS Genet.* 9(6): e1003515 (2013) covered 1.7M CpG sites. Using the method disclosed herein, with a 10x read depth, from a single lane of GA, 2.6M CpG sites were covered.

[0127] **Figure 13** shows that the sites detected by the UNG/Endo IV method are almost completely different than those detected by *Msp*I RRBS. This is a tremendous expansion of the sites detected by the new method, when compared to the currently available strategy. Use of UNG/Endo IV RRBS demonstrates significant expansion of CpG sites not containing a *Msp*I recognition sequence. 2.5M CpG sites, encompassing 1.4M methylated CpG sites, would have been missed by *Msp*I RRBS that are detected with UNG/Endo IV. The results show that the scope of *Msp*I RRBS and UNG/Endo IV method are entirely different.

[0128] The foregoing detailed description of embodiments refers to the accompanying drawings, which illustrate specific embodiments of the present disclosure. Other embodiments having different structures and operations do not depart from the scope of the present disclosure. The term "the invention" or the like is used with reference to certain specific examples of the many alternative aspects or embodiments of the applicants' invention set forth in this specification, and neither its use nor its absence is intended to limit the scope of the applicants' invention or the scope of the claims. This specification is divided into sections for the convenience of the reader only. Headings should not be construed as limiting of the scope of the invention. The definitions are intended as a part of the description of the invention. It will be understood that various details of the present invention may be changed without departing from the scope of the present invention. Furthermore, the foregoing description is for the purpose of illustration only, and not for the purpose of limitation.

[0129] All publications, including patent documents and scientific articles, referred to in this application and the bibliography and attachments are incorporated by reference for the referenced materials and in their entireties for all purposes to the same extent as if each individual publication were individually incorporated by reference.

[0130] Citation of the above publications or documents is not intended as an admission that any of the foregoing is pertinent prior art, nor does it constitute any admission as to the contents or date of these publications or documents.

[0131] Although the present invention has been fully described in connection with embodiments thereof with reference to the accompanying drawings, it is to be noted that various changes and modifications will become apparent to those skilled in the art. Such changes and modifications are to be understood as being included within the scope of the present invention. The various embodiments of the invention should be understood that they have been presented by way of example only, and not by way of limitation. Likewise, the various diagrams may depict an example architectural or other configuration for the invention, which is done to aid in understanding the features and functionality that can be included in the invention. The invention is not restricted to the illustrated example architectures or configurations, but can be implemented using a variety of alternative architectures and configurations. Additionally, although the invention is described above in

terms of various exemplary embodiments and implementations, it should be understood that the various features and functionality described in one or more of the individual embodiments are not limited in their applicability to the particular embodiment with which they are described. They instead can, be applied, alone or in some combination, to one or more of the other embodiments of the invention, whether or not such embodiments are described, and whether or not such features are presented as being a part of a described embodiment. Thus the breadth and scope of the invention should not be limited by any of the above-described exemplary embodiments.

[0132] Terms and phrases used in this document, and embodiments thereof, unless otherwise expressly stated, should be construed as open ended as opposed to limiting. As examples of the foregoing: the term "including" should be read as meaning "including, without limitation" or the like; the term "example" is used to provide exemplary instances of the item in discussion, not an exhaustive or limiting list thereof; and adjectives such as "conventional," "traditional," "normal," "standard," "known", and terms of similar meaning, should not be construed as limiting the item described to a given time period, or to an item available as of a given time. But instead these terms should be read to encompass conventional, traditional, normal, or standard technologies that may be available, known now, or at any time in the future. Likewise, a group of items linked with the conjunction "and" should not be read as requiring that each and every one of those items be present in the grouping, but rather should be read as "and/or" unless apparent from the context or expressly stated otherwise. Similarly, a group of items linked with the conjunction "or" should not be read as requiring mutual exclusivity among that group, but rather should also be read as "and/or" unless it is apparent from the context or expressly stated otherwise. Furthermore, although items, elements or components of the invention may be described or claimed in the singular, the plural is contemplated to be within the scope thereof unless limitation to the singular is explicitly stated. For example, "at least one" may refer to a single or plural and is not limited to either. The presence of broadening words and phrases such as "one or more," "at least," "but not limited to", or other like phrases in some instances shall not be read to mean that the narrower case is intended or required in instances where such broadening phrases may be absent.

WHAT IS CLAIMED IS:

1. A method for preparing a sample for sequencing, comprising:

treating nucleic acid molecules in the sample to convert at least a portion of unmethylated cytosine residues into uracil residues; and

cleaving the nucleic acid molecules at at least a portion of the uracil residues to obtain nucleic acid fragments.

2. The method of Claim 1, wherein the sample is treated with bisulfite to convert at least a portion of unmethylated cytosine residues into uracil residues.

3. The method of any one of Claim 1 or 2, wherein cleaving the nucleic acid molecules at the at least a portion of uracil residues comprises treating the nucleic acid molecules with a uracil DNA glycosylase resulting in a plurality of abasic residues in place of at least a portion of uracil residues.

4. The method of Claim 3, further comprising treating the nucleic acid molecules with an endonuclease after treating with said uracil DNA glycosylase, wherein said endonuclease cleaves said nucleic acid molecules at at least a portion of the plurality of abasic residues.

5. The method of Claim 4, wherein said uracil DNA glycosylase and said endonuclease are present in a single reaction mixture.

6. The method of Claim 3, further comprising treating the nucleic acid molecules with heat after treating with said uracil DNA glycosylase, wherein said nucleic acid molecules are broken at at least a portion of the plurality of abasic residues.

7. The method of Claim 6, wherein the treatments with said uracil DNA glycosylase and heat are conducted simultaneously.

8. The method of any one of Claims 3-7, wherein the treatment with said uracil DNA glycosylase is conducted for about 1 minute to about 240 minutes.

9. The method of Claim 8, wherein the treatment with said uracil DNA glycosylase is conducted for about 1 minute to about 10 minutes.

10. The method of Claim 8, wherein the treatment with said uracil DNA glycosylase is conducted for about 1 minute to about 5 minutes.

11. The method of any one of Claims 6-10, wherein the treatment with heat is conducted at 70 °C.

12. The method of any one of Claims 3-11, wherein said uracil DNA glycosylase is human uracil-DNA glycosylase (UNG).

13. The method of Claim 12, wherein said UNG is at a concentration of about 0.02 U/1 μg of DNA to about 1 U/1 μg of DNA.

14. The method of Claim 12, wherein said UNG is at a concentration of about 0.04 U/1 μg of DNA to about 0.2 U/1 μg of DNA.

15. The method of Claim 12, wherein said UNG is at a concentration of about 0.04 U/1 μg of DNA to about 0.1 U/1 μg of DNA.

16. The method of any one of Claims 4-15, wherein said endonuclease is endonuclease IV (Endo IV).

17. The method of Claim 16, wherein said Endo IV is at a concentration of about 2 U/1 μg of DNA.

18. The method of Claim 16, wherein said Endo IV is at a concentration of about 0.2 U/1 μg of DNA.

19. The method of Claim 16, wherein said Endo IV is at a concentration of about 0.08 U/1 μg of DNA.

20. The method of any one of Claims 1-19, further comprising selecting the nucleic acid fragments based on their size.

21. The method of Claim 20, wherein said selecting the nucleic acid fragments based on their size comprises a bead-based method.

22. The method of any one of Claims 1 to 21, wherein the portion of unmethylated cytosine residues is at least 50%, 60%, 70%, 80%, 90%, 95%, or 100%.

23. The method of any one of Claims 1 to 22, wherein the portion of the uracil residues is at least 50%, 60%, 70%, 80%, 90%, 95%, or 100%.

24. The method of any one of Claims 4 to 23, wherein the portion of the plurality of abasic residues is at least 50%, 60%, 70%, 80%, 90%, 95%, or 100%.

25. A method for sequencing nucleic acid fragments, comprising:

obtaining the nucleic acid fragments from a sample treated with a method of any one of Claims 1-24; and

sequencing the nucleic acid fragments.

26. The method of Claim 25, further comprising amplifying said nucleic acid fragments by PCR.

27. The method of Claim 26, wherein said PCR comprises about 10 cycles to about 15 cycles.

28. The method of Claim 26, wherein said PCR comprises about 12 cycles.

29. The method of any one of Claims 25-28, wherein said sequencing is sequencing by synthesis (SBS).

30. The method of any one of Claims 25-29, wherein the sequences of said nucleic acid fragments are compared to a reference sequence.

31. The method of any one of Claims 25-30, comprising identifying methylated cytosines in said nucleic acid fragments.

32. The method of Claim 30 or 31, wherein the reference sequence comprises a methylome.

33. The method of any one of Claims 30-32, wherein said methylated cytosines are identified using Bismark.

34. The method of any one of Claims 30-33, wherein said methylated cytosines are analyzed using an integrated genome viewer.

35. The method of any one of Claims 30-34, wherein said methylated cytosines comprise methylated CpG, methylated CHG and/or methylated CHH.

36. The method of any one of Claims 30-35, wherein the methylation calls are significantly enriched in comparison to the methylation calls in a sample treated with bisulfite to convert unmethylated cytosine residues into uracil residues, but without cleaving the nucleic acid molecules at the uracil residues.

