Each diversity base in a diversity adaptor comprising more than one diversity base can be either a "D" base or an "R" base. In some cases, a pool of diversity adaptors is used in a method herein, wherein the pool comprises diversity adaptors comprising no diversity bases, diversity adaptors comprising 1 "D" base, diversity adaptors comprising 2 "D" bases, and diversity adaptors comprising 1 R base and 2 D bases. The diversity adaptor with 1 R base and two D bases can have the R base at the 5', middle, or 3' position.

[0065] In some cases, the diversity bases can be added to the ends of target polynucleotides as shown in **FIG. 35**, using primers comprising said diversity bases or complements thereof instead of adaptors, whereby the introduction is performed using polymerase extension reactions (e.g., PCR). The primers can be designed to target specific sequences in a polynucleotide such as specific genomic regions (e.g., 16S rRNA sequences in microbial species) such that the primers comprise tails that comprise the diversity bases (i.e., extender sequences) or complements thereof. The adaptors comprising diversity base(s) (i.e., extender sequence) can be designed such that the diversity base or bases are immediately adjacent to sequence derived from a polynucleotide or polynucleotide insert from a library. In other words, the diversity base or bases are the first non-target sequence polynucleotide base or bases present at the junction of a complex comprising a target sequence from a polynucleotide and non-target sequence derived sequence. Also as provided herein, the adaptor comprising diversity bases can comprise a sequence or complement thereof that is capable of hybridizing a sequencing primer and this sequence or complement thereof can be located adjacent to the diversity base or bases such that the diversity base or bases are the bases initially sequenced in a sequencing reaction. In some cases, there can also be additional sequence elements located between the sequence or complement thereof that is capable of hybridizing a sequencing primer and the diversity base or bases such as barcodes (e.g., inline barcodes), universal sequence, or random sequence. In some cases, the additional sequence elements (e.g., barcode(s), universal sequence and/or random sequence) can be located after the sequence or complement thereof that is capable of hybridizing a sequencing primer and distal to the diversity base or bases. In some cases, the additional sequence elements (e.g., barcode(s), universal sequence and/or random sequence) can be located after the sequence or complement thereof that is capable of hybridizing a sequencing primer and distal to the diversity base or bases as well as between the sequence or complement thereof that is capable of hybridizing a sequencing primer and the diversity base or bases such as barcodes (e.g., inline barcodes), universal sequence, or random sequence. An example of a polynucleotide library insert generated to possess one or more additional sequence elements after the sequence that is capable of hybridizing a sequencing primer and therefore distal to the end of the

IDT Exhibit 1020.3717

polynucleotide library insert is shown in **FIGs. 32** and **37**. The insert generated in **FIGs. 32** and **37** can be generated by introducing an adaptor or primer as provided herein.

[0066] As provided herein, the polynucleotide to which the adaptor attached can be a nucleic acid fragment generated using any fragmentation method known in the art. The fragmentation method can be one described herein such as enzymatic fragmentation. The nucleic acid fragment can be double stranded and the adaptor can be attached to both ends of the double stranded nucleic acid fragment. The term "adaptor", as used herein, can refer to an oligonucleotide of known sequence or an oligonucleotide comprising a portion or portions of known sequence, the ligation of which to a target polynucleotide or a target polynucleotide strand of interest enables the generation of amplification-ready products of the target polynucleotide or the target polynucleotide strand of interest. Various adaptor designs can be envisioned. Various ligation processes and reagents are known in the art and can be useful for carrying out the methods described herein. For example, blunt end ligation can be employed. Similarly, a single dA nucleotide can be added to the 3′-end of a double-stranded, blunted DNA product, by a polymerase lacking 3′-exonuclease activity and can anneal to an adaptor comprising a dT overhang (or the reverse). This design allows the hybridized components to be subsequently ligated (e.g., by T4 DNA ligase). Other ligation strategies and the corresponding reagents known in the art and kits and reagents for carrying out efficient ligation reactions are commercially available (e.g, from New England Biolabs, Roche).

[0067] In some embodiments, an adaptor comprises a duplex adaptor. In some cases, an adapter comprises a partial duplex adaptor. The partial duplex adaptor can be a forked adaptor. The forked adaptor can be a forked adaptor available from a commercial source and comprise known sequence. The forked adaptor can be compatible with specific NGS sequencing systems that utilize cluster amplification (e.g, Illumina forked adaptors for Illumina based sequencing by synthesis). The forked adaptors can comprise a duplex region where a first strand and a second strand share complementary sequence. The duplex region can be of varying length. The duplex region can be 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30 or more nucleotides in length. Outside of the duplex region, the forked adaptor can have single stranded non-complementary regions such that the first strand comprises a 5' single stranded region that is non-complementary to a 3' single-stranded region of the second strand. In some cases, one or more cytosine bases in the forked adaptor can be substituted with a cytosine analog resistant to bisulfite treatment. These types of forked adaptors can be useful in the generation of RRBS libraries using the methods provided herein. In some cases, the cytosine analog is 5-methylcytosine. In some cases, the cytosine analog is 5-hydroxymethylcytosine. In some cases, the cytosine analog is 5-

propynylcytosine. The 3' and 5' ends of the duplex region of a forked adaptor used in the methods, compositions, and kits herein can comprise no or one or more bases of extender sequence (i.e., diversity bases) as described herein or complements thereof. In some cases, ligation of a forked adaptor as provided herein to both ends of a double stranded polynucleotide can generate forked adaptor-ligated polynucleotide complexes comprising non-complementary adaptor sequence on the 5' and 3' ends of each strand of the double stranded polynucleotide. The non-complementary sequence can be useful for further manipulation and downstream processing of the ligated polynucleotide complexes (e.g., PCR amplification, solid-state amplification, and/or sequencing).

[0068] In some cases, the partial duplex adaptor useful in the methods, compositions and kits provided herein can comprise a long strand and a short strand wherein the short strand comprises sequence complementary to the long strand. The short strand can hybridize completely with the long strand. The short strand can hybridize to a portion of the long strand, and thereby comprise either a 5' or 3' overhang. In some cases, an adaptor for use in the methods provided herein comprises a partial duplex adaptor as shown in **FIGs 2A-B** or **3**. In some cases, an adapter for use in the methods described herein comprises a partial duplex adapter wherein the long strand comprises adaptor sequence comprising one or more sequence elements, while the short strand comprises sequence complementary to a portion of the sequence of the long strand as well as an overhang complementary to sequence in an overhang found in a nucleic acid following cleavage of said nucleic acid.

[0069] The nucleic acid can be cleaved by a restriction endonuclease (RE). The RE can be any restriction endonuclease described herein. In some cases, the RE is MspI.

[0070] In some cases, an adaptor for use in the methods provided herein comprises a ligation strand and a non-ligation strand. In some cases, an adaptor for use in the methods provided herein comprises a partial duplex wherein the long strand comprises a ligation strand, and a short strand comprises a non-ligation strand.

[0071] In some cases, an adaptor (e.g., diversity adaptor) for use in the methods provided herein comprises one or more identifier sequences. The one or more identifier sequences can be present in a long strand and/or a short strand of a partial duplex diversity adaptor as provided herein. The one or more identifier sequences can be present in a ligation strand and/or a non-ligation strand of a partial duplex diversity adapter as provided herein. The identifier sequence can be a barcode sequence, a flow cell sequence, and/or an index sequence. In some cases, the index sequence is a Truseq primer sequence compatible with the NGS platform produced by Illumina. In some cases, the 3' and/or 5' ends of the non-ligation strand comprise a blocking group or

IDT Exhibit 1020.3719

moiety and are enzymatically unreactive. The blocking group or moiety can be a dideoxynucleotide (ddCMP, ddAMP, ddTMP, or ddGMP), various modified nucleotides (e.g. phosphorothioate-modified nucleotides), or non-nucleotide chemical moieties. In some cases, the blocking group comprises a nucleotide analog that comprises a blocking moiety. The blocking moiety can mean a part of the nucleotide analog that inhibits or prevents the nucleotide analog from forming a covalent linkage to a second nucleotide or nucleotide analog. For example, in the case of nucleotide analogs having a pentose moiety, a reversible blocking moiety can prevent formation of a phosphodiester bond between the 3′ oxygen of the nucleotide and the 5′ phosphate of the second nucleotide. Reversible blocking moieties can include phosphates, phosphodiesters, phosphotriesters, phosphorothioate esters, and carbon esters. In some cases, a blocking moiety can be attached to the 3′ position or 2′ position of a pentose moiety of a nucleotide analog. A reversible blocking moiety can be removed with a deblocking agent. The 3' end of the non-ligation strand can be modified to comprise a blocking group, for example, a dideoxynucleotide (ddCMP, ddAMP, ddTMP, or ddGMP) to prevent polymerase extension. The blocking group at the 3' end of the non-ligation strand can be a nucleotide terminator. In some cases, the block at the 3' end of the non-ligation strand comprises a terminal dideoxycytosine. The 5' end of the non-ligation strand can be modified to comprise a blocking group. The blocking group at the 5' end of the non-ligation strand can be a spacer (C3 phosphoramidite, triethylene glycol (TEG), photo-cleavable, hexa-ethyleneglycol), inverted dideoxy-T, biotin, thiol, dithiol, hexanediol, digoxigenin, an azide, alkynes, or an amino modifier. The biotin blocking group can be photocleavable biotin, biotin-triethylene glycol (TEG), biotin-dT, desthiobiotin-TEG, biotin-azide, or dual biotin. In some cases, the block at the 5' end of the non-ligation strand comprises a biotin moiety. In some cases, the 5' end of the non-ligation strand does not comprise a 5' phosphate. The 5' end can be removed by treatment with an enzyme. The enzyme can be a phosphatase. In some cases, the 5' end of the non-ligation strand is dephosphorylated by treatment with alkaline phosphatase. In some cases, the 5' end of the non-ligation strand does comprise a 5' phosphate, wherein the 3' end of the polynucleotide lacks a free 3' hydroxyl. In some cases, the non-ligation strand comprises a block at the 3' end comprising terminal dideoxycytosine and a block at the 5' end comprising a biotin moiety. In some cases, one or more cytosine bases in the ligation strand of a duplex adaptor can be substituted with a cytosine analog resistant to bisulfite treatment. The ligation strand comprising the one or more cytosine analogs resistant to bisulfite treatment can be the long strand in a partial duplex adaptor as provided herein. In some cases, one or more cytosine bases in the non-ligation strand of a duplex adaptor can be substituted with a cytosine analog resistant to bisulfite

IDT Exhibit 1020.3720

treatment. The non-ligation strand comprising the one or more cytosine analogs resistant to bisulfite treatment can be the short strand in a partial duplex adaptor as provided herein. These types of duplex or partial duplex adaptors can be useful in the generation of RRBS libraries using the methods provided herein. In some cases, the cytosine analog is 5-methylcytosine. In some cases, the cytosine analog is 5-hydroxymethylcytosine. In some cases, the cytosine analog is 5-propynylcytosine.

[0072] Provided herein is a method for generating the diversity adaptors useful for generating RRBS nucleic acid libraries comprising increased sequence diversity as provided herein. **FIG. 2A-B** illustrates a method for generating the diversity adaptors useful in the methods, compositions, and kits proved herein. **FIG. 2A** shows a long strand of nucleotides (R01446) and 4 different short strands (i.e., Ben383, Ben384, Ben385, and Ben341) such that each short strand comprises stretch of 12 nucleotides that comprise bases complementary to the 3' end of the long strand (R01446). Each of the four short strands differ in their 5' end sequences such that Ben341 comprises a 5'-CG-3' dinucleotide, while Ben385 comprises an additional H base 3' to the 5'-CG-3' dinucleotide, Ben384 an additional two H bases 3' to the 5'-CG-3' dinucleotide, and Ben383 an additional three H bases 3' to the 5'-CG-3' dinucleotide. The H bases in the short strands can be either and adenine (A), cytosine (C), or thymine (T). **FIG. 2B** illustrates the method used to generate a diversity adaptor using the long strand (R01446) and a representative short strand (Ben383) from **FIG. 2A**, although it should be noted that any of the four short strands can be substituted for the Ben383 short strand in **FIG. 2B.** Each of the short strands (Ben383, 384, 385, or 341) in **FIG. 2A** comprise a 5' terminal GC sequence with a Biosg label on the terminal cytosine residue, while the 3' terminus comprises a terminal 3' dideoxy cytosine residue.

[0073] In **FIG. 2B**, the long strand of nucleotides (R01446) can be annealed to a short strand of nucleotides (Ben383) such that a partial duplex is formed. The partial duplex can comprise a duplex region and a 5' overhang on both ends of the duplex, wherein both the long strand (R01446) and the short strand (Ben383) contribute sequence to the 5' overhang on the opposite strand. The 5' overhang comprising sequence from the short strand of the partial duplex can comprise a 5' terminal GC sequence with a Biosg label on the terminal cytosine residue, while the 3' terminus comprises a terminal 3' dideoxy cytosine residue. Additionally, the 5' overhang from the short strand comprises three "H" bases. If Ben384 were used in the method illustrated in **FIG. 2B**, the 5' overhang would further comprise two H bases, Ben385 would comprise one H base, and Ben341 would comprise no H bases. The "H" bases can be either an adenosine, cytosine or thymine residue, but cannot be a guanine residue. As shown in **FIG. 2B**, following

IDT Exhibit 1020.3721

annealing, the partial duplex adapter can be subjected to an extension reaction using a dNTP mix containing only dATP, dGTP, and dTTP (i.e., no dCTP), whereby a partial duplex adapter comprising a 5' overhang comprising at least one "H" (e.g., R01446 hybridized to Ben 383, 384, or 385) is extended to produce a partial duplex adapter containing a two base 5' overhang with the sequence GC at one end. In some cases, the diversity adaptors generated in **FIG. 2B** are ligated to polynucleotides fragmented with MspI. The diversity adaptors generated in **FIG. 2B** can be mixed into a single pool thereby comprising a mixture of diversity adaptors comprising 0, 1, 2, or 3 D bases (D0-D3) and ligated to the MspI polynucleotide fragments. In some cases, the MspI polynucleotide fragments were separated into different pools, wherein a first pool was ligated to diversity adaptors with no D bases (D0), a second pool was ligated to diversity adapters with one D base (D1), a third pool was ligated diversity adapters comprising two D bases (D2), a fourth pool was ligated to adapters comprising three D bases (D3), and a fifth pool was ligated to a mixture of D0, D1, D2, and D3 diversity adapters. The methods illustrated in **FIG. 2A-B** can be used to generate diversity adaptors useful for ligating to polynucleotide fragments generated by digestion with any restriction enzyme. In these cases, the nucleotide base compositions of the non-H base overhang can be adjusted to be complementary to overhang sequence produced by enzymatic digestion, while the base identity of the H bases can be altered to so as to be distinguishable from the recognition sequence of the restriction enzyme used to generate the site upon sequencing of the fragments. In some cases, diversity adaptors or pools of diversity adaptors can be ligated to polynucleotide ends using blunt end ligation. The ends of the polynucleotide can be blunt due to digestion with an enzyme that generates blunt ends, or can be blunted using an end repair reaction as described herein and known in the art. In some cases, the diversity adaptors are configured such that each strand is capable of ligation to an end of a polynucleotide. In some cases, the diversity adaptors where both strands are ligation capable differ in that one strand comprises cytosine analogs resistant to bisulfite treatment, while the complementary strand comprises cytosine bases sensitive to bisulfite treatment such that bisulfite treatment of polynucleotides comprising said diversity adaptors ligated thereto generate polynucleotides comprising non-complementary termini.

III. Polynucleotides, oligonucleotides, primers, samples, and nucleotides

[0074] A bisulfite converted polynucleotide library (e.g., reduced representation bisulfite library) can be generated from a polynucleotide obtained from a source of polynucleotides. The polynucleotides can be single-stranded or double stranded. In some cases, the polynucleotide is DNA. The DNA can be obtained and purified using standard techniques in the art and include DNA in purified or unpurified form. The DNA can be mitochondrial DNA, cell-free DNA,

IDT Exhibit 1020.3722

complementary DNA (cDNA), or genomic DNA. In some cases, the polynucleotide is genomic DNA. The DNA can be plasmid DNA, cosmid DNA, bacterial artificial chromosome (BAC), or yeast artificial chromosome (YAC). The DNA can be derived from one or more chromosomes. For example, if the DNA is from a human, the DNA can derived from one or more of chromosome 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, or Y. In some cases, the DNA is double-stranded DNA. In some cases, the double-stranded DNA is genomic DNA. In some cases, the DNA is cDNA. In some cases, the cDNA is double-stranded cDNA. In some cases, the cDNA is derived from RNA, wherein the RNA is subjected to first strand synthesis followed by second strand synthesis. The RNA can be obtained and purified using standard techniques in the art and include RNAs in purified or unpurified form, which include, but are not limited to, mRNAs, tRNAs, snRNAs, rRNAs, retroviruses, small non-coding RNAs, microRNAs, polysomal RNAs, pre-mRNAs, intronic RNA, viral RNA, cell free RNA and fragments thereof. The non-coding RNA or ncRNA can include snoRNAs, microRNAs, siRNAs, piRNAs and long nc RNAs. First strand synthesis can be performed using any number of RNA dependent DNA polymerases known in the art.

[0075] The source of polynucleotides for use in the methods described herein can be a sample comprising the polynucleotides. The polynucleotides can be isolated from the sample and purified by any of the methods known in the art for purifying the nucleic acid from the sample. The sample can be derived from a non-cellular entity comprising polynucleotides (e.g., a virus) or from a cell-based organism (e.g., member of archaea, bacteria, or eukarya domains). In some cases, the sample is obtained from a swab of a surface, such as a door or bench top.

[0076] The sample can from a subject, e.g., a plant, fungi, eubacteria, archeabacteria, protest, or animal. The subject can be an organism, either a single-celled or multi-cellular organism. The subject can be cultured cells, which can be primary cells or cells from an established cell line, among others. The sample can be isolated initially from a multi-cellular organism in any suitable form. The animal can be a fish, e.g., a zebrafish. The animal can be a mammal. The mammal can be, e.g., a dog, cat, horse, cow, mouse, rat, or pig. The mammal can be a primate, e.g., a human, chimpanzee, orangutan, or gorilla. The human can be a male or female. The sample can be from a human embryo or human fetus. The human can be an infant, child, teenager, adult, or elderly person. The female can be pregnant, suspected of being pregnant, or planning to become pregnant.

[0077] The sample can be from a subject (e.g., human subject) who is healthy. In some cases, the sample is taken from a subject (e.g., an expectant mother) at at least 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or 26 weeks of gestation. In some cases, the

IDT Exhibit 1020.3723

subject is affected by a genetic disease, a carrier for a genetic disease or at risk for developing or passing down a genetic disease, where a genetic disease is any disease that can be linked to a genetic variation such as mutations, insertions, additions, deletions, translocation, point mutation, trinucleotide repeat disorders and/or single nucleotide polymorphisms (SNPs).

[0078] The sample can be from a subject who has a specific disease, disorder, or condition, or is suspected of having (or at risk of having) a specific disease, disorder or condition. For example, the sample can be from a cancer patient, a patient suspected of having cancer or a patient at risk of having cancer. The cancer can be, e.g., acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), adrenocortical carcinoma, Kaposi Sarcoma, anal cancer, basal cell carcinoma, bile duct cancer, bladder cancer, bone cancer, osteosarcoma, malignant fibrous histiocytoma, brain stem glioma, brain cancer, craniopharyngioma, ependymoblastoma, ependymoma, medulloblastoma, medulloeptithelioma, pineal parenchymal tumor, breast cancer, bronchial tumor, Burkitt lymphoma, Non-Hodgkin lymphoma, carcinoid tumor, cervical cancer, chordoma, chronic lymphocytic leukemia (CLL), chronic myelogenous leukemia (CML), colon cancer, colorectal cancer, cutaneous T-cell lymphoma, ductal carcinoma in situ, endometrial cancer, esophageal cancer, Ewing Sarcoma, eye cancer, intraocular melanoma, retinoblastoma, fibrous histiocytoma, gallbladder cancer, gastric cancer, glioma, hairy cell leukemia, head and neck cancer, heart cancer, hepatocellular (liver) cancer, Hodgkin lymphoma, hypopharyngeal cancer, kidney cancer, laryngeal cancer, lip cancer, oral cavity cancer, lung cancer, non-small cell carcinoma, small cell carcinoma, melanoma, mouth cancer, myelodysplastic syndromes, multiple myeloma, medulloblastoma, nasal cavity cancer, paranasal sinus cancer, neuroblastoma, nasopharyngeal cancer, oral cancer, oropharyngeal cancer, osteosarcoma, ovarian cancer, pancreatic cancer, papillomatosis, paraganglioma, parathyroid cancer, penile cancer, pharyngeal cancer, pituitary tumor, plasma cell neoplasm, prostate cancer, rectal cancer, renal cell cancer, rhabdomyosarcoma, salivary gland cancer, Sezary syndrome, skin cancer, nonmelanoma, small intestine cancer, soft tissue sarcoma, squamous cell carcinoma, testicular cancer, throat cancer, thymoma, thyroid cancer, urethral cancer, uterine cancer, uterine sarcoma, vaginal cancer, vulvar cancer, Waldenstrom Macroglobulinemia, or Wilms Tumor. The sample can be from the cancer and/or normal tissue from the cancer patient.

[0079] The sample can be from a subject who is known or suspected to have a genetic disease, disorder or condition. The sample can be from a subject who is known or suspected to have gene silencing, imprinting, X-chromosome inactivation, phenotypic variation or genetic disease susceptibility. In some cases, the subject is known to be wild-type or mutant for a gene, or portion of a gene, e.g., CFTR, Factor VIII (F8 gene), beta globin, hemachromatosis, G6PD,

IDT Exhibit 1020.3724

neurofibromatosis, GAPDH, beta amyloid, or pyruvate kinase gene. In some cases, the status of the subject is either known or not known, and the subject is tested for the presence of a mutation or genetic variation of a gene, e.g., CFTR, Factor VIII (F8 gene), beta globin, hemachromatosis, G6PD, neurofibromatosis, GAPDH, beta amyloid, or pyruvate kinase gene.

[0080] The sample can be aqueous humour, vitreous humour, bile, whole blood, blood serum, blood plasma, breast milk, cerebrospinal fluid, cerumen, enolymph, perilymph, gastric juice, mucus, peritoneal fluid, saliva, sebum, semen, sweat, tears, vaginal secretion, vomit, feces, or urine. The sample can be obtained from a hospital, laboratory, clinical or medical laboratory. The sample can be taken from a subject.

[0081] The sample can comprise nucleic acid. The sample can comprise a whole genome, and thus be considered a whole genome sample. The sample can be a nucleic acid or polynucleotide library. The nucleic acid can be, e.g., mitochondrial DNA, genomic DNA, mRNA, siRNA, miRNA, cRNA, single-stranded DNA, double-stranded DNA, single-stranded RNA, double-stranded RNA, tRNA, rRNA, or cDNA. The sample can comprise cell-free nucleic acid. The sample can be a cell line, genomic DNA, cell-free plasma, formalin fixed paraffin embedded (FFPE) sample, or flash frozen sample. A formalin fixed paraffin embedded sample can be deparaffinized before nucleic acid is extracted. The sample can be from an organ, e.g., heart, skin, liver, lung, breast, stomach, pancreas, bladder, colon, gall bladder, brain, etc. Nucleic acids can be extracted from a sample by means available to one of ordinary skill in the art.

[0082] The sample can be processed to render it competent for fragmentation, ligation, denaturation, and/or amplification. Exemplary sample processing can include lysing cells of the sample to release nucleic acid, purifying the sample (e.g., to isolate nucleic acid from other sample components, which can inhibit enzymatic reactions), diluting/concentrating the sample, and/or combining the sample with reagents for further nucleic acid processing. In some examples, the sample can be combined with a restriction enzyme, reverse transcriptase, or any other enzyme of nucleic acid processing.

[0083] The methods described herein can be used for analyzing or detecting one or more target polynucleotides. The term polynucleotide, or grammatical equivalents, can refer to at least two nucleotides covalently linked together. A polynucleotide described herein can contain phosphodiester bonds, although in some cases, as outlined below (for example in the construction of primers and probes such as label probes), nucleic acid analogs are included that can have alternate backbones, comprising, for example, phosphoramide (Beaucage et al., Tetrahedron 49(10):1925 (1993) and references therein; Letsinger, J. Org. Chem. 35:3800 (1970); Sprinzl et al., Eur. J. Biochem. 81:579 (1977); Letsinger et al., Nucl. Acids Res. 14:3487

IDT Exhibit 1020.3725

(1986); Sawai et al, Chem. Lett. 805 (1984), Letsinger et al., J. Am. Chem. Soc. 110:4470 (1988); and Pauwels et al., Chemica Scripta 26:141 91986)), phosphorothioate (Mag et al., Nucleic Acids Res. 19:1437 (1991); and U.S. Pat. No. 5,644,048), phosphorodithioate (Briu et al., J. Am. Chem. Soc. 111:2321 (1989), O-methylphosphoroamidite linkages (see Eckstein, Oligonucleotides and Analogues: A Practical Approach, Oxford University Press), and peptide nucleic acid (also referred to herein as "PNA") backbones and linkages (see Egholm, J. Am. Chem. Soc. 114:1895 (1992); Meier et al., Chem. Int. Ed. Engl. 31:1008 (1992); Nielsen, Nature, 365:566 (1993); Carlsson et al., Nature 380:207 (1996), all of which are incorporated by reference). Other analog nucleic acids include those with bicyclic structures including locked nucleic acids (also referred to herein as "LNA"), Koshkin et al., J. Am. Chem. Soc. 120.13252 3 (1998); positive backbones (Denpcy et al., Proc. Natl. Acad. Sci. USA 92:6097 (1995); non-ionic backbones (U.S. Pat. Nos. 5,386,023, 5,637,684, 5,602,240, 5,216,141 and 4,469,863; Kiedrowshi et al., Angew. Chem. Intl. Ed. English 30:423 (1991); Letsinger et al., J. Am. Chem. Soc. 110:4470 (1988); Letsinger et al., Nucleoside & Nucleotide 13:1597 (1994); Chapters 2 and 3, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui and P. Dan Cook; Mesmaeker et al., Bioorganic & Medicinal Chem. Lett. 4:395 (1994); Jeffs et al., J. Biomolecular NMR 34:17 (1994); Tetrahedron Lett. 37:743 (1996)) and non-ribose backbones, including those described in U.S. Pat. Nos. 5,235,033 and 5,034,506, and Chapters 6 and 7, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui and P. Dan Cook. Nucleic acids containing one or more carbocyclic sugars are also included within the definition of nucleic acids (see Jenkins et al., Chem. Soc. Rev. (1995) pp 169 176). Several nucleic acid analogs are described in Rawls, C & E News Jun. 2, 1997 page 35. "Locked nucleic acids" are also included within the definition of nucleic acid analogs. LNAs are a class of nucleic acid analogues in which the ribose ring is "locked" by a methylene bridge connecting the 2'-O atom with the 4'-C atom. All of these references are hereby expressly incorporated by reference. These modifications of the ribose-phosphate backbone can be done to increase the stability and half-life of such molecules in physiological environments. For example, PNA:DNA and LNA-DNA hybrids can exhibit higher stability and thus can be used in some cases. The polynucleotides can be single stranded or double stranded, as specified, or contain portions of both double stranded or single stranded sequence. Depending on the application, the nucleic acids can be DNA (including, e.g., genomic DNA, mitochondrial DNA, and cDNA), RNA (including, e.g., mRNA and rRNA) or a hybrid, where the nucleic acid contains any combination of deoxyribo- and ribo-nucleotides, and any combination of bases,

IDT Exhibit 1020.3726

including uracil, adenine, thymine, cytosine, guanine, inosine, xathanine hypoxathanine, isocytosine, isoguanine, etc.

[0084] "Unmodified nucleotide" or "unmodified dNTP" can refer to the four deoxyribonucleotide triphosphates dATP (deoxyadenosine triphosphate), dCTP (deoxycytidine triphosphate), dGTP (deoxyguanosine triphosphate) and dTTP (deoxythymidine triphosphate) that can normally be used as building blocks in the synthesis of DNA.

[0085] "Modified nucleotide," "modified dNTP," or "nucleotide analog," can refer to any molecule suitable for substituting one corresponding unmodified nucleotide. The modified nucleotide or dNTP render the polynucleotide more or less susceptible to degradation or alteration by a suitable degrading or altering agent. In some cases, the modified nucleotide substitutes for cytosine, which in its unmodified state undergoes conversion to uracil when subjected to bisulfite treatment. In some cases, the modified nucleotide substituting for cytosine is 5-methylcytosine. In some cases, the modified nucleotide substituting for cytosine is 5-hydroxymethylcytosine. In some cases, the modified nucleotide is 5-propynylcytosine.

[0086] As used herein, "oligonucleotide" can refer to a polynucleotide chain, typically less than 200 residues long, most typically between 15 and 100 nucleotides long, but also intended to encompass longer polynucleotide chains. Oligonucleotides may be single-or double-stranded. The terms "primer", and "oligonucleotide primer", as used herein, can refer to an oligonucleotide capable of hybridizing to a complementary nucleotide sequence.

[0087] As used herein, "hybridization"/ "hybridizing" and "annealing" can be used interchangeably and refer to the pairing of complementary nucleic acids or sequences within nucleic acids. "Primer", as used herein, can refer to an oligonucleotide, generally with a free 3' hydroxyl group, that is capable of hybridizing with a template (such as a target polynucleotide or target sequence within a polynucleotide, target DNA, target RNA or a primer extension product) and is also capable of promoting polymerization of a polynucleotide complementary to the template. A primer may contain a non-hybridizing sequence that constitutes a tail of the primer. A primer may still be hybridizing to a target even though its sequences are not fully complementary to the target. In some cases, a first and/or second primer used to amplify a nucleic acid or polynucleotide fragment comprising diversity sequences on the 5' and/or 3' ends generated by the methods as provided herein further comprise one or more identifier sequences. In some cases, the identifier sequences comprise a non-hybridizable tail on the first and/or second primer. The identifier sequence can be a barcode sequence, a flow cell sequence, and/or an index sequence. In some cases, the index sequence is a Truseq primer sequence compatible with the next generation sequencing (NGS) platform produced by Illumina. In some cases, the

IDT Exhibit 1020.3727

first and/or second primer can bind to a solid surface. The solid surface can be a planar surface or a bead. The planar surface can be the surface of a chip, microarray, well, or flow cell. In some cases, the first and/or second primer comprises one or more sequence elements wherein products of the amplification reaction (i.e. amplification products) bind to a solid surface, whereby the one or more sequences elements are complementary to one or more capture probes attached to a solid surface.

[0088] The primers described herein can generally be oligonucleotides that are employed in an extension reaction by a polymerase along a template (such as a target polynucleotide or target sequence within a polynucleotide, target DNA, target RNA or a primer extension product), such as in PCR or cDNA synthesis, for example. An oligonucleotide primer can be a synthetic polynucleotide that is single stranded, containing a sequence at its 3′-end that is capable of hybridizing with a sequence of a target polynucleotide. Normally, the 3′ region of the primer that hybridizes with the target nucleic acid has at least 80%, 90%, 95%, 100%, complementarity to a sequence or primer binding site.

[0089] In some cases, one or more primers can be used in the methods provided herein to generate nucleic acid libraries comprising increased diversity relative to the same nucleic acid libraries being generated without the increased diversity. The primers used to generate the nucleic acid libraries comprising increased diversity can comprise extender sequence diversity sequence as provided herein. The primers can comprise sequence complementary to a target sequence in a polynucleotide as well as sequence that is non-complementary to the target sequence. The non-complementary sequence can be a tail and can comprise one or more diversity bases or extender sequence as described herein. The primers comprising one or more diversity bases can be introduced to one or more ends of the target sequence from a polynucleotide using a polymerase extension reaction (e.g., PCR) as provided herein. The primers comprising diversity base(s) (i.e., extender sequence) can be designed such that the diversity base or bases are immediately adjacent to the target sequence derived from a polynucleotide following introduction. In other words, the diversity base or bases are the first non-target sequence polynucleotide base or bases present at the junction of a complex comprising a target sequence from a polynucleotide and non-target sequence derived sequence. Also as provided herein, the primers comprising diversity bases can comprise a sequence or complement thereof that is capable of hybridizing a sequencing primer and this sequence or complement thereof can be located adjacent to the diversity base or bases such that the diversity base or bases are the bases initially sequenced in a sequencing reaction. In some cases, there can also be additional sequence elements located between the sequence or complement thereof that is

IDT Exhibit 1020.3728

capable of hybridizing a sequencing primer and the diversity base or bases such as barcodes (e.g., inline barcodes), universal sequence, index sequence, or random sequence. In some cases, the additional sequence elements (e.g., barcode(s), universal sequence and/or random sequence) can be located after the sequence or complement thereof that is capable of hybridizing a sequencing primer and distal to the diversity base or bases. In some cases, the additional sequence elements (e.g., barcode(s), universal sequence and/or random sequence) can be located after the sequence or complement thereof that is capable of hybridizing a sequencing primer and distal to the diversity base or bases as well as between the sequence or complement thereof that is capable of hybridizing a sequencing primer and the diversity base or bases such as barcodes (e.g., inline barcodes), universal sequence, or random sequence. In some cases, the primers comprising diversity bases used in the methods provided herein can be used to generate low diversity libraries such as 16S rRNA libraries or RRBS libraries such that the low diversity libraries comprise increased diversity versus low diversity libraries generated without the use of primers comprising diversity bases. The primers comprising diversity bases can be mixed into a single pool thereby comprising a mixture of primers comprising diversity bases of varying lengths, wherein some of the primers in the pool comprise no diversity bases, while some comprise 1 diversity bases, some comprise 2 diversity bases, some comprise 3 diversity bases, some comprise 4 diversity bases, some comprise 5 diversity bases, some comprise 6 diversity bases, some comprise 7 diversity bases, some comprise 8 diversity bases, some comprise 9 diversity bases, some comprise 10 diversity bases, some comprise 11 diversity bases, and some comprise 12 diversity bases. Each primer in the pool of primers can comprise sequence complementary to the same target sequence. Each primer in the pool of primers can further comprise sequence that can be hybridized by a sequencing primer or a primer that can introduce a common sequencing primer target sequence. The single pool of primers comprising diversity bases of varying lengths can be introduced to a target sequence (e.g., 16S rRNA) present in polynucleotides from multiple samples. As described herein, the diversity bases in a primer comprising diversity bases can be "D" bases, "R" bases, complements of "D" bases, or complements of "R" bases or combinations thereof.

[0090] "Complementary", as used herein, can refer to complementarity to all or only to a portion of a sequence. The number of nucleotides in the hybridizable sequence of a specific oligonucleotide primer can be such that stringency conditions used to hybridize the oligonucleotide primer can prevent excessive random non-specific hybridization. In some cases, the number of nucleotides in the hybridizing portion of the oligonucleotide primer can be at least as great as the defined sequence on the target polynucleotide that the oligonucleotide primer

IDT Exhibit 1020.3729

hybridizes to, namely, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least about 20, and generally from about 6 to about 10 or 6 to about 12 of 12 to about 200 nucleotides, usually about 10 to about 50 nucleotides. In general, the target polynucleotide can be larger than the oligonucleotide primer or primers as described previously.

[0091] In some cases, the identity of the investigated target polynucleotide sequence is known, and hybridizable primers can be synthesized precisely according to the antisense sequence of the aforesaid target polynucleotide sequence. In other cases, when the target polynucleotide sequence is unknown, the hybridizable sequence of an oligonucleotide primer is a random sequence. Oligonucleotide primers comprising random sequences may be referred to as "random primers", as described below. In yet other cases, an oligonucleotide primer such as a first primer or a second primer comprises a set of primers such as for example a set of first primers or a set of second primers. In some cases, the set of first or second primers may comprise a mixture of primers designed to hybridize to a plurality (e.g. 2, 3, 4, about 6, 8, 10, 20, 40, 80, 100, 125, 150, 200, 250, 300, 400, 500, 600, 800, 1000, 1500, 2000, 2500, 3000, 4000, 5000, 6000, 7000, 8000, 10,000, 20,000, 25,000, 30,000, 40,000, 50,000, 60,000, 70,000, 80,000, 90,000, 100,000 or more) of target sequences. In some cases, the plurality of target sequences may comprise a group of related sequences, random sequences, a whole transcriptome or fraction (e.g. substantial fraction) thereof, or any group of sequences such as mRNA.

[0092] "Barcode" can refer to a known polynucleotide sequence that allows some feature of a polynucleotide with which the barcode is associated to be identified. In some cases, the feature of the polynucleotide to be identified is the sample from which the polynucleotide is derived. In some cases, barcodes are at least 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, or more nucleotides in length. In some cases, barcodes are shorter than 10, 9, 8, 7, 6, 5, or 4 nucleotides in length. A oligonucleotide (e.g., primer or adaptor) can comprise about, more than, less than, or at least 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 different barcodes. In some cases, barcodes associated with some polynucleotides are of different length than barcodes associated with other polynucleotides. Barcodes can be of sufficient length and comprise sequences that can be sufficiently different to allow the identification of samples based on barcodes with which they are associated. In some cases, a barcode, and the sample source with which it is associated, can be identified accurately after the mutation, insertion, or deletion of one or more nucleotides in the barcode sequence, such as the mutation, insertion, or deletion of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more nucleotides. In some cases, each barcode in a plurality of barcodes differ from every other barcode in the plurality at at least three nucleotide positions, such as at least 3, 4, 5, 6, 7, 8, 9, 10, or more

IDT Exhibit 1020.3730

positions. In some cases, an oligonucleotide (e.g., adaptor or primer) comprises at least one of a plurality of barcode sequences. In some cases, barcodes for a second oligonucleotide (e.g., adaptor or primer) are selected independently from barcodes for a first oligonucleotide (e.g., adaptor or primer). In some cases, first oligonucleotides (e.g., adaptor or primer) and second oligonucleotides (e.g., adaptor or primer) having barcodes are paired, such that oligonucleotides (e.g., adaptors or primers) of the pair comprise the same or different one or more barcodes. In some cases, the methods described herein further comprise identifying the sample from which a target polynucleotide is derived based on a barcode sequence to which the target polynucleotide is joined. A barcode can comprise a oligonucleotide sequence that when joined to a target polynucleotide serves as an identifier of the sample from which the target polynucleotide was derived.

[0093] "Extender sequence" can refer to a stretch of nucleotides that can vary in length and base composition and can serve to increase the distance between the 3' end of a sequencing primer hybridized to a sequence complementary to the sequencing primer and the first nucleotide of a target sequence in a polynucleotide to be read during a nucleotide sequencing reaction. "Extender sequence" can be used interchangeably with "diversity sequence" or "diversity base."

[0094] As provided herein, the sequence complementary to the sequencing primer and the extender sequence can be appended or introduced to the first nucleotide of a target sequence in a polynucleotide through ligation or amplification such that the extender sequence is immediately upstream of the first base with respect to the direction of sequencing and thus will be sequenced prior to the first base. The extender sequence can be 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more nucleotides in length. In some cases, the extender sequence and the sequence complementary to a sequencing primer are present in the same oligonucleotide sequence. As provided herein, the oligonucleotide sequence can be present in a primer or adaptor. The extender sequence and the sequence complementary to a sequencing primer can be immediately adjacent to each other or can be separated by a stretch of nucleotides. The stretch of nucleotides can be an additional sequence element. The additional sequence element can be one or more barcodes as provided herein, universal sequence, linker sequence, or random sequence. The nucleotide base composition of each nucleotide in an extender sequence comprising more than one nucleotide can be identical or different. The nucleotide base composition of each nucleotide in an extender sequence can be randomly assigned. The nucleotide base composition of each nucleotide in an extender sequence can be semi-randomly assigned such that each nucleotide in an extender sequence (or diversity sequence) can only comprise 2 or 3 of the four possible canonical bases (i.e., thymine, guanine, cytosine, or adenine) such that one or two

IDT Exhibit 1020.3731

specific bases can be excluded. The nucleotide base composition of each nucleotide in an extender sequence can have a 25% chance of being one of the four unmodified bases (i.e., thymine, guanine, cytosine, or adenine). The nucleotide base composition of each nucleotide in an extender sequence can be the same nucleotide base. The nucleotide base composition of an extender sequence comprising more than one nucleotide can be constructed such that one or more nucleotide positions can have a fixed base composition, while the remaining nucleotides can have a randomly or semi-randomly assigned base composition as described herein.

[0095] RNA-dependent DNA polymerases for use in the methods and compositions provided herein can be capable of effecting extension of a primer according to the methods provided herein. Accordingly, an RNA-dependent DNA polymerase can be one that is capable of extending a nucleic acid primer along a nucleic acid template that is comprised at least predominantly of ribonucleotides. Suitable RNA-dependent DNA polymerases for use in the methods, compositions, and kits provided herein include reverse transcriptases (RTs). RTs are well known in the art. Examples of RTs include, but are not limited to, Moloney murine leukemia virus (M- MLV) reverse transcriptase, human immunodeficiency virus (HIV) reverse transcriptase, rous sarcoma virus (RSV) reverse transcriptase, avian myeloblastosis virus (AMV) reverse transcriptase, rous associated virus (RAV) reverse transcriptase, and myeloblastosis associated virus (MAV) reverse transcriptase or other avian sarcoma-leukosis virus (ASLV) reverse transcriptases, and modified RTs derived therefrom. See e.g. US7056716. Many reverse transcriptases, such as those from avian myeoloblastosis virus (AMV-RT), and Moloney murine leukemia virus (MMLV-RT) comprise more than one activity (for example, polymerase activity and ribonuclease activity) and can function in the formation of the double stranded cDNA molecules. However, in some instances, it is preferable to employ a RT which lacks or has substantially reduced RNase H activity. RTs devoid of RNase H activity are known in the art, including those comprising a mutation of the wild type reverse transcriptase where the mutation eliminates the RNase H activity. Examples of RTs having reduced RNase H activity are described in US20100203597. In these cases, the addition of an RNase H from other sources, such as that isolated from E. coli, can be employed for the degradation of the starting RNA sample and the formation of the double stranded cDNA. Combinations of RTs can also contemplated, including combinations of different non-mutant RTs, combinations of different mutant RTs, and combinations of one or more non-mutant RT with one or more mutant RT.

[0096] DNA-dependent DNA polymerases for use in the methods and compositions provided herein can be capable of effecting extension of a primer according to the methods provided herein. Accordingly, a DNA-dependent DNA polymerase can be one that is capable of

IDT Exhibit 1020.3732

extending a nucleic acid primer along a first strand cDNA in the presence of the RNA template or after selective removal of the RNA template. Exemplary DNA dependent DNA polymerases suitable for the methods provided herein include but are not limited to Klenow polymerase, with or without 3'-exonuclease, Bst DNA polymerase, Bca polymerase, .phi.29 DNA polymerase, Vent polymerase, Deep Vent polymerase, Taq polymerase, T4 polymerase, and E. coli DNA polymerase 1, derivatives thereof, or mixture of polymerases. In some cases, the polymerase does not comprise a 5'-exonuclease activity. In other cases, the polymerase comprises 5' exonuclease activity. In some cases, the primer extension can be performed using a polymerase comprising strong strand displacement activity such as for example Bst polymerase. In other cases, the primer extension can be performed using a polymerase comprising weak or no strand displacement activity. One skilled in the art can recognize the advantages and disadvantages of the use of strand displacement activity during the primer extension step, and which polymerases can be expected to provide strand displacement activity (see e.g., New England Biolabs Polymerases). For example, strand displacement activity can be useful in ensuring whole transcriptome coverage during the random priming and extension step. Strand displacement activity can further be useful in the generation of double stranded amplification products during the priming and extension step. Alternatively, a polymerase which comprises weak or no strand displacement activity can be useful in the generation of single stranded nucleic acid products during primer hybridization and extension that can be hybridized to the template nucleic acid.

[0097] In some cases, the double stranded products generated by the methods described herein can be end repaired to produce blunt ends for the adapter ligation applications described herein. Generation of the blunt ends on the double stranded products can be generated by the use of a single strand specific DNA exonuclease such as for example exonuclease 1, exonuclease 7 or a combination thereof to degrade overhanging single stranded ends of the double stranded products. Alternatively, the double stranded products can be blunt ended by the use of a single stranded specific DNA endonuclease for example but not limited to mung bean endonuclease or S1 endonuclease. Alternatively, the double stranded products can be blunt ended by the use of a polymerase that comprises single stranded exonuclease activity such as for example T4 DNA polymerase, any other polymerase comprising single stranded exonuclease activity or a combination thereof to degrade the overhanging single stranded ends of the double stranded products. In some cases, the polymerase comprising single stranded exonuclease activity can be incubated in a reaction mixture that does or does not comprise one or more dNTPs. In other cases, a combination of single stranded nucleic acid specific exonucleases and one or more polymerases can be used to blunt end the double stranded products of the primer extension

IDT Exhibit 1020.3733

reaction. In still other cases, the products of the extension reaction can be made blunt ended by filling in the overhanging single stranded ends of the double stranded products. For example, the fragments can be incubated with a polymerase such as T4 DNA polymerase or Klenow polymerase or a combination thereof in the presence of one or more dNTPs to fill in the single stranded portions of the double stranded products. Alternatively, the double stranded products can be made blunt by a combination of a single stranded overhang degradation reaction using exonucleases and/or polymerases, and a fill-in reaction using one or more polymerases in the presence of one or more dNTPs.

[0098] In another embodiment, the adapter ligation applications described herein can leave a nick between a non-ligation strand of the adapters and a strand of the double stranded product. In these instances, a gap repair or fill-in reaction can be used to append the double stranded product with the sequence complementary to the ligation strand of the adapter. Gap repair can be performed with any number of DNA dependent DNA polymerase described herein. In some cases, gap repair can be performed with a DNA dependent DNA polymerase with strand displacement activity. In some cases, gap repair can be performed using a DNA dependent DNA polymerase with weak or no strand displacement activity. In some cases, the ligation strand of the adapter can serve as the template for the gap repair or fill-in reaction. In some cases, gap repair can be performed using *Taq* DNA polymerase.

[0099] Various ligation processes and reagents are known in the art and can be useful for carrying out the methods provided herein. For example, the NuGEN Ovation ® Ultralow Methyl-Seq Library System can be used. In some cases, blunt end ligation is employed. Similarly, a single dA nucleotide can be added to the 3′-end of the double-stranded DNA product, by a polymerase lacking 3′-exonuclease activity and can anneal to an adaptor (e.g., diversity adaptor as provided herein) comprising a dT overhang (or the reverse). This design can allow the hybridized components to be subsequently ligated (e.g., by T4 DNA ligase). Other ligation strategies and the corresponding reagents known in the art can be employed and kits and reagents for carrying out efficient ligation reactions are commercially available (e.g, from New England Biolabs, Roche) and can be used for the methods provided herein.

[00100] "Bisulfite" as used herein can encompass all types of bisulfites, such as sodium bisulfite that are capable of chemically converting a cytosine (C) to a uracil (U) without chemically modifying a methylated cytosine and therefore can be used to differentially modify a DNA sequence based on the methylation status of the DNA.

[00101] "Low diversity" can refer to a polynucleotide. For example, the methods, compositions, and kits described herein can be used with pools of polynucleotides that have low diversity at

IDT Exhibit 1020.3734

one or more positions at one or both ends of polynucleotides in the pool. A "position" can be a single nucleotide, or 2 or more nucleotides. A "position" can be 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 nucleotides in length. A "position" can be a nucleotide sequence repeat such as CpG or a conserved region of a gene.

[00102] A polynucleotide can have low diversity at a position among polynucleotides if the position among polynucleotides has only one, two, or three nucleotide bases out of the four canonical DNA nucleotide bases. In some cases, a polynucleotide can have low diversity at a position among polynucleotides if it lacks one of the types of nucleotide bases present in the polynucleotide. For example, a pool of polynucleotides can one type of nucleotide base, e.g., only A, only T, only C, or only G at a position in the polynucleotides in the pool. In some cases, a pool of polynucleotides has only two types of polynucleotides at a position, e.g., only A or C, only A or G, only A or T, only C or T, only G or T, or only C or G. In some cases, a pool of polynucleotides has only three types of polynucleotides at a position, e.g., only A, C, and G; only A, C, and T; only C, T, and G.

[00103] In some cases, a pool of polynucleotides can have low diversity at a position if all types of nucleotide bases (e.g., A, T, C, and G) are at the position among polynucleotides in the pool, but the percentage of polynucleotides with one, two, or three types of bases at the position is low. For example, a pool of polynucleotides can have all four canonical bases found in DNA at the position, but one of the types of nucleotide bases is present at that position in less than 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1% of the polynucleotides. For example, a pool polynucleotides can have A, T, C, and G at a position, where 50% of polynucleotides have an A at the position, 20% have a T at the position, 20% have a C at the position, and only 10% have a G at the position.

[00104] In some cases, a pool of polynucleotides can have all four canonical bases found in DNA at a position among the polynucleotides in the pool, but each of two of the types of nucleotide bases is present at that position in less than 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1% of the polynucleotides. For example, a pool polynucleotides can have A, T, C, and G at a position, where 50% of polynucleotides have an A at the position, 40% have a T at the position, 5% have a C at the position, and 5% have a G at the position.

[00105] A pool of polynucleotides can have low diversity at a position if all four canonical bases found in DNA are at the position among the polynucleotides in the pool, but each of three of the types of nucleotide bases is present at that position in less than 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1% of the polynucleotides. For example, a pool polynucleotides can have A, T, C,

IDT Exhibit 1020.3735

and G at a position, where 94% of polynucleotides have an A at the position, 2% have a T at the position, 2% have a C at the position, and 2% have a G at the position.

[00106] A pool of polynucleotides can have any of the characteristics of a position described herein at 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or more positions. A pool of polynucleotides can have any of the characteristics of a position described herein at 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or more positions immediately 5' to base that anneals to the 3' end of a primer, e.g., a sequencing primer. A pool of polynucleotides can have any of the characteristics of a position described herein at 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or more positions immediately 5' to base that anneals to the 3' end of a primer, e.g., a sequencing primer.

[00107] In some cases, a pool of polynucleotides with low diversity at one or both ends is generated using restriction enzyme. In some cases the restriction enzyme has a palindromic recognition sequence. In some cases, a recognition site comprises a methylation site. In some cases, a palindromic recognition site comprises a methylation site. For example, a pool of polynucleotides can be generated by using the restriction enzyme MspI, which can have the recognition sequence CCGG, where CG can be a methylation site.

[00108] In some cases, a pool of polynucleotides with low sequence diversity is generated using an amplification based method, for example, polymerase chain reaction (PCR).

IV. Methods of Fragmentation

[00109] The methods provided herein can comprise fragmenting a polynucleotide to produce polynucleotide fragments. In some cases, the polynucleotides are double-stranded and fragmentation of the polynucleotides generates double stranded polynucleotide fragments. In some cases, fragmentation can be achieved through methods known in the art. Fragmentation can be through physical fragmentation methods and/or enzymatic fragmentation methods. Physical fragmentation methods can include nebulization, sonication, and/or hydrodynamic shearing. In some cases, the fragmentation can be accomplished mechanically comprising subjecting the nucleic acid to acoustic sonication. In some cases, the fragmentation comprises treating the nucleic acid with one or more enzymes under conditions suitable for the one or more enzymes to generate breaks in the double-stranded nucleic acid. Examples of enzymes useful in the generation of nucleic acid fragments include sequence specific and non-sequence specific nucleases. Non-limiting examples of nucleases include DNase I, Fragmentase, restriction endonucleases, variants thereof, and combinations thereof. Reagents for carrying out enzymatic fragmentation reactions are commercially available (e.g, from New England Biolabs). For example, digestion with DNase I can induce random double-stranded breaks in DNA in the absence of Mg^{++} and in the presence of Mn^{++}. In some cases, fragmentation comprises treating

IDT Exhibit 1020.3736

DNA with one or more restriction endonucleases. Fragmentation can produce fragments having 5' overhangs, 3' overhangs, blunt ends, or a combination thereof. In some cases, such as when fragmentation comprises the use of one or more restriction endonucleases, cleavage of the DNA leaves overhangs having a predictable sequence. In some cases, the methods provided herein are used on to generate nucleic acid libraries from polynucleotides fragmented by digestion with restriction enzymes. In some cases, the method includes the step of size selecting the fragments via standard methods known in the art such as column purification or isolation from an agarose gel.

[00110] In some cases, the methods, compositions, and kits provided herein are used to generate a reduced representation bisulfite sequencing (RRBS) library. An initial step in the generation of an RRBS library (see **FIG. 1**, **FIG. 3,** or **FIG. 36**) can comprise fragmenting a polynucleotide or a sample comprising polynucleotides by treating the polynucleotide or sample comprising polynucleotides with a methylation insensitive restriction enzyme. The methylation insensitive restriction enzyme can be AatII, AccB7I, AccIII, Acc65I, ApaI, AvaI, BanII, BbuI, BclI, BamHI, BalI, BglII, Bsp12861, BssHII, BstEII, BstOI, BstXI, ClaI, CaspI, Csp45I, DdeI, Eco47III, EcoRI, HaeIII, HhaI, HincII, HindIII, KpnI, MboII, MluI, MspI, NarI, NdeII, NruI, PstI, PvuII, SacI, ScaI, SinI, SnaBI, SphI, StuI, TaqI, XbaI, XhoII, or XmaI. In some cases, fragmenting a polynucleotide or a sample comprising polynucleotides comprises treating the polynucleotide or the sample comprising polynucleotides with MspI, thereby generating polynucleotide fragments comprising a 5'-CG-3" overhang at each end. The polynucleotides treated with MspI can be from mammalian sources, whereby the CCGG recognition site of MspI is not methylated at the first cytosine residue.

[00111] In some cases, the polynucleotide, for example DNA, can be fragmented into a population of fragmented polynucleotides of one or more specific size range(s). In some cases, the fragments can have an average length from about 10 to about 10,000 nucleotides or base pairs. In some cases, the fragments have an average length from about 50 to about 2,000 nucleotides or base pairs. In some cases, the fragments have an average length from about 100 to about 2,500, about 10 to about1000, about 10 to about 800, about 10 to about 500, about 50 to about 500, about 50 to about 250, or about 50 to about 150 nucleotides or base pairs. In some cases, the fragments have an average length less than 10,000 nucleotides or bp, less than 7,500 nucleotides or bp, less than 5,000 nucleotides or bp, less than 2,500 nucleotides or bp, less than 2,000 nucleotides or bp, less than 1,500 nucleotides or bp, less than 1,000 nucleotides or bp, less than 500 nucleotides or bp, less than 400 nucleotides or bp, less than 300 nucleotides or bp, less than 200 nucleotides or bp, or less than 150 nucleotides or bp. In some cases, the polynucleotide

IDT Exhibit 1020.3737

fragments have an average length of about, more than, less than, or at least 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, or 10,000 nucleotides or base pairs.

[00112] In some cases, polynucleotide fragments generated by fragmentation are subjected to end repair. End repair can include the generation of blunt ends, non-blunt ends (i.e. sticky or cohesive ends), or single base overhangs such as the addition of a single dA nucleotide to the 3′-end of the double-stranded nucleic acid product by a polymerase lacking 3′-exonuclease activity. In some cases, end repair is performed on the double stranded polynucleotide fragments to produce blunt ends wherein the ends of the polynucleotide fragments contain 5' phosphates and 3' hydroxyls. End repair can be performed using any number of enzymes and/or methods known in the art. An overhang can comprise about, more than, less than, or at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 nucleotides.

[00113] In some cases, the polynucleotide fragments are captured using a binding agent directed against an epigenetic modification within the sequence of the polynucleotide fragments. The epigenetic modification can be methylation. In some cases, the double stranded polynucleotide fragments are captured using a binding agent directed against 5-methylcytosine residues in the double-stranded polynucleotide fragments. The binding agent can be an antibody, or the binding domain of a protein directed against 5-methylcytosine residues. The protein can be a methyl-CpG-binding domain (MBD) protein. The MBD protein can be methyl-CpG-binding domain protein 1, 2, 4, or MECP2. In some cases, the double stranded polynucleotide fragments are captured using the binding domain of MBD2. In some cases, the double stranded polynucleotide fragments are captured using the binding domain of MECP2.

V. Methods of Amplification

[00114] The methods, compositions and kits described herein can be useful to generate amplification-ready products for downstream applications such as massively parallel sequencing (i.e. next generation sequencing methods) or hybridization platforms. As provided herein, the method, compositions and kits can be used to generate nucleic acid library (e.g., bisulfite converted libraries) such that each nucleic acid insert in the library comprises non-complementary sequence in opposing termini. Primer pairs comprising primers designed to target the non-complementary sequence can be used to amplify the nucleic acid inserts of a nucleic acid library prepared using the methods provided herein using the methods of amplification as provided herein. Methods of amplification are well known in the art. In some

embodiments, the amplification is exponential, e.g. in the enzymatic amplification of specific double stranded sequences of DNA by a polymerase chain reaction (PCR). In other embodiments the amplification method is linear. In other embodiments the amplification method is isothermal. Examples of PCR techniques that can be used include, but are not limited to, quantitative PCR, quantitative fluorescent PCR (QF-PCR), multiplex fluorescent PCR (MF-PCR), real time PCR(RT-PCR), single cell PCR, restriction fragment length polymorphism PCR(PCR-RFLP), PCR-RFLP/RT-PCR-RFLP, hot start PCR, nested PCR, in situ polony PCR, in situ rolling circle amplification (RCA), bridge PCR, picotiter PCR, digital PCR, droplet digital PCR, and emulsion PCR. Other suitable amplification methods include the ligase chain reaction (LCR), transcription amplification, molecular inversion probe (MIP) PCR, self-sustained sequence replication, selective amplification of target polynucleotide sequences, consensus sequence primed polymerase chain reaction (CP-PCR), arbitrarily primed polymerase chain reaction (AP-PCR), degenerate oligonucleotide-primed PCR (DOP-PCR) and nucleic acid based sequence amplification (NABSA), single primer isothermal amplification (SPIA, see e.g. U.S. Pat. No. 6,251,639), Ribo-SPIA, or a combination thereof. Other amplification methods that can be used herein include those described in U.S. Pat. Nos. 5,242,794; 5,494,810; 4,988,617; and 6,582,938. Amplification of target nucleic acids can occur on a bead. In other embodiments, amplification does not occur on a bead. Amplification can be by isothermal amplification, e.g., isothermal linear amplification. A hot start PCR can be performed wherein the reaction is heated to 95°C. for two minutes prior to addition of the polymerase or the polymerase can be kept inactive until the first heating step in cycle 1. Hot start PCR can be used to minimize nonspecific amplification. Other strategies for and aspects of amplification are described in U.S. Patent Application Publication No. 2010/0173394 A1, published Jul. 8, 2010, which is incorporated herein by reference. In some cases, the amplification methods can be performed under limiting conditions such that only a few rounds of amplification (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 etc.), such as for example as is commonly done for cDNA generation. The number of rounds of amplification can be about 1-30, 1-20, 1-15, 1-10, 5-30, 10-30, 15-30, 20-30, 10-30, 15-30, 20-30, or 25-30.

[00115] Techniques for amplification of target and reference sequences are known in the art and include the methods described in U.S. Pat. No. 7,048,481. Briefly, the techniques can include methods and compositions that separate samples into small droplets, in some instances with each containing on average less than about 5, 4, 3, 2, or one target nucleic acid molecule (polynucleotide) per droplet, amplifying the nucleic acid sequence in each droplet and detecting the presence of a target nucleic acid sequence. In some cases, the sequence that is amplified is

present on a probe to the genomic DNA, rather than the genomic DNA itself. In some cases, at least 200, 175, 150, 125, 100, 90, 80, 70, 60, 50, 40, 30, 20, 10, or 0 droplets have zero copies of a target nucleic acid.

[00116] PCR can involve in vitro amplification based on repeated cycles of denaturation, oligonucleotide primer annealing, and primer extension by thermophilic template dependent polynucleotide polymerase, which can result in the exponential increase in copies of the desired sequence of the polynucleotide analyte flanked by the primers. In some cases, two different PCR primers, which anneal to opposite strands of the DNA, can be positioned so that the polymerase catalyzed extension product of one primer can serve as a template strand for the other, leading to the accumulation of a discrete double stranded fragment whose length is defined by the distance between the 5' ends of the oligonucleotide primers.

[00117] LCR uses a ligase enzyme to join pairs of preformed nucleic acid probes. The probes can hybridize with each complementary strand of the nucleic acid analyte, if present, and ligase can be employed to bind each pair of probes together resulting in two templates that can serve in the next cycle to reiterate the particular nucleic acid sequence.

[00118] SDA (Westin et al 2000, Nature Biotechnology, 18, 199-202; Walker et al 1992, Nucleic Acids Research, 20, 7, 1691-1696), can involve isothermal amplification based upon the ability of a restriction endonuclease such as HincII or BsoBI to nick the unmodified strand of a hemiphosphorothioate form of its recognition site, and the ability of an exonuclease deficient DNA polymerase such as Klenow exo minus polymerase, or Bst polymerase, to extend the 3'-end at the nick and displace the downstream DNA strand. Exponential amplification results from coupling sense and antisense reactions in which strands displaced from a sense reaction serve as targets for an antisense reaction and vice versa.

[00119] Some aspects of the methods described herein can utilize linear amplification of nucleic acids or polynucleotides. Linear amplification can refer to a method that involves the formation of one or more copies of the complement of only one strand of a nucleic acid or polynucleotide molecule, usually a nucleic acid or polynucleotide analyte. Thus, the primary difference between linear amplification and exponential amplification is that in the latter process, the product serves as substrate for the formation of more product, whereas in the former process the starting sequence is the substrate for the formation of product but the product of the reaction, i.e. the replication of the starting template, is not a substrate for generation of products. In linear amplification the amount of product formed increases as a linear function of time as opposed to exponential amplification where the amount of product formed is an exponential function of time.

IDT Exhibit 1020.3740

[00120] In some cases, the amplification is exponential, e.g. in the enzymatic amplification of specific double stranded sequences of DNA by a polymerase chain reaction (PCR). In other embodiments the amplification method is linear. In other embodiments the amplification method is isothermal.

VI. Generation of Bisulfite Converted Libraries

[00121] In one aspect, a method is provided for generating a directional, bisuflite-converted nucleic acid library using duplex-forming adaptors comprising diversity bases or extender sequence as described herein. The nucleic acid library generated using the duplex-forming adapters comprising diversity bases can maintain directional (strandedness) information of the original nucleic acid sample. In some cases, the original nucleic acid is DNA. In some cases, the DNA is double-stranded DNA. In some cases, the double-stranded DNA is genomic DNA. In some cases, the nucleic acid or polynucleotide is from a whole genome sample. In some cases, the methods provided herein comprise generating a plurality of bisulfite converted libraries. One of the plurality of bisulfite converted libraries can be generated using a duplex-forming adaptor lacking diversity bases, while others of the plurality of bisulfite converted libraries can be generated using a duplex-forming adaptor comprising a variable number of diversity bases. One of the others of the plurality of bisulfite converted libraries can be generated using a duplex-forming adaptor comprising 1 diversity base. One of the others of the plurality of bisulfite converted libraries can be generated using a duplex-forming adaptor comprising 2 diversity bases. One of the others of the plurality of bisulfite converted libraries can be generated using a duplex-forming adaptor comprising 3 diversity bases.

[00122] As provided herein, the diversity base or bases can be introduced to adjacent to the first nucleotide of sequence derived from the polynucleotide to which it is attached such that the diversity base is sequenced immediately prior to the sequencing of the polynucleotide to which it is attached. In some cases, each of the plurality of bisulfite converted libraries can comprises a polynucleotide comprising the same or substantially similar sequence across at least a portion of each polynucleotide. Each of the plurality of bisulfite converted libraries comprising the polynucleotide comprising the same or substantially similar sequence across at least a portion of each polynucleotide sequence can be pooled and subsequently processed (e.g., amplified and/or solid-state amplified) and sequenced in parallel. In some cases, the bisulfite converted libraries can be generated using the method described herein by introducing (e.g., ligating) a pool of adaptors comprising adaptors comprising no diversity bases as well as adaptors comprising a varying number of diversity bases as described herein to a sample comprising polynucleotides comprising the same or substantially similar sequence across at least a portion of each

IDT Exhibit 1020.3741

polynucleotide sequence. The sequencing can be performed using a commercially available NGS system such as Illumina. The presence of diversity bases can aid in cluster determination in sequencing systems such as Illumina sequencing which can suffer from poor cluster identification in sequencing samples with low sequence diversity in their initial nucleotide bases as shown in **FIG. 31**.

[00123] The method can comprise fragmenting a double stranded polynucleotide to produce double stranded polynucleotide fragments. In some cases, fragmentation can be achieved through methods known in the art. In some cases, fragmentation can be achieved through methods known in the art and provided herein. In some cases, the fragmentation comprises treating the nucleic acid with one or more enzymes under conditions suitable for the one or more enzymes to generate breaks in the double-stranded nucleic acid. In some cases, fragmentation comprises treating DNA with one or more restriction endonucleases. Fragmentation can produce fragments having 5' overhangs, 3' overhangs, blunt ends, or a combination thereof. In some cases, such as when fragmentation comprises the use of one or more restriction endonucleases, cleavage of the DNA leaves overhangs having a predictable sequence. In some cases, the method includes the step of size selecting the fragments via standard methods known in the art such as column purification or isolation from an agarose gel. In some cases, the methods provided herein are used to generate RRBS libraries such that a sample comprising polynucleotides is fragmented using a methylation insensitive enzyme such as MspI in order to generate fragments with a predictable 5'-CG-3' overhang and subsequently introduced (e.g., through ligation) to an adaptor as shown in **FIGs 1, 3** and **36**.

[00124] In some cases, polynucleotide fragments generated by fragmentation are subjected to end repair. End repair can include the generation of blunt ends, non-blunt ends (i.e. sticky or cohesive ends), or single base overhangs such as the addition of a single dA nucleotide to the 3′-end of the double-stranded nucleic acid product by a polymerase lacking 3′-exonuclease activity. In some cases, end repair is performed on the double stranded nucleic acid fragments to produce blunt ends wherein the ends of the polynucleotide fragments contain 5' phosphates and 3' hydroxyls. End repair can be performed using any number of enzymes and/or methods known in the art. An overhang can comprise about, more than, less than, or at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 nucleotides.

[00125] In some cases, double stranded polynucleotide fragments are captured using a binding agent directed against an epigenetic modification within the sequence of the polynucleotide fragments. The epigenetic modification can be methylation. In some cases, the double stranded polynucleotide fragments are captured using a binding agent directed against 5-methylcytosine

residues in the double-stranded polynucleotide fragments. The binding agent can be an antibody, or the binding domain of a protein directed against 5-methylcytosine residues. The protein can be a methyl-CpG-binding domain (MBD) protein. The MBD protein can be methyl-CpG-binding domain protein 1, 2, 4, or MECP2. In some cases, the double stranded polynucleotide fragments are captured using the binding domain of MBD2. In some cases, the double stranded polynucleotide fragments are captured using the binding domain of MECP2.

[00126] The method can further comprise ligating an adapter comprising either no diversity bases or a variable number of diversity bases as provided herein to the double-stranded polynucleotide fragments. Ligation can be blunt end ligation or sticky or cohesive end ligation. In some cases, the ligation is performed using cohesive end ligation as shown in **FIG. 3**. The ligation can be performed with any of the enzymes known in the art for performing ligation (e.g. T4 DNA ligase). The adapter can be any type of adapter known in the art including, but not limited to, a conventional duplex or double stranded adapter. The adapter can comprise DNA, RNA, or a combination thereof. The adapters can be about, less than about, or more than about 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 90, 100, or 200 nucleotides in length. The adapters can be a duplex adaptor, partial duplex adaptor, a forked adaptor or single stranded adaptor. In some cases, the adapter is a duplex adapter. In some cases, the duplex adapters comprises about, less than about, or more than about 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 90, 100, or 200 nucleotides in length. In some cases, the adapter is a partial duplex adapter, wherein the adapter comprises a long strand and a short strand. In some cases, a partial duplex adapter has overhangs of about, more than, less than, or at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 nucleotides. In some cases, the overhang is a 5' overhang. In some cases, the overhang is a 3' overhang. In some cases, the partial duplex adapter comprises a 5' and 3' overhang. In some cases, the adapter comprises duplexed sequence. In some cases, the adapters comprise about, more than, less than, or at least 5, 6, 7, 8, 9, 10, 12, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 90, 100, 200, or more of base paired or duplexed sequence. In some cases, the adapter comprises a single stranded adapter. In some cases, a single-stranded adapter comprises about, more than, less than, or at least 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 90, 100, or 200 nucleotides in length. In some cases, the single-stranded adapter forms a stem-loop or hairpin structure. In some cases, the stem of the hairpin adapter is about, less than about, or more than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 45, 50, 75, 100, or more nucleotides in length. In some cases, the loop sequence of a hairpin adapter is about, less than about, or more than about 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, or more nucleotides in

length. The adapter can further comprise known or universal sequence and, thus, allow generation and/or use of sequence specific primers for the universal or known sequence. In some cases, an adapter comprises one or more barcodes. In some cases, the one or more barcodes are in a stem and/or a loop.

[00127] In some cases, an adapter is marked via incorporation of at least one modified dNTP. In some cases, the modified dNTP comprises a nucleotide analog resistant to conversion by treatment with a converting agent. The nucleotide analog can be a cytosine analog. The converting agent can be any biological, biochemical, and/or chemical agent capable of altering the base composition of a dNTP. In some cases, the converting agent is a chemical. In some cases, the converting agent is the chemical compound bisulfite or sodium bisulfite. In some cases, the adapter comprises a cytosine analog resistant to conversion by bisulfite treatment. In some cases, the long strand of a partial duplex adapter comprises cytosine analog residues in place of cytosine residues, which are protected from bisulfite conversion, while the short strand of the partial duplex adapter does not comprise cytosine analog residues in place of cytosine residues. In some cases, the short strand of a partial duplex adapter comprises cytosine analog residues in place of cytosine residues, which are protected from bisulfite conversion, while the long strand of the partial duplex adapter does not comprise cytosine analog residues in place of cytosine residues. In some cases, both the long and short strand of a partial duplex adapter comprises cytosine analog residues in place of cytosine residues. In some cases, the cytosine analog is 5-methylcytosine. In some cases, the cytosine analog is 5-hydroxymethylcytosine. In some cases, the cytosine analog is 5-propynylcytosine. A strand can comprise a modified cytosine at about, more than, less than, or at least every 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 40, 50, 65, 75, 85, 90, 95, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 225, or 250 nucleotides. In some cases, ligation of an adapter to a double stranded polynucleotide is by blunt end ligation. In some cases, ligation of an adapter to a double stranded polynucleotide is by cohesive or sticky end ligation, wherein an overhang in the adapter hybridizes to an overhang in the double stranded polynucleotide comprising complementary sequence such as shown in **FIGs. 3**. In some cases, an adapter comprising a modified dNTP (e.g. a cytosine analog resistant to bisulfite treatment) comprises a ligation strand or first strand capable of ligation to a 5'end of the polynucleotide fragments and a non-ligation strand or second strand incapable of ligation to either end of the polynucleotide fragments. In some cases, the duplex adapter is a partial duplex adapter, wherein the adapter comprises a long strand and a short strand, and wherein the long strand is the ligation strand or first strand, while the short strand is the non-ligation strand or second strand. In some cases, the partial duplex has strands of unequal length. In some cases,

IDT Exhibit 1020.3744

the partial duplex comprises an overhang at one end of the adapter and a blunt end at another end of the adapter. The overhang can be at the 3' end or the 5' end. In some cases, the partial duplex comprises an overhang at each end of the adapter. The overhang can be of equal length or unequal length. In some cases, the 5' end of the ligation strand does not comprise a 5' phosphate group. In some cases, the 5' end of the ligation strand does comprise a 5' phosphate, wherein the 3' end of the polynucleotide lacks a free 3' hydroxyl.

[00128] In some cases, the 3' and/or 5' ends of the non-ligation strand comprise a blocking group and are enzymatically unreactive. The blocking group or moiety on the 3' and/or 5' end can be any blocking group or moiety as described herein and/or known in the art. In some cases, the block at the 5' end of the non-ligation strand comprises a biotin moiety. In some cases, the 5' end of the non-ligation strand does not comprise a 5' phosphate. The 5' end can be removed by treatment with an enzyme. The enzyme can be a phosphatase. In some cases, the 5' end of the non-ligation strand is dephosphorylated by treatment with alkaline phosphatase. In some cases, the 5' end of the non-ligation strand does comprise a 5' phosphate, wherein the 3' end of the polynucleotide lacks a free 3' hydroxyl. In some cases, the non-ligation strand comprises a block at the 3' end comprising terminal dideoxycytosine and a block at the 5' end comprising a biotin moiety. In some cases, distinct adapters as described herein are ligated to a 5' end of a double strand polynucleotide.

[00129] In some cases, the adapter is a hairpin adapter comprising a stem-loop, wherein both strands of the stem comprise a modified dNTP (e.g. a cytosine analog resistant to bisulfite treatment). In some cases, the stem-loop adapter comprises a ligation or first strand and a non-ligation or second strand as described herein. In some cases, the 3' end of the stem comprises the ligation strand, while the 5' end of the stem comprises the non-ligation strand. In some cases, the 5' end of the stem does not comprise a 5' phosphate. In some cases, the 5' end of the stem comprises a 5' phosphate, while the 3' ends of the double strand polynucleotide lacks a free 3' hydroxyl. In some cases, the 5' end of the stem comprises a blocking group. The blocking group can be any of the blocking groups described herein. In some cases, the stem comprises an overhang. The overhang can be a 5' overhang or a 3' overhang and can comprise DNA, RNA, or both. A stem-loop adapter can be ligated to a double stranded polynucleotide by the methods described herein. In some cases, a stem loop adapter comprises a replication block. The replication block can be a non-replicable base or region in the loop or in a region of the stem adjacent to the loop comprising abasic sites. The replication block can comprise an inverted repeat. Abasic sites can be generated in the stem-loop by any of the methods known in the art, which can include, but is not limited to, incorporation of dUTP during generating of the adapter

IDT Exhibit 1020.3745

followed by treatment with dU-glycosylase (which is also referred to as Uracyl-DNA Glycosylase or UDG). In some cases, the replication block is removable or cleavable.

[00130] In some cases, the adapter comprises a ligation or first strand as described herein, and a non-ligation or second strand, wherein the non-ligation or second strand comprises RNA residues. In some cases, the adapter comprises a ligation or first strand as described herein, and a non-ligation or second strand, wherein the ligation or first strand comprises RNA residues.

[00131] In some cases, the ligation of an adapter to a first strand of a double stranded polynucleotide fragments creates a nick or break in the backbone between the non-ligation strand of the adapter and the 3' end of the second strand of the double-stranded polynucleotide fragments, wherein the non-ligation strand is not joined to the 3' end of the second strand of the polynucleotide fragments. In this case, the 5' end of the ligation strand does not comprise a 5' phosphate group. Further to this case, ligation of an adapter to the polynucleotide fragment can generate a polynucleotide fragment comprising the ligation strand comprising a cytosine analog joined to a first and second 5' end of the polynucleotide fragments. In some cases, the 5' end of the ligation strand comprises a 5' phosphate group, and the 3' ends of the polynucleotide fragment lacks a free 3' hydroxyl. Further to this case, ligation of an adapter to the polynucleotide fragment can generate a polynucleotide fragment comprising the ligation strand comprising a cytosine analog joined to a first and second 5' end of the polynucleotide fragments.

[00132] The method can further comprise performing an extension reaction. The extension reaction can be performed using any number of methods known in the art including, but not limited to, the use of a DNA dependent DNA polymerase with strand displacement activity or a DNA polymerase with exonuclease activity and all four dNTPs (i.e. dATP, dTTP, dCTP, and dGTP), wherein the dNTPs are unmodified. In some cases, adaptors (e.g., duplex or partial duplex) as described herein comprising cytosine analogs resistant to bisulfite treatment in a ligation strand as well as no or one or more diversity bases are ligated to a polynucleotide and further subject to an extension reaction that is performed with a DNA polymerase and unmodified dNTPs (i.e. dATP, dTTP, dCTP, and dGTP). In some cases, adaptors (e.g., duplex or partial duplex) as described herein comprising cytosine bases sensitive to bisulfite treatment in a ligation strand as well as no or one or more diversity bases are ligated to a polynucleotide and further subject to an extension reaction that is performed with a DNA polymerase and a dNTP mixture comprising dATP, dTTP, dGTP, and a dCTP analog resistant to bisulfite treatment. In some cases, the extension reaction extends the 3' ends of the polynucleotide fragments, whereby a non-ligation strand of an adapter is removed. The non-ligation strand can be removed by being displaced, degraded, or denatured. In some cases, the non-ligation strand of the joined adapter is

IDT Exhibit 1020.3746

removed by heat denaturation, and the 3' ends of the polynucleotide fragment are extended with a polymerase without strand displacement activity. In some cases, the melting temperature of the non-ligation strand bound to the ligation strand can be lower than the melting temperature of the two strands of the polynucleotide fragment to which the ligation strand of the adapter is joined. In some cases, the non-ligation strand is displaced by a polymerase comprising strand displacement activity during extension of the 3' ends of the double stranded polynucleotide fragment. In some cases, the adapter is a hairpin adapter and the extension reaction displaces the non-ligation strand of the stem. In some cases, the displaced strand of the stem adapter remains connected to the ligation strand of the stem via the loop. In some cases, the loop comprises a cleavage site for an enzyme (i.e. restriction endonuclease). In some cases, the cleavage site is within a replication block. In some cases, the cleavage site is cleaved, thereby removing the non-ligation strand of the stem. In these cases, the ligation strand of the stem can comprise the modified nucleotide (i.e. nucleotide with cytosine analog resistant to bisulfite treatment). In some cases, the ligation strand serves as the template, wherein the extension reaction generates sequence complementary to the ligation strand. In some cases a single adapter is ligated to the 5' ends of the double stranded polynucleotide fragment, whereby extension of the 3' ends of the polynucleotide fragment generates polynucleotide fragments comprising complementary adapter sequences at the 3' and 5' ends. In some cases, the adapter ligated to the polynucleotide fragments comprises a non-ligation strand comprising RNA thereby forming a DNA/RNA heteroduplex with the ligation strand, wherein the extension reaction extends the 3' ends of the polynucleotide fragments following degradation of the RNA in the non-ligation strand using an agent capable of degrading RNA in a DNA/RNA heteroduplex. The agent can be an enzyme. The enzyme can be RNase H. In this embodiment, the ligation or first strand serves as the template, wherein the extension reaction generates sequence complementary to the ligation or first strand, thereby generating polynucleotide fragments comprising complementary adapter sequences at the 3' and 5' ends.