37. The method of Claim 36, wherein the enriched methylation calls comprise methylated CpG.

38. The method of Claim 36, wherein the enriched methylation calls comprise methylated CHG and/or methylated CHH.

39. The method of any one of Claims 36-38, wherein the methylation calls comprise sites with a read depth of greater than about 10.

40. The method of Claim 30, wherein the cytosine sites are significantly enriched in comparison to the cytosine sites in a sample treated with bisulfite to convert unmethylated

cytosine residues into uracil residues, but without cleaving the nucleic acid molecules at the uracil residues.

41. The method of Claim 40, wherein the cytosine sites comprise CpG.

42. The method of Claim 40, wherein the cytosine sites comprise CHG and/or CHH.

43. A nucleic acid sample comprising fragmented nucleic acid molecules wherein substantially all cytosine residues in said fragmented nucleic acid molecules are methylated.

44. The nucleic acid sample of Claim 43, wherein the first three residues of some of said fragmented nucleic acid molecules are not cytosine-guanine-guanine and/or the last four residues of some of said fragmented nucleic acid molecules are not thymine-cytosine-guanine-guanine.

45. The nucleic acid sample of Claim 43 or 44, wherein the fragmented nucleic acid molecules comprise an apurinic/apyrimidinic site at the 5' or 3' end.

46. The nucleic acid sample of any one of Claims 43-45, wherein the fragmented nucleic acid molecules have a size from about 10 bp to about 2,000 bp.

47. The nucleic acid sample of any one of Claims 43-45, wherein the fragmented nucleic acid molecules have a size from about 50 bp to about 500 bp.

48. The nucleic acid sample of any one of Claims 43-45, wherein the fragmented nucleic acid molecules have a size from about 100 bp to about 200 bp.

49. The nucleic acid sample of any one of Claims 43-48, wherein the fragmented nucleic acid molecules have a mean size of about 140 bp to about 170 bp.

50. The nucleic acid sample of any one of Claims 43-49, wherein a portion of the CpG sites in said fragmented nucleic acid molecules are methylated.

51. The nucleic acid sample of any one of Claims 43-50, wherein at least about 0.1% of the CHG sites in said fragmented nucleic acid molecules are methylated.

52. The nucleic acid sample of any one of Claims 43-51, wherein at least about 0.1% of the CHH sites in said fragmented nucleic acid molecules are methylated.

53. A population of nucleic acid fragments resulting from a sample treated with a method of any one of Claims 1-24.

54. A kit comprising at least one container means, wherein the at least one container means comprises a reagent that cleaves a nucleic acid molecule at a uracil residue.

55. The kit of Claim 54, wherein the reagent comprises UNG and endonuclease IV (Endo IV).

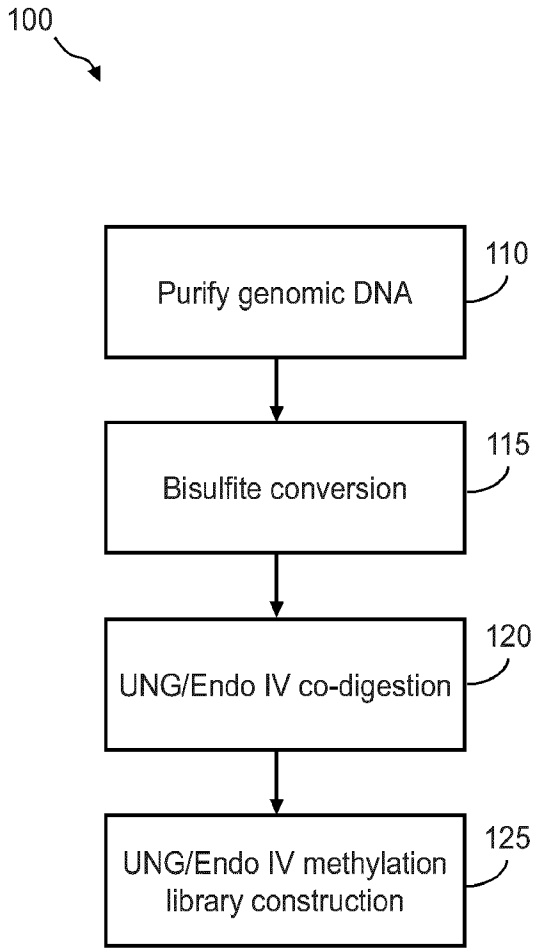
100
Purify genomic DNA
110
Bisulfite conversion
115
UNG/Endo IV co-digestion
120
UNG/Endo IV methylation library construction
125

Figure 1

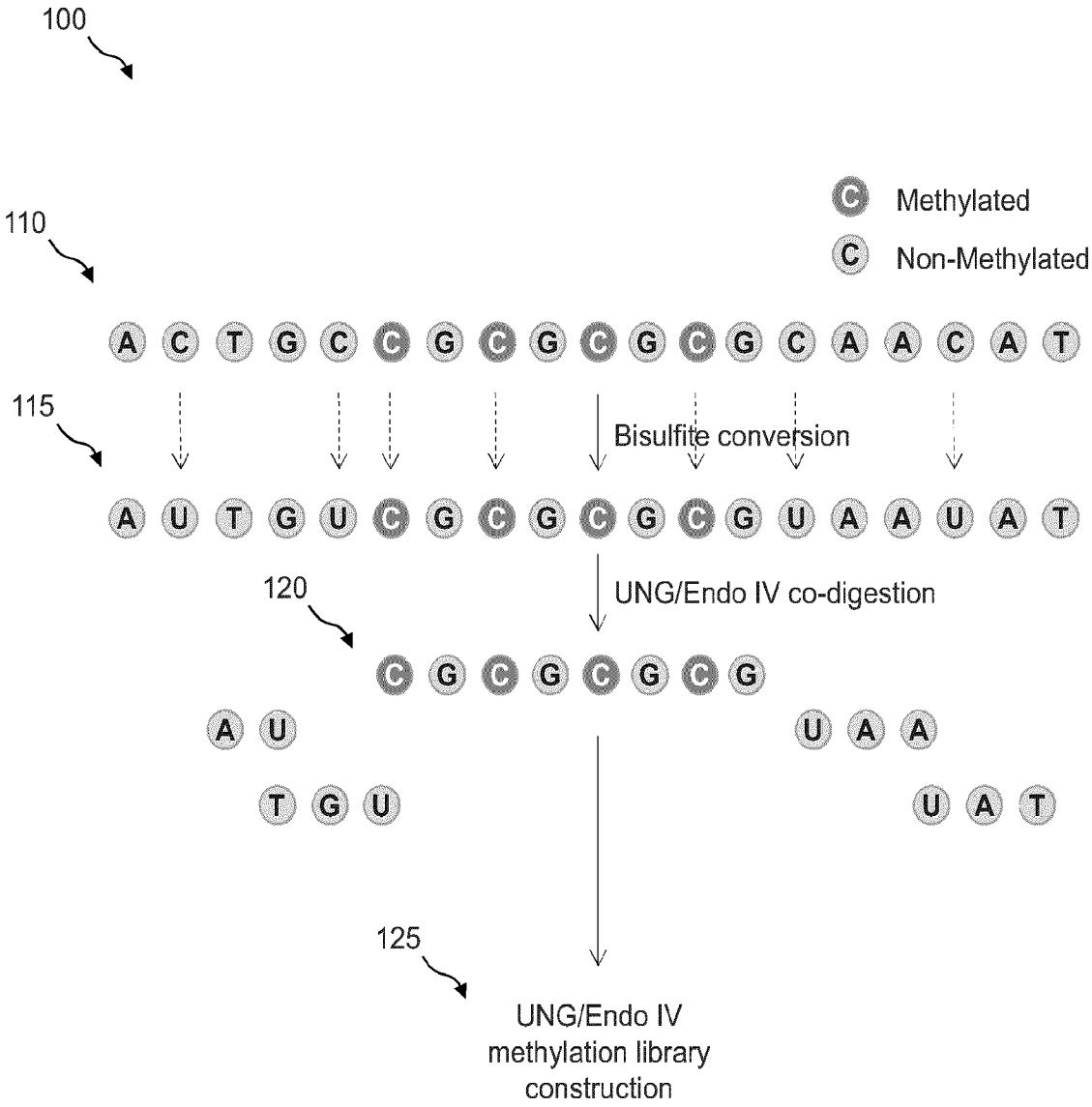
100
110
C Methylated
C Non-Methylated
A C T G C C G C G C G C G C A A C A T
115
Bisulfite conversion
A U T G U C G C G C G C G U A A U A T
120
UNG/Endo IV co-digestion
C G C G C G C G
A U
U A A
T G U
U A T
125
UNG/Endo IV
methylation library
construction

Figure 2

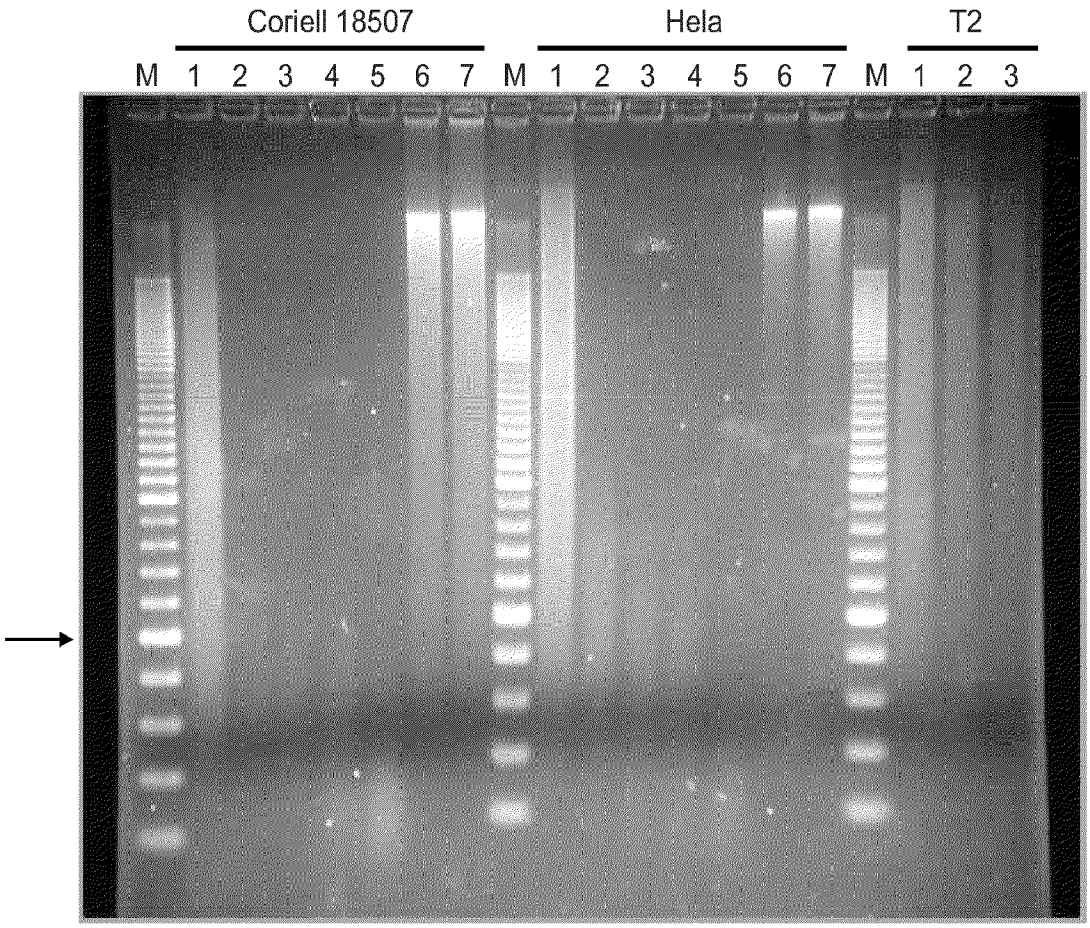

Figure 3

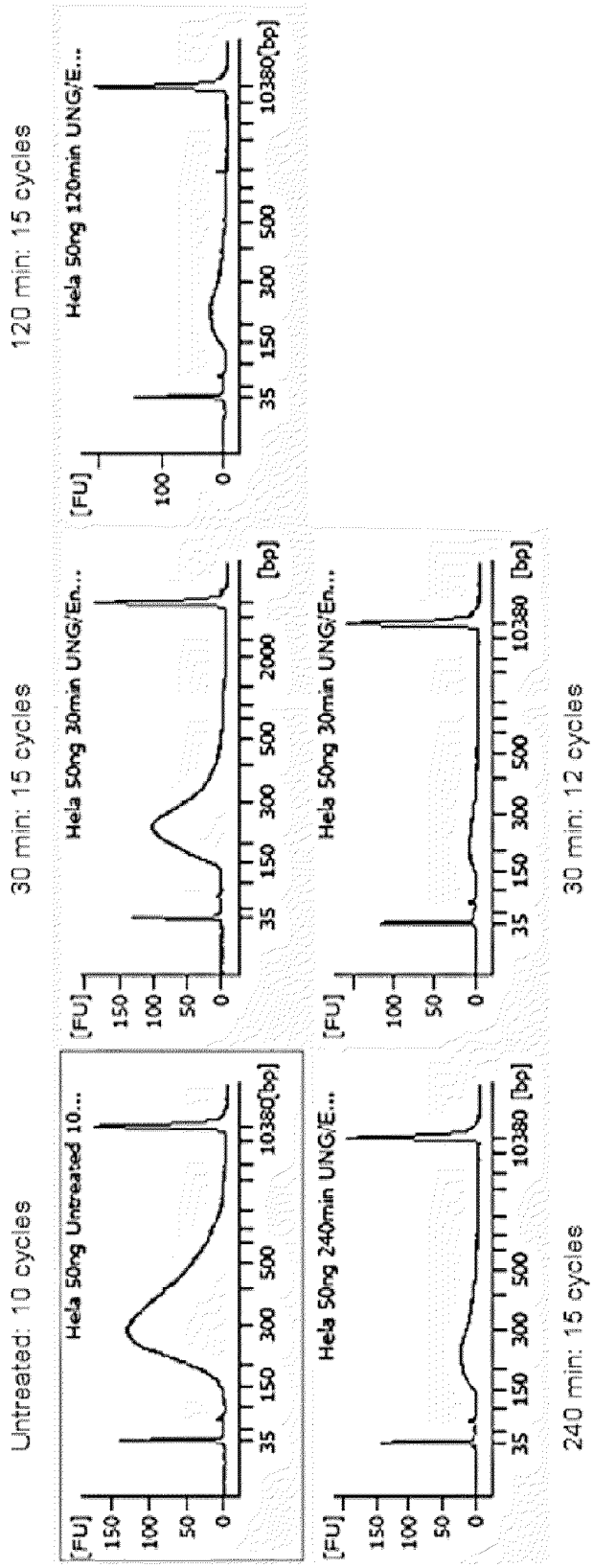
Untreated: 10 cycles
30 min: 15 cycles
120 min: 15 cycles
Hela 50ng Untreated 10...
Hela 50ng 30min UNG/En...
Hela 50ng 120min UNG/E...
Hela 50ng 240min UNG/E...
Hela 50ng 30min UNG/En...
[FU]
[bp]
10380[bp]
240 min: 15 cycles
30 min: 12 cycles

Figure 4

Status	07-Sequencing Complete
Requested by	Burgess,Joshua
Seq. Project Title	DolosV2
Urgency	Normal
Run Purpose	Reasearch
Sequencer Type	GA-GenomeAnalyzer
Run Type	PE Paired-End (no index read)
Cycles/read	76
Read Tiles	Quick 24 tiles
Lanes per Sample	1 Sample/Lane
Seq. Read Primers (R1,R2,Idx)	ScriptSeq v2 (HP6,HP7,HP8)
Source Nucleic Acid	Human1
Fastq	Default:multiple fastqs-no trim
Primary Analyst	Burgess,Joshua
Post Run Analysis	Primary Analysis Metrics (PAM)
Comparison Samples for Diff.Exp.	N/A
Sample ID	1)Hela Untreated 357bp 2.66 2)Hela 30 min UNG/Endo 275bp 1.76 3)Hela 120 min UNG/Endo 269bp 0.453 4)Hela 240 min UNG/Endo 296bp 0.545 5)Coriell Untreated 363bp 1.81 6)Coriell 30 min UNG/Endo 325bp 1.86 7)Coriell 120 min UNG/Endo 325bp 0.481 8)Coriell 240 min UNG/Endo 314bp 0.523 Notebook Reference: ILMN 1634-p177
Additional Communications	Original samples replaced with new (reduced PCR cycles) 66B4YAAXX Discarded 140321_HARRY_00070_FC66B7LAAXX FAIL. Error Rates Last 10 cycles of Read 2 quality diminished (Median Qscore less than 30) REPEAT
Sample Disposition	Please return any unused sample(s)
Run Folder	140327_HARRY_00072_FC66B7BAAXX