[00133] In some cases, the duplex adapter is a partial duplex adapter, wherein the adapter comprises a long strand and a short strand, wherein both the long strand and the short strand are capable of ligation. In some cases, the long strand comprises a modified dNTP (e.g. a cytosine analog resistant to bisulfite treatment). In some cases, the short strand comprises a modified dNTP (e.g. a cytosine analog resistant to bisulfite treatment). In these cases, the partial duplex adapter comprises a 5' overhang and a blunt end, or both a 5' and 3' overhang. In order to reduce the formation of primer dimers, the 3' end of the short arm of the adapter can comprise a blocking group and can be enzymatically unreactive. The blocking group can be any of the

IDT Exhibit 1020.3747

blocking groups described herein. In some cases, the short arm of the adapter comprises a reversible blocking group, wherein the reversible blocking group can be removed following ligation of the adapter to the double stranded polynucleotide. In some cases, unligated adapter is removed by washing and/or degradation following ligation and prior to removal of the reversible blocking group. In some cases, the method can further comprise performing an extension reaction. The extension reaction can be performed using any number of methods known in the art including, but not limited to, the use of a DNA dependent DNA polymerase with strand displacement activity and all four dNTPs (i.e. dATP, dTTP, dCTP, and dGTP), wherein the dNTPs are unmodified. In some cases, the extension reaction is performed with a DNA polymerase and unmodified dNTPs (i.e. dATP, dTTP, dCTP, and dGTP). In some cases, the extension reaction extends the 3' ends of short strand of the adapters ligated to the ends of the double stranded polynucleotide fragments, thereby generating polynucleotide fragments comprising complementary adapter sequences at the 3' and 5' ends. The cytosine analog resistant to bisulfite treatment can be 5-methylcytosine, 5-hydroxymethylcytosine or 5-propynylcytosine.

[00134] In some cases, double stranded polynucleotide fragments comprising adapter sequence at the 3' and 5' ends are captured prior to treatment with a converting agent. In some cases, the double stranded polynucleotide fragments are captured using a binding agent directed against modified dNTPs in the double-stranded polynucleotide fragments with adapters. The modified dNTP can be a modified dCTP comprising a cytosine analog. The cytosine analog can be 5-methylcytosine, 5-hydroxymethylcytosine or 5-propynylcytosine. The binding agent can be an antibody, or the binding domain of a protein directed against a cytosine analog. In some cases, the binding domain is directed against 5-methylcytosine residues. The binding domain can be from a methyl-CpG-binding domain (MBD) protein. The MBD protein can be methyl-CpG-binding domain protein 1, 2, 4, or MECP2. In some cases, the double stranded polynucleotide fragments are captured using the binding domain of MBD2. In some cases, the double stranded polynucleotide fragments are captured using the binding domain of MECP2. In some cases, one or both strands of the adapter sequence on the end(s) of the double stranded polynucleotide fragments comprise a cytosine analog other than 5-methylcytosine, wherein the double stranded polynucleotide fragments are captured using the binding domain of a methyl-CpG-binding domain (i.e. MBD2 or MECP2). The cytosine analog other than 5-methylcytosine can be 5-hydroxymethylcytosine or 5-propynylcytosine.

[00135] In some cases, the method further comprises a denaturing step, wherein the polynucleotide fragments comprising adaptor sequences at the 3' and 5' ends are denatured.

IDT Exhibit 1020.3748

Denaturation can be achieved using any of the methods known in the art which can include, but are not limited to, heat denaturation, and/or chemical denaturation. Heat denaturation can be performed by raising the temperature of the reaction mixture to be above the melting temperature of the polynucleotide fragments comprising adapter sequence at both ends. The melting temperature can be about, more than, less than, or at least 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, or 95 degrees C. The temperature can be raised above the melting temperature by about, more than, less than, or at least 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 degrees C. Chemical denaturation can be performed using bases (i.e. NaOH), and/or competitive denaturants (i.e. urea, or formaldehyde). In some cases, denaturation generates single stranded polynucleotide fragments comprising complementary adaptor sequence at the 3' and 5' ends such that the juncture between the fragments and the adaptors comprise one or more diversity bases. In some cases, denaturation generates single stranded polynucleotides fragments comprising distinct adapter sequence at the 3' and 5' ends such that the juncture between the fragments and the adaptors comprise one or more diversity bases.

[00136] In some cases, single stranded polynucleotide fragments comprising adapter sequence at the 3' and 5' ends are captured prior to treatment with a converting agent. In some cases, the polynucleotide fragments are captured by a binding agent directed against one or more modified dNTPs present in the adapter sequence. In some cases, the modified dNTP is a nucleotide base analog. In some cases, the binding agent is a binding protein. In some cases, the binding protein is an antibody directed against the modified dNTP. In some cases, the binding protein is an antibody directed against the modified dNTP, wherein the modified dNTP is a nucleotide analog. In some cases, the single stranded polynucleotide fragments comprising adapter sequence at the 3' and 5' ends are captured prior to treatment with a bisulfite. In some cases, the nucleic acid fragments (polynucleotides) are captured by a binding agent directed against one or more elements present in the adapter sequence. In some cases, the one or more elements comprise a cytosine analog. In some cases, the cytosine analog is 5-methylcytosine. In some cases, the binding agent is a 5-methylcytosine binding protein. In some cases, the binding protein is an anti-5-methylcytosine antibody. In some cases, 5-methylcytosine capture is performed prior to bisulfite treatment, wherein the cytosine analog resistant to bisulfite treatment is a cytosine analog other than 5-methylcytosine. In some cases, the cytosine analog can be 5-hydroxymethylcytosine or 5-propynylcyotsine. The one or more elements can be introduced during the extension reaction. In some cases, a modified nucleotide can be incorporated during

IDT Exhibit 1020.3749

the extension reaction, wherein the modified nucleotide contains a tag. The tag can be a biotin moiety. In some cases, the binding agent is avidin, streptavidin, or an anti-biotin antibody.

[00137] Following denaturation, and optional capture by a binding agent, the single-stranded polynucleotide fragments comprising adapter sequence at the 3' and 5' ends can be treated with a converting agent. In some cases, treatment of the single-stranded polynucleotide fragments with a converting agent alters the sequence of the complement of the ligation strand as well as the first and second strands of the double stranded polynucleotide fragment, while leaving the sequence of the ligation or first strand unchanged. In some cases, a single adapter is ligated to the 5' ends of the polynucleotide fragments, whereby treatment with a converting agent generates single stranded polynucleotide fragments comprising non-complementary sequence at the 5' and 3' ends. In some cases, distinct adapters are ligated to the 5' ends of the polynucleotide fragments, whereby treatment with a converting agent generates single stranded polynucleotide fragments wherein the non-ligation strands of the distinct adapters is altered to be non-complementary to the ligation strands of the distinct adapters. In some embodiments, the sequence of the ligation or first strand of the adapter marks the 5' end of the polynucleotide fragments, thereby maintaining the strandedness of the polynucleotide fragment and thus providing information on directionality.

[00138] In some cases, the single-stranded nucleic acid fragments are treated with a converting agent wherein the converting agent is bisulfite. In some cases, treatment of the single-stranded polynucleotide fragments converts cytosine residues in the polynucleotide fragment and the complement of the ligation or first strand to uracil residues while the cytosine analogs in the ligation or first strand are resistant to conversion. In some cases, treatment of the single stranded polynucleotide fragments with bisulfite generates single stranded polynucleotide fragments comprising non-complementary adapter sequence at the 5' and 3' ends. In some cases, the sequence of the ligation strand of the adapter unaltered by bisulfite treatment marks the 5' end of the polynucleotide fragments, thereby maintaining the strandedness of the polynucleotide fragment and thus providing information on directionality. In some cases, distinct adapters are ligated to the 5' ends of the polynucleotide fragments, whereby treatment with a bisulfite generates single stranded polynucleotide fragments wherein cytosine residues in the non-ligation strands of the distinct adapters are converted to uracil residues, whereby the sequence of the non-ligation strand is no longer complementary to the ligation strands of the distinct adapters.

[00139] In some cases, the method further comprises amplifying the single-stranded polynucleotide fragments comprising adapter sequences at the 3' and 5' ends. In some cases, amplification of the single-stranded polynucleotide fragments comprising adapter sequence at

IDT Exhibit 1020.3750

the 3' and 5' ends generates directional polynucleotide libraries. In some cases, one end of the polynucleotide fragment marks the orientation of the original polynucleotide strand to which it is appended due to its resistance to conversion by the converting agent, whereby the sequence in said end is resistant to conversion to a different sequence by treatment with the converting agent. In some cases, amplification of the single-stranded polynucleotide fragments comprising adapter sequence at the 3' and 5' ends generates directional polynucleotide libraries wherein one end of the polynucleotide fragments marks the orientation of the original polynucleotide strand to which it is appended due to its resistance to conversion by bisulfite treatment. In some cases, the cytosine residues present in said end are resistant to conversion to uracil residues by bisulfite treatment.

[00140] In some cases, amplifying the single stranded polynucleotide fragments comprising adapter sequence at the 3' and 5' ends comprises the use of a first primer and a second primer. In some cases, the first primer is directed against sequence complementary to the ligation or first strand of an adapter altered following treatment with a converting agent. In some cases, the second primer is directed against sequence complementary to the ligation or first strand of an adapter, wherein the ligation or first strand to which said complementary sequence is complementary is not altered by treatment with the converting agent. In some cases, the converting agent is bisulfite, whereby treatment with bisulfite converts cytosine residues in the sequence complementary to the ligation or first strand to uracil residues. In some cases, the first primer is directed against sequence complementary to the ligation or first strand of the adapter comprising uracil residues following bisulfite treatment. In some cases, the second primer is directed against sequence complementary to the ligation or first strand of the adapter, wherein the ligation or first strand to which said complementary sequence is complementary to does not contain uracil residues following bisulfite treatment. The single stranded polynucleotide fragments comprising adapter sequence at the 3' and 5' ends can represent a first strand of a double stranded polynucleotide fragment or a second strand of a double stranded polynucleotide fragment.

[00141] In some cases, a single adapter is ligated to the 5' ends of the polynucleotide, whereby the first and second strands can comprise non-complementary sequence following treatment with the converting agent (e.g., bisulfite treatment). In some cases, distinct adapters are ligated to the 5' ends of the polynucleotide fragments, whereby treatment with bisulfite generates single stranded polynucleotide fragments from a first strand of a double stranded polynucleotide fragment or a second strand of a double stranded polynucleotide fragment, wherein cytosine residues in the non-ligation strands of the distinct adapters are converted to uracil residues. In

IDT Exhibit 1020.3751

these cases, the sequence of the non-ligation strand is no longer complementary to the ligation strands of the distinct adapters. Amplifying the single stranded polynucleotide fragments comprising adapter sequence at the 3' and 5' ends can produce amplification products from either or both of the first and second strand of the double stranded polynucleotide fragment following treatment with the converting agent (i.e. bisulfite). In some cases, the first and/or second primer further comprises one or more identifier sequences. In some cases, the identifier sequences comprise a non-hybridizable tail on the first and/or second primer. The identifier sequence can be a barcode sequence, a flow cell sequence, and/or an index sequence. In some cases, the index sequence is a Truseq primer sequence compatible with the next generation sequencing platform produced by Illumina. In some cases, the first and/or second primer can bind to a solid surface. The solid surface can be a planar surface or a bead. The planar surface can be the surface of a chip, microarray, well, or flow cell. In some cases, the first and/or second primer comprises one or more sequence elements products of the amplification reaction (i.e. amplification products) to a solid surface, wherein the one or more sequences are complementary to one or more capture probes attached to a solid surface.

[00142] In some cases, methods for generating a polynucleotide library using modified duplex-forming adapters described herein further comprise determining the methylation status of the input double stranded polynucleotide. In some cases, the input polynucleotide is genomic DNA and the amplification of single-stranded polynucleotide fragments comprising non-complementary sequence at the 3' and 5' ends is followed by sequencing. The non-complementary sequence at the 3' ends of the single-stranded polynucleotide fragments can be complementary to reverse read sequencing primers, while the 5' ends of the single-stranded polynucleotide fragments can be complementary to forward read sequencing primers. This differentially complementarity to forward and reverse sequencing primers can result from the bisulfite treatment differentially affecting the sequence of the 3' and 5' ends based on the presence or absence of cytosine analogs resistant to bisulfite treatment in the 3' and 5' ends as described herein. Further to this embodiment, the methylation status of the genomic DNA can be determined by comparing the sequence obtained from the sequencing of the single-stranded polynucleotide fragments comprising non-complementary sequence at the 3' and 5' ends representing either or both of the first and second strand of the double stranded polynucleotide following treatment with converting agent (i.e. bisulfite treatment) generated by the methods described herein against a reference sequence. The reference sequence can be the sequence of the genomic DNA (either or both strands) not subjected to alteration by treatment with the converting agent. The comparing can be performed on a computer. The comparing can be done

IDT Exhibit 1020.3752

on a computer using a sequence alignment tool or program. The sequence alignment tool or program can map bisulfite treated sequencing reads to a genome of interest and perform methylation calls. The bisulfite sequencing mapping tool can be the Bismark program. In some cases, the comparing comprises performing a nucleotide alignment between the sequence obtained from the sequencing of the single-stranded DNA fragments comprising non-complementary sequence at the 3' and 5' ends generated by the methods described herein with a reference sequence on a computer using any of the nucleotide alignment programs known in the art (e.g. Bismark). In some cases, the methods described herein can be used to determine the methylation status of a specific locus or region of genomic DNA or the entire genome (i.e. the methylome). In some cases, following bisulfite treatment, the methylation status of a given cytosine residue is inferred by comparing the sequence to an unmodified reference sequence.

[00143] Sequencing can be any method of sequencing, including any of the next generation sequencing (NGS) methods described herein. In some cases, the NGS method comprises sequencing by synthesis. In some embodiments, sequencing is performed with primers directed against known or universal sequence introduced into the nucleic acid fragments by the adapter ligated to the nucleic acid fragments. In some cases, the primers used for sequencing are directed against adapter sequence unaltered by treatment with a converting agent. In some cases, primers used for sequencing are directed against adapter sequence altered by treatment with a converting agent. The converting agent can be bisulfite, wherein bisulfite treatment converts cytosine residues to uracil residues. In some cases, the sequencing primers are directed against adapter sequence comprising thymine residues following bisulfite treatment and amplification. In some cases, the sequencing primers are directed against adapter sequence wherein the adapter sequence is resistant to conversion by bisulfite treatment. In this embodiment, the adapter sequence to which the sequencing primers are directed does not comprise thymine residues following bisulfite treatment and amplification. In some cases, sequencing is performed with primers directed against identifier sequence introduced into the polynucleotide fragments by the first and/or second primer used to amplify single-stranded polynucleotide fragments comprising non-complementary sequence at the 3' and 5' ends. The identifier sequence can be a barcode sequence, a flow cell sequence, and/or index sequence. In some cases, the index sequence is a Truseq primer sequence compatible with the next generation sequencing platform produced by Illumina.

[00144] As described herein, the methods, compositions and kits provided herein can be used to generate reduced representation bisulfite (RRBS) libraries. As shown in **FIG. 23**, generation of an RRBS library can result in representation of about 1% of the genome that is enriched for 40-

IDT Exhibit 1020.3753

300 bp fragments comprising microsatellite bands and aid *in silico* assignment of CpG sites into genomic classes using far less sequencing than can be required using whole genome bisulfite sequencing approach as shown in **FIG. 25**. A comparison of RRBS library generation with other DNA methylation profiling technologies (**FIG. 24**), as reproduced from Gu et al. *Nature Protocols*. 6(4): 468-481(2011), which is herein incorporated by reference in its entirety, reveals that RRBS provides about 9% actual genomic coverage at a resolution of 1 bp with an input of only 0.01-0.3 μg. However, as described herein, RRBS can sometimes create a lack of color balance that can be an issue for some sequencing platforms. The methods, compositions, and kits provided herein can be used to eliminate or substantially reduce these issues.

[00145] A schematic exemplary of an embodiment of the methods described herein for generating an RRBS library using partial duplex-forming adapters comprising diversity bases is shown in **FIG. 3**. The method can comprise digesting a double stranded polynucleotide (e.g., genomic DNA) at one or more recognition sites for a methylation insensitive restriction endonuclease (RE). In some cases, the one or more recognition sites comprise the sequence "CCGG", and the methylation insensitive RE is MspI. The methylation insensitive restriction enzyme can be DpnII, HpaII, Sall-HF®, ScrFI, and the recognition site can be any recognition site recognized by said restriction endonucleases. Cleavage with the RE produces one or more polynucleotide fragments with or without overhangs. As shown in **FIG. 3**, cleavage with MspI generates a double-stranded polynucleotide fragment comprising a 5' overhang dinucleotide comprising 5'-CG-3'. A pool of diversity adaptors as shown in **FIG. 2A-B** can be appended to the double stranded polynucleotide generated by cleavage with MspI. The appending can be through ligation. Ligation can be performed with any ligation method known in the art. The ligation can be cohesive end ligation such as shown in **FIG. 3**. For illustrative purposes, ligation of a single diversity adaptor to both ends of the MspI cleaved double stranded polynucleotide is shown. The cleaved polynucleotide can be ligated on each end by a diversity adaptor comprising a different number of diversity bases than on an opposite end of the same cleaved polynucleotide. The diversity adapter in **FIG. 3** comprises a short strand whose 5' end comprises a terminal "CG" dinucleotide which immediately follows three 'H' nucleotides, which can be either a adenine (A), cytosine (C), or thymine (T) base, but cannot be a guanine (G) base. The long strand of the diversity adaptor terminates with a "DDD" trinucleotide comprising all "D" bases, wherein the "DDD" is the complement of the HHH triplet on the short strand and, therefore can be an A, G, or T base, but cannot be a C base. The long strand of the diversity adapter comprises non-methylated cytosine residues. In some cases, the long strand of the diversity adaptor can comprise cytosine analogs resistant to bisulfite treatment. In some cases,

IDT Exhibit 1020.3754

the cytosine analog resistant to bisulfite treatment is 5-methyl-cytosine. As can be seen in **FIG. 2A-B**, the short strand can comprise 0, 1, 2, or 3 "H" bases, and thus, the long strand can comprise either 0, 1, 2, or 3 "D" bases. **FIG 2A-B** also illustrates that the short strand comprises both a 5' (e.g., Biosg), and 3' (e.g., ddC) group. One or both groups can serve to block ligation and/or extension of the short strand. Following ligation of the diversity adapters in **FIG. 3**, the 3' ends of the polynucleotide fragments can be extended in the presence of a nucleotide mixture comprising dTTP, dGTP, dATP, and 5-methyl dCTP using the long strand of the diversity adapter as template. Extension can produce polynucleotide fragments comprising the long strand diversity adapter sequence comprising non-methylated cytosine residues at the 5' end and sequence complementary to the long strand diversity adapter sequence comprising methylated cytosine residues at the 3' end. In some cases, the long strand of the diversity adaptor comprises 5-methylcytosine and following ligation of the diversity adapters in **FIG. 3**, the 3' ends of the polynucleotide fragments can be extended in the presence of a nucleotide mixture comprising dTTP, dGTP, dATP, and dCTP using the long strand of the diversity adapter as template. Thus, extension can produce polynucleotide fragments comprising the long strand diversity adapter sequence comprising methylated cytosine residues at the 5' end and sequence complementary to the long strand diversity adapter sequence comprising non-methylated cytosine residues at the 3' end. In some cases, both the long strand and the short strand of the diversity adaptor comprises 5-methylcytosines and can ligate to each end of the MspI digested polynucleotide and following ligation of the diversity adapters , the 3' ends of the short strand of the D-adaptor can be extended in the presence of a nucleotide mixture comprising dTTP, dGTP, dATP, and dCTP using the long strand of the diversity adapter as template. Thus, extension can produce polynucleotide fragments comprising the long strand diversity adapter sequence comprising methylated cytosine residues at the 5' end and sequence complementary to the long strand diversity adapter sequence comprising non-methylated cytosine residues at the 3' end. In some cases, both the long strand and the short strand of the diversity adaptor comprises non-methylated cytosines and can ligate to each end of the MspI digested polynucleotide and following ligation of the diversity adapters, the 3' ends of the short strand of the diversity adaptor can be extended in the presence of a nucleotide mixture comprising dTTP, dGTP, dATP, and 5-methyl dCTP using the long strand of the diversity adapter as template. Thus, extension can produce polynucleotide fragments comprising the long strand diversity adapter sequence comprising non-methylated cytosine residues at the 5' end and sequence complementary to the long strand diversity adapter sequence comprising methylated cytosine residues at the 3' end. 5-methyl cytosine in the diversity adaptor and/or the reaction mixture used for extension can be

IDT Exhibit 1020.3755

substituted with any cytosine analog resistant to bisulfite treatment known in the art. The cytosine analog can be 5-hydroxymethylcytosine. The cytosine analog can be 5-propynlcytosine. The polynucleotide fragment resulting from digestion with MspI and ligation with one or more diversity adaptors to each end can be denatured and then subjected to bisulfite treatment as shown in **FIG. 3**. In some cases, fragments of a desired size are selected prior to bisulfite treatment and PCR ampilfication by running the fragments comprising the adapters in a gel and size selecting. Bisulfite treatment can generate single stranded polynucleotide fragments comprising non-complementary diversity adaptor sequence at each terminus. As shown in **FIG. 3**, bisulfite treatment can generates single-stranded polynucleotide fragments comprising long strand diversity adapter sequence at the 5' end wherein each cytosine residue has been converted to a uracil, and sequence which remains complementary to the long strand of the diversity adapter sequence present in the fragment prior to treatment at the 3' end. Any non-methylated cytosine residues present in the polynucleotide fragment between the adapter sequence can also be converted to uracil residues. Following bisulfite treatment, the fragments can be subjected to PCR using primers directed to the adapter sequences following conversion. PCR amplification can generate polynucleotide fragments with distinct adapter sequences on each end. The "D," "R," and "H" bases present in the diversity adapters can retain their original base identity following bisulfite treatment and subsequent amplification. The resultant bisulfite converted library can be further subjected to sequencing. Sequencing can be performed by any of the known next generation sequencing platforms. As can be seen in **FIG. 3**, sequencing the RRBS library generated by the method depicted in **FIG3** can generate predictable forward and reverse reads in the initial cycles. The long strand of the adapter can further comprise additional sequence elements such as barcodes and/or random sequence. Additional sequence elements can also be introduced to RRBS nucleic acid inserts through amplification of the inserts with primers comprising the additonal sequence elements. The additional sequence elements can be one or more barcodes, random sequence, and/or flow cell sequences that can be used to hybridize and thus capture an insert on a solid surface for subsequent solid-phase amplification. **FIG. 27A** illustrates a polynucleotide insert from an RRBS generated using diversity adapters as provided herein. **FIG. 37** illustrates a polynucleotide insert from an RRBS generated using diversity adapters as provided herein that is subsequently attached to a solid surface. The solid surface can be a flow cell surface or a bead. As shown in **FIG. 37**, a capture probe on a flow cell surface can be used to capture (via hybridization) a library insert such as the one shown in **FIG. 27A**. The "N" in **FIG. 37** or **FIG. 27A** can be a D base or a mixture of D and H bases as described herein. **FIG. 27B** illustrates the increased color balance in the initial sequence reads of an RRBS

IDT Exhibit 1020.3756

library generated using diversity adapters as provided herein and comprising polynucleotide inserts as shown in **FIG. 27A**. The increased sequence diversity (as shown in **FIG. 27B** by the increase in nucleotide base composition in the initial sequencing cycles) present in one or more polynucleotide inserts from an RRBS library generated using diversity adaptors can serve to substantially lower levels of genomic, higher diversity control libraries (e.g., PhiX control libraries) typically employed when sequencing RRBS libraries. The amount of higher diversity control libraries can be reduced by about, more than, or less than 1%, 5%, 10 %, 20%, 30%, 40%, 50%, 60%, 70%, 80% , 90%, 95%, or 99%. The amount of higher diversity control libraries can be reduced by about 1-5%, 5-10%, 10-20%, 30-40%, 40-50%, 50-60%, 60-70%, 70-80%, 80-90%, 90-95% or 95-99%. The amount of higher diversity control libraries can be reduced by between 1-5%, 5-10%, 10-20%, 30-40%, 40-50%, 50-60%, 60-70%, 70-80%, 80-90%, 90-95% or 95-99%. **FIG. 19** illustrates the total theoretical diversity possible for introducing to the ends of a polynucleotide upon introduction of a pool of diversity adaptors comprising diversity adapters comprising 0 to 3 "D" bases (D0-D3).

[00146] In some cases, no genomic, higher diversity control libraries (e.g., PhiX control libraries) are required to be added or spiked into sequencing reactions from the RRBS libraries generated by the methods provided herein.

[00147] In some cases, the methods depicted in **FIG. 3** for generating RRBS libraries can be combined wtih a dark sequencing approach as illustrated in **FIG. 26 (**see Boyle et al., *Genome Biology*. 13(R92):1-10 (2012), which is herein incorporated by reference in its entirety), which can serve to further reduce a lack of color balance. In some cases, the methods depicted in **FIG. 3** for generating RRBS libraries can be combined with spiking in a high diversity control library (e.g., Phi X) during sequencing, which can serve to further reduce a lack of color balance.

[00148] The method depicted in **FIG. 3** and described herein can also be used in the context of generating whole genome bisulfite sequencing libraries from samples spanning a 100-fold range of inputs, down to as little as 1ng of genomic DNA. Library performance metrics such as mapping, complexity, and 5-methylC content can be altered using the methods provided herein. In some cases, the methods provided herein improve Library performance metrics. The Library performance metrics can be mapping, complexity, and 5-methylC content.

VII. Downstream Applications

[00149] The methods and compositions disclosed herein can be efficiently and cost-effectively utilized for downstream analyses, such as next generation sequencing or hybridization platforms, with minimal loss of biological material of interest. The methods described herein can be useful

IDT Exhibit 1020.3757

for generating high throughput reduced representation sequencing libraries from bisulfite-converted DNA, for methylation analysis at specific regions across the genome or methylome.

[00150] For example, the methods described herein can be useful for sequencing by the method commercialized by Illumina, as described U.S. Pat. Nos. 5,750,341; 6,306,597; and 5,969,119. Directional (strand-specific) cDNA libraries are prepared using the methods described herein, and the selected single-stranded nucleic acid is amplified, for example, by PCR. The resulting nucleic acid can then be denatured and the single-stranded amplified polynucleotides can be randomly attached to the inside surface of flow-cell channels. Unlabeled nucleotides can be added to initiate solid-phase bridge amplification to produce dense clusters of double-stranded DNA. To initiate the first base sequencing cycle, four labeled reversible terminators, primers, and DNA polymerase can be added. After laser excitation, fluorescence from each cluster on the flow cell can be imaged. The identity of the first base for each cluster can then be recorded. Cycles of sequencing are performed to determine the fragment sequence one base at a time. The identity of the bases of the first few cycles (4-5) can then be used to discern separate clusters and facilitate sequencing of individual clusters. Additionally, clusters that share identical or substantially similar sequences at one or more ends as other clusters can experience problems with focusing of the laser and subsequent detection of fluorescence when sequencing such clusters in parallel due to the presence of the same base at each cycle. In some cases, the introduction of one or more diversity bases allows clear identification of individual clusters. For library inserts comprising low diversity portions at one or more ends of the inserts, the introduction of diversity bases at the low diversity ends can allow identification of individual clusters and allow different nucleotide bases to be incorporated in each cycle between inserts with identical or substantially similar sequences at their ends. The incorporation in each cycle of different bases between inserts with identical or substantially similar sequences at their ends can facilitate enhanced laser focusing during each sequencing cycle and thus accurate sequencing of individual clusters.

[00151] In some embodiments, the methods described herein are useful for preparing target polynucleotides for sequencing by the sequencing by ligation methods commercialized by Applied Biosystems (e.g., SOLiD sequencing). In other embodiments, the methods are useful for preparing target polynucleotides for sequencing by synthesis using the methods commercialized by 454/Roche Life Sciences, including but not limited to the methods and apparatus described in Margulies et al., *Nature*(2005) 437:376-380 (2005); and U.S. Pat. Nos. 7,244,559; 7,335,762; 7,211,390; 7,244,567; 7,264,929; and 7,323,305. In other embodiments, the methods are useful for preparing target polynucleotide(s) for sequencing by the methods commercialized by Helicos

IDT Exhibit 1020.3758

BioSciences Corporation (Cambridge, Mass.) as described in U.S. application Ser. No. 11/167,046, and U.S. Pat. Nos. 7,501,245; 7,491,498; 7,276,720; and in U.S. Patent Application Publication Nos. US20090061439; US20080087826; US20060286566; US20060024711; US20060024678; US20080213770; and US20080103058. In other embodiments, the methods are useful for preparing target polynucleotide(s) for sequencing by the methods commercialized by Pacific Biosciences as described in U.S. Pat. Nos. 7,462,452; 7,476,504; 7,405,281; 7,170,050; 7,462,468; 7,476,503; 7,315,019; 7,302,146; 7,313,308; and US Application Publication Nos. US20090029385; US20090068655; US20090024331; and US20080206764.

[00152] Another example of a sequencing technique that can be used in the methods described hereinis nanopore sequencing (see e.g. Soni G V and Meller A. (2007) *Clin Chem* 53: 1996-2001). A nanopore can be a small hole of the order of 1 nanometer in diameter. Immersion of a nanopore in a conducting fluid and application of a potential across it can result in a slight electrical current due to conduction of ions through the nanopore. The amount of current that flows can be sensitive to the size of the nanopore. As a DNA molecule passes through a nanopore, each nucleotide on the DNA molecule obstructs the nanopore to a different degree. Thus, the change in the current passing through the nanopore as the DNA molecule passes through the nanopore can represent a reading of the DNA sequence.

[00153] Another example of a sequencing technique that can be used in the methods described herein is semiconductor sequencing provided by Ion Torrent (e.g., using the Ion Personal Genome Machine (PGM)). Ion Torrent technology can use a semiconductor chip with multiple layers, e.g., a layer with micro-machined wells, an ion-sensitive layer, and an ion sensor layer. Nucleic acids can be introduced into the wells, e.g., a clonal population of single nucleic can be attached to a single bead, and the bead can be introduced into a well. To initiate sequencing of the nucleic acids on the beads, one type of deoxyribonucleotide (e.g., dATP, dCTP, dGTP, or dTTP) can be introduced into the wells. When one or more nucleotides are incorporated by DNA polymerase, protons (hydrogen ions) are released in the well, which can be detected by the ion sensor. The semiconductor chip can then be washed and the process can be repeated with a different deoxyribonucleotide. A plurality of nucleic acids can be sequenced in the wells of a semiconductor chip. The semiconductor chip can comprise chemical-sensitive field effect transistor (chemFET) arrays to sequence DNA (for example, as described in U.S. Patent Application Publication No. 20090026082). Incorporation of one or more triphosphates into a new nucleic acid strand at the 3′ end of the sequencing primer can be detected by a change in current by a chemFET. An array can have multiple chemFET sensors.

IDT Exhibit 1020.3759

[00154] Another example of a sequencing technique that can be used in the methods described herein is DNA nanoball sequencing (as performed, e.g., by Complete Genomics; see e.g., Drmanac et al. (2010) Science 327: 78-81). DNA can be isolated, fragmented, and size selected. For example, DNA can be fragmented (e.g., by sonication) to a mean length of about 500 bp. Adapters (Adl) can be attached to the ends of the fragments. For example, DNA can be fragmented with MspI and size selected to a mean length of about 500 bp. Adapters (Adl) can be attached to the ends of the fragments. The adapters can be used to hybridize to anchors for sequencing reactions. DNA with adapters bound to each end can be PCR amplified. The adapter sequences can be modified so that complementary single strand ends bind to each other forming circular DNA. The adapter sequences can be further modified so that they comprise diversity bases. In some cases, the adaptors comprise a pool of adaptors such that some of the adaptors in the pool comprise no diversity bases, some adaptors comprise 1 diversity base, some adaptors comprise 2 diversity bases, and some comprise 3 diversity bases. The DNA can be methylated to protect it from cleavage by a type IIS restriction enzyme used in a subsequent step. An adapter (e.g., the right adapter) can have a restriction recognition site, and the restriction recognition site can remain non-methylated. The non-methylated restriction recognition site in the adapter can be recognized by a restriction enzyme (e.g., AcuI), and the DNA can be cleaved by AcuI 13 bp to the right of the right adapter to form linear double stranded DNA. A second round of right and left adapters (Ad2) can be ligated onto either end of the linear DNA, and all DNA with both adapters bound can be PCR amplified (e.g., by PCR). Ad2 sequences can be modified to allow them to bind each other and form circular DNA. The DNA can be methylated, but a restriction enzyme recognition site can remain non-methylated on the left Adl adapter. A restriction enzyme (e.g., AcuI) can be applied, and the DNA can be cleaved 13 bp to the left of the Adl to form a linear DNA fragment. A third round of right and left adapter (Ad3) can be ligated to the right and left flank of the linear DNA, and the resulting fragment can be PCR amplified. The adapters can be modified so that they can bind to each other and form circular DNA. A type III restriction enzyme (e.g., EcoP15) can be added; EcoP15 can cleave the DNA 26 bp to the left of Ad3 and 26 bp to the right of Ad2. This cleavage can remove a large segment of DNA and linearize the DNA once again. A fourth round of right and left adapters (Ad4) can be ligated to the DNA, the DNA can be amplified (e.g., by PCR), and modified so that they bind each other and form the completed circular DNA template. Rolling circle replication (e.g., using Phi 29 DNA polymerase) can be used to amplify small fragments of DNA. The four adapter sequences can contain palindromic sequences that can hybridize and a single strand can fold onto itself to form a DNA nanoball (DNB(TM)) which can be approximately 200-300 nanometers in diameter on

IDT Exhibit 1020.3760

average. A DNA nanoball can be attached (e.g., by adsorption) to a microarray (sequencing flowcell). The flow cell can be a silicon wafer coated with silicon dioxide, titanium and hexamehtyldisilazane (HMDS) and a photoresist material. Sequencing can be performed by unchained sequencing by ligating fluorescent probes to the DNA. The color of the fluorescence of an interrogated position can be visualized by a high resolution camera. The identity of nucleotide sequences between adapter sequences can be determined.

[00155] In some cases, the sequencing technique can comprise paired-end sequencing in which both the forward and reverse template strand can be sequenced. In some cases, the sequencing technique can comprise mate pair library sequencing. In mate pair library sequencing, DNA can be fragments, and 2-5 kb fragments can be end-repaired (e.g., with biotin labeled dNTPs). The DNA fragments can be circularized, and non-circularized DNA can be removed by digestion. Circular DNA can be fragmented and purified (e.g., using the biotin labels). Purified fragments can be end-repaired and ligated to sequencing adapters.

[00156] In some cases, a sequence read is about, more than about, less than about, or at least about 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412,

IDT Exhibit 1020.3761

413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 525, 550, 575, 600, 625, 650, 675, 700, 725, 750, 775, 800, 825, 850, 875, 900, 925, 950, 975, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, or 3000 bases. In some cases, a sequence read is about 10 to about 50 bases, about 10 to about 100 bases, about 10 to about 200 bases, about 10 to about 300 bases, about 10 to about 400 bases, about 10 to about 500 bases, about 10 to about 600 bases, about 10 to about 700 bases, about 10 to about 800 bases, about 10 to about 900 bases, about 10 to about 1000 bases, about 10 to about 1500 bases, about 10 to about 2000 bases, about 50 to about 100 bases, about 50 to about 150 bases, about 50 to about 200 bases, about 50 to about 500 bases, about 50 to about 1000 bases, about 100 to about 200 bases, about 100 to about 300 bases, about 100 to about 400 bases, about 100 to about 500 bases, about 100 to about 600 bases, about 100 to about 700 bases, about 100 to about 800 bases, about 100 to about 900 bases, or about 100 to about 1000 bases.

[00157] The number of sequence reads from a sample can be about, more than about, less than about, or at least about 100, 1000, 5,000, 10,000, 20,000, 30,000, 40,000, 50,000, 60,000, 70,000, 80,000, 90,000, 100,000, 200,000, 300,000, 400,000, 500,000, 600,000, 700,000, 800,000, 900,000,
1,000,000, 2,000,000, 3,000,000, 4,000,000, 5,000,000, 6,000,000, 7,000,000, 8,000,000, 9,000,000, or 10,000,000.

[00158] The depth of sequencing of a sample can be about, more than about, less than about, or at least about lx, 2x, 3x, 4x, 5x, 6x, 7x, 8x, 9x, 10x, 11x, 12x, 13x, 14x, 15x, 16x, 17x, 18x, 19x, 20x, 21x, 22x, 23x, 24x, 25x, 26x, 27x, 28x, 29x, 30x, 31x, 32x, 33x, 34x, 35x, 36x, 37x, 38x, 39x, 40x, 41x, 42x, 43x, 44x, 45x, 46x, 47x, 48x, 49x, 50x, 51x, 52x, 53x, 54x, 55x, 56x, 57x, 58x, 59x, 60x, 61x, 62x, 63x, 64x, 65x, 66x, 67x, 68x, 69x, 70x, 71x, 72x, 73x, 74x, 75x, 76x, 77x, 78x, 79x, 80x, 81x, 82x, 83x, 84x, 85x, 86x, 87x, 88x, 89x, 90x, 91x, 92x, 93x, 94x, 95x, 96x, 97x, 98x, 99x, l00x, 110x, 120x, 130x, 140x, 150x, 160x, 170x, 180x, 190x, 200x, 300x, 400x, 500x, 600x, 700x, 800x, 900x, l000x, 1500x, 2000x, 2500x, 3000x, 3500x, 4000x, 4500x, 5000x, 5500x, 6000x, 6500x, 7000x, 7500x, 8000x, 8500x, 9000x, 9500x, or 10,000x. The depth of sequencing of a sample can about 1x to about 5x, about 1x to about 10x, about 1x to about 20x, about 5x to about 10x, about 5x to about 20x, about 5x to about 30x, about 10x to about 20x, about 10x to about 25x, about 10x to about 30x, about 10x to about 40x, about 30x to

IDT Exhibit 1020.3762

about 100x, about 100x to about 200x, about 100x to about 500x, about 500x to about 1000x, about 1000x, to about 2000x, about 1000x to about 5000x, or about 5000x to about 10,000x. Depth of sequencing can be the number of times a sequence (e.g., a genome) is sequenced. In some cases, the Lander/Waterman equation is used for computing coverage. The general equation can be: C = LN/G, where C = coverage; G = haploid genome length; L = read length; and N = number of reads.

[00159] In some cases, the methods described herein can be useful for enhancing sequence diversity in a sample. For example, the sample can be a sequencing library comprising polynucleotides sharing the same sequencing terminal 1, 2, or 3 bases before sequencing inserts. The polynucleotides can be fragments digested by MspI. 0, 1, 2, 3 diversity bases can be added before the shared sequencing terminal bases. For example, the diversity bases can be added between a sequencing primer and a sequencing insert. In some cases, the sequencing terminal bases of the polynucleotides are the diversity bases, not the shared bases. Thus, sequence diversity can be enhanced in the library. The library can be used in next generation sequencing. For example, when used for Illumina sequencing, the library with enhanced diversity can allow better cluster identification, phasing and color matrix calculation by the sequencer.

[00160] In some cases, the methods described herein can be useful for *in silico* analysis of sequence reads. In one aspect, computer-implemented software can be used to remove certain bases (e.g., bases added from an adaptor) from a sequence read. In some cases, Trim Galore (www.bioinformatics.babraham.ac.uk/projects/trim_galore) can be used to trim an adaptor sequence from a sequence read. In another case, python scriptscan be used to trim (i.e., remove) diversity bases from a sequence read. The computer implemented software can follow rules for removing certain bases from sequencing reads such as the rules shown in **FIG. 18A-C**. As shown in **FIG. 18A**, the rules can be used on library inserts that comprise adaptors (e.g., diversity adaptors or adaptors comprising extender sequence) on each end. The adaptors on each end can be different lengths (i.e., different numbers of nucleotides). As shown in **FIG. 18A**, the computer implemented software can be designed to look from a YGG triplet of bases at the 5' end of the sequence and trim or remove 0-2 D bases and/or an R base before the YGG triplet. The Y in the YGG triplet can be a cytosine (C) or thymine base (T). Within the 6 nucleotides at the 3' end of the forward read, the computer implemented software can look for a TCG triplet and trim it as well as the 0-3 "H" bases after it. Outside of the D, R, and H bases, the computer implemented software can be further designed to remove any additional adaptor sequence from or a plurality of sequence reads. In some cases, the MspI fragment can be larger than the read length and thus the software can look for TC at the 3' end and remove it. If TC is not found, then

the software can look for a "T" and remove it. As shown in **FIG. 18B**, with regards to a reverse read, the computer-implemented software can look for a CGA triplet within the 6 nucleotides at the 5' end and trim the 0-3 D bases as well as the CGA triplet since this contains no genomic info other than the fact that it is an MspI site. Within the 6 nucleotides at the 3' end of the reverse reads, the computer implemented software can look for CCR and trim the 0-3 H bases after it. Again, the computer implemented software can be further designed to remove any additional adaptor sequence from or a plurality of sequence reads. With regards to the reverse read, if the software cannot find a CCR, the MspI fragment may be larger than the read length. In general, the number of D (or D and R base combination) removed from the 5' end of forward reads can be the same as the number of H bases removed from 3' end of the reverse read (if the read spans across the entire MspI fragment). Additionally, the number of H bases removed from the 3' end of a forward read (if the read spans across the entire MspI fragment) can be the same as the number of D bases (or D and R base combination) removed from the 3' end of the reverse read. In some cases, following trimming, a library insert can align with respect to MspI sites in a genome from which the insert was derived as shown in **FIG. 18C**.

[00161] In another aspect, the methods provided herein can be useful for removing duplicated sequence reads. In these cases, a stretch of nucleotides whose base compositions is selected at random is inserted within one or both ends of a nucleic acid library insert generated using the method provided herein. This stretch of nucleotides whose base compositions is selected at random can be inserted adjacent to another sequence element (e.g., a barcode) and can be located proximally to a sequence that is capable of binding a sequencing primer. An example of a library insert generated using the methods provided herein for generating an RRBS library that comprises this stretch of nucleotides whose base compositions is selected at random is shown in **FIG. 32** and **FIG. 37**. This stretch of nucleotides whose base compositions is selected at random can be present in an adaptor or pool of adaptors comprising no or varying numbers of diversity bases (or extender sequence) and can be introduced to the end or ends of a polynucleotide upon introduction of the adaptor or pool of adaptors to the end or ends of the polynucleotide using the methods provided herein. This stretch of nucleotides whose base compositions is selected at random can be present in a primer or pool of primers comprising no or varying numbers of diversity bases (or extender sequence) and can be introduced to the end or ends of a polynucleotide upon introduction of the primer or pool of primers to the end or ends of the polynucleotide using the methods provided herein. This stretch of nucleotides whose base compositions is selected at random can be considered an additional sequence element as described herein. In some cases, this stretch of nucleotides whose base compositions is selected

IDT Exhibit 1020.3764

at random is introduced to the end or ends of a nucleic acid library insert during amplification of the nucleic acid library inserts using a primer or set of primers comprising the stretch of nucleotides whose base compositions is selected at random. The length of the stretch of nucleotides whose base compositions is selected at random can comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 9, 10, 11, or 12 bases. This random sequence can be common to each adaptor in a pool of adaptors or can be different between each adaptor within a pool of adaptors. This random sequence can be different between pools of adaptors. This random sequence can be common to each primer in a pool of primers or can be different between each primer within a pool of primers. This random sequence can be different between pools of primers. The presence of this random sequence can be utilized by computer-implemented software during the analysis of sequence reads. The computer-implemented software can be used to analyze the random sequence associated with all sequence reads that map to the same genomic location. Any reads that both map to the same genomic location and contain the same random sequence are considered duplicates, and all but one copy are removed.

[00162] In some cases, different barcodes can be added to polynucleotides in different samples (e.g., by using primers and/or adapters), and the different samples can be pooled and analyzed in a multiplexed assay. The barcode can allow the determination of the sample from which a polynucleotide originated.

[00163] The compositions, kits, and methods provided herein can be used to treat, prevent, diagnose, and/or prognose a variety of methylation related diseases. Such methylation related diseases can be cancer, mental retardation, neurodegenerative disorders, imprinting disorders, and syndromes involving chromosomal abnormalities. Such methylation related diseases can be Immunodeficiency-centromeric instability-facial anomalies syndrome (ICF), Rett syndrome, Beckwith-Wiedemann Syndrome (BWS), ATRX-linked mental retardation, fragile X syndrome. The cancer can be breast, ovarian, lung, head and neck, testicular, colon, or brain cancer. The cancer can be medulloblastoma, hepatoblastoma, uterine leiomyosarcomata, cervical carcinoma, renal cell carcinoma, rhadbomyosarcoma, gliomas, colorectal cancer, Wilm's tumour, Burkitt's lymphoma, or leukemia. In some cases, the methods described herein are used to determine the status of one or more genes associated with methylation related disorders. The status can include the presence or absence of a nucleic acid modification (i.e. methylation) at one or more bases in a nucleic acid sequence. In some cases, the methods disclosed herein are used to determine or recommend a course of treatment or administration of a therapy based on the status of one or more genes. The therapy can reduce one or more signs or symptoms of a methylation related disease. The therapy can prevent one or more signs or symptoms of any methylation related

IDT Exhibit 1020.3765

diseases. In some cases, the methods disclosed herein are used to determine the outcome or progress of a course of treatment or administration of a therapy based on the status of one or more genes. Genes associated with methylation related diseases can be, but are not limited to *Socs1, Cdkn1c, Slc22a1l, Bmp3b, Wit1, Rassf1a, Brca1, p16, Dapk, Mgmt, D4z4, Nbl2, H19, Igf2, G6pd, Rasgrf1 Sybl1, Ar, Pgk1, Dyz2, or Fmr1*. In some cases, the status of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20 of any of the genes associated with methylation related diseases are analyzed.

[00164] The methods, kits, and compositions described herein can be used to prevent the development of one or more signs and/or symptoms of methylation related diseases or reduce the severity of one or more signs and/or symptoms of methylation related diseases. The severity of the sign and/or symptom can be reduced by about, or more than about, or at least about, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or 100 percent. The severity of the sign or symptom can be decreased by about 1 percent to about 10 percent, about 1 percent to about 20 percent, about 1 percent to about 30 percent, about 1 percent to about 50 percent, about 1 percent to about 90 percent, about 1 percent to about 99 percent, about 10 percent to about 20 percent, about 10 percent to about 30 percent, about 10 percent to about 50 percent, about 50 percent to about 75 percent, about 75 percent to about 90 percent, about 75 percent to about 99 percent. The severity of the sign and/or symptom can be reduced by about, more than about, or at least about 2-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, 10-fold, 11-fold, 12-fold, 13-fold, 14-fold, 15-fold, 16-fold, 17-fold, 18-fold, 19-fold, 20-fold, 21-fold, 22-fold, 23-fold, 24-fold, 25-fold, 30-fold, 35-fold, 40-fold, 45-fold, 50-fold, 55-fold, 60-fold, 65-fold, 70-fold, 75-fold, 80-fold, 85-fold, 90-fold, 95-fold, 100-fold, 200-fold, 300-fold, 400-fold, 500-fold, 600-fold, 700-fold, 800-fold, 900-fold, or 1000-fold. The severity of the sign and/or symptom can be reduced by about 2-fold to 10-fold, about 2-fold to about 50-fold, about 2-fold to about 100-fold, about 10-fold to about 20-fold, about 10-fold to about 50-fold, about 10-fold to about 75-fold, about 10-fold to about 100-fold, about 50-fold to about 75-fold, about 50-fold to about 100-fold, about 100-fold to about 500-fold, about 100-fold to about 1000-fold, or about 500-fold to about 1000-fold.

[00165] The methods, kits, and compositions described herein can be used to decrease the likelihood that a subject will develop one or more signs and/or symptoms of methylation related diseases. The decrease in likelihood can be about, or more than about, or at least about 1, 2, 3, 4,

IDT Exhibit 1020.3766

5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or 100 percent. The decrease in likelihood can be about 1 percent to about 10 percent, about 1 percent to about 20 percent, about 1 percent to about 30 percent, about 1 percent to about 50 percent, about 1 percent to about 90 percent, about 1 percent to about 99 percent, about 10 percent to about 20 percent, about 10 percent to about 30 percent, about 10 percent to about 50 percent, about 50 percent to about 75 percent, about 75 percent to about 90 percent, about 75 percent to about 99 percent. The decrease in likelihood can be about, more than about, or at least about 1-fold, 2-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, 10-fold, 11-fold, 12-fold, 13-fold, 14-fold, 15-fold, 16-fold, 17-fold, 18-fold, 19-fold, 20-fold, 21-fold, 22-fold, 23-fold, 24-fold, 25-fold, 30-fold, 35-fold, 40-fold, 45-fold, 50-fold, 55-fold, 60-fold, 65-fold, 70-fold, 75-fold, 80-fold, 85-fold, 90-fold, 95-fold, 100-fold, 200-fold, 300-fold, 400-fold, 500-fold, 600-fold, 700-fold, 800-fold, 900-fold, or 1000-fold. The decrease in likelihood can be about 2-fold to 10-fold, about 2-fold to about 50-fold, about 2-fold to about 100-fold, about 10-fold to about 20-fold, about 10-fold to about 50-fold, about 10-fold to about 75-fold, about 10-fold to about 100-fold, about 50-fold to about 75-fold, about 50-fold to about 100-fold, about 100-fold to about 500-fold, about 100-fold to about 1000-fold, or about 500-fold to about 1000-fold.

[00166] A diagnosis and/or prognosis of a methylation associated neurological in a subject can be made by a health care provider, e.g., a developmental-behavioral pediatrician, a neurologist, a pediatric psychologist, or a psychiatrist. A diagnosis and/or prognosis of a neurological condition can be made or supported by a genetic test performed by a diagnostic laboratory. In some cases, a neurological assessment is administered to a subject by an individual trained and certified to administer a neurological assessment.

[00167] In some cases, a procedure can be performed to diagnose a methylation associated neurological condition in a subject, e.g., angiography, biopsy, a brain scan (e.g., computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET)), cerebrospinal fluid analysis (by, e.g., lumbar puncture or spinal tap), discography, intrathecal contrast-enhanced CT scan (cisternograhpy), electronencephalography (EEG), electromyography (EMG), nerve conduction velocity (NCV) test, electronystagmography (ENG), evoked potentials (evoked response; e.g., auditory evoked potentials, visual evoked potentials, somatosensory evoked potentials), myelography, polysomnogram, single photon emission computed tomography (SPECT), thermography, or ultrasound imaging (e.g., neurosonography, transcranial

IDT Exhibit 1020.3767

Doppler ultrasound). One or more procedures that can diagnose a neurological condition can be performed on a subject.

[00168] Instruments that can be used in neurological examination can include, e.g., a tuning fork, flashlight, reflex hammer, ophthalmoscope, X-ray, fluoroscope, or a needle.

[00169] The methods, kits, and compositions provided herein can be used to treat, prevent, diagnose, and/or prognose a methylation associate disease or condition in a subject. The subject can be a male or female. The subject can have, or be suspected of having, a methylation associated disease. The subject can have a relative (e.g., a brother, sister, monozygotic twin, dizygotic twin, father, mother, cousin, aunt, uncle, grandfather, grandmother) that was diagnosed with a methylation associate disease. The subject can be, for example, a newborn (birth to about 1 month old), an infant (about 1 to 12 months old), a child (about 1 year old to 12 years old), a teenager (about 13 years old to 19 years old), an adult (about 20 years old to about 64 years old), or an elderly person (about 65 years old and older). The subject can be, for example, about 1 day to about120 years old, about 1 day to about 110 years old, about 1 day to about 100 years old, about 1 day to about 90 years old, about 1 day to about 80 years old, about 1 day to about 70 years old, about 1 day to about 60 years old, about 1 day to about 50 years old, about 1 day to about 40 years old, about 1 day to about 30 years old, about 1 day to about 20 years old, about 1 day to about 15 years old, about 1 day to about 10 years old, about 1 day to about 9 years old, about 1 day to about 8 years old, about 1 day to about 7 years old, about 1 day to about 6 years old, about 1 day to about 5 years old, about 1 day to about 4 years old, about 1 day to about 3 years old, about 1 year to about 2 years old, about 3 years to about 15 years old, about 3 years to about 10 years old, about 3 years to about 7 years old, or about 3 years to about 5 years old. The subject can be about, more than about, at least about, or less than about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, or 120 years old.

[00170] The methods for generating polynucleotide libraries as described herein can be used for detecting the presence of fetal DNA in a maternal sample. In some cases, the method comprises: (a) generating bisulfite treated DNA libraries as described herein using a sample obtained from a pregnant woman comprising maternal and fetal DNA; (b) detecting the methylation status of DNA sequence of one or more genes from the sample comprising maternal and fetal DNA; and (c) comparing the methylation status the one or more genes from the sample comprising

maternal and fetal DNA to a reference maternal DNA sample comprising only maternal DNA. In some cases, step (b) of the method comprises an amplification process. In some cases, the amplification process is a polymerase chain reaction (PCR), such as real-time PCR. In other embodiments, step (b) determines the quantity of the DNA sequence. In some cases, the methods provided herein can be used to determine the Rhesus D (RhD) blood group compatibility between a pregnant woman and a fetus. In some cases, the methods for generating directional polynucleotide libraries as described herein can be used for diagnosing, monitoring, or risk assessment of a number of prenatal conditions. For example, the prenatal conditions can include, but are not limited to, beta-thalassemia, cystic fibrosis, congenital adrenal hyperplasia, chromosomal aneuploidies, preeclampsia, preterm labor, and intrauterine growth retardation (IUGR). In some cases, the method comprises (a) generating directional, bisulfite treated DNA libraries as described herein using a sample obtained from a pregnant woman comprising maternal and fetal DNA; (b) detecting the amount of DNA sequence of one or more genes from the sample comprising maternal and fetal DNA; and (c) comparing the amount of the DNA sequence with a standard control, wherein an increase from the control indicates the presence of or an increased risk for developing the pregnancy-associated condition. In some cases, step (b) of the method comprises an amplification process, which can be accomplished by various means, including polymerase chain reaction (PCR), such as real-time PCR. The one or more genes can be *RASSF1A, APC, CASP8, RARB, SCGB3A1, DAB2IP, PTPN6, THY1, TMEFF2, or PYCARD.* The sample can be whole blood, plasma, serum, urine, or saliva. The DNA can be cell-free DNA and/or DNA derived from maternal and fetal cells present in the sample from the pregnant woman. "Standard control value" as used herein refers to a predetermined amount of a genomic sequence that is originated from a fetus and is present in an established sample. The standard control value is suitable for the use of a method described herein, in order for comparing the amount of a gene of interest (or a non-coding sequence) that is present in a test sample. The standard control can provide an average amount of a fetal gene of interest that is typical for a defined time (e.g., first trimester) during pregnancy in the blood of an average, healthy pregnant woman carrying a normal fetus, both of whom are not at risk of developing any pregnancy-associated disorders or complications. A standard control value can vary depending on the genomic sequence of interest and the nature of the sample.

[00171] The methods for generating polynucleotide libraries as described herein can be combined with one or more methods for measuring DNA methylation at specific genomic loci. For example, the methods for measuring DNA methylation can include, but are not limited to,

IDT Exhibit 1020.3769

immunoprecipitation of methylated DNA, methyl-binding protein enrichment of methylated fragments, and/or digestion with methylation-sensitive restriction enzymes.

[00172] The methods for generating polynucleotide libraries as described herein can be combined with one or more methods for profiling methylation status of the whole genome, i.e. the methylome. For example, the methods provided herein can be combined with reduced representation bisuflite sequencing (RRBS). RRBS involves digestion of a DNA sample with a methylation-insensitive restriction endonuclease that has CpG dinucleotide as a part of its recognition site, followed by bisulfite sequencing of the selected fragments (Meissner *et al.*, *Nucleic Acids Res.* 33(18):5868-5877, 2005).

VIII. Compositions and Reaction mixtures

[00173] The present methods further provide one or more compositions or reaction mixtures. In some cases, the reaction mixture comprises: (a) a pool of duplex adaptors comprising duplex adaptors comprising 0, 1, 2, or 3 diversity bases in a ligation strand, wherein the ligation strand comprises cytosine analogs resistant to bisulfite treatment and a non-ligation strand wherein the non-ligation strand is blocked at the 3' and 5' ends and is enzymatically unreactive; (b) a strand displacing polymerase; (c) unmodified dNTPs; and (d) bisulfite. The duplex adaptors in the pool can be partial duplex adaptors comprising a long strand and a short strand as provided herein, wherein the long strand comprises the ligation strand, while the short strand comprises the non-ligation strand. In some cases, the strand displacing polymerase in the reaction mixture is substituted with a polymerase comprising exonuclease activity (e.g., Taq polymerase). In some cases, the reaction mixture further comprises (e) amplification primers directed to unique priming sites created at each end of the DNA fragments following bisulfite treatment. In some cases, at least one of the amplification primers is directed against adaptor sequence following bisulfite treatment, whereby cytosine residues have been converted to uracil residues. In some cases, the reaction mixture further comprises (f) sequencing primers directed against sequences present in the adaptor sequence. In some cases, at least one of the sequencing primers is directed against adaptor sequence following bisulfite treatment, whereby cytosine residues have been converted to uracil residues and subsequently replaced with thymine residues following amplification.

[00174] In some cases, the reaction mixture comprises: (a) a pool of duplex adaptors comprising duplex adaptors comprising 0, 1, 2, or 3 diversity bases in a ligation strand and a non-ligation strand wherein the non-ligation strand is blocked at the 3' and 5' ends and is enzymatically unreactive; (b) a strand displacing polymerase; (c) modified dCTP (i.e. 5-methyl-dCTP, 5-hydroxymethyl-dCTP, or 5-propynyl-dCTP); (d) dATP, dGTP, and dTTP; and (e) bisulfite. The

duplex adaptors in the pool can be partial duplex adaptors comprising a long strand and a short strand as provided herein, wherein the long strand comprises the ligation strand, while the short strand comprises the non-ligation strand. In some cases, the strand displacing polymerase in the reaction mixture is substituted with a polymerase comprising exonuclease activity (e.g., Taq polymerase). In some cases, the reaction mixture further comprises (f) amplification primers directed to unique priming sites created at each end of the DNA fragments following bisulfite treatment. In some cases, at least one of the amplification primers is directed against adaptor sequence following bisulfite treatment, whereby cytosine residues have been converted to uracil residues. In some cases, the reaction mixture further comprises (g) sequencing primers directed against sequences present in the adaptor sequence. In some cases, at least one of the sequencing primers is directed against adaptor sequence following bisulfite treatment, whereby cytosine residues have been converted to uracil residues and subsequently replaced with thymine residues following amplification.

IX. Kits

[00175] Any of the compositions described herein may be comprised in a kit. In a non-limiting example, the kit, in a suitable container, comprises: an adaptor or several adaptors, one or more of oligonucleotide primers and reagents for ligation, primer extension and amplification. In some cases, the adaptors are diversity adaptors. The diversity adaptors can be any adapter has generated by the method depicted in **FIG. 2B**. The diversity adaptors can be any adaptor that comprises a partial duplex as depicted in **FIG. 3**. The adapters can be the adapters depicted in **FIG. 3**. The kit may also comprise means for purification, such as a bead suspension, and nucleic acid modifying enzymes.

[00176] The containers of the kits can include at least one vial, test tube, flask, bottle, syringe or other containers, into which a component may be placed, and preferably, suitably aliquotted. Where there is more than one component in the kit, the kit also can contain a second, third or other additional container into which the additional components may be separately placed. Various combinations of components may be comprised in a container.

[00177] When the components of the kit are provided in one or more liquid solutions, the liquid solution can be an aqueous solution. The components of the kit may be provided as dried powder(s). When reagents and/or components are provided as a dry powder, the powder can be reconstituted by the addition of a suitable solvent.

[00178] The present methods provide kits containing one or more compositions described herein and other suitable reagents suitable for carrying out the methods described herein. The methods described herein provide, e.g., diagnostic kits for clinical or criminal laboratories, or nucleic acid

IDT Exhibit 1020.3771

amplification or analysis kits for general laboratory use. The present methods thus include kits which include some or all of the reagents to carry out the methods described herein, e.g., sample preparation reagents, oligonucleotides, binding molecules, stock solutions, nucleotides, polymerases, enzymes, positive and negative control oligonucleotides and target sequences, test tubes or plates, fragmentation reagents, detection reagents, purification matrices, and an instruction manual. In some cases, the kit comprises a binding molecule, wherein the binding molecule is a nucleotide analog binding protein. In some cases, the nucleotide analog binding protein comprises a methylcytosine binding protein. In some cases, the methylcyotsine binding protein comprises an anti-5-methylcytosine antibody. In some cases, the kit contains a modified nucleotide. Suitable modified nucleotides include any nucleotides provided herein including but not limited to a nucleotide analog. In some cases, the nucleotide analog can be a cytosine analog. In some cases, the cytosine analogs can be 5-methyl dCTP, 5-hydroxymethyl dCTP, and/or 5-propynl dCTP. In some cases, the kit comprises a converting agent. In some cases, the converting agent is bisulfite or its equivalent.

[00179] In some cases, the kit can contain one or more reaction mixture components, or one or more mixtures of reaction mixture components. In some cases, the reaction mixture components or mixtures thereof can be provided as concentrated stocks, such as 1.1x, 1.5 x, 2 x, 2.5 x, 3 x, 4 x, 5 x, 6 x, 7 x, 10 x, 15 x, 20 x, 25 x, 33 x, 50 x, 75 x, 100 x or higher concentrated stock. The reaction mixture components can include any of the compositions provided herein including but not limited to buffers, salts, divalent cations, azeotropes, chaotropes, dNTPs, labeled nucleotides, modified nucleotides, dyes, fluorophores, biotin, enzymes (such as endonucleases, exonucleases, glycosylases), or any combination thereof.

[00180] In some cases, the kit can contain one or more oligonucleotide primers, such as the oligonucleotide primers described herein. The oligonucleotide primers can be one or more primers comprising diversity bases or extender sequence as described herein. In some cases, the kit comprises a pool of primers, wherein some of the primers comprise no diversity bases, while the remaining primers in the pool of primers comprise a variable number of diversity bases. The primers can comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more diversity bases. In some cases, the pool of primers is used to generate RRBS libraries such that the base composition of the diversity base is chosen so as not to recreate the YGG (Y is cytosine or thymine) motif at the 5' end of a nucleic insert in the RRBS library. The kit can contain one or more oligonucleotide primers comprising sequence directed against the ligation strand of an adapter or its complement and/or sequence directed against the ligation strand of an adapter or its complement whose sequence is altered by treatment with a converting agent. In some cases, the

IDT Exhibit 1020.3772

converting agent is bisulfite. In some cases the kit can contain tailed primers comprising a 3'-portion hybridizable to the target nucleic acid and a 5'-portion which is not hybridizable to the target nucleic acid. In some cases, the kit can contain chimeric primers comprising an RNA portion and a DNA portion. In some cases, the 5' portion of the tailed primers comprises one or more barcodes or other identifier sequences. In some cases, the identifier sequences comprise flow cell sequences, TruSeq primer sequence, and/or second read barcode sequences. In some cases, the identifier sequence is a stretch of random sequences such that the random sequences are read during an index or barcode read and are used as mark to facilitate identification and removal of PCR duplicates during subsequent *in silico* analysis of sequencing reads.

[00181] In some cases, the kit can contain one or more polymerases or mixtures thereof. In some cases, the one or more polymerases or mixtures thereof can comprise strand displacement activity and/or exonuclease activity. Suitable polymerases include any of the polymerases provided herein. The kit can further contain one or more polymerase substrates such as for example dNTPs, non-canonical or modified nucleotides, or nucleotide analogs.

[00182] In some cases, the kit can contain one or more means for purification of the nucleic acid products, removing of the fragmented products from the desired products, or combination of the above. Suitable means for the purification of the nucleic acid products include but are not limited to single stranded specific exonucleases, affinity matrices, nucleic acid purification columns, spin columns, ultrafiltration or dialysis reagents, or electrophoresis reagents including but not limited acrylamide or agarose, or any combination thereof.

[00183] In some cases, the kit can contain one or more reagents for producing blunt ends. For example, the kit can contain one or more of single stranded DNA specific exonucleases including but not limited to exonuclease 1 or exonuclease 7; a single stranded DNA specific endonucleases such as mung bean exonuclease or S1 exonuclease, one or more polymerases such as for example T4 DNA polymerase or Klenow polymerase, or any mixture thereof. Alternatively, the kit can contain one or more single stranded DNA specific exonucleases, endonucleases and one or more polymerases, wherein the reagents are not provided as a mixture. Additionally, the reagents for producing blunt ends can comprise dNTPs.

[00184] In some cases, the kit can contain one or more reagents for preparing the double stranded products for ligation to adapter molecules. For example, the kit can contain dATP, dCTP, dGTP, dTTP, or any mixture thereof. In some cases, the kit can contain a polynucleotide kinase, such as for example T4 polynucleotide kinase. Additionally, the kit can contain a polymerase suitable for producing a 3' extension from the blunt ended double stranded DNA fragments. Suitable polymerases can be included, for example, exo-Klenow polymerase.

IDT Exhibit 1020.3773

[00185] In some cases, the kit can contain one or more adaptor molecules such as any of the adapter molecules provided herein. Suitable adapter molecules include single or double stranded nucleic acid (DNA or RNA) molecules or derivatives thereof, stem-loop nucleic acid molecules, forked nucleic acid molecules, double stranded molecules comprising one or more single stranded overhangs of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 bases or longer, proteins, peptides, aptamers, organic molecules, small organic molecules, or any adapter molecules known in the art that can be covalently or non-covalently attached, such as for example by ligation, to the double stranded DNA fragments. In some cases, the adapters can be duplex adapters wherein one strand comprises nucleotide analogs resistant to conversion by a converting agent, while the other strand comprises a 5' and 3' block. In some cases, the adapters can be duplex adapters wherein both strands comprise nucleotide analogs resistant to conversion by a converting agent. In some cases, the adapters can be duplex adapters wherein both strands comprise nucleotides sensitive to conversion by a converting agent. In a further embodiment, the duplex adapter is a partial duplex adapter. In some cases, the partial duplex adapter comprises a long strand comprising nucleotide analogs resistant to conversion by a converting agent, and a short strand comprising a 5' and 3' block. In some cases, the partial duplex adapter comprises a long strand comprising nucleotides sensitive to conversion by a converting agent, and a short strand comprising a 5' and 3' block. In some cases, the nucleotide analog is a cytosine analog. In some cases, the nucleotides sensitive to a converting agent is dCTP. In some cases, the cytosine analogs present in the adapter can be 5-methylcytosine, 5-hydroxymethylcytosine, and/or 5-propynlcytosine. In some cases, the 5' block comprises a biotin moiety. In some cases, the 3' block is blocked with a terminal dideoxycytosine. In some cases, one or more adaptors in the kit comprise one or more diversity bases or extender sequence as described herein.

[00186] In some cases, the kit can contain one or more reagents for performing nick or fill-in repair on the ligation complex formed between the adaptors and the double stranded products of the methods described herein. The kit can contain a polymerase suitable for performing nick repair. Suitable polymerases can be included, for example, *Taq* DNA polymerase.

[00187] The kit can further contain instructions for the use of the kit. For example, the kit can contain instructions for generating RRBS libraries, directional cDNA libraries or directional cDNA libraries representing the methylome or the methylation status of a specific genomic region or locus useful for large scale analysis of including but not limited to e.g., , pyrosequencing, sequencing by synthesis, sequencing by hybridization, single molecule sequencing, nanopore sequencing, and sequencing by ligation, high density PCR, digital PCR, massively parallel Q-PCR, and characterizing amplified nucleic acid products generated by the

IDT Exhibit 1020.3774

methods described herein, or any combination thereof. The kit can further contain instructions for mixing the one or more reaction mixture components to generate one or more reaction mixtures suitable for the methods described herein. The kit can further contain instructions for hybridizing the one or more oligonucleotide primers to a nucleic acid template. The kit can further contain instructions for extending the one or more oligonucleotide primers with for example a polymerase and/or nucleotide analogs. The kit can further contain instructions for treating the DNA products with a converting agent. In some cases, the converting agent is bisulfite. The kit can further contain instructions for purification of any of the products provided by any of the steps of the methods provided herein. The kit can further contain instructions for producing blunt ended fragments, for example by removing single stranded overhangs or filling in single stranded overhangs, with for example single stranded DNA specific exonucleases, polymerases, or any combination thereof. The kit can further contain instructions for phosphorylating the 5' ends of the double stranded DNA fragments produced by the methods described herein. The kit can further contain instructions for ligating one or more adapter molecules to the double stranded DNA fragments. The kit can further contain instructions for analysing sequence reads obtained from libraries generated using the methods provided herein and subsequently sequenced. The instructions can describe how to trim diversity bases present in sequence reads *in silico*. The instructions can describe how to remove duplicates present in sequence reads *in silico*.

[00188] A kit will can include instructions for employing, the kit components as well the use of any other reagent not included in the kit. Instructions can include variations that can be implemented.

[00189] Unless otherwise specified, terms and symbols of genetics, molecular biology, biochemistry and nucleic acid used herein follow those of standard treatises and texts in the field, e.g. Kornberg and Baker, *DNA Replication*, Second Edition (W.H. Freeman, New York, 1992); Lehninger, *Biochemistry*, Second Edition (Worth Publishers, New York, 1975); Strachan and Read, Human *Molecular Genetics*, Second Edition (Wiley-Liss, New York, 1999); Eckstein, editor, *Oligonucleotides and Analogs: A Practical Approach* (Oxford University Press, New York, 1991); Gait, editor, *Oligonucleotide Synthesis: A Practical Approach* (IRL Press, Oxford, 1984); and the like.

EXAMPLES

Example 1: Generation of a RRBS library using Diversity Adapters:

[00190] In this example, an RRBS library was generated using the diversity adapters shown in **FIG. 2A-B** and the sequencing results were subsequently analyzed. As shown in **FIG.**

IDT Exhibit 1020.3775

2A-B, the diversity adapters can be partial duplex adapters comprising a long strand (e.g., R01446 in **FIG. 2A**), hybridized with a short strand (e.g., Ben383, 384, 385, or 381 in **FIG. 2A**) in such a manner that both ends of the partial duplex comprise a 5' overhang. The 5' overhang comprising sequence from the short strand of the partial duplex comprises a 5' terminal GC sequence with a Biosg label on the terminal cytosine residue, while the 3' terminus comprises a terminal 3' dideoxy cytosine residue. As can be seen in **FIG. 2A-B**, the 5' overhang comprising short strand sequence contains either no (D0, e.g., Ben341 in **FIG. 2A**), 1 (D1, e.g., Ben385 in **FIG. 2A**), 2 (D2, e.g., Ben384 in **FIG. 2A**), or 3 ((D3, e.g., Ben383 in **FIG. 2A**) "H" bases. The "H" bases can be either an adenosine, cytosine or thymine residue, but cannot be a guanine residue. As shown in **FIG. 2A-B**, for use in the methods provided herein as well as this Example, the partial duplex adapter is subjected to an extension reaction using a dNTP mix containing only dATP, dGTP, and dTTP (i.e., no dCTP), whereby a partial duplex adapter comprising a 5' overhang comprising at least one "H" (e.g., R01446 hybridized to Ben 383, 384, or 385) is extended to produce a partial duplex adapter containing a two base 5' overhang with the sequence GC at one end. In this Example, 5 ng of human gDNA was digested with MspI and the resulting fragments were separated into different diversity adaptor pools as shown in FIG. 2A-B, wherein a first pool was ligated to the D0 diversity adaptors, a second pool was ligated to the D1 diversity adapters, a third pool was ligated the D2 diversity adapters, a fourth pool was ligated to the D3 adapters, and a fifth pool was ligated to a mixture of the D0, D1, D2, and D3 diversity adapters. The ligations and all subsequent steps were performed according to the protocol for the NuGEN Ovation ® Ultralow Methyl-Seq Library System, which is herein incorporated by reference in its entirety. Briefly, following digestion of the gDNA with MspI to produce DNA fragments, the DNA fragments are combined with a master mix containing the appropriate diversity adapters comprising 0, 1, 2, or 3 diveristy "D" bases (i.e., D3 ,D2, D1, D0, or a mix of D0-D3 in **FIGs. 4A-B, 5A-B, 6A-B, 7A-B, and 9A-B, respectively**), ligation buffer, water and ligation enzyme mix, placed in a thermal cycler and run through a program on the thermal cycler that comprised of a single cycle comprising incubations at 25 C for 30 minutes, 70 C for 10 minutes, and then holding the mixture at 4 C. Following ligation, the ligation complexes were subjected to a post-ligation purification followed by a final repair reaction entailing incubation of the ligation complexes with a DNA repair enzyme in a thermal cycler programmed for a single cycle of 60 C incubation for 10 minutes. Following final repair, the samples were subjected to bisulfite conversion, wherein 20 uL of the products from the final repair reactions were inputted directly into a Qiagen EpiTect Fast DNA Bisulfite Kit according to manufacturer protocols and subjected to two cycles of 95 C for 5 minutes, and 60 C for 20

IDT Exhibit 1020.3776

minutes on a thermal cycler followed by elution of the purified, bisulfite converted DNA libraries.. The bisulfite converted DNA libraries were then PCR amplified and purified. Following library amplification and purification, the libraries were pooled and sequenced on an Illumina MiSeq sequencer according to manufacturer instructions with a 40 nucleotide forward read, 6 nucleotide index read, and 25 nucleotide reverse read. **FIGs. 4A-B, 5A-B, 6A-B, 7A-B, 8, and 9A-B** show a high correlation between the expected forward and reverse sequencing reads and the actual nucleotide distributions at read positions 1-40 in the forward reads (FIGs. 4A, 5A, 6A, 7A, and 9A) and 1-25 in the reverse reads (FIGs. 4B, 5B, 6B, 7B, and 9B)for nucleic acid fragments appended with diversity adapters with three diversity "D" bases (**FIG. 4A-B**), diversity adapters with 2 diversity "D" bases (**FIG. 5A-B**), diversity adapters with 1 diversity "D" base (**FIG. 6A-B**), diversity adapters with no diversity bases (**FIG. 7A-B**), and the mix of diversity adapters comprising zero, 1, 2, or 3 diversity "D" bases (**FIGs. 8 & 9A-B**). In some cases, the diversity base in the 5' position of a diversity adaptor comprising three diversity bases can be an "R" base while the remaining 2 diversity bases can be "D" bases as shown in **FIG. 35**. The "R" base can be either an adenine or a guanine nucleotide base, while the "D" bases can be either an adenine, guanine, or thymine nucleotide base as shown in **FIG. 35** and as described herein. In some cases, an "R" base is substituted for a "D" base in a pool of diversity adaptors such that library inserts generated using the methods described herein for generating an RRBS library following MspI digestion using the pool of diversity adaptors with an "R" base comprise an "R" base as the first base read in a forward sequence read (e.g., **FIG. 35**).

Example 2: Sequencing of RRBS library generated using Diversity Adapters without Phi X control:

[00191] In this example, RRBS libraries were generated using diversity adapters as described in Example 1; however, sequencing of the RRBS libraries generated using diversity adapters with a 40 nucleotide forward read, 6 nucleotide index read, and 25 nucleotide reverse read was conducted without running a dedicated control high diversity Phi X control lane. As can be seen in **FIGs. 10-12**, the RRBS library sequenced without a PhiX control showed a Q-score distribution whereby > 95% of the Q-scores were > 30 (**FIG. 10**), an anticipated distribution of bases (**FIGs. 10** and **11**) whereby a diversity of bases existed at the start of the sequencing read while roughly equal numbers of "A" and "G" bases were reported further into the read, and the percentage of > Q20 scores being above 95% at PCR amplification cycles of less than 40 (**FIG. 12**) , which indicated a low probability of an incorrect base calling. **FIG. 13** shows a Q-score heatmap over 71 sequencing cycles that include 40 forward read cycles, 6 index read cycles, and 25 reverse read sequencing cycles. **FIG. 16** shows a summary of the run for this

IDT Exhibit 1020.3777

Example showing a high percentage (>95%) of total reads with a Q score above 30, with a forward read (read 1) having a Q 30 score of 95%, an index read (read 2) having a Q 30 score of >90%, anda reverse read (read 3) having a Q30 score of >90%. In summary, the use of the diversity adapters in generating RRBS libraries with diversity adaptors comprising D bases as described in Example 1 produced high quality sequencing reads with or without the use of PhiX control libraries.