Fig. 5A

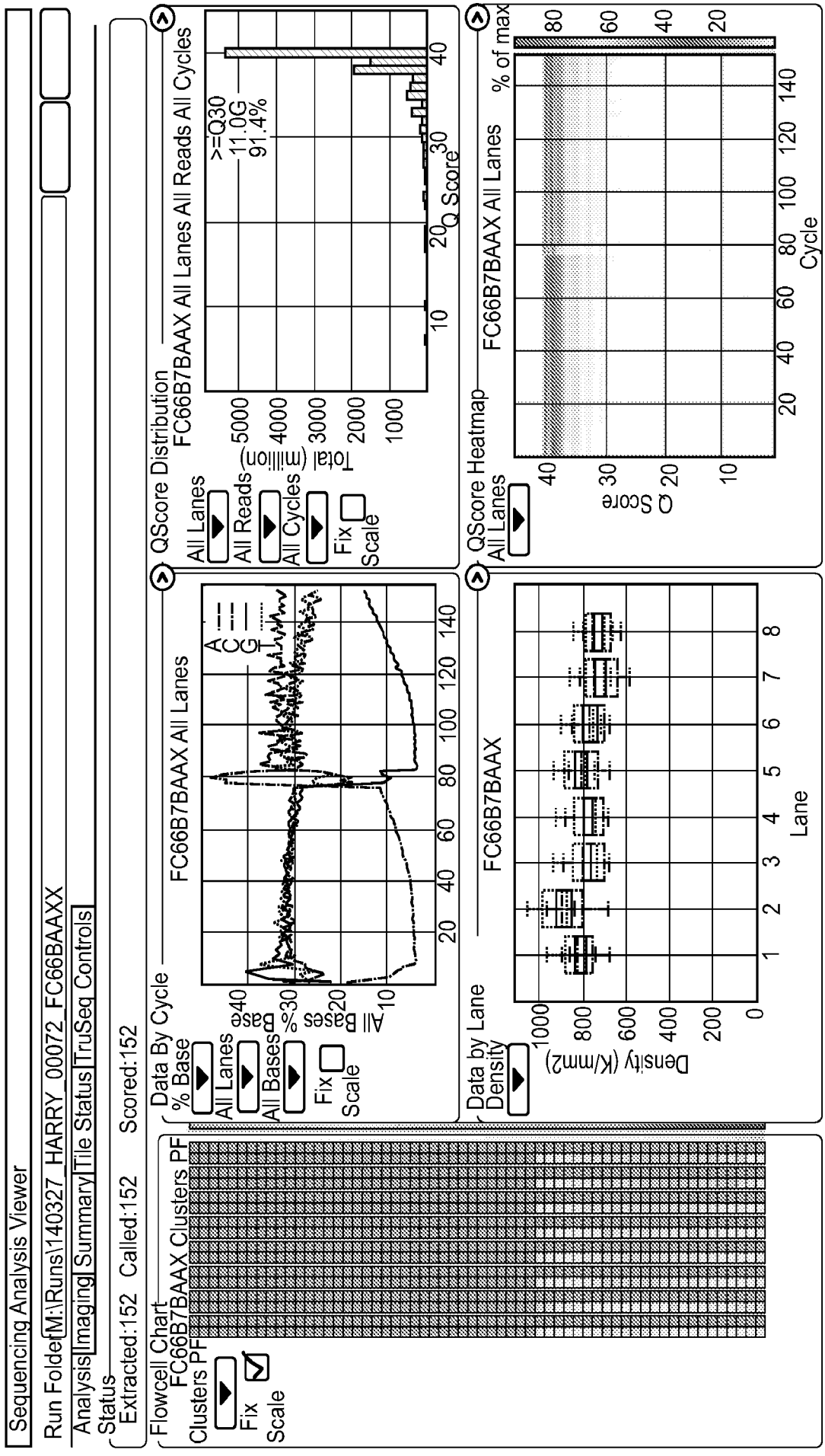
Sequencing Analysis Viewer
Run Folder M:\Runs\140327_HARRY_00072_FC66BAAXX
Analysis Imaging Summary Tile Status TruSeq Controls
Status
Extracted:152 Called:152 Scored:152
Flowcell Chart
FC66B7BAAX Clusters PF
Clusters PF
Fix
Scale
Data By Cycle
FC66B7BAAX All Lanes
% Base
All Lanes
All Bases
Fix
Scale
A
C
G
T
All Bases % Base
Data by Lane
Density
FC66B7BAAX
Density (K/mm2)
Lane
QScore Distribution
FC66B7BAAX All Lanes All Reads All Cycles
All Lanes
All Reads
All Cycles
Fix
Scale
>=Q30
11.0G
91.4%
Total (million)
Q Score
QScore Heatmap
FC66B7BAAX All Lanes
% of max
Q Score
Cycle

Fig. 5B

Run Summary

Level	Yield Total (G)	Projected Total Yield	Yield Perfect (G)	Yield<= 3 errors (G)	Aligned (%)	% Perfect [Num Cycles]	% <=3 errors [Num Cycles]	Error Rate (%)	Intensity Cycle 1	% Intensity Cycle 20	% >=Q30
Read 1	6.0	6.0	0.0	0.0	0.23	67.3[75]	97.9[75]	0.85	1928	83.6	94.1
Read 2	6.0	6.0	0.0	0.0	0.19	34.2[75]	91.6[75]	1.98	2121	84.8	88.6
Total	11.9	11.9	0.0	0.0	0.21	52.5	95.1	1.41	2024	84.2	91.4

Read 1

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >=Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle(%)	Error Rate 75 cycle(%)
1	25	834 +/- 56	93.59 +/- 1.40	0.190/0.254	10.88	10.18	94.0	0.8	75	0.2 +/- 0.0	1.01 +/- 0.46	0.60 +/- 0.63	1.01 +/- 0.46
2	25	931 +/- 65	91.88 +/- 3.31	0.185/0.260	12.15	11.16	93.2	0.8	75	0.2 +/- 0.0	0.94 +/- 0.55	0.67 +/- 0.49	0.94 +/- 0.55
3	25	801 +/- 62	94.88 +/- 1.17	0.189/0.282	10.45	9.91	95.6	0.7	75	0.2 +/- 0.0	0.49 +/- 0.35	0.46 +/- 0.45	0.49 +/- 0.35
4	25	796 +/- 63	95.53 +/- 1.02	0.196/0.303	10.39	9.92	95.8	0.7	75	0.2 +/- 0.0	0.48 +/- 0.32	0.35 +/- 0.37	0.48 +/- 0.32
5	25	836 +/- 56	92.58 +/- 4.04	0.198/0.274	10.91	10.09	93.0	0.8	75	0.2 +/- 0.0	1.29 +/- 1.02	0.86 +/- 0.58	1.29 +/- 1.02
6	25	798 +/- 55	93.28 +/- 2.65	0.195/0.298	10.41	9.70	93.0	0.7	75	0.2 +/- 0.0	1.04 +/- 0.78	0.58 +/- 0.50	1.04 +/- 0.78
7	25	752 +/- 54	93.24 +/- 3.92	0.196/0.296	9.81	9.16	93.1	0.7	75	0.2 +/- 0.0	1.13 +/- 0.68	0.88 +/- 0.62	1.13 +/- 0.68
8	25	748 +/- 53	95.68 +/- 0.55	0.200/0.317	9.76	9.33	95.6	0.7	75	0.3 +/- 0.0	0.44 +/- 0.24	0.26 +/- 0.21	0.44 +/- 0.24

Read 2

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >=Q30	Yield (G)	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle(%)	Error Rate 75 cycle(%)
1	25	834 +/- 56	93.59 +/- 1.40	0.171/0.264	10.88	10.18	90.6	0.8	75	0.2 +/- 0.0	2.07 +/- 0.74	1.56 +/- 0.64	2.07 +/- 0.74
2	25	931 +/- 65	91.88 +/- 3.31	0.197/0.254	12.15	11.16	88.1	0.8	75	0.1 +/- 0.0	2.09 +/- 0.59	1.57 +/- 0.52	2.09 +/- 0.59
3	25	801 +/- 62	94.88 +/- 1.17	0.208/0.244	10.45	9.91	89.9	0.7	75	0.2 +/- 0.0	1.68 +/- 0.59	1.54 +/- 0.70	1.68 +/- 0.59
4	25	796 +/- 63	95.53 +/- 1.02	0.205/0.265	10.39	9.92	88.9	0.7	75	0.2 +/- 0.0	2.03 +/- 0.76	1.72 +/- 0.81	2.03 +/- 0.76
5	25	836 +/- 56	92.58 +/- 4.04	0.229/0.289	10.91	10.09	87.9	0.8	75	0.1 +/- 0.0	2.32 +/- 1.18	1.82 +/- 0.95	2.32 +/- 1.18
6	25	798 +/- 55	93.28 +/- 2.65	0.236/0.272	10.41	9.70	85.2	0.7	75	0.2 +/- 0.1	2.24 +/- 1.25	1.79 +/- 1.21	2.24 +/- 1.25
7	25	752 +/- 54	93.24 +/- 3.92	0.233/0.286	9.81	9.16	88.2	0.7	75	0.2 +/- 0.0	1.84 +/- 0.71	1.52 +/- 0.65	1.84 +/- 0.71
8	25	748 +/- 53	95.68 +/- 0.55	0.194/0.255	9.76	9.33	89.8	0.7	75	0.3 +/- 0.0	1.67 +/- 0.87	1.42 +/- 0.90	1.67 +/- 0.87