[00192] As a comparison, inserts from a Human Blood RNA sequencing library generated for ~ 100 ng of total RNA from adult human whole blood according to the manufacturer's instructions for the NuGEN Ovation® Human Blood RNA-Seq Library System, which is herein incorporated by reference in its entirety, were sequenced with a 40 nucleotide forward read, 6 nucleotide index read, and 25 nucleotide reverse read. Briefly, ~100 ng of total RNA from adult human whole blood was subjected to first strand synthesis using random primers, followed by second strand synthesis in the presence of a non-canonical nucleotide (i.e., dUTP), fragmented using the Covaris sonication, end repaired, ligated with adapters marked with a non-canonical nucleotide (i.e., dUTP) and comprising barcode sequences, strand selected, adapter cleavage to remove unwanted sequences (e.g., rRNA and Globin RNA), and finally, library amplification using PCR with primers directed against the adapter sequences, thereby allowing for 16-plex sequencing capability. **FIGs. 14** and **15** show intensity by cycle, cluster density, Q-score distribution and heat-map data from the 40 cycle forward read by 6 cycle index read by 25 cycle reverse read run performed on the Blood-Seq library that was similar to the data obtained from the 40 cycle forward read by 6 cycle index read by 25 cycle reverse read run performed on the RRBS library generated without Phi X controls (see **FIGs. 10 and 13**). **FIG. 17** shows the run summary for the 40 cycle forward read by 6 cycle index read by 25 cycle reverse read run performed on the Blood-Seq library. A comparison between **FIGs 16** and **17** showed that sequencing of the RRBS library generated with diversity adapters as described in Example 1 produced high quality sequencing data without requiring a Phi X high diversity control.

Example 3: Comparison of the CpGs analyzed by sequencing a RRBS library generated using Diversity Adapters vs. Whole Genome Bisulfite Sequencing (WGBS):

[00193] In this example, a comparison of the CpGs analyzed by sequencing a RRBS library generated by the methods provided herein vs. WGBS was conducted. The RRBS library was generated from MspI fragmented genomic DNA (gDNA) from IMR90 human cell line was ligated to a mixture of DO-D3 diversity adapters as described in Example 1 using either 400, 100, 25, 6.25, or 1.56 ng of gDNA as input for generation of the RRBS library. The library used for WGBS was generated by first subjecting 50 ng of human gDNA to Covaris sonication to an

average size of 200 bp, and then generating a library for WGBS using the fragmented gDNA as input into the NuGEN Ovation ® Ultralow Methyl-Seq Library System, wherein the library for WGBS was generated as directed by the manufacturer. The RRBS and WGBS libraries were each separately sequenced on an Illumina GAIIx sequencer with a 40 nucleotide single end (40nt SE) and 6 nucleotide index read. For the RRBS sequences, the D1-D3 adaptor sequences were trimmed off, and the resulting reads, which started with "YGG", were analyzed with the Bismark program for mapping bisulfite treated sequencing reads. Mapped reads were further analyzed by grouping into categories of CpG islands, CpG shores, and promoters. **FIG. 20** shows that the total CpGs analyzed using as little as 1.56 ng of starting gDNA was considerably greater than the total CpGs analyzed by WGBS. Moreover, the average CpG per read and average CpG per uniquely aligned read were at least 6 X and 7X greater than WGBS, respectively (**FIG. 21**). **FIG. 22** shows that the feature coverage (i.e., CpG Islands and Shores, as well as Promoters) using RRBS libraries generated using the methods provided herein are very close to the theoretical maximum predicted by an *in silico* analysis for a 40 nt SE read for a RRBS library (**FIGs. 22 and 25**). Overall, RRBS libraries generated using diversity adapters as provided herein can be easily sequenced (e.g., with commercially available NGS sequencers such Illumina systems), and provide CpG-dense data.

Example 4: Kit Components

[00194] A kit can be used for RRBS library formation.

[00195] A. Reagents

[00196] The kit can comprise the components in Table 1:

Table 1. Reagents

MspI Buffer Mix
MspI Enzyme Mix
Ligation Buffer Mix
Ligation Adaptor Mixes
Ligation Enzyme Mix
Final Repair Buffer Mix
Final Repair Enzyme Mix
Amplification Primer Mix
Amplification Enzyme Mix
25 μM Sequencing Primer

IDT Exhibit 1020.3779

Nuclease-free Water
Agencourt RNAClean XP Beads
DNA Resuspension Buffer

[00197] B. Additional Equipment, Reagents and Labware

[00198] Additional materials that can be used include the following Equipment, Reagents, and Labware.

[00199] Equipment includes: Agilent 2100 Bioanalyzer, High Sensitivity DNA Kit or materials andequipment for electrophoretic analysis of nucleic acids; Qubit® 2.0 Fluorometer and dsDNA HS Assay Kit (Life Technologies); microcentrifuge for individual 1.5 mL and 0.5 mL tubes; 0.5–10 µL pipette, 2–20 µL pipette, 20–200 µL pipette, 200–1000 µL pipette; vortexer**;** thermal cycler with 0.2 mL tube heat block, heated lid, and 100 µL reaction capacity.

[00200] Reagents include: EpiTect Fast DNA Bisulfite Kit (QIAGEN Cat. #59824 for 50 preps or Cat. #59826 for 200 preps); Ethanol (Sigma-Aldrich, Cat. #E7023), for purification steps; OPTIONAL: EvaGreen® Dye, 20X in water (Biotium, Cat. #31000).

[00201] Supplies and Labware include: Nuclease-free pipette tips; 1.5 mL and 0.5 mL RNase-free microcentrifuge tubes; 0.2 mL individual thin-wall PCR tubes or 8 X 0.2 mL strip PCR tubes or 0.2 mL thin-wall PCR plates; magnetic separation device for the Bead Purification of the Amplified Material protocol; disposable gloves; kimwipes; ice bucket; cleaning solutions such as DNA-OFF™ (MP Biomedicals, Cat. #QD0500); OPTIONAL: PhiX Control (Illumina, Cat. Cat. #FC-110-3001); OPTIONAL: Real-time PCR system.

Example 5: Planning an experiment

[00202] A. Input DNA

[00203] Methods, compositions, and kits provided herein can work with inputs of 100 ng of intact human genomic DNA. Quantitation of gDNA by a dsDNA assay, such as PicoGreen® or the Qubit System, can be performed. In some cases, OD260 readings for quantitation of input material are not relied upon. DNA samples can be free of contaminating proteins, RNA, organic solvents (including phenol and ethanol) and salts. A commercially available system for gDNA isolation can be used. The A260:A280 ratio for DNA samples can be in excess of 1.8. Using DNA samples with lower ratios can compromise results. RRBS libraries can be generated from less than 100 ng of gDNA, or from degraded gDNA, such as DNA extracted from formalin fixed, paraffin embedded (FFPE) specimens. If such samples are used, library amplification can

require additional cycles of PCR. One additional cycle of PCR can be used for each two fold decrease in starting material. To accurately determine the number of PCR cycles to perform, the optional real-time PCR protocol in Example 7. An increase in adaptor artifacts can be observed, as well as decreased library complexity, in libraries made from less than 100 ng or degraded samples, and additional cycles of PCR amplification can be performed. Adaptor artifacts can appear as a peak of approximately 145 bp. These artifacts can be reduced by performing a second bead purification. One volume of beads can be added to the purified library.

[00204] Signs of reduced complexity can be monitored by sequencing unique 6 random bases appended to the index on each adaptor molecule, then a Duplicate Marking tool can be used to identify true PCR duplicates. The Duplicate Marking tool utilizes information provided by the unique 6 random bases sequence to discriminate between true PCR duplicates and independent adaptor ligation events to fragments with the same start site resulting in the recovery of more usable data. The Duplicate Marking tool and instructions for use can be obtained by contacting NuGEN Technical Support (techserv@nugen.com). After removing true PCR duplicates, a reliable measure of library complexity can be made. Sequencing the additional 6 bases after the 6-base barcode can be useful when attempting RRBS on rare or degraded samples. See Example 7 for information on index structure and sequencing.

[00205] B. Using compositions and kits described herein on Illumina NGS Platforms

[00206] Methods described herein can use the same approach to multiplexing found in the standard Illumina method. These libraries can be sequenced using the Illumina protocol for multiplex sequencing. The dedicated read (DR) barcode sequences, found in Table 2, can be entered into the Illumina software prior to analysis.

[00207] Barcode sequences and multiplex guidelines for adaptors used herein can be found in Table 2. These 6-nucleotide barcode sequences can be input into the Illumina Sequencing System prior to parsing of the data. You may combine anywhere from 2–16 barcoded libraries to allow for a range of multiplex sequencing. The barcodes can be chosen for their ability to parse properly and for color balancing and can have strict pairing requirements when performing 2-plex multiplexing. Users wishing to perform greater than a 2-plex multiplexing can choose a Duplex Set (as defined in Table 9), combined with any of the remaining barcoded libraries. The barcode sequences can be separated by an edit distance of three. For further details on the barcode design strategy, please refer to Faircloth BC and Glenn TC (2012). *PLoS ONE* 7(8): e42543. doi:10.1371/journal.pone.0042543, which is herein incorporated by reference in its entirety. NuGEN RRBS libraries can be mixed with other libraries if the barcodes are compatible (i.e. can be parsed). One mismatch allowed can be parsed with if the spike in library barcodes are

IDT Exhibit 1020.3781

an edit distance of 3 or greater from the RRBS barcodes used in that lane. The PhiX library from Illumina does not contain an index. As a result, the sequencer canproduce a low-quality index read from PhiX clusters. To remove PhiX reads prior to parsing, filter by index read quality and remove reads with quality less than 20.

[00208] Barcode sequences for dedicated read (DR) adaptors used in herein can be found in Table 2.

Table 2 Barcode sequences for dedicated read (DR) adaptors

6 NT BARCODE SEQUENCE AS READ BY THE SEQUENCER	BARCODE PAIRING (2-PLEX)
AACCAG	Duplex Set 1
TGGTGA	
AGTGAG	Duplex Set 2
GCACTA	
ACCTCA	Duplex Set 3
GTGCTT	
AAGCCT	Duplex Set 4
GTCGTA	
AAGAGG	Duplex Set 5
GGAGAA	
AGCATG	Duplex Set 6
GAGTCA	
CGTAGA	Duplex Set 7
TCAGAG	
CACAGT	Duplex Set 8
TTGGCA	

One of the duplex sets can be used in combination with any of the other remaining 14 individual barcodes.

[00209] The barcode sequences used can be chosen for their ability to parse properly and for color balancing.

[00210] Libraries generated using methods, compositions, and kits described herein can be sequenced using a custom Read 1 sequencing primer, e.g., MetSeq Primer 1, which can be included in a kit at a concentration of 25 μM. The Standard Read 1 Primer can be used when

IDT Exhibit 1020.3782

multiplexing with PhiX or other libraries. The standard primers provided in the Illumina sequencing kit can be sufficient for sequencing the barcodes and, if desired, the reverse read.

[00211] Methods, compositions, and kits provided herein can be used to produce directional bisulfite-converted libraries. The Read 1 sequencing primer can sequence the C-to-T converted strand and the overall nucleotide balance for the Read 1 sequencing primer can show a low proportion of C bases. Illumina can recommend how to obtain high-quality base calls from libraries containing unbalanced nucleotide ratios. These recommendations can differ by instrument. Illumina software can produce higher-quality reads from low-diversity samples such as RRBS. High quality results can be made if the sequencing instrument is running the following versions or later:

HiSeq- HCS v2.2.38 (includes RTA v1.18.61)

MiSeq- RTA v1.17.28

[00212] These software upgrades, in combination with the sequence diversity incorporated in the adaptors described herein, can produce high-quality RRBS reads without the need to spike in balanced library such as PhiX. However, spiking in 5% PhiX or another, previously characterized high-quality library can be useful for troubleshooting purposes in the event of a failed run, but can reduce the data output by 5%. Such a control can be used to tell if the sequencing run failed in general, or if there is a problem specific to the particular RRBS library being sequenced.

[00213] The first base of the read RRBS method contains a CpG methylation measurement. In principle, reads need only be long enough to accurately map them to the genome. However, mapping rates can be directly affected by sequence read length, and use of longer reads can potentially lead to more uniquely mapping reads and coverage of a greater number of CpG loci. Table 3 lists the percent of MspI fragments 40–300 nt in length that are uniquely or non-uniquely mappable to the hg19 reference genome. In this theoretical analysis, the 1,545,560 MspI fragments of hg19 that are 40–300 nt in length are bisulfite-converted in silico (assuming 0% methylation and 100% bisulfite conversion), trimmed to the indicated read length, then mapped using Bismark (www.bioinformatics.bbsrc.ac.uk/projects/bismark/).

[00214] Table 3. Percent of MspI fragments 40–300 nt in length that are uniquely or nonuniquely mappable to the hg19 reference genome

SINGLE END READ LENGTH	40 nt	50 nt	100 nt
Sequences mapping uniquely	65.2%	69.6%	83.9%
Sequences mapping non-uniquely	34.8%	30.4%	16.1%

IDT Exhibit 1020.3783

[00215] In addition to mappability, how read length affects CpG loci coverage can be considered. Many MspI fragments contain internal CpG's, so longer reads can sequence more CpGs. However, many MspI fragments can be smaller than 100 bp, and even smaller than 50 bp. For these fragments, long sequencing reads, or paired end reads, can provide no additional CpG data. Table 4 lists the number of CpG loci that can be expected to be covered by various sequencing protocols.

[00216] Table 4. Number of CpG loci expected to be covered by various sequencing protocols

SINGLE END READ LENGTH	40 nt	50 nt	100 nt
CpG loci covered	2,649,855	3,281,511	5,916,268

[00217] C. Amplified Library Storage

[00218] Amplified libraries can be stored at –20°C.

[00219] D. Data Analysis and Parsing Multiplex Libraries

[00220] The sequences of the barcodes can be entered prior to parsing (the sequences are listed in Table 3). With bisulfite-converted libraries, a slightly higher rate of unmatched barcodes relative to non- bisulfite-converted libraries can be observed. The methods can use edit distance 3 barcodes, which can allow one mismatch during parsing. This can significantly reduce the fraction of unmatched barcode reads.

[00221] The Duplicate Marking feature mentioned above can built into the NuGEN RRBS system, by adding an additional 6 nt to the index read. See Example 7 for more information on this feature.

[00222] Once the data have been parsed according to sample index, the reads can be trimmed before attempting alignment. Trimming can be done in two steps. First, any low-quality bases and adaptor sequences can be removed, then the sequence diversity provided by adaptors described herein can be removed. Example 8 contains detailed recommendations on trimming. At this point reads can be ready for downstream analysis, such as mapping to the genome and determining methylation status.

Example 6: Protocol

[00223] A. Overview

IDT Exhibit 1020.3784

[00224] As shown in **FIG. 38**, an RRBS library preparation process can be performed in five stages: 1. DNA digestion with MspI: 1.0 hour; 2. Adaptor ligation: 0.75 hours; 3. final repair: 0.25 hours; 4. bisulfite conversion: 3.0 hours; 5. amplification and purification: 1.5 hours. Total time to prepare amplified library: 6.5 hours.

[00225] Components in a kit can be color coded, with each color linked to a specific stage of the process. Performing each stage can involve making a master mix, then adding it to the reaction, followed by incubation. Master mixes can be prepared by mixing components provided for that stage.

[00226] Each library can be produced independently when performing multiplex sequencing. In some cases, adaptors are not mixed during the actual library construction protocol. Multiplexing can be achieved by mixing the amplified libraries prior to cluster formation. The barcode sequences in can be carefully chosen for their ability to parse properly and for color balancing.

[00227] B. Protocol Notes

[00228] Water provided with a kit or an alternate source of nuclease-free water can be used. In some cases, DEPC-treated water is used; in some cases, DEPC-treated water is not used. A minimum of four reactions can be set up at a time to avoid pipetting very small volumes. Components used in each step can be thawed and immediately placed on ice. In some cases, all reagents at not thawed once. Thawed reagents and reaction tubes can be kept on ice. After thawing and mixing buffer mixes, if any precipitate is observed, it can be re-dissolved completely prior to use. The buffer mix can be gently warmed for 2 minutes at room temperature followed by brief vortexing. In some cases, any enzyme, primer, or adaptor mixes is not warmed.

[00229] Small amounts of reagents can be placed into a reaction mix, pipetted up and down several times to ensure complete transfer from the pipet tip into the reaction mix. Pipetting a mix can involve gently aspirating and dispensing a volume that is at least half of the total volume of the reaction mix.

[00230] The thermal cycler can be allowed to reach the initial incubation temperature prior to placing the tubes or plates in the block. Fresh ethanol stocks can be used make 70% ethanol for the purification protocols. The ethanol mixes can be made fresh, carefully measuring both the ethanol and water with pipettes. Lower concentrations of ethanol in wash solutions can result in loss of yield as the higher aqueous content will dissolve the DNA and wash it off the beads or column.

[00231] C. Agencourt® RNAClean® XP Purification Beads

IDT Exhibit 1020.3785

[00232] The bead purification processes can include 1. binding of DNA to RNAClean XP beads; 2. magnetic separation of beads from supernatant; 3. ethanol wash of bound beads to remove salts, etc.; 4. elution of bound DNA from beads. Beads can be removed from 4°C and left at room temperature for at least 15 minutes before use. Beads can completely reach room temperature before using. Cold beads can reduce recovery. Beads can be fully resuspended by vortexing before adding to sample. The ratio of Agencourt RNAClean XP bead volume to sample volume can vary among different steps in the protocol. Furthermore, the bead:sample ratios used can differ from the standard Agencourt protocol. The beads can separate on the magnet for a full 5 minutes. Removing binding buffer before the beads have completely separated can impact DNA yields. After completing the binding step, bead loss can be minimized when removing the binding buffer. With the samples placed on the magnet, the amount of the binding buffer specified at each of the individual purification steps can be removed. Some liquid can remain at the bottom of the tube, helping to minimize bead loss at this step. Bead loss can be minimized throughout the procedure to maximize DNA yield. An ethanol wash can be freshly prepared from fresh ethanol stocks at the indicated concentration. Lower percent ethanol mixes can reduce recovery. During the ethanol washes, the samples can be kept on the magnet. Dispersion of the beads can be avoided. The magnet can keep the beads on the walls of sample wells or tubes. All residual ethanol can be removed prior to continuing with the next step. When removing the final ethanol wash, most of the ethanol can be removed first, then excess ethanol is collected at the bottom of the tube before removing the remaining ethanol. This process can reduce the required bead air drying time. After drying the beads for the time specified in the protocol, each tube can be carefully inspected to determine whether all the ethanol has evaporated before proceeding.

[00233] Strip tubes or partial plates can be firmly placed within the magnetic plate. Individual tubes can be difficult to position stably on the magnetic plates.

[00234] D. Programming the Thermal Cycler

[00235] A thermal cycler with a heat block designed for 0.2 mL tubes can be used, equipped with a heated lid, and with a capacity of 100 µL reaction volume. The programs can be prepared as shown in Table 5, following the operating instructions provided by the thermal cycler manufacturer. For thermal cyclers with an adjustable heated lid, the lid temperature can be set to 100°C when sample temperature reaches above 30°C. For thermal cyclers with a fixed temperature heated lid (e.g., ABI GeneAmp® PCR 9600 and 9700 models), default settings can be used (e.g., 100°C to 105°C).

IDT Exhibit 1020.3786

[00236] Table 5. Thermal cycler programming

Program 1 (MspI Digestion)	37°C – 60 min, hold at 4°C
Program 2 (Ligation)	25°C – 30 min, 70°C – 10 min, hold at 4°C
Program 3 (Final Repair)	60°C – 10 min, 70°C – 10 min, hold at 4°C
Program 4 (Bisulfite Conversion)	2 cycles of (95°C – 5 min, 60°C – 20 min) hold at 20°C
Program 5 (Library Amplification)	95°C – 2 min, 12 cycles of (95°C – 15 sec, 60°C – 1 min, 72°C – 30 sec), hold at 10°C

[00237] If less than 100 ng of intact gDNA is used, or if DNA is severely degraded (such as with FFPE), more than 12 cycles of PCR may be performed. As an estimate, one additional cycle of PCR for each two-fold reduction of input can be performed. For example, for 25 ng of intact gDNA, 14 cycles of PCR can be used.

[00238] E. MspI Digestion

[00239] The MspI Buffer Mix, MspI Enzyme Mix, and Nuclease-free Water can be removed from –20°C storage. Contents of MspI Enzyme Mix can be spun down and placed on ice. MspI buffer mix and nuclease-free water can be thawed at room temperature. MspI buffer mix can be mixed by vortexing, spun and placed on ice. Each 100 ng gDNA sample can be adjusted to 8.5 µL with nuclease-free water. A master mix can be prepared by combining MspI buffer mix and MspI enzyme mix in a 0.5 mL capped tube, according to the volumes shown in Table 6.

[00240] Table 6. MspI Master Mix (volumes listed are for a single reaction)

MspI Buffer Mix	MspI Enzyme Mix
1.0 µL	0.5 µL

[00241] 1.5 µL of the MspI master mix can be added to each sample tube. Mixing can be by pipetting; tubes can be capped, spun and placed on ice. Tubes can be placed in a thermal cycler programmed to run Program 1 (MspI Digestion; see Table 5): 37°C – 60 min, hold at 4°C. Tubes can be removed from the thermal cycler, spin to collect condensation and place on ice.

[00242] F. Ligation

[00243] A Ligation Buffer Mix, Ligation Adaptor Mixes, and Ligation Enzyme Mix can be removed from –20°C storage. Ligation Enzyme mix can be spun down and placed on ice.

IDT Exhibit 1020.3787

Ligation Buffer Mix (L1) and Ligation Adaptor Mixes (L2) can be thawed at room temperature. The solution can be mixed by vortexing, spun and placed on ice. The Ligation Buffer Mix (L1) can be extremely viscous. Care can be taken to ensure it is well mixed after thawing (it can be helpful to alternately vortex the tube right-side up and upside down). In some cases, Ligation Adaptor mixes are not warmed above room temperature. Heating can degrade performance.

[00244] Three µL of appropriate L2 Ligation Adaptor Mix can be added to each sample and mixed thoroughly by pipetting. Prior to use, a master mix can be made by combining nuclease-free water, L1, Ligation Enzyme Mix in a 0.5 mL capped tube according to the volumes in Table 7. The solutions can be mixed by pipetting slowly, without introducing bubbles, spun and placed on ice. The mix can be used immediately.

[00245] The L1 Ligation Buffer Mix can be very viscous. This reagent can be pipetted slowly, and care can be taken to ensure that the Ligation Master Mix and the ligation reactions are well mixed (visually observe that the solutions become homogeneous).

[00246] Table 7. Ligation Master Mix (volumes listed are for a single reaction)

WATER	2.0 µL
LIGATION BUFFER MIX	4.0 µL
LIGATION ENZYME MIX	1.0 µL

[00247] 7 µL Ligation Master Mix can be added to each reaction tube. The solution can be mixed thoroughly by pipetting slowly, gently spun and placed on ice. Incubation can be preceded immediately. Tubes can be placed in a pre-warmed thermal cycler programmed to run Program 2 (Ligation; see Table 5): 25°C – 30 min, 70°C – 10 min, hold at 4°C. The tubes can be removed from the thermal cycler, spun to collect condensation, and placed on ice.

[00248] **G. Final Repair**

[00249] The Final Repair Buffer Mix (FR1) and Final Repair Enzyme Mix (FR2) can be removed from –20°C storage. Contents of FR2 can be spun down and placed on ice. FR1 can be thawed at room temperature. The solutions can be mixed by vortexing, spun and placedon ice. A master mix can be prepared by combining FR1 and FR2 in a 0.5 mL capped tube, according to the volumes shown in Table 8.

[00250] Table 8. Final Repair Master Mix (volumes listed are for a single reaction)

FINAL REPAIR BUFFER	FINAL REPAIR ENZYME MIX	WATER

IDT Exhibit 1020.3788

MIX		
6.0 µL	0.5 µL	13.5µL

[00251] The solutions can be mixed by pipetting slowly, without introducing bubbles, spun for 2 seconds and place on ice. The mix can be used immediately.20 µL of the Final Repair Master Mix can be added to each of the 20 µL ligation reactions. The solution can be mixed by pipetting. The tubes can be capped, and spun, and placed on ice. The tubes can then be placed in a thermal cycler pre-heated to 60°C and programmed to run Program 3 (Final Repair; see Table 5): 60°C – 10 min, 70°C – 10 min, hold at 4°C. The tubes can be removed from the thermal cycler, spun to collect condensation and placed on ice.

[00252] H. Bisulfite Conversion

[00253] 1. QIAGEN EpiTect Fast DNA Bisulfite Kit (QIAGEN Cat. #59824 for 50 preps or Cat. #59826 for 200 preps) can be used for bisulfite conversion. Other commercial bisulfite conversion kits can also be used. The 40 µL product of the Final Repair reaction can be input directly into the bisulfite conversion kit. The input can be performed using the QIAGEN EpiTect Fast protocol "Bisulfite Conversion of Unmethylated Cytosines in DNA", on page 19, following the guidelines for 1–500 ng in a maximum volume of 40 µL (low concentration). 40 µL of input and 15 µL of DNA Protect Buffer can be used. The following thermal cycler conditions can be used: 95°C – 5 min, 60°C – 20 min, 95°C – 5 min, 60°C – 20 min, hold at 20°C. After thermal cycling, a blue ring at the meniscus can be observed. This can be normal and may not affect the results. The rest of the QIAGEN protocol can then be followed. Carrier RNA in the BL Buffer can be used. The purified, bisulfite-converted DNA can be eluted in 23 µL of EB. This can yield 20 µL of bisulfite-converted DNA ready for amplification. If necessary, the final eluted volume of bisulfite-converted DNA can be adjusted to 20 µL with nuclease-free water.

[00254] I. Library Amplification

[00255] The real-time protocol in Example 7 can be consulted to determine the number of cycles of PCR to perform (such as when starting with degraded DNA or significantly less than 100 ng of high-quality DNA). The Amplification Primer Mix (P2) and Amplification Enzyme Mix (P3)m can be removed from –20°C storage. P3 can be spun and placed on ice. P2 can be thawed at room temperature. The solutions can be mixed by vortexing, spun and placed on ice. The 20 µL product from the Bisulfite Conversion protocol can be obtained. A master mix can be made by combining P2 and P3 in an appropriately sized capped tube according to the volumes shown in Table 9. The solutions can be mixed by pipetting (care can be taken to avoid bubbles), spun and placed on ice.

[00256] Table 9. Amplification Master Mix (volumes listed are for a single reaction)

AMP PRIMER MIX (P2)	AMP ENZYME MIX (P3)
5 µL	25 µL

[00257] On ice, 30 µL of the Amplification Master Mix can be added to each sample. Tubes can be placed in a pre-warmed thermal cycler programmed to run Program 5 (Library Amplification; see Table 5): 95°C – 2 min, 12 cycles of (95°C – 15 sec, 60°C – 1 min, 72°C – 30 sec), hold at 10°C. If less than 100 ng of intact gDNA is used, or if the DNA is severely degraded (such as with FFPE), more than 12 cycles of PCR can be performed. Methods provided in Example 8 can be used in such situations. Tubes can be removed from the thermal cycler, spun to collect condensation and placed on ice.

[00258] J. Amplified Library Purification

[00259] The Agencourt RNAClean XP Beads and 70% ethanol set aside previously can be used if they are still at room temperature. The beads can be resuspended by vortexing the tube and added to the sample. In some cases, the beads are not spun after resuspending. At room temperature, 50 µL (1 volume) of the bead suspension can be added to each reaction. The bead suspension can be quite viscous; therefore slow pipetting can be performed to enhance accuracy. The solutions can be mixed thoroughly by pipetting 10 times. A multichannel pipettor can be used to make the incubation times uniform. The mixture can be incubated at room temperature for 10 minutes. The tubes can be transferred to the magnet and let stand 5 minutes to clear the solution of beads. 90 µL of the binding buffer can be carefully removed and discarded. Some of the volume can be left behind to reduce bead loss at this step. In some cases, the beads are not disperse; instead, they can stay on the walls of the tubes. Significant loss of beads at this stage will impact the amount of purified DNA, so in some cases, beads are not removed with the binding buffer or the wash. The 70% ethanol wash in the library purification step can involve removing samples from the magnet, fully resuspending the beads in the 70% ethanol wash, and re-magnetizing until the beads have cleared from the solution. The beads can be washed in following steps: the plate can be removed from the magnet, 150 µL of freshly prepared 70% ethanol can be added and the mixture can be pipetted to resuspend the beads; the beads can be cleared by re-magnetizing and a 70% ethanol wash can be performed using a pipette. The washing steps can be repeated twice. In some cases, the ethanol can be removed as much as possible with the final wash. At least 2 pipetting steps can be performed to allow excess ethanol to collect at the bottom of the tubes after removing most of the ethanol in the first pipetting step.

IDT Exhibit 1020.3790

The beads can be air dried on the magnet for a minimum of 10 minutes. In some cases each tube can be carefully to ensure that all the ethanol has evaporated. All residual ethanol can be removed prior to continuing. The tubes can then be removed from the magnet. 20 µL DNA Resuspension Buffer can be added to the dried beads. The beads can be resuspended by thorough mixing. The tubes can be transferred to the magnet and let stand for 5 minutes. 18 µL of the eluate can be carefully removed, ensuring as few beads as possible are carried over, and transferred to a fresh set of tubes. When pipetting any portion of this eluted library downstream, the library can be let stand briefly on a magnet to reduce bead carryover. The eluate can then be used to quantitatively and qualitatively assess the library. After quantitation, barcoded libraries that will be run in the same flow cell can be mixed in equimolar ratios prior to processing on the flow cell, cBot or Cluster Station.

[00260] K. Quantitative and Qualitative Assessment of the Library

[00261] Samples can be diluted 1:3 with water. 1 µL of the diluted sample can be loaded on the Bioanalyzer High Sensitivity DNA Chip. In some cases, fragment distribution can be as shown in **FIG. 28**. The three peaks at 200 bp, 265 bp, and 330 bp are due to MspI-containing micro-satellite repeats, and are characteristic of RRBS libraries made from human DNA. For qPCR for quantification, 250 bp can be used as library size for calculations.

Example 7 - PCR amplification when using degraded DNA or less than 100 ng of high-quality DNA

[00262] Standard RRBS libraries can be generated using 100 ng of high-quality, MspI-digested genomic DNA. Library Amplification can be performed by preparing an Amplification Master Mix and adding 30 µL of this to 20 µL of bisulfite-converted sample for a total PCR volume of 50 µL. In some cases, the following two options can be used for degraded DNA or less than 100 ng of high-quality DNA: A) If high-quality DNA is used and the amount of DNA is known, the number of PCR cycles can be adjusted to obtain sufficient material for analysis and sequencing. One additional cycle of PCR can be performed for each two-fold reduction of input. For example, for 25 ng of intact gDNA, 14 cycles of PCR can be used. B) If degraded DNA is used or if the amount and/or quality of the DNA is unknown, a 1/5 scale real-time PCR with an aliquot of your sample can be performed as follows: A master mix can be made by combining P2 and P3 in an appropriately sized capped tube according to the volumes shown in Table 10. P3 Enzyme Mix can be added at the last moment and mixed well by pipetting taking care to avoid bubbles, spun and placed on ice. EvaGreen (Biotium, Cat. #31000) is less inhibitory to PCR than SYBR Green®, however SYBR Green can also be used at a 1X final concentration in the real-

IDT Exhibit 1020.3791

time PCR reaction. For each library, 6 µL of the Amplification Master Mix can be aliquoted into a well of on a real-time PCR plate. 4 µL of sample can be added to each well (after bisulfite conversion but before PCR amplification) for a total real-time PCR volume of 10 µL. Real-time PCR can be performed with the following cycling conditions: 95°C – 2 min, 30 cycles of (95°C – 15 sec, 60°C – 1 min, 72°C – 30 sec). The SYBR Green channel can be monitored. In some cases, ROX normalization is not used. After thermal cycling is complete, the amplification plot can be examined for each sample to determine the number of PCR cycles to perform during the Library Amplification step. A cycle in the late exponential phase can be selected. For example, in the amplification plot shown in **FIG. 29**, cycle 9 was chosen. This provides sufficient amplification without entering the plateau phase.

[00263] Table 10. Amplification Master Mix (volumes listed are for a single reaction)

AMP PRIMER MIX (P2)	20X EVAGREEN	AMP ENZYME MIX (P3)
1.0 µL	0.5 µL	4.5 µL

[00264] If using an Applied Biosystems real-time PCR instrument, Plot > Rn vs. Cycle, not Plot > deltaRn vs. Cycle is selected. Selecting Plot > deltaRn vs. Cycle may give unexpected results similar to those shown in **FIG. 30**.

[00265] After determining the optimal number of PCR cycles to perform, the remaining 16 µL of bisulfite-converted library can be amplified as follows: A master mix can be made by combining P2 and P3 in an appropriately sized capped tube according to the volumes shown in Table 11. P3 Enzyme Mix can be added at the last moment and mix well by pipetting taking care to avoid bubbles, spun and placed on ice. For each library, 24 µL of the Amplification Master Mix can be aliquoted into a well of on a real-time PCR plate. 16 µL of sample can be added to each well. PCR can be performed with the following cycling conditions, where N = number of cycles determined from the above real-time PCR assay: 95°C – 2 min, N cycles of (95°C – 15 sec, 60°C – 1 min, 72°C – 30 sec, hold at 10°C. Proceed with the Amplified Library Purification protocol in Example 6, Section J, by adding 1 volume (40 µL) of Agencourt RNAClean XP Beads to each amplified sample and following the remaining steps of the purification protocol.

[00266] Table 11. Amplification Master Mix (volumes listed are for a single reaction)

AMP PRIMER MIX (P2)	AMP ENZYME MIX (P3)
4.0 µL	20 µL

IDT Exhibit 1020.3792

Example 8 – Index structure and index read recommendations

[00267] The 6-base barcodes can be used for sample multiplexing. In addition to the 6-base barcode, the adaptor can contain 6 random bases immediately following the 6-base barcode, for a total of 12 bases. The additional 6 bases can be used for duplicate read determination using the Duplicate Marking Tool. The libraries can be sequenced using 12 cycles for the index read. The 6-base barcode sequences can be found in Table 2. In some cases, the MiSeq and HiSeq instruments do not provide a simple way to obtain the sequence information contained in the 12-base pair index read including the 6 random bases that can be necessary for duplicate read determination. Several methods can be used to generate the necessary index fastq file are provided below.

[00268] MiSeq Instrument

[00269] Parsing multiplex runs using the MiSeq built-in Illumina software can replace the barcode sequence from each library with a numerical substitute, which removes the duplicate information provided by the N6 sequence present after the barcode. To retrieve this information using the MiSeq instrument, one of the options can be used: Option 1: contact Illumina Technical Support and request a modification of the MiSeq config file to allow generation of an index fastq file during data analysis. Option 2: Modify the MiSeq config file to allow generation of an index fastq file during data analysis: 1. Stop the MiSeq Reporter process. 2. Locate the "MiSeq Reporter.exe.config" file located in C:/Illumina/MiSeq Reporter. 3. Open config file and search for a line that reads:

<add key="CreateFastqForIndexReads" value="0"/>

1. If this line is present, change the value from "0" to "1".
2. If this line is not present, add the line to the config file using the add keys function under the app Settings tab with the value set as "1".

[00270] 4. Restart the MiSeq reporter process. 5. Re-queue the run for data analysis if required. The 6-base barcodes followed by NNNNNN can be entered into the sample sheet to enable proper multiplex library parsing.

[00271] HiSeq Instrument

[00272] When setting up the HiSeq run, 12 bases of index sequencing (no sample sheet is required) can be specified. However, if a sample sheet to be included, the 6 bases of the actual barcode can be specified. In some cases, the N6 in your sample sheet is not included. The method described below can be used to parse and generate the N6 index fastq files for HiSeq using CASAVA: 1. Browse to the location of the run folder (called "RunFolder" in this

IDT Exhibit 1020.3793

example). 2. Open the "Data" folder in your run (/RunFolder/Data). 3. Run CASAVA. Use the "--use-bases-mask Y*, I6Y*" option to generate an Index fastq file along with the forward read (for paired end reads use "--use-bases-mask Y*,I6Y*, Y*"). For example:

"/illumina/pipeline/CASAVA-1.8.2/bin/configureBclToFastq.pl --input-dir /data/Runs/RunFolder/Data/Intensities/Basecalls/ --fastq-cluster-count 0 --use-basesmask
Y*,I6Y* --ignore-missing-control".

[00273] To generate the read and index fastq files without parsing, the –usebases- mask option to "--use-bases-mask Y*,Y*" can be modified. The generated fastq files can then be parsed using alternative software. 4. Open /RunFolder/Unaligned. 5. Type "nohup make -j 12" (this command will produce the actual fastq files, the 12 indicates how many threads to use). 6. The fastq files will be located in /RunFolder/Unaligned/Project_UniqueID

Example 9 -- Diversity trimming

[00274] Following sequencing and parsing, adaptors can be trimmed and the additional sequence added by the diversity adaptors to the 5' and 3' ends of the insert can be removed before downstream analysis.

[00275] Removing Adaptor Sequences

[00276] To accurately identify the diversity sequence and MspI (C^CGG) site, adaptor sequences present on the 3' end of your reads can be trimmed. Trim Galore (www.bioinformatics.babraham.ac.uk/projects/trim_galore/) can be used for this purpose. Other programs can also be used. Trim Galore may also trim some or all of a read due to low quality. In some cases, the program is run with default parameters, and the --RRBS option is not used. Single end reads can be trimmed with the following command:

trim_galore sampleR1.fq

[00277] Paired end reads can be trimmed with the following command:

trim_galore --adaptor2 AAATCAAAAAAAC sampleR1.fq sampleR2.fq

[00278] Diversity Trimming and Filtering

[00279] Additional filtering and trimming can be performed by a custom python script provided by NuGEN. The script can be obtained from NuGEN Technical Support at techserv@nugen.com. The script removes reads that do not contain an MspI site signature (YGG) at the 5' end. For paired end data an MspI site signature can be required at the 5' end of both sequences. The script will generate new file(s) with "_trimmed.fq" appended to the filename. In it, all reads have been trimmed at the 5' end to remove the diversity sequence (0–

3bases), and all reads can begin with YGG, where Y is C or T. On the 3' end, 5 bases are trimmed from every read (6 bases are trimmed for paired-end to prevent alignment issues).

[00280] Examples of Trimming the 5' Ends of Forward Reads

[00281] Underlined bases denote sequence derived from the adaptor. In this example, the fragment is derived from the genomic sequence, starting and ending with MspI sites:

5' CCGGAGTT…AAGGGCCGG 3'
3' GGCCTCAA…TTCCCGGCC 5'

[00282] After MspI digestion:

5' CGGAGTT…AAGGGC 3'
3' CTCAA…TTCCCGGC 5'

[00283] After ligation to adaptors, both with three bases of diversity:

5' RDDCGGAGTT…AAGGGCCGHHY 3'
3' YHHGCCTCAA…TTCCCGGCDDR 5'

[00284] After bisulfite conversion and PCR amplification of the top strand:

5' RDDYGGAGTT…AAGGGTCGHHY 3'
3' YHHRCCTCAA…TTCCCAGCDDR 5'

[00285] Assuming the insert is smaller than the read length, the forward read after Trim Galore is used to trim the adaptor from the 3' end will be:

5' RDDYGGAGTT…AAGGGTCGHHY 3'

[00286] The result of the NuGEN diversity trim of the forward read (if it's a single-end read) will be:

5' YGGAGTT…AAGGGT 3'

[00287] The reverse read after Trim Galore is used will be:

5' RDDCGACCCTT…AACTCCRHHY 3'

[00288] The result of the NuGEN diversity trim of the reverse read:

5' ACCCTT…AACT 3'

[00289] The adaptor can contain between 0 and 3 bases of diversity.

[00290] Tables 12–15 show how the script trims all types of adaptor variation.

[00291] Table 12. Trimming of the 5' ends of forward reads

READ BEFORE TRIMMING	READ AFTER TRIMMING
5' YGGAGTT...	5' YGGAGTT...
5' DYGGAGTT...	5' YGGAGTT...
5' DDYGGAGTT...	5' YGGAGTT...
5' RDDYGGAGTT...	5' YGGAGTT...

Note: If YGG is not found in the first 6 bases, the read is discarded.

[00292] Table 13. Trimming of the 3' ends of forward reads

READ BEFORE TRIMMING	READ AFTER TRIMMING
5' ...AAGGGTCG	5' ...AAG
5' ...AAGGGTCGH	5' ...AAGG
5' ...AAGGGTCGHH	5' ...AAGGG
5' ...AAGGGTCGHHY	5' ...AAGGGT

[00293] Table 14. Trimming of the 5' ends of reverse reads

READ BEFORE TRIMMING	READ AFTER TRIMMING
5' CGACCCTT…	5' ACCCTT…
5' DCGACCCTT…	5' ACCCTT…
5' DDCGACCCTT…	5' ACCCTT…
5' RDDCGACCCTT…	5' ACCCTT…

Note: If CGR is not found in the first 6 bases, the read is discarded.

[00294] Table 15. Trimming of the 3' ends of reverse reads

READ BEFORE TRIMMING	READ AFTER TRIMMING
5' …AACTCCRHHY	5' ...AACTC
5' …AACTCCRHH	5' ...AACT
5' …AACTCCRH	5' ...AAC
5' …AACTCCR	5' ...AA

[00295] The script accepts as input one or two fastq file strings, given either as complete filenames or as a pattern in quotes. When a pattern is given, the script will find all the filenames matching a specified pattern according to the rules used by the Unix shell (*,?).

[00296] Example usage for single end reads with a complete filename:

python trimRRBSdiversityAdaptCustomers.py -1 sample_R1.fq

[00297] with a pattern:

python trimRRBSdiversityAdaptCustomers.py -1 "s_R1_1_edit3_6bp_BC??.fq"

[00298] Example usage for paired end reads with a complete filename:

python trimRRBSdiversityAdaptCustomers.py -1 sample_R1.fq -2 sample_R2.fq

[00299] with a pattern:

python trimRRBSdiversityAdaptCustomers.py -1 "s_R1*.fq" -2 “s_R2*.fq”

[00300] The help option of this script can be accessed for more details (-h). The script outputs a trimmed fastq file that can be used for downstream analysis including Bismark.

[00301] Following trimming, an additional algorithm can be used to remove sequencing duplicates. The Duplicate Marking tool can utilize information provided by the unique N6 sequence to discriminate between true PCR duplicates and independent adaptor ligation events to fragments with the same start site resulting in the recovery of more usable data. Using the index sequence reads, the Duplicate Marking tool can utilize the N6 random sequence found adjacent to the barcode sequence (as shown in **FIGs. 32** and **37**) in the sequence appended to nucleic acid inserts generated using the methods outlined in Example 4-8 to unambiguously identify other PCR duplicates comprising the same N6–barcode sequence and removing all but one of these reads.

Example 10 -- A simplified workflow for reduced representation bisulfite sequencing

[00302] RRBS can be used to obtain quantitative DNA methylation information across many features of the genome with approximately 50-fold fewer sequencing reads than can be needed with whole genome bisulfite sequencing which can result in substantial sequencing cost reduction. The methods described herein can be used for generating RRBS libraries in less than one day, without gel purification. The methods can entail digesting samples, ligating adaptors, performing a final repair, bisulfite converting, and PCR amplifying to generate libraries ready for NGS sequencing platforms such as Illumina. Additional benefits of the methods provided herein can include improved sequencing performance without the need for PhiX spike in, and the ability to mark PCR duplicates.

IDT Exhibit 1020.3797

[00303] DNA methylation can be a central component of epigenetic regulation, and can play a role in development, environmental exposure, and diseases including cancer. In humans, the predominant form of DNA methylation can be 5-methyl cytosine. When present, 5-mC can occur in the context of the CpG sequence motif. RRBS can utilize the restriction enzyme MspI, which can recognize and cleave the sequence CCGG regardless of the methylation state of the central CpG.

[00304] Reducing Representation to ~2%

[00305] After digestion with MspI, ~2% of the genome can be represented in fragments under 500 bp. However, many of these fragments can be from CpG-rich regions. By focusing RRBS sequencing on small fragments, the methylation status of these regions can be determined with much less sequencing than would be required using a whole genome bisulfite sequencing approach.

[00306] Challenges with Sequencing RRBS Libraries

[00307] On sequencers such as provided by Illumina, cluster identification as well as phasing and color matrix calculations can be negatively impacted by a lack of sequence complexity in the sample. RRBS libraries all begin with YGG (Y = C or T), which can make these libraries challenging to sequence. These issues can be mitigated by reducing cluster density and/or mixing in a high diversity library, such as a PhiX control library. This approach can reduce the number of useful reads as PhiX levels of 40% or higher may be needed to ensure a successful run.

[00308] Adding Diversity to RRBS Reads

[00309] The challenges described above can be overcome by adding diversity in the form of 0 to 3 bases of sequence between the sequencing primer and the insert. This can ensure all 4 bases are present during the first few cycles of sequencing. In addition, the YGG signature can be de-phased so that no cycle contains only G across all clusters.

[00310] The methods described herein can be multiplexed up to 16 samples per lane, can require no gel purification steps, and can have many advantages, including: simplified workflow that can be completed in a single day; adaptor features (i.e., sequence diversity) which can provide enhanced sequence quality without PhiX spike in, the ability to identify and remove PCR duplicates; and highly reproducible data that can displays excellent concordance with published methylation profiles such as shown in **FIG. 34**.

Example 11 -- Generation of and analysis of reduced representation bisulfite sequencing library from IMR-90 human genomic DNA using diversity adaptors

IDT Exhibit 1020.3798

[00311] In this example, an RRBS nucleic acid library from 25 ng of genomic DNA (gDNA) from the IMR-90 human cell line was generated using the kit, protocol, considerations, and analytical software outlined and described in Examples 4-9. Briefly, following digestion of 25 ng of gDNA from IMR-90 human cells with MspI to produce DNA fragments, the DNA fragments are combined with a master mix containing the appropriate diversity adapters (i.e., D0 ,D1, D2, D3, or a mix of D0-D3), ligation buffer, water and ligation enzyme mix, placed in a thermal cycler and run through a program on the thermal cycler that comprised of a single cycle comprising incubations at 25 C for 30 minutes, 70 C for 10 minutes, and then holding the mixture at 4 C. Following ligation, the ligation complexes were subjected to a post-ligation purification followed by a final repair reaction entailing incubation of the ligation complexes with a DNA repair enzyme in a thermal cycler programmed for a single cycle of 60 C incubation for 10 minutes. Following final repair, the samples were subjected to bisulfite conversion, wherein 20 uL of the products from the final repair reactions were inputted directly into a Qiagen EpiTect Fast DNA Bisulfite Kit according to manufacturer protocols and subjected to two cycles of 95 C for 5 minutes, and 60 C for 20 minutes on a thermal cycler followed by elution of the purified, bisulfite converted DNA libraries.. The bisulfite converted DNA libraries were then PCR amplified and purified. Following library amplification and purification, the libraries were pooled and sequenced on an Illumina MiSeq sequencer according to manufacturer instructions with a 40 nucleotide forward read, 6 nucleotide index read, and 25 nucleotide reverse read. The sequence reads were analyzed using the *in silico* diversity trimming and duplicate marking computer programs described in Example 9. This procedure was repeated two more times to generate technical replicates. The sequencing analysis was then compared between the replicates.

[00312] Following sequencing and *in silico* analysis of sequence reads obtained from the sequencing, a high coverage overlap can be seen between the three RRBS technical replicates as shown in **FIG. 33**. Moreover, as shown in Table 16, each of the technical replicates produced similar high levels of total mapping, as well as other measures.

[00313] Table 16. Replicate RRBS libraries from 25 ng IMR90 gDNA

	Replicate 1	Replicate 2	Replicate 3
Uniquely Mapping	67.60%	67.47%	66.98%
Non-Uniquely Mapping	23.12%	22.70%	23.93%

IDT Exhibit 1020.3799

Total Mapping	90.72%	90.18%	90.91%
Methyl CG	44.05%	44.14%	43.17%
Methyl CHG	0.37%	0.37%	0.36%
Methyl CHH	0.31%	0.31%	0.30%

[00314] Additionally, **FIG. 34** shows a concordance in methylation levels between IMR-90 RRBS replicates (left) and IMR-90 whole genome bisulfite sequencing (WGBS) libraries (right).

[00315] While preferred embodiments are shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the instant disclosure. It should be understood that various alternatives to the embodiments described herein may be employed in practicing the disclosure. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

IDT Exhibit 1020.3800

CLAIMS

WHAT IS CLAIMED IS:

1. A method for increasing diversity of sequences at ends of a plurality of polynucleotides in a sample, the method comprising introducing a pool of oligonucleotide sequences to 3' ends of the plurality of polynucleotides,
   a. wherein each oligonucleotide sequence in the pool of oligonucleotide sequences comprises a sequence complementary to a sequencing primer,
   b. wherein the pool of oligonucleotide sequences comprises a first sequence, wherein a terminal 3' nucleotide base of the sequencing primer is capable of annealing to a terminal 5' nucleotide base of the first sequence, wherein following introduction the 5' end of the first sequence is adjacent to a 3' end of a polynucleotide from the plurality of polynucleotides,
   c. wherein the pool of oligonucleotide sequences comprises a second sequence, wherein the second sequence terminates at its 5' end with an additional base relative to the 5' end of the first sequence, wherein following introduction the 5' end of the second sequence is adjacent to a 3' end of a polynucleotide from the plurality of polynucleotides;
   d. wherein the pool of oligonucleotide sequences comprises a third sequence, wherein the third sequence terminates at its 5' end with two additional bases relative to the 5' end of the first sequence, wherein following introduction the 5' end of the third sequence is adjacent to a 3' end of a polynucleotide from the plurality of polynucleotides;
   e. wherein the pool of oligonucleotide sequences comprises a fourth sequence, wherein the fourth sequence terminates at its 5' end with three additional bases relative to the 5' end of the first sequence, wherein following introduction the 5' end of the fourth sequence is adjacent to a 3' end of a polynucleotide from the plurality of polynucleotides; and
   f. wherein the introducing generates a plurality of oligonucleotide-polynucleotide complexes, thereby increasing diversity of sequences at ends of the plurality of polynucleotides.
2. The method of claim 1, wherein the sample is a whole genome sample.
3. The method of claim 1, wherein the sample comprises a polynucleotide library.

IDT Exhibit 1020.3801

4. The method of any of claims 1-3, further comprising fragmenting the polynucleotides prior to introducing the pool of oligonucleotide sequences to the 3' ends of the plurality of polynucleotides.
5. The method of claim 4, wherein the fragmenting is performed by contacting the polynucleotides with an enzyme.
6. The method of claim 5, wherein the enzyme is a restriction enzyme.
7. The method of claim 6, wherein the restriction enzyme is a methylation insensitive restriction enzyme.
8. The method of claim 7, wherein the methylation insensitive restriction enzyme is MspI.
9. The method of claim 8, wherein the additional base of the second sequence is an adenine, thymine, or cytosine base.
10. The method of claim 8 or 9, wherein each of the two additional bases of the third sequence is an adenine, thymine, or cytosine base.
11. The method of any of claims 8-10 wherein a 3' most base of the three additional bases of the fourth sequence comprises a thymine or cytosine base, and each of the two other bases of the three additional bases of the fourth sequence comprises an adenine, thymine, or cytosine base.
12. The method of any of claims 1-11, wherein each oligonucleotide sequence in the pool of oligonucleotide sequences is within a strand of a duplexed adaptor.
13. The method of claim 12, wherein the duplexed adaptor is a partial duplexed adaptor comprising a short strand and a long strand.
14. The method of claim 12, wherein the duplexed adaptor is a partial duplexed adaptor comprising two strands of equal length.
15. The method of claim 13, wherein the short strand is a non-ligation strand of the duplexed adaptor, and the long strand is a ligation strand of the duplexed adaptor.
16. The method of claim 15, wherein the non-ligation strand comprises a 5' overhang.
17. The method of claim 16, wherein the 5' overhang comprises two bases.
18. The method of claim 17, wherein the two bases are 3'-GC-5'.
19. The method of any of the above claims, wherein the introducing comprises a ligation reaction.
20. The method of claim 19, wherein the ligation reaction comprises cohesive end ligation.

IDT Exhibit 1020.3802

21. The method of claim 19, wherein the ligation reaction comprises blunt end ligation.

22. The method of claim 20, wherein the plurality of polynucleotides are strands of double-stranded nucleic acids, wherein each 5' end of the double-stranded nucleic acids comprises a ligation end and each 3' end comprises a non-ligation end, wherein each of the non-ligation ends of the double-stranded nucleic acid comprises part of the polynucleotide sequence.

23. The method of claim 22, wherein the 3'-GC-5' of the non-ligation strand of the duplexed adaptor anneals to a 5'-CG-3' sequence of the ligation end of the double-stranded sequence comprising the polynucleotide sequence.

24. The method of claim 23, wherein the 3'-GC-5' of the non-ligation strand of the duplexed adaptor is 5' to the terminal 5' nucleotide base of the first sequence.

25. The method of claim 23 or 24, wherein the 3'-GC-5' of the non-ligation strand of the duplexed adaptor is 5' to the additional base of the second sequence relative to the 5' end of the first sequence.

26. The method of any of claims 23-25, wherein the 3'-GC-5' of the non-ligation strand of the duplexed adaptor is 5' to the two additional bases of the third sequence relative to the 5' end of the first sequence.

27. The method of any of claims 23-26, wherein the 3'-GC-5' of the non-ligation strand of the duplexed adaptor is 5' to the three additional bases of the fourth sequence relative to the 5' end of the first sequence.

28. The method of any of claims 16-20, wherein the ligation strand is capable of ligating to a 5' end of the plurality of polynucleotides, and wherein a nick is formed between the non-ligation strand and the end of the plurality of polynucleotides.

29. The method of claim 28, further comprising an extending step, wherein the non-ligation end of the plurality of polynucleotides is extended using a polymerase in the presence of a mixture of dNTPs using the ligation strand of the adaptor as template, thereby introducing the pool of oligonucleotide sequences to the 3' ends of the plurality of polynucleotides.

30. The method of claim 29, wherein the ligation strand of each adaptor duplex in the pool comprise cytosine analogs resistant to bisulfite treatment and the mixture of dNTPs comprises dATP, dTTP, dGTP, and dCTP, whereby the extending generates the plurality of oligonucleotide sequence-polynucleotide complexes, wherein each of the plurality of oligonucleotide sequence-polynucleotide complexes comprises complementary ends,

IDT Exhibit 1020.3803

wherein one end comprises cytosine bases sensitive to bisulfite treatment, while the opposite end comprises cytosine analogs resistant to bisulfite treatment.

31. The method of claim 29, wherein the ligation strand of each adaptor duplex in the pool of adaptor duplexes comprise cytosine bases sensitive to bisulfite treatment and the mixture of dNTPs comprises dATP, dTTP, dGTP, and a dCTP analog resistant to bisulfite treatment, whereby the extending generates the plurality of oligonucleotide sequence-polynucleotide complexes, wherein each of the plurality of oligonucleotide sequence-polynucleotide complexes comprises complementary ends, wherein one end comprises cytosine bases sensitive to bisulfite treatment, while the opposite end comprises cytosine analogs resistant to bisulfite treatment.

32. The method of any of claims 1-3, wherein each oligonucleotide sequence in the pool of oligonucleotide sequences is a primer, wherein the pool of oligonucleotide sequences comprises the primer and a complement thereof.

33. The method of claim 32, wherein the introducing comprises an amplification reaction.

34. The method of claim 33, wherein the amplification reaction comprises polymerase chain reaction.

35. The method of claim 33 or 34, wherein each of the primers in the pool comprises cytosine bases sensitive to bisulfite treatment, wherein the amplification is conducted in the presence of a mixture of dNTPs comprising dATP, dTTP, dGTP, and a dCTP analog resistant to bisulfite treatment, thereby generating the plurality of oligonucleotide sequence-polynucleotide complexes, wherein each of the plurality of oligonucleotide sequence-polynucleotide complexes comprises complementary ends, wherein one end comprises cytosine bases sensitive to bisulfite treatment, while the opposite end comprises cytosine analogs resistant to bisulfite treatment.

36. The method of claim 33 or 34, wherein each of the primers in the pool of primers comprises cytosine analogs resistant to bisulfite treatment, wherein the amplification is conducted in the presence of a mixture of dNTPs comprising dATP, dTTP, dGTP, and dCTP, thereby generating the plurality of oligonucleotide sequence-polynucleotide complexes, wherein each of the plurality of oligonucleotide sequence-polynucleotide complexes comprises complementary ends, wherein one end comprises cytosine bases sensitive to bisulfite treatment, while the opposite end comprises cytosine analogs resistant to bisulfite treatment.

IDT Exhibit 1020.3804

37. The method of claims 30, 31, 35, or 36, wherein the cytosine analog resistant to bisulfite treatment is 5-methylcytosine, 5-hydroxymethylcytosine, or 5-propynylcytosine.

38. The method of claims 30, 31, 35, or 36, wherein the cytosine analog resistant to bisulfite treatment is 5-methylcytosine.

39. The method of any of the above claims, further comprising amplifying the plurality of oligonucleotide sequence-polynucleotide complexes, wherein the amplifying comprises solid-phase nucleic acid amplification to generate a plurality of clusters, wherein each cluster represents a plurality of copies of each of the plurality of oligonucleotide sequence-polynucleotide complexes.

40. The method of claim 39, further comprising treating the plurality of oligonucleotide sequence-polynucleotides complexes with bisulfite prior to the amplification, wherein the treating generates a plurality of oligonucleotide sequence-polynucleotide complexes comprising non-complementary ends.

41. The method of claim 40, wherein the treating with bisulfite prior to amplification generates the sequence complementary to a sequencing primer in each oligonucleotide sequence in the pool of oligonucleotide sequences comprises.

42. The method of any of claims 39-41, further comprising sequencing each of the plurality of clusters, wherein the sequencing comprises hybridizing the sequencing primer to the sequence complementary to a sequencing primer, and conducting sequencing by synthesis.

43. The method of claim 42, wherein the introducing of the pool of oligonucleotide sequences to the polynucleotides improves identification of each of the plurality of clusters generated from the plurality of oligonucleotide sequence-polynucleotide complexes relative to the identification of each cluster from a plurality of clusters generated from the plurality of polynucleotides not introduced to the pool of oligonucleotide sequences.

44. The method of claim 42, wherein the sequencing comprises paired-end sequencing.

45. A method for generating a bisulfite converted library, the method comprising:

a.) introducing a pool of oligonucleotide sequences to each end of a plurality of polynucleotides from a sample,

i. wherein each oligonucleotide sequence in the pool of oligonucleotide sequences comprises a sequence complementary to a sequencing primer and a

IDT Exhibit 1020.3805

sequence element adjacent to the sequence complementary to a sequencing primer,

ii. wherein the pool of oligonucleotide sequences comprises a first sequence, wherein the first sequence terminates at its 3' end with the sequence element, wherein following ligation the 3' end of the sequence element is adjacent to a 5' end of a polynucleotide from the plurality of polynucleotides,

iii. wherein the pool of oligonucleotide sequences comprises a second sequence, wherein the second oligonucleotide sequence terminates at its 3' end with an additional base relative to the 3' end of the first sequence, wherein following ligation the 3' end of the second sequence is adjacent to a 5' end of a polynucleotide from the plurality of polynucleotides;

iv. wherein the pool of oligonucleotide sequences comprises a third sequence, wherein the third sequence terminates at its 3' end with two additional bases relative to the 3' end of the first sequence, wherein following ligation the 3' end of the third sequence is adjacent to a 5' end of a polynucleotide from the plurality of polynucleotides;

v. wherein the pool of oligonucleotide sequences comprises a fourth sequence, wherein the fourth sequence terminates at its 3' end with three additional bases relative to the 3' end of the first sequence, wherein following ligation the 3' end of the fourth sequence is adjacent to a 5' end of a polynucleotide from the plurality of polynucleotides;

vi. wherein the introducing generates a plurality of oligonucleotide sequence-polynucleotide complexes; and

b.) treating the plurality of oligonucleotide sequence-polynucleotide complexes with bisulfite, thereby generating a bisulfite converted polynucleotide library.

46. The method of claim 45, wherein the sample is a whole genome sample.

47. The method of claim 45, wherein the sample comprises a polynucleotide library.

48. The method of any of claims 45-47, further comprising fragmenting the polynucleotides prior to introducing the pool of oligonucleotide sequences to the 3' ends of the plurality of polynucleotides.

49. The method of claim 48, wherein the fragmenting is performed by contacting the polynucleotides with an enzyme.

50. The method of claim 49, wherein the enzyme is a restriction enzyme.

IDT Exhibit 1020.3806

51. The method of claim 50, wherein the restriction enzyme is a methylation insensitive restriction enzyme.
52. The method of claim 51, wherein the methylation insensitive restriction enzyme is MspI.
53. The method of claim 52, wherein the additional base of the second sequence is an adenine, thymine, or guanine base,
54. The method of claim 52 or 53, wherein each of the two additional bases of the third sequence is an adenine, thymine, or guanine base.
55. The method of any of claims 52-54 wherein a 5' most base of the three additional bases of the fourth sequence comprises an adenine or guanine base, and each of the two other bases of the three additional bases of the fourth sequence comprises an adenine, thymine, or guanine base.
56. The method of any of claims 45-55, wherein each oligonucleotide sequence in the pool of oligonucleotide sequences is within a strand of a duplexed adaptor.
57. The method of claim 56, wherein the duplexed adaptor is a partial duplexed adaptor.
58. The method of claim 57, wherein the strand is a ligation strand of the duplexed adaptor.
59. The method of claim 58, wherein the non-ligation strand comprises a 5' overhang.
60. The method of claim 59, wherein the 5' overhang comprises two bases.
61. The method of claim 60, wherein the two bases are 3'-GC-5'.
62. The method of any of claims 45-61, wherein the introducing comprises a ligation reaction.
63. The method of claim 62, wherein the ligation reaction comprises cohesive end ligation.
64. The method of claim 62, wherein the ligation reaction comprises blunt end ligation.
65. The method of claim 63, wherein the plurality of polynucleotides are double-stranded nucleic acids, wherein each 5' end of the double-stranded nucleic acids comprises a ligation end and each 3' end comprises a non-ligation end.
66. The method of any of claims 63-65, wherein each duplexed adaptor in the pool comprises a ligation strand and a non-ligation strand, wherein the ligation strand is capable of ligating to the ligation end of the plurality of polynucleotides, and wherein a

IDT Exhibit 1020.3807

nick is formed between the non-ligation strand and the non-ligation end of the plurality of polynucleotides.

67. The method of claim 66, further comprising an extending step, wherein the non-ligation end of the plurality of polynucleotides is extended using a polymerase in the presence of a mixture of dNTPs using the ligation strand of the adaptor as template, thereby introducing the pool of oligonucleotide sequences to the 3' ends of the plurality of polynucleotides.

68. The method of claim 67, wherein the ligation strand of each adaptor duplex in the pool comprise cytosine analogs resistant to bisulfite treatment and the mixture of dNTPs comprises dATP, dTTP, dGTP, and dCTP, whereby the extending generates the plurality of oligonucleotide sequence-polynucleotide complexes, wherein each of the plurality of oligonucleotide sequence-polynucleotide complexes comprises complementary ends, wherein one end comprises cytosine bases sensitive to bisulfite treatment, while the opposite end comprises cytosine analogs resistant to bisulfite treatment.

69. The method of claim 68, wherein the ligation strand of each adaptor duplex in the pool of adaptor duplexes comprise cytosine bases sensitive to bisulfite treatment and the mixture of dNTPs comprises dATP, dTTP, dGTP, and a dCTP analog resistant to bisulfite treatment, whereby the extending generates the plurality of oligonucleotide sequence-polynucleotide complexes, wherein each of the plurality of oligonucleotide sequence-polynucleotide complexes comprises complementary ends, wherein one end comprises cytosine bases sensitive to bisulfite treatment, while the opposite end comprises cytosine analogs resistant to bisulfite treatment.

70. The method of any of claims 45-47, wherein each oligonucleotide sequence in the pool of oligonucleotide sequences is within a primer, wherein the pool of oligonucleotide sequences comprises the primer and a complement thereof.

71. The method of claim 70, wherein the introducing comprises an amplification reaction.

72. The method of claim 71, wherein the amplification reaction comprises polymerase chain reaction.

73. The method of claim 71 or 72, wherein each of the primers in the pool comprises cytosine bases sensitive to bisulfite treatment, wherein the amplification is conducted in the presence of a mixture of dNTPs comprising dATP, dTTP, dGTP, and a dCTP analog resistant to bisulfite treatment, thereby generating the plurality of oligonucleotide sequence -polynucleotide complexes, wherein each of the plurality of oligonucleotide

IDT Exhibit 1020.3808

sequence -polynucleotide complexes comprises complementary ends, wherein one end comprises cytosine bases sensitive to bisulfite treatment, while the opposite end comprises cytosine analogs resistant to bisulfite treatment.

74. The method of claim 71 or 72, wherein each of the primers in the pool of primers comprises cytosine analogs resistant to bisulfite treatment, wherein the amplification is conducted in the presence of a mixture of dNTPs comprising dATP, dTTP, dGTP, and dCTP, thereby generating the plurality of oligonucleotide sequence-polynucleotide complexes, wherein each of the plurality of oligonucleotide sequence-polynucleotide complexes comprises complementary ends, wherein one end comprises cytosine bases sensitive to bisulfite treatment, while the opposite end comprises cytosine analogs resistant to bisulfite treatment.

75. The method of claims 68, 69, 73, or 74, wherein the cytosine analog resistant to bisulfite treatment is 5-methylcytosine, 5-hydroxymethylcytosine, or 5-propynylcytosine.

76. The method of claims 68, 69, 73, or 74, wherein the cytosine analog resistant to bisulfite treatment is 5-methylcytosine.

77. The method of any of claims 45-76, further comprising amplifying the plurality of amplifying the plurality of oligonucleotide sequence-polynucleotide complexes.

78. The method of claim 77, wherein the amplifying comprises solid-phase nucleic acid amplification to generate a plurality of clusters, wherein each cluster represents a plurality of copies of each of the plurality of oligonucleotide sequence-polynucleotide complexes.

79. The method of claim 77, further comprising treating the plurality of oligonucleotide sequence-polynucleotides complexes with bisulfite prior to the amplification, wherein the treating generates a plurality of oligonucleotide sequence-polynucleotide complexes comprising non-complementary ends.

80. The method of claim 79, wherein the sequence complementary to a sequencing primer in one end of the plurality of oligonucleotide sequence-polynucleotide complexes comprising non-complementary ends is complementary to a reverse read sequencing primer, and wherein the sequence complementary to a sequencing primer in an opposite end of the non-complementary ends in the plurality of oligonucleotide sequence-polynucleotide complexes comprising non-complementary ends is complementary to a forward read sequencing primer.

81. The method of any of claims 78-80, further comprising sequencing each of the plurality of clusters, wherein the sequencing comprises hybridizing the sequencing

IDT Exhibit 1020.3809

primer to the sequence complementary to a sequencing primer, and conducting sequencing by synthesis, wherein the introducing of the pool of oligonucleotide sequences to the polynucleotides improves identification of each of the plurality of clusters generated from the plurality of oligonucleotide sequence-polynucleotide complexes relative to the identification of each cluster from a plurality of clusters generated from the plurality of polynucleotides not introduced to the pool of oligonucleotide sequences.

82. The method of any of claims 45-81, wherein the sequence element comprises a barcode, a universal sequence, a linker sequence, or a random sequence.

83. A kit comprising a pool of oligonucleotide sequences,

a. wherein each oligonucleotide sequence in the pool of oligonucleotide sequences comprises a sequence complementary to a sequencing primer,

b. wherein the pool of oligonucleotide sequences comprises a first oligonucleotide sequence, wherein a terminal 3' nucleotide base of the sequencing primer is capable of annealing to a terminal 5' nucleotide base of the first oligonucleotide sequence,

c. wherein the pool of oligonucleotide sequences comprises a second oligonucleotide sequence, wherein the second oligonucleotide sequence terminates at a 5' end with an additional base relative to the 5' end of the first oligonucleotide sequence;

d. wherein the pool of oligonucleotide sequences comprises a third oligonucleotide sequence, wherein the third oligonucleotide sequence terminates at a 5' end with two additional bases relative to the 5' end of the first oligonucleotide sequence; and

e. wherein the pool of oligonucleotide sequences comprises a fourth oligonucleotide sequence, wherein the fourth oligonucleotide sequence terminates at a 5' end with three additional bases relative to the 5' end of the first oligonucleotide sequence.

84. The kit of claim 83, wherein each oligonucleotide sequence in the pool of oligonucleotide sequences is within a primer, wherein the pool of oligonucleotide sequences comprises the primer and a complement thereof

85. The kit of claim 84, wherein the primer comprises cytosine analogs resistant to bisulfite conversion.

86. The kit of claim 84, wherein the primer comprises cytosine bases sensitive to bisulfite conversion.

IDT Exhibit 1020.3810

87. The kit of claim 83, wherein each oligonucleotide sequence in the pool of oligonucleotide sequences is within a strand of a duplexed adaptor.

88. The kit of claim 87, wherein each duplexed adaptor in the pool comprises a ligation strand and a non-ligation strand.

89. The kit of claim 88, wherein the ligation strand and the non-ligation strand of each of duplexed adaptor in the pool comprise cytosine bases sensitive to bisulfite treatment.

90. The kit of claim 89, further comprising a mixture of dNTPs comprising dATP, dTTP, dGTP, and a dCTP analog resistant to bisulfite treatment.

91. The kit of claim 88, wherein the ligation strand and the non-ligation strand of each of duplexed adaptor in the pool comprise cytosine analogs resistant to bisulfite treatment.

92. The kit of claim 91, further comprising a mixture of dNTPs comprising dATP, dTTP, dGTP, and a dCTP analog sensitive to bisulfite treatment.

93. The kit of any of claims 83-92, further comprising a restriction enzyme.

94. The kit of claim 93, wherein the restriction enzyme is MspI.

95. The kit of any of claims 83-94, further comprising a DNA polymerase.

96. The kit of claim 95, wherein the DNA polymerase comprises exonuclease activity.

97. The kit of claim 95, wherein the DNA polymerase comprises strand displacement activity.

98. The kit of any of claims 88, 89, or 91, wherein the non-ligation strand comprises a block at a 5' and 3' end, whereby the non-ligation strand is enzymatically unreactive.

99. The kit of claim 98, wherein the block at the 5' end comprises a biotin moiety.

100. The kit of claim 98 or 99, wherein the block at the 3' end comprises a dideoxy residue.

101. The kit of claims 85, 90, or 91, wherein the cytosine analog resistant to bisulfite treatment is 5-methylcytosine, 5-hydroxymethylcytosine, or 5-propynylcytosine.

102. The kit of claims 85, 90, or 91, wherein the cytosine analog resistant to bisulfite treatment is 5-methylcytosine.

103. A method for analyzing sequence reads, the method comprising:

a. generating a plurality of sequence reads, wherein the generating comprises conducting sequencing by synthesis on a plurality of clusters, wherein each cluster of the plurality of clusters comprises copies of one oligonucleotide sequence-polynucleotide from a

IDT Exhibit 1020.3811

plurality of oligonucleotide sequence-polynucleotides, wherein the sequencing by synthesis comprises hybridizing a sequencing primer to sequence complementary to the sequencing primer present in each cluster of the plurality of clusters, wherein the plurality of clusters are produced from solid-phase nucleic acid amplification of the plurality of oligonucleotide sequence-polynucleotides, wherein each of the plurality of oligonucleotide sequence-polynucleotides comprises a single nucleotide base extender sequence, a dinucleotide base extender sequence, a trinucleotide base extender sequence or no nucleotide base extender sequence between the sequence complementary to the sequencing primer of an oligonucleotide sequence and an end of a polynucleotide; and

b. analyzing each of the plurality of sequence reads, wherein the analyzing comprises trimming or removing the extender sequence *in silico* from each of the plurality of sequence reads.

104. The method of claim 103, wherein the plurality of oligonucleotide sequence-polynucloetides are generated by introducing a pool of oligonucleotide sequences to 3' ends of the plurality of polynucleotides from a sample,

a. wherein each oligonucleotide sequence in the pool of oligonucleotide sequences comprises a sequence complementary to a sequencing primer,

b. wherein the pool of oligonucleotide sequences comprises a first sequence, wherein a terminal 3' nucleotide base of the sequencing primer is capable of annealing to a terminal 5' nucleotide base of the first sequence, wherein following introduction the 5' end of the first sequence is adjacent to a 3' end of a polynucleotide from the plurality of polynucleotides,

c. wherein the pool of oligonucleotide sequences comprises a second sequence, wherein the second sequence terminates at its 5' end with an additional base relative to the 5' end of the first sequence comprising the single nucleotide base extender, wherein following introduction the 5' end of the second sequence is adjacent to a 3' end of a polynucleotide from the plurality of polynucleotides;

d. wherein the pool of oligonucleotide sequences comprises a third sequence, wherein the third sequence terminates at its 5' end with two additional bases relative to the 5' end of the first sequence comprising the dinucleotide base extender, wherein following introduction the 5' end of the third sequence is adjacent to a 3' end of a polynucleotide from the plurality of polynucleotides;

e. wherein the pool of oligonucleotide sequences comprises a fourth sequence, wherein the fourth sequence terminates at its 5' end with three additional bases

IDT Exhibit 1020.3812

relative to the 5' end of the first sequence comprising the trinucleotide base extender, wherein following introduction the 5' end of the fourth sequence is adjacent to a 3' end of a polynucleotide from the plurality of polynucleotides; and

f. wherein the introducing generates a plurality of oligonucleotide-polynucleotide complexes, thereby increasing diversity of sequences at ends of the plurality of polynucleotides.

105. The method of claim 104, wherein the sample is a whole genome sample.

106. The method of claim 104, wherein the sample comprises a polynucleotide library.

107. The method of any of claims 104-106, further comprising fragmenting the polynucleotides prior to introducing the pool of oligonucleotide sequences to the 3' ends of the plurality of polynucleotides.

108. The method of claim 107, wherein the fragmenting is performed by contacting the polynucleotides with an enzyme.

109. The method of claim 108, wherein the enzyme is a restriction enzyme.

110. The method of claim 109, wherein the restriction enzyme is a methylation insensitive restriction enzyme.

111. The method of claim 110, wherein the methylation insensitive restriction enzyme is MspI.

112. The method of claim 111, wherein the additional base of the second sequence is an adenine, thymine, or cytosine base.

113. The method of claim 111 or 112, wherein each of the two additional bases of the third sequence is an adenine, thymine, or cytosine base.

114. The method of any of claims 111-113 wherein a 3' most base of the three additional bases of the fourth sequence comprises a thymine or cytosine base, and each of the two other bases of the three additional bases of the fourth sequence comprises an adenine, thymine, or cytosine base.

115. The method of any of claims 104-114, wherein each oligonucleotide sequence in the pool of oligonucleotide sequences is within a strand of a duplexed adaptor.

116. The method of claim 115, wherein the duplexed adaptor is a partial duplexed adaptor comprising a long strand and a short strand.

117. The method of claim 116, wherein the short strand is a non-ligation strand of the duplexed adaptor and the long strand is a ligation strand.

118. The method of claim 117, wherein the non-ligation strand comprises a 5' overhang.

IDT Exhibit 1020.3813

119. The method of claim 118, wherein the 5' overhang comprises two bases.

120. The method of claim 119, wherein the two bases are 3'-GC-5'.

121. The method of any of claims 104-120, wherein the introducing comprises a ligation reaction.

122. The method of claim 121, wherein the ligation reaction comprises cohesive end ligation.

123. The method of claim 121, wherein the ligation reaction comprises blunt end ligation.

124. The method of claim 122, wherein the plurality of polynucleotides are strands of double-stranded nucleic acids, wherein each 5' end of the double-stranded nucleic acids comprises a ligation end and each 3' end comprises a non-ligation end, wherein each of the non-ligation ends of the double-stranded nucleic acid comprises part of the polynucleotide sequence.

125. The method of claim 124, wherein the 3'-GC-5' of the non-ligation strand of the duplexed adaptor anneals to a 5'-CG-3' sequence of the ligation end of the double-stranded sequence comprising the polynucleotide sequence.

126. The method of claim 125, wherein the 3'-GC-5' of the non-ligation strand of the duplexed adaptor is 5' to the terminal 5' nucleotide base of the first sequence,

127. The method of claim 125 or 126, wherein the 3'-GC-5' of the non-ligation strand of the duplexed adaptor is 5' to the additional base of the second sequence relative to the 5' end of the first sequence.

128. The method of any of claims 125-127, wherein the 3'-GC-5' of the non-ligation strand of the duplexed adaptor is 5' to the two additional bases of the third sequence relative to the 5' end of the first sequence.

129. The method of any of claims 125-128, wherein the 3'-GC-5' of the non-ligation strand of the duplexed adaptor is 5' to the three additional bases of the fourth sequence relative to the 5' end of the first sequence.

130. The method of any of claims 118-122, wherein each duplexed adaptor in the pool comprises a ligation strand and a non-ligation strand, wherein the ligation strand is capable of ligating to a 5' end of the plurality of polynucleotides, and wherein a nick is formed between the non-ligation strand and the 3' end of the plurality of polynucleotides.

131. The method of claim 130, further comprising an extending step, wherein the 3' end of the plurality of polynucleotides is extended using a polymerase in the presence of a mixture of dNTPs using the ligation strand of the adaptor as template, thereby

IDT Exhibit 1020.3814

introducing the pool of oligonucleotide sequences to the 3' ends of the plurality of polynucleotides.

132. The method of claim 131, wherein the ligation strand of each adaptor duplex in the pool comprise cytosine analogs resistant to bisulfite treatment and the mixture of dNTPs comprises dATP, dTTP, dGTP, and dCTP, whereby the extending generates the plurality of oligonucleotide-polynucleotide complexes, wherein each of the plurality of oligonucleotide-polynucleotide complexes comprises complementary ends, wherein one end comprises cytosine bases sensitive to bisulfite treatment, while the opposite end comprises cytosine analogs resistant to bisulfite treatment.

133. The method of claim 131, wherein the ligation strand of each adaptor duplex in the pool of adaptor duplexes comprise cytosine bases sensitive to bisulfite treatment and the mixture of dNTPs comprises dATP, dTTP, dGTP, and a dCTP analog resistant to bisulfite treatment, whereby the extending generates the plurality of oligonucleotide-polynucleotide complexes, wherein each of the plurality of oligonucleotide-polynucleotide complexes comprises complementary ends, wherein one end comprises cytosine bases sensitive to bisulfite treatment, while the opposite end comprises cytosine analogs resistant to bisulfite treatment.

134. The method of any of claims 104-106, wherein each oligonucleotide in the pool of oligonucleotides is a primer, wherein the pool of oligonucleotides comprises the primer and a complement thereof.

135. The method of claim 134, wherein the introducing comprises an amplification reaction.

136. The method of claim 135, wherein the amplification reaction comprises polymerase chain reaction.

137. The method of claim 135 or 136, wherein each of the primers in the pool comprises cytosine bases sensitive to bisulfite treatment, wherein the amplification is conducted in the presence of a mixture of dNTPs comprising dATP, dTTP, dGTP, and a dCTP analog resistant to bisulfite treatment, thereby generating the plurality of oligonucleotide sequence-polynucleotide complexes, wherein each of the plurality of oligonucleotide sequence-polynucleotide complexes comprises complementary ends, wherein one end comprises cytosine bases sensitive to bisulfite treatment, while the opposite end comprises cytosine analogs resistant to bisulfite treatment.

138. The method of claim 135 or 136, wherein each of the primers in the pool of primers comprises cytosine analogs resistant to bisulfite treatment, wherein the

IDT Exhibit 1020.3815

amplification is conducted in the presence of a mixture of dNTPs comprising dATP, dTTP, dGTP, and dCTP, thereby generating the plurality of oligonucleotide sequence-polynucleotide complexes, wherein each of the plurality of oligonucleotide sequence-polynucleotide complexes comprises complementary ends, wherein one end comprises cytosine bases sensitive to bisulfite treatment, while the opposite end comprises cytosine analogs resistant to bisulfite treatment.

139. The method of claims 132, 133, 137, or 138, wherein the cytosine analog resistant to bisulfite treatment is 5-methylcytosine, 5-hydroxymethylcytosine, or 5-propynylcytosine.

140. The method of claim 137-139, further comprising treating the plurality of oligonucleotide sequence-polynucleotides complexes with bisulfite prior to the amplification, wherein the treating generates a plurality of oligonucleotide sequence-polynucleotide complexes comprising non-complementary ends.

141. The method of claim 140, wherein the treating with bisulfite prior to amplification generates the sequence complementary to a sequencing primer in each oligonucleotide sequence in the pool of oligonucleotide sequences comprises.

142. The method of any of claims 103-141, wherein the generating the plurality of sequence reads comprises paired-end sequencing.

143. A method for increasing diversity of sequences at ends of a plurality of polynucleotide fragments in a library, the method comprising:

a. fragmenting polynucleotides to generate polynucleotide fragments;

b. ligating a pool of adaptor sequences to 3' ends of the plurality of polynucleotide fragments, wherein each adaptor sequence in the pool of adaptor sequences comprises a sequence complementary to a sequencing primer, wherein a 3' end of the sequencing primer anneals at least four bases from the 5' end of the adaptor sequence, and wherein each of four different types of nucleotide bases are present in each of the at least three terminal nucleotides at the 5' ends of the adaptors, thereby increase diversity of sequences at ends of the plurality of polynucleotide fragments.

IDT Exhibit 1020.3816

FIG. 1

unknown methylation status
Non-methylated
Methylated
digest with MspI
End repair
Ligate adaptor, bisulfite convert, PCR amplify
sequence
5'Forward (C or T)GG - insert TTG Reverse 3' (top strand shown)
Forward Read: 5' (C or T)GG (bisulfite converted insert C-to-T strand)
Reverse Read: 5' CAA (bisulfite converted insert G-to-A strand)

FIG. 2A

```
Ben383    /5Biosg/CGHHHAGATCGGAAGAG/3ddC/     (H = A,C, or T, (no G))
Ben384     /5Biosg/CGHHAGATCGGAAGAG/3ddC/
Ben385      /5Biosg/CGHAGATCGGAAGAG/3ddC/
Ben341       /5Biosg/CGAGATCGGAAGAG/3ddC/

R01446    CAAGCAGAAGACGGCATACGTGATCTGGTTGTGACTGGAGTTCTGACGTGTGCTCTTCCGATCT
```

FIG. 2B

1. Anneal

```
                    R01446
                      ↓
5'-CAAGCAGAAGACGGCATACGTGATCTGGTTGTGACTGGAGTTCTGACGTGTGCTCTTCCGATCT
                                                    ddCGAAGAGCTAGAAHHHCG/Biosg5
                                                      ↑
                                                   Ben383
```

2. Extend with D mix (dATP, dGTP, dTTP (no dCTP))

```
                    R01446                    (D = T,G, or A, (no C))
                      ↓
5'-CAAGCAGAAGACGGCATACGTGATCTGGTTGTGACTGGAGTTCTGACGTGTGCTCTTCCGATCTDDD
                                                    ddCGAAGAGCTAGAAHHHCG/Biosg5
                                                      ↑
                                                   Ben383
```

FIG. 3

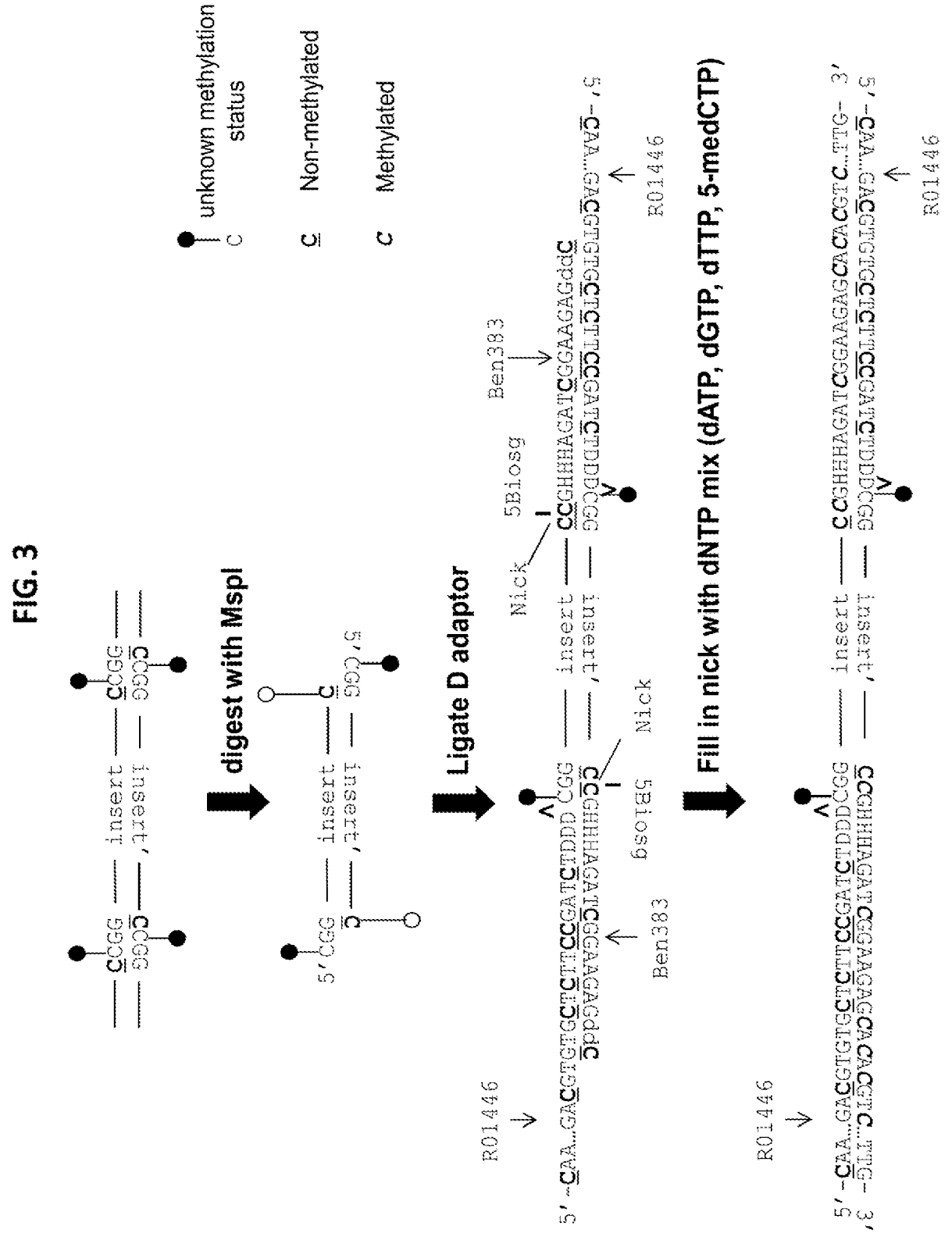

unknown methylation status
Non-methylated
Methylated
digest with MspI
Ligate D adaptor
Fill in nick with dNTP mix (dATP, dGTP, dTTP, 5-medCTP)
Nick
5Biosg
Ben383
R01446
insert

FIG. 3 (continued)

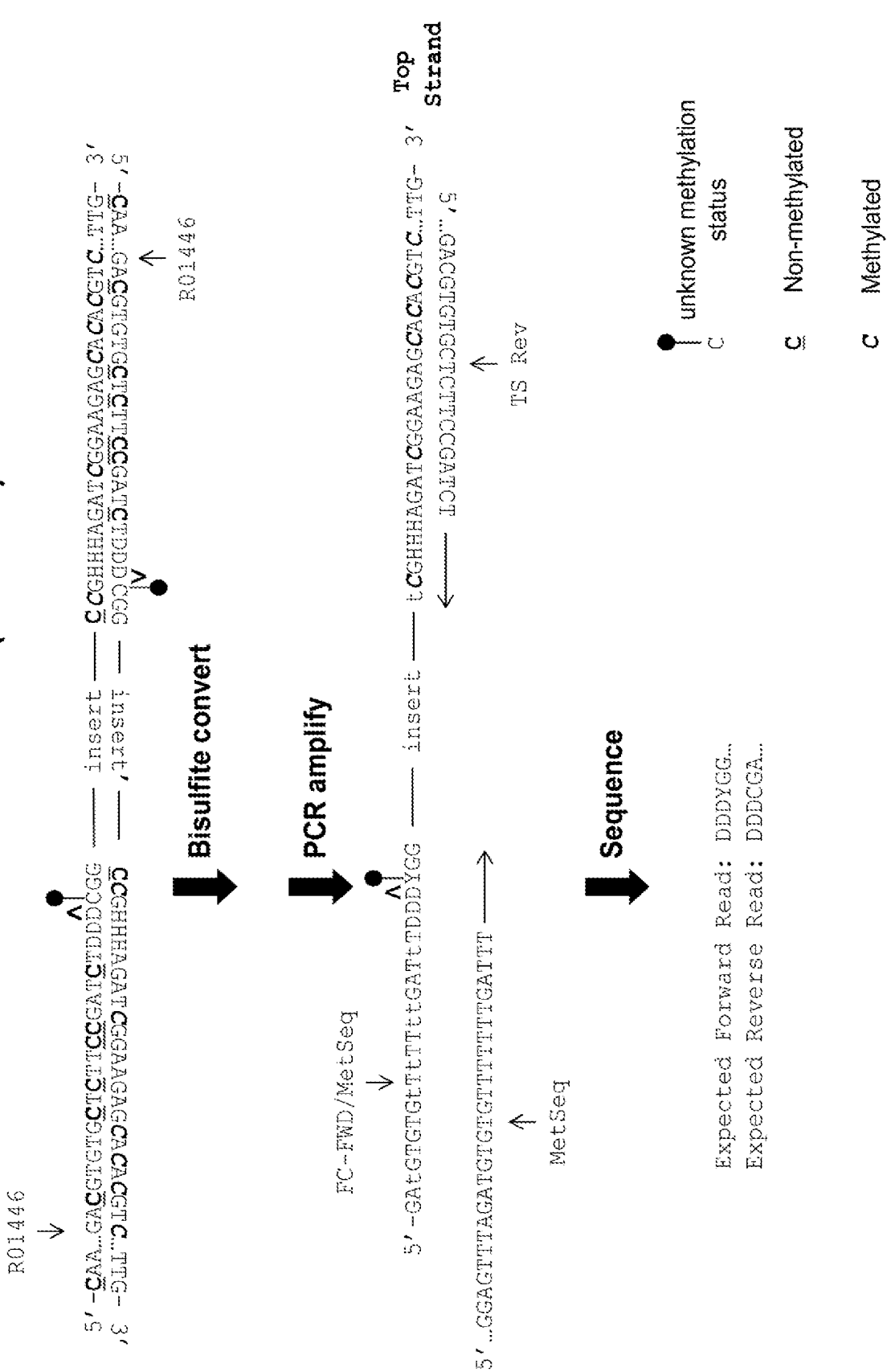


FIG. 4A

DR Rev Helper Style Ed3 BC

R01446 Ben383 D3 1

Expected Forward Read:

DDDYGG...

H = A,C, or T

D = A,G, or T

Y = C or T

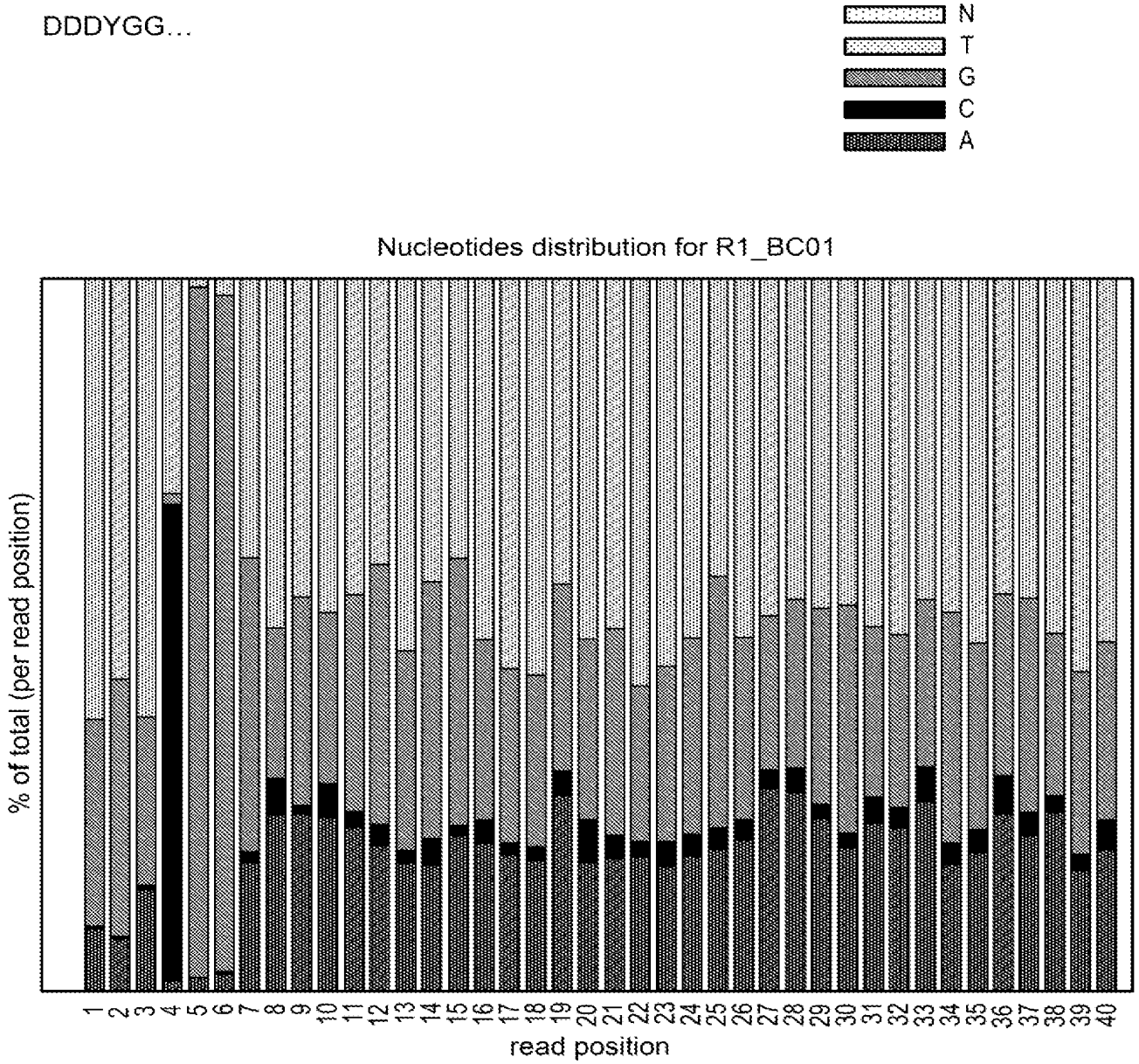


FIG. 4B

DR Rev Helper Style Ed3 BC

R01446 Ben383 D3 1

Expected Reverse Read:

DDDCGA…

H = A,C, or T

D = A,G, or T

Y = C or T

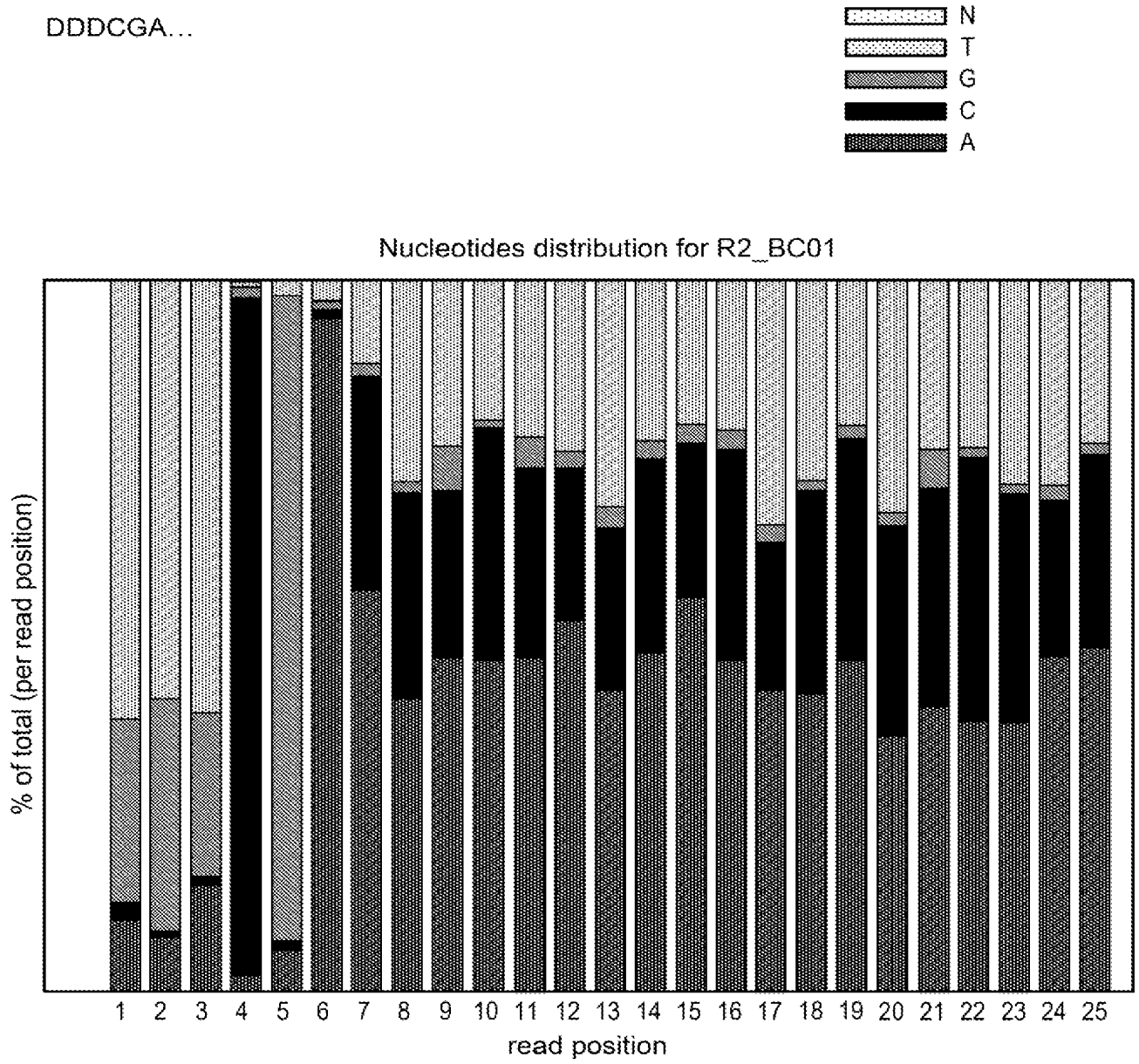


FIG. 5A

DR Rev Helper Style Ed3 BC

R01447 Ben384 D2 2

Expected Forward Read:

DDYGG...

H = A,C, or T

D = A,G, or T

Y = C or T

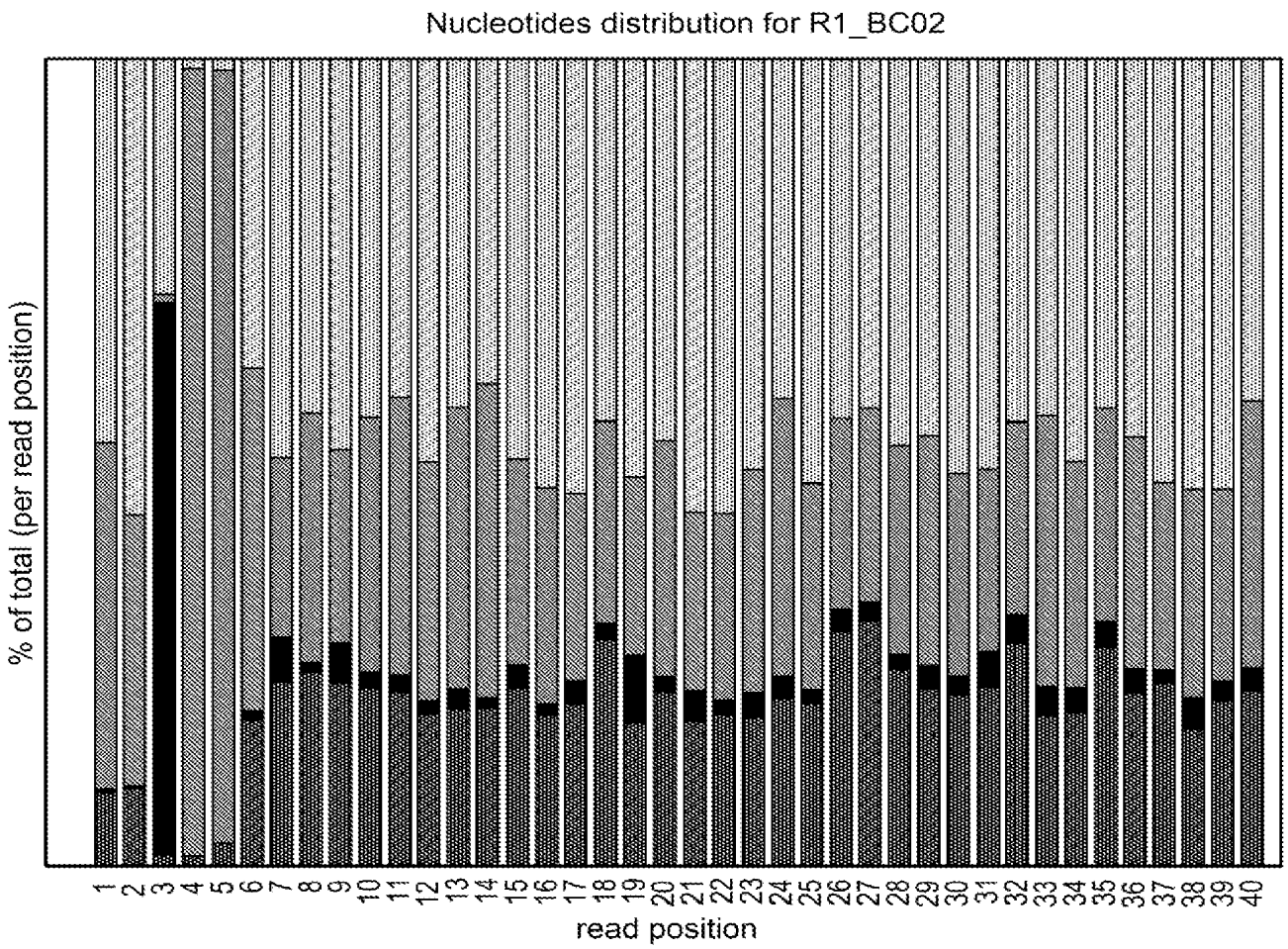


FIG. 5B

DR Rev Helper Style Ed3 BC

R01447 Ben384 D2 2

Expected Reverse Read:

DDCGA...

H = A,C, or T

D = A,G, or T

Y = C or T

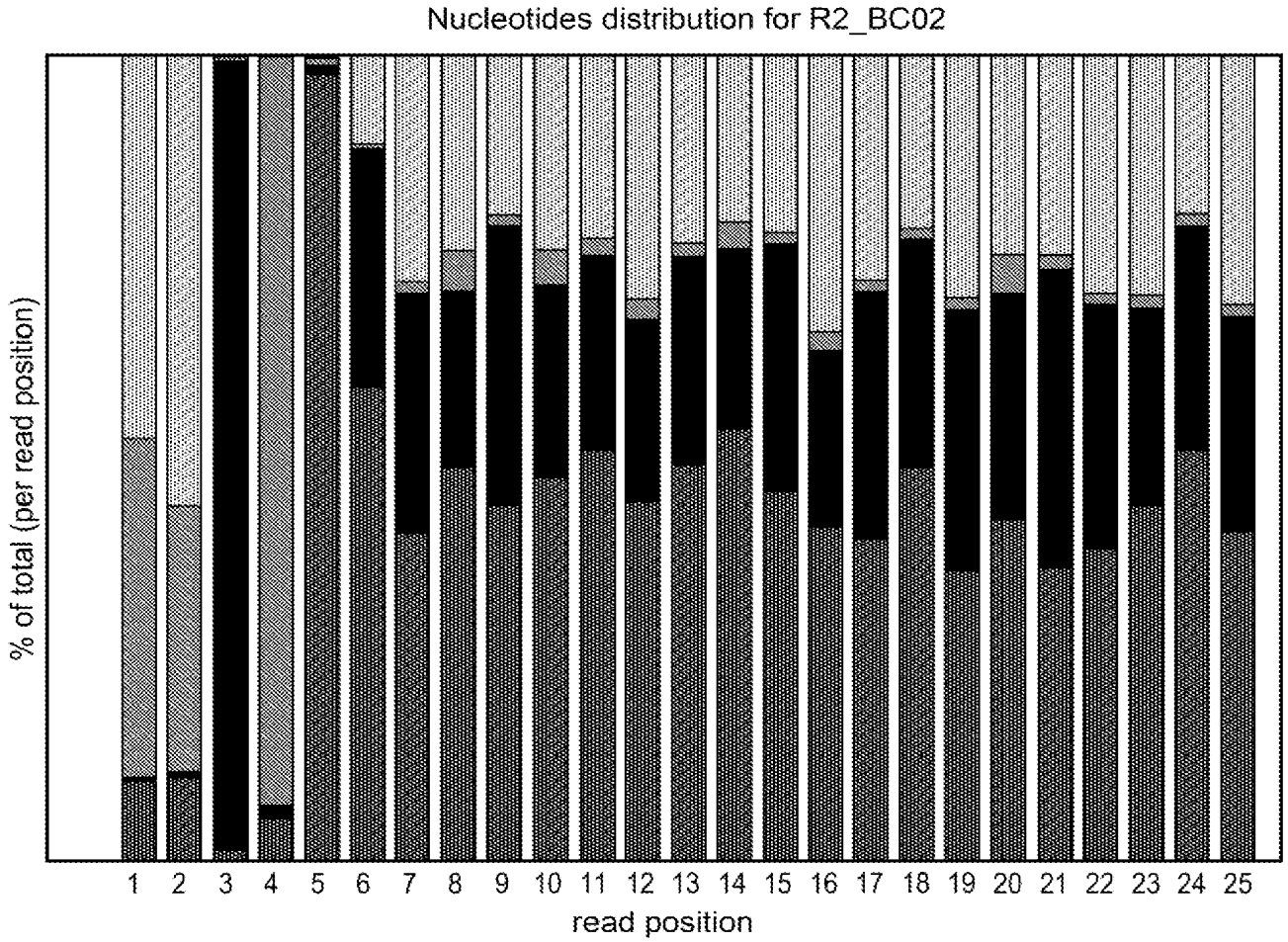


FIG. 6A

DR Rev Helper Style Ed3 BC

R01448 Ben385 D1 3

Expected Forward Read:

DYGG...

H = A,C, or T

D = A,G, or T

Y = C or T

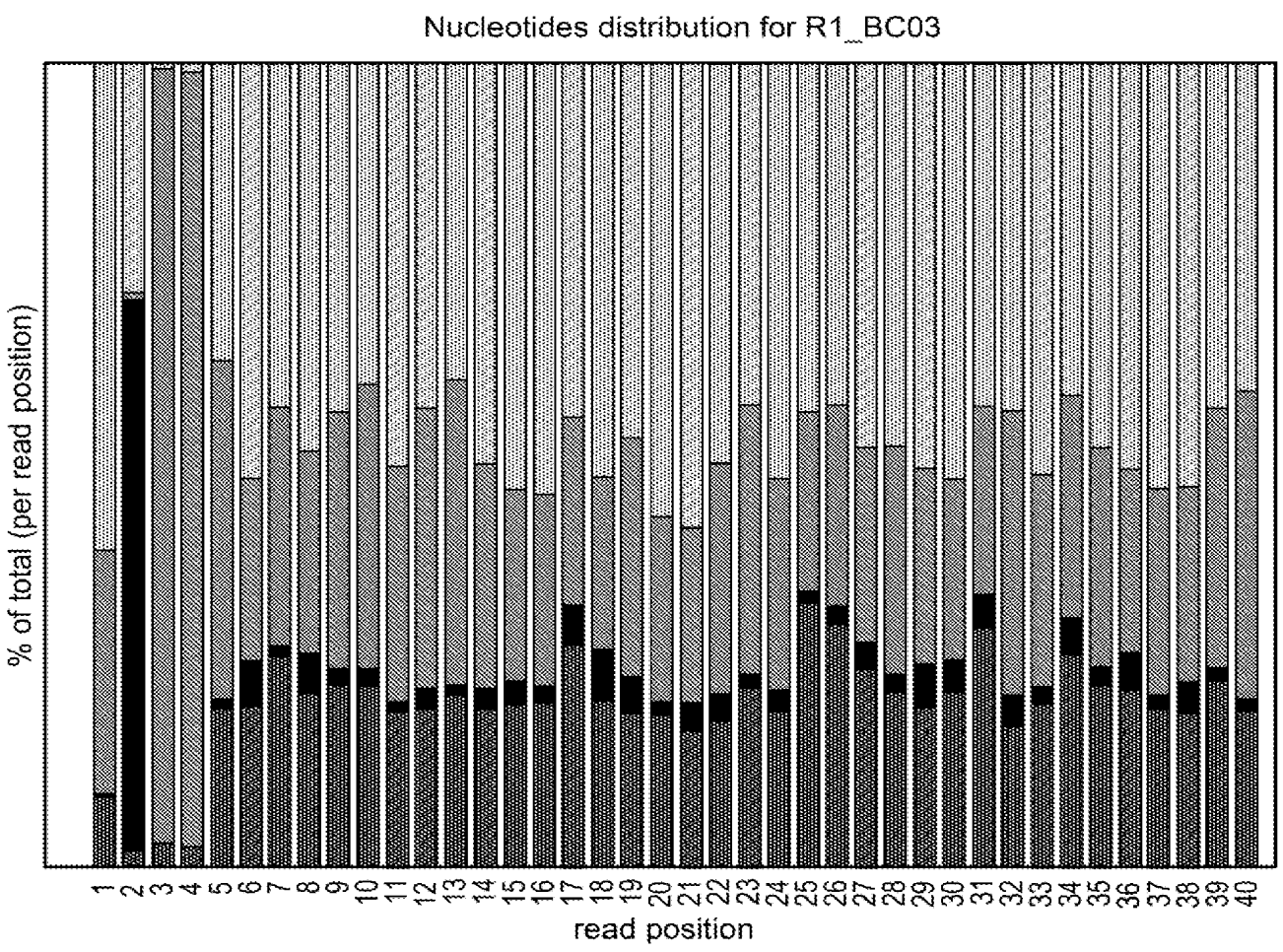


FIG. 6B

DR Rev Helper Style Ed3 BC

R01448 Ben385 D1 3

Expected Reverse Read:

DCGA...

H = A,C, or T

D = A,G, or T

Y = C or T

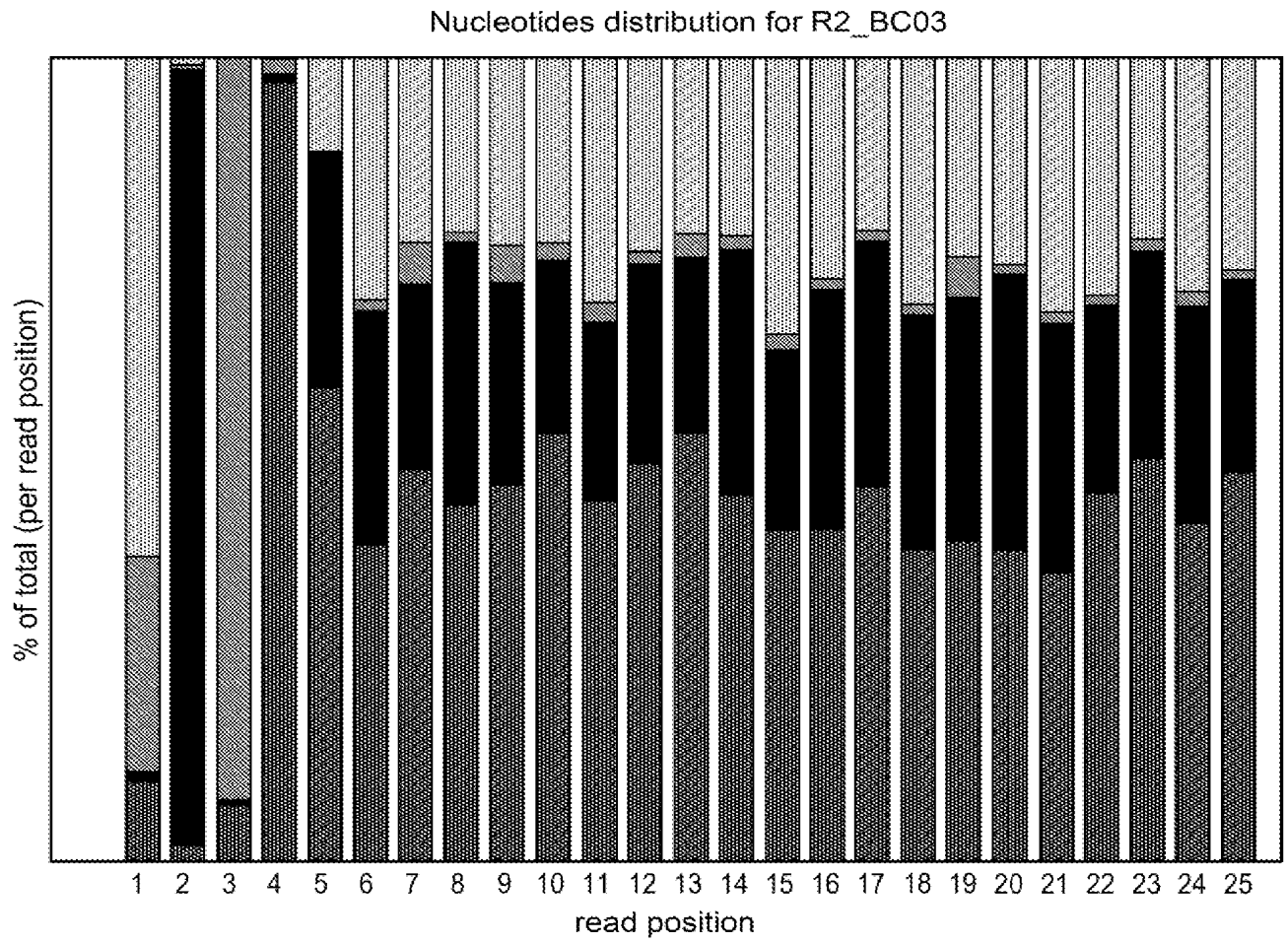


FIG. 7A

DR Rev Helper Style Ed3 BC

R01449 Ben341 D0 1

Expected Forward Read:

YGG...

H = A,C, or T

D = A,G, or T

Y = C or T

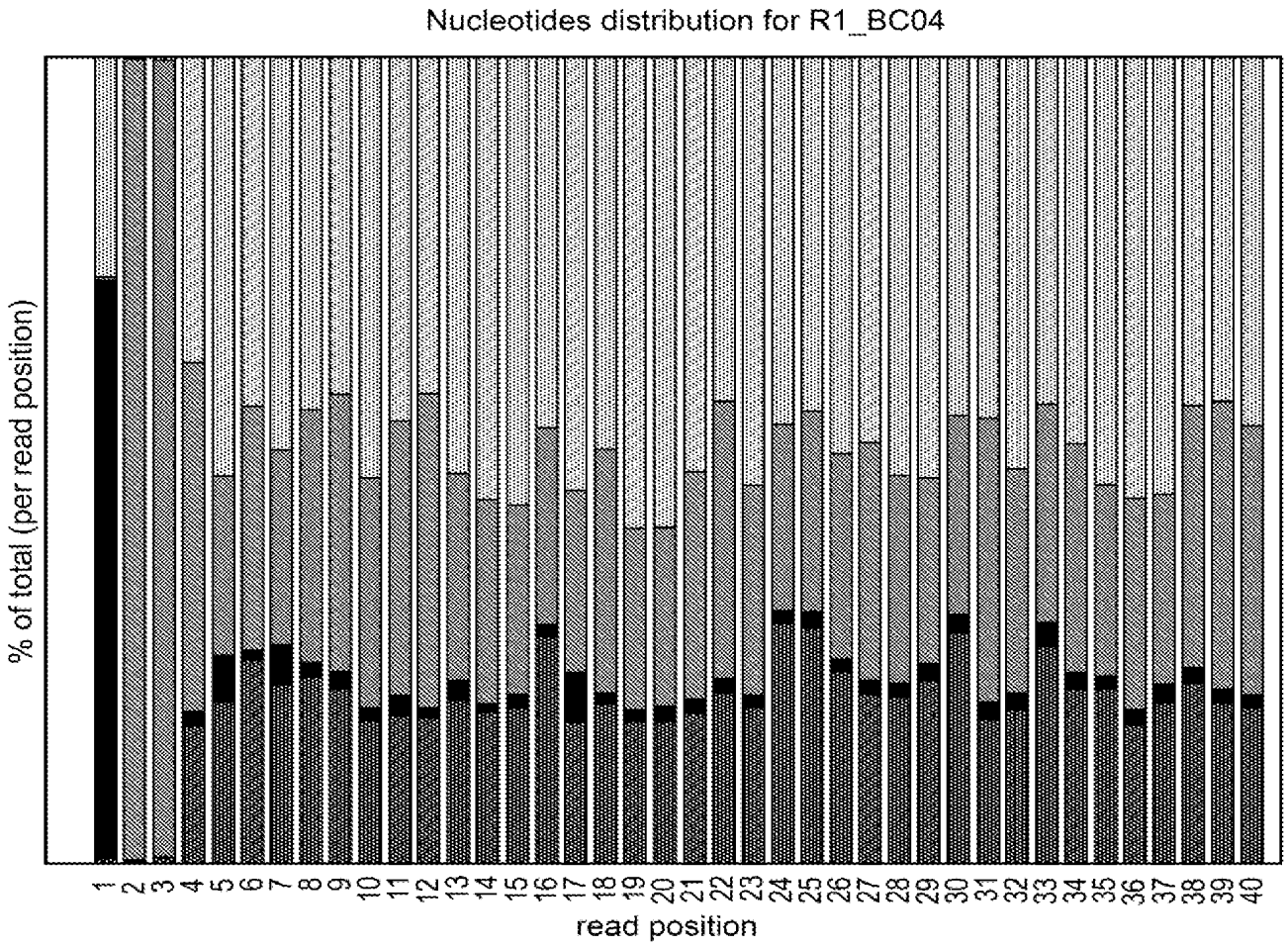


FIG. 7B

DR Rev Helper Style Ed3 BC

R01449 Ben341 D0 4

Expected Reverse Read:

CGA...

H = A,C, or T

D = A,G, or T

Y = C or T

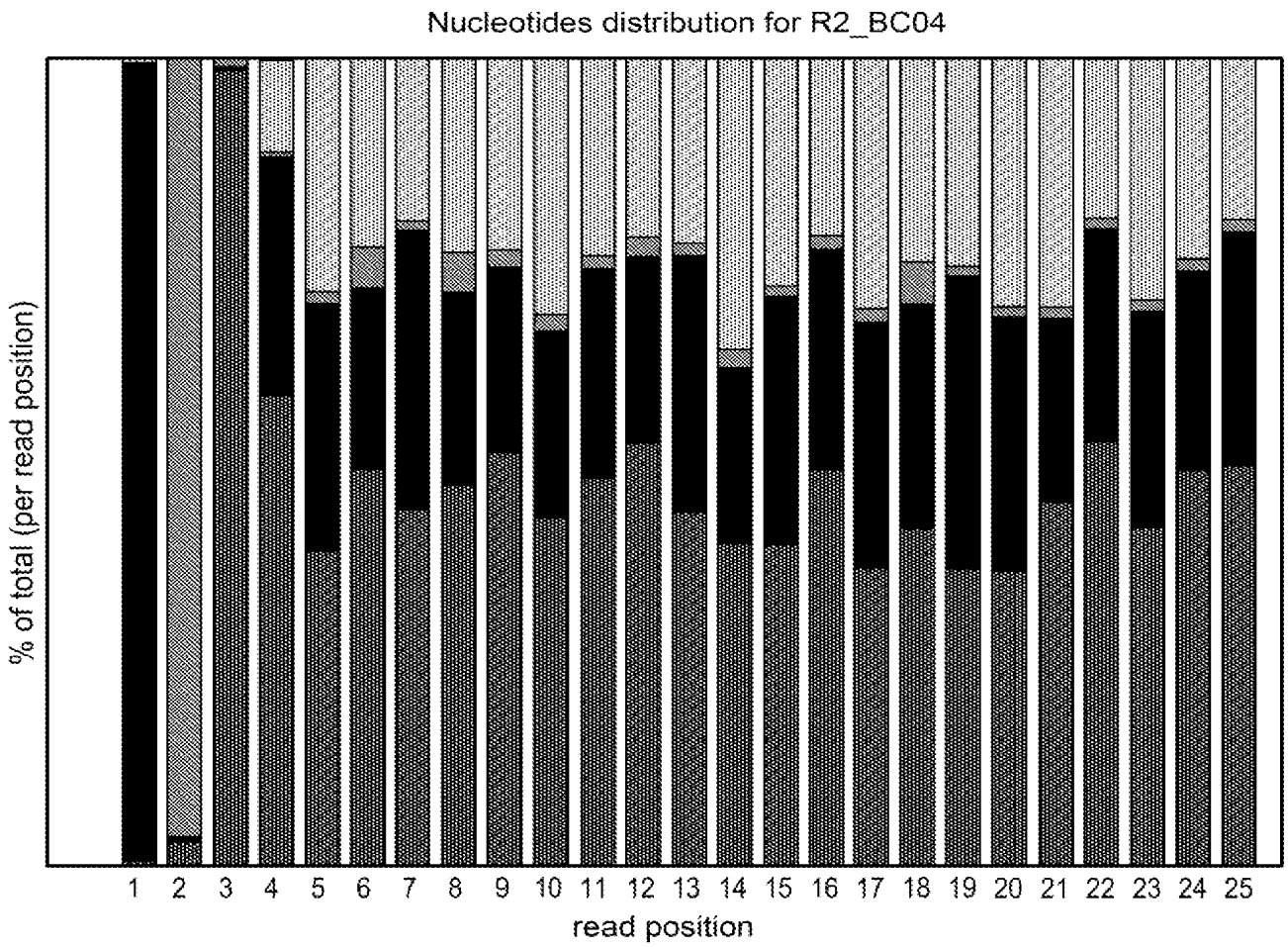


FIG. 8

DR Rev Helper Style Ed3 BC
R01450 mix4 MIX 5

H = A,C, or T
D = A,G, or T
Y = C or T
B = C,G, or T

Expected Forward Read
sum of:
YGG…
DYGG…
DDYGG…
DDDYGG…
NNNNNN…

Expected Reverse Read
sum of:
CGA…
DCGA…
DDCGA…
DDDCGA…
NNNNNN…

Expected Forward Read
sum of:
YGG…
DYGG…
DDYGG… (if don't have DDDYGG)
NNBNN… (missing A at position 3)

IDT Exhibit 1020.3829

FIG. 9A

DR Rev Helper Style Ed3 BC

R01450 mix4 MIX 5

Expected Forward Read:

NNNNNN...

H = A,C, or T

D = A,G, or T

Y = C or T

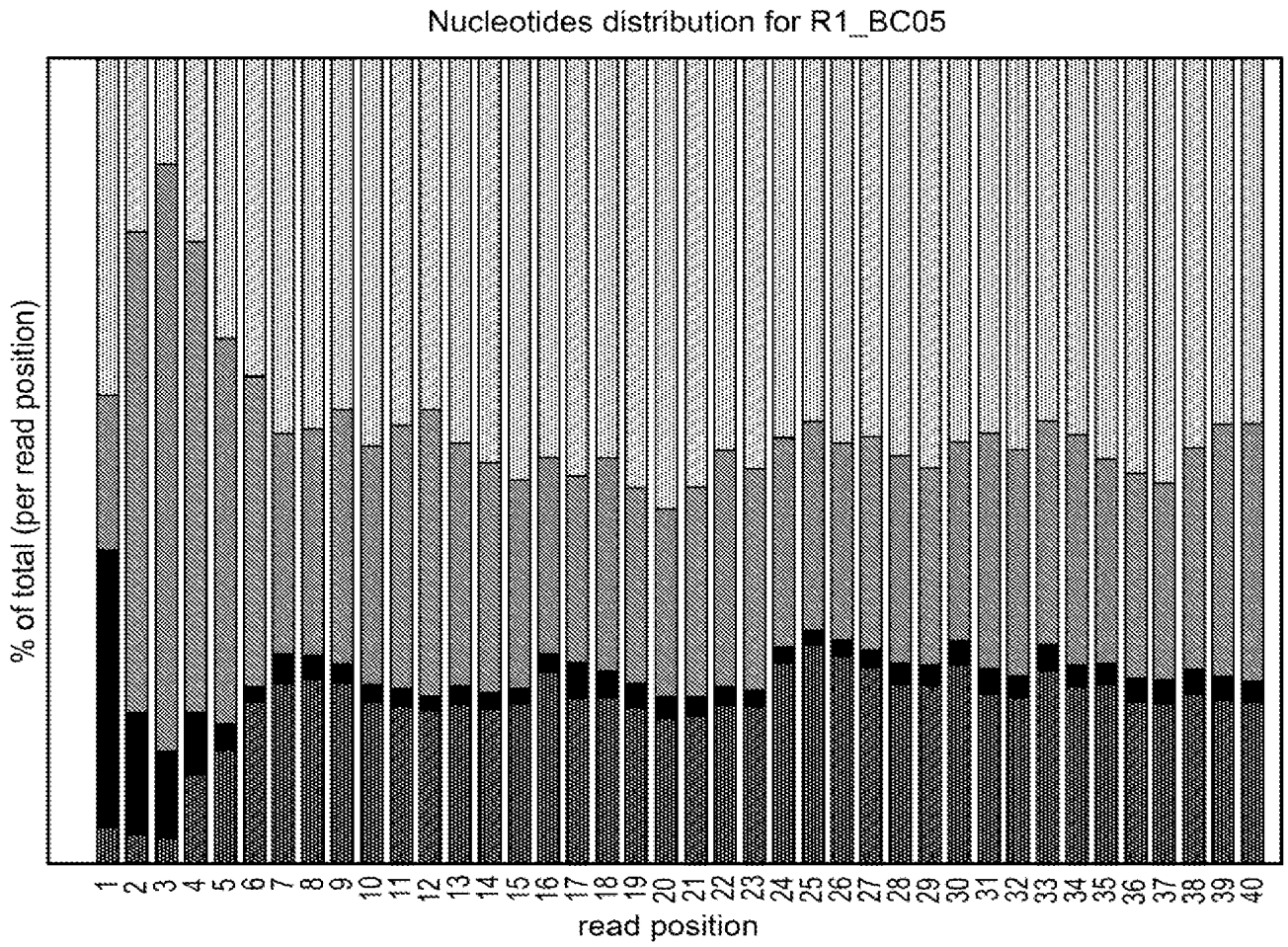


FIG. 9B

DR Rev Helper Style Ed3 BC

R01450 MIX4 5

Expected Reverse Read

NNNNNN...

H = A,C, or T

D = A,G, or T

Y = C or T

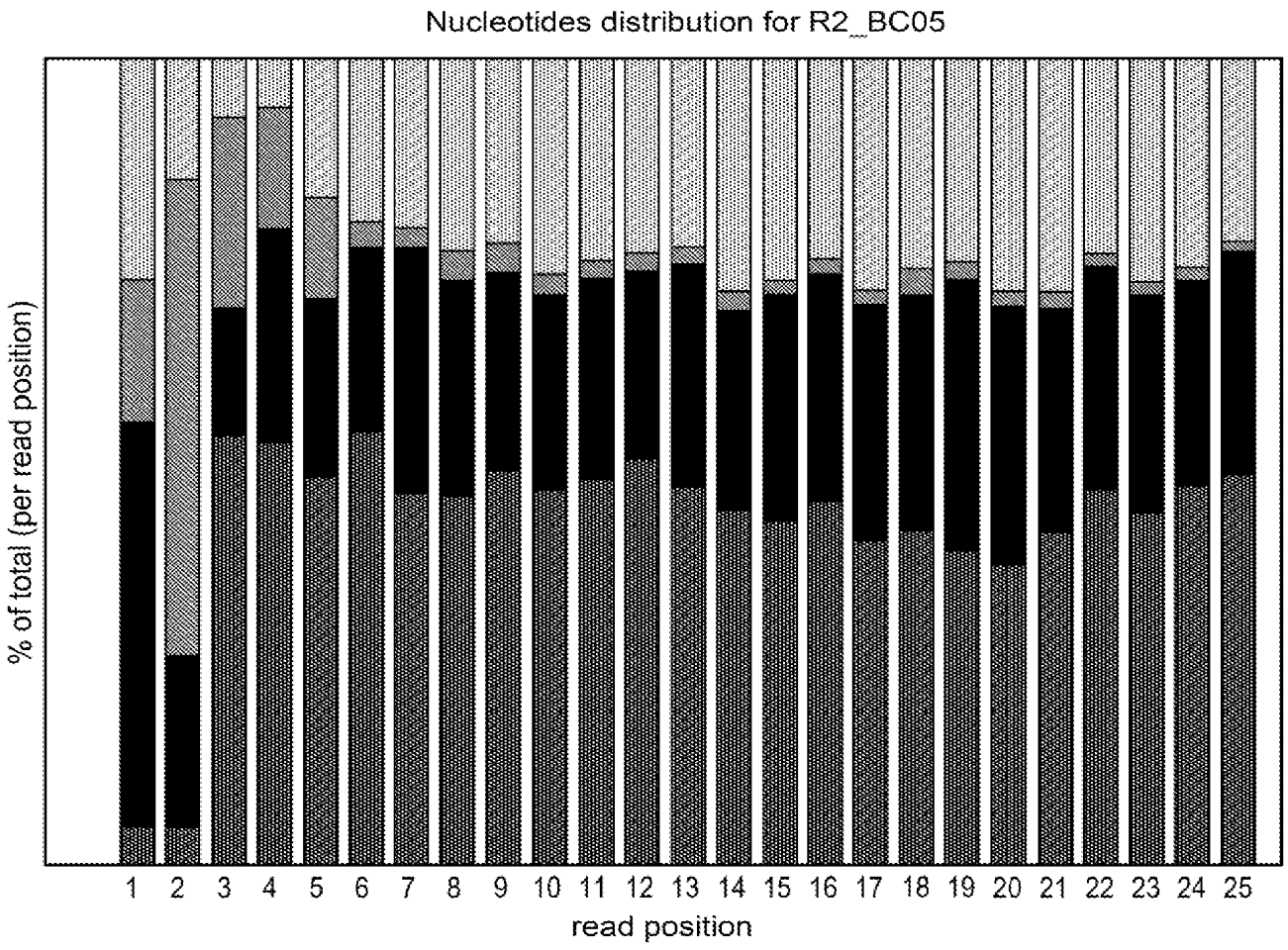


FIG. 10

40+6+25 run

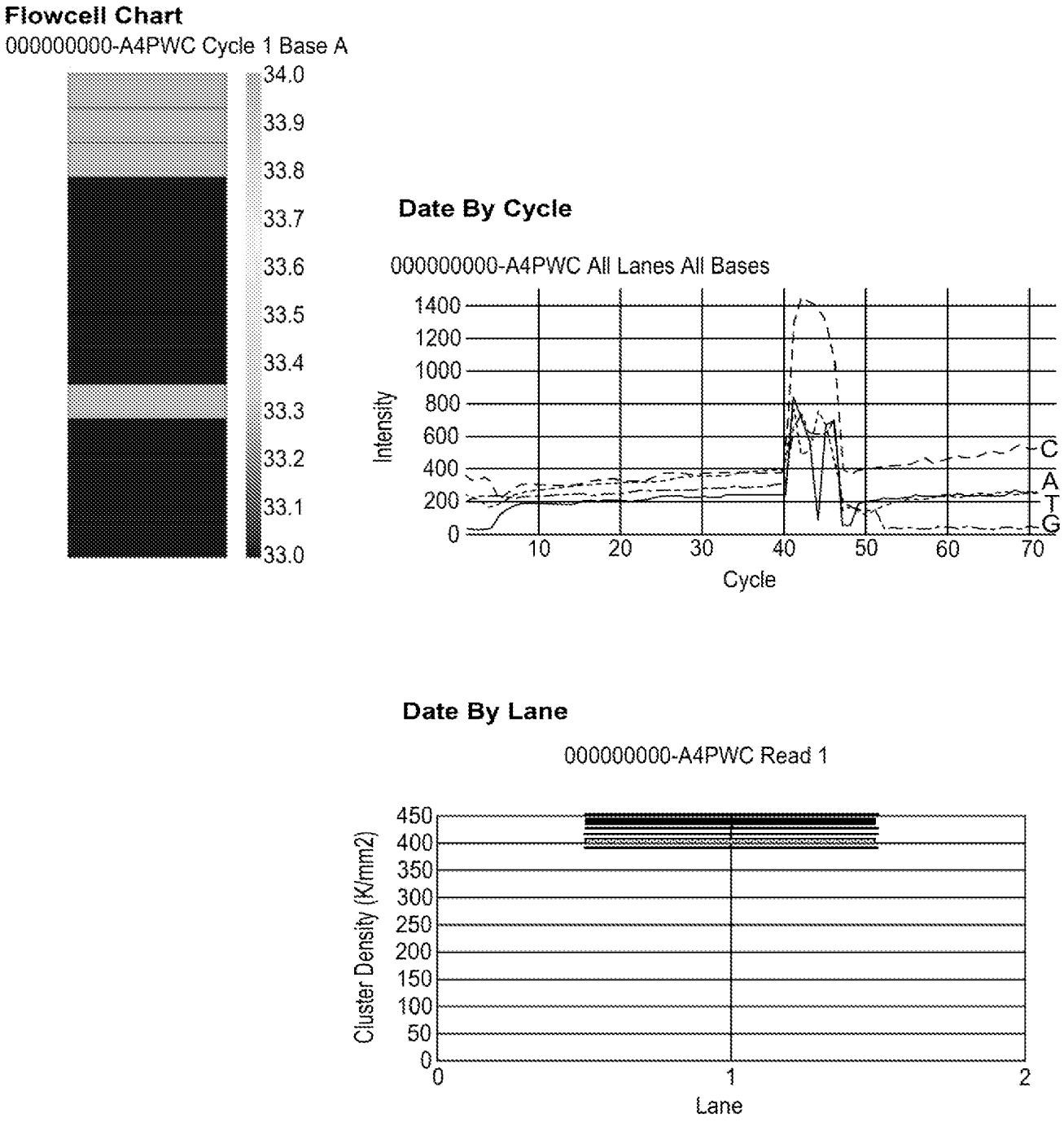
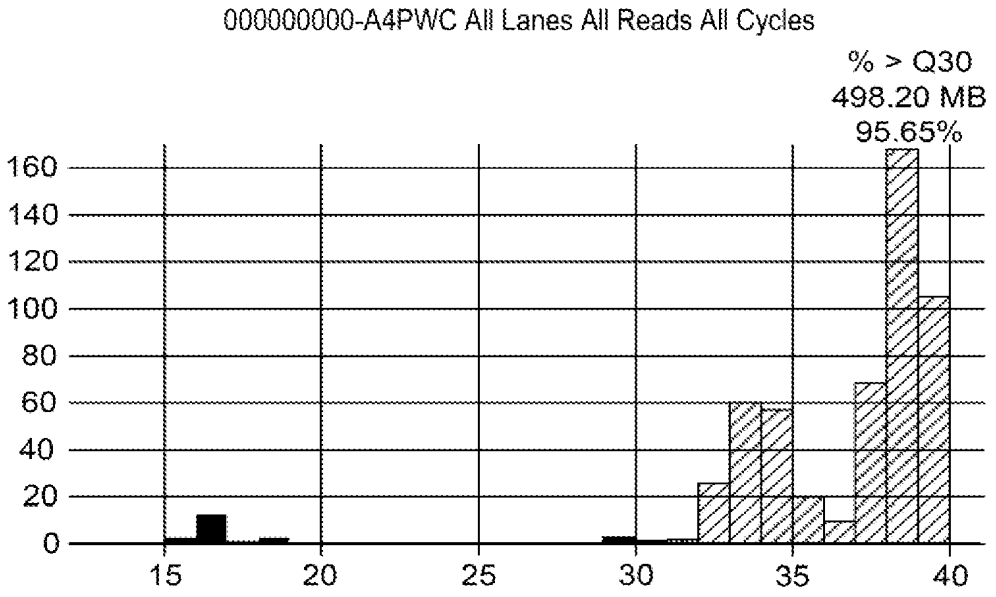


FIG. 10
(Continued)

QScore Distribution



QScore Heatmap

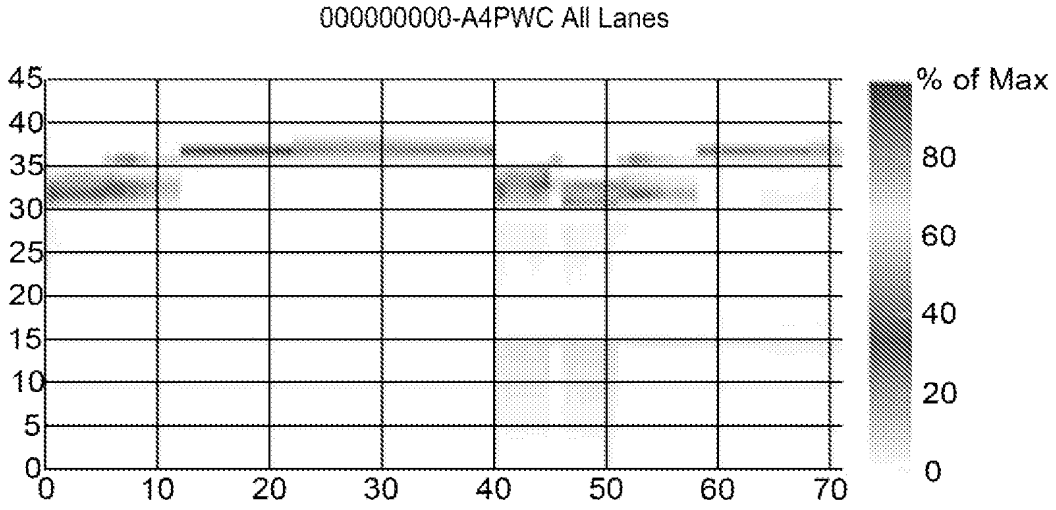


FIG. 11

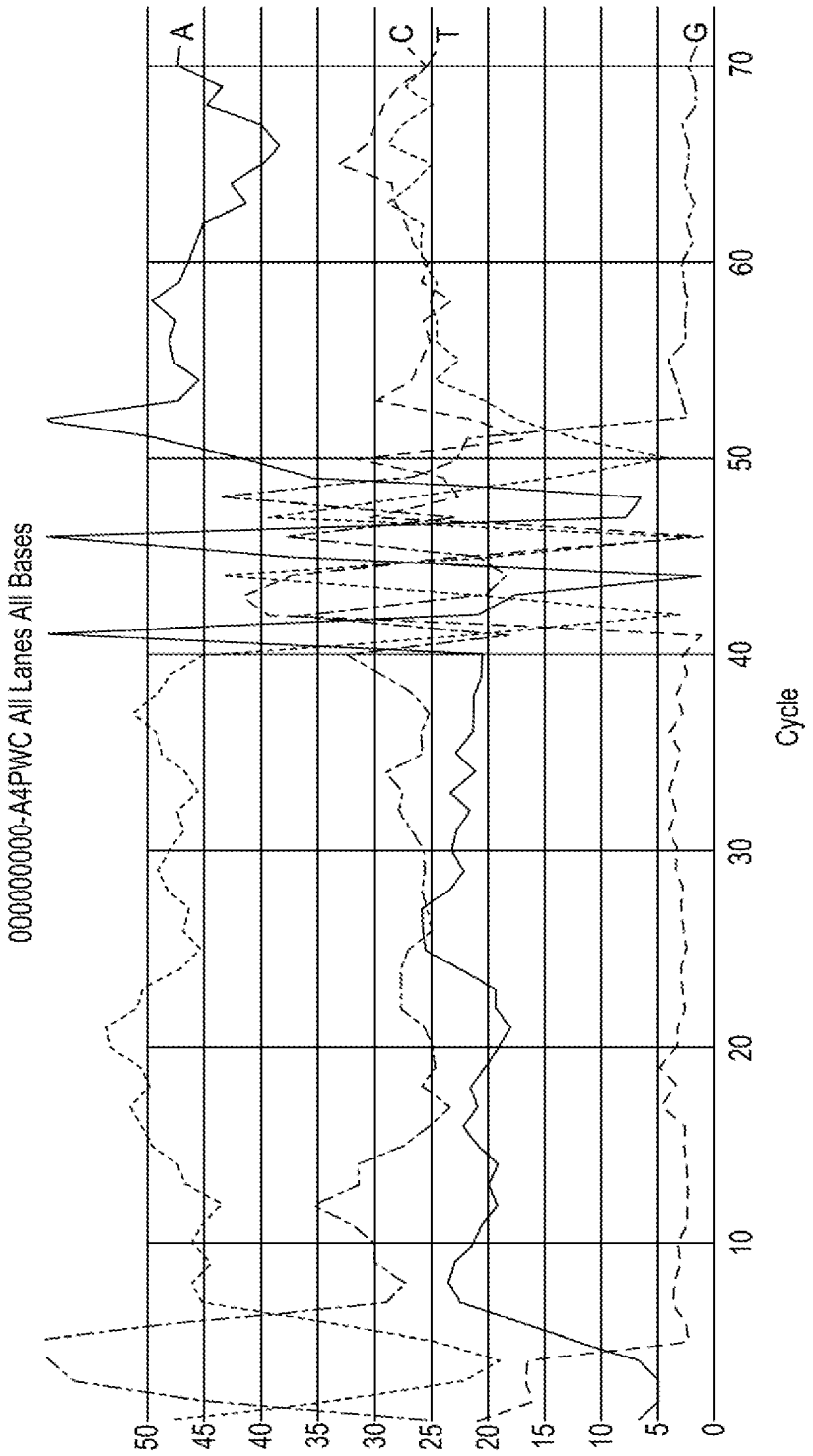

000000000-A4PWC All Lanes All Bases
A
C
T
G
Cycle
0
5
10
15
20
25
30
35
40
45
50
10
20
30
40
50
60
70

FIG. 12

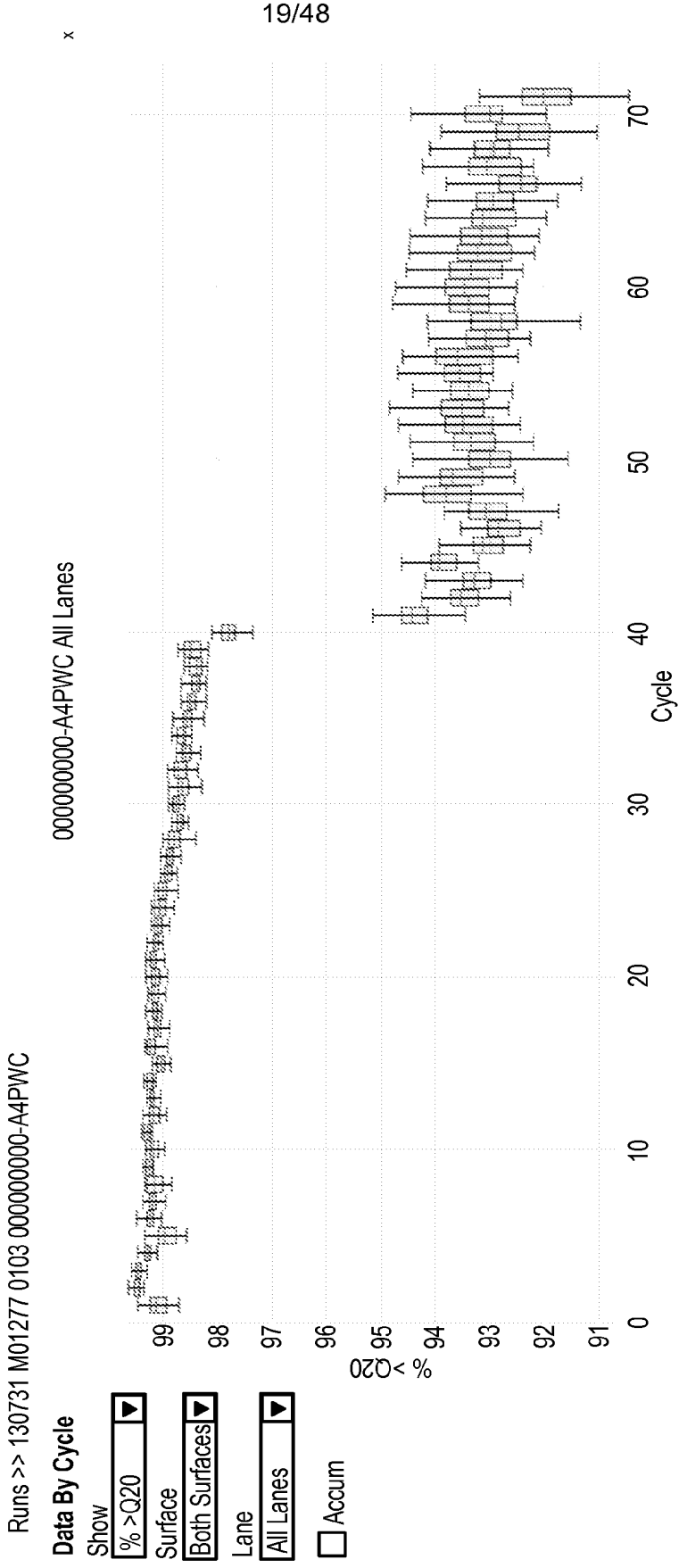

Runs >> 130731 M01277 0103 000000000-A4PWC
Data By Cycle
Show
% >Q20
Surface
Both Surfaces
Lane
All Lanes
Accum
000000000-A4PWC All Lanes
% >Q20
99
98
97
96
95
94
93
92
91
0
10
20
30
40
50
60
70
Cycle

FIG. 13

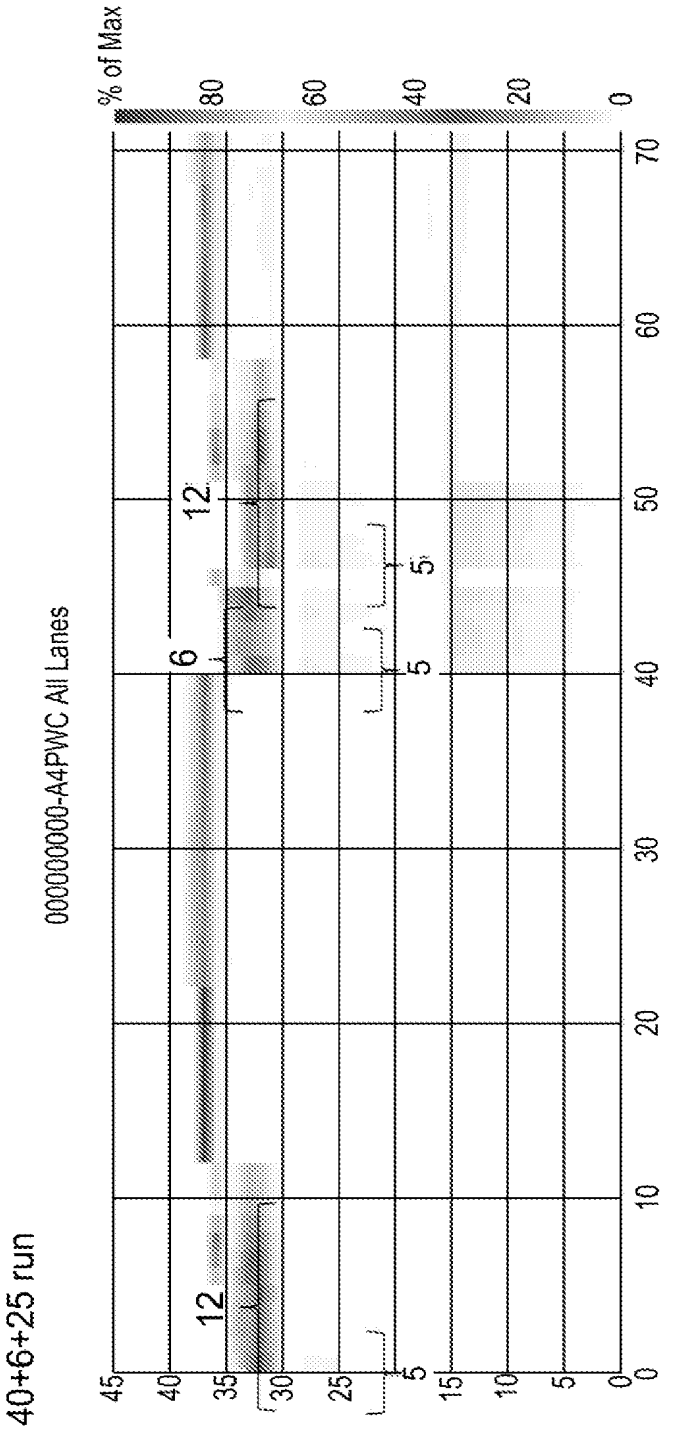

40+6+25 run
000000000-A4PWC All Lanes
% of Max
0
20
40
60
80
0
10
20
30
40
50
60
70
0
5
10
15
25
30
35
40
45
12
6
12
5
5
5

FIG. 14

40+6+25 run

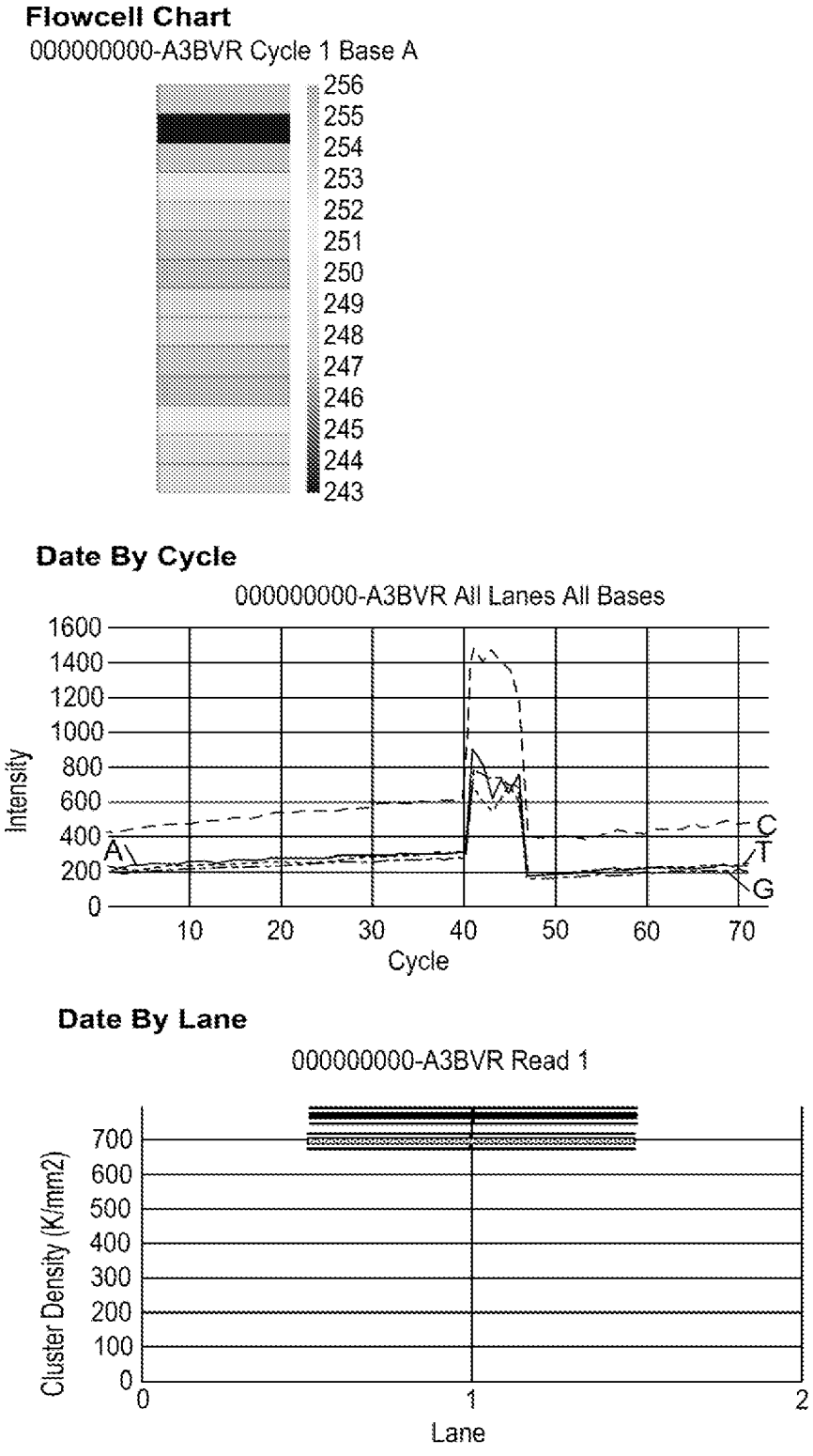
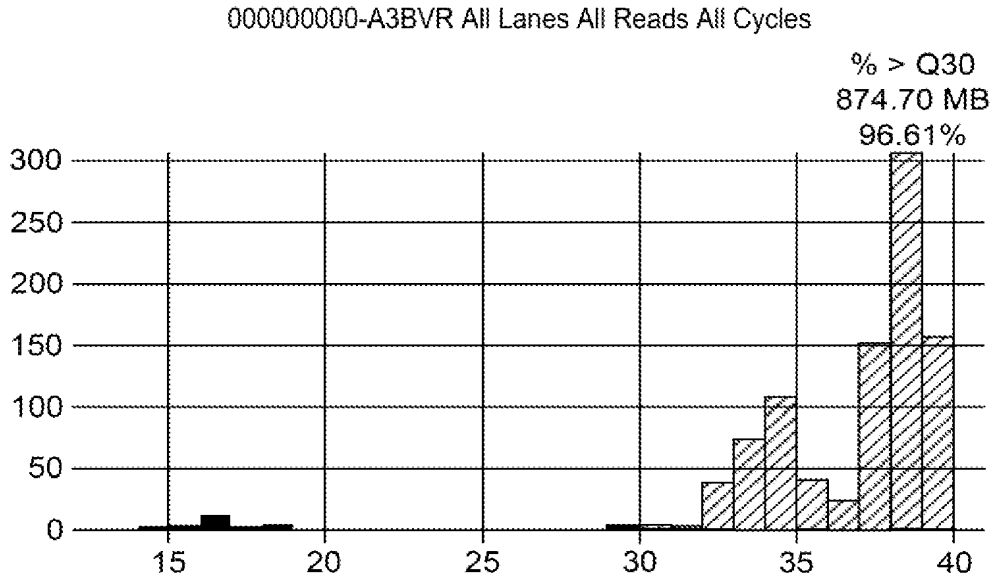


FIG. 14
(Continued)

QScore Distribution



QScore Heatmap

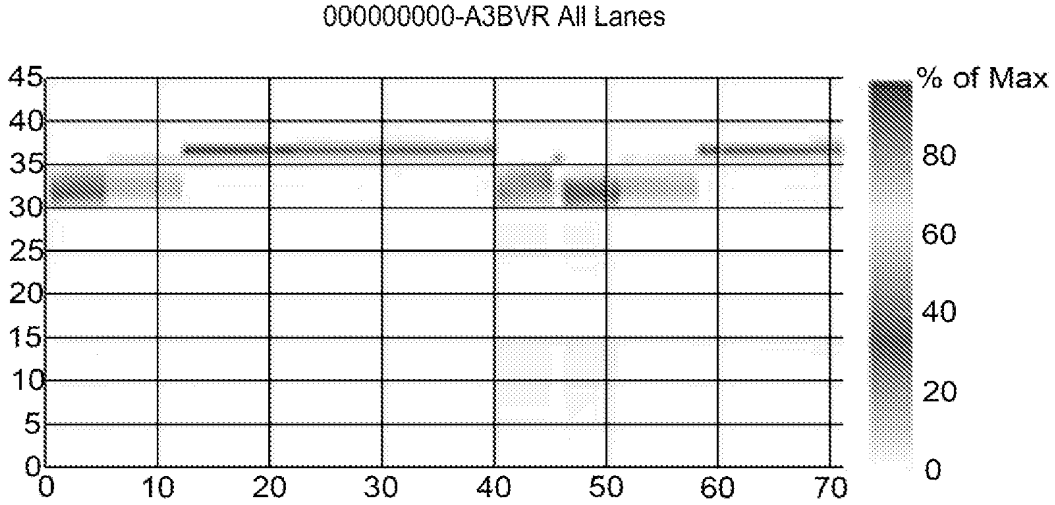


FIG. 15

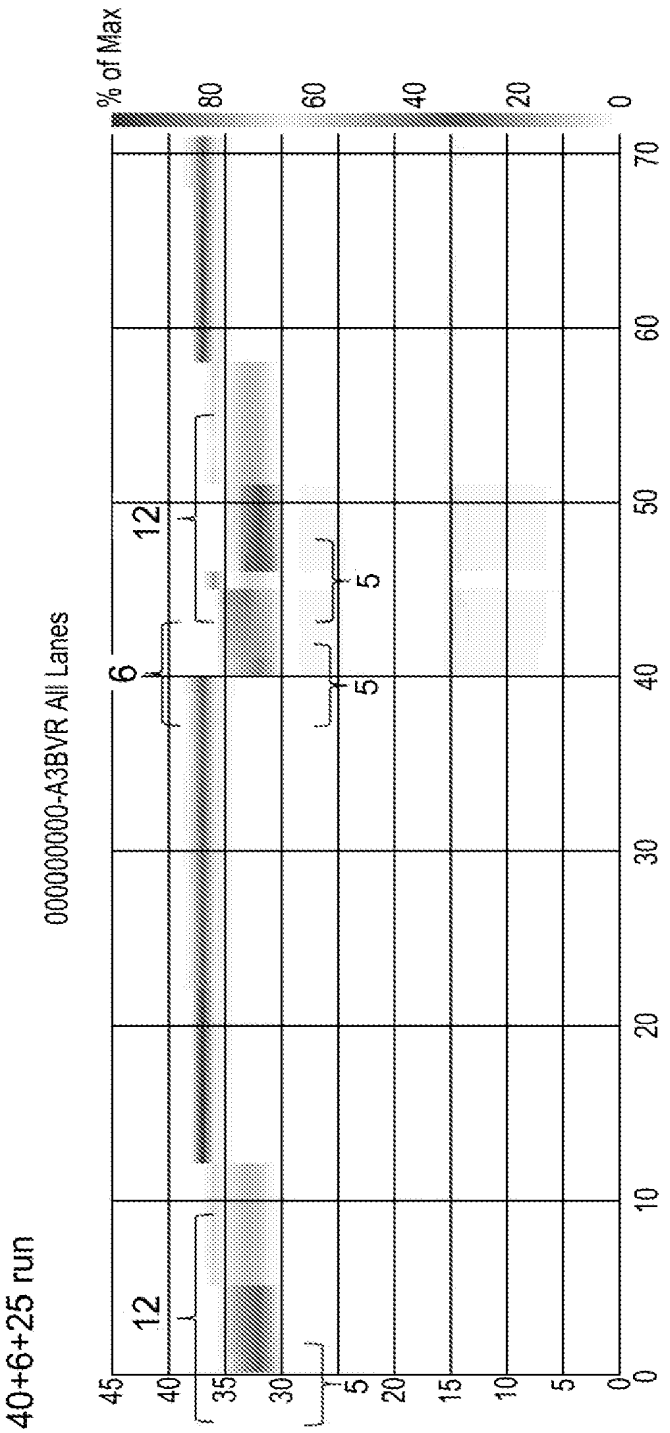

40+6+25 run
000000000-A3BVR All Lanes
% of Max
12
6
12
5
5
0
5
10
15
20
25
30
35
40
45
0
10
20
30
40
50
60
70
0
20
40
60
80

FIG. 16

Runs >> 130731 MO1277 0103 000000000-A4PWC

Run Summary

Level	Cycles	Yield Total	Projected Total Yield	Yield Perfect	Yield <= 3 errors	Aligned (%)	% Perfect [Non Usable Cycles]	% <= 3 errors [Num Usable Cycles]	Error Rate (%)	Intensity Cycle 1	% Intensity Cycle 20	% >= Q30
Read 1	40	300.0 M	300.0 M	0.0	0.0	0.00	0.0 [39]	0.0 [39]	0.00	31	632.9	98.5
Read 2 (I)	6	38.5 M	38.5 M	0.0	0.0	0.00	0.0 [5]	0.0 [5]	0.00	850	0.0	90.9
Read 3	25	184.6 M	184.6 M	0.0	0.0	0.00	0.0 [24]	0.0 [24]	0.00	53	447.1	92.2
Total	71	523.1 M	523.1 M	0.0	0.0	0.00	0.0	0.0	0.00	311	540.0	95.7

Lane	Tiles	Density (K/ mm2)	Cluster PF (%)	Phas/ Prephas (%)	Reads (M)	Reads PF (M)	% > Q30	Yield	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1	% Intensity Cycle 20
Read 1																
1	28	446 +/- 7	91.95 +/- 0.41	0.202 / 0.227	8.36	7.69	98.5	300.0 M	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	31 +/- 2	632.9 +/- 24.3
Read 2 (I)																
1	28	446 +/- 7	91.95 +/- 0.41	0.653 / 0.129	8.36	7.69	90.9	38.5 M	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	850 +/- 77	0.0 +/- 0.0
Read 3																
1	28	446 +/- 7	91.95 +/- 0.41	0.259 / 0.245	8.36	7.69	92.2	184.6 M	0	0.0 +/- 0.0	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	53 +/- 4	447 +/- 22.6

FIG. 17

Runs >> 130314 MO1277 0048 00000000-A3BVR

Run Summary

Level	Cycles	Yield Total	Projected Total Yield	Yield Perfect	Yield <= 3 errors	Aligned (%)	% Perfect [Non Usable Cycles]	% <= 3 errors [Num Usable Cycles]	Error Rate (%)	Intensity Cycle 1	% Intensity Cycle 20	% >= Q30
Read 1	40	521.5 M	521.5 M	2.7 M	2.9 M	0.56	92.8 [39]	99.9 [39]	0.21	228	117.6	97.9
Read 2 (I)	6	66.9 M	66.9 M	0.0	0.0	0.00	0.0 [5]	0.0 [5]	0.00	894	0.0	93.9
Read 3	25	320.9 M	320.9 M	1.7 M	1.8 M	0.55	95.0 [24]	100.0 [24]	0.21	185	119.3	95.2
Total	71	909.3 M	909.3 M	4.4 M	4.7 M	0.55	83.9	100.0	0.21	436	118.4	96.6

Lane	Tiles	Density (K/ mm2)	Cluster PF (%)	Phas/ Prephas (%)	Reads (M)	Reads PF (M)	% > Q30	Yield	Cycles Err Rated	Aligned (%)	Error Rate (%)	Error Rate 35 cycle (%)	Error Rate 75 cycle (%)	Error Rate 100 cycle (%)	Intensity Cycle 1	% Intensity Cycle 20
Read 1																
1	28	770 +/- 14	91.02 +/- 0.27	0.118 / 0.249	14.69	13.37	97.9	521.5 M	39	0.6 +/- 0.0	0.21 +/- 0.02	0.20 +/- 0.02	0.00 +/- 0.0	0.00 +/- 0.0	228 +/- 22	117.6 +/- 2.6
Read 2 (I)																
1	28	770 +/- 14	91.02 +/- 0.27	0.000 / 0.000	14.69	13.37	93.9	66.9 M	0	0.0 +/- 0.0	0.00 +/- 0.0	0.00 +/- 0.0	0.00 +/- 0.0	0.00 +/- 0.0	894 +/- 88	0.0 +/- 0.0
Read 3																
1	28	770 +/- 14	91.02 +/- 0.27	0.126 / 0.266	14.69	13.37	95.2	320.9 M	24	0.6 +/- 0.0	0.21 +/- 0.02	0.00 +/- 0.0	0.00 +/- 0.0	0.00 +/- 0.0	185 +/- 17	119.3 +/- 3.1

IDT Exhibit 1020.3841

FIG. 18A

Adaptors at each end can be different lengths.