Fig. 5C

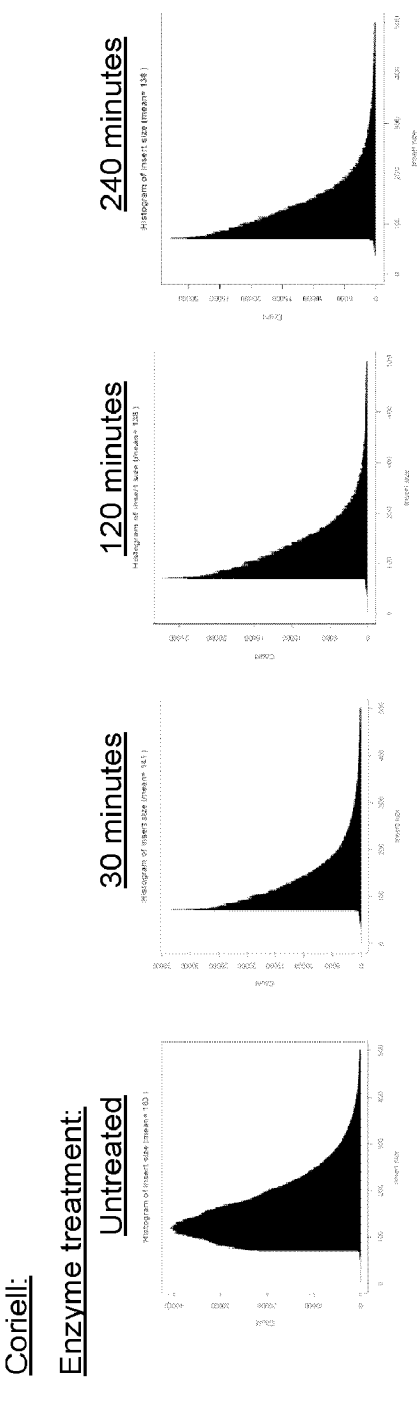

Figure 6A

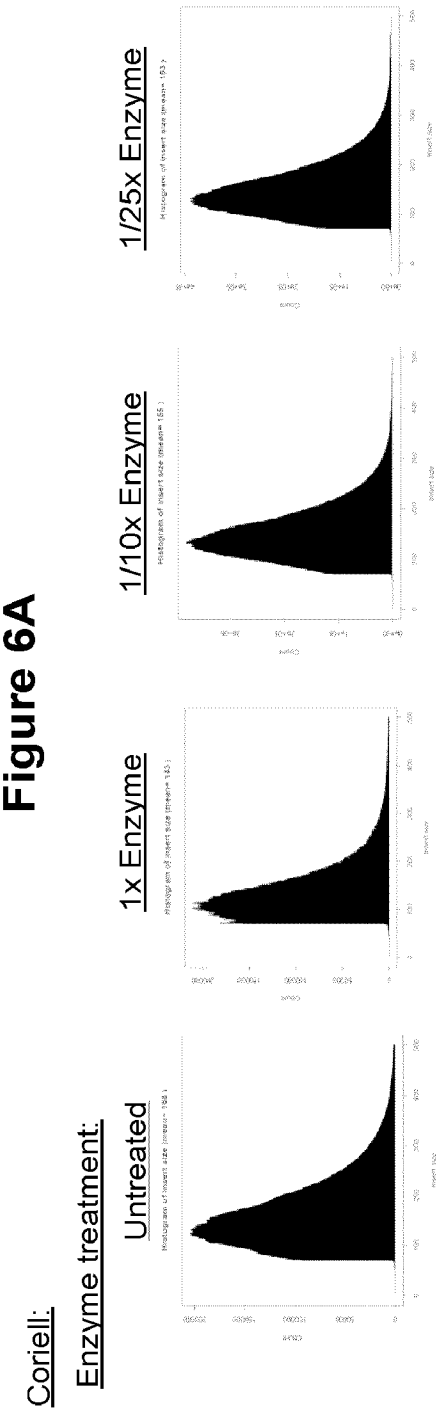

Figure 6B

Hela		Sequencer	DNA Source	Preparation	Unique Alignment	Notes	CpG Methylation	CHg Methylation	CHH Methylation	Insert Size
EpiGnome	Sample 1	24 tile GA	Hela	No treatment	66.90%		94.80%	0.60%	0.60%	164bp
	Sample 2	24 tile GA	Hela	30 min enzymes	38.00%	GC content off in FASTQC	93.50%	3.50%	4.80%	122bp
	Sample 3	24 tile GA	Hela	120 min enzymes	17.50%		93.20%	70.00%	90.20%	124bp
	Sample 4	24 tile GA	Hela	240 min enzymes	20.60%		95.00%	87.00%	95.80%	143bp
Coriell										
EpiGnome	Sample 5	24 tile GA	Coriell	No treatment	54.60%		52.00%	0.50%	0.60%	163bp
	Sample 6	24 tile GA	Coriell	30 min enzymes	21.30%		78.30%	71.20%	91.30%	141bp
	Sample 7	24 tile GA	Coriell	120 min enzymes	19.10%		77.40%	64.80%	88.30%	138bp
	Sample 8	24 tile GA	Coriell	240 min enzymes	24.00%		81.10%	74.70%	93.40%	138bp

Fig. 7

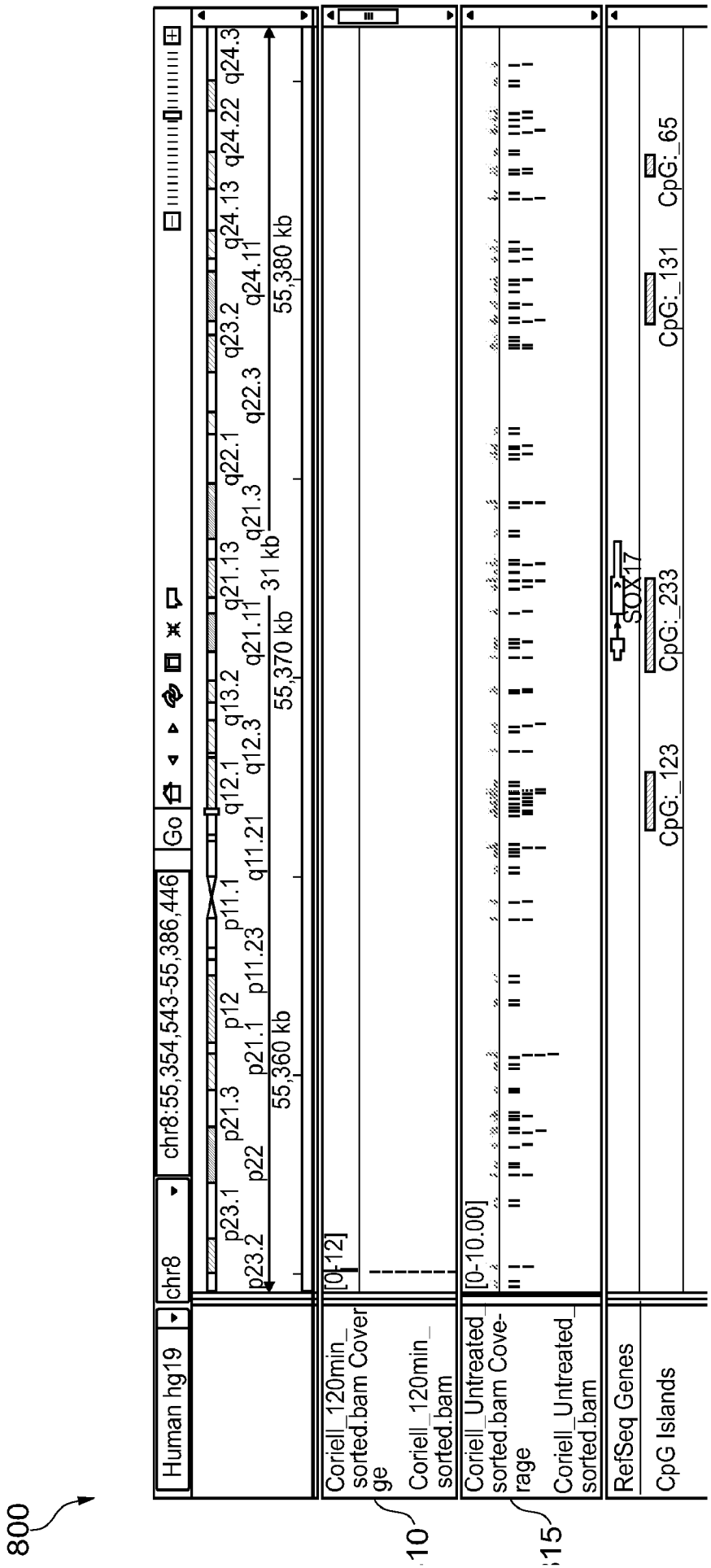
800
Human hg19
chr8
chr8:55,354,543-55,386,446
Go
p23.1
p23.2
p22
p21.3
p21.1
p12
p11.23
p11.1
q11.21
q12.1
q12.3
q13.2
q21.11
q21.13
q21.3
q22.1
q22.3
q23.2
q24.11
q24.13
q24.22
q24.3
55,360 kb
55,370 kb
31 kb
55,380 kb
[0-12]
[0-10.00]
Coriell_120min_sorted.bam Coverage
Coriell_120min_sorted.bam
Coriell_Untreated_sorted.bam Coverage
Coriell_Untreated_sorted.bam
RefSeq Genes
CpG Islands
SOX17
CpG:_123
CpG:_233
CpG:_131
CpG:_65
10
15