H = A,C, or T
D = A,G, or T
Y = C or T

FC-FWD/MetSeq 5′-…GAtGTGTGtTtTTttGATtTDDDYGG —— insert —— t***C***GHHHAGAT***C***GGAAGAG***C***A***C***A***C***GT***C***…TTG-3′
MetSeq 5′…GGAGTTTAGATGTGTGTTTTTTTGATTT

Forward Read
Within the 6nt at 5′ end, look for YGG and trim the 0-3 D's before it:
YGG…insert (what if YGG not found?)
DYGG…insert (what to do with non-D bases?)
DDYGG…insert
DDDYGG…insert (but in this case, DDD can = TGG, so how to tell if adaptor or real insert?)

Within the 6nt at 3′ end (after adaptor trimming is done), look for TCG and trim it and the 0-3 H's after it:
insert…TCGHHH
insert…TCGHH
insert…TCGH
insert…TCG (what to do with non-H bases?)
It's OK if TCG not found, the MspI fragment can be larger than the read length.
If TCG not found, look for TC at 3′ end and remove it, and if TC not found, look for T at 3′ end and remove it.

properly trimmed Forward read:
5′ YGG…insert…

C Methylated

FIG. 18B

H = A, C, or T
D = A, G, or T
Y = C or T
R = A or G

Adaptors at each end can be different lengths.

FC-FWD/MetSeq 5′-…GAtGTGTGttTTttGATtTDDDYGG — insert — tCGHHHAGATCGGAAGAGCACACGTC…TTG-3′
TS Rev 5′,…GACGTGTGCTCTTCCGATCT

Reverse Read
Within the 6nt at 5′ end, look for CGA and trim the 0-3 D's and the CGA (since this contains no genomic info other than it's an MspI site) :
CGA…insert
DCGA…insert
DDCGA…insert
DDDCGA…insert

Within the 6nt at 3′ end (after adaptor trimming is done), look for CCR and trim the 0-3 H's after it:
insert…CCRHHH
insert…CCRHH
insert…CCRH
insert…CCR
It's OK if CCR is not found, the MspI fragment can be larger than the read length.

properly trimmed Reverse read:
5′ insert…(may or may not end in CCR, depending on the MspI fragment size and readlength)

c Methylated

FIG. 18C

Adaptors at each end can be different lengths.

C Methylated

FC-FWD/MetSeq 5′ -...GAtGTGTGttTTttGATtTD$_{(0-3)}$YGG — insert — tCHHH$_{(0-3)}$AGATCGGAAGAGCACACGTC...TTG-3′

REV 5′ ...GACGTGTGCTCTTCCGATCT

i) Number of D's removed from 5' end of Forward read can be the same as the number of H's removed from 3' end of Reverse read (if the read spans across the entire MspI fragment)

ii) Number of H's removed from 3' end of Forward read (if the read spans across the entire MspI fragment) can be the same as the number of D's removed from 3' end of Reverse read

iii) Insert can align with respect to MspI sites in genome as shown:

(part shown in BOLD may or may not be present in the read depending on the MspI fragment length and read length)

```
Top strand fragment case
genome reference:              5′ ...CCGG... - insert - CCGG...
properly trimmed Forward read: 5′ YGG...    - insert -
properly trimmed Reverse read: 3′ RCC...    - insert -

Bottom strand fragment case
genome reference:              5′ ...CCGG... - insert - CCGG...
properly trimmed Forward read:                - insert - ...GGY 5′
properly trimmed Reverse read:                - insert - ...CCR 3′
```

FIG. 19

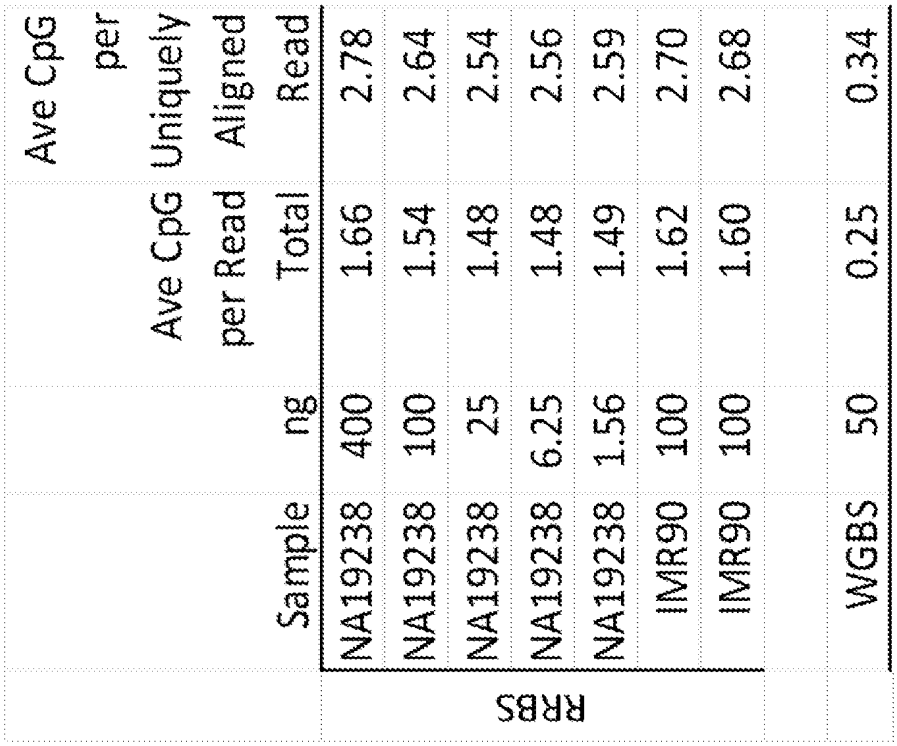
D0 = 1
D1 = 3
D2 = 3x3 = 9
D3 = 3x3x3 = 27

total diversity = 40

FIG. 20

Sample	ng	Sequences analysed in total:	Mapping efficiency:	No alignments:	Did not map uniquely:	CT/CT:	CT/GA:	mCG	mCHG	mCHH	Total CpGs analysed
NA19238	400	28,467,191	59.58%	7.52%	32.91%	8,464,806	8,495,775	28.6%	0.5%	0.3%	**47,208,201**
NA19238	100	27,239,653	58.42%	6.94%	34.64%	7,944,860	7,968,954	32.7%	0.5%	0.3%	**42,071,026**
NA19238	25	26,732,815	58.28%	6.81%	34.92%	7,774,534	7,804,757	34.9%	0.5%	0.4%	**39,632,176**
NA19238	6.25	28,820,811	57.86%	7.09%	35.05%	8,324,948	8,351,549	34.9%	0.6%	0.4%	**42,654,227**
NA19238	1.56	25,889,539	57.46%	7.46%	35.08%	7,425,771	7,451,201	34.5%	0.6%	0.5%	**38,504,765**
IMR90	100	27,431,639	59.94%	6.59%	33.47%	8,213,675	8,228,851	30.5%	0.5%	0.3%	**44,343,781**
IMR90	100	29,671,768	59.59%	6.75%	33.67%	8,833,296	8,846,892	31.0%	0.5%	0.3%	**47,371,339**
WGBS	50	21,577,677	75.66%	6.5%	17.8%	8,161,875	8,163,270	76.7%	1.2%	1.2%	**5,495,435**

FIG. 21



	Sample	ng	Ave CpG per Read Total	Ave CpG per Uniquely Aligned Read
RRBS	NA19238	400	1.66	2.78
	NA19238	100	1.54	2.64
	NA19238	25	1.48	2.54
	NA19238	6.25	1.48	2.56
	NA19238	1.56	1.49	2.59
	IMR90	100	1.62	2.70
	IMR90	100	1.60	2.68
	WGBS	50	0.25	0.34

FIG. 22

		CpG Islands			CpG Shores			Promoters		
Sample	ng	1x depth	5x depth	20x depth	1x depth	5x depth	20x depth	1x depth	5x depth	20x depth
NA19238	400	1,504,475	1,266,783	667,786	656,096	461,332	170,336	1,664,663	1,324,874	660,439
NA19238	100	1,476,388	1,260,019	432,222	659,484	482,788	110,350	1,647,854	1,334,295	430,005
NA19238	25	1,450,314	1,236,153	306,773	654,881	486,022	89,159	1,624,751	1,318,058	312,095
NA19238	6.25	1,436,633	1,228,801	399,082	644,323	479,246	125,823	1,603,990	1,306,739	408,945
NA19238	1.56	1,374,221	1,149,018	345,608	596,378	436,296	107,778	1,516,399	1,215,205	353,982
IMR90	100	1,463,177	1,257,887	552,607	653,059	472,682	145,786	1,625,626	1,323,880	546,676
IMR90	100	1,469,576	1,272,816	619,104	658,688	483,909	170,471	1,636,068	1,343,959	615,871
insilico 40nt SE 40-300bp		1,213,017			507,061			1,305,452		
insilico 40nt SE all sizes		1,665,942			852,847			2,025,019		

FIG. 23

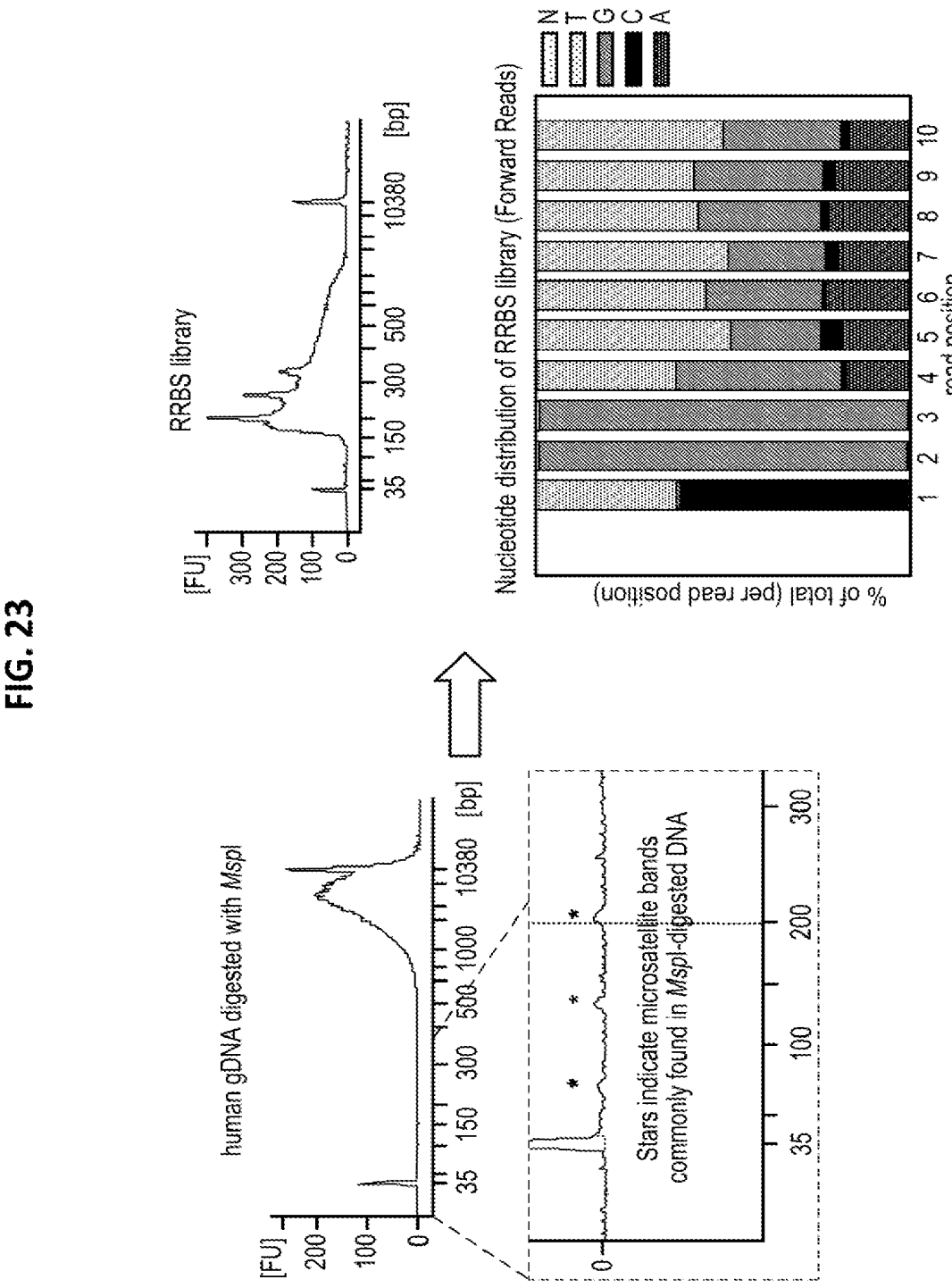

human gDNA digested with MspI
[FU]
200
100
0
35 150 300 500 1000 10380 [bp]
Stars indicate microsatellite bands
commonly found in MspI-digested DNA
0
35 100 200 300
RRBS library
[FU]
300
200
100
0
35 150 300 500 10380 [bp]
Nucleotide distribution of RRBS library (Forward Reads)
% of total (per read position)
N
T
G
C
A
1 2 3 4 5 6 7 8 9 10
read position

FIG. 24

TABLE 1 | Comparison of genome-wide DNA methylation profiling technologies[a].

	RRBS	MeDIP-seq	MethylCap-seq/MBD-seq	WGBS	Infinium 27K
Method for enriching CpGs	CCGG/MspI	Anti-^{m}C antibody	^{m}C-binding protein	N/A	Predefined set
Genomic DNA input	0.01–0.3 μg	0.3–5 μg	1–3 μg	5 μg	0.5–1 μg
Maximum resolution (bp)	1	>100	>100	1	1
Theoretical genomic coverage	10%	100%	100%	100%	0.1%
Actual coverage (≥10 reads[b])	9%	9%	14%	76%	0.1%
Illumina reads per sample	~10 million	~50 million	~50 million	>500 million[c]	Microarray based

RRBS is appropriate for analysis of small amounts of DNA, e.g., from FACS-purified cell populations or embryonic tissue; analysis of clinical samples in FFPE format or laser-microdissected tissue material; studies with a focus on CpG islands and gene promoters (while also sampling a sizable percentage of other genomic elements); studies investigating small effects (e.g., for non-cancer diseases or environmental effects) that require high sequencing depth and many samples to achieve sufficient sensitivity; and multicenter studies that require robustness and reproducibility across several labs. ^{m}C, 5-methylcytosine. [b]Adapted from reference 21. [c]Quantitative measurement of DNA methylation requires a minimum number of overlapping reads. [d]Very deep sequencing is necessary to achieve the same read coverage per CpG as RRBS (which is the key determinant of method sensitivity and accuracy).

FIG. 25

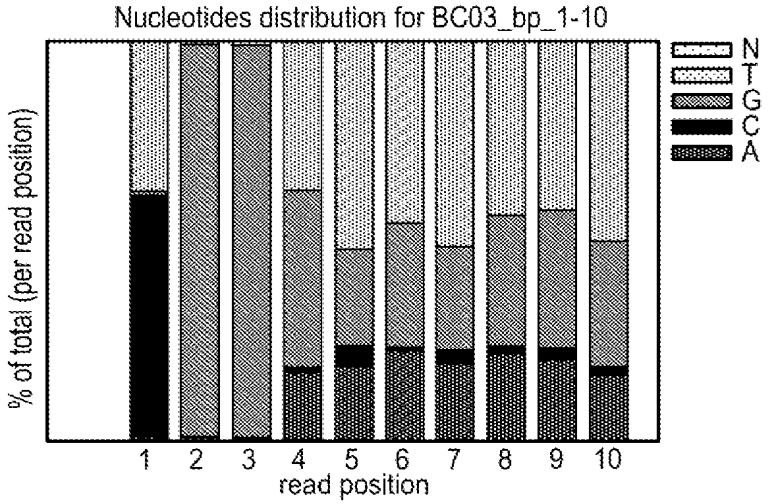
	Class	CpG island	CpG shores	Promoters	Enhancers	Coding Exons	Introns
Genome-wide	# regions in Class	27,718	45,494	82,145	1,339	377,744	404,405
	Average Size	764.1	2,225.6	2,334.9	1,522.2	165.6	6,422.8
	Total # CpG	2,089,538	2,287,717	4,148,299	22,160	1,951,024	26,571,158
	CpG per Kbp	97.3	20.9	20.3	11.4	28.7	14.0
	Stranded CpG	4,179,076	4,575,434	8,296,598	44,320	3,902,048	53,142,316
RRBS	100nt SE	2,187,671	1,003,815	2,431,738	3,860	984,468	5,022,090
	50nt SE	1,452,825	606,959	1,564,822	2,578	629,482	2,786,556
	40nt SE	1,213,017	507,061	1,305,452	2,270	528,990	2,277,500

RRBS analysis considers 40-300bp *Msp* I fragments and assumes 0% methyl-C and 100% bisulfite conversion.

1 CpG = 2 "Stranded CpG" measurements

FIG. 26

RRBS



DARK SEQUENCING

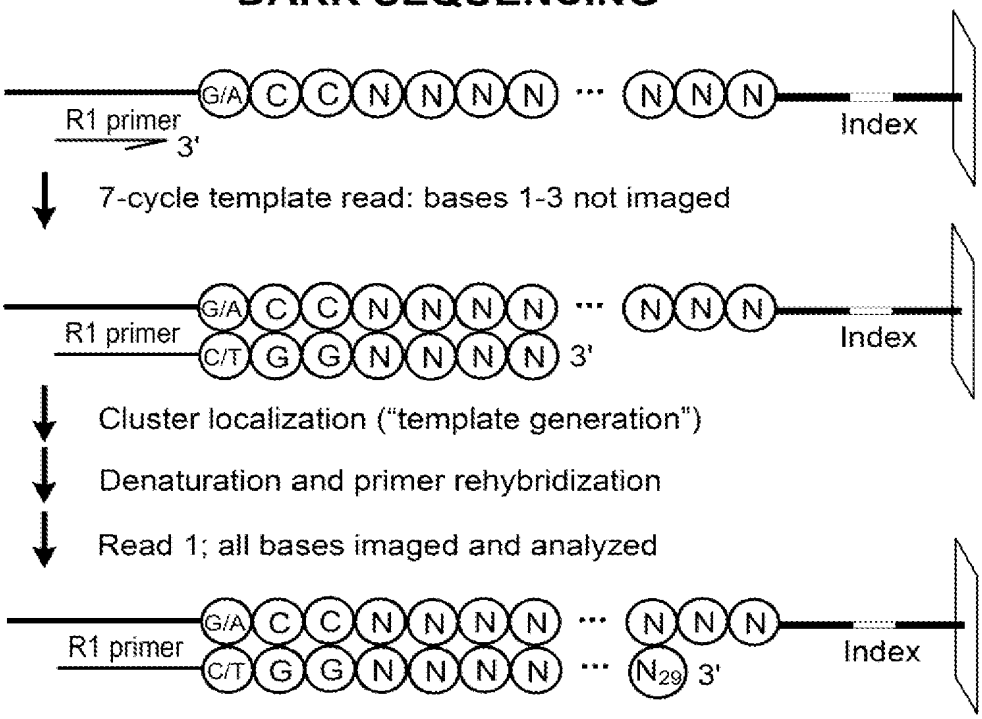


FIG. 27A

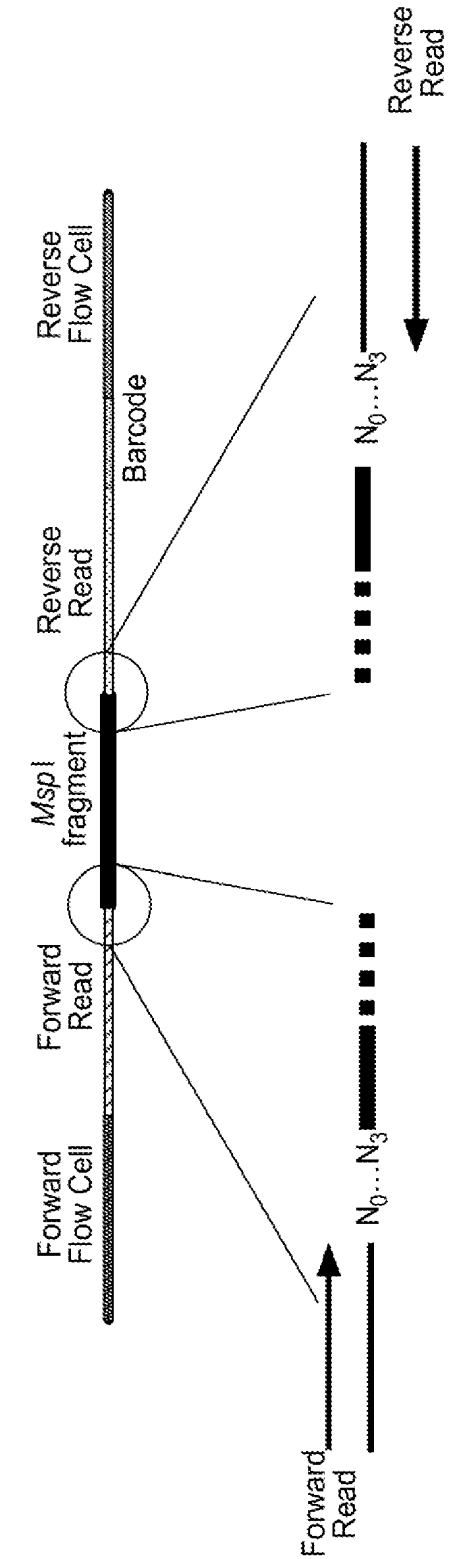


FIG. 27B

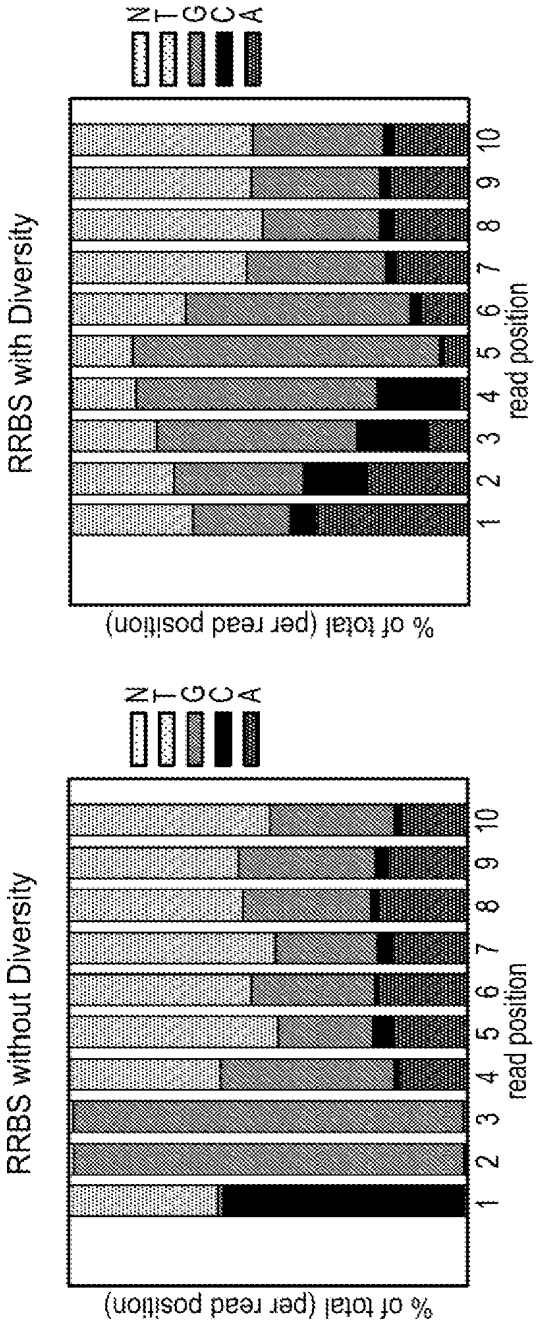
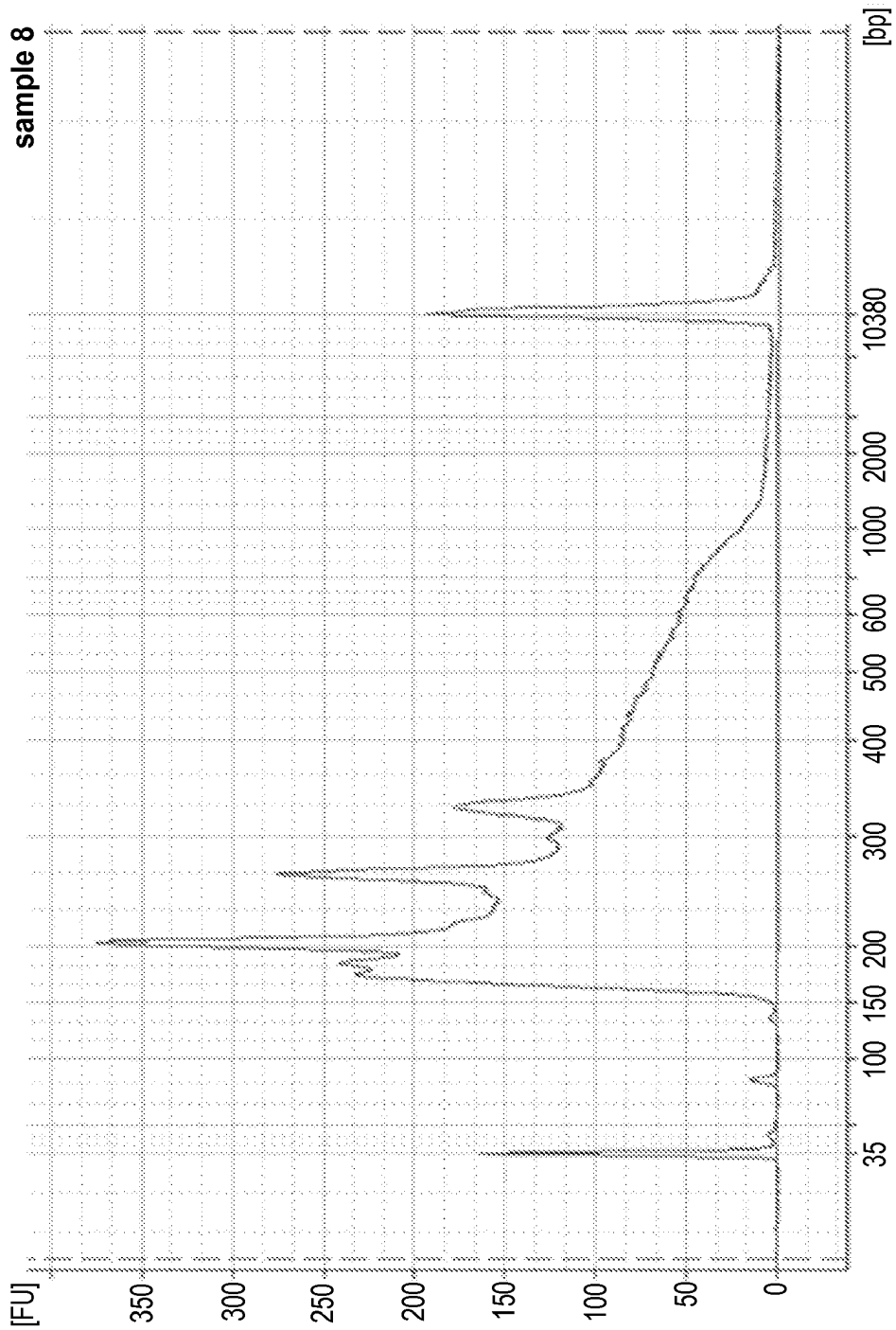
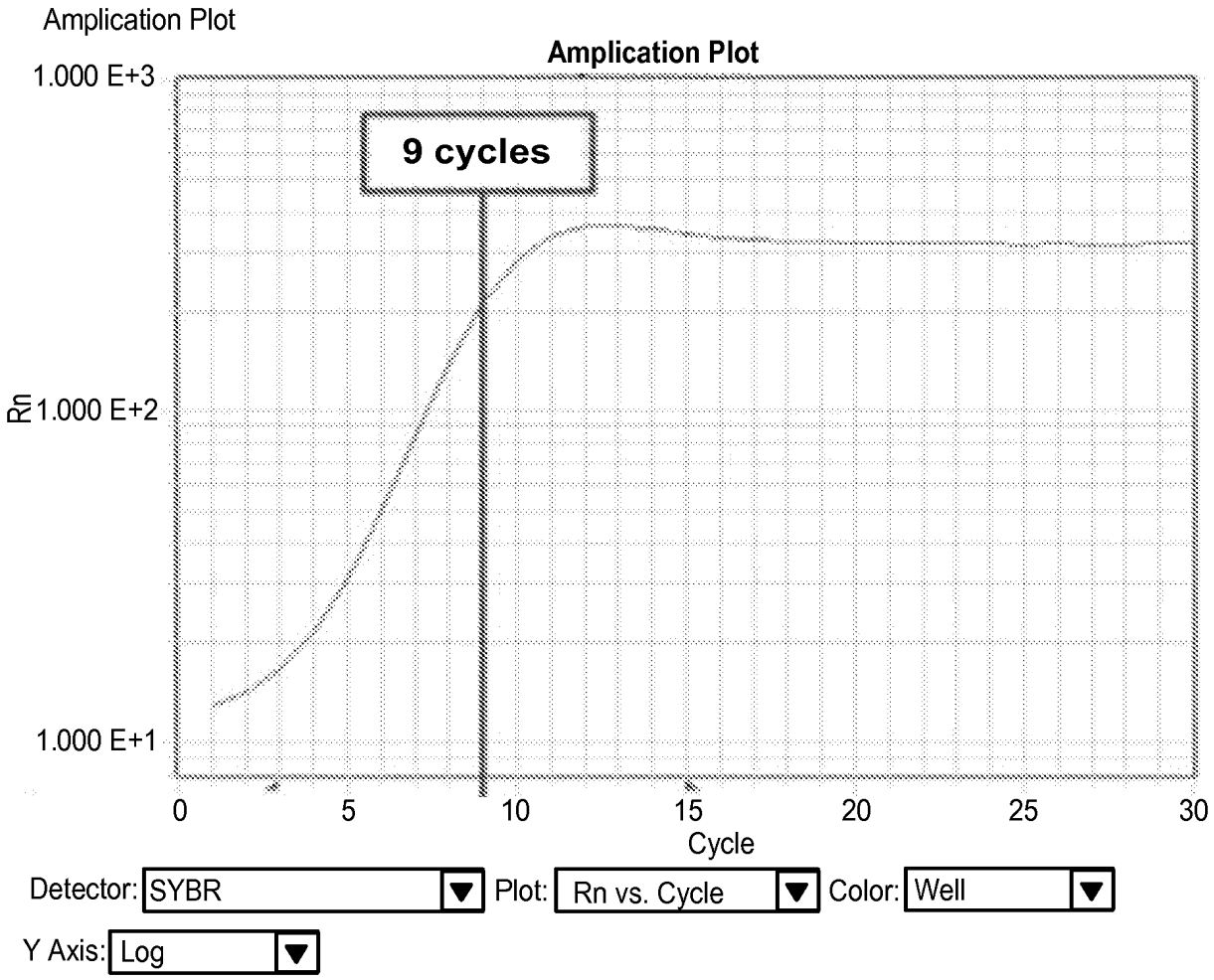


FIG. 28



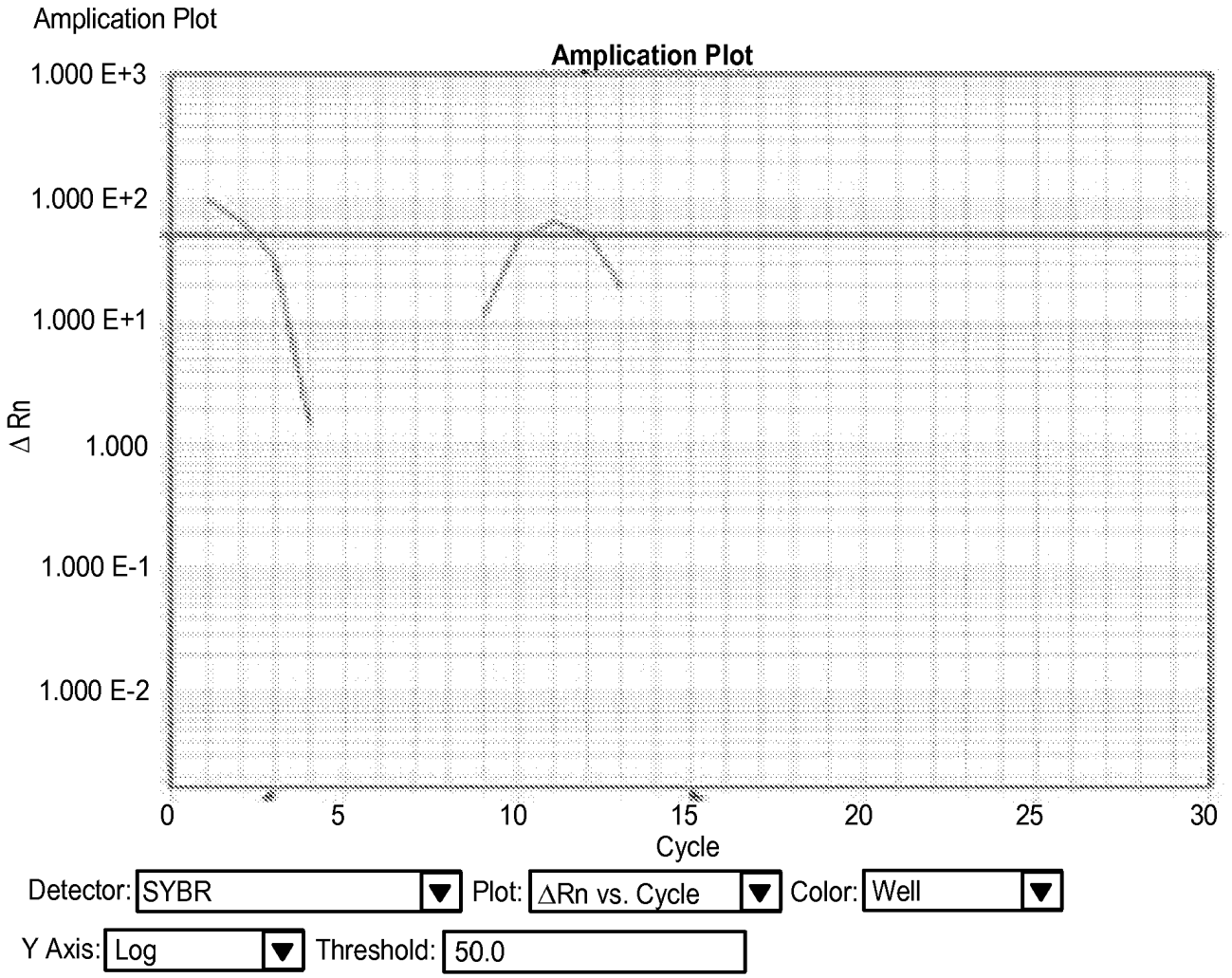
IDT Exhibit 1020.3854

FIG. 29



IDT Exhibit 1020.3855

FIG. 30



IDT Exhibit 1020.3856

FIG. 31

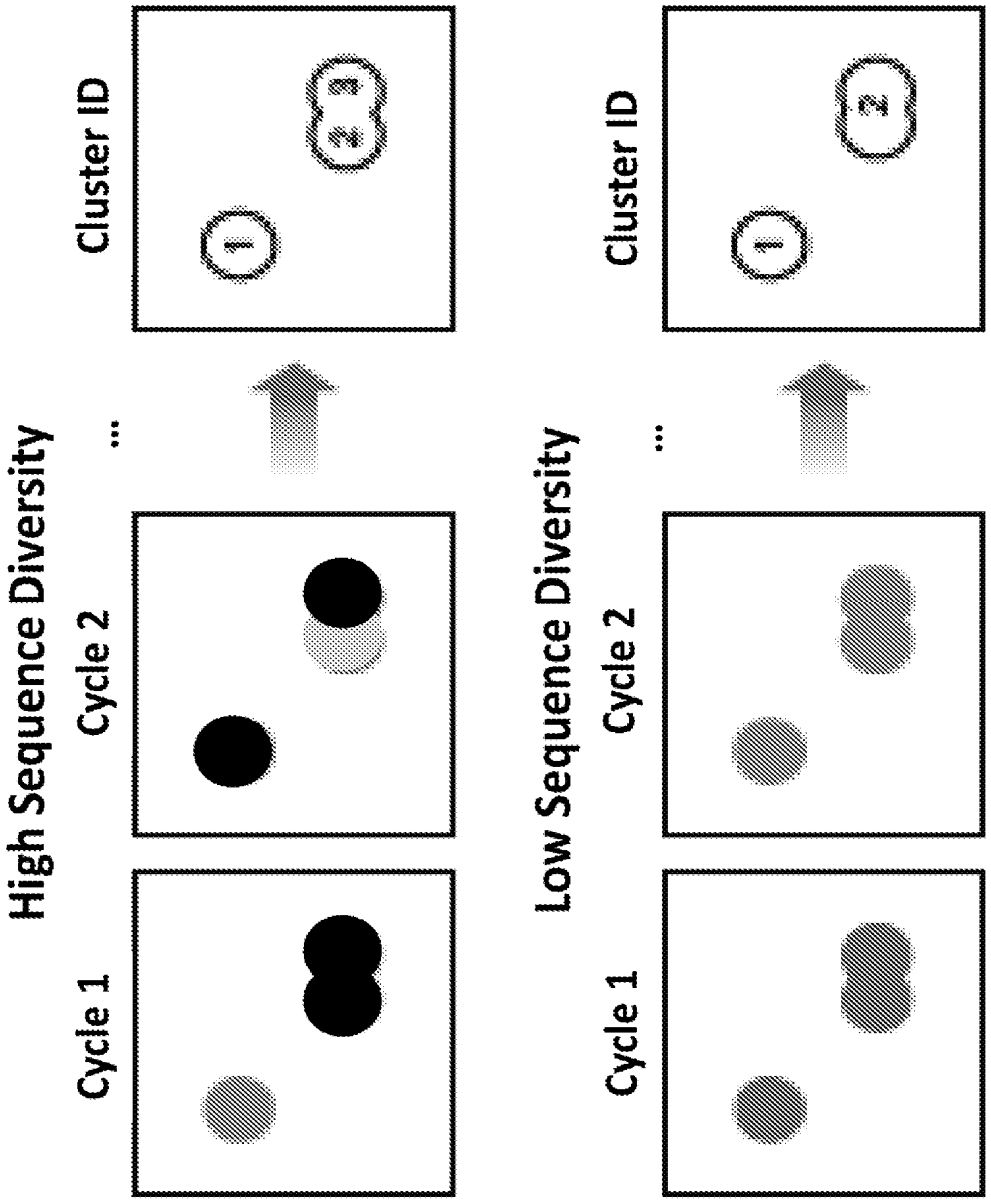

High Sequence Diversity
Cycle 1
Cycle 2
...
Cluster ID
1
2 3
Low Sequence Diversity
Cycle 1
Cycle 2
...
Cluster ID
1
2

FIG. 32

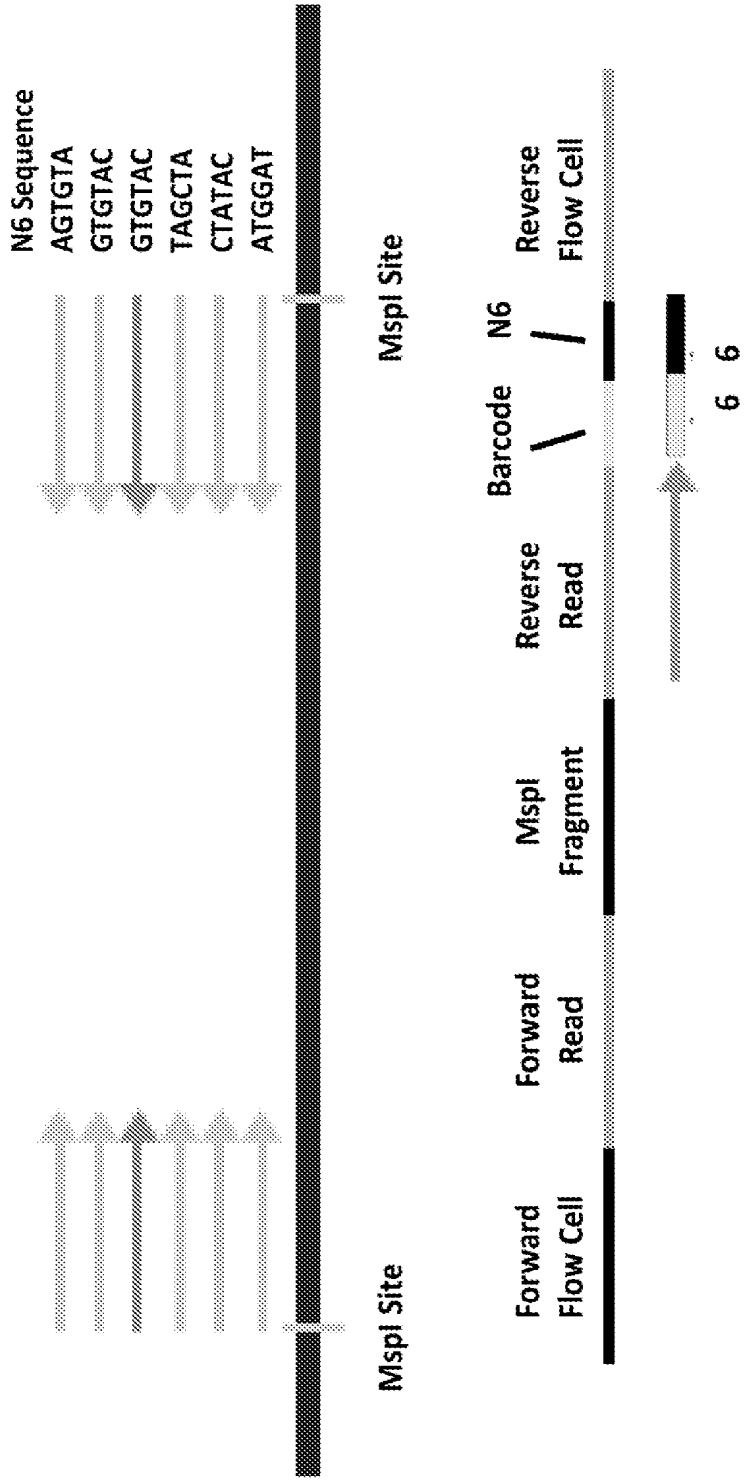

N6 Sequence
AGTGTA
GTGTAC
GTGTAC
TAGCTA
CTATAC
ATGGAT
MspI Site
MspI Site
Forward Flow Cell
Forward Read
MspI Fragment
Reverse Read
Barcode
N6
Reverse Flow Cell
6 6

FIG. 33

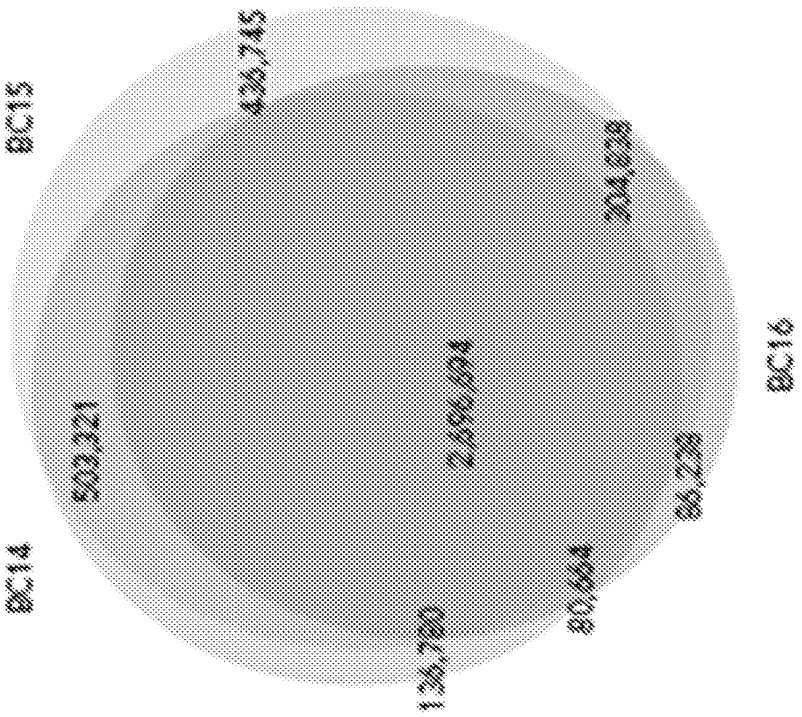

426,745
204,638
2,496,894
86,228
80,364
136,780
203,321
GC15
GC16
GC14

FIG. 34

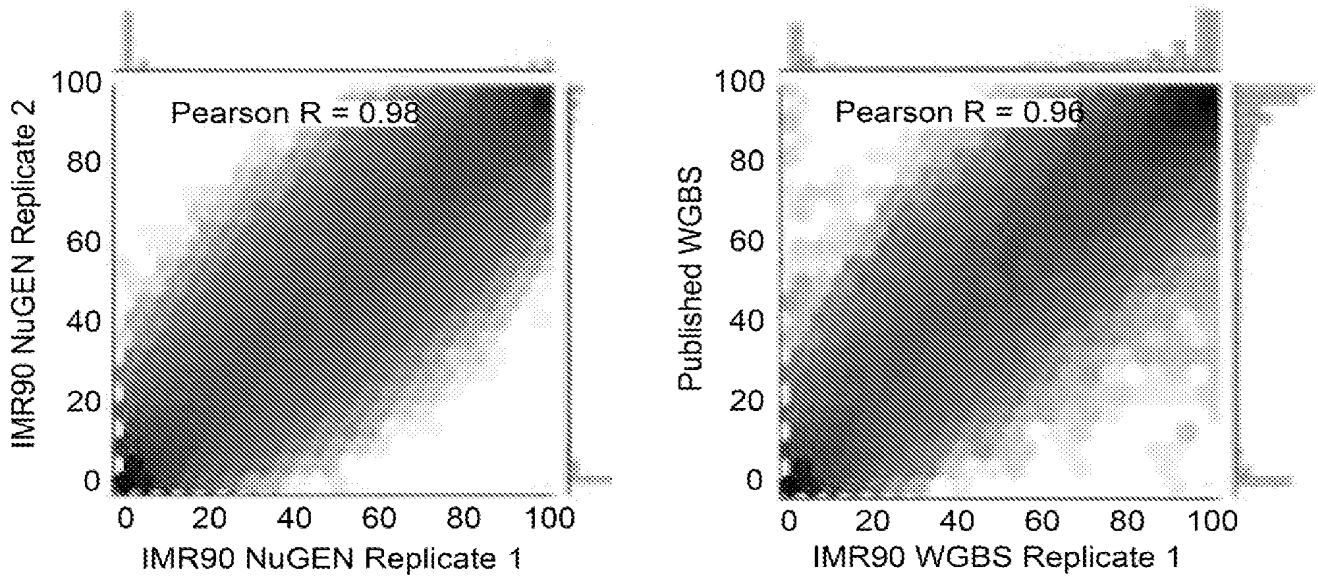

Pearson R = 0.98
IMR90 NuGEN Replicate 2
IMR90 NuGEN Replicate 1
0
20
40
60
80
100
Pearson R = 0.96
Published WGBS
IMR90 WGBS Replicate 1

FIG. 35

*D and R bases comprise DNA sequence derived from library construction

0 diversity bases added	5'	YGG(BCO)	BCO = Bisulfite-Converted Original genomic strand
1 diversity base added	5'	DYGG(BCO)	Y = C or T, depending on methylation status
2 diversity bases added	5'	DDYGG(BCO)	D = A, G, or T
3 diversity bases added	5'	RDDYGG(BCO)	R = A or G

FIG. 36

5' ...CCGG(OT)CCGG... 3'
3' ...GGCC(OB)GGCC... 5'
OT = Original Top genomic strand
OB = Original Bottom genomic strand
CG motifs that may contain 5-methyl C are underlined.
MspI digestion
5' pCGG(OT)C 3'
3' C(OB)GGCp 5'
Adaptor ligation
5' (Ad)CGG(OT)CCG(Ad) 3'
3' (Ad)GCC(OB)GGC(Ad) 5'
DNA sequence derived from library construction is shown in bold. Ad= Adaptor
Bisulfite conversion
5' (Ad)YGG(BCOT)CCG(Ad) 3'
and
3' (Ad)GCT(BCOB)GGY(Ad) 5'
BCOT = Bisulfite Converted Original Top genomic strand
BCOB = Bisulfite Converted Original Bottom genomic strand
Y = C or T, depending on methylation status

FIG. 37

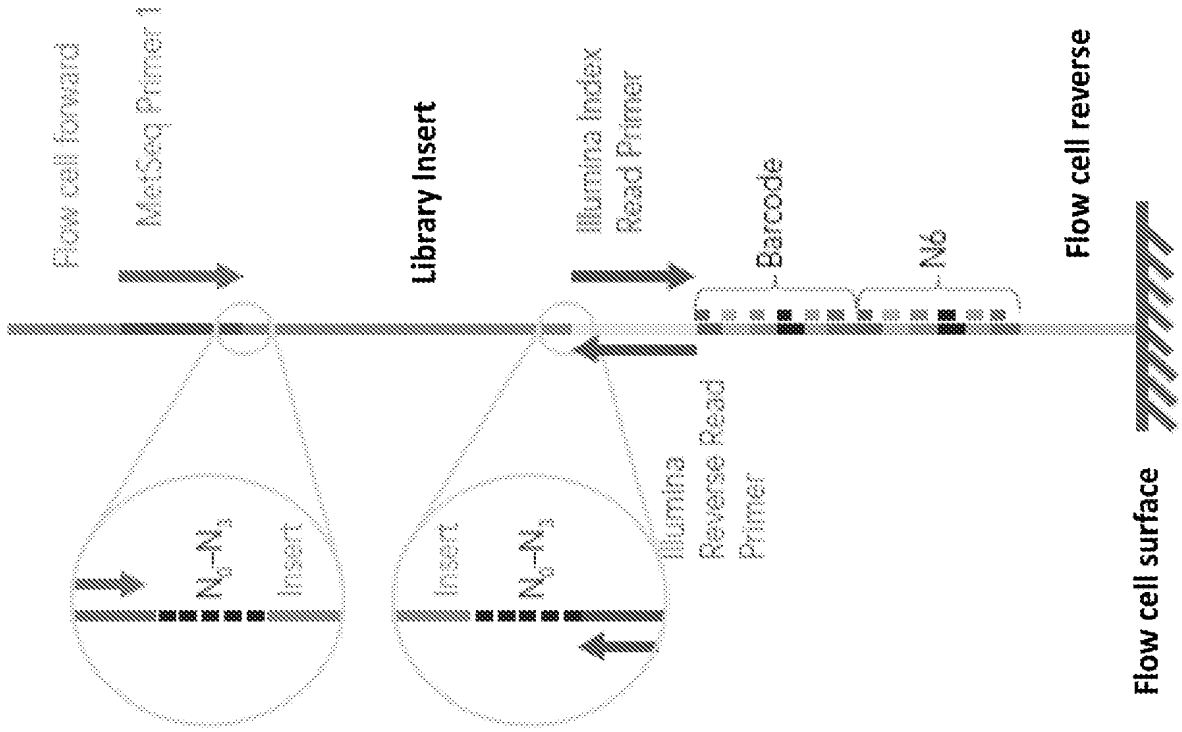
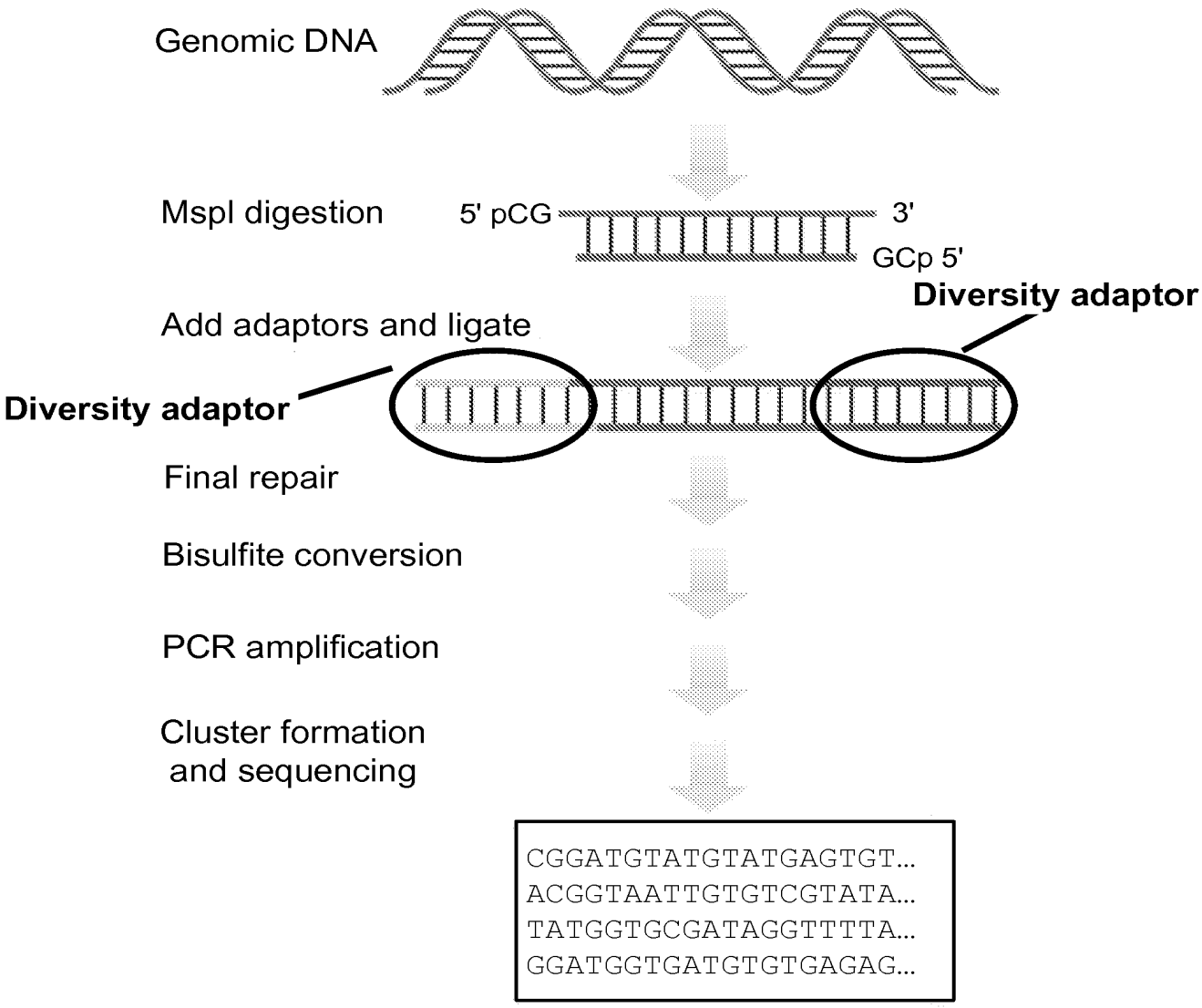

Flow cell forward
MetSeq Primer 1
Library Insert
Illumina Index Read Primer
Barcode
N6
Flow cell reverse
Flow cell surface
Illumina Reverse Read Primer
Insert
N_1–N_3
Insert
N_1–N_3

FIG. 38



IDT Exhibit 1020.3864

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2015/018112

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C12P 19/34 (2015.01)
CPC - C12Q 1/6874 (2015.04)
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)
IPC(8) - C12N 15/10; C12P 19/34; C40B 50/06; G06F 19/10 (2015.01)
CPC - C12Q 1/6874, 1/6806, 1/6837 (2015.04)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
CPC - C12Q 1/6874, 1/6806, 1/6837 (2015.04) (keyword delimited)
US Classes - 435/6.11, 91.2; 506/16

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

PatBase, Google Patents, Google Scholar, Google, PubMed
Search terms used: bisulfite sequencing, methylation, epigenetic, diversity, nucleotide, oligonucleotide, polynucleotide, end, terminus, termini, terminal, sequence, extender, overhang, surface, cluster, in silico

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2010/063711 A1 (ISTITUTO DI RICHERCHE DI BIOLOGIA MOLECOLARE P. ANGELETTI S. P. A. et al) 10 June 2010 (10.06.2010) entire document	1-9, 32-34, 45-53, 70-72, 83-94, 98, 99, 101-112, 134-136, 143
A	US 2012/0157322 A1 (MYLLYKANGAS et al) 21 June 2012 (21.06.2012) entire document	1-9, 32-34, 45-53, 70-72, 83-94, 98, 99, 101-112, 134-136, 143
A	WO 2013/177220 A1 (THE SCRIPPS RESEARCH INSTITUTE et al) 28 November 2013 (28.11.2013) entire document	1-9, 32-34, 45-53, 70-72, 83-94, 98, 99, 101-112, 134-136, 143
P, X	US 2014/0378345 A1 (10X TECHNOLOGIES INC et al) 25 December 2014 (25.12.2014) entire document	1-9, 32-34, 45-53, 83-94, 98, 99, 101-102
P, X	CHEN et al. "BisQC: an Operational Pipeline for Multiplexed Bisulfite Sequencing," BMC Genomics, 30 April 2014 (30.04.2014), Vol. 15, No. 290, Pgs. 1-22. entire document	1, 45, 83, 103, 143

☐ Further documents are listed in the continuation of Box C. ☐

* 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
08 May 2015	18 JUN 2015

Name and mailing address of the ISA/US	Authorized officer:
Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Blaine R. Copenheaver PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2015/018112

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 10-31, 35-44, 54-69, 73-82, 95-97, 100, 113-133, 137-142
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

IDT Exhibit 1020.3866

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 June 2007 (28.06.2007)

PCT

(10) International Publication Number
WO 2007/073165 A1

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

(21) International Application Number:
PCT/NL2006/000648

(22) International Filing Date:
20 December 2006 (20.12.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/752,590 22 December 2005 (22.12.2005) US

(71) Applicant *(for all designated States except US)*: **KEYGENE N.V.** [NL/NL]; 90, Agro Business Park, NL-6708 PW Wageningen (NL).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **VAN EIJK, Michael, Josephus, Theresia** [NL/NL]; 12, Pastoor Strijboschstraat, NL-5373 EJ Herpen (NL). **SØRENSEN, Anker, Preben** [DK/NL]; 140, Bram Streeflandweg, NL-6871 HZ Renkum (NL). **VAN SCHRIEK, Marco, Gerardus, Maria** [NL/NL]; 44, Wildekamp, NL-6721 JE Bennekom (NL).

(74) Agent: **DE LANG, R.J.**; Exter Polak & Charlouis B.V., P.O.Box 3241, NL-2280 GE Rijswijk (NL).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 2007/073165 A1

(54) Title: METHOD FOR HIGH-THROUGHPUT AFLP-BASED POLYMORPHISM DETECTION

(57) Abstract: The invention relates to a method for the high throughput discovery, detection and genotyping of one or more genetic markers in one or more samples, comprising the steps of restriction endonuclease digest of DNA, adaptor-ligation, optional pre-amplif ication, selective amplification, pooling of the amplified products, sequencing the libraries with sufficient redundancy, clustering followed by identification of the genetic markers within the library and/or between libraries and determination of (co-) dominant genotypes of the genetic markers.

IDT Exhibit 1020.3867

METHOD FOR HIGH-THROUGHPUT AFLP-BASED POLYMORPHISM DETECTION.

Technical Field

The present invention relates to the fields of molecular biology and genetics. The invention relates to rapid discovery, detection and large-scale genotyping of polymorphisms in a nucleic acid sample or between samples. The identified polymorphisms may be used as genetic markers.

Background of the invention

Exploration of genomic DNA has long been desired by the scientific, in particular medical, community. Genomic DNA holds the key to identification, diagnosis and treatment of diseases such as cancer and Alzheimer's disease. In addition to disease identification and treatment, exploration of genomic DNA may provide significant advantages in plant and animal breeding efforts, which may provide answers to food and nutrition problems in the world.

Many diseases are known to be associated with specific genetic components, in particular with polymorphisms in specific genes. The identification of polymorphisms in large samples such as genomes is at present a laborious and time-consuming task. However, such identification is of great value to areas such as biomedical research, developing pharmacy products, tissue typing, genotyping and population studies.

Markers, i.c. genetic markers, have been used for a very long time as a genetic typing method, i.e. to connect a phenotypic trait to the presence, absence or amount of a particular part of DNA (gene). One of the most versatile genetic typing technologies is AFLP, already around for many years and widely applicable to any organism (for reviews see Savelkoul *et al.* J. Clin. Microbiol, 1999, 37(10), 3083-3091; Bensch *et al.* Molecular Ecology, 2005, 14, 2899-2914)

The AFLP technology (Zabeau & Vos, 1993; Vos *et al.*, 1995) has found widespread use in plant breeding and other field since its invention in the early nineties. This is due to several characteristics of AFLP, of which the most important is that no prior

IDT Exhibit 1020.3868

sequence information is needed to generate large numbers of genetic markers in a reproducible fashion. In addition, the principle of selective amplification, a cornerstone of AFLP, ensures that the number of amplified fragments can be brought in line with the resolution of the detection system, irrespective of genome size or origin.

Detection of AFLP fragments is commonly carried out by electrophoresis on slab-gels (Vos *et al.*, 1995) or capillary electrophoresis (van der Meulen *et al.*, 2002). The majority of AFLP markers scored in this way represent (single nucleotide) polymorphisms occurring either in the restriction enzyme recognition sites used for AFLP template preparation or their flanking nucleotides covered by selective AFLP primers. The remainder of the AFLP markers are insertion/deletion polymorphisms occurring in the internal sequences of the restriction fragments and a very small fraction on single nucleotide substitutions occurring in small restriction fragments (< approximately 100 bp), which for these fragments cause reproducible mobility variations between both alleles; these AFLP markers can be scored co-dominantly without having to rely on band intensities.

In a typical AFLP fingerprint, the AFLP markers therefore constitute the minority of amplified fragments (less than 50 percent but often less than 20 percent), while the remainder are commonly referred to as constant AFLP fragments. The latter are nevertheless useful in the gel scoring procedure as they serve as anchor points to calculate fragments mobilities of AFLP markers and aid in quantifying the markers for co-dominant scoring. Co-dominant scoring (scoring for homo- or heterozygosity) of AFLP markers currently is restricted to the context of fingerprinting a segregating population. In a panel of unrelated lines, only dominant scoring is possible.

Although the throughput of AFLP is very high due to high multiplexing levels in the amplification and detection steps, the rate limiting step is the resolving power of electrophoresis. Electrophoresis allows unique identification of the majority of amplified fragments based on the combination of restriction enzyme combinations (EC), primer combinations (PC) and mobility, but ideally, the detection system should be capable of determining the entire sequence of the amplified fragments to capture all polymorphisms.

Detection by sequencing instead of mobility determination will

IDT Exhibit 1020.3869

increase throughput because:

1) polymorphisms located in the internal sequences will be detected in most (or all) amplified fragments; this will increase the number of markers per PC considerably.

2) no loss of AFLP markers due to co-migration of AFLP markers and constant bands.

3) co-dominant scoring does not rely on quantification of band intensities and is independent of the relatedness of the individuals fingerprinted.

So far, detection of AFLP markers/sequences by sequencing has not been economically feasible due to, among other limitations, cost limitations of Sanger dideoxy sequencing technology and other conventional sequencing technologies.

Accordingly, it is one of the goals of the present invention to provide for economically feasible methods for the detection of AFLP markers or other genetic markers such as SNP markers based on sequencing.

An important problem further associated with detection of a collection of AFLP or SNP containing fragments via sequencing for genotyping (i.e. diagnostic) purposes is that of sampling variation. Specifically, this means that when a collection of fragments is analyzed and particular fragments are not observed, one has to make sure that this is not due to the fact that the fragments involved were not sampled at the detection step, although they are present in the fragment mixture, because this would lead to false-negative scoring of the marker. This limitation does not apply to detection by electrophoresis because position information on the gel is available.

Accordingly, it is one of the further goals of the present invention provide a method that solves the problem of sample variation or at least reduces the errors caused by sample variation to an acceptable minimum.

Summary of the invention

The present inventors have found that sequencing is within reach for the detection of AFLP and SNP markers with the use of AFLP in certain adapted procedures for high throughput sequencing. The invention thus provides a method or strategy which combines the power and generic applicability of AFLP with certain high throughput sequencing technologies to establish a generically applicable polymorphism scoring system. In this strategy, the issue of sampling

IDT Exhibit 1020.3870

variation is also addressed to ensure genotyping with high accuracy and maximizing chances for datasets with minimal numbers of missing genotypes.

Definitions

In the following description and examples a number of terms are used. In order to provide a clear and consistent understanding of the specification and claims, including the scope to be given such terms, the following definitions are provided. Unless otherwise defined herein, all technical and scientific terms used have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. The disclosures of all publications, patent applications, patents and other references are incorporated herein in their entirety by reference.

Polymorphism: polymorphism refers to the presence of two or more variants of a nucleotide sequence in a population. A polymorphism may comprise one or more base changes, an insertion, a repeat, or a deletion. A polymorphism includes e.g. a simple sequence repeat (SSR) and a single nucleotide polymorphism (SNP), which is a variation, occurring when a single nucleotide: adenine (A), thymine (T), cytosine (C) or guanine (G) - is altered. A variation must generally occur in at least 1% of the population to be considered a SNP. SNPs make up e.g. 90% of all human genetic variations, and occur every 100 to 300 bases along the human genome. Two of every three SNPs substitute Cytosine (C) with Thymine (T). Variations in the DNA sequences of e.g. humans or plants can affect how they handle diseases, bacteria, viruses, chemicals, drugs, etc.

Nucleic acid: a nucleic acid according to the present invention may include any polymer or oligomer of pyrimidine and purine bases, preferably cytosine, thymine, and uracil, and adenine and guanine, respectively (*See* Albert L. Lehninger, *Principles of Biochemistry*, at 793-800 (Worth Pub. 1982) which is herein incorporated by reference in its entirety for all purposes). The present invention contemplates any deoxyribonucleotide, ribonucleotide or peptide nucleic acid component, and any chemical variants thereof, such as methylated, hydroxymethylated or glycosylated forms of these bases, and the like. The polymers or oligomers may be heterogenous or homogenous in composition, and may be isolated from naturally occurring sources or may be artificially or synthetically produced. In addition, the nucleic acids may be DNA or RNA, or a mixture thereof, and may exist

IDT Exhibit 1020.3871

permanently or transitionally in single-stranded or double-stranded form, including homoduplex, heteroduplex, and hybrid states.

Complexity reduction: the term complexity reduction is used to denote a method wherein the complexity of a nucleic acid sample, such as genomic DNA, is reduced by the generation of a subset of the sample. This subset can be representative for the whole (i.e. complex) sample and is preferably a reproducible subset. Reproducible means in this context that when the same sample is reduced in complexity using the same method, the same, or at least comparable, subset is obtained. The method used for complexity reduction may be any method for complexity reduction known in the art. A preferred example of a method for complexity reduction includes for example AFLP® (Keygene N.V., the Netherlands; see e.g. EP 0 534 858, US6045994), the methods described by Dong (see e.g. WO 03/012118, WO 00/24939), indexed linking (Unrau et al., vide infra), linker-PCR (WO90/008821), and SALSA-PCR (WO00/23620)Schouten et al) etc. The complexity reduction methods used in the present invention have in common that they are reproducible. Reproducible in the sense that when the same sample is reduced in complexity in the same manner, the same subset of the sample is obtained, as opposed to more random complexity reduction such as microdissection or the use of mRNA (cDNA) which represents a portion of the genome transcribed in a selected tissue and for its reproducibility is depending on the selection of tissue, time of isolation etc.

AFLP: AFLP refers to a method for selective amplification of DNA based on digesting a nucleic acid with one or more restriction endonucleases to yield restriction fragments, ligating adaptors to the restriction fragments and amplifying the adaptor-ligated restriction fragments with at least one primer that is (part) complementary to the adaptor, (part) complementary to the remains of the restriction endonuclease, and that further contains at least one randomly selected nucleotide from amongst A,C, T, or G (or U as the case may be). AFLP does not require any prior sequence information and can be performed on any starting DNA. In general, AFLP comprises the steps of:

(a) digesting a nucleic acid, in particular a DNA or cDNA, with one or more specific restriction endonucleases, to fragment the DNA into a corresponding series of restriction fragments;

(b) ligating the restriction fragments thus obtained with

IDT Exhibit 1020.3872

a double-stranded synthetic oligonucleotide adaptor, one end of which is compatible with one or both of the ends of the restriction fragments, to thereby produce adapter-ligated, preferably tagged, restriction fragments of the starting DNA;

(c) contacting the adapter-ligated, preferably tagged, restriction fragments under hybridizing conditions with one or more oligonucleotide primers that contain selective nucleotides at their 3'-end;

(d) amplifying the adapter-ligated, preferably tagged, restriction fragment hybridised with the primers by PCR or a similar technique so as to cause further elongation of the hybridised primers along the restriction fragments of the starting DNA to which the primers hybridised; and

(e) detecting, identifying or recovering the amplified or elongated DNA fragment thus obtained.

AFLP thus provides a reproducible subset of adaptor-ligated fragments. AFLP is described in EP 534858, US 6045994 and in Vos *et al.* Reference is made to these publications for further details regarding AFLP. The AFLP is commonly used as a complexity reduction technique and a DNA fingerprinting technology. Within the context of the use of AFLP as a fingerprinting technology, the concept of an AFLP marker has been developed.

AFLP marker: An AFLP marker is an amplified adaptor-ligated restriction fragment that is different between two samples that have been amplified using AFLP (fingerprinted), using the same set of primers. As such, the presence or absence of this amplified adaptor-ligated restriction fragment can be used as a marker that is linked to a trait or phenotype. In conventional gel technology, an AFLP marker showes up as a band in the gel located at a certain mobility. Other electrophoretic techniques such as capillary electrophoresis may not refer to this as a band, but the concept remains the same, i.e. a nucleic acid with a certain length and mobility. Absence or presence of the band may be indicative of (or associated with) the presence or absence of the phenotype. AFLP markers typically involve SNPs in the restriction site of the endonuclease or the selective nucleotides. Occasionally, AFLP markers may involve indels in the restriction fragment.

SNP marker: a SNP marker is a marker that is based on an identified single nucleotide polymorphism at a certain position. SNP markers can be located at identical positions to AFLP markers, but

IDT Exhibit 1020.3873

SNP markers can also be located in the restriction fragment itself. As such the genus SNP markers thus encompasses the species AFLP markers.

Constant band: a constant band in the AFLP technology is an amplified adaptor-ligated restriction fragment that is relatively invariable between samples. Thus, a constant band in the AFLP technology will, over a range of samples, show up at about the same position in the gel, i.e. has the same length/mobility. In conventional AFLP these are typically used to anchor the lanes corresponding to samples on a gel or electropherograms of multiple AFLP samples detected by capillary electrophoresis. Typically, a constant band is less informative than an AFLP marker. Nevertheless, as AFLP markers customary involve SNPs in the selective nucleotides or the restriction site, constant bands may comprise SNPs in the restriction fragments themselves, rendering the constant bands an interesting alternative source of genetic information that is complementary to AFLP markers.

Selective base: Located at the 3' end of the primer that contains a part that is complementary to the adaptor and a part that is complementary to the remains of the restriction site, the selective base is randomly selected from amongst A, C, T or G. By extending a primer with a selective base, the subsequent amplification will yield only a reproducible subset of the adaptor-ligated restriction fragments, i.e. only the fragments that can be amplified using the primer carrying the selective base. Selective nucleotides can be added to the 3'end of the primer in a number varying between 1 and 10. Typically 1-4 suffice. Both primers may contain a varying number of selective bases. With each added selective base, the subset reduces the amount of amplified adaptor-ligated restriction fragments in the subset by a factor of about 4. Typically, the number of selective bases used in AFLP is indicated by +N+M, wherein one primer carries N selective nucleotides and the other primers carries M selective nucleotides. Thus, an Eco/Mse +1/+2 AFLP is shorthand for the digestion of the starting DNA with EcoRI and MseI, ligation of appropriate adaptors and amplification with one primer directed to the EcoRI restricted position carrying one selective base and the other primer directed to the MseI restricted site carrying 2 selective nucleotides.

Clustering: with the term "clustering" is meant the comparison of two or more nucleotide sequences based on the presence of short or

IDT Exhibit 1020.3874

long stretches of identical or similar nucleotides. Several methods for alignment of nucleotide sequences are known in the art, as will be further explained below. Sometimes the terms "assembly" or "alignment" are used as synonyms.

Tag: a short sequence that can be added to a primer or included in its sequence or otherwise used as label to provide a unique identifier. Such a sequence identifier can be a unique base sequence of varying but defined length uniquely used for identifying a specific nucleic acid sample. For instance 4 bp tags allow 4(exp4) = 256 different tags. Typical examples are ZIP sequences, known in the art as commonly used tags for unique detection by hybridization (Iannone *et al.* Cytometry 39:131-140, 2000). Using such a tag, the origin of a PCR sample can be determined upon further processing. In the case of combining processed products originating from different nucleic acid samples, the different nucleic acid samples are generally identified using different tags. In the case of the present invention, the addition of a unique sequence tag serves to identify the coordinates of the individual plant in the pool of sequences amplification products. Multiple tags can be used.

Tagging: the term tagging refers to the addition of a tag to a nucleic acid sample in order to be able to distinguish it from a second or further nucleic acid sample. Tagging can e.g. be performed by the addition of a sequence identifier during complexity reduction or by any other means known in the art. Such sequence identifier can e.g. be a unique base sequence of varying but defined length uniquely used for identifying a specific nucleic acid sample. Typical examples thereof are for instance ZIP sequences. Using such a tag, the origin of a sample can be determined upon further processing. In case of combining processed products originating from different nucleic acid samples, the different nucleic acid samples should be identified using different tags.

Tagged library: the term tagged library refers to a library of tagged nucleic acids.

Sequencing: The term sequencing refers to determining the order of nucleotides (base sequences) in a nucleic acid sample, e.g. DNA or RNA.

High-throughput screening: High-throughput screening, often abbreviated as HTS, is a method for scientific experimentation especially relevant to the fields of biology and chemistry. Through a combination of modern robotics and other specialised laboratory

IDT Exhibit 1020.3875

hardware, it allows a researcher to effectively screen large amounts of samples simultaneously.

Restriction endonuclease: a restriction endonuclease or restriction enzyme is an enzyme that recognizes a specific nucleotide sequence (target site) in a double-stranded DNA molecule, and will cleave both strands of the DNA molecule at every target site.

Restriction fragments: the DNA molecules produced by digestion with a restriction endonuclease are referred to as restriction fragments. Any given genome (or nucleic acid, regardless of its origin) will be digested by a particular restriction endonuclease into a discrete set of restriction fragments. The DNA fragments that result from restriction endonuclease cleavage can be further used in a variety of techniques and can for instance be detected by gel electrophoresis.

Gel electrophoresis: in order to detect restriction fragments, an analytical method for fractionating double-stranded DNA molecules on the basis of size can be required. The most commonly used technique for achieving such fractionation is (capillary) gel electrophoresis. The rate at which DNA fragments move in such gels depends on their molecular weight; thus, the distances travelled decrease as the fragment lengths increase. The DNA fragments fractionated by gel electrophoresis can be visualized directly by a staining procedure e.g. silver staining or staining using ethidium bromide, if the number of fragments included in the pattern is sufficiently small. Alternatively further treatment of the DNA fragments may incorporate detectable labels in the fragments, such as fluorophores or radioactive labels.

Ligation: the enzymatic reaction catalyzed by a ligase enzyme in which two double-stranded DNA molecules are covalently joined together is referred to as ligation. In general, both DNA strands are covalently joined together, but it is also possible to prevent the ligation of one of the two strands through chemical or enzymatic modification of one of the ends of the strands. In that case the covalent joining will occur in only one of the two DNA strands.

Synthetic oligonucleotide: single-stranded DNA molecules having preferably from about 10 to about 50 bases, which can be synthesized chemically are referred to as synthetic oligonucleotides. In general, these synthetic DNA molecules are designed to have a unique or desired nucleotide sequence, although it is possible to synthesize families of molecules having related sequences and which have

IDT Exhibit 1020.3876

different nucleotide compositions at specific positions within the nucleotide sequence. The term synthetic oligonucleotide will be used to refer to DNA molecules having a designed or desired nucleotide sequence.

Adaptors: short double-stranded DNA molecules with a limited number of base pairs, e.g. about 10 to about 30 base pairs in length, which are designed such that they can be ligated to the ends of restriction fragments. Adaptors are generally composed of two synthetic oligonucleotides which have nucleotide sequences which are partially complementary to each other. When mixing the two synthetic oligonucleotides in solution under appropriate conditions, they will anneal to each other forming a double-stranded structure. After annealing, one end of the adaptor molecule is designed such that it is compatible with the end of a restriction fragment and can be ligated thereto; the other end of the adaptor can be designed so that it cannot be ligated, but this need not be the case (double ligated adaptors).

Adaptor-ligated restriction fragments: restriction fragments that have been capped by adaptors.

Primers: in general, the term primers refer to DNA strands which can prime the synthesis of DNA. DNA polymerase cannot synthesize DNA *de novo* without primers: it can only extend an existing DNA strand in a reaction in which the complementary strand is used as a template to direct the order of nucleotides to be assembled. We will refer to the synthetic oligonucleotide molecules which are used in a polymerase chain reaction (PCR) as primers.

DNA amplification: the term DNA amplification will be typically used to denote the *in vitro* synthesis of double-stranded DNA molecules using PCR. It is noted that other amplification methods exist and they may be used in the present invention without departing from the gist.

Selective hybridisation: relates to hybridisation, under stringent hybridisation conditions, of a nucleic acid sequence to a specified nucleic acid target sequence to a detectably greater degree (e.g., at least 2-fold over background) than its hybridisation to non-target nucleic acid sequences and to the substantial exclusion of non-target nucleic acids. The terms "stringent conditions" or "stringent hybridisation conditions" includes reference to conditions under which a probe will hybridise to its target sequence, to a detectably greater degree than other sequences (e.g., at least 2-fold

IDT Exhibit 1020.3877

over background). Stringent conditions are sequence-dependent and will be different in different circumstances. By controlling the stringency of the hybridisation and/or washing conditions, target sequences can be identified which are 100% complementary to the probe (homologous probing). Alternatively, stringency conditions can be adjusted to allow some mismatching in sequences so that lower degrees of similarity are detected (heterologous probing). Generally, a probe is less than about 100 nucleotides in length, optionally no more than 50, or 25 nucleotides in length. Typically, stringent conditions will be those in which the salt concentration is less than about 1.5 M Na ion, typically about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about is 30 °C. for short probes (e.g., 10 to 50 nucleotides) and at least about 60 °C. for long probes (e.g., greater than 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilising agents such as formamide. Exemplary low stringency conditions include hybridisation with a buffer solution of 30 to 35% formamide, 1 M NaCl, 1% SDS (sodium dodecylsulphate) at 37 °C, and a wash in 1* to 2*SSC (20*SSC=3.0 M NaCl/0.3 M trisodium citrate) at 50 to 55 °C. Exemplary moderate stringency conditions include hybridisation in 40 to 45% formamide, 1 M NaCl, 1% SDS at 37 °C., and a wash in 0.5* to 1*SSC at 55 to 60 °C. Exemplary high stringency conditions include hybridisation in 50% formamide, 1 M NaCl, 1% SDS at 37 °C., and a wash in 0.1*SSC at 60 to 65 °C. Specificity is typically the function of post-hybridisation washes, the critical factors being the ionic strength and temperature of the final wash solution. For DNA-DNA hybrids, the Tm can be approximated from the equation of Meinkoth and Wahl, Anal. Biochem., 138:267-284 (1984): Tm=81.5 °C.+16.6 (log M)+0.41 (% GC)-0.61 (% form)-500/L; where M is the molarity of monovalent cations, % GC is the percentage of guanosine and cytosine nucleotides in the DNA, % form is the percentage of formamide in the hybridisation solution, and L is the length of the hybrid in base pairs. The Tm is the temperature (under defined ionic strength and pH) at which 50% of a complementary target sequence hybridises to a perfectly matched probe. Tm is reduced by about 1 °C. for each 1% of mismatching; thus, Tm, hybridisation and/or wash conditions can be adjusted to hybridise to sequences of the desired identity. For example, if sequences with >90% identity are sought, the Tm can be decreased 10 °C. Generally, stringent conditions are selected to be about 5 °C. lower than the thermal melting point (Tm) for the

IDT Exhibit 1020.3878

specific sequence and its complement at a defined ionic strength and pH. However, severely stringent conditions can utilise a hybridisation and/or wash at 1, 2, 3, or 4 °C. lower than the thermal melting point (Tm); moderately stringent conditions can utilise a hybridisation and/or wash at 6, 7, 8, 9, or 10 °C. lower than the thermal melting point (Tm); low stringency conditions can utilise a hybridisation and/or wash at 11, 12, 13, 14, 15, or 20 °C. lower than the thermal melting point (Tm). Using the equation, hybridisation and wash compositions, and desired Tm, those of ordinary skill will understand that variations in the stringency of hybridisation and/or wash solutions are inherently described. If the desired degree of mismatching results in a Tm of less than 45 °C. (aqueous solution) or 32 °C. (formamide solution) it is preferred to increase the SSC concentration so that a higher temperature can be used. An extensive guide to the hybridisation of nucleic acids is found in Tijssen, Laboratory Techniques in Biochemistry and Molecular Biology-Hybridisation with Nucleic Acid Probes, Part 1, Chapter 2 "Overview of principles of hybridisation and the strategy of nucleic acid probe assays", Elsevier, N.Y. (1993); and Current Protocols in Molecular Biology, Chapter 2, Ausubel, et al., Eds., Greene Publishing and Wiley-Interscience, New York (1995).

Detailed description of the invention

In a first aspect the present invention relates to a method for the high throughput discovery, detection and large-scale genotyping of one or more genetic markers in one or more samples, comprising the steps of:

(a) providing DNA from one or more samples;

(b) restricting the DNA with at least one restriction endonuclease to produce restriction fragments;

(c) ligating adaptors to the restriction fragments to produce adaptor-ligated restriction fragments;

(d) optionally, amplifying the adaptor-ligated restriction fragments with a primer pair that is at least complementary to the adaptors to produce pre-amplified adaptor-ligated restriction fragments;

(e) amplifying the (optionally pre-amplified) adaptor-ligated restriction fragments with a primer pair, wherein at least one of the primers contains an identifier tag at the 5'

IDT Exhibit 1020.3879

end of the primer to produce a library of tagged amplified subsets of adaptor-ligated restriction fragments for each sample;

(f) optionally, pooling the libraries derived from multiple samples;

(g) sequencing the libraries using high throughput sequencing technology;

(h) clustering the sequences per library, using the identifier tag;

(i) identify genetic markers by comparing clustered sequences within a library and/or between the libraries.

(j) determine (co-)dominant genotypes of the genetic markers in the one or more libraries, preferably for all samples and for all identified markers.

The method relates to the discovery, detection and genotyping of one or more genetic markers in one or more samples. In certain embodiments, the methods relates to presence/absence scoring of the genetic markers of interest. In certain embodiments the method relates to determination of (co-)dominant genotypes of one more more samples for one or more genetic markers. This may require normalisation of the observed number of marker- or marker allele sequences between samples.

In the first step (a) of the method, DNA is to be provided. This can be done by methods known in the art per se. The isolation of DNA is generally achieved using common methods in the art such as the collection of tissue from a member of the population, DNA extraction (for instance using the Q-Biogene fast DNA kit), quantification and normalisation to obtain equal amounts of DNA per sample. The DNA can be from a variety of sources (Genomic, RNA, cDNA, BAc, YAC etc.) and organisms (human, mammal, plant, microorganisms, etc.). The isolated DNA may be pooled.

The DNA is restricted in step (b) using at least one restriction endonuclease. Depending on the case, i.e. size of genome, more endonucleases can be used. In certain embodiments, 2 or more endonucleases can be used. For most genomes 2 endonucleases are sufficient and this is hence most preferred. In certain embodiments, especially for large or complex genomes, more endonucleases can be used. Preferably the endonuclease provides for relatively short restriction fragments in the order of 250-500 bp, but this is not essential. Typically, at least one frequent cutting endonuclease is

IDT Exhibit 1020.3880

preferred, i.e. endonucleases that have a 4 or 5 base pair recognition sequence. One such enzyme is MseI, but numerous others are commercially available and can be used. Also enzymes that cut outside their recognition sequence can be used (IIs type), or enzymes that provide blunt ended restriction fragments. A preferred combination uses one rare (6 and more base pair recognition sequence, for example EcoRI) and one frequent cutter.

After restriction of the pooled DNAs, or simultaneously therewith, adaptors are ligated to the restriction fragments to provide for adaptor-ligated restriction fragments. One or more different adaptors may be used, for instance two adaptors, one forward, one reverse adaptor. Alternatively one adaptor may be used for all fragments or sets of adaptors may be used that at the overhanging end of the adaptor contain permutations of nucleotides such as to provide for indexing linkers that may allow for a pre-selection step (Unrau et al., Gene, 1994, 145, 163-169). Alternatively, blunt ended adaptors can be used, in the case of blunt ended restriction fragments. Adaptor-ligation is well known in the art and is described *inter alia* in EP 534858.

One useful variant of the AFLP technology uses no selective nucleotides (i.c. +0/+0 primers) and is sometimes called linker PCR. As with Salsa PCR , the selection step is provided by the use of restriction enzymes, different restriction enzymes yields different subsets. This is sometimes also denoted as a pre-amplification wherein primers are used that are at least complementary to the adapters and optionally also to part of the remains of the recognition sequence of the restriction endonuclease. Pre-amplification may serve to (further) normalize the amount of DNA from each sample, or to increase the total amount of DNA to allow for multiple analysis (i.e. splitting up samples) and to enhance the signal-to-noise ratio. Pre-amplification may also be used to introduce tags that allow pooling prior to selective amplification. By the introduction of nucleotide tags (for instance 4 bp) at the 5'end of the primer, restriction fragments for a distinct sample can be tagged and at the end of the process can be retrieved by using the tag.

The adaptor-ligated restriction fragments are, after the optional pre-amplification, amplified in step (d) of the method of the invention with a pair of primers. One of the primers is complementary to at least part of the adaptor and may further be

IDT Exhibit 1020.3881

complementary to part of the remainder of the recognition sequence of the endonuclease and may further contain (randomly selected) selective nucleotides at its 3'-end, similar as is described in EP534858. Preferably the primers are capable of selectively hybridising under stringent hybridisation conditions. The selective amplification can also be performed with primers that carry a 5' tag to identify the origin of the sample, similar as above. The result is a library of (tagged) subsets of amplified adaptor-ligated restriction fragments.

The selectively amplified fragments in the libraries prepared from multiple samples can optionally be pooled at this point This may be useful in case markers are sought which are specific for certain groups of samples, such as those sharing certain phenotypic characteristics. Screening pooled samples is commonly referred to as bulked segregant analysis (BSA; Michelmore, Paran and Kesseli, 1991). In certain embodiments, pooling can also be performed before DNA extraction in the sampling stage, reducing the number of DNA preparations. Pooling of the DNA further serves to normalise the DNAs prior to PCR amplification to provide for a more equal representation in the libraries for sequencing.

The, optionally pooled, libraries of selectively amplified adaptor-ligated restriction fragments are now sequenced using high throughput sequencing technology.

The sequencing may in principle be conducted by any means known in the art, such as the dideoxy chain termination method (Sanger sequencing). It is however preferred and more advantageous that the sequencing is performed using high-throughput sequencing methods, such as the methods disclosed in WO 03/004690, WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007, and WO 2005/003375 (all in the name of 454 Life Sciences), by Seo et al. (2004) Proc. Natl. Acad. Sci. USA 101:5488-93, and technologies of Helios, Solexa, US Genomics, etcetera, which are herein incorporated by reference. It is most preferred that sequencing is performed using the apparatus and/or method disclosed in WO 03/004690, WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007, and WO 2005/003375 (all in the name of 454 Life Sciences), which are herein incorporated by reference. The technology described currently allows sequencing of up to 40 million bases in a single run and is 100 times faster and cheaper than competing technology based on Sanger sequencing and using currently available capillary electrophoresis instruments such

IDT Exhibit 1020.3882

as MegaBACE (GE Healthcare) or ABI3700(xl) (Applied Biosystems). This will increase with increasing read length per reaction and/or increasing numbers of parallel reactions. The sequencing technology roughly consists of 5 steps: 1) fragmentation of DNA and ligation of specific adaptor to create a library of single-stranded DNA (ssDNA); 2) annealing of ssDNA to beads, emulsification of the beads in water-in-oil microreactors and performing emulsion PCR to amplify the individual ssDNA molecules on beads; 3) selection of /enrichment for beads containing amplified ssDNA molecules on their surface 4) deposition of DNA carrying beads in a PicoTiterPlate®; and 5) simultaneous sequencing in 100,000 wells by generation of a pyrophosphate light signal.

In a preferred embodiment, the sequencing comprises the steps of:

(1) annealing sequencing-adaptor-ligated fragments to beads, each bead annealing with a single fragment;

(2) emulsifying the beads in water-in-oil micro reactors, each water-in-oil micro reactor comprising a single bead;

(3) performing emulsion PCR to amplify adaptor-ligated fragments on the surface of beads

(4) selecting / enriching beads containing amplified adaptor-ligated fragments;

(5) loading the beads in wells, each well comprising a single bead; and

(6) generating a pyrophosphate signal.

In the first step (1), the adaptors that are present in the adaptor ligated restriction fragments are annealed to the beads. As outlined herein before, the sequencing adaptor includes at least a "key" region for annealing to a bead, a sequencing primer region and a PCR primer region. In particular, the amplified adaptor-ligated restriction fragments now contain at one of the ends the following sequence 5'-Sequence primer binding site---Tag---PCR primer sequence-3', while at the other end a segment is present that may be as follows: 5'-Bead annealing sequence--- Tag---Adaptor specific sequence---restriction site-specific sequence (optional)---(randomly) selective sequence (optional)- 3'. It may be clear that the Sequence primer binding site and the Bead annealing sequence may be interchanged. This Bead annealing sequence can now be used for annealing the fragments to the bead, the bead carrying a nucleotide sequence to that end.

IDT Exhibit 1020.3883

Thus, adapted fragments are annealed to beads, each bead annealing with a single adapted fragment. To the pool of adapted fragments, beads are added in excess as to ensure annealing of one single adapted fragment per bead for the majority of the beads (Poisson distribution).

In a preferred embodiment, to increase the efficiency of the screening further, it is beneficial to amplify the PCR product directionally onto the bead for sequencing. This can be accomplished to perform the PCR with adaptor-tailed PCR primers of which one strand of the adaptor on the MseI (or other restriction enzyme) side is complementary to the oligonucleotide coupled to the sequence beads.

In a next step, the beads are emulsified in water-in-oil microreactors, each water-in-oil microreactor comprising a single bead. PCR reagents are present in the water-in-oil microreactors allowing a PCR reaction to take place within the microreactors. Subsequently, the microreactors are broken, and the beads comprising DNA (DNA positive beads) are enriched.

In a following step, the beads are loaded in wells, each well comprising a single bead. The wells are preferably part of a PicoTiter™Plate allowing for simultaneous sequencing of a large amount of fragments.

After addition of enzyme-carrying beads, the sequence of the fragments is determined using pyrosequencing. In successive steps, the PicoTiter™Plate and the beads as well as the enzyme beads therein are subjected to different deoxyribonucleotides in the presence of conventional sequencing reagents, and upon incorporation of a deoxyribonucleotide a light signal is generated which is recorded. Incorporation of the correct nucleotide will generate a pyrosequencing signal which can be detected.

Pyrosequencing itself is known in the art and described inter alia on www.biotagebio.com; www.pyrosequencing.com / section technology. The technology is further applied in e.g. WO 03/004690, WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007, and WO 2005/003375 (all in the name of 454 Life Sciences), which are herein incorporated by reference.

After sequencing, the sequences of the fragments that are directly obtained from the sequencing step may be trimmed, preferably *in silico*, to remove any bead annealing sequence, sequencing primer, adaptor or primer-related sequence information.

IDT Exhibit 1020.3884

Typically, the alignment or clustering is performed on sequence data that have been trimmed for any added adaptors/primer sequences i.e. using only the sequence data from the fragments that originate from the nucleic acid sample, together with the optional identifier tag.

Methods of alignment of sequences for comparison purposes are well known in the art. Various programs and alignment algorithms are described in: Smith and Waterman (1981) Adv. Appl. Math. 2:482; Needleman and Wunsch (1970) J. Mol. Biol. 48:443; Pearson and Lipman (1988) Proc. Natl. Acad. Sci. USA 85:2444; Higgins and Sharp (1988) Gene 73:237-244; Higgins and Sharp (1989) CABIOS 5:151-153; Corpet et al. (1988) Nucl. Acids Res. 16:10881-90; Huang et al. (1992) Computer Appl. in the Biosci. 8:155-65; and Pearson et al. (1994) Meth. Mol. Biol. 24:307-31, which are herein incorporated by reference . Altschul et al. (1994) Nature Genet. 6:119-29 (herein incorporated by reference) present a detailed consideration of sequence alignment methods and homology calculations.

The NCBI Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) is available from several sources, including the National Center for Biological Information (NCBI, Bethesda, Md.) and on the Internet, for use in connection with the sequence analysis programs blastp, blastn, blastx, tblastn and tblastx. It can be accessed at <http://www.ncbi.nlm.nih.gov/BLAST/>. A description of how to determine sequence identity using this program is available at <http://www.ncbi.nlm.nih.gov/BLAST/blast_help.html>. The database preferably comprises EST sequences, genomic sequences of the species of interest and/or the non-redundant sequence database of GenBank or similar sequence databases.

High throughput sequencing methods can be used as described in Shendure et al. *Science*, Vol 309, Issue 5741, 1728-1732. Examples thereof are microelectrophoretic sequencing, Hybridization sequencing/sequencing by hybridization (SBH), cyclic-array sequencing on amplified molecules, cyclic-array sequencing on single molecules, Non-cyclical, single-molecule, real-time methods, such as polymerase sequencing, exonuclease sequencing, nanopore sequencing.

Within the library the presence of a genetic marker and/or the genotype of the sample for a genetic marker can now be determined.

The method of the present invention can be used for the identification, detection of genotype determination AFLP markers, but also for the identification, detection and genotyping of SNP markers

IDT Exhibit 1020.3885

contained in constant bands.

To provide a solution to the problem of sampling variation which affects the accuracy of genotyping genetic markers by sequencing allelic (marker) fragments contained in a library of multiple fragments, the present inventors have also found that detection of AFLP markers via sequencing is preferably performed with sufficient redundancy (depth) to sample all amplified fragments at least once and accompanied by statistical means which address the issue of sampling variation in relation to the accuracy of the genotypes called. Furthermore, just as with AFLP scoring, in the context of a segregating population, the simultaneous scoring of the parent individuals in one experiment, will aid in determining the statistical threshold, because all possible alleles in the sample will be scored in either parent 1 or parent 2. Note that it is suggested to sample parent individuals with higher redundancy than individuals of segregating populations.

Thus, in certain embodiments, the redundancy of the tagged amplified adaptor-ligated restriction fragments is at least 6, preferably at least 7, more preferably at least 8 and most preferably at least 9. In certain embodiments, the sequence of each adaptor-ligated restriction fragment is determined at least 6, preferably at least 7, more preferably at least 8 and most preferably at least 9 fold. In certain embodiments, the redundancy is selected such, assuming a 50/50 overall chance of identifying the locus correctly as homozygous, that the chance of correct identification of the locus is more than 95%, 96%, 97%, 98%, 99%, 99.5%.

In certain embodiments, the number of samples can be varied between 1 and 100.000, this also largely depends on the size of the genome to be analysed and the number of selectively amplified fragments. Usually, the capacity of the sequencing technology employed provides the most limiting factor in this respect.

Brief description of the drawings

Figure 1A shows a fragment according to the present invention annealed onto a bead ('454 bead') and the sequence of primer used for pre-amplification of the two pepper lines. 'DNA fragment' denotes the fragment obtained after digestion with a restriction endonuclease, 'keygene adaptor' denotes an adaptor providing an annealing site for the (phosphorylated) oligonucleotide primers used to generate a library, 'KRS' denotes an identifier sequence (tag), '454 SEQ.

IDT Exhibit 1020.3886

Adaptor' denotes a sequencing adaptor, and '454 PCR adaptor' denotes an adaptor to allow for emulsion amplification of the DNA fragment. The PCR adaptor allows for annealing to the bead and for amplification and may contain a 3'-T overhang.

Figure 1B shows a schematic primer used in the complexity reduction step. Such a primer generally comprises a recognition site region indicated as (2), a constant region that may include a tag section indicated as (1) and one or more selective nucleotides in a selective region indicated as (3) at the 3'-end thereof.).

Figure 2 shows DNA concentration estimation using 2% agarose gel-electrophoresis. S1 denotes PSP11; S2 denote PI201234. 50, 100, 250 and 500 ng denotes respectively 50 ng, 100 ng, 250 ng and 500 ng to estimate DNA amounts of S1 and S2. Fig 2C and 2D show DNA concentration determination using Nanodrop spectrophotometry.

Figure 3 shows the results of intermediate quality assessments of example 3.

Figure 4 shows flow charts of the sequence data processing pipeline, i.e. the steps taken from the generation of the sequencing data to the identification of putative SNPs, SSRs and indels, via steps of the removal of known sequence information in Trimming & Tagging resulting in trimmed sequence data which are clustered and assembled to yield contigs and singletons (fragments that cannot be assembled in a contig) after which putative polymorphisms can be identified and assessed. Figure 4B further elaborates on the process of polymorphisms mining.

Figure 5: Multiple alignment "10037_CL989contig2" of pepper AFLP fragment sequences, containing a putative single nucleotide polymorphism (SNP). Note that the SNP (indicated by an the black arrow) is defined by an A allele present in both reads of sample 1 (PSP11), denoted by the presence of the MS1 tag in the name of the top two reads, and a G allele present in sample 2 (PI201234), denoted by the presence of the MS2 tag in the name of the bottom two reads. Read names are shown on the left. The consensus sequence of this multiple alignment is (5'- 3'):

TAACACGACTTTGAACAAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACA[A/G]TGTTGGTTTT GGTGCTAACTTCAACCCCACTACTGTTTTGCTCTATTTTTG.

Figure 6: Graphic representation of the probability of correct classification of the genotype based on the number of observed reads per locus.

IDT Exhibit 1020.3887

Examples

The method is exemplified as follows:

1) AFLP templates are prepared according to a modified protocol of Vos *et al.* which involves a heat-denaturation step for 20 min at 80 °C between the restriction and ligation steps. After incubation for 20 min at 80 °C, the restriction enzyme digest is cooled to room temperature and DNA ligase is added. The denaturation step leads to dissociation of the complementary strands of restriction fragments up to 120 bp such that no adaptors will be ligated to the ends. As a result, fragments smaller than 120 bp will not be amplified, hence size selection is achieved.

2) Pre-amplification reactions, if applicable, are performed as in conventional AFLP.

3) The last (selective) amplification step is performed using AFLP primers with unique identifier tags for every sample in the population/experiment, (using a unique 4 bp identifier sequence; KIS). The KIS are located at the 5'end of the selective AFLP primers. One additional selective nucleotide will be used in comparison with the number of selective bases used in conventional AFLP detection by electrophoresis, e.g. +4/+3 for an EcoRI/MseI fingerprint in pepper (gel detection +3/+3) and +4/+4 for and EcoRI/MseI fingerprint in maize (gel detection +4/+3). The number of selective nucleotides that are applied needs to be determined empirically; it may be so that the same number of selective nucleotides can be applied as used for gel detection. This number further depends on the number of samples included in the experiment, since the numbers of sequence traces is assumed to be fixed 200,000 at the current status of sequencing technology, but this may and probably will increase. Preferred starting point is to achieve 10-fold sampling of AFLP fragments per sample library.

4) The collection of samples prepared according to steps 1-4 is subjected to sequencing via 454 Life Sciences technology. This means that individual AFLP fragments are cloned on beads, PCR amplified and sequenced. An output of 200,000 sequences of 100 bp length is expected. For a collection of 100 samples, this equals an average of 2000 sequences traces/sample, traceable to sample nr. via the 5' tag.

5) Assuming the amplification of 100 AFLP fragments per PC when 1 additional selective nucleotide is used compared to the number used with gel detection, of which 90 percent are constant bands, the AFLP

IDT Exhibit 1020.3888

fragments are sampled with 20-fold average redundancy per fragment. However, since sequencing is non directional and most bands are > 200 bp, sequencing redundancy will be slightly over 10-fold for each fragment end.

6) All sequences are clustered per sample using the KRS tag. Given a 10-fold over sampling, this means that 200 different sequence traces are expected per sample, representing 200 x 100 bp = 20 kb sequence/sample. When 10 percent of these sequences are derived from AFLP markers (i.e. 1 allele is amplified and the other is absent in the PCR reaction), 90 percent (18 kb) of the sequences are derived from constant bands.

7) Two types of genetic markers are scored:

A) AFLP markers: these are sequences which are observed in some samples, but absent in others.

Inspection of the frequency of sequences in the collection of samples will reveal this category. Dominant scoring is performed depending on the presence/absence observation of these sequences in every sample. Reliable scoring of AFLP markers requires a statistical threshold to be set regarding the frequency with which other AFLP sequences are observed in the experiment. I.e. an AFLP marker can be scored as present (dominant) if the AFLP marker sequence is observed in the sample, but the reliability of the absent score depends on the (average) frequency of (constant) AFLP fragments. Statistical threshold levels are required such that presence/absence scoring is performed with preferably at least 99,5 % accuracy, depending on the acceptable level needed for the specific application. If a segregating population and its parents is analysed, these markers can possibly be scored co-dominantly as well by defining frequency categories of the marker sequences. The latter may actually be complicated by the influence of sampling variation of the AFLP marker which differs between samples.

B) (SN) polymorphisms in constant AFLP fragments.

This is the most interesting (and abundant) category of genetic markers. The essence is that SNP markers contained in the internal sequences of constant AFLP fragments are scored as co-dominant SNP markers. Again, this preferably requires applying a statistical threshold level for accurate calling of the presence or absence of an allele. A 10-fold sequencing redundancy of the fragment library is expected to be sufficient but a statistical analysis method is needed

IDT Exhibit 1020.3889

to determine accuracy of the SNP marker genotypes depending on the number each allele sequence is observed. The rationale is that when a constant band contains a SNP and one allele is observed e.g. 5 times while (the sequence containing the) other allele is not observed, it is highly likely that the sample is homozygous for the observed allele. Consequently, when both alleles are observed, the sample is scored heterozygous for the SNP marker, irrespective of their frequencies.

8) The result will be a genotyping table containing the genotypes of (co-)dominantly scored AFLP markers and co-dominantly scored SNPs, along with probabilities for correctness of the genotypes for all markers. Alternatively, a dataset is generated which contains genotypes which have surpassed the set statistical threshold level.

The approach assumes 10-fold over sampling of AFLP fragments per sample, yielding 18 kb of constant sequence/sample and 2 kb of AFLP marker sequences.

The numbers of genetic markers observed depends on the SNP rate in the germplasm investigated. Below, estimates of the numbers of genetic markers are provided at different germplasm SNP rates, when sampling 20 kb sequence. The average length of AFLP markers/fragments is assumed to be 200 bp:

Table 1. Expected numbers of genetic markers scored by sequencing AFLP fragments using

454 Life sciences technology assuming 10-fold over sampling, 200,000 sequence traces, 90 percent constant bands / 10 percent AFLP markers at various SNP rates.

SNP rate	AFLP markers (2 kb)	SNPs in constant bands (18 kb)*
1/ 250 bp	8	72
1/ 1000 bp	2	18
1/ 2000 bp	1	9
1/ 5000 bp	0,4	3,6

* As the AFLP fragments may be sequenced from both ends, a proportion of the observed SNP can be derived from the same loci.

It is important to note that the numbers provided in table 1

IDT Exhibit 1020.3890

are averages, which may differ between combinations of different primers. Analogous to conventional AFLP typing, identification of top primer combinations (PC) may yield higher numbers of markers per PC. In addition, the numbers presented in Table 1 may change depending on the required level of over sampling needed in order to reach the required accuracy level.

The calculation of the correct classification of the genotype is as follows:

$$P(correct)=P(aa) + P(AA) + P(Aa) * [1-0.5*exp(n-1))]$$

Wherein P(aa) is the fraction of the population with genotype aa (in the enclosed graph, fig 9, set at 0.25. P(AA) is the fraction of the population with genotype AA (set at 0.25. P(Aa)is the fraction of the population with genotype Aa (in fig 6 and table below, set at 0.5. n equals the number of individuals.

Table

n	P
1	0,5
2	0,75
3	0,875
4	0,9375
5	0,96875
6	0,984375
7	0,992188
8	0,996094
9	0,998047
10	0,999023

Example 1 PEPPER

DNA from the *Pepper* lines PSP-11 and PI201234 was used to generate AFLP product by use of AFLP *Keygene Recognition Site* specific primers. (These AFLP primers are essentially the same as conventional AFLP primers, e.g. described in EP 0 534 858, and will generally contain a *recognition site region*, a constant region and one or more selective nucleotides in a selective region.

From the pepper lines PSP-11 or PI201234 150ng of DNA was digested with the restriction endonucleases *EcoRI* (5U/reaction) and *MseI* (2U/reaction) for 1 hour at 37°C following by inactivation for 10 minutes at 80°C. The obtained restriction fragments were ligated with double-stranded synthetic oligonucleotide adapter, one end of witch is compatible with one or both of the ends of the EcoRI and/or

IDT Exhibit 1020.3891

MseI restriction fragments. The restriction ligation mixture was 10 times diluted and 5 microliter of each sample was pre-amplified (2) with EcoRI +1(A) and MseI +1(C) primers (set I). After amplification the quality of the pre-amplification product of the two pepper samples was checked on a 1% agarose gel. The preamplification products were 20 times diluted, followed by a KRSEcoRI +1(A) and KRSMseI +2(CA) AFLP pre-amplification. The KRS (identifier)sections are underlined and the selective nucleotides are in bold at the 3'-end in the primersequence SEQ ID 1-4 below. After amplification the quality of the pre-amplification product of the two pepper samples was checked on a 1% agarose gel and by an EcoRI +3(A) and MseI +3(C) (3) AFLP fingerprint (4). The pre-amplification products of the two pepper lines were separately purified on a QiagenPCR column (5). The concentration of the samples was measured on a NanoDrop® ND-1000 Spectrophotometer. A total of 5 micrograms PSP-11 and 5 micrograms PI201234 PCR products were mixed and sequenced.

Primer set I used for preamplification of PSP-11

E01LKRS1 5'-CGTCAGACTGCGTACCAATTC**A**-3' [SEQ ID 1]

M15KKRS1 5'-TGGTGATGAGTCCTGAGTAA**CA**-3' [SEQ ID 2]

Primer set II used for preamplification of PI201234

E01LKRS2 5'-CAAGAGACTGCGTACCAATTC**A**-3' [SEQ ID 3]

M15KKRS2 5'-AGCCGATGAGTCCTGAGTAA**CA**-3' [SEQ ID 4]

(1) EcoRI/MseI restriction ligation mixture

Restriction mix (40ul/sample)

DNA	6µl (±300ng)
ECoRI (5U)	0.1µl
MseI(2U)	0.05µl
5xRL	8µl
MQ	25.85µl
Totaal	40µl

Incubation during 1h. at 37°C

Addition of:

Ligation mix (10µl/sample)

10mM ATP	1µl
T4 DNA ligase	1µl
ECoRI adapt. (5pmol/µl)	1µl

IDT Exhibit 1020.3892

MseI adapt. . (50pmol/µl) 1µl
5xRL 2µl
MQ 4µl
Totaal 10µl
Incubation during 3h. at 37°C

EcoRI-adaptor

91M35/91M36: *-CTCGTAGACTGCGTACC :91M35 [SEQ ID 5]
± bio CATCTGACGCATGGTTAA :91M36 [SEQ ID 6]

MseI-adaptor

92A18/92A19: 5-GACGATGAGTCCTGAG-3 :92A18 [SEQ ID 7]
3-TACTCAGGACTCAT-5 :92A19 [SEQ ID 8]

(2) Pre-amplification

Preamplification (A/C):

RL-mix (10x) 5µl
EcoRI-pr E01L(50ng/ul) 0.6µl
MseI-pr M02K(50ng/ul) 0.6µl
dNTPs (25mM) 0.16µl
Taq.pol.(5U) 0.08µl
10XPCR 2.0µl
MQ 11.56µl
Total 20µl/reaction

Pre-amplification thermal profile

Selective pre amplification was done in a reaction volume of 50µl. The PCR was performed in a PE GeneAmp PCR System 9700 and a **20** cycle profile was started with a 94°C denaturation step for 30 seconds, followed by an annealing step of 56°C for 60 seconds and an extension step of 72°C for 60 seconds.

EcoRI +1(A)[1]
E01 L 92R11: 5-AGACTGCGTACCAATTCA-3 [SEQ ID 9]
MseI +1(C)[1]
M02k 93E42: 5-GATGAGTCCTGAGTAAC-3 [SEQ ID 10]

Preamplification A/CA:

PA+1/+1-mix (20x) :5µl
EcoRI-pr :1.5µl
MseI-pr. :1.5µl
dNTPs (25mM) :0.4µl
Taq.pol.(5U) :0.2µl
10XPCR :5µl
MQ :36.3µl

IDT Exhibit 1020.3893

Total :50µl

Selective pre amplification was done in a reaction volume of 50µl. The PCR was performed in a PE GeneAmp PCR System 9700 and a **30** cycle profile was started with a 94°C denaturation step for 30 seconds, followed by an annealing step of 56°C for 60 seconds and an extension step of 72°C for 60 seconds.

(3) KRSEcoRI +1(A)and KRSMseI +2(CA)^2

05F212	E01LKRS1	CGTCAGACTGCGTACCAATTC**A**	-3' [SEQ ID 11]
05F213	E01LKRS2	CAAGAGACTGCGTACCAATTC**A**	-3' [SEQ ID 12]
05F214	M15KKRS1	TGGTGATGAGTCCTGAGTAA**CA**	-3' [SEQ ID 13]
05F215	M15KKRS2	AGCCGATGAGTCCTGAGTAA**CA**	-3' [SEQ ID 14]

selective nucleotides in bold and tags (KRS) underlined

Sample PSP11 : E01LKR**S1**/M15KKR**S1**

Sample PI120234 : E01LKR**S2**/M15KKR**S2**

(4) AFLP protocol

Selective amplification was done in a reaction volume of 20µl. The PCR was performed in a PE GeneAmp PCR System 9700. A **13** cycle profile was started with a 94°C denaturation step for 30 seconds, followed by an annealing step of 65°C for 30 seconds, with a touchdown phase in witch the annealing temperature was lowered 0.7°C in each cycle, and an extension step of 72°C for 60 seconds. This profile was followed by a 23 cycle profile with a 94°C denaturation step for 30 seconds, followed by an annealing step of 56°C for 30 seconds and an extension step of 72°C for 60 seconds.

EcoRI +3(**AAC**) and MseI +3(**CAG**)

E32 92S02: 5-GACTGCGTACCAATTC**AAC**-3 [SEQ ID 15]

M49 92G23: 5-GATGAGTCCTGAGTAA**CAG**-3 [SEQ ID 16]

(5) Qiagen column

The AFLP product was purified by using the QIAquick PCR Purification Kit (QIAGEN) following the QIAquick® Spin Handbook 07/2002 page 18 and the concentration was measured with a NanoDrop® ND-1000 Spectrophotometer. A total of 5µg of +1/+2 PSP-11 AFLP product and 5µg of +1/+2 PI201234 AFLP product was put together and solved in 23.3µl TE. Finally a mixture with a concentration of 430ng/µl +1/+2 AFLP product was obtained.

IDT Exhibit 1020.3894

Sequence library preparation and high-throughput sequencing

Mixed amplification products from both pepper lines were subjected to high-throughput sequencing using 454 Life Sciences sequencing technology as described by Margulies et al., (Margulies et al., *Nature* 437, pp. 376-380 and Online Supplements). Specifically, the AFLP PCR products were first end-polished and subsequently ligated to adaptors to facilitate emulsion-PCR amplification and subsequent fragment sequencing as described by Margulies and co-workers. 454 adaptor sequences, emulsion PCR primers, sequence-primers and sequence run conditions were all as described by Margulies and co-workers. The linear order of functional elements in an emulsion-PCR fragment amplified on Sepharose beads in the 454 sequencing process was as follows as exemplified in figure 1A:

454 PCR adaptor - 454 sequence adaptor - 4 bp AFLP primer tag 1 - AFLP primer sequence 1 including selective nucleotide(s) - AFLP fragment internal sequence - AFLP primer sequence 2 including selective nucleotide(s), 4 bp AFLP primers tag 2 - 454 sequence adaptor - 454 PCR adaptor - Sepharose bead

Two high-throughput 454 sequence runs were performed by 454 Life Sciences (Branford, CT; United States of America).

454 sequence run data-processing.

Sequence data resulting from one 454 sequence run were processed using a bio-informatics pipeline (Keygene N.V.). Specifically, raw 454 basecalled sequence reads were converted in FASTA format and inspected for the presence of tagged AFLP adaptor sequences using a BLAST algorithm. Upon high-confidence matches to the known tagged AFLP primer sequences, sequences were trimmed, restriction endonuclease sites restored and assigned the appropriate tags (sample 1 EcoRI (ES1), sample 1 MseI (MS1), sample 2 EcoRI (ES2) or sample 2 MseI (MS2), respectively). Next, all trimmed sequences larger than 33 bases were clustered using a megaBLAST procedure based on overall sequence homologies. Next, clusters were assembled into one or more contigs and/or singletons per cluster, using a CAP3 multiple alignment algorithm. Contigs containing more than one sequence were inspected for the sequence mismatches, representing putative polymorphisms. Sequence mismatches were assigned quality scores based on the following criteria:

* the numbers of reads in a contig

IDT Exhibit 1020.3895

* the observed allele distribution

The above two criteria form the basis for the so called Q score assigned to each putative SNP/indel. Q scores range from 0 to 1; a Q score of 0.3 can only be reached in case both alleles are observed at least twice.

* location in homopolymers of a certain length (adjustable; default setting to avoid polymorphism located in homopolymers of 3 bases or longer).

* number of contigs in cluster.

* distance to nearest neighboring sequence mismatches (adjustable; important for certain types of genotyping assays probing flanking sequences)

* the level of association of observed alleles with sample 1 or sample 2; in case of a consistent, perfect association between the alleles of a putative polymorphism and samples 1 and 2, the polymorphism (SNP) is indicated as an "elite" putative polymorphism (SNP). An elite polymorphism is thought to have a high probability of being located in a unique or low-copy genome sequence in case two homozygous lines have been used in the discovery process. Conversely, a weak association of a polymorphism with sample origin bears a high risk of having discovered false polymorphisms arising from alignment of non-allelic sequences in a contig.

Sequences containing SSR motifs were identified using the MISA search tool (MIcroSAtellelite identification tool; available from http://pgrc.ipk-gatersleben.de/misa/

Overall statistics of the run is shown in the Table below.

Table. Overall statistics of a 454 sequence run for SNP discovery in pepper.

Enzyme combination	Run
Trimming	
All reads	254308
Fault	5293 (2 %)
Correct	249015 (98%)
Concatamers	2156 (8.5 %)
Mixed tags	1120 (0.4 %)
Correct reads	
Trimmed one end	240817 (97%)
Trimmed both ends	8198 (3 %)

IDT Exhibit 1020.3896

Number of reads sample 1	136990 (55%)
Number of reads sample 2	112025 (45 %)
Clustering	
Number of contigs	21918
Reads in contigs	190861
Average number reads per contig	8.7
SNP mining	
SNPs with Q score ≥ 0.3 *	1483
Indel with Q score ≥ 0.3 *	3300
SSR mining	
Total number of SSR motifs identified	359
Number of reads containing one or more SSR motifs	353
Number of SSR motif with unit size 1 (homopolymer)	0
Number of SSR motif with unit size 2	102
Number of SSR motif with unit size 3	240
Number of SSR motif with unit size 4	17

* SNP / indel mining criteria were as follows:

No neighbouring polymorphisms with Q score larger than 0.1 within 12 bases on either side, not present in homopolymers of 3 or more bases. Mining criteria did not take into account consistent association with sample 1 and 2, i.e. the SNPs and indels are not necessarily elite putative SNPs/indels

An example of a multiple alignment containing an elite putative single nucleotide polymorphism is shown in Figure 5.

Example 2 : Maize

DNA from the *Maize* lines B73 and M017 was used to generate AFLP product by use of AFLP *Keygene Recognition Site* specific primers. (These AFLP primers are essentially the same as conventional AFLP primers, e.g. described in EP 0 534 858, and will generally contain a *recognition site region*, a constant region and one or more selective nucleotides at the 3'-end thereof.).

DNA from the pepper lines B73 or M017 was digested with the restriction endonucleases *TaqI* (5U/reaction) for 1 hour at 65°C and *MseI* (2U/reaction) for 1 hour at 37°C following by inactivation for 10 minutes at 80°C. The obtained restriction fragments were ligated

IDT Exhibit 1020.3897

with double-stranded synthetic oligonucleotide adapter, one end of witch is compatible with one or both of the ends of the TaqI and/or MseI restriction fragments.

AFLP preamplification reactions (20µl/reaction) with +1/+1 AFLP primers were performed on 10 times diluted restriction-ligation mixture. PCR profile:20*(30 s at 94°C + 60 s at 56°C + 120 s at 72°C). Additional AFLP reactions (50µl/reaction) with different +2 *TaqI* and *MseI* AFLP Keygene Recognition Site primers (Table below, tags are in bold, selective nucleotides are underlined.) were performed on 20 times diluted +1/+1 TaqI/MseI AFLP preamplification product. PCR profile: 30*(30 s at 94°C + 60 s at 56°C + 120 s at 72°C). The AFLP product was purified by using the QIAquick PCR Purification Kit (QIAGEN) following the QIAquick® Spin Handbook 07/2002 page 18 and the concentration was measured with a NanoDrop® ND-1000 Spectrophotometer. A total of 1.25µg of each different B73 +2/+2 AFLP product and 1.25µg of each different M017 +2/+2 AFLP product was put together and solved in 30µl TE. Finally a mixture with a concentration of 333ng/µl +2/+2 AFLP product was obtained.

Table

SEQ ID	PCR Primer	Primer sequence	Maize	AFLP Reaction
[SEQ ID 17]	05G360	**ACGT**GTAGACTGCGTACCGAAA	B73	1
[SEQ ID 18]	05G368	**ACGT**GATGAGTCCTGAGTAACA	B73	1
[SEQ ID 19]	05G362	**CGTA**GTAGACTGCGTACCGAAC	B73	2
[SEQ ID 20]	05G370	**CGTA**GATGAGTCCTGAGTAACA	B73	2
[SEQ ID 21]	05G364	**GTAC**GTAGACTGCGTACCGAAG	B73	3
[SEQ ID 22]	05G372	**GTAC**GATGAGTCCTGAGTAACA	B73	3
[SEQ ID 23]	05G366	**TACG**GTAGACTGCGTACCGAAT	B73	4
[SEQ ID 24]	05G374	**TACG**GATGAGTCCTGAGTAACA	B73	4
[SEQ ID 25]	05G361	**AGTC**GTAGACTGCGTACCGAAA	M017	5
[SEQ ID 26]	05G369	**AGTC**GATGAGTCCTGAGTAACA	M017	5
[SEQ ID 27]	05G363	**CATG**GTAGACTGCGTACCGAAC	M017	6
[SEQ ID 28]	05G371	**CATG**GATGAGTCCTGAGTAACA	M017	6
[SEQ ID 29]	05G365	**GAGC**GTAGACTGCGTACCGAAG	M017	7
[SEQ ID 30]	05G373	**GAGC**GATGAGTCCTGAGTAACA	M017	7
[SEQ ID 31]	05G367	**TGAT**GTAGACTGCGTACCGAAT	M017	8
[SEQ ID 32]	05G375	**TGAT**GATGAGTCCTGAGTAACA	M017	8

IDT Exhibit 1020.3898

Finally the 4 P1-samples and the 4 P2-samples were pooled and concentrated. A total amount of 25µl of DNA product and a final concentration of 400ng/ul (total of 10µg) was obtained. Intermediate quality assessments are given in Figure 3.

SEQUENCING BY 454

Pepper and maize AFLP fragment samples as prepared as described hereinbefore were processed by 454 Life Sciences as described (Margulies et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437 (7057):376-80. Epub July 31, 2005).

DATA PROCESSING

Processing Pipeline:

Input data

raw sequence data were received for each run:

- 200,000 - 400,000 reads
- base calling quality scores

Trimming and tagging

These sequence data are analyzed for the presence of Keygene Recognition Sites (KRS) at the beginning and end of the read. These KRS sequences consist of both AFLP-adaptor and sample label sequence and are specific for a certain AFLP primer combination on a certain sample. The KRS sequences are identified by BLAST and trimmed and the restriction sites are restored. Reads are marked with a tag for identification of the KRS origin. Trimmed sequences are selected on length (minimum of 33 nt) to participate in further processing.

Clustering and assembly

A *MegaBlast* analysis is performed on all size-selected, trimmed reads to obtain clusters of homologous sequences. Consecutively all clusters are assembled with *CAP3* to result in assembled contigs. From both steps unique sequence reads are identified that do not match any other reads. These reads are marked as singletons.

The processing pipeline carrying out the steps described herein before is shown in Figure 4A

Polymorphism mining and quality assessment

The resulting contigs from the assembly analysis form the basis of polymorphism detection. Each 'mismatch' in the alignment of each

IDT Exhibit 1020.3899

cluster is a potential polymorphism. Selection criteria are defined to obtain a quality score:

- number of reads per contig
- frequency of 'alleles' per sample
- occurrence of homopolymer sequence
- occurrence of neighbouring polymorphisms

SNPs and indels with a quality score above the threshold are identified as putative polymorphisms. For SSR mining we use the MISA (MIcroSAtellite identification) tool (http://pgrc.ipk-gatersleben.de/misa). This tool identifies di-, tri-, tetranucleotide and compound SSR motifs with predefined criteria and summarizes occurrences of these SSRs.

The polymorphism mining and quality assignment process is shown in Figure 4B

RESULTS

The table below summarizes the results of the combined analysis of sequences obtained from 2 454 sequence runs for the combined pepper samples and 2 runs for the combined maize samples.

	Pepper	Maize
Total number of reads	457178	492145
Number of trimmed reads	399623	411008
Number singletons	105253	313280
Number of contigs	31863	14588
Number of reads in contigs	294370	97728
Total number of sequences containing SSRs	611	202
Number of different SSR-containing sequences	104	65
Number of different SSR motifs (di, tri, tetra and compound)	49	40
Number SNPs with Q score ≥ 0.3 *	1636	782
Number of indels *	4090	943

* both with selection against neighboring SNPs, at least 12 bp flanking sequence and not occurring in homopolymer sequences larger than 3 nucleotides.

Example 3. SNP validation by PCR amplification and Sanger sequencing

IDT Exhibit 1020.3900

In order to validate the putative A/G SNP identified in example 1, a sequence tagged site (STS) assay for this SNP was designed using flanking PCR primers. PCR primer sequences were as follows:

Primer_1.2f: 5'- AAACCCAAACTCCCCCAATC-3', [SEQ ID 33] and

Primer_1.2r:5'- AGCGGATAACAATTTCACACAGGACATCAGTAGTCACACTGGTA CAAAAATAGAGCAAAACAGTAGTG -3' [SEQ ID 34]

Note that primer 1.2r contained an M13 sequence primer binding site and length stuffer at its 5 prime end. PCR amplification was carried out using +A/+CA AFLP amplification products of PSP11 and PI210234 prepared as described in example 4 as template. PCR conditions were as follows:

For 1 PCR reaction the following components were mixed:

5 µl 1/10 diluted AFLP mixture (app. 10 ng/µl)
5 µl 1pmol/µl primer 1.2f (diluted directly from a 500 µM stock)
5 µl 1pmol/µl primer 1.2r (diluted directly from a 500 µM stock)

5 µl PCR mix
- 2 µl 10 x PCR buffer
- 1 µl 5 mM dNTPs
- 1.5 µl 25 mM $MgCl_2$
- 0.5 µl H_2O

5 µl Enzyme mix
- 0.5 µl 10 x PCR buffer (Applied Biosystems)
- 0.1 µl 5U/µl AmpliTaq DNA polymerase (Applied Biosystems)
- 4.4 µl H_2O

The following PCR profile was used:

Cycle 1	2';	94° C
Cycle 2-34	20'';	94° C
	30'';	56° C
	2'30'';	72° C
Cycle 35	7';	72° C
	∞;	4° C

PCR products were cloned into vector pCR2.1 (TA Cloning kit ; Invitrogen) using the TA Cloning method and transformed into INVαF' competent *E. coli* cells. Transformants were subjected to blue/white screening. Three independent white transformants each for PSP11 and PI-201234 were selected and grown O/N in liquid selective medium for plasmid isolation.

IDT Exhibit 1020.3901

Plasmids were isolated using the QIAprep Spin Miniprep kit (QIAGEN). Subsequently, the inserts of these plasmids were sequenced according to the protocol below and resolved on the MegaBACE 1000 (Amersham). Obtained sequences were inspected on the presence of the SNP allele. Two independent plasmids containing the PI-201234 insert and 1 plasmid containing the PSP11 insert contained the expected consensus sequence flanking the SNP. Sequence derived from the PSP11 fragment contained the expected A (underlined) allele and sequence derived from PI-201234 fragment contained the expected G allele (double underlined):

PSP11 (sequence 1): (5'-3')

AAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAATGTTGGTTTTGGTGCTAACTTCAA CCCCACTACTGTTTTGCTCTATTTTTGT [SEQ ID 35]

PI-201234 (sequence 1): (5'-3')

AAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAGTGTTGGTTTTGGTGCTAACTTCAA CCCCACTACTGTTTTGCTCTATTTTTG [SEQ ID 36]

PI-201234 (sequence 2): (5'-3')

AAACCCAAACTCCCCCAATCGATTTCAAACCTAGAACAGTGTTGGTTTTGGTGCTAACTTCAA CCCCACTACTGTTTTGCTCTATTTTTG [SEQ ID 37]

This result indicates that the putative pepper A/G SNP represents a true genetic polymorphism detectable using the designed STS assay.

REFERENCES

1.Zabeau,M. and Vos,P. (1993) Selective restriction fragment amplification; a general method for DNA fingerprinting. EP 0534858-A1, B1, B2; US patent 6045994.

2.Vos,P., Hogers,R., Bleeker,M., Reijans,M., van de Lee,T., Hornes,M., Frijters,A., Pot,J., Peleman,J., Kuiper,M. *et al.* (1995) AFLP: a new technique for DNA fingerprinting. *Nucl. Acids Res.*, **21**, 4407-4414.

3. M. van der Meulen, J. Buntjer, M. J. T. van Eijk, P. Vos, and R. van Schaik. (2002). Highly automated AFLP® fingerprint analysis on the MegaBACE capillary sequencer. *Plant, Animal and Microbial Genome X,* San Diego, CA, January 12-16, P228, pp. 135.

4. Margulies *et al.*, 2005. Genome sequencing in microfabricated

IDT Exhibit 1020.3902

high-density picolitre reactions. *Nature advanced online publication 03959*, August 1.

5. R.W. Michelmore, I. Paran, and R.V. Kesseli. (1991).Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc. Natl. Acad. Sci USA 88(21):9828-32.

6.Shendure *et al.*, 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Scienceexpress Report*, August 4.

IDT Exhibit 1020.3903

Claims

1. Method for the high throughput discovery, detection and genotyping of one or more genetic markers in one or more samples, comprising the steps of:

(a) providing DNA from one or more samples;

(b) restricting the DNA with at least one restriction endonuclease to produce restriction fragments;

(c) ligating adaptors to the restriction fragments to produce adaptor-ligated restriction fragments;

(d) optionally, amplifying the adaptor-ligated restriction fragments with a primer pair that is complementary to the adaptors to produce pre-amplified adaptor-ligated restriction fragments;

(e) amplifying the (optionally pre-amplified) adaptor-ligated restriction fragments with a primer pair, wherein at least one of the primers contains an identifier tag at the 5' end of the primer to produce a library of tagged amplified subsets of adaptor-ligated restriction fragments for each sample;

(f) optionally, pooling the libraries;

(g) sequencing the libraries using high throughput sequencing technology;

(h) clustering the sequences per library, using the identifier tag;

(i) identify genetic markers within the library and/or between libraries

(j) determine (co-)dominant genotypes of the genetic markers in one or more libraries.

2. Method according to claim 1, wherein the genetic marker is an AFLP marker or an SNP marker.

3. Method according to claim 1 or 2, wherein sequencing is based on sequencing by synthesis, preferably pyrosequencing.

4. Method according to claims 1-3, wherein sequencing is performed on a solid support such as a bead.

5. Method according to claims 1-4, wherein the sequencing comprises the steps of:

- annealing amplified adaptor-ligated restriction fragments to beads, each bead annealing with a single adaptor-ligated fragment;

IDT Exhibit 1020.3904

- emulsifying the beads in water-in-oil microreactors, each water-in-oil microreactor comprising a single bead;
- performing emulsion PCR to amplify the adaptor -ligated restriction fragments on the surface of the beads;
- loading the beads in wells, each well comprising a single bead; and
- generating a pyrophosphate signal.

6. Method according to claims 1-4, wherein the average redundancy of the tagged amplified adaptor-ligated restriction fragments is at least 6, preferably at least 7, more preferably at least 8 and most preferably at least 9.

7. Method according to claims 1-5, wherein the sequence of each adaptor-ligated restriction fragment is determined at least 6, preferably at least 7, more preferably at least 8 and most preferably at least 9 fold.

8. Method according to claims 1-6, wherein between endonuclease restriction and adaptor ligation a size selection is performed by a denaturation step.

9. Method according to claims 1-8, wherein the DNA is selected from the group consisting of genomic DNA, RNA, cDNA, BACs, YACs, whole-genome amplified DNA, PCR product.

10. Method according to claims 1-9, wherein the adaptor is a double stranded synthetic oligonucleotide adaptor having one end that is compatible with one or both ends of the restriction fragments.

11. Method according to claims 1-10, wherein the DNA is restricted with two, preferably three, or more restriction endonucleases.

12. Method according to claims 1-11, wherein the DNA is restricted with two restriction endonucleases.

13. Method according to claims 1-2, wherein at least one of the restriction endonucleases is a rare cutter.

14. Method according to claims 1-13, wherein at least one of the restriction endonucleases is a frequent cutter.

15. Method according to claims 1-14, wherein the primer contains from one to 10 (preferably randomly selected from amongst A, C, T or G) selective nucleotides, more preferably from one to 5 nucleotides.

16. Method according to any of the previous claims wherein the DNA is restricted using a combination of three or more restriction

IDT Exhibit 1020.3905

endonucleases.

17. Use of the method as defined in any of the above claims for co-dominant scoring of AFLP and/or SNP marker sequences.

18. Use of high throughput sequencing methodology for the detection of polymorphisms in of the method as defined in any of the above method claims, for genotyping purposes including genetic mapping, QTL mapping, fine mapping genes/traits, linkage disequilibrium (LD) mapping, marker-assisted back-crossing, genetic distance analysis, discovery of markers linked to traits or phenotypes, diagnostic genotyping of patient samples etc.

IDT Exhibit 1020.3906

Figure 1

Primer set I used for preamplification of PSP-11

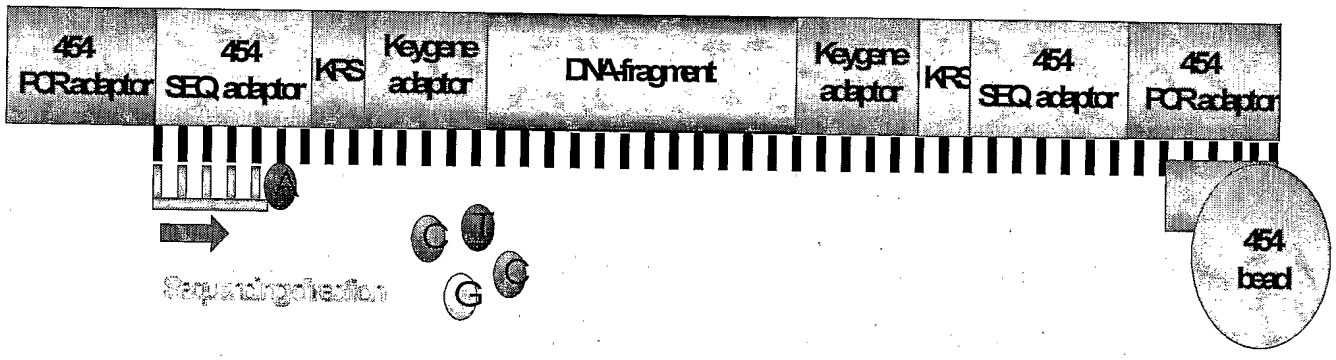
E01LKRS1 5'-CGTCAGACTGCGTACCAATTCA-3'

M15KKRS1 5'-TGGTGATGAGTCCTGAGTAACA-3'

Primer set II used for preamplification of PI201234

E01LKRS2 5'-CAAGAGACTGCGTACCAATTCA-3'

M15KKRS2 5'-AGCCGATGAGTCCTGAGTAACA-3'



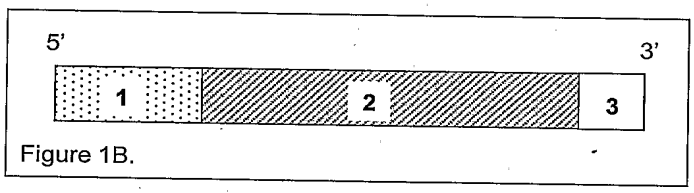


Figure 1B.

IDT Exhibit 1020.3907

Figure 2

DNA quality control on a 1% agarose gel

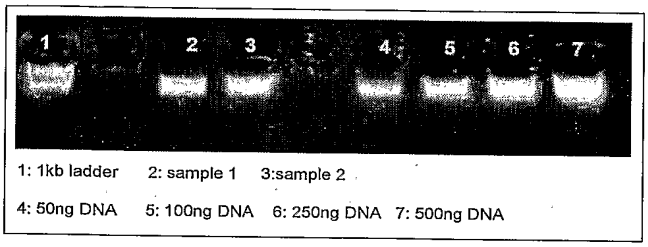


Figure 2A. Short gel electrophoresis

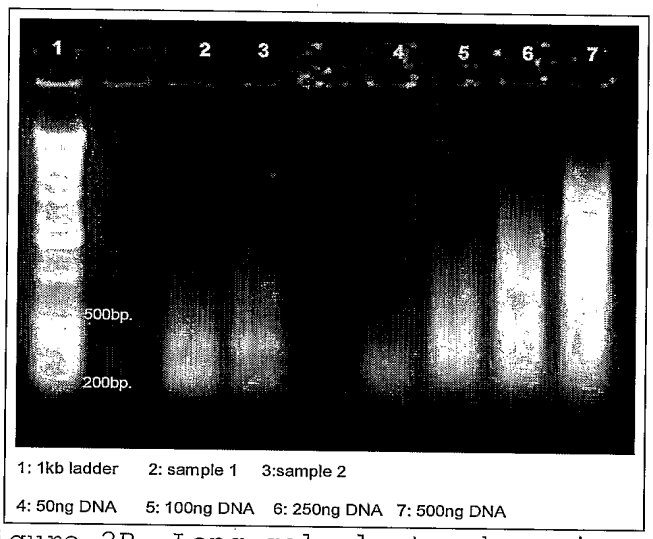


Figure 2B. Long gel electrophoresis

DNA concentration measured with the Nanodrop

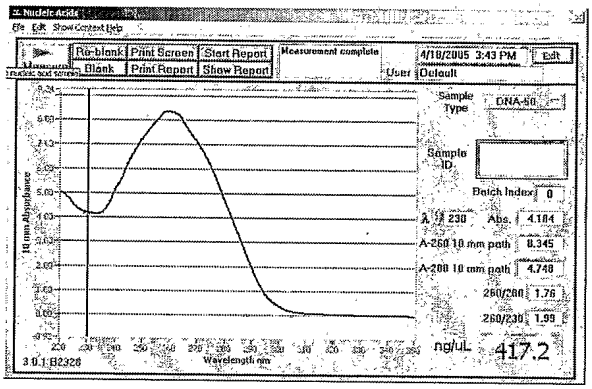


Figure 2C. Concentration sample 1.

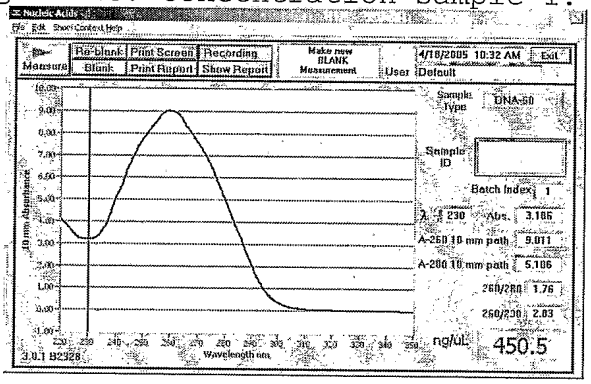


Figure 2D Concentration sample 2

IDT Exhibit 1020.3908

Figure 3

DNA quality control on a 1% agarose gel

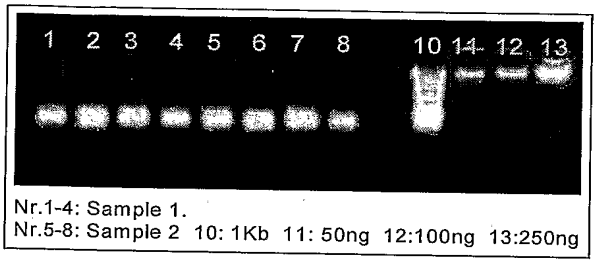


Figure 3A. Short gel electrophoresis

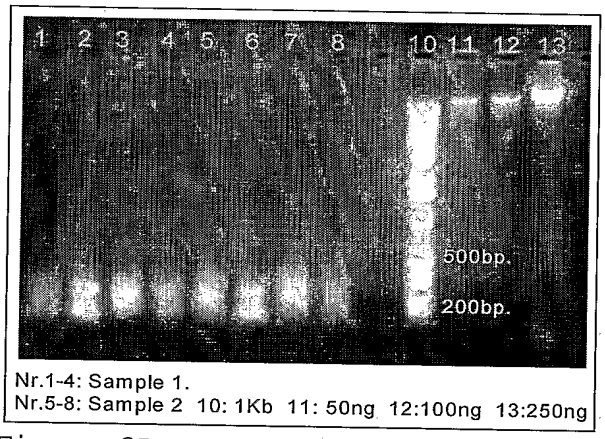


Figure 3B. Long gel electrophoresis

DNA concentrations measured on the Nanodrop

Nr.	Sample ID	ng/uL	A260	260/280	260/230	Constant
1	P1.1	22.61	0.452	1.5	1.81	50
2	P1.2	19.08	0.382	1.67	2.49	50
3	P1.3	18.05	0.361	1.63	2.35	50
4	P1.4	15.19	0.304	1.71	2.1	50

Nr.	Sample ID	ng/uL	A260	260/280	260/230	Constant
5	P2.1	17.5	0.35	1.66	2.01	50
6	P2.2	16.67	0.333	1.96	2	50
7	P2.3	22.03	0.441	1.81	2.28	50
8	P2.4	9.8	0.196	1.78	1.98	50

IDT Exhibit 1020.3909

Figure 4A. Sequence data processing pipeline

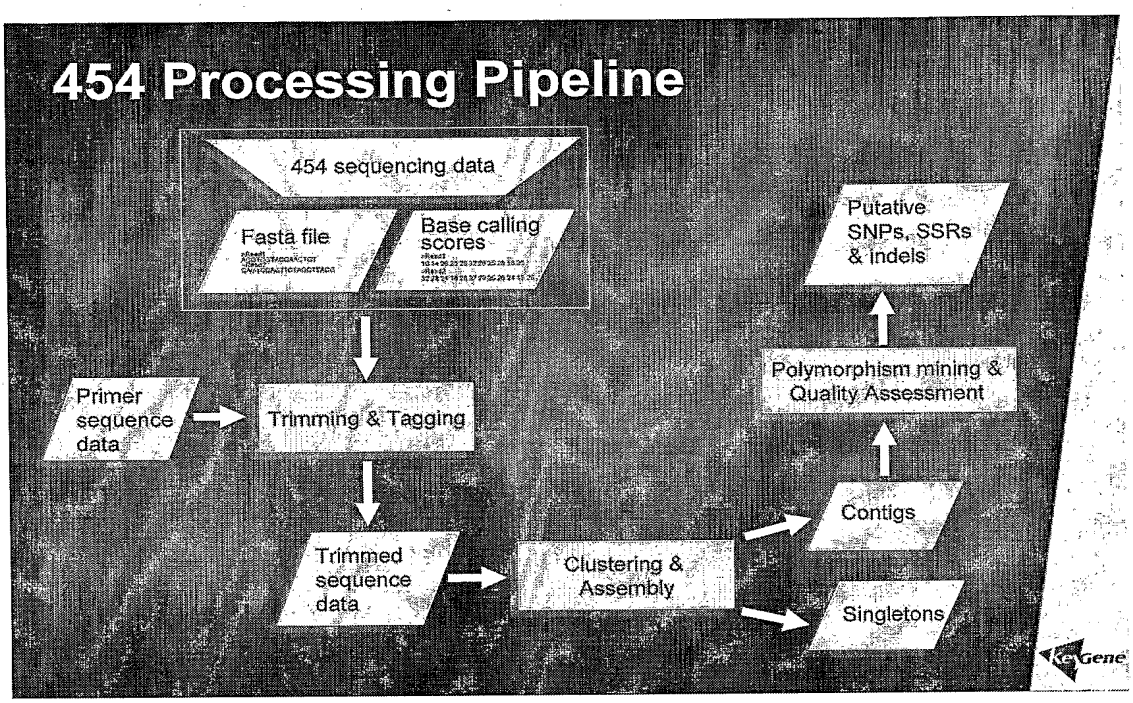
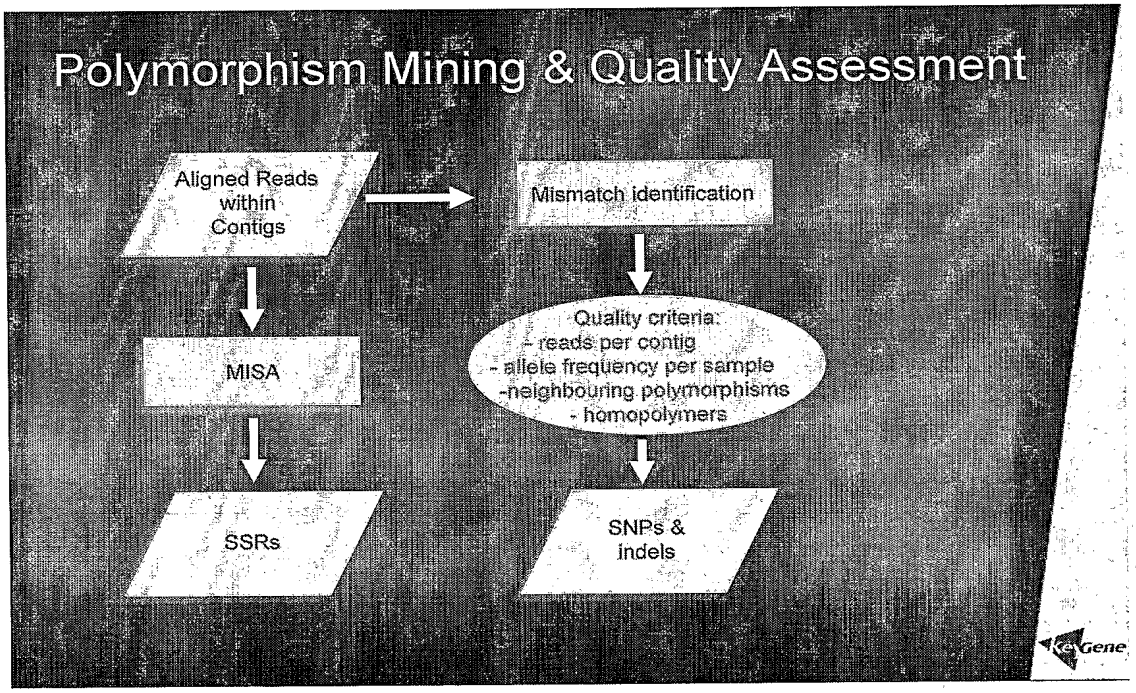
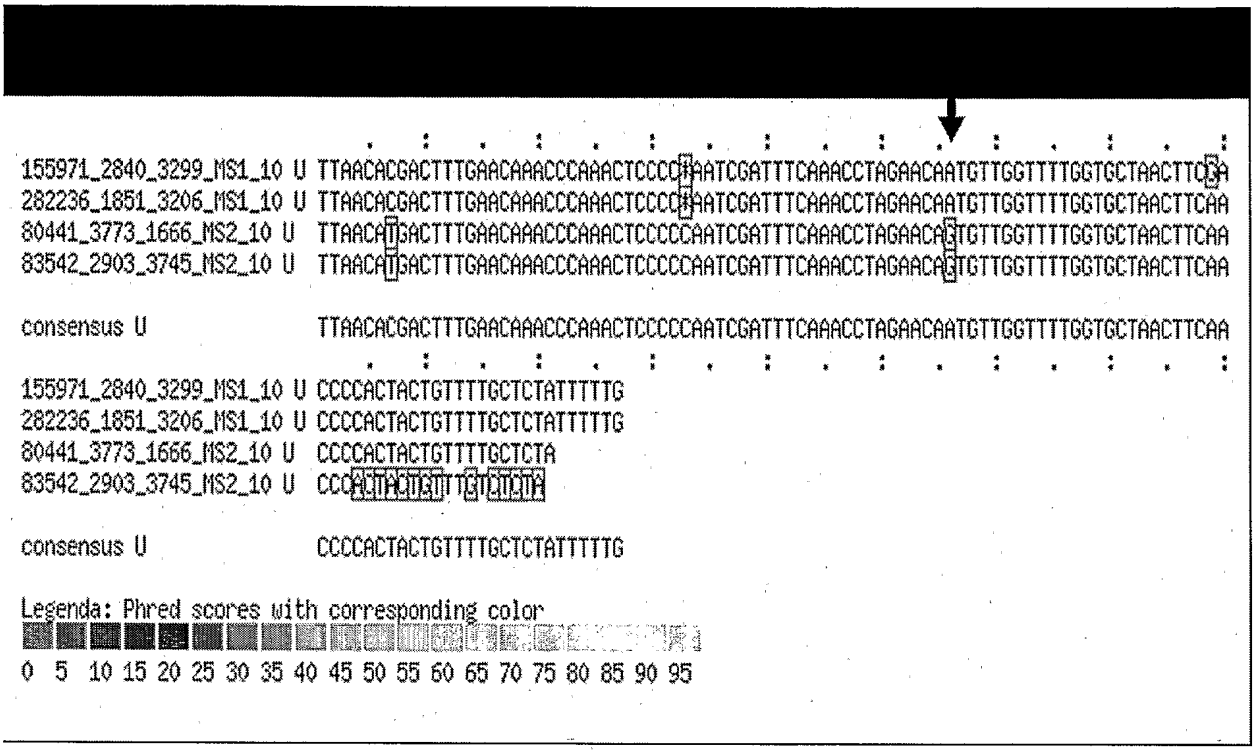


Figure 4B. Polymorphism mining and quality assignment process



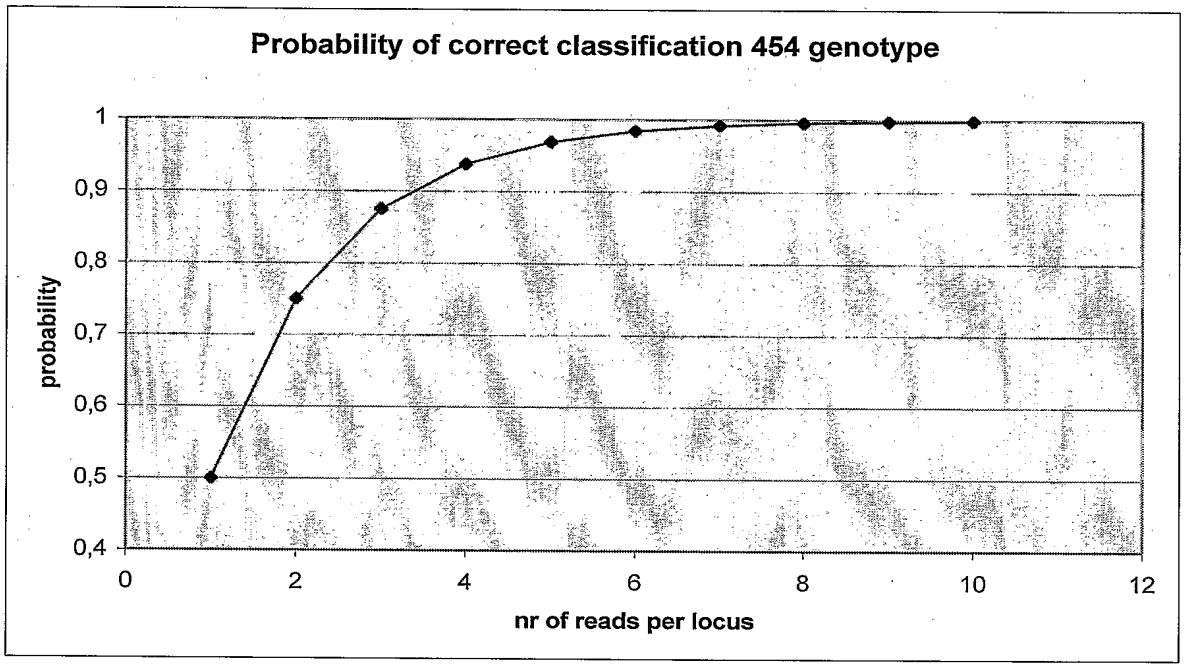
IDT Exhibit 1020.3910

FIG 5



IDT Exhibit 1020.3911

FIG 6



IDT Exhibit 1020.3912

SEQUENCE LISTING

<110> Keygene NV

<120> METHOD FOR HIGH-THROUGHPUT AFLP-BASED POLYMORPHISM DETECTION.