Fig. 8

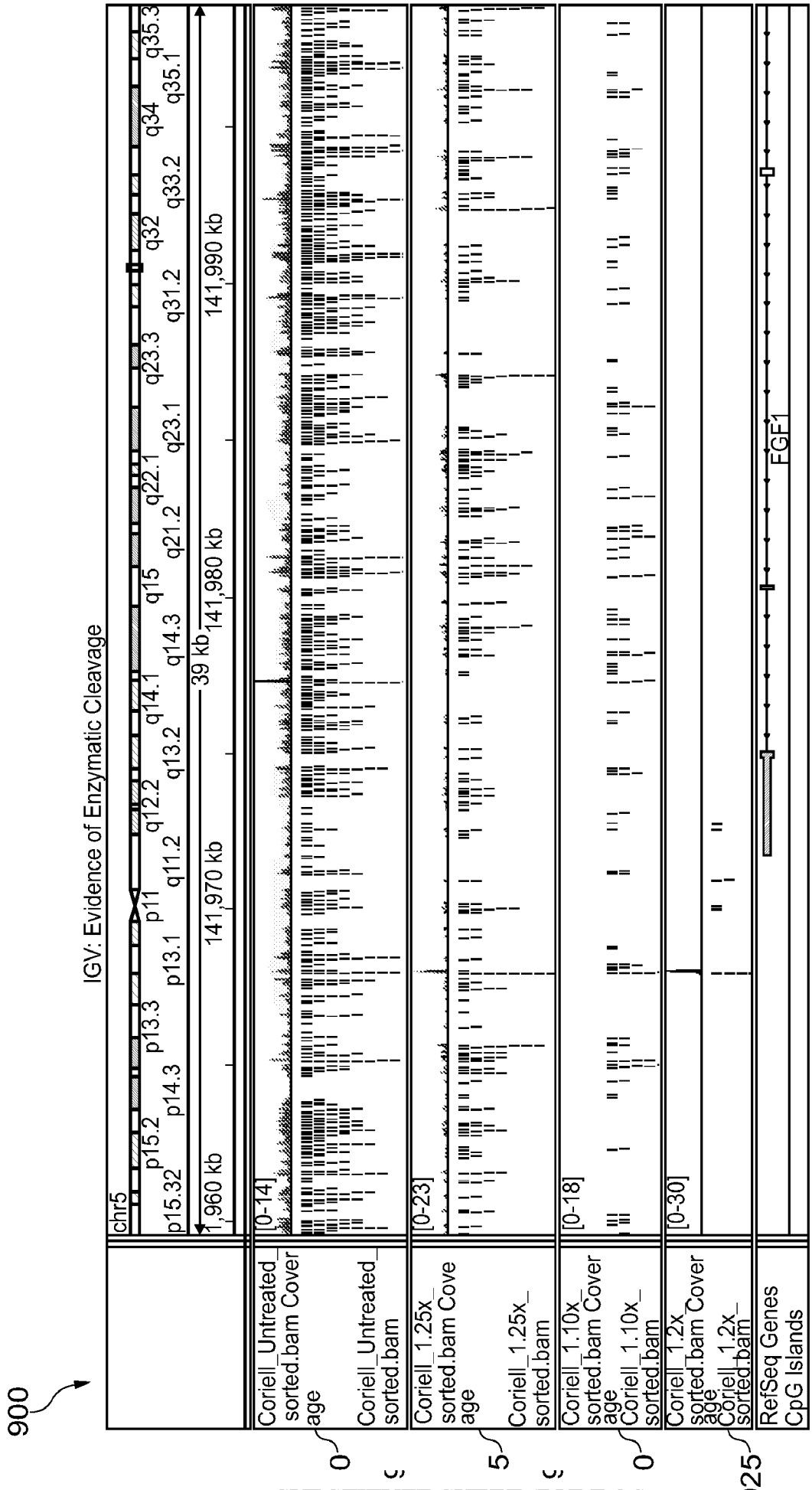
900
IGV: Evidence of Enzymatic Cleavage
chr5
p15.32
p15.2
p14.3
p13.3
p13.1
p11
q11.2
q12.2
q13.2
q14.1
q14.3
q15
q21.2
q22.1
q23.1
q23.3
q31.2
q32
q33.2
q34
q35.1
q35.3
1,960 kb
141,970 kb
39 kb
141,980 kb
141,990 kb
Coriell_Untreated_sorted.bam Coverage
Coriell_Untreated_sorted.bam
[0-14]
0
Coriell_1.25x_sorted.bam Coverage
Coriell_1.25x_sorted.bam
[0-23]
5
Coriell_1.10x_sorted.bam Coverage
Coriell_1.10x_sorted.bam
[0-18]
0
Coriell_1.2x_sorted.bam Coverage
Coriell_1.2x_sorted.bam
[0-30]
925
RefSeq Genes
CpG Islands
FGF1