<130> P27959PC00

<150> us 60/752,590
<151> 2005-12-22

<160> 37

<170> PatentIn version 3.3

<210> 1
<211> 22
<212> DNA
<213> Artificial

<220>
<223> Primer

<400> 1
cgtcagactg cgtaccaatt ca 22

<210> 2
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 2
tggtgatgag tcctgagtaa ca 22

<210> 3
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 3
caagagactg cgtaccaatt ca 22

<210> 4
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 4
agccgatgag tcctgagtaa ca 22

<210> 5
<211> 17
<212> DNA
<213> Artificial

IDT Exhibit 1020.3913

<220>
<223> adaptor

<400> 5
ctcgtagact gcgtacc 17

<210> 6
<211> 18
<212> DNA
<213> Artificial

<220>
<223> adaptor

<400> 6
aattggtacg cagtctac 18

<210> 7
<211> 16
<212> DNA
<213> Artificial

<220>
<223> adaptor

<400> 7
gacgatgagt cctgag 16

<210> 8
<211> 14
<212> DNA
<213> Artificial

<220>
<223> adaptor

<400> 8
tactcaggac tcat 14

<210> 9
<211> 18
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 9
agactgcgta ccaattca 18

<210> 10
<211> 17
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 10
gatgagtcct gagtaac 17

IDT Exhibit 1020.3914

<210> 11
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 11
cgtcagactg cgtaccaatt ca 22

<210> 12
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 12
caagagactg cgtaccaatt ca 22

<210> 13
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 13
tggtgatgag tcctgagtaa ca 22

<210> 14
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 14
agccgatgag tcctgagtaa ca 22

<210> 15
<211> 19
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 15
gactgcgtac caattcaac 19

<210> 16
<211> 19
<212> DNA
<213> Artificial

<220>

IDT Exhibit 1020.3915

<223> primer

<400> 16
gatgagtcct gagtaacag 19

<210> 17
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 17
acgtgtagac tgcgtaccga aa 22

<210> 18
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 18
acgtgatgag tcctgagtaa ca 22

<210> 19
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 19
cgtagtagac tgcgtaccga ac 22

<210> 20
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 20
cgtagatgag tcctgagtaa ca 22

<210> 21
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 21
gtacgtagac tgcgtaccga ag 22

<210> 22

IDT Exhibit 1020.3916

<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 22
gtacgatgag tcctgagtaa ca 22

<210> 23
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 23
tacggtagac tgcgtaccga at 22

<210> 24
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 24
tacggatgag tcctgagtaa ca 22

<210> 25
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 25
agtcgtagac tgcgtaccga aa 22

<210> 26
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 26
agtcgatgag tcctgagtaa ca 22

<210> 27
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

IDT Exhibit 1020.3917

<400> 27
catggtagac tgcgtaccga ac 22

<210> 28
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 28
catggatgag tcctgagtaa ca 22

<210> 29
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 29
gagcgtagac tgcgtaccga ag 22

<210> 30
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 30
gagcgatgag tcctgagtaa ca 22

<210> 31
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 31
tgatgtagac tgcgtaccga at 22

<210> 32
<211> 22
<212> DNA
<213> Artificial

<220>
<223> primer

<400> 32
tgatgatgag tcctgagtaa ca 22

<210> 33
<211> 20
<212> DNA

IDT Exhibit 1020.3918

<213> artificial

<220>
<223> primer

<400> 33
aaacccaaac tcccccaatc 20

<210> 34
<211> 68
<212> DNA
<213> artificial

<220>
<223> primer

<400> 34
agcggataac aatttcacac aggacatcag tagtcacact ggtacaaaaa tagagcaaaa 60
cagtagtg 68

<210> 35
<211> 91
<212> DNA
<213> artificial

<220>
<223> probe

<400> 35
aaacccaaac tcccccaatc gatttcaaac ctagaacaat gttggttttg gtgctaactt 60
caacccccact actgttttgc tctatttttg t 91

<210> 36
<211> 90
<212> DNA
<213> artificial

<220>
<223> PI-201234 SNP containing sequence

<400> 36
aaacccaaac tcccccaatc gatttcaaac ctagaacagt gttggttttg gtgctaactt 60
caacccccact actgttttgc tctatttttg 90

<210> 37
<211> 90
<212> DNA
<213> artificial

<220>
<223> PI-201234 SNP

<400> 37
aaacccaaac tcccccaatc gatttcaaac ctagaacagt gttggttttg gtgctaactt 60
caacccccact actgttttgc tctatttttg 90

IDT Exhibit 1020.3919

INTERNATIONAL SEARCH REPORT

International application No
PCT/NL2006/000648

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68

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 practical, search terms used)
EPO-Internal, WPI Data, BIOSIS, EMBASE

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2005/003375 A (454 CORP [US]; LEAMON JOHN H [US]; LOHMAN KENTON [US]; ROTHBERG JONATH) 13 January 2005 (2005-01-13) cited in the application the whole document	1-18
X	WO 00/61800 A2 (KEYGENE NV [NL]; EIJK MICHIEL JOSEPHUS THERESIA [NL]; WITSENBOER HANNE) 19 October 2000 (2000-10-19) the whole document	1,2,4, 9-15,17, 18
X	WO 00/24939 A (AFFYMETRIX INC [US]; DONG SHOULIAN [US]; LIPSHUTZ ROBERT J [US]; LOCKH) 4 May 2000 (2000-05-04) the whole document	1,2,4, 8-15,17, 18

-/--

[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 document 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
28 March 2007	14/05/2007

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, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Werner, Andreas

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

IDT Exhibit 1020.3920

INTERNATIONAL SEARCH REPORT

International application No
PCT/NL2006/000648

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	NICOD J-C ET AL: "SNPs by AFLP (SBA): a rapid SNP isolation strategy for non-model organisms" NUCLEIC ACIDS RESEARCH, OXFORD UNIVERSITY PRESS, SURREY, GB, vol. 31, no. 5, 1 March 2003 (2003-03-01), pages E19-1-E19-5, XP002398230 ISSN: 0305-1048 pages E19-1 - pages E19-3	1,2, 9-15,17, 18
E	WO 2006/137733 A (KEYGENE NV [NL]; VAN EIJK MICHAEL JOSEPHUS THER [NL]; VAN DER POEL HEN) 28 December 2006 (2006-12-28) the whole document	1-18
E	WO 2006/137734 A (KEYGENE NV [NL]; VAN EIJK MICHAEL JOSEPHUS THER [NL]; SOERENSEN ANKER) 28 December 2006 (2006-12-28) the whole document	1-18
A	SHENDURE JAY ET AL: "Accurate multiplex polony sequencing of an evolved bacterial genome" SCIENCE (WASHINGTON D C), vol. 309, no. 5741, September 2005 (2005-09), pages 1728-1732, XP002427180 ISSN: 0036-8075 abstract; figure 1	1-18
A	BREYNE PETER ET AL: "Transcriptome analysis during cell division in plants." PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, vol. 99, no. 23, 12 November 2002 (2002-11-12), pages 14825-14830, XP002427181 ISSN: 0027-8424 the whole document	1-18
A	US 6 013 445 A (ALBRECHT GLENN [US] ET AL) 11 January 2000 (2000-01-11)	1-18
A	MARGULIES MARCEL ET AL: "Genome sequencing in microfabricated high-density picolitre reactors" NATURE, NATURE PUBLISHING GROUP, LONDON, GB, vol. 437, no. 7057, 15 September 2005 (2005-09-15), pages 376-380, XP002398505 ISSN: 0028-0836 the whole document	6,7

-/--

IDT Exhibit 1020.3921

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	LINDSTEDT B A ET AL: "A variation of the amplified-fragment length polymorphism (AFLP) technique using three restriction endonucleases, and assessment of the enzyme combination BglII-MfeI for AFLP analysis of Salmonella enterica subsp. enterica isolates" FEMS MICROBIOLOGY LETTERS, AMSTERDAM, NL, vol. 189, no. 1, 1 August 2000 (2000-08-01), pages 19-24, XP002221164 ISSN: 0378-1097	16

IDT Exhibit 1020.3922

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/NL2006/000648

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2005003375 A	13-01-2005	AU 2004209001 A1	19-08-2004
		AU 2004209416 A1	19-08-2004
		AU 2004209426 A1	19-08-2004
		AU 2004254552 A1	13-01-2005
		CA 2513535 A1	19-08-2004
		CA 2513541 A1	19-08-2004
		CA 2513889 A1	19-08-2004
		CA 2513899 A1	13-01-2005
		EP 1594980 A2	16-11-2005
		EP 1594981 A2	16-11-2005
		EP 1594950 A2	16-11-2005
		EP 1590477 A2	02-11-2005
		JP 2006515522 T	01-06-2006
		WO 2004069849 A2	19-08-2004
		WO 2004070005 A2	19-08-2004
		WO 2004070007 A2	19-08-2004
WO 0061800 A2	19-10-2000	AT 278034 T	15-10-2004
		AU 773855 B2	10-06-2004
		AU 4150800 A	14-11-2000
		AU 773854 B2	10-06-2004
		AU 4150900 A	14-11-2000
		CA 2366365 A1	19-10-2000
		CA 2366374 A1	19-10-2000
		DE 60014350 D1	04-11-2004
		DE 60014350 T2	13-10-2005
		WO 0061801 A2	19-10-2000
		US 7169552 B1	30-01-2007
WO 0024939 A	04-05-2000	AT 316152 T	15-02-2006
		AU 2144000 A	15-05-2000
		CA 2345441 A1	04-05-2000
		DE 69929542 T2	14-09-2006
		EP 1124990 A1	22-08-2001
		JP 2002528096 T	03-09-2002
WO 2006137733 A	28-12-2006	NONE	
WO 2006137734 A	28-12-2006	NONE	
US 6013445 A	11-01-2000	NONE	

IDT Exhibit 1020.3923

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
5 April 2007 (05.04.2007)

PCT

(10) International Publication Number
WO 2007/037678 A2

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

(21) International Application Number:
PCT/NL2006/000467

(22) International Filing Date:
21 September 2006 (21.09.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/721,528 29 September 2005 (29.09.2005) US

(71) Applicant *(for all designated States except US)*: **KEYGENE N.V.** [NL/NL]; 90, Agro Business Park, NL-6708 PW Wageningen (NL).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **VAN EIJK, Michael, Josephus, Theresia** [NL/NL]; 12, Pastoor Strijboschstraat, NL-5373 EJ Herpen (NL). **VAN TUNEN, Adrianus, Johannes** [NL/NL]; 31, Wim Sonneveldstraat, NL-6708 NA Wageningen (NL).

(74) Agent: **DE LANG, R.-J.**; Exter Polak & Charlouis B.V., P.O. Box 3241, NL-2280 GE Rijswijk (NL).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 2007/037678 A2

(54) Title: HIGH THROUGHPUT SCREENING OF MUTAGENIZED POPULATIONS

(57) Abstract: Efficient methods are disclosed for the high throughput identification of mutations in genes in members of mutagenized populations. The methods comprise DNA isolation, pooling, amplification, creation of libraries, high throughput sequencing of libraries, preferably by sequencing-by-synthesis technologies, identification of mutations and identification of the member of the population carrying the mutation and identification of the mutation.

IDT Exhibit 1020.3924

HIGH THROUGHPUT SCREENING OF MUTAGENIZED POPULATIONS

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention, in the fields of molecular biology and genetics relates to improved strategies for identifying mutations in populations, based on the use of high throughput sequencing technologies. The invention further provides for kits that can be applied in the methods.

Description of the Background Art

Populations carrying mutations, either induced or naturally occurring are used in modern genomics research to identify genes affecting traits of importance by reverse genetics approaches. This is in particular applicable for plants and crops of agronomic importance, but such populations are also useful, for other organisms such as yeast, bacteria etc. Other organisms, such as animals, birds, mammals etc can also be used, but these populations are typically more cumbersome to obtain or to control. Nevertheless, it is observed that the invention described herein is of a very general nature, and can be applied also to such organisms.

Mutagenized populations represent complementary tools for gene discovery, as such populations are commonly used to screen known genes for loss-of-function mutations or assessing phenotype changes in organisms with the mutated gene. The rate-limiting step is the screening work associated with identification of, respectively, organisms carrying a mutation in the gene of interest. Below, the principles of such populations and the screening methods are described in more detail and more efficient screening methods are presented which increase the value of these tools for gene-discovery.

A technology that uses mutagenized populations is known as TILLING (Targeted Induced Local Lesions In Genomes) (McCallum *et al.*, Nat. Biotechnol 2000, 18, 455-457, McCallum *et al.*, Plant Physiology, 2000,123, 439-442; Till *et al.* Genome Research 2003, 13, 524-530) relies on random introduction of large numbers of mutations (mostly nucleotide substitutions) into the genome by treatment with ethyl methane sulfonate (EMS) or by ionizing radiation (fast neutron bombardment,) (Li *et al.* The Plant Journal, 2001, 27, 235-42). Every plant in the population carries several hundred (or thousand) mutations, some of which affect normal development, morphology or otherwise confer a phenotype due to loss-of-function

IDT Exhibit 1020.3925

(knock-out, knock-down) of one or multiple genes or their regulatory sequences. A TILLING population generally contains a sufficient number of plants to cover all genes with multiple independent mutations (5-20 per gene). A mutagenized plant population used in TILLING therefore usually consist of 3000-10,000 plants and can be used in two ways:

Reverse Genetics

"Reverse Genetics" is the most common way of using TILLING populations. A gene of interest is identified, *e.g.*, by transcript profiling or a candidate gene approach, and the question to be answered is whether this gene affects a particular phenotypic trait of interest. The challenge therefore is to identify one (or several) plants with loss-of-function mutations in this gene. This is commonly performed in a multi-step screening process, typically comprising the following steps:

1. Genomic DNA of a large number of (pooled) M2 plants (*e.g.*, 3072) of the TILLING population is isolated.
2. Pools of equal amounts of DNA from 8 to 32 plants per pool are assembled, with the pooling level depending on the sensitivity of the CEL I screening system (see below). This results in a total of 96- to 384 pooled DNA samples in case of 3072 plants.
3. Labeled PCR primers are used to amplify parts of the gene from all pooled DNAs. Overlapping PCR fragments are used to cover the entire gene (*e.g.*, 3 * 600 bp PCR fragments are amplified from a 1500 bp gene).
4. Heteroduplexes of the PCR products obtained from the pooled DNA samples are prepared and incubated with CEL I or another enzyme which recognizes and cuts single nucleotide sequence mismatches (*e.g.*, mung bean nuclease, S1 nuclease, Surveyor *etc.*) and the treated samples are resolved on a denaturing (sequencing) gel or by capillary electrophoresis.
5. Pools containing a plant carrying a mutation in the gene are identified by observing bands of digestion products resulting from CEL I treatment.

To identify the plant carrying the mutation, PCRs are repeated on individual DNAs of the plants in the positive pools, followed by bi-directional Sanger sequencing.

Plants harboring a mutation are grown and out-crossed to wild-type to establish causal relationship between the mutation and the observed phenotype change.

The advantage of CEL I screening (steps 3-5 above) is that pre-screening the pooled samples saves costs over sequencing all plants individually by Sanger sequencing.

IDT Exhibit 1020.3926

However, a limitation of CEL I screening is that not all identified mutations affect gene function (*e.g.*, silent substitutions) and this is not known until the PCR products of individual plants in a positive pool are sequenced. Nevertheless, the CEL I mediated screening method is cost-saving compared to sequencing PCR products of all plants separately.

Another limitation is that CEL I screening involves running gels and scoring, a relatively cumbersome process that requires confirmation of mutations from the second strand as gel-patterns are not always clear-cut.

A third disadvantage is that CEL I screening is relatively insensitive to mutation detection at the termini of the PCR product which may lead to some mutations going undetected. Further disadvantages of CEL I are that it has been found that the enzyme is extremely sensitive to reaction conditions such as salt concentrations. This makes that the enzyme can only be used in a limited number of buffers, thereby hampering the broad use of CEL I. Another practical disadvantage associated with the application of CEL I is that the enzyme is not reliable in cutting all mismatched heteroduplexes.

Finally, CEL I screening is incapable of distinguishing missense mutations (which are the most prevalent) from non-sense mutations, causing a great deal of screening work carried out on positive pools without yielding interesting mutations.

Forward Genetics

Plants of the mutagenized population are grown and phenotyped for traits of interest. Plants with an interesting phenotype are then crossed to a wild-type plant to out-cross mutations that are not linked to the phenotype of interest. Finally, the mutated gene responsible for the phenotype of interest is identified by positional cloning (using genetic markers), analogous to mapping QTL in conventional genetic mapping populations (F2, RIL etc). Although theoretically possible, mutagenized populations are not commonly used this way.

The present invention was made in part improve the existing strategies for screening of mutagenized populations. It is an object of the invention to provide efficient methods for screening large populations for the presence of mutations and to improve efficient assessment of the mutations for impact on gene function, *i.e.*, to reduce the amount of effort expended on screening mutations that do not lead to altered gene functions. The present methods were designed to avoid the use of the CEL I enzyme or its equivalents.

SUMMARY OF THE INVENTION

IDT Exhibit 1020.3927

The present inventors found that using high throughput sequencing strategies, the above-mentioned goals were achieved and mutagenized populations, such as TILLING populations, populations wherein mutations have been introduced using (synthetic) mutagenic or DNA damaging oligonucleotides or, i.e. by Targeted Nucleotide Exchange (TNE) or by Region Targeted Mutagenesis (RTM), or populations that contain naturally occurring mutations such as Single nucleotide polymorphisms (SNPs), small insertions and deletions, and variations in microsatellite repeat number could be efficiently screened for the presence of mutations of interest.

Definitions

In the following description and examples, a number of terms are used. To provide a clear and consistent understanding of the specification and claims, including the scope to be given such terms, the following definitions are provided. Unless otherwise defined herein, all technical and scientific terms used have the same meaning as commonly understood by those of ordinary skill in the art to which this invention belongs. The disclosures of all publications, patent applications, patents and other references are incorporated by reference herein in their entirety.

"TILLING" or "Targeting induced local lesions in genomes" is a general reverse genetic strategy providing an allelic series of induced (point) mutations by random chemical or physical mutagenesis in combination with PCR-based screening to identify point mutations in a region of interest. In TILLING screening, regions of interest are amplified by PCR. Heteroduplexes between wild-type fragments and fragments harboring an induced mutation are formed by denaturing and reannealing PCR products. These heteroduplexes are cleaved by CEL I and cleaved products are resolved. Throughput can be increased by pooling. Following discovery of PCR products harboring sequence differences in a pool, PCR products included in the pool are commonly screened again by Sanger sequencing of individual PCR products, thereby identifying the mutant plant and the exact sequence difference in the mutated gene.

"Mutagenized Population" refers to a population of organisms (usually plants, but other organisms, including animals such as Drosophila and mice may be used to create a mutagenized populations; Schimenti *et al.*, 1998, *Genome Research 8:*698-710) that have been subjected to mutagenesis (chemical or physical) to yield a library of mutants. TILLING populations may vary widely in size, and for certain purposes, partial TILLING populations can be used that contain 90, 80 70, 60, 50, 40 30 or even only 20% of the original population. As an alternative to

IDT Exhibit 1020.3928

mutagenized populations, populations can be used wherein the population is not mutagenized but comprises sub-populations that contain naturally occurring mutations such as Single nucleotide polymorphisms (SNPs), small insertions and deletions, and variations in microsatellite repeat number. These populations are particularly advantageous when mutagenized populations are not readily accessible (humans) or where already large germplasms are available. See for instance Comai *et al.* , The Plant Journal, 2004, 37, 778-786. Such a population can be used in combination with a 'reference DNA'.

"Targeted Nucleotide Exchange" or "TNE". Targeted nucleotide exchange (TNE) is a process by which a synthetic oligonucleotide, partially complementary to a site in a chromosomal or an episomal gene directs the reversal of a single nucleotide at a specific site. TNE has been described using a wide variety of oligonucleotides and targets. Some of the reported oligonucleotides are RNA/DNA chimeras, contain terminal modifications to impart nuclease resistance.

"Region targeted mutagenesis" or "RTM". Region targeted mutagenesis is a process by which double-strand breaks at a predefined target site in the genomic DNA are artificially created, resulting in repair of the break by one of various available cellular repair mechanisms, mostly leading to mutations at the site of the break. Double-strand breaks may be created by introduction into the cell nucleus of zinc-finger nucleases (e.g. see Lloyd *et al.*, 2005), meganucleases such as I-SceI (Epinat *et al.*, 2003), or triplex-forming oligonucleotides coupled to mutagenic chemical groups (Havre *et al.*, 1993).

"Nucleic acid": A nucleic acid, as used herein, may include any polymer or oligomer of nucleotides with pyrimidine and purine bases, preferably cytosine, thymine (or uracil), adenine and guanine, respectively (*See* Lehninger, *Principles of Biochemistry*, at 793-800 (Worth Pub. 1982) which is herein incorporated by reference in its entirety for all purposes). Any deoxyribonucleotide, ribonucleotide or peptide nucleic acid component, and any chemical variant thereof, such as those with methylated, hydroxymethylated or glycosylated forms of these bases, and the like, are included. The polymers or oligomers may be heterogeneous or homogeneous in composition, and may be isolated from naturally occurring sources or may be artificially or synthetically produced. A nucleic acid may be DNA or RNA, or a mixture thereof, and may exist permanently or transiently in single-stranded or double-stranded form, including homoduplexes, heteroduplexes, and hybrid states.

"Tagging" refers to the addition of a tag or label to a nucleic acid in order to be able to distinguish it from a second or further nucleic acid. Tagging can be performed, for example, by

IDT Exhibit 1020.3929

the addition of a sequence identifier during amplification by using tagged primers or by any other means known in the art. Such a sequence identifier can be a unique base sequence of varying but defined length uniquely used for identifying a specific nucleic acid sample. Typical example are ZIP sequences. Using such a tag, the origin of a sample can be determined upon further processing. In the case of combining processed products originating from different nucleic acid samples, the different nucleic acid samples are generally identified using different tags.

"Tagged library" refers to a library of tagged nucleic acids.

"Sequencing" refers to determining the order of nucleotides (base sequences) in a nucleic acid sample, *e.g.*, DNA or RNA.

"Aligning and alignment" mean the comparison of two or more nucleotide sequences based on the presence of short or long stretches of identical or similar nucleotides. Several methods for alignment of nucleotide sequences are known in the art, as will be further explained below. Sometimes the terms "assembly" or "clustering" are used as synonyms.

"High-throughput screening" (HTS) is a method of scientific experimentation especially relevant to the fields of biology and chemistry. Through a combination of modern robotics and other specialized laboratory hardware, HTS allows an investigator to effectively screen large numbers of samples simultaneously (or virtually simultaneously).

"Primers" in general refers to DNA strands which can prime the synthesis of DNA. DNA polymerase cannot synthesize DNA *de novo* without primers: it can only extend an existing DNA strand in a reaction in which the complementary strand is used as a template to direct the order of nucleotides to be assembled. The synthetic oligonucleotide molecules which are used in a polymerase chain reaction (PCR) are referred to herein as primers.

"Primers with increased affinity" are primers with modified nucleotides such as PNA or LNA, which increases their thermal stability and allows for allele-specific amplification based on single nucleotide sequence differences. In order to achieve this, one or several modified nucleotides are often included, preferably at the 3'-end of the primer.

"DNA amplification" is typically used to denote the *in vitro* synthesis of double-stranded DNA molecules using PCR. It is noted that other amplification methods exist and they may also be used in the present invention.

BRIEF DESCRIPTION OF THE DRAWINGS

IDT Exhibit 1020.3930

Fig 1: Schematic representation of clustered sequences resulting from shotgun sequencing a gene to identify EMS-induced mutations. Mutations are lighter, sequence errors darker colored. Sequence errors are expected to be observed randomly and most often just once.

Fig 2: Schematic representation of clustered tagged sequencing resulting from a 100 bp gene region amplified with 4 bp –tagged PCR primers from a 3-D pooled library. Mutations are lighter, sequence errors darker colored. Plant IDs are known for mutations identified by 3 tags (1,2,3) and (4,5,6) but not for those identified by less than 2 tag (7,8). Sequence errors are expected to be observed randomly and just once.

Fig 3: Illustration of the system of long and short PCR primers to use in tagging the sequences.

Fig. 4. Agarose gel estimation of the PCR amplification yield of eIF4E exon 1 amplification for each of the 28 3D pools.

DETAILED DESCRIPTION OF THE INVENTION

In one aspect the invention is directed to a method for the detection of a mutation in a target sequence in a member of a mutagenized population comprising the steps of:

(a) Isolating genomic DNA of each member of the mutagenized population to provide for DNA samples of each member in the population;

(b) pooling the DNA obtained in step (a);

(c) amplifying the target sequence with a pair of (optionally labeled) primers from the DNA pools;

(d) pooling the amplification products of step (c) to create a library of amplification products;

(e) optionally, fragmenting the amplification products in the library;

(f) determining the nucleotide sequence of the products and/or fragments using high throughput sequencing;

(g) identifying mutations by clustering (aligning) the sequences of the fragments;

(h) screening the identified mutations for a modified function of the target sequence;

(i) designing a primer directed to hybridize to the identified mutation;

(j) amplifying the library of step (d) with the primer of step (i) and one of the primers of step (c);

IDT Exhibit 1020.3931

(k) identifying the member(s) carrying the mutation;

(l) optionally, confirming the mutation by amplifying the target sequence from the member(s) of step (k) using the primers of step (c) and determining the sequence of the amplified product.

The isolation of DNA is generally achieved using common methods in the art such as the collection of tissue from a member of the population, DNA extraction (for instance using the Q-Biogene fast DNA kit), quantification and normalization to obtain equal amounts of DNA per sample. As an example, the present invention is illustrated based on a TILLING population of 3072 plants and a gene of 1500bp.

The pooling of the isolated DNA can for instance be achieved using a 3-dimensional pooling scheme (Vandenbussche *et al.,* 2003, *The Plant Cell, 15:* 2680-93). The pooling is achieved preferably using equal amounts of DNA. The 3D-pooling scheme may comprise 15x15x14, resulting in 44 pools (15+15+14) containing 3072/14=219 or 3072/15 =205 different DNA samples per pool. Other pooling schemes can be used.

The pooling step typically serves to identify the plant containing an observed mutation after one round of PCR screening. Pooling of the DNA further serves to normalize the DNAs prior to PCR amplification to provide for a more equal representation in the libraries for sequencing. The additional advantage of the pooling of the DNA is that not all sequences have to be determined separately, but that the pools allow for rapid identification of the sequences of interest, in particular when tagged libraries are used. This facilitates the screening of large or complex populations in particular.

The amplification of the target sequence with a pair of optionally labeled primers from the pools can be achieved by using a set of primers that have been designed to amplify the gene of interest. As stated, the primers may be labeled to visualize the amplification product of the gene of interest.

The amplification products are pooled, preferably in equal or normalized amounts to thereby create a library of amplification products. Exemplary, the complexity of the library will be 3072 plants X 1500 bp gene sequence = 4.6 Mb sequence.

The amplification products in the library may be randomly fragmented prior to sequencing of the fragments in case the PCR product length exceeds the average length of the sequence traces. Fragmentation can be achieved by physical techniques, *i.e.,* shearing, sonication or other random fragmentation methods. In step (f), at least part, but preferably the entire, nucleotides sequence of at least part of, but preferably of all the fragments contained in the

IDT Exhibit 1020.3932

libraries is determined. In certain embodiments, the fragmentation step is optional. For instance, when the read length of the sequencing technique and the PCR fragments length are about the same, there is no need for fragmentation. Also in the case of larger PCR products this may not be necessary if it is acceptable that only part of the PCR product is sequenced for instance in case of 1500 bp PCR product and read length of 400 (from each side) 700 bp remain unsequenced.

The sequencing may in principle be conducted by any means known in the art, such as the dideoxy chain termination method (Sanger sequencing), but this is less preferred given the large number of sequences that have to be determined. It is however preferred and more advantageous that the sequencing is performed using high-throughput sequencing methods, such as the methods disclosed in WO 03/004690, WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007, and WO 2005/003375 (all in the name of 454 Life Sciences), by Seo *et al.* (2004) *Proc. Natl. Acad. Sci. USA 101*:5488-93, and technologies of Helios, Solexa, US Genomics, etcetera, which are herein incorporated by reference. It is most preferred that sequencing is performed using the apparatus and/or method disclosed in WO 03/004690, WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007, and WO 2005/003375 (all in the name of 454 Life Sciences), which are herein incorporated by reference. The technology described allows sequencing of 40 million bases in a single run and is 100 times faster and cheaper than competing technology. The sequencing technology roughly consists of 5 steps: 1) fragmentation of DNA and ligation of specific adaptor to create a library of single-stranded DNA (ssDNA); 2) annealing of ssDNA to beads, emulsification of the beads in water-in-oil microreactors and performing emulsion PCR to amplify the individual ssDNA molecules on beads; 3) selection of /enrichment for beads containing amplified ssDNA molecules on their surface 4) deposition of DNA carrying beads in a PicoTiterPlate®; and 5) simultaneous sequencing in at least 100,000 wells by generation of a pyrophosphate light signal. The method will be explained in more detail below.

In a preferred embodiment, the sequencing comprises the steps of:

(a) annealing adapted fragments to beads, with a single adapted fragment being annealed to each bead;

(b) emulsifying the beads in water-in-oil microreactors, each water-in-oil microreactor comprising a single bead;

(c) loading the beads in wells, each well comprising a single bead; and generating a pyrophosphate signal.

IDT Exhibit 1020.3933

In the first step (a), sequencing adaptors are ligated to fragments within the library. The sequencing adaptor includes at least a "key" region for annealing to a bead, a sequencing primer region and a PCR primer region. Thus, adapted fragments are obtained.

In a second step, adapted fragments are annealed to beads, each bead annealing with a single adapted fragment. To the pool of adapted fragments, beads are added in excess as to ensure annealing of one single adapted fragment per bead for the majority of the beads (Poisson distribution).

In a next step, the beads are emulsified in water-in-oil microreactors, each water-in-oil microreactor comprising a single bead. PCR reagents are present in the water-in-oil microreactors allowing a PCR reaction to take place within the microreactors. Subsequently, the microreactors are broken, and the beads comprising DNA (DNA positive beads) are enriched.

In a following step, the beads are loaded in wells, each well comprising a single bead. The wells are preferably part of a PicoTiter™Plate allowing for simultaneous sequencing of a large amount of fragments.

After addition of enzyme-carrying beads, the sequence of the fragments is determined using pyrosequencing. In successive steps, the PicoTiter™Plate and the beads as well as the enzyme beads therein are subjected to different deoxyribonucleotides in the presence of conventional sequencing reagents, and upon incorporation of a deoxyribonucleotide a light signal is generated which is recorded. Incorporation of the correct nucleotide will generate a pyrosequencing signal which can be detected.

Pyrosequencing itself is known in the art and described inter alia on www.biotagebio.com; www.pyrosequencing.com / section technology. The technology is further applied in *e.g.*, WO 03/004690, WO 03/054142, WO 2004/069849, WO 2004/070005, WO 2004/070007, and WO 2005/003375 (all in the name of 454 Life Sciences), which are herein incorporated by reference.

The mutations are identified by clustering of the sequenced fragments in the amplified library. Identification of the mutations is achieved by aligning the determined sequences of the fragments of the libraries. The majority of the sequences are wild-type (not mutated) but the induced mutations and occasional sequencing errors are also observed. As the amplification libraries are sequenced with multifold redundancy (typically about 4- to 5-fold redundant), multiple observations of the same sequence change is indicative of a mutation rather than a sequencing error. See Fig. 1.

IDT Exhibit 1020.3934

The clustering provides alignments of the fragments in the amplified library. In this way for each PCR product in the library, a cluster is generated from sequenced fragments, *i.e.*, a contig of the fragments, is build up from the alignment of the sequence of the various fragments obtained from the fragmenting in step (e).

Methods of alignment of sequences for comparison purposes are well known in the art. Various programs and alignment algorithms are described in: Smith and Waterman (1981) Adv. Appl. Math. 2:482; Needleman and Wunsch (1970) J. Mol. Biol. 48:443; Pearson and Lipman (1988) Proc. Natl. Acad. Sci. USA 85:2444; Higgins and Sharp (1988) Gene 73:237-244; Higgins and Sharp (1989) CABIOS 5:151-153; Corpet *et al.* (1988) Nucl. Acids Res. 16:10881-90; Huang *et al.* (1992) *Computer Appl. in the Biosci. 8*:155-65; and Pearson *et al.* (1994) *Meth. Mol. Biol.* 24:307-31, which are herein incorporated by reference . Altschul *et al.* (1994) *Nature Genet. 6*:119-29 (herein incorporated by reference) present a detailed consideration of sequence alignment methods and homology calculations.

The NCBI Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990) is available from several sources, including the National Center for Biological Information (NCBI, Bethesda, Md.) and on the Internet, for use in connection with the sequence analysis programs blastp, blastn, blastx, tblastn and tblastx. It can be accessed at <http://www.ncbi.nlm.nih.gov/BLAST/>. A description of how to determine sequence identity using this program is available at <http://www.ncbi.nlm.nih.gov/BLAST/blast_help.html>.

In the analysis of mutagenized populations, after the mutations have been identified, the identified mutations are assessed for a modified function of the associated gene, for instance the introduction of a stop codon. This assessment is performed on the sequence itself, for example by six-frame translation. Once the interesting mutations have been identified, the mutations are further investigated to identify the associated member of the population.

For each mutation that has been classified as an interesting mutation, an allele specific primer is designed that targets the mutation of interest. Thus, the allele specific primer is then used in combination with one of the primers used in the amplification of the pooled DNA samples (either the reverse or the forward primer). One or both of the primers may be labeled. The set of primers is used to amplify the pools of DNA. The positive pools are identified and the mutant plant is identified. In the above-mentioned 3D pooling scheme, the allele specific PCR with the set of primers to screen the 3D pooled DNA sample plates results in the identification of 3 positive pools (one in each dimension), which specifies the library address of the mutant plant.

IDT Exhibit 1020.3935

In certain embodiments, the allele-specific primers comprise alternative nucleotides such as Locked Nucleic Acids (LNA) or Peptide Nucleic Acids (PNA) to increase their specificity. Such nucleic acids are widely known in the art and are commercially available from a choice of suppliers.

Confirmation of the mutation is achieved by amplification of the target sequence from the identified mutant plant. This amplification is performed with the primers from step (c). The nucleotide sequence of the amplified product is determined and by comparison with the consensus sequence, the mutation is identified. The sequencing is preferably performed Sanger sequencing.

In one aspect the invention pertains to a method for the detection of a mutation in a target sequence in a member of a mutagenized population comprising the steps of:

(a) isolating genomic DNA of each member of the mutagenized population to provide DNA samples of each member in the population;

(b) pooling the DNA obtained in step (a);

(c) amplifying a part or segment of the target sequence with a pair of tagged (optionally labeled) primers from the DNA pools, preferably wherein at least one of the primers comprise a gene-specific section, a tag and a sequence primer binding site;

(d) pooling the amplification products of step (c) to create a library of amplification products;

(d) determining the nucleotide sequence of the amplification products using high throughput sequencing;

(f) identifying mutations by clustering (aligning) the sequences of the fragments;

(g) identifying the member(s) having the mutation using the tags;

(h) optionally, confirming the mutation by amplifying the target sequence from the member(s) of step (g) using the primers of step (c) and determining the sequence of the amplified product.

The isolation of genomic DNA of the members of the mutagenized population and the pooling of the isolated DNA can be carried out essentially as described above.

A part or segment of the target sequence is amplified using a pair of tagged primers that may be labeled. Preferably, for each pool of each dimension, a different primer is used. In the above illustration this means that 44 forward and 44 reverse primers are preferred. Preferably, each of the forward and reverse primers comprises

(i) a sequence primer binding site that can be used in the following sequencing step,

IDT Exhibit 1020.3936

(ii) a tag that serves to link the primer (and the resulting amplification product) to the original member of the population, and

(iii) a gene specific sequence that is capable of annealing to the target sequence of interest (*i.e.*, the gene).

In a typical embodiment the primer has the following order:

5'-Sequence primer binding site---Tag---Gene specific PCR primer sequence-3'

The length of the sequence primer binding site and the gene specific PCR primer sequence are those that are conventional in common PCR use, *i.e.*, independently from about 10 to about 30bp with a preference for from 15 to 25bp. Preferably the part or segment of the sequence that is amplified corresponds to a length that can be sequenced in one run using the high throughput sequencing technologies described below. In certain embodiments the part or segment has a length of between about 50 bp to about 500 bp, preferably from about 75 bp to about 300 bp and more preferably between about 90bp and about 250 bp. As stated above, this length may vary with the sequencing technology employed including those yet to be developed.

By using primers (forward and/or reverse) containing a tag sequence that is unique for each of the primers representing all pool dimensions, the specific plant origin of each tag sequence is known as the sequence primer anneals upstream of the tag and as a consequence, the tag sequence is present in each amplification product. In certain embodiments, both forward and reverse primers are tagged. In other embodiments, only on of the forward or reverse primers is tagged. The choice between one or two tags depends on the circumstances and depends on the read length of the high throughput sequencing reaction and/ or the necessity of independent validation. In the case of, *e.g.*, a 100bp PCR product that is sequenced unidirectionally, only one tag is needed. In the case of a 200bp PCR product and a 100bp read-length, double tagging is useful in combination with bi-directional sequencing as it improves efficiency 2-fold. It further provides the possibility of independent validation in the same step. When a 100bp PCR product is sequenced bi-directionally with two tagged primers, all traces, regardless of orientation, will provide information about the mutation. Hence both primers provide "address information" about which plant contains which mutation.

The tag can be any number of nucleotides, but preferably contains 2, 3, 4 or 5 nucleotides. With 4 nucleotides permuted, 256 tags are possible, whereas 3 nucleotides permuted provide 64 different tags. In the illustration used, the tags preferably differ by >1 base, so preferred tags are 4 bp in length. Amplification using these primers results in a library of tagged amplification products.

IDT Exhibit 1020.3937

In certain embodiments, a system of tags can be used wherein the amplification process includes

(1) a long PCR primer comprising (a) a 5'-constant section linked to (b) a degenerate tag section (NNNN) linked to (c) a gene specific section- 3' and

(2) a short PCR primer in subsequent amplifications that consists of (a) the 5'-contact section linked to (b) non-degenerate tag section-3' (*i.e.*, a selection amongst NNNN).

The non-degenerate tag section can be unique for each sample, for example, ACTG for sample 1, AATC for sample 2, *etc.* The short primer anneals to a subset of the long primer. The constant section of the primer can be used as a sequence primer. See Fig. 3.

The library preferably comprises equal, amounts of PCR products from all amplified pools. In the illustrative example, the library contains 3072 plants x 100bp= 307kb sequence to be determined.

The PCR products in the library are subjected to a sequencing process as disclosed above. In particular, the PCR products are attached to beads using the sequence primer binding site that corresponds to the sequence linked to the bead. Thus the present embodiment does not require fragmentation and adapter ligation. Rather, in this embodiment, the adapters have been introduced earlier via the PCR primer design. This improves the reliability of the method. Following the annealing to the beads, sequencing is performed as described above, *i.e.*, (1) emulsification of the beads in water-in-oil microreactors, (2) emulsion PCR to amplify the individual ssDNA molecules on beads; (3) selection of /enrichment for beads containing amplified ssDNA molecules on their surface, (4) transfer of the DNA carrying beads to a PicoTiterPlate®; and (5) simultaneous sequencing in 100,000 wells by a method that generates a pyrophosphate light signal. Typical output is about 200.000x100-200 bp sequences, representing a 66 fold coverage of all PCR products in the library.

Clustering and alignment is performed essentially as described above. The individual plant containing the mutation can be identified using the tags. In the examples, the combination of the 3 tags denotes the positive pools and the consequently the coordinates of the individual plant in the pools.

Confirmation of the mutation by re-sequencing of the PCR product of the identified mutant sample is as described above.

Various pooling strategies can be used with the present invention, examples of which are multidimensional pooling (including 3D pooling) or column-, row- or plate pooling.

IDT Exhibit 1020.3938

High throughput sequencing methods that can be used here are described, for example, in Shendure *et al., Science 309*:1728-32. Examples include microelectrophoretic sequencing, hybridization sequencing/sequencing by hybridization (SBH), cyclic-array sequencing on amplified molecules, cyclic-array sequencing on single molecules, non-cyclical, single-molecule, real-time methods, such as , polymerase sequencing, exonuclease sequencing, or nanopore sequencing.

For optimal results, fragments or amplified products should be sequenced with sufficient redundancy. Redundancy permits distinction between a sequencing error and a genuine possible mutation. In certain embodiments, the redundancy of the sequencing is preferable at least 4, more preferably at least 5, but, as can be seen from the Examples, redundancies of more than 10, preferably more than 25 or even more than 50 are considered advantageous, although not essential for this invention..

Advantages of the methods of the present invention reside *inter alia* in the fact that mutations can be assessed *in silico* for their impact on gene function, meaning that a selection is made for the active mutations. Mutations conferring only silent substitutions can be selected against, thereby making the overall process more economical and efficient. This is a particular advantage with regard to the known CEL I based TILLING technology because the majority of CEL I mutations are C/G to T/A transitions, of which only 5% commonly create stop codons (Colbert *et al.* 2001). The vast majority are missense mutations of reduced interest. Efficient recognition of members in a population with stop codon mutations economizes the process and obviates the need for additional screening of individual members of positive pools.

All mutations can be found with equal probability, irrespective of their position in the PCR product, in particular when the whole target sequence is screened.

The method further avoids the use of CEL I digestion, heteroduplex formation and cumbersome gel scoring. The invention is therefore insensitive to pooling limitations associated with CEL I technology.

The invention further relates to kits that may contain one or more compounds selected form the group consisting of: one or more (labeled) primers for a particular gene or trait, mutation- or allele-specific primers. The kits may further contain beads, sequencing primers, software, descriptions for pooling strategies and other components that are known for kits per se. In certain embodiments, kits are provided that are dedicated to find specific mutations, for instance disease-related mutations.

IDT Exhibit 1020.3939

The invention is now illustrated here in below.

EXAMPLES

Screening a TILLING population can be advanced by using novel high-throughput sequencing methods, such as that of 454 Life Sciences (Margulies *et al.*, 2005) or Polony Sequencing (Shendure *et al.*, 2005). With the current state-of-the-art, 454 Life Sciences technology produces approximately 20 Mb sequence in a single sequencing run. Read lengths are approximately 100 bp per read. Assuming the screening of a population consisting of 3072 plants for mutations in a 1500 bp gene (as described in the above-cited reference in Chapter 2), two approaches are envisaged and described in more detail below.

(1) an approach where the entire 1500 bp gene is investigated for the presence of EMS-induced mutations; and

(1) an approach where one or several 100 bp stretches are investigated for the presence of EMS-induced mutations.

EXAMPLE I

Screening the Entire 1500 bp Region:

Genomic DNA of 3072 plants of the TILLING population is isolated. A 3-D pooling scheme of equal amounts of DNA per plant is set up (*e.g.*, 15 x 15 x 14), resulting in 44 pools (15+15+14 = 44) containing 3072/14= 219 or 3072/15 = 205 different DNA samples (Vandenbussche *et al.*, *supra*).

This pooling step serves to permit identification of a plant containing an observed mutation after one round of PCR screening (step 8). Pooling of genomic DNAs further serves to normalize DNAs prior to PCR amplification to increase the probability that all DNAs are represented equally in the sequence library.

The 1500 bp gene is amplified from the pooled DNA samples using 1 pair of unlabelled PCR primers.

Equal amounts of PCR products from all pools wells are pooled to create a pooled PCR products library (complexity 3072 plants x 1500bp = 4.6 Mb sequence).

The pooled PCR product library is subjected to shotgun sequencing using conventional technologies (such as those provided by 454 Life Sciences) wherein PCR products are randomly fragmented, amplified on individual beads and sequenced on the bead. Output is approximately 200,000 100bp sequences, representing 4- to 5-fold coverage of all PCR products in the library).

IDT Exhibit 1020.3940

All sequences are clustered. The majority of sequences are wild-type but EMS-induced mutations (and sequence errors) are observed as well. Since PCR products are sequenced with 4-5 fold redundancy, multiple observations of the same sequence change is indicative of a mutation rather than a sequencing error (Figure 1).

Mutations are assessed for their impact on gene function such as introduction of a stop-codon.

An allele-specific primer targeting a mutation of interest (with 3' Locked Nucleic Acid; LNA; or Peptide Nucleic Acid; PNA) is designed to be used in combination with either the forward or reverse primer used in step 3 to screen the 3-D pooled DNA sample plate. Allele-specific PCR will result in three positive pools (one of each dimension), which specifies the library address of the mutant plant.

The mutation is confirmed by amplifying the 1500 bp gene using the primers of step 3, followed by (bi-directional) Sanger sequencing.

EXAMPLE II

Screening 100 bp Stretches

(100 bp is the read length of one 454 sequence run)

Genomic DNA of 3072 plants of the TILLING population is isolated. A 3-D pooling scheme of equal amounts of DNA per plant is set up (*e.g.*, 15 x 15 x 14), resulting in 44 pools (15+15+14 = 44) containing 3072/14= 219 or 3072/15 = 205 different DNA samples (Vandenbussche *et al.*, *supra*).

This pooling step serves to permit identification of the plant containing an observed mutation directly from the sequence data. Pooling of genomic DNAs further serves to normalize DNAs prior to PCR amplification to increase the probability that all DNAs are represented equally in the sequence library.

A 100bp (or 200bp) region of the gene is amplified from a the pools by PCR using tagged unlabelled PCR primers. This requires 44 forward and 44 reverse primers (one for each pool of each dimension) with the following configuration:

5'-Sequence primer binding site---4 bp Tag---Gene specific primer sequence-3'.

By using tailed forward and reverse primers containing a 4 bp sequence tag that is different for each of the 44 primers representing all pool dimensions, the specific plant origin of each sequence is known as the sequence primer anneals upstream of the tag. Hence the tag sequence in present in each sequence trace. A 4 bp tag allows $4^4 = 256$ different tags. A 3bp tag

IDT Exhibit 1020.3941

allows 64 different tag sequences - sufficient to distinguish 44 tags- but tag sequences differing by more than 1 base are preferred.

Equal amounts of PCR products from all pools wells are pooled to create a pooled PCR products library (complexity 3072 plants x 100 bp = 307 kb sequence).

The pooled PCR product library is provided to 454 for sequencing, *i.e.*, PCR products are amplified and sequenced on the beads. Output is approximately 200,000 100bp sequences, representing 66-fold coverage of all PCR products in the library.

All sequences (from either direction) are clustered; the majority of sequences are wild-type sequences but EMS-induced mutations (and sequence errors) are observed as well. Since PCR products are sequenced with 66 fold redundancy, multiple observations of the same sequence change are indicative of a mutation rather than a sequencing error (Figure 1).

The coordinates of the individual plant containing the mutation will be known immediately based on the unique combination of 3 tags sequences that occur in the sequence traces harboring the mutation (Figure 2).

The mutation is confirmed by amplifying the 1500 bp gene using the primers of step 3, followed by (bi-directional) Sanger sequencing.

Example III. Identifying specific mutations in a mutant library of tomato

Mutant library of tomato

This example describes the screening of a mutant library of tomato by massive parallel sequencing in order to identify point mutations in a specific locus (target gene). The mutant library used is an isogenic library of inbred determinate tomato cultivar M82 consisting of 5075 M2 families derived from EMS mutagenesis treatments. Seeds of each of the 5075 M2 families were stored at 10% RH and 7°C. The origin and characteristics of the library are described in Menda *et al.* (*Plant J.* **38**: 861-872, 2004) and in a database on http://zamir.sgn.cornell.edu/mutants/index.html.

DNA isolation

Leaf material was harvested from 5 individual greenhouse-grown plants of each of 3072 M2 families randomly chosen from the library. As any mutation occurring in the library will segregate in a Mendelian fashion in the M2 offspring, the pooling of the leaf material of 5 individual M2 plants reduced the likelihood of overlooking any mutation as a consequence of segregation to less than 0.1%. Genomic DNA was isolated from the pooled leaf material using a

IDT Exhibit 1020.3942

modified CTAB procedure described by Stuart and Via (*Biotechniques*, **14**: 748-750, 1993). DNA samples were diluted to a concentration of 100 ng/µl in TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA) and stored at -20 °C in 96-well microtitre plates.

3D pooling of the DNA samples

The isolated DNA samples were normalized to a concentration of 20 ng/µl and subsequently pooled 4-fold resulting in 768 samples comprised in eight 96-well microtitre plates. Subsequently, these eight microtitre plates were subjected to a 3D pooling strategy, resulting in 28 pools of DNA. The 3D pooling strategy consisted of pooling together all DNAs in three different manners, thus ensuring that each single 4-fold pool occurs only once in an X-coordinate pool, only once in a Y-coordinate pool and only once in a Z-coordinate pool.

X-pools were assembled by pooling all DNA samples together per column of eight wells (e.g. A1-H1) from all eight microtitre plates, resulting in 12 X-pools. Each X-pool therefore held 8 (wells in a column) × 8 (plates) = 64 samples of 4-fold pools, representing 256 M2 families.

Y-pools were assembled by pooling all DNA samples together per row of twelve wells (e.g. A1-A12) from all eight microtitre plates, resulting in 8 Y-pools. Each Y-pool therefore held 12 (wells in a row) × 8 (plates) = 96 samples of 4-fold pools, representing 384 M2 families.

Z-pools were assembled by pooling all DNA samples together from an entire microtitre plate, resulting in 8 Z-pools. Each Z-pool therefore held 12 × 8 = 96 samples of 4-fold pools, representing 384 M2 families.

Target locus

The target locus in this example was part of the tomato gene for eucaryotic initiation factor 4E (eIF4E). This gene has been shown to be involved in susceptibility to infection of potyviruses in Arabidopsis (Duprat *et al.*, *Plant J.* **32**: 927-934, 2002), lettuce (Nicaise *et al. Plant Physiol.* **132**: 1272-1282, 2003) and Solanaceae (Ruffel *et al.*, *Plant J.* **32**: 1067-1075, 2002; *Mol.Gen.Genomics* **274**: 346-353, 2005), and specific mutations in this gene are associated with recessive potyvirus resistance. The mutation screening described in this example was aimed to identify additional mutations in the tomato eIF4E gene as possible sources of new potyvirus resistance. For the tomato eIF4E, only the cDNA sequence was known (NCBI accession numbers AY723733 and AY723734). Using a PCR approach using primers designed on the basis of the cDNA sequence, fragments of the genomic sequence of the eIF4E locus of tomato cultivar Moneyberg were amplified and sequenced. This resulted in a sequence of most of the

IDT Exhibit 1020.3943

genomic locus of tomato eIF4E. The locus consists of 4 exons and 3 introns. For the mutation screening, exon 1 of the gene was chosen as the target sequence (SEQ ID 57).

SEQ ID 57: Sequence of exon 1 of tomato Moneyberg eIF4E :

ATGGCAGCAGCTGAAATGGAGAGAACGATGTCGTTTGATGCAGCTGAGAAGTTGAAGGCCGCCG
ATGGAGGAGGAGGAGAGGTAGACGATGAACTTGAAGAAGGTGAAATTGTTGAAGAATCAAATGA
TACGGCATCGTATTTAGGGAAAGAAATCACAGTGAAGCATCCATTGGAGCATTCATGGACTTTT
TGGTTTGATAACCCTACCACTAAATCTCGACAAACTGCTTGGGGAAGCTCACTTCGAAATGTCT
ACACTTTCTCCACTGTTGAAAATTTTTGGGG

Primer design for target locus amplification

Primers were designed for the PCR amplification of exon 1 of tomato eIF4E. The forward primers were designed to correspond to the ATG start codon of the Open Reading Frame of exon 1, with 5' of the ATG a tag sequence of four bases, providing a unique identifier for each of the 28 pools. At the far 5' end of the forward PCR primers, a 5'-C was added. All primers were phosphorylated at their 5'end to facilitate subsequent ligation of adaptors. The sequence and names of the 28 forward primers are listed in Table 1. The tag sequences are underlined.

Table 1. Forward primers, sequences and pool identification for exon 1 amplification.

name	sequence	3D pool	SEQ ID #
06I009	CACACATGGCAGCAGCTGAAATGG	X1	1
06I010	CACAGATGGCAGCAGCTGAAATGG	X2	2
06I011	CACGAATGGCAGCAGCTGAAATGG	X3	3
06I012	CACGTATGGCAGCAGCTGAAATGG	X4	4
06I013	CACTCATGGCAGCAGCTGAAATGG	X5	5
06I014	CACTGATGGCAGCAGCTGAAATGG	X6	6
06I015	CAGACATGGCAGCAGCTGAAATGG	X7	7
06I016	CAGAGATGGCAGCAGCTGAAATGG	X8	8
06I017	CAGCAATGGCAGCAGCTGAAATGG	X9	9
06I018	CAGCTATGGCAGCAGCTGAAATGG	X10	10
06I019	CAGTCATGGCAGCAGCTGAAATGG	X11	11
06I020	CAGTGATGGCAGCAGCTGAAATGG	X12	12
06I021	CATCGATGGCAGCAGCTGAAATGG	Y1	13

IDT Exhibit 1020.3944

06I022	CATGCATGGCAGCAGCTGAAATGG	Y2	14
06I023	CTACGATGGCAGCAGCTGAAATGG	Y3	15
06I024	CTAGCATGGCAGCAGCTGAAATGG	Y4	16
06I025	CTCACATGGCAGCAGCTGAAATGG	Y5	17
06I026	CTCAGATGGCAGCAGCTGAAATGG	Y6	18
06I027	CTCGAATGGCAGCAGCTGAAATGG	Y7	19
06I028	CTCGTATGGCAGCAGCTGAAATGG	Y8	20
06I029	CTCTCATGGCAGCAGCTGAAATGG	Z1	21
06I030	CTCTGATGGCAGCAGCTGAAATGG	Z2	22
06I031	CTGACATGGCAGCAGCTGAAATGG	Z3	23
06I032	CTGAGATGGCAGCAGCTGAAATGG	Z4	24
06I033	CTGCAATGGCAGCAGCTGAAATGG	Z5	25
06I034	CTGCTATGGCAGCAGCTGAAATGG	Z6	26
06I035	CTGTCATGGCAGCAGCTGAAATGG	Z7	27
06I036	CTGTGATGGCAGCAGCTGAAATGG	Z8	28

The reverse primers were designed to correspond to basepair position 267 to 287 of exon 1 in the non-coding strand. Again, 5' of the priming part the same series of tag sequences of four bases were included, providing a identifier for each of the 28 pools. At the far 5' end of the reverse PCR primers, a 5'-C was added. All primers were phosphorylated at their 5'end to facilitate subsequent ligation of adaptors. The sequence and names of the 28 reverse primers are listed in Table 2. The tags are underlined.

Table 2. Reverse primers, sequences and pool identification for exon 1 amplification.

name	sequence	3D pool	SEQ ID #
06I037	CACACCCCCAAAAATTTTCAACAGTG	X1	29
06I038	CACAGCCCCAAAAATTTTCAACAGTG	X2	30
06I039	CACGACCCCAAAAATTTTCAACAGTG	X3	31
06I040	CACGTCCCCAAAAATTTTCAACAGTG	X4	32
06I041	CACTCCCCCAAAAATTTTCAACAGTG	X5	33
06I042	CACTGCCCCAAAAATTTTCAACAGTG	X6	34
06I043	CAGACCCCCAAAAATTTTCAACAGTG	X7	35

IDT Exhibit 1020.3945

06I044	CAGAGCCCCAAAAATTTTCAACAGTG	X8	36
06I045	CAGCACCCCAAAAATTTTCAACAGTG	X9	37
06I046	CAGCTCCCCAAAAATTTTCAACAGTG	X10	38
06I047	CAGTCCCCCAAAAATTTTCAACAGTG	X11	39
06I048	CAGTGCCCCAAAAATTTTCAACAGTG	X12	40
06I049	CATCGCCCCAAAAATTTTCAACAGTG	Y1	41
06I050	CATGCCCCCAAAAATTTTCAACAGTG	Y2	42
06I051	CTACGCCCCAAAAATTTTCAACAGTG	Y3	43
06I052	CTAGCCCCCAAAAATTTTCAACAGTG	Y4	44
06I053	CTCACCCCCAAAAATTTTCAACAGTG	Y5	45
06I054	CTCAGCCCCAAAAATTTTCAACAGTG	Y6	46
06I055	CTCGACCCCAAAAATTTTCAACAGTG	Y7	47
06I056	CTCGTCCCCAAAAATTTTCAACAGTG	Y8	48
06I057	CTCTCCCCCAAAAATTTTCAACAGTG	Z1	49
06I058	CTCTGCCCCAAAAATTTTCAACAGTG	Z2	50
06I059	CTGACCCCCAAAAATTTTCAACAGTG	Z3	51
06I060	CTGAGCCCCAAAAATTTTCAACAGTG	Z4	52
06I061	CTGCACCCCAAAAATTTTCAACAGTG	Z5	53
06I062	CTGCTCCCCAAAAATTTTCAACAGTG	Z6	54
06I063	CTGTCCCCCAAAAATTTTCAACAGTG	Z7	55
06I064	CTGTGCCCCAAAAATTTTCAACAGTG	Z8	56

Target locus amplification

The exon 1 of the target locus was amplified from the 3D pooled DNAs using the forward and reverse primers described above. For each PCR reaction, a forward and a reverse primer were used with identical tags. For the amplification of exon 1 from each of the 28 3D pools, a different set of forward and reverse primers was used.

The PCR amplification reaction conditions for each sample were as follows:

25 µl DNA (= 50 ng); 5 µl RNase-mix ; 10 µl 5x Herculase PCR-buffer; 0.6 µl of the four dNTPs (20mM); 1.25 µl forward primer (50 ng/µl); 1.25 µl reverse primer (50 ng/µl); 0.5 µl Herculase DNA polymerase; 28.9 µl milliQ-purified water. The RNase-mix consisted of 157,5 µl milliQ-purified water + 17,5 µl RNase.

IDT Exhibit 1020.3946

PCR amplifications were performed in a PE9600 thermocycler with a gold or silver block using the following conditions: 2 minutes hot-start of 94°C, followed by 35 cycles of 30 sec at 94°C, 30 sec at 53°C, 1 min at 72°C, and a final stationary temperature of 4°C. The PCR amplification efficiency was checked by analysis of 10 µl of PCR products on a 1% agarose gel. Fig. 4 shows the efficient amplification of exon 1 PCR products from each of the 28 3D pools in comparison to a concentration range of lambda DNA on the same gel.

Following amplification, equal amounts of PCR products were mixed and purified using the QIAquick PCR Purification Kit (QIAGEN), according to the QIAquick® Spin handbook (page 18). On each column a maximum of 100 µl of product was loaded. Products were eluted in 10 mM Tris-EDTA.

Sequence library preparation and high-throughput sequencing

Mixed amplification products from the 3D pools were subjected to high-throughput sequencing on a GS20 sequencer using 454 Life Sciences sequencing technology as described by Margulies *et al.* (*Nature* **437**: 376-380, 2005, and Online Supplements). Specifically, the PCR products were ligated to adaptors to facilitate emulsion-PCR amplification and subsequent fragment sequencing as described by Margulies *et al.*. The 454 adaptor sequences, emulsion PCR primers, sequence primers and sequence run conditions were all as described by Margulies *et al.*. The linear order of functional elements in an emulsion-PCR fragment amplified on Sepharose beads in the 454 sequencing process was as follows:

454 PCR adaptor - 454 sequence adaptor - C-nucleotide - 4 bp tag – target amplification primer sequence 1 - target fragment internal sequence – target amplification primer sequence 2 - 4 bp tag – G-nucleotide - 454 sequence adaptor - 454 PCR adaptor - Sepharose bead.

454 sequence run data-processing.

After base calling with 454 software for each region of the microtiter plate a file with FASTA formatted sequences was produced. These were concatenated into one file. Within this file a search was conducted with a regular expression to a 100% match of the forward primer preceded with 5 nucleotides (C plus four bp tag sequence). The same was done with the reverse primer extended with 5 nucleotides (C plus tag sequence). All sequences were then grouped by their tag sequence (pool indentifiers) in separate files. Each file was analysed with the ssahaSNP tool and the known exon 1 nucleotide sequence as a reference. The ssahaSNP tool reported about all

IDT Exhibit 1020.3947

single nucleotide sequence differences and "indels" (single base insertions or deletions as a result of either mutagenesis or erroneous base-calling) of the 454 sequences versus the reference genome. These single nucleotide sequence difference and indel statistics were saved in a database and used for error rate analysis and point mutation identification.

454 sequencing error rate

The total number of correct sequences obtained from the data processing for all 28 pools combined was 247,052. The sequences were divided in two groups, those that aligned with the forward primer and coding strand (5' end) of the exon 1 PCR product (128,594 = 52%), and those that aligned with the reverse primer and the complementary strand of the PCR product (118,458 = 48%). The number of sequences obtained from each of the different pools and alignment groups ranged from 69 to 7269. On average, each of the 3072 M2 families should be represented 80 times in the total collection of sequences, and each allele 40 times.

Within the alignment group corresponding to the forward primer, 1338 sequences out of 128,594 (1.2%) showed one or more single nucleotide sequence differences in relation to the eIF4E reference sequence along a stretch of 63 bases of aligned target sequence. For the reverse primer group, 743 sequences out of 118,458 (0.6%) showed one or more single nucleotide sequence differences in relation to the eIF4E reference sequence along a stretch of 102 bases of aligned target sequence. Therefore, the single base substitution error rate for both sequence groups combined equals 0.84% for a 165 base stretch, or 0.0051% per base position (0.5 errors per 10,000 bases). This error rate is similar to the one reported by Margulies *et al.* of 0.004% for individual read substitution errors in test sequences, but much lower than for whole-genome resequencing (0.68%).

A similar analysis of the occurence of indels in both alignment groups revealed an indel incidence of 3883 (forward primer group) and 3829 (reverse primer group) in a total of 247,052 sequences (is 3.1% in a 165 bp stretch). The indel occurence rate therefore equals 0.01891% per base position (1.89 indels per 10,000 bases). The indel rate is significant higher than the base substitution error rate. Both types of sequencing error combined occur on average at a frequency of 2.39 per 10,000 bases, or 0.024 per base position. This error rate is much lower than reported by Margulies *et al.*, and may be explained by the absence of long homopolymer stretches in the eIF4e exon 1 sequence.

IDT Exhibit 1020.3948

Detection of a mutation in the target locus

Because the objective of this screen is the identification of (EMS)-induced point mutations (preferentially C→T and G→A mutations), all sequences representing indels in comparison to the reference sequence were discarded for the sake of the analysis in this example. Most of the single base substitutions occurred only once in any given 3D pool, some occurred 2 or 3 times, or rarely more often. Since these single base substitutions occur more or less uniformly at every position of the aligned sequence, and at a more or less uniform frequency of 0.005% per base, they were assumed to represent sequencing errors, and not specific mutations that exist in the mutant library. However, at a few specific base positions in the scanned sequence, a much higher incidence of a specific single base sequence difference occurs. Such single base sequence differences reveal mutations in the library, when the following criteria are fulfilled:

1. the single base sequence difference represents an C→T or G→A mutation;
2. the incidence is higher than 20 per 10,000 sequence reads per 3D pool;
3. the single base sequence difference occurs in precisely one and not more than one X-pool, Y-pool and Z-pool.

In this example, one such mutation was found in the alignment group corresponding to the reverse primer, at base position 221 of the eIF4E exon 1 sequence. This mutation, a G→A mutation (corresponding to C→T in the complementary strand) occured in pool X12 at a frequency of 70 per 10,000 sequences, in pool Y3 at a frequency of 33 per 10,000 and in pool Z6 at 62 per 10,000 sequences. This same mutation at the same position did not occur in any of the other pools, not even at background error rates.

The unique occurrence of this G221A mutation in only the three pools allowed the identification of the original 4-fold pool of DNA, representing four M2 families. DNA of each of these four M2 families was amplified individually with the primers 06F598 and 06F599 that are identical to the forward and reverse primers of Tables 1 and 2, but without the 5' five base sequence tags. The amplified PCR products were subjected to conventional Sanger sequencing. The sequence of the eIF4E gene in one of the four families (coded "24") revealed a dual peak at position 221, corresponding to an overlapping G and A. This is indicative of an M2 family pool, in which half the alleles are wild-type, and the other half carry the G221A point mutation (Fig. 2). The sequences of the other M2 families around base position 221 were according to the reference (wild-type).

IDT Exhibit 1020.3949

The mutation causes an arginine to glutamine substitution. Seeds of this particular M2 family were planted in the greenhouse in order to select for homozygous mutant individuals, that will be used for phenotyping.

In a similar manner, two other point mutations were identified in the 454 sequence reads. An estimation of the mutation density of the M82 tomato mutant library therefore equals 3 mutations per 165 bp scanned sequence, or 18 mutations per 1000 bases in 3072 M2 families. This corresponds to mutation densities reported for *Arabidopsis* (Greene *et al.*, *Genetics* **164**: 731-740, 2003).

References

Colbert *et al.*.2001. High-throughput screening for induced point mutations. *Plant Physiology* 126: 480-484.

Duprat *et al.*, 2002. The *Arabidopsis* eukaryotic initiation factor (iso)4E is dispensable for plant growth but required for susceptibility to potyviruses. *Plant J.* 32: 927-934.

Epinat *et al.*, 2003. A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic Acids Research*, **31**(11): 2952-2962.

Havre *et al.*, 1993. Targeted mutagenesis of DNA using triple helix-forming oligonucleotides linked to psoralen. *Proc.Natl.Acad.Sci. USA* **90**: 7879-7883.

McCallum *et al.*, 2000. Targeted screening for induced mutations. *Nature Biotechnology* **18:** 455-457.

Greene *et al.*, 2003. Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in Arabidopsis. *Genetics* **164:** 731-740.

Lloyd *et al.*, 2005. Targeted mutagenesis using zinc-finger nucleases in *Arabidopsis*. *Proc.Natl.Acad.Sci. USA* **102**: 2232-2237.

Margulies *et al.*, 2005. Genome sequencing in microfabricated high-density picolitre reactions. *Nature* **437**: 376-380.

IDT Exhibit 1020.3950

Menda *et al.*, 2004. *In silico* screening of a saturated mutation library of tomato. *Plant J.* **38**: 861-872.

Nicaise *et al.*, 2003. The eukaryotic translation initiation factor 4E controls lettuce susceptibility to the p otyvirus l*ettuce mosaic virus*1. *Plant Physiol.* **132**: 1272-1282.

Ruffel *et al.*, 2002. A natural recessive resistance gene against potato virus Y in pepper corresponds to the eukaryotic initiation factor 4E (eIF4E). *Plant J.* **32**: 1067-1075.

Ruffel *et al.*, 2005. The recessive potyvirus resistance gene *pot-1* is the tomato orthologue of the pepper *pvr2-eIF4E* gene. *Mol.Gen.Genomics* **274**: 346-353.

Shendure *et al.*, 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Scienceexpress Report*, August 4.

Stuart and Via, 1993. A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. *Biotechniques*, **14**: 748-750.

Vandenbussche *et al.*, 2003. Toward the analysis of the petunia MADS box gene family by reverse and forward transposon insertion mutagenesis approaches: B, C, and D floral organ identity functions require SEPALLATA-like MADS box genes in petunia. *The Plant Cell* **15**:2680-2693.

IDT Exhibit 1020.3951

CLAIMS

1. A method for the detection of a mutation in a target sequence in a member of a mutagenized population, comprising the steps of:

(a) isolating genomic DNA of each member of the mutagenized population to provide DNA samples of each population member;

(b) pooling the DNA obtained in step (a);

(c) amplifying the target sequence with a pair of (optionally labeled) primers from the DNA pools;

(d) pooling the amplification products of step (c) to create a library of amplification products;

(e) optionally, fragmenting the amplification products in the library;

(f) determining the nucleotide sequence of the amplified products or fragments using high throughput sequencing;

(g) identifying mutations by clustering/aligning the sequences of the fragments;

(h) screening the identified mutations for a modified function of the target sequence;

(i) designing a primer directed to hybridize to the identified mutation;

(j) amplifying the library of step (d) with the primer of step (i) and one of the primers of step (c);

(k) identify the population member(s) carrying the mutation;

(l) optionally, confirming the mutation by amplifying the target sequence from the member(s) of step (k) using the primers of step (c) and determining the sequence of the amplified product.

2. A method for the detection of a mutation in a target sequence in a member of a mutagenized population, comprising the steps of:

(a) isolating genomic DNA of each member of the mutagenized population to provide for DNA samples of each population member;

(b) pooling the DNA obtained in step (a);

(c) amplifying part of the target sequence with a pair of tagged (optionally labeled) primers from the DNA pools, wherein at least one of the primers preferably comprises a gene-specific section, a tag and a sequence primer binding site;

(d) pooling the amplification products of step (c) to create a library of amplification products;

IDT Exhibit 1020.3952

(e) determining the nucleotide sequence of the amplification products using high throughput sequencing;

(f) identifying mutations by clustering/aligning the sequences of the fragments;

(g) identifying the member(s) having the mutation using the tags;

(h) optionally, confirming the mutation by amplifying the target sequence from the member(s) of step (g) using the primers of step (c) and determining the sequence of the amplified product.

3. The method according to claim 1 or 2, wherein the mutagenized population is obtained by treating the genome with one or more selected from the group consisting of mutation-inducing chemicals, ionizing radiation, targeted nucleotide exchange or region targeted muatgenesis.

4. The method according to claim 1 or 2, wherein the population is not mutagenized but comprises sub-populations that contain naturally occurring mutations.

5. The method according to claims 1-4, wherein the pooling is a 3D-pooling strategy.

6. The method according to claims 1-5, wherein the high throughput sequencing is performed by synthesis, preferably pyrosequencing.

7. The method according to any of claims 1-6, wherein sequencing is performed on a solid support such as a bead.

8. The method according to any of claims 1-7, wherein the sequencing is based on High Throughput Sequencing, preferably Sequencing-by-Synthesis.

9. The method according to any of claims 1-7, wherein the sequencing is performed by synthesis, preferably pyrosequencing.

10. The method according to claim 6, wherein sequencing comprises the steps of:

(i) ligating sequencing adaptors to the fragments;

(ii) annealing said adaptor-ligated fragments to beads, each bead annealing with a single fragment;

(iii) emulsifying the beads in water-in-oil microreactors such that each microreactor contains a single bead;

IDT Exhibit 1020.3953

(iv) performing emulsion PCR to amplify adaptor-ligated fragments on the surface of the beads

(v) selecting/enriching beads to which are attached said amplified adaptor-ligated fragments

(vi) placing the beads in wells such that each well comprises a single bead; and

(vii) generating a pyrophosphate signal.

11. The method according to any of claims 1-10, wherein the primers in steps (c) and/or step (i) contain nucleotides with improved binding affinity.