Fig. 9

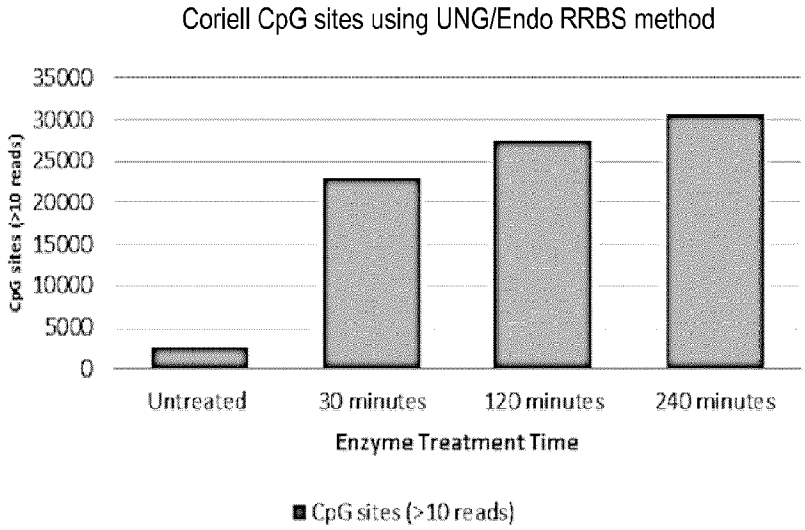

Figure 10A

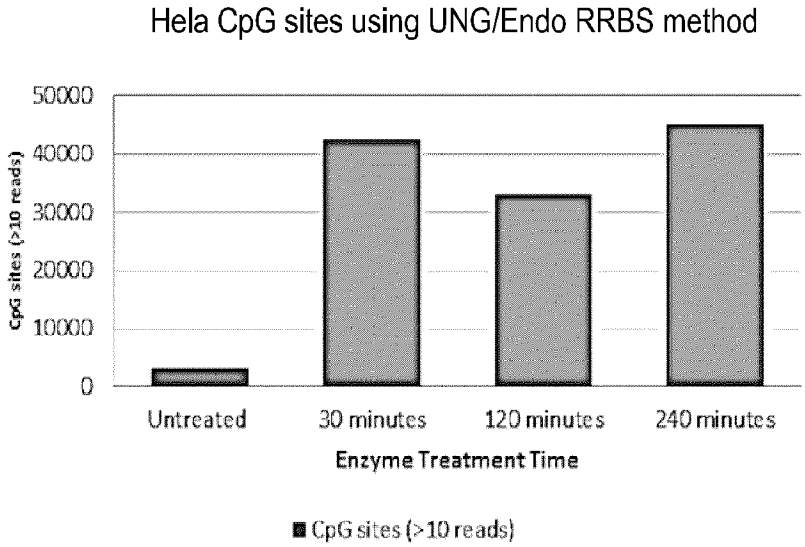

Figure 10B

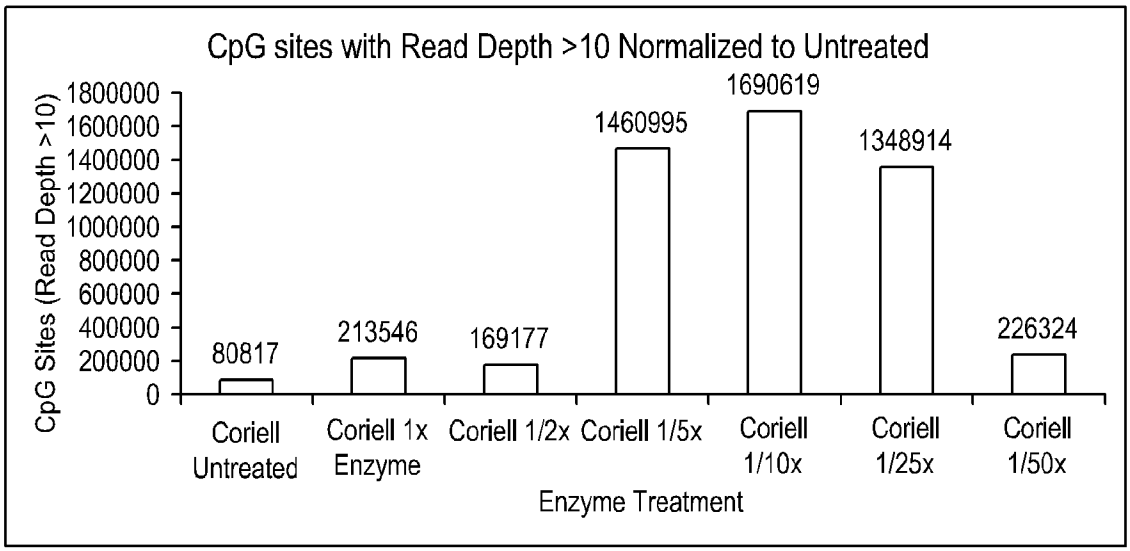

Figure 11A

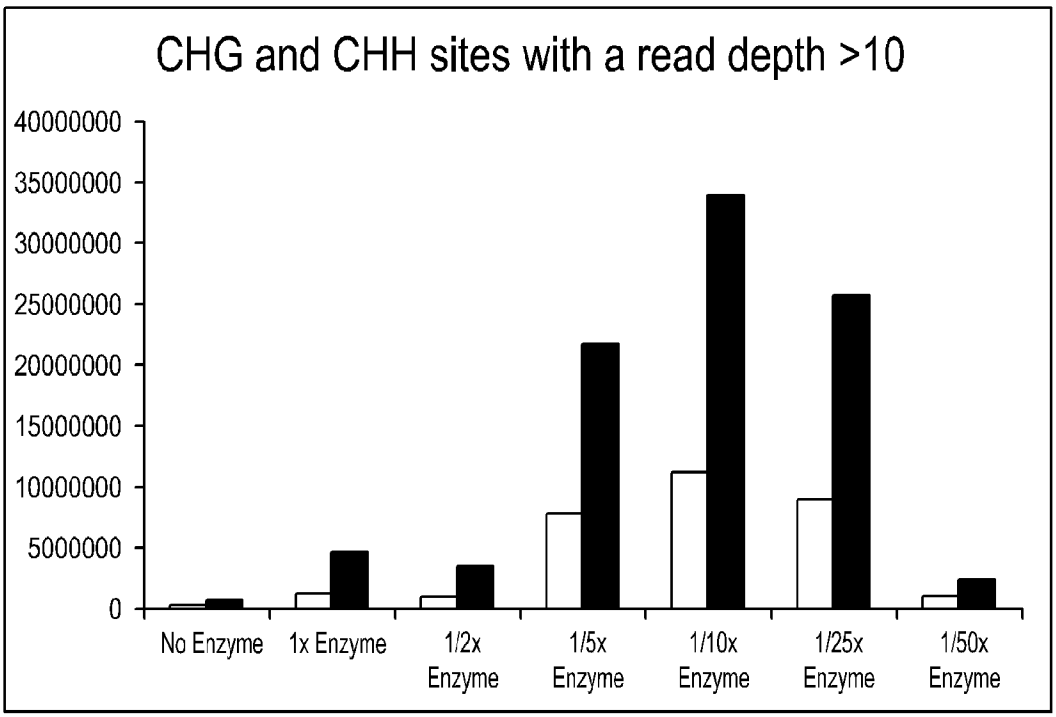

Figure 11B

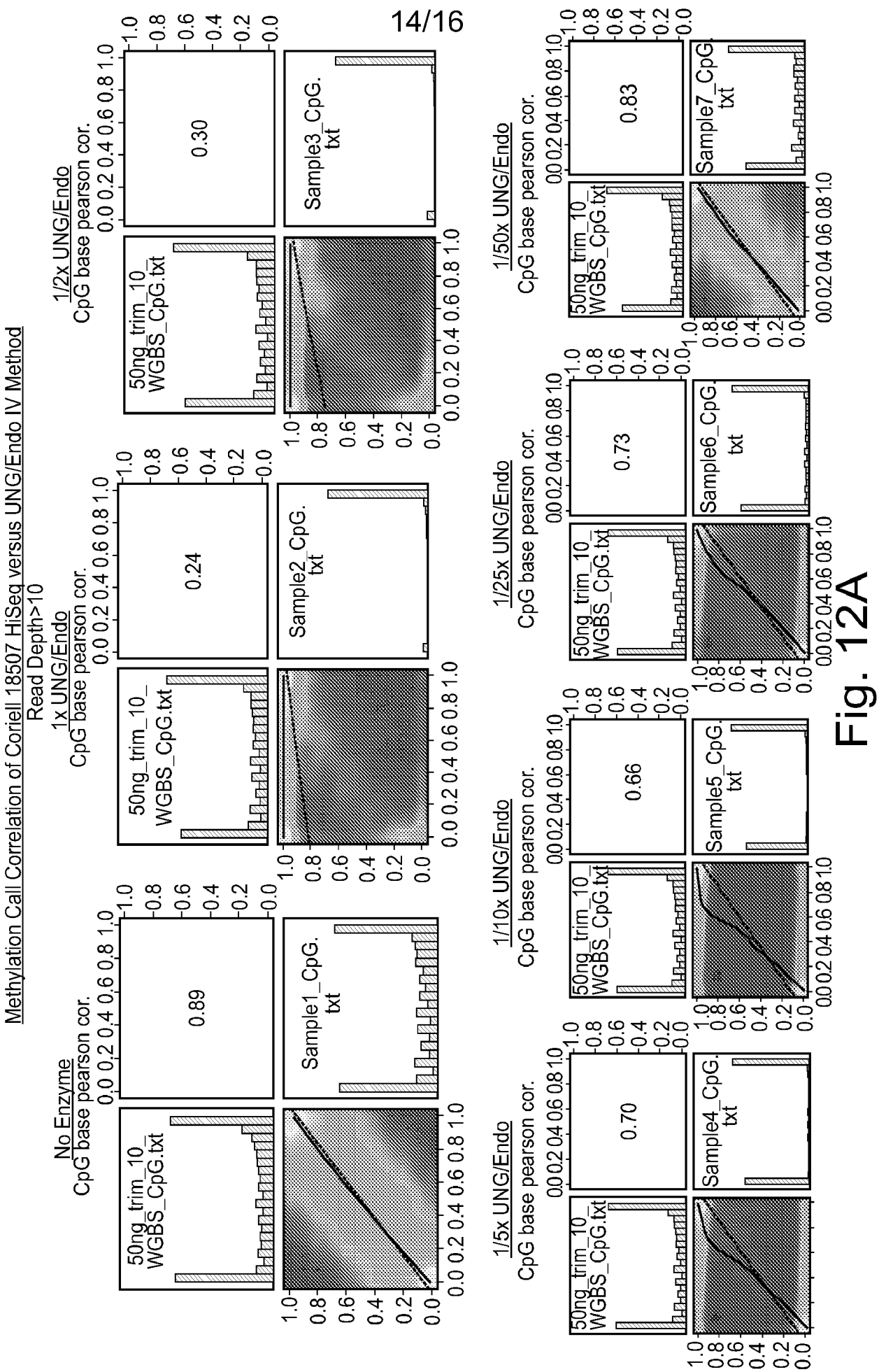
Methylation Call Correlation of Coriell 18507 HiSeq versus UNG/Endo IV Method
Read Depth>10
No Enzyme
1x UNG/Endo
1/2x UNG/Endo
1/5x UNG/Endo
1/10x UNG/Endo
1/25x UNG/Endo
1/50x UNG/Endo
CpG base pearson cor.
50ng_trim_10_WGBS_CpG.txt
Sample1_CpG.txt
Sample2_CpG.txt
Sample3_CpG.txt
Sample4_CpG.txt
Sample5_CpG.txt
Sample6_CpG.txt
Sample7_CpG.txt
0.89
0.24
0.30
0.70
0.66
0.73
0.83

Fig. 12A

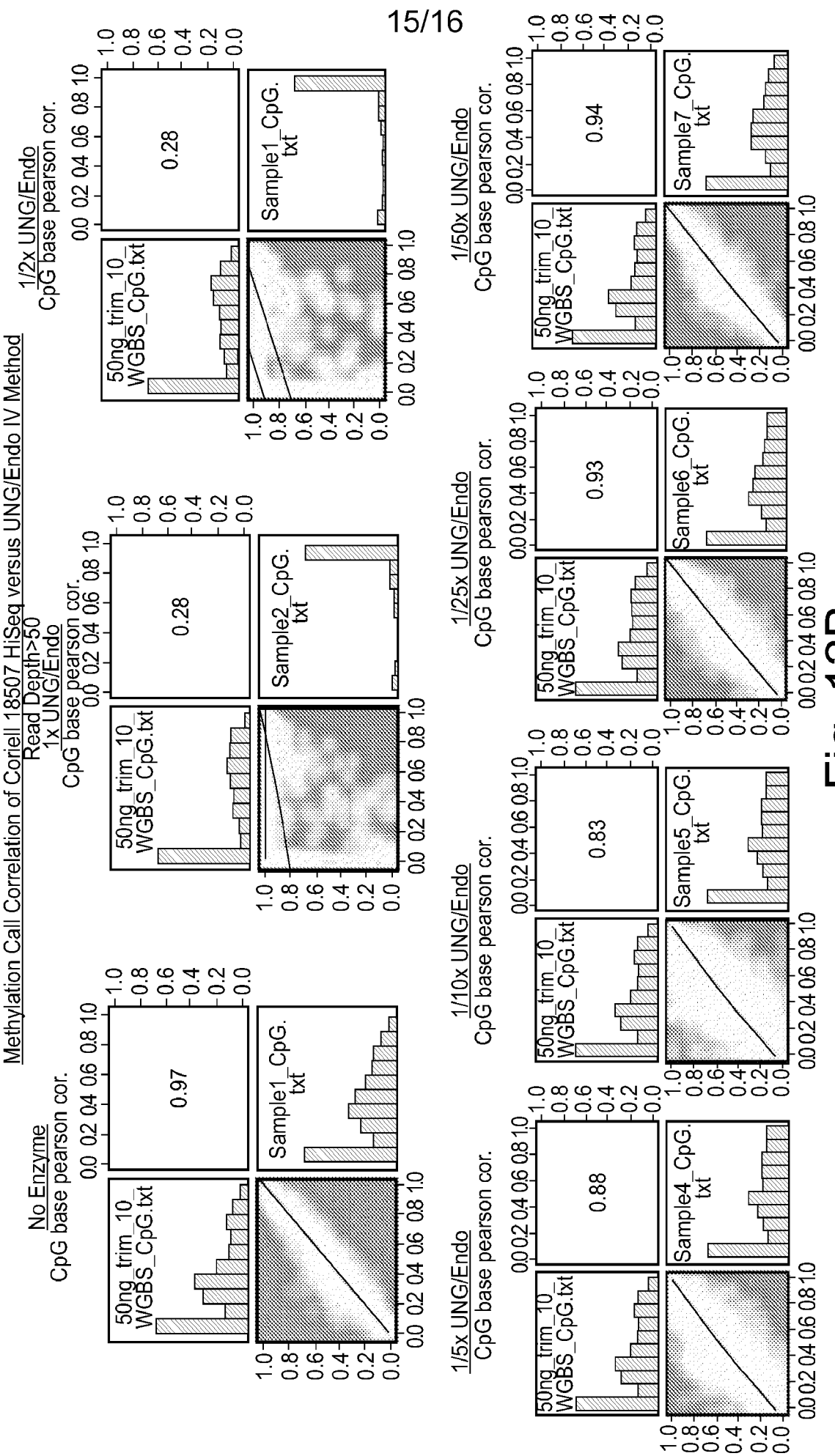
Methylation Call Correlation of Coriell 18507 HiSeq versus UNG/Endo IV Method
Read Depth>50
No Enzyme
CpG base pearson cor.
0.97
50ng_trim_10_WGBS_CpG.txt
Sample1_CpG.txt
1x UNG/Endo
0.28
Sample2_CpG.txt
1/2x UNG/Endo
0.28
Sample1_CpG.txt
1/5x UNG/Endo
0.88
Sample4_CpG.txt
1/10x UNG/Endo
0.83
Sample5_CpG.txt
1/25x UNG/Endo
0.93
Sample6_CpG.txt
1/50x UNG/Endo
0.94
Sample7_CpG.txt
0.0 0.2 0.4 0.6 0.8 1.0

Fig. 12B

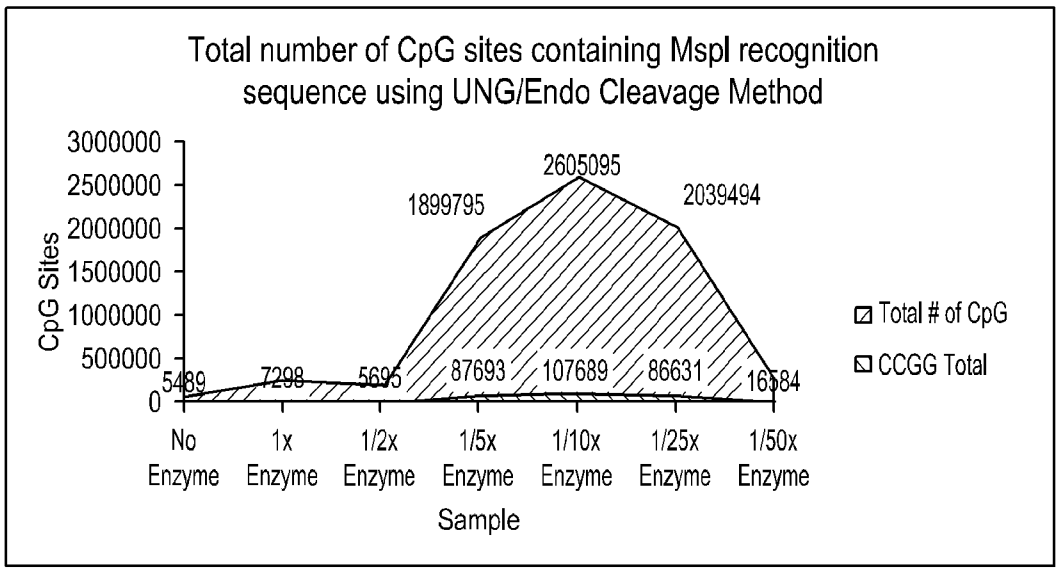

Figure 13A

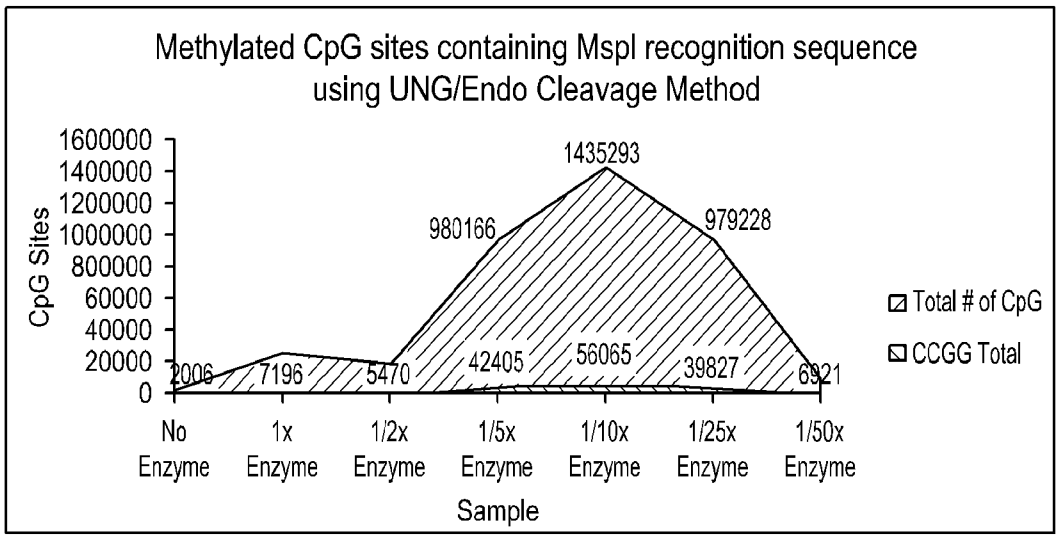

Figure 13B

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2015/049385

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, BIOSIS, EMBASE, FSTA, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	US 2009/233804 A1 (KURN NURITH [US] ET AL) 17 September 2009 (2009-09-17) para. 18, 63, 100-102, 105, 106, 150, 161, 180-182, 186, 188, 194; table 1	1-4,6, 8-53 5
X	Anonymous: "Uracil-DNA Excision Mix - Cat No UEM04100", Epicentre , 1 October 2012 (2012-10-01), XP055232599, Retrieved from the Internet: URL:http://www.epibio.com/docs/default-source/protocols/uracil-dna-excision-mix.pdf?sfvrsn=8 [retrieved on 2015-12-01]	54,55
Y	p. 2, para. 1	5

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
2 December 2015	14/12/2015

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Ripaud, Leslie

2

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2012/171691 A1 (PEITZ INGMAR [DE]) 5 July 2012 (2012-07-05) para. 1, 79, 85, 87, 130 -----	1-3,6,7, 20,43, 45,50-53
X	HUANG RONG ET AL: "A novel combined bisulfite UDG assay for selective 5-methylcytosine detection", TALANTA, vol. 117, 15 December 2013 (2013-12-15), pages 445-448, XP028767799, ISSN: 0039-9140, DOI: 10.1016/J.TALANTA.2013.09.026 abstract; p. 446, para. 2.2-2.4; scheme 1 -----	1-3,6,8, 20,43, 45,50-53
A	US 2005/112590 A1 (BOOM DIRK V D [US] ET AL) 26 May 2005 (2005-05-26) para. 37-40, 181-184, 189 -----	1-55

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2015/049385

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2009233804 A1	17-09-2009	AU 2004311882 A1	21-07-2005
		CA 2552007 A1	21-07-2005
		EP 1711591 A2	18-10-2006
		JP 2007524407 A	30-08-2007
		US 2005208538 A1	22-09-2005
		US 2009233804 A1	17-09-2009
		WO 2005065321 A2	21-07-2005
US 2012171691 A1	05-07-2012	CN 102549168 A	04-07-2012
		EP 2480685 A1	01-08-2012
		JP 5702789 B2	15-04-2015
		JP 2013505031 A	14-02-2013
		US 2012171691 A1	05-07-2012
		WO 2011036609 A1	31-03-2011
US 2005112590 A1	26-05-2005	AU 2003298733 A1	23-06-2004
		CA 2507189 A1	17-06-2004
		CN 1774511 A	17-05-2006
		EP 1613723 A2	11-01-2006
		HK 1087436 A1	21-02-2014
		JP 4786904 B2	05-10-2011
		JP 2006515987 A	15-06-2006
		US 2005112590 A1	26-05-2005
		WO 2004050839 A2	17-06-2004