12. The method according to any of claims 2-11, wherein the part of the target sequence that is amplified in step (c) is from about 80 to about 400 bp.

13. The method according to claim 12, wherein the part of the target sequence that is amplified in step (c) is from about 90 to about 300 bp.

14. The method according to claim 12, wherein the part of the target sequence that is amplified in step (c) is from 100 to about 200 bp.

15. Kit comprising one or more (labeled) primers for a particular gene or trait

16. Kit according to claim 15, further comprising mutation- or allele-specific primers

17. Kit according to claim 15 or 16 further comprising beads and/or sequencing primers.

IDT Exhibit 1020.3954

FIGURE 1

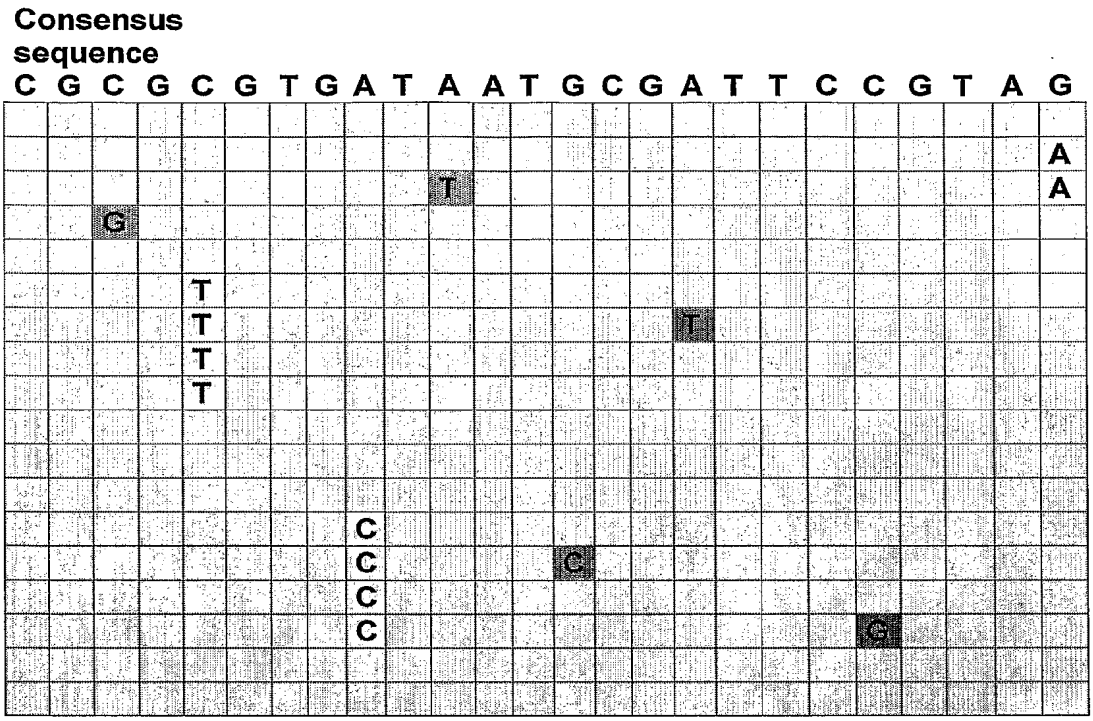


FIGURE 2

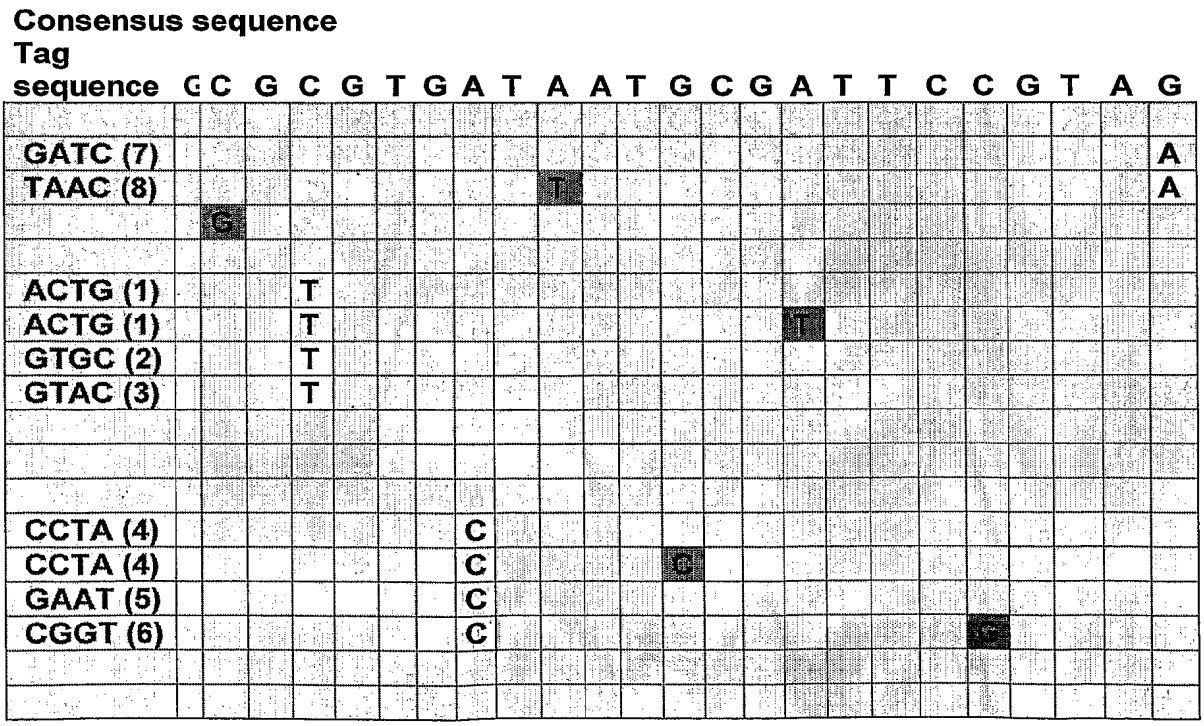


Fig. 3

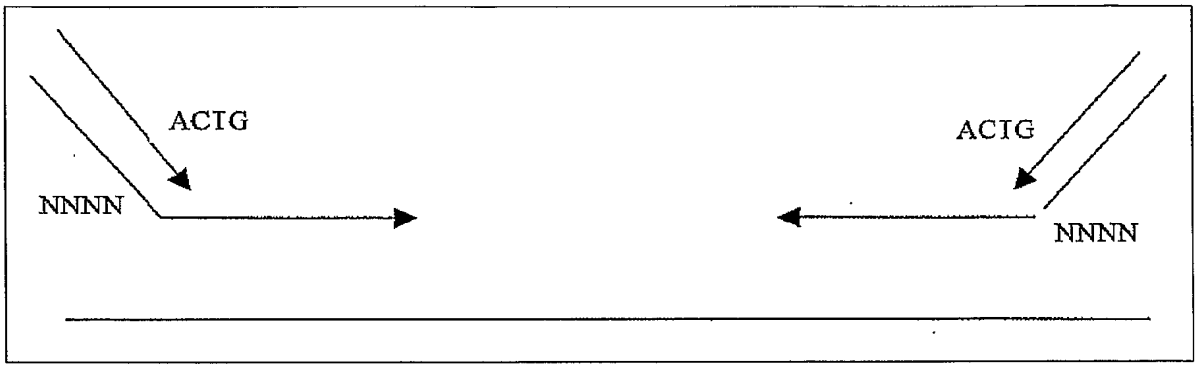
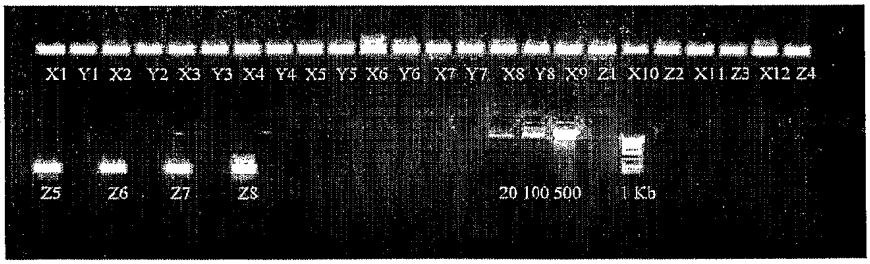

ACTG
NNNN
ACTG
NNNN

Fig 4


X1 Y1 X2 Y2 X3 Y3 X4 Y4 X5 Y5 X6 Y6 X7 Y7 X8 Y8 X9 Z1 X10 Z2 X11 Z3 X12 Z4
Z5
Z6
Z7
Z8
20 100 500
1 Kb

SEQUENCE LISTING

<110> Keygene NV

<120> HIGH THROUGHPUT SCREENING OF MUTAGENIZED POPULATIONS

<130> P27926PC00

<150> 60/721,528
<151> 2005-09-29

<160> 57

<170> PatentIn version 3.3

<210> 1
<211> 24
<212> DNA
<213> artificial;

<400> 1
cacacatggc agcagctgaa atgg 24

<210> 2
<211> 24
<212> DNA
<213> artificial;

<400> 2
cacagatggc agcagctgaa atgg 24

<210> 3
<211> 24
<212> DNA
<213> artificial;

<400> 3
cacgaatggc agcagctgaa atgg 24

IDT Exhibit 1020.3957

<210> 4
<211> 24
<212> DNA
<213> artificial;

<400> 4
cacgtatggc agcagctgaa atgg 24

<210> 5
<211> 24
<212> DNA
<213> artificial;

<400> 5
cactcatggc agcagctgaa atgg 24

<210> 6
<211> 24
<212> DNA
<213> artificial;

<400> 6
cactgatggc agcagctgaa atgg 24

<210> 7
<211> 24
<212> DNA
<213> artificial;

<400> 7
cagacatggc agcagctgaa atgg 24

<210> 8
<211> 24
<212> DNA
<213> artificial;

IDT Exhibit 1020.3958

<400> 8
cagagatggc agcagctgaa atgg 24

<210> 9
<211> 24
<212> DNA
<213> artificial;

<400> 9
cagcaatggc agcagctgaa atgg 24

<210> 10
<211> 24
<212> DNA
<213> artificial;

<400> 10
cagctatggc agcagctgaa atgg 24

<210> 11
<211> 24
<212> DNA
<213> artificial;

<400> 11
cagtcatggc agcagctgaa atgg 24

<210> 12
<211> 24
<212> DNA
<213> artificial;

<400> 12
cagtgatggc agcagctgaa atgg 24

IDT Exhibit 1020.3959

<210> 13
<211> 24
<212> DNA
<213> artificial;

<400> 13
catcgatggc agcagctgaa atgg 24

<210> 14
<211> 24
<212> DNA
<213> artificial;

<400> 14
catgcatggc agcagctgaa atgg 24

<210> 15
<211> 24
<212> DNA
<213> artificial;

<400> 15
ctacgatggc agcagctgaa atgg 24

<210> 16
<211> 24
<212> DNA
<213> artificial;

<400> 16
ctagcatggc agcagctgaa atgg 24

<210> 17
<211> 24
<212> DNA
<213> artificial;

IDT Exhibit 1020.3960

<400> 17
ctcacatggc agcagctgaa atgg 24

<210> 18
<211> 24
<212> DNA
<213> artificial;

<400> 18
ctcagatggc agcagctgaa atgg 24

<210> 19
<211> 24
<212> DNA
<213> artificial;

<400> 19
ctcgaatggc agcagctgaa atgg 24

<210> 20
<211> 24
<212> DNA
<213> artificial;

<400> 20
ctcgtatggc agcagctgaa atgg 24

<210> 21
<211> 24
<212> DNA
<213> artificial;

<400> 21
ctctcatggc agcagctgaa atgg 24

<210> 22

IDT Exhibit 1020.3961

<211> 24
<212> DNA
<213> artificial;

<400> 22
ctctgatggc agcagctgaa atgg 24

<210> 23
<211> 24
<212> DNA
<213> artificial;

<400> 23
ctgacatggc agcagctgaa atgg 24

<210> 24
<211> 24
<212> DNA
<213> artificial;

<400> 24
ctgagatggc agcagctgaa atgg 24

<210> 25
<211> 24
<212> DNA
<213> artificial;

<400> 25
ctgcaatggc agcagctgaa atgg 24

<210> 26
<211> 24
<212> DNA
<213> artificial;

<400> 26

IDT Exhibit 1020.3962

ctgctatggc agcagctgaa atgg 24

<210> 27
<211> 24
<212> DNA
<213> artificial;

<400> 27
ctgtcatggc agcagctgaa atgg 24

<210> 28
<211> 24
<212> DNA
<213> artificial;

<400> 28
ctgtgatggc agcagctgaa atgg 24

<210> 29
<211> 26
<212> DNA
<213> artificial;

<400> 29
cacaccccca aaaattttca acagtg 26

<210> 30
<211> 26
<212> DNA
<213> artificial;

<400> 30
cacagcccca aaaattttca acagtg 26

<210> 31
<211> 26

IDT Exhibit 1020.3963

<212> DNA
<213> artificial;

<400> 31
cacgacccca aaaattttca acagtg 26

<210> 32
<211> 26
<212> DNA
<213> artificial;

<400> 32
cacgtcccca aaaattttca acagtg 26

<210> 33
<211> 26
<212> DNA
<213> artificial;

<400> 33
cactccccca aaaattttca acagtg 26

<210> 34
<211> 26
<212> DNA
<213> artificial;

<400> 34
cactgcccca aaaattttca acagtg 26

<210> 35
<211> 26
<212> DNA
<213> artificial;

<400> 35
cagaccccca aaaattttca acagtg 26

IDT Exhibit 1020.3964

<210> 36
<211> 26
<212> DNA
<213> artificial;

<400> 36
cagagcccca aaaattttca acagtg 26

<210> 37
<211> 26
<212> DNA
<213> artificial;

<400> 37
cagcacccca aaaattttca acagtg 26

<210> 38
<211> 26
<212> DNA
<213> artificial;

<400> 38
cagctcccca aaaattttca acagtg 26

<210> 39
<211> 26
<212> DNA
<213> artificial;

<400> 39
cagtccccca aaaattttca acagtg 26

<210> 40
<211> 26
<212> DNA

IDT Exhibit 1020.3965

<213> artificial;

<400> 40
cagtgcccca aaaattttca acagtg 26

<210> 41
<211> 26
<212> DNA
<213> artificial;

<400> 41
catcgcccca aaaattttca acagtg 26

<210> 42
<211> 26
<212> DNA
<213> artificial;

<400> 42
catgcccca aaaattttca acagtg 26

<210> 43
<211> 26
<212> DNA
<213> artificial;

<400> 43
ctacgcccca aaaattttca acagtg 26

<210> 44
<211> 26
<212> DNA
<213> artificial;

<400> 44
ctagcccca aaaattttca acagtg 26

IDT Exhibit 1020.3966

<210> 45
<211> 26
<212> DNA
<213> artificial;

<400> 45
ctcaccccca aaaattttca acagtg 26

<210> 46
<211> 26
<212> DNA
<213> artificial;

<400> 46
ctcagcccca aaaattttca acagtg 26

<210> 47
<211> 26
<212> DNA
<213> artificial;

<400> 47
ctcgacccca aaaattttca acagtg 26

<210> 48
<211> 26
<212> DNA
<213> artificial;

<400> 48
ctcgtcccca aaaattttca acagtg 26

<210> 49
<211> 26
<212> DNA
<213> artificial;

IDT Exhibit 1020.3967

<400> 49
ctctccccca aaaattttca acagtg 26

<210> 50
<211> 26
<212> DNA
<213> artificial;

<400> 50
ctctgcccca aaaattttca acagtg 26

<210> 51
<211> 26
<212> DNA
<213> artificial;

<400> 51
ctgaccccca aaaattttca acagtg 26

<210> 52
<211> 26
<212> DNA
<213> artificial;

<400> 52
ctgagcccca aaaattttca acagtg 26

<210> 53
<211> 26
<212> DNA
<213> artificial;

<400> 53
ctgcacccca aaaattttca acagtg 26

IDT Exhibit 1020.3968

<210> 54
<211> 26
<212> DNA
<213> artificial;

<400> 54
ctgctcccca aaaattttca acagtg 26

<210> 55
<211> 26
<212> DNA
<213> artificial;

<400> 55
ctgtccccca aaaattttca acagtg 26

<210> 56
<211> 26
<212> DNA
<213> artificial;

<400> 56
ctgtgcccca aaaattttca acagtg 26

<210> 57
<211> 287
<212> DNA
<213> Lycopersicon esculentum

<400> 57
atggcagcag ctgaaatgga gagaacgatg tcgtttgatg cagctgagaa gttgaaggcc 60

gccgatggag gaggaggaga ggtagacgat gaacttgaag aaggtgaaat tgttgaagaa 120

tcaaatgata cggcatcgta tttagggaaa gaaatcacag tgaagcatcc attggagcat 180

tcatggactt tttggtttga taaccctacc actaaatctc gacaaactgc ttggggaagc 240

IDT Exhibit 1020.3969

tcacttcgaa atgtctacac tttctccact gttgaaaatt tttgggg 287

IDT Exhibit 1020.3970

Espacenet

Bibliographic data: WO2012013932 (A1) — 2012-02-02

IMPROVED RACE

Inventor(s): JOHNSTON IAN ALISTAIR [GB]; BOWER NEIL [AU] ± (JOHNSTON, IAN, ALISTAIR, ; BOWER, NEIL)

Applicant(s): UNIV ST ANDREWS [GB]; JOHNSTON IAN ALISTAIR [GB]; BOWER NEIL [AU] ± (UNIVERSITY COURT OF THE UNIVERSITY OF ST ANDREWS, ; JOHNSTON, IAN, ALISTAIR, ; BOWER, NEIL)

Classification: **- international:** ***C12N15/10***
- cooperative: **C12N15/1096**

Application number: WO2011GB01131 20110728 Global Dossier

Priority number (s): GB20100012748 20100729

Abstract of WO2012013932 (A1)

The present invention relates to a method of generating a cDNA molecule or pool of cDNA molecules which comprise adapter sequences 5' and 3' of the sequence corresponding to the 5' or 3' ends of the original mRNA molecule. There is further provided a method of amplifying a desired 5' or 3' end of said modified cDNA or pool of cDNA molecules whilst minimising the generation of false positives and/or non-specific amplification. There is also provided a kit for carrying out the method of the present invention.

IDT Exhibit 1020.3971

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
2 February 2012 (02.02.2012)

PCT

(10) International Publication Number
WO 2012/013932 A1

(51) International Patent Classification:
C12N 15/10 (2006.01)

(21) International Application Number:
PCT/GB2011/001131

(22) International Filing Date:
28 July 2011 (28.07.2011)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1012748.8 29 July 2010 (29.07.2010) GB

(71) Applicant: **UNIVERSITY COURT OF THE UNIVERSITY OF ST ANDREWS** [GB/GB]; College Gate, North Street, St Andrews KY16 9AJ (GB).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **JOHNSTON, Ian, Alistair** [GB/GB]; University of St Andrews, School of Biology, Scottish Oceans Institute, East Sands, St Andrews, Fife KY16 8LB (GB). **BOWER, Neil** [GB/AU]; Institute for Molecular Bioscience, 306 Carmody Road, Building 80, The University of Queensland, St Lucia, Brisbane, QLD 4072 (AU).

(74) Agent: **NAISMITH, Robert, Stewart**; Marks & Clerk LLP, Aurora, 120 Bothwell Street, Glasgow G2 7JS (GB).

(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, 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, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, 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, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, 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, ML, MR, NE, SN, TD, TG).

Published:

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

— *with sequence listing part of description (Rule 5.2(a))*

WO 2012/013932 A1

(54) **Title**: IMPROVED RACE

(57) **Abstract**: The present invention relates to a method of generating a cDNA molecule or pool of cDNA molecules which comprise adapter sequences 5' and 3' of the sequence corresponding to the 5' or 3' ends of the original mRNA molecule. There is further provided a method of amplifying a desired 5' or 3' end of said modified cDNA or pool of cDNA molecules whilst minimising the generation of false positives and/or non-specific amplification. There is also provided a kit for carrying out the method of the present invention.

IDT Exhibit 1020.3972

Improved RACE

Field of the Invention

The present invention relates to a method of generating a cDNA molecule or pool of cDNA molecules which comprise adapter sequences 5' and 3' of the sequence corresponding to the 5' or 3' ends of the original mRNA molecule. There is further provided a method of amplifying a desired 5' or 3' end of said modified cDNA or pool of cDNA molecules whilst minimising the generation of false positives and/or non-specific amplification. There is also provided a kit for carrying out the method of the present invention.

INTRODUCTION

The genomes of an increasing number of organisms have been sequenced, however the exact structure of many transcripts remain unresolved, particularly their 5' untranslated regions (UTRs). Many open reading frames are predicted and need to be supported by experimental data to obtain an accurate annotation of the genome. For example, Salehi-Ashtiani et al (1) performed large scale rapid amplification of cDNA ends (RACE) on *C. elegans*, identifying UTRs, new exons and redefined previously annotated exons. The lack of experimental data supporting gene structure is in part due to the methodologies employed in cloning gene sequences. Cloning of novel cDNAs typically results in sequences that only partially represent the mRNA species, with the 3' ends usually well represented and the 5' end of the sequence often incomplete. This tendency can be due to RNA degradation, or is a result of the cDNA synthesis reaction not reaching completion. When working with novel organisms,

where the genome has not been sequenced, obtaining the 5' end of the cDNA, although in theory quite simple, can often prove to be technically challenging.

The unknown 5' sequence from the ends of cDNA molecules can be cloned using a PCR technique know as rapid amplification of cDNA ends (RACE) (2). In Classic RACE, homopolymer tailing of the cDNA is used to append a linker sequence to the cDNA terminus. Extension of unknown regions of the cDNA is then achieved through PCR using a gene specific primer and a primer that can bind and prime DNA synthesis from the linker sequence.

Different versions of RACE have been developed, making significant improvements to the original methodology, including New RACE (3), and Cap finder RACE (5, 6). In New RACE, an anchor primer is ligated to the 5' end of the mRNA molecule prior to reverse transcription, which subsequently synthesised cDNA molecules incorporate into their sequence (3). Therefore only full length cDNA molecules are amplified by PCR using the anchor primer and gene specific primer. The Cap-switch RACE (Cap finder) protocol exploits the Murine moloney leukaemia virus (MMLV) reverse transcriptase ability to add an extra 2-4 cytosine residues to the 3' ends of cDNA molecules after reaching the cap structure at the 5' end of the mRNA template. When a primer with multiple guanine residues at its 3' end is present in the reaction mix, annealing of the poly G to poly C sequences occurs, allowing for copying of the annealed primer sequence to the cDNA, adding an adapter sequence to the end of the cDNA (5, 6). These methods, have been developed into commercially available RACE kits including Clontech (Cap switching Smart RACE), Invitrogen (5'RACE system) and Ambion (First Choice RLM-RACE kit).

Many of the commercially available kits offer protocols that allow the amplification of 5' cDNA ends of many genes from a single cDNA synthesis reaction. Despite claims of specific amplification of the desired 5' region of a target gene, a common problem that occurs in RACE reactions is obtaining non-specific amplification products (6, 7, 8). The use of a second nested PCR reaction has been shown to increase the specificity of the reaction, however, non-specific amplification and false positive products are still often obtained. When cDNA synthesis is primed using an oligo-dT primer, all mRNA species are reverse transcribed, and the 3' primer site will subsequently be incorporated into the 3' end of all newly transcribed cDNAs. When cDNA templates such as these are used for PCR, non-specific amplification is likely due to the incorporation of the primer sequence in every cDNA molecule in the universal pool of full length cDNAs.

Prevention of PCR carryover contamination by incorporation of dUTP during PCR and uracil-DNA glycosylase treatment is a widely used technique (9) (covered by patents US5035996, US 5683896 and US5945313). dUTP is incorporated into all amplicons by substituting dUTP for dTTP in PCR reactions. Treatment of PCR products with UDG prior to thermal cycling cleaves DNA at uracil bases thereby inhibiting amplification of carryover contamination from previous PCR reactions (10). This function can be exploited to remove unwanted DNA sequences from reaction mixtures. For example, the use of dUTP and UDG to increase the specificity of 3' RACE reactions has been reported (11, see also US5334515), however, no methods using dUTP/UDG for 5' RACE are currently available.

It is amongst the objects of the present invention to obviate and/or mitigate one or more of the aforementioned disadvantages.

Summary of the Invention

The present invention is based on the development of a modified version of the RACE protocol which enables the amplification reaction to be simplified by degrading non-target modified cDNA sequences using uracil DNA glycosylase (UDG). Essentially, a double stranded cDNA is synthesised to incorporate dUTP, and this then serves as template for subsequent amplification of cDNA ends of any target gene. During synthesis of the modified double stranded cDNA, PCR primer sites (where dUTP replaces dTTP) are incorporated at the 3' and 5' ends of each cDNA molecule. These primer sites are incorporated at the 3' and 5' ends by way of adapter sequences. The modified double stranded cDNA, incorporating dUTP, now serves as a universal pool for PCR amplification of cDNA ends for any gene of interest.

Thus, in a first aspect there is provided a method for generating a modified cDNA molecule comprising sequence corresponding to the 5' and 3' ends of an mRNA molecule comprising the steps of:

a) annealing a first adapter oligonucleotide to an mRNA molecule, said first adapter comprising a 3' portion which is substantially complimentary to the 3' poly A tail of the mRNA molecule and a 5' portion which comprises one or more uracil residues;

b) extending from said first adapter, utilizing dNTPs which include dUTP, so as to generate a first modified cDNA strand, which comprises one or more, typically a plurality of uracil residues in its sequence;

c) incorporating a second adapter, which is different to the first adapter, at the 3' end of the first cDNA strand, wherein the second adapter comprises at least a 3' portion which is not specific to the mRNA molecule and which comprises one or more uracil residues in the sequence;

d) copying said first modified cDNA strand by extending from a first primer which comprises one or more uracils and comprises at least a 3' portion which is substantially complementary to a portion of the second adapter sequence, using dNTPs which include dUTP, thereby generating a modified double stranded cDNA molecule, which includes one or more, typically a plurality of uracil residues in the first and second strands and adapter sequences 5' and 3' of sequence corresponding to the 5'and 3' ends of the mRNA molecule.

Conveniently the method may be conducted on a pool of mRNA molecules, thereby generating a single pool of cDNA molecules, each cDNA molecule from the pool comprising the same 5' and 3' adapter sequences respectively

It is to be understood that the term "modified cDNA molecule" relates to a cDNA molecule which comprises one or more, typically many uracil molecules throughout its sequence. Thus, the modified cDNA molecules of the present invention comprise uracils in place of many thymidine residues. However, not all thymidine residues need to be replaced by a uracil and indeed the present inventors have observed that it can be desirable not to replace all thymidine residues with a uracil. Thus, typically at least 1 in 1, to 1 in 20, such as 1 in 1 to 1, in 4, especially 1 in 1.05 to 1 in 2 thymidine residues are replaced by a uracil residue.

The first adapter oligonucleotide typically comprises a 3' portion which comprises a stretch of T residues, so as to be able to complimentary base pair with the poly A tail of the mRNA molecule. Typically the stretch will be 10-40 Ts in length, such as 15-30 Ts in length. The 3' end of the first adapter may be a T or may incorporate other residues to ensure proper annealing and extension from the first adapter. Indeed a pool of first adapters may be provided which include possible variations at the 3' end so as to ensure proper annealing and extension from the adapter. The 5' end of the first adapter sequence is designed to be not complimentary to eukaryotic sequences based on blast homology searches and generally not to be specific to the mRNA sequence.

The term "substantially complimentary" as used throughout the description, is to be understood to mean that complimentary is such that specific binding at high stringency, (see for example ref 3) for what is generally understood as high stringency, but the skilled addressee is well aware that relevant sequences need not be 100% complimentary. Thus, a small degree of variation may be permitted. Typically any nucleotide sequence as described herein will comprise a complimentary portion which is at least 95%, such as 97-99% identical with the sequence to which it is to bind.

After denaturation, typically by heat, of the RNA, the poly dT sequence of the first adapter base pairs to the polyA sequence of the mRNA.

The extension from the first adapter is carried out using dNTPs which comprise dUTP, but as mentioned above, conveniently dTTP may also be present, albeit in much lower (e.g. <10%) amounts than dUTP, in the dNTP mix. The extension reaction may be catalysed using a variety of reverse transcriptases, such as avian myeloblastosis virus (AMV) or Moloney murine leukaemia virus (M-MLV). However, in one embodiment, the reverse transcriptase is M-MLV reverse transciptase. As mentioned previously M-

MLV reverse transcriptase adds 2-4 cytosine residues to the 3' ends of cDNA molecules and by using the cap switching techniques previously described (5, 6), it is possible to incorporate a second adapter (containing uracil) at the 3' end of the first cDNA strand.

Other reverse transcriptases may be employed and the second adapter incorporated at 3' end of the first modified cDNA strand using enzymes such as terminal transferase (2). An alternative approach to incorporate a uracil containing adapter in the 3' end of the newly synthesised first strand cDNA is to use the RNA ligase method previously described (3), however to make this compatible with the technique described herein, dUTP should replace dTTP in the dNTP mix used for first strand cDNA synthesis.

Before copying the first cDNA strand, RNA of the RNA/cDNA hybrid may be removed by way of RNAse H treatment to degrade RNA of the RNA/cDNA hybrid and the cDNA purified from degraded RNA and unincorporated adapter molecules by way of affinity purification (e.g. silica-based purification) using a suitable column purification method, such as by using QIA quick (Qiagen Inc., Chatsworth, CA, USA).

The second cDNA strand may be easily generated by copying the first cDNA strand using a first primer (where uracil replaces thymidine) which is able to specifically bind to the second adapter sequence at 3' end of the first cDNA strand, using a suitable DNA polymerase, such as Taq polymerase and a dNTP mix comprising dUTP.

The resulting double stranded cDNA can be purified by way of affinity purification (e.g. silica-based purification) using a suitable column purification method, such as by using QIA quick (Qiagen Inc., Chatsworth, CA, USA) and then used as a template for the specific amplification of the ends of target genes. However, the modified resulting

double stranded cDNA molecule(s) can be amplified further by using said first primer and a second primer which is able to specifically bind to the sequence at the 3' end of the second strand cDNA strand and which second primer comprises one or more uracil residues, and the dNTP mix contains dUTP, so as to generate an amplified cDNA molecule(s), which comprises a plurality of uracil residues within the first and second strands of the amplified cDNA molecule.

Further rounds of amplification may be carried out, such as 20-25 cycles of PCR, so as to increase the cDNA copies.

To specifically amplify the ends of target genes, an asymmetric amplification reaction using a single gene specific primer and standard dNTPs (without dUTP) is performed. The reaction mixture is then treated with UDG to simplify the PCR reaction mixture by degrading non-target cDNAs. A subsequent amplification reaction using a nested gene specific primer in combination with a primer which is complimentary to an adapter sequence which has been incorporated into the 5' or 3' end of the cDNA, during cDNA synthesis, results in specific amplification of an end of the cDNA. This technique provides a novel and reliable method to specifically amplify the 5' and/or 3' ends of cDNAs from a single pool of cDNA.

Thus, in a further aspect, the present invention provides a method of amplifying a 5' or 3' end of an mRNA molecule for enabling identification of sequence at a respective 5' or 3' end of the mRNA molecule, the method comprising:

a) providing the double stranded cDNA molecule or pool of cDNA molecules according to the first aspect, or amplified copies thereof;

b) asymmetrically amplifying the 5' or 3' end of said cDNA molecule or one of said cDNA molecules in the pool using a gene specific primer which is substantially complementary to an internal portion of the cDNA sequence, using dNTPs which do not include dUTP, so as to generate an amplified gene specific single stranded DNA molecule;

c) adding UDG so as to degrade sequences which comprise uracil thereby leaving only an amplified gene specific single stranded DNA molecule;

d) optionally generating an amplified double stranded DNA gene specific product comprising a sequence corresponding to the 5' or 3' end of a specific mRNA molecule using a gene nested specific primer and a primer which comprises a 3' portion which is substantially complementary to the 5' or 3' end of the amplified gene specific single stranded DNA molecule.

Step d) is preferable in order to increase specificity and yield of the amplified 5' or 3' end to be detected

.

The gene specific primers are generally comprised only of DNA bases (i.e. A, C, G and T) and may be designed to hybridise specifically to a desired mRNA sequence. The gene specific primers may be designed and synthesised based on a known mRNA sequence, or may be based on a consensus sequence ascertained from previously sequenced gene sequences from other organisms.

The skilled addressee is very familiar with asymmetric amplification or PCR and a number of cycles (e.g. 10-40) may be performed to linearly amplify the sequence of

interest and hence the 5' or 3' end of the cDNA. It is to be appreciated that the gene specific primer can be chosen to anneal to either particular strand of the modified cDNA molecule and hence generate amplified copies of the 5' or 3' end of a particular mRNA, as desired.

UDG is added so as to degrade the modified cDNA molecule(s) present once the asymmetric amplification step has been completed. Thus only the amplified single stranded DNA molecule which includes a complimentary DNA copy representative of the desired 5' or 3' end of the mRNA molecule remains as a viable target for a second exponential PCR amplification.

Exponential PCR amplification of the desired amplified single stranded DNA molecule is achieved using a second nested gene specific primer and a primer which is able to specifically bind to the sequence at the 5' or 3' end of the amplified single stranded DNA molecule. Further rounds of amplification may be carried out as necessary.

Preferably the 5' end of the mRNA molecule is amplified.

Once an amplified gene specific product comprising a sequence corresponding to the 5' or 3' end of an mRNA molecule has been generated, the appropriately amplified 5' or 3' end sequence can easily be ascertained by conventional cloning and gene sequencing techniques known to the skilled addressee, utilizing a suitable primer.

As well as the methods described hereinabove, the present invention also provides a kit for use in generating the modified cDNA molecule(s) and/or the amplified DNA molecule which includes the amplified 5' or 3' end. A kit for generating suitable modified cDNA molecules may comprise first and second adapter molecules, a dNTP

mix which includes dUTP and first and second primer molecules for amplifying the modified cDNA molecule. A kit for generating an amplified DNA molecule which comprises a sequence corresponding to a 5' or 3' end of a mRNA molecule may comprise UDG and a primer for use in amplifying the 5' or 3' end. Such kits may also comprise other suitable reagents such as buffers, salts solutions and the like, as well as instructions for carrying out the methods as described hereinabove.

Detailed Description

The present invention will now be further described with reference to the figures which show:

Figure 1 shows a diagrammatic representation of a double stranded template (T-RACE cDNA) synthesis protocol according to the present invention. First strand cDNA synthesis using a dNTP mix which includes dUTP is primed using an oligo-dT adapter-A primer. After reaching the 5' end of the mRNA (dashed line), oligo(dC) is added to the end of the cDNA by the M-MLV reverse transcriptase. Base pairing between the oligo(dC) and the oligo(dG) of adapter-B occurs, and through the process of template switching, the reverse compliment of adapter-B (adapter B') is incorporated into the 3' end of the newly synthesised cDNA as previously described (5, 6). Adapter A and B both contain dUTP instead of dTTP. Adapter A and B serve as priming sites for PCR amplification using primers A and B, which also have dTTP replaced by dUTP. PCR is performed using a dNTP mix which contains dUTP, so the resulting double stranded cDNA (T-RACE ready cDNA) contains dUTP residues throughout and dUTP containing adapters at the 3' and 5' ends.

Figure 2 shows a diagrammatic representation of the 5' specific amplification of cDNA ends using T-RACE. T-RACE ready cDNA serves as template for an asymmetric PCR reaction using a single gene specific primer (GSP1) and standard dNTPs (without dUTP). This reaction produces single stranded target cDNAs which include the adapter sequence comprised of standard dNTPs. UDG treatment degrades all non-target cDNAs and adapters containing dUTP. Subsequent PCR using a nested gene specific primer 2 (GSP2) with the 5' T-RACE primer results in the specific amplification of target cDNA ends.

Figure 3 shows (A) Specific amplification of 5' cDNA ends from Atlantic salmon (Ss) and Zebrafish (Dr) fast muscle. Lane 1 STAC3 (Ss), Lane 2 IGF-I (Ss) Lane 3 CMYA5 (Dr), Lane 4 MAFbx (Ss), Lane 5 STAC3 (Ss). DNA size marker of 5000, 2000, 850, 400 and 200 bp is shown in Lane M. In all cases, a single amplicon was produced, except for STAC3 (Lane 1) where two alternative splice variants were obtained. 5' T-RACE successfully amplified IGF-I (Lane 2) which is expressed at low levels in Atlantic salmon muscle, and a long transcript, CMYA5 (Lane 3), which is over 6kb in length. To demonstrate that long amplicons can be amplified using the T-RACE method, we amplified ~1.9 kb from STAC3 (Lane 5). (B) To demonstrate that the single pool of cDNA can be used for both 5' and 3' amplification of cDNA ends, we amplified the 3' ends of Zebrafish CMYA5 (Lane 1) and Atlantic salmon MAFbx (Lane 2) and STAC3 (Lane 3). DNA size marker of 5000, 2000, 850, 400 and 200 bp is shown in Lane M.

Figure 4 shows for background purposes non-specific amplification of products without UDG treatment. To demonstrate the increased specificity obtained by UDG treatment, we performed the same reactions to those in figure 1A but without UDG treatment prior to thermal cycling.

MATERIALS AND METHODS

RNA extraction

Total RNA was extracted from 100mg of Atlantic salmon (*Salmo salar L.*) or zebrafish (*Danio rerio*) fast skeletal muscle tissue using Lysing matrix D (Qbiogene, Irvine, CA) with 1ml Tri Reagent (Sigma, Gillingham,Dorset, UK) and homogenised using a Fast Prep instrument (Qbiogene, Irvine, CA). Total RNA was quantified based on absorbance at 260nm. Genomic DNA contamination was removed by treatment with Turbo DNA-free (Ambion, Austin, Texas, USA), and the integrity of purified RNA confirmed by agarose gel electrophoresis.

T-RACE ready cDNA synthesis

The T-RACE ready protocol is depicted in figure 1 and the primers used listed in table 1.

First strand cDNA synthesis incorporating dUTP and 5' and 3' adapter sequences

First strand cDNA synthesis was performed using a modified dNTP mixture which contained dUTP and a MMLV reverse transcriptase. cDNA synthesis was primed with Adapter A (table 1) using 1µg total RNA. The RNA, Adapter A and dNTPs were incubated at 65 °C for 5 minutes to denature the RNA and then chilled on ice before addition of the remaining components in the following reaction mixture (20 µl): 2.5 uM Adapter A, 50 mM Tris-HCl pH 8.3, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, 1.0 mM dUTP, 0.5 mM dATP, 0.5 mM dCTP, 0.5 mM dGTP, 0.1mM dTTP, 40 U RNAse inhibitor (RNase OUT, Invitrogen, Carlsbad, CA, USA), 200 U Superscript II (Invitrogen, Carlsbad, CA, USA) and incubated for 1 h at 42 °C. After incubation, 0.4 µl of 100 mM $MnCl_2$, and 1 µl of Adapter B (10 µM) was added to the reaction which was further

incubated for 15 min at 42°C. The reaction was then heat inactivated at 70 °C for 10 minutes. Degradation of the RNA in the RNA/cDNA hybrid was achieved by addition of 1µl RNAse H (Invitrogen, Carlsbad, CA, USA) and incubation at 37 °C for 15 minutes. The reaction was purified through a QIAquick PCR purification column and eluted in 30 µl elution buffer as per manufacturer's guidelines (Qiagen Inc., Chatsworth, CA, USA).

Second strand synthesis incorporating dUTP by PCR

Second strand synthesis was achieved by PCR using the priming sites incorporated at the 3' (Adapter B) and 5' (Adapter A) ends of the first strand cDNA. 2 µl of the purified first strand cDNA was used as template, and amplified in the following reaction (50 µl): 2.5 mM MgCl2, 1.0 mM dUTP, 0.5 mM dATP, 0.5 mM dCTP, 0.5 mM dGTP, 0.1mM dTTP, 0.5 µM primer A, 0.5 µM primer B (table 1). The reaction mixture was mixed and heated to 95 °C before addition of 5 units of Taq DNA polymerase (Bioline, London, UK). The reaction mixture was then incubated at 95 °C for a further 2 minutes, followed by 20 cycles of 95 °C denaturing for 30 s, annealing at 60 °C for 30 s, and extension at 72 °C 12 minutes. The reaction was then chilled on ice and purified through a QIAquick PCR purification column (Qiagen Inc., Chatsworth, CA, USA) and eluted in 100ul elution buffer as per manufacturer's guidelines. The purified double stranded cDNA, incorporating dUTP and 5' and 3' adapter sequences (T-RACE ready cDNA) serves as template for the following PCR reactions.

T-RACE PCR

The PCR reactions for the targeted amplification of cDNA ends are shown in figure 2, and the primers used listed in table 1.

Asymmetric PCR using standard dNTPs

Asymmetric PCR was performed using a gene specific primer (GSP1) and standard dNTPs (dATP, dTTP, dGTP, dCTP) in the following hot start 20 µl reaction: 0.5 µM GSP1, 2µl T-RACE ready cDNA, 16 mM $(NH_4)_2SO_4$, 67 mM Tris-HCl, 3 mM $MgCl_2$, 0.25 mM dNTPs. The reaction mixture was heated to 95 °C for 2.5 min before addition of 1 unit of Taq DNA polymerase (Bioline, London, UK), and then cycled 20 times with the following conditions: 95 °C 30 s, 60 °C 30 s, 72 °C 2 min. The reaction was chilled on ice, purified through a MinElute PCR purification column (Qiagen Inc., Chatsworth, CA, USA) and eluted in 11µl water. The length of time for the 72 °C extension step of the PCR should be varied according to expected amplicon length.

Uracil DNA glycosylase degradation of non target dUTP containing cDNAs

The entire asymmetric reaction is used in the following steps using a nested gene specific primer (GSP2), and the primers used for PCR do not contain dUTP. cDNAs incorporating dUTP were removed by UDG treatment to leave only those cDNAs synthesised with standard dNTPs during the asymmetric PCR. The following 17.8 µl reaction was performed: 10 µl asymmetric PCR product, 2 µl (2 units) Uracil DNA glycosylase, 2 µl 10 x PCR buffer (160 mM $(NH_4)_2SO_4$, 670 mM Tris-HCl), 1.2 µl 50 mM $MgCl_2$, 1 µl nested GSP2, 1µl of 10 µM T-RACE primer (table 1), 0.6 µl water, and incubated at 37 °C for 30 minutes. Depending on whether 3' or 5' cDNA ends are targeted, the 3' or 5' T-RACE primer is used for amplification.

Targeted amplification of cDNA ends (T-RACE PCR)

The UDG treated cDNA mixture was heated to 95 °C before addition of 2 µl 2.5 mM standard dNTPs and 0.2 µl (1unit) of Taq DNA polymerase (Bioline, London, UK). The reaction was denatured at 95 °C for 5 min followed by 35 cycles of 95 °C 30 s, 60 °C 30s, 72 °C 2 min, followed by a final extension of 7 min at 72 °C. The length of time for

the 72 °C extension step of the PCR should be varied according to expected amplicon length.

Cloning

PCR products were separated on a 1.2% (m/v) agarose gel, and the bands excised and purified using a QIAquick gel extraction kit (Qiagen Inc., Chatsworth, CA, USA). The PCR products were cloned into a T/A pCR4-TOPO vector (Invitrogen, Carlsbad, CA, USA) and transformation of chemically competent TOP10 *Escherichia coli* cells (Invitrogen, Carlsbad, CA, USA). Individual colonies were grown and plasmids purified using a QIAprep Spin Miniprep Kit (Qiagen Inc., Chatsworth, CA, USA). Sequencing was performed using T3 and T7 sequencing primers with Big Dye terminator v3.1 sequencing chemistry (Applied Biosystems, Foster City, CA, USA) at the University of Dundee Sequencing Facility.

Table 1. Primer and adapter sequences used for first and second strand cDNA synthesis and specific amplification of cDNA ends.

Primer	Sequence (5' – 3')	Application
Adapter A	AUAUCUCGAGUUCGCGCCGGAUCC$(T)_{25}$VN	cDNA synthesis and adapter-A incorporation
Adapter B	AUAUGCACUGCCGCGUCUGAGGGGGGGG	Cap finder adapter-B incorporation
Primer A	AUAUCUCGAGUUCGCGCCGGAU	2nd strand cDNA synthesis
Primer B	AUAUGCACUGCCGCGUCUGA	2nd strand cDNA synthesis
5' T-RACE primer	ATATGCACTGCCGCGTCTGA	5' Specific Amplification of cDNA ends

3' T-RACE primer	ATATCTCGAGTTCGCGCCGGAT	3' Specific amplification of cDNA ends

RESULTS AND DISCUSSION

We report the development of a novel method for acquiring the ends especially the 5' ends of cDNA molecules using a modification of previously published methods. Using this approach, we have identified the 5' ends of cDNA molecules expressed in muscle of zebrafish and Atlantic salmon (Fig. 3A). The improved method involves the production of double stranded cDNA incorporating dUTP, and the addition of 5' and 3' adapters which also contain dUTP (Fig. 1). By producing the second strand of cDNA, a priming site for a gene specific primer for 5' RACE is introduced, allowing for an asymmetric PCR reaction to be performed using standard dNTP mix. This asymmetric PCR using only a gene specific primer produces a cDNA fragment which includes the 5' adapter sequence, and is comprised of standard dNTPs (Fig. 2). Degradation of dUTP containing cDNA molecules by Uracil DNA glycosylase (UDG) treatment leaves only the molecules comprised of standard dNTPs primed by the gene specific primer (Fig. 2). This now serves as the template for PCR amplification using a nested gene specific primer and a 5' T-RACE primer, thus specifically amplifying the 5'end of the desired cDNA sequence.

To demonstrate the effectiveness of our approach, we amplified the 5' ends of several genes of differing lengths whose mRNAs are present in varying abundance (Fig. 3A). For example, we amplified the 5' ends of zebrafish *myospryn* (*CMYA5*), a gene over 6kb in length, Atlantic salmon *MAFbx*, and *insulin-like growth factor-I* (*IGF-I*) which is expressed at low levels in muscle (Fig 3A). We also demonstrated that long amplicons

can be generated using this method by amplifying the ~1.9 kb 5' end of *STAC3* (Fig. 3A). Using GSPs and the 5' T-RACE primer, in all cases we were able to obtain a single amplicon (except for *STAC3* where we identified two alternatively spliced transcripts) of the correct size (Fig. 3A), the identity of which was confirmed by sequencing. We also used the same cDNA pool to amplify the 3' ends for a number of genes (Fig. 3B), demonstrating that a single cDNA pool can be used for both 5' and 3' amplification of cDNA ends.

The current method has several advantages over existing methodologies. Standard RACE methods often result in non specific amplification giving false positive results (6, 7, 8). When cDNA is synthesised using oligo dT, the subsequent incorporation of the adapter sequence for 5' RACE (through terminal transferase tailing or cap finder methods) occurs in every cDNA molecule. Due to this, mispriming events during PCR will often lead to exponential amplification, and smears, rather than distinct bands are often obtained. The use of a nested PCR reaction should increase the specificity of the RACE reaction, however non-specific products are frequently obtained. The T-RACE method described herein degrades all of the non-target cDNA molecules, except for those that were primed by the gene specific primer and synthesised using standard dNTPs, thereby significantly simplifying the templates present in the PCR reaction. As the gene specific primer is used in an asymmetric PCR, where higher annealing temperatures are permitted, rather than during cDNA synthesis, the specificity of the reaction is increased. By using a nested primer in the second PCR reaction, specific amplification of the desired molecule is achieved.

During the development of this method, we made several changes to our original protocol which improved the efficiency of the reactions. Taq polymerase is less efficient at incorporating dUTP than dTTP (10) which could impact on the double

stranded cDNA synthesis. It was observed that a ratio of dUTP:dTTP of 8:1, improved the efficiency of the T-RACE reaction, but still resulted in sufficient degradation of non-target cDNAs to give specific amplification. We also observed an improvement in the specificity of the PCR reactions by adding the $MnCl_2$ and adapter-A after cDNA synthesis has completed. Presumably this modification decreases the chances of adapter-A priming cDNA synthesis during the first strand cDNA reaction which could lead to spurious amplification products in subsequent reactions. Performing all reactions as a hot-start (adding Taq DNA polymerase after reactions have reached 95°C) improved specificity and this could also be achieved using a commercially available hot-start enzyme. Originally our method did not use a PCR (exponential) amplification step during double stranded cDNA synthesis, and we were unable to amplify rare transcripts such as IGF-I from this template. This is likely to be due to the fact that the asymmetric PCR performed using standard dNTPs, which is critical to the success of this method, does not result in exponential amplification of products. However, performing 20 cycles of PCR for the second strand synthesis improved the outcomes for amplifying rare transcripts, most likely by greatly increasing the number of targets available for the asymmetric PCR. The use of Uracil DNA glycosylase and dUTP has been widely used to control PCR product carryover contamination (9, 10), and its effectiveness in simplifying 5'RACE reactions is clearly demonstrated here (Fig. 3A and 4). Nested PCR reactions, which were not treated with UDG prior to thermal cycling resulted in smears and non specific amplification products (Fig. 3), whereas treatment with UDG resulted in specific amplification (Fig. 3A).

The method described herein offers several advantages over other protocols that are currently in use. Using UDG treated PCR products as template results in specific amplification, which in most cases results in a single amplicon of the correct size. We

have now used this technique to amplify the 5' ends of several cDNAs, each one from the same T-RACE ready cDNA pool using total RNA as starting material.

REFERENCES

1. Salehi-Ashtiani, K., Lin, C.W., Hao, T., Shen, Y., Szeto, D., Yang, X.P., Ghamsari, L., Lee, H., Fan, C.Y., Murray, R.R. et al. (2009) Large-scale RACE approach for proactive experimental definition of C. elegans ORFeome. *Genome Res.*, **19**, 2334-2342.

2. Frohman, M.A., Dush, M.K. and Martin, G.R. (1988) Rapid Production of Full-Length Cdnas from Rare Transcripts - Amplification Using a Single Gene-Specific Oligonucleotide Primer. *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 8998-9002.

3. Fromont-Racine, M., Bertrand, E., Pictet, R. and Grange, T. (1993) A Highly Sensitive Method for Mapping the 5' Termini of mRNAs. *Nucleic Acids Res.*, **21**, 1683-1684.

4. Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989) In *Molecular cloning: a laboratory manual*, 2d ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp. 1.74–1.84, 7.37–7.52.

5. Schmidt, W.M. and Mueller, M.W. (1999) CapSelect: a highly sensitive method for 5' CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. *Nucleic Acids Res.*, **27**, e31.

6. Schramm, G., Bruchhaus, I. and Roeder, T. (2000) A simple and reliable 5'-RACE approach. Nucleic Acids Res., **28**, e96

7. Schaefer, B.C. (1995) Revolutions in Rapid Amplification of Cdna Ends - New Strategies for Polymerase Chain-Reaction Cloning of Full-Length Cdna Ends. *Anal Biochem.*, **227**, 255-273.

8. Lipovich, L., Lynch, E.D., Lee, M.K. and King, M.C. (2001) A novel sodium bicarbonate cotransporter-like gene in an ancient duplicated region: SLC4A9 at 5q31 . *Genome Biol.*, **2**:research0011.1-0011.13..

9. Kwok, S. and Higuchi, R. (1989) Avoiding False Positives with Pcr. *Nature.*, **339**, 237-238.

10. Longo, M.C., Berninger, M.S. and Hartley, J.L. (1990) Use of Uracil DNA Glycosylase to Control Carry-over Contamination in Polymerase Chain-Reactions. *Gene.*, **93**, 125-128.

11. Buchman, G.W., Schuster, D.M. and Rashtchian, A. (1993) Selective Rna Amplification - a Novel Method Using Dump-Containing Primers and Uracil DNA Glycosylase. *PCR Methods Appl.*, **3**, 28-31.

Claims

1. a method for generating a modified cDNA molecule comprising sequence corresponding to the 5' and 3' ends of an mRNA molecule comprising the steps of:

e) annealing a first adapter oligonucleotide to an mRNA molecule, said first adapter comprising a 3' portion which is substantially complimentary to the 3' poly A tail of the mRNA molecule and a 5' portion which comprises one or more uracil residues;

f) extending from said first adapter, utilizing dNTPs which include dUTP, so as to generate a first modified cDNA strand, which comprises one or more, typically a plurality of uracil residues in its sequence;

g) incorporating a second adapter, which is different to the first adapter, at the 3' end of the first cDNA strand, wherein the second adapter comprises at least a 3' portion which is not specific to the mRNA molecule and which comprises one or more uracil residues in the sequence;

h) copying said first modified cDNA strand by extending from a first primer which comprises one or more uracils and comprises at least a 3' portion which is substantially complementary to a portion of the second adapter sequence, using dNTPs which include dUTP, thereby generating a modified double stranded cDNA molecule, which includes one or more, typically a plurality of uracil residues in the first and second strands and adapter sequences 5' and 3' of sequence corresponding to the 5'and 3' ends of the mRNA molecule.

2. The method according to claim 1 wherein the method may be conducted on a pool of mRNA molecules, thereby generating a single pool of cDNA molecules, each cDNA molecule from the pool comprising the same 5' and 3' adapter sequences respectively.

3. The method according to either of claims 1 or 2 wherein at least 1 in 1, to 1 in 20, such as 1 in 1 to 1, in 4, especially 1 in 1.05 to 1 in 2 thymidine residues of the cDNA are replaced by a uracil residue.

4. The method according to any preceding claim wherein the first adapter oligonucleotide comprises a 3' portion which comprises a stretch of T residues, so as to be able to complimentary base pair with the poly A tail of the mRNA molecule.

5. The method according to any preceding claim wherein the extension reaction is catalysed by a reverse transcriptase, such as avian myeloblastosis virus (AMV) or Moloney murine leukaemia virus (M-MLV).

6. The method according to claim 5 wherein the reverse transcriptase is MMLV and the second adapter is added using a cap switching technique.

7. The method according to claim 5 wherein the second adapter is added using terminal transferase or RNA ligase.

8. The method according to any preceding claim wherein the second cDNA strand is generated by copying the first cDNA strand using a first primer (where uracil replaces thymidine) which is able to specifically bind to the second adapter sequence at 3' end of the first cDNA strand, using a suitable DNA polymerase, such as Taq polymerase and a dNTP mix comprising dUTP.

9. A method of amplifying a 5' or 3' end of an mRNA molecule for enabling identification of sequence at a respective 5' or 3' end of the mRNA molecule, the method comprising:

e) providing the double stranded cDNA molecule or pool of cDNA molecules according to any preceding claim, or amplified copies thereof;

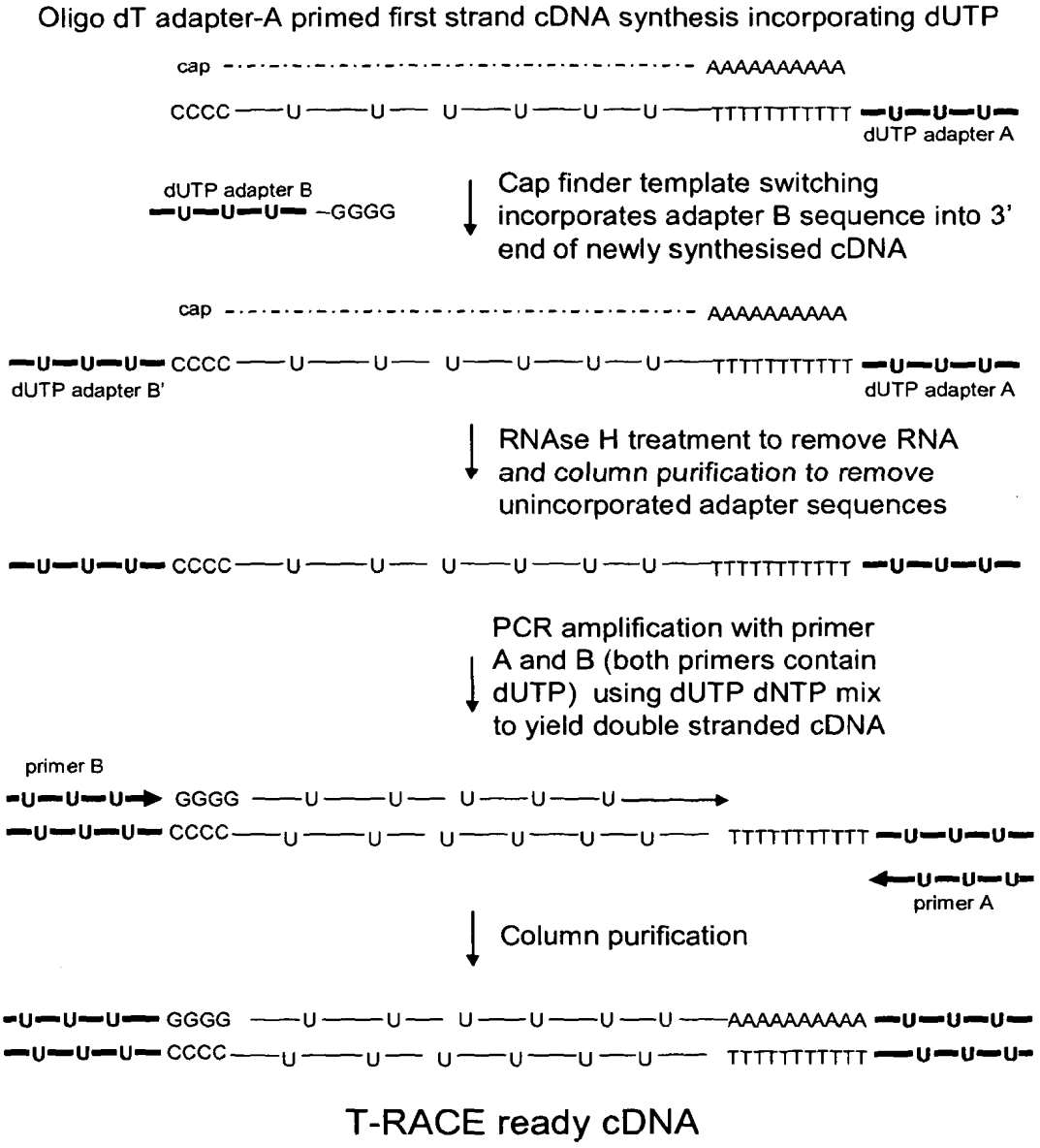
f) asymmetrically amplifying the 5' or 3' end of said cDNA molecule or one of said cDNA molecules in the pool using a gene specific primer which is substantially complementary to an internal portion of the cDNA sequence, using dNTPs which do not include dUTP, so as to generate an amplified gene specific single stranded DNA molecule;

g) adding UDG so as to degrade sequences which comprise uracil thereby leaving only an amplified gene specific single stranded DNA molecule;

h) optionally generating an amplified double stranded DNA gene specific product comprising a sequence corresponding to the 5' or 3' end of a specific mRNA molecule using a gene nested specific primer and a primer which comprises a 3' portion which is substantially complementary to the 5' or 3' end of the amplified gene specific single stranded DNA molecule.

10. The method according to claim 9 wherein **[step?]** (d) is included.

11. The method according to claims 9 or 10 wherein exponential amplification of the desired amplified single stranded DNA molecule is achieved using a second nested



T-RACE ready cDNA

Figure 1

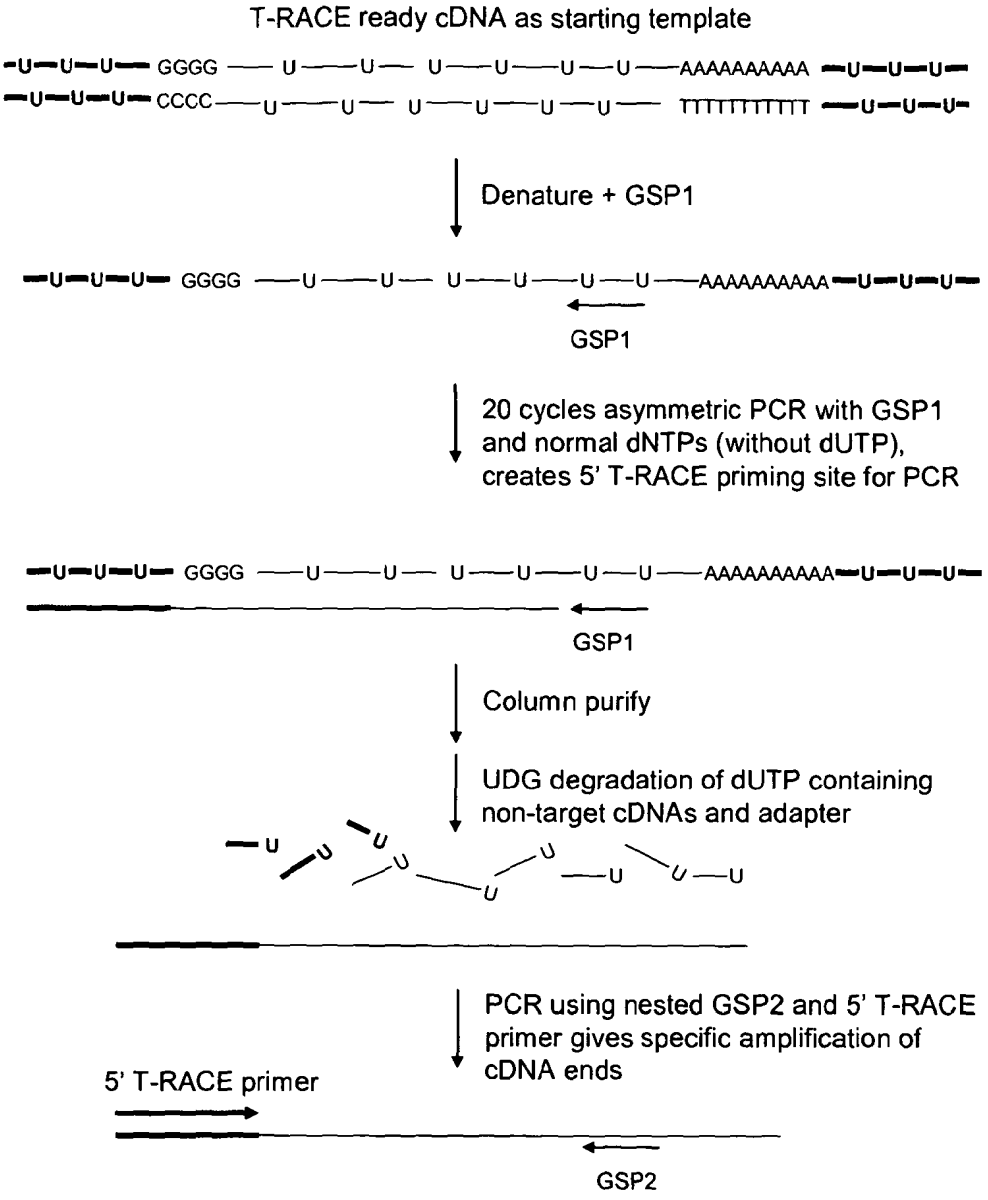

T-RACE ready cDNA as starting template
GGGG
AAAAAAAAAA
CCCC
TTTTTTTTTT
Denature + GSP1
GSP1
20 cycles asymmetric PCR with GSP1 and normal dNTPs (without dUTP), creates 5' T-RACE priming site for PCR
GSP1
Column purify
UDG degradation of dUTP containing non-target cDNAs and adapter
PCR using nested GSP2 and 5' T-RACE primer gives specific amplification of cDNA ends
5' T-RACE primer
GSP2

Figure 2

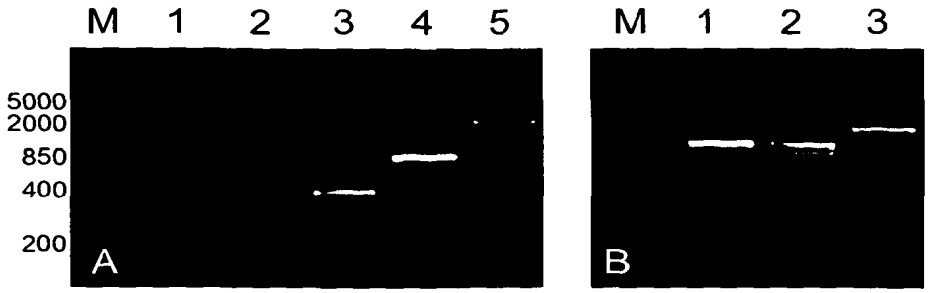


Figure 3

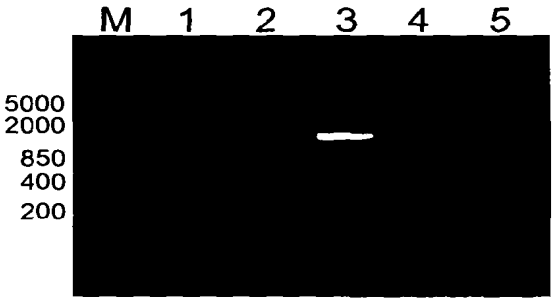


Figure 4

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2011/001131

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12N15/10
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)
C12N

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 practical, search terms used)
EPO-Internal, EMBASE, BIOSIS, WPI Data, PAJ

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	SCHRAMM G ET AL: "A simple and reliable 5'-RACE approach.", NUCLEIC ACIDS RESEARCH, vol. 28, no. 22, E96, 15 November 2000 (2000-11-15), pages 1-4, XP002661255, ISSN: 1362-4962 cited in the application figure 1 ----- -/--	1-10

[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 document 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
14 October 2011	15/11/2011

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	Weinberg, Suzanna

4

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

IDT Exhibit 1020.4000

INTERNATIONAL SEARCH REPORT

International application No.
PCT/GB2011/001131

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.: 11
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
see FURTHER INFORMATION sheet PCT/ISA/210

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

IDT Exhibit 1020.4001

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Continuation of Box II.2

Claims Nos.: 11

Claim 11 is incomplete and hence lacks clarity to the extent that no additional feature can be determined, which feature can be searched.

The applicant's attention is drawn to the fact that claims relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure. If the application proceeds into the regional phase before the EPO, the applicant is reminded that a search may be carried out during examination before the EPO (see EPO Guideline C-VI, 8.2), should the problems which led to the Article 17(2) declaration be overcome.

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	LONGO M C ET AL: "Use of uracil DNA glycosylase to control carry-over contamination in polymerase chain reactions", GENE, vol. 93, no. 1, 1 September 1990 (1990-09-01), pages 125-128, XP023545615, ELSEVIER, AMSTERDAM, NL ISSN: 0378-1119, DOI: 10.1016/0378-1119(90)90145-H [retrieved on 1990-09-01] cited in the application page 126, left-hand column	1-10
X,P	BOWER NEIL I ET AL: "Targeted rapid amplification of cDNA ends (T-RACE)--an improved RACE reaction through degradation of non-target sequences.", NUCLEIC ACIDS RESEARCH, vol. 38, no. 21, E194, November 2010 (2010-11), pages 1-7, XP002661009, ISSN: 1362-4962, DOI: 10.1093/nar/gkq816 the whole document	1-10
A	BUCHMAN G W ET AL: "Selective RNA amplification: a novel method using dUMP-containing primers and uracil DNA glycosylase.", PCR METHODS AND APPLICATIONS, vol. 3, no. 1, August 1993 (1993-08), pages 28-31, XP002661010, ISSN: 1054-9803 cited in the application figure 1	1-10
A	WO 92/18521 A1 (LIFE TECHNOLOGIES INC [US]) 29 October 1992 (1992-10-29) page 22 - page 23; figure 4	1-10
A	SCHMIDT W M ET AL: "CapSelect: a highly sensitive method for 5' CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs.", NUCLEIC ACIDS RESEARCH, vol. 27, no. 21, E31, 1 November 1999 (1999-11-01), pages I-IV, XP002661269, ISSN: 1362-4962 cited in the application abstract; figure 1	1-10

4

IDT Exhibit 1020.4003

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/GB2011/001131

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9218521 A1	29-10-1992	US 5334515 A	02-08-1994

Form PCT/ISA/210 (patent family annex) (April 2005)

IDT Exhibit 1020.4004

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 March 2014 (13.03.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/039556 A1

(51) **International Patent Classification**:
C12Q 1/68 (2006.01) *C12N 15/11* (2006.01)

(21) **International Application Number**:
PCT/US2013/058061

(22) **International Filing Date**:
4 September 2013 (04.09.2013)

(25) **Filing Language**: English

(26) **Publication Language**: English

(30) **Priority Data**:

61/696,734	4 September 2012 (04.09.2012)	US
61/704,400	21 September 2012 (21.09.2012)	US
61/793,997	15 March 2013 (15.03.2013)	US
61/845,987	13 July 2013 (13.07.2013)	US

(71) **Applicant** *(for all designated States except US)*: **GUARDANT HEALTH, INC.** [US/US]; 2686 Middlefield Road, Suite D, Redwood City, CA 94063 (US).

(72) **Inventors; and**
(71) **Applicants** *(for US only)*: **TALASAZ, AmirAli** [US/US]; 2181 Camino A Los Cerros, Menlo Park, CA 94025 (US). **ELTOUKHY, Helmy** [US/US]; 2 Barry Lane, Atherton, CA 94027 (US).

(74) **Agents**: **ALEMOZAFAR, Ali, R.** et al.; Wilson Sonsini Goodrich & Rosati, 650 Page Mill Road, Palo Alto, CA 94304-1050 (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, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, 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).

Published:

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

WO 2014/039556 A1

(54) **Title**: SYSTEMS AND METHODS TO DETECT RARE MUTATIONS AND COPY NUMBER VARIATION

(57) **Abstract**: The present disclosure provides a system and method for the detection of rare mutations and copy number variations in cell free polynucleotides. Generally, the systems and methods comprise sample preparation, or the extraction and isolation of cell free polynucleotide sequences from a bodily fluid; subsequent sequencing of cell free polynucleotides by techniques known in the art; and application of bioinformatics tools to detect rare mutations and copy number variations as compared to a reference. The systems and methods also may contain a database or collection of different rare mutations or copy number variation profiles of different diseases, to be used as additional references in aiding detection of rare mutations, copy number variation profiling or general genetic profiling of a disease.

IDT Exhibit 1020.4005

SYSTEMS AND METHODS TO DETECT RARE MUTATIONS AND COPY NUMBER VARIATION

CROSS-REFERENCE

[0001] This application claims priority to U.S. Provisional Patent Application No. 61/696,734, filed September 4, 2012, U.S. Provisional Patent Application No. 61/704,400, filed September 21, 2012, U.S. Provisional Patent Application No. 61/793,997, filed March 15, 2013, and U.S. Provisional Patent Application No. 61/845,987, filed July 13, 2013, each of which is entirely incorporated herein by reference for all purposes.

BACKGROUND OF THE INVENTION

[0002] The detection and quantification of polynucleotides is important for molecular biology and medical applications such as diagnostics. Genetic testing is particularly useful for a number of diagnostic methods. For example, disorders that are caused by rare genetic alterations (e.g., sequence variants) or changes in epigenetic markers, such as cancer and partial or complete aneuploidy, may be detected or more accurately characterized with DNA sequence information.

[0003] Early detection and monitoring of genetic diseases, such as cancer is often useful and needed in the successful treatment or management of the disease. One approach may include the monitoring of a sample derived from cell free nucleic acids, a population of polynucleotides that can be found in different types of bodily fluids. In some cases, disease may be characterized or detected based on detection of genetic aberrations, such as a change in copy number variation and/or sequence variation of one or more nucleic acid sequences, or the development of other certain rare genetic alterations. Cell free DNA ("cfDNA") has been known in the art for decades, and may contain genetic aberrations associated with a particular disease. With improvements in sequencing and techniques to manipulate nucleic acids, there is a need in the art for improved methods and systems for using cell free DNA to detect and monitor disease.

SUMMARY OF THE INVENTION

[0004] The disclosure provides for a method for detecting copy number variation comprising: a) sequencing extracellular polynucleotides from a bodily sample from a subject, wherein each of the extracellular polynucleotide are optionally attached to unique barcodes; b) filtering out reads that fail to meet a set threshold; c) mapping sequence reads obtained from step (a) to a reference

IDT Exhibit 1020.4006

sequence; d) quantifying/counting mapped reads in two or more predefined regions of the reference sequence; e) determining a copy number variation in one or more of the predefined regions by (i) normalizing the number of reads in the predefined regions to each other and/or the number of unique barcodes in the predefined regions to each other; and (ii) comparing the normalized numbers obtained in step (i) to normalized numbers obtained from a control sample.

[0005] The disclosure also provides for a method for detecting a rare mutation in a cell-free or substantially cell free sample obtained from a subject comprising: a) sequencing extracellular polynucleotides from a bodily sample from a subject, wherein each of the extracellular polynucleotide generate a plurality of sequencing reads; b) sequencing extracellular polynucleotides from a bodily sample from a subject, wherein each of the extracellular polynucleotide generate a plurality of sequencing reads; sequencing extracellular polynucleotides from a bodily sample from a subject, wherein each of the extracellular polynucleotide generate a plurality of sequencing reads; c) filtering out reads that fail to meet a set threshold; d) mapping sequence reads derived from the sequencing onto a reference sequence; e) identifying a subset of mapped sequence reads that align with a variant of the reference sequence at each mappable base position; f) for each mappable base position, calculating a ratio of (a) a number of mapped sequence reads that include a variant as compared to the reference sequence, to (b) a number of total sequence reads for each mappable base position; g) normalizing the ratios or frequency of variance for each mappable base position and determining potential rare variant(s) or mutation(s); h) and comparing the resulting number for each of the regions with potential rare variant(s) or mutation(s) to similarly derived numbers from a reference sample.

[0006] Additionally, the disclosure also provides for a method of characterizing the heterogeneity of an abnormal condition in a subject, the method comprising generating a genetic profile of extracellular polynucleotides in the subject, wherein the genetic profile comprises a plurality of data resulting from copy number variation and/or other rare mutation (e.g., genetic alteration) analyses.

[0007] In some embodiments, the prevalence/concentration of each rare variant identified in the subject is reported and quantified simultaneously. In other embodiments, a confidence score, regarding the prevalence/concentrations of rare variants in the subject, is reported.

[0008] In some embodiments, extracellular polynucleotides comprise DNA. In other embodiments, extracellular polynucleotides comprise RNA. Polynucleotides may be fragments or

IDT Exhibit 1020.4007

fragmented after isolation. Additionally, the disclosure provides for a method for circulating nucleic acid isolation and extraction.

[0009] In some embodiments, extracellular polynucleotides are isolated from a bodily sample that may be selected from a group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool and tears.

[0010] In some embodiments, the methods of the disclosure also comprise a step of determining the percent of sequences having copy number variation or other rare genetic alteration (e.g., sequence variants) in said bodily sample.

[0011] In some embodiments, the percent of sequences having copy number variation in said bodily sample is determined by calculating the percentage of predefined regions with an amount of polynucleotides above or below a predetermined threshold.

[0012] In some embodiments, bodily fluids are drawn from a subject suspected of having an abnormal condition which may be selected from the group consisting of, mutations, rare mutations, single nucleotide variants, indels, copy number variations, transversions, translocations, inversion, deletions, aneuploidy, partial aneuploidy, polyploidy, chromosomal instability, chromosomal structure alterations, gene fusions, chromosome fusions, gene truncations, gene amplification, gene duplications, chromosomal lesions, DNA lesions, abnormal changes in nucleic acid chemical modifications, abnormal changes in epigenetic patterns, abnormal changes in nucleic acid methylation infection and cancer.

[0013] In some embodiments, the subject may be a pregnant female in which the abnormal condition may be a fetal abnormality selected from the group consisting of, single nucleotide variants, indels, copy number variations, transversions, translocations, inversion, deletions, aneuploidy, partial aneuploidy, polyploidy, chromosomal instability, chromosomal structure alterations, gene fusions, chromosome fusions, gene truncations, gene amplification, gene duplications, chromosomal lesions, DNA lesions, abnormal changes in nucleic acid chemical modifications, abnormal changes in epigenetic patterns, abnormal changes in nucleic acid methylation infection and cancer

[0014] In some embodiments, the method may comprise comprising attaching one or more barcodes to the extracellular polynucleotides or fragments thereof prior to sequencing, in which the barcodes comprise are unique. In other embodiments barcodes attached to extracellular polynucleotides or fragments thereof prior to sequencing are not unique.

IDT Exhibit 1020.4008

[0015] In some embodiments, the methods of the disclosure may comprise selectively enriching regions from the subject's genome or transcriptome prior to sequencing. In other embodiments the methods of the disclosure comprise selectively enriching regions from the subject's genome or transcriptome prior to sequencing. In other embodiments the methods of the disclosure comprise non-selectively enriching regions from the subject's genome or transcriptome prior to sequencing.

[0016] Further, the methods of the disclosure comprise attaching one or more barcodes to the extracellular polynucleotides or fragments thereof prior to any amplification or enrichment step.

[0017] In some embodiments, the barcode is a polynucleotide, which may further comprise random sequence or a fixed or semi-random set of oligonucleotides that in combination with the diversity of molecules sequenced from a select region enables identification of unique molecules and be at least a 3, 5, 10, 15, 20 25, 30, 35, 40, 45, or 50mer base pairs in length.

[0018] In some embodiments, extracellular polynucleotides or fragments thereof may be amplified. In some embodiments amplification comprises global amplification or whole genome amplification.

[0019] In some embodiments, sequence reads of unique identity may be detected based on sequence information at the beginning (start) and end (stop) regions of the sequence read and the length of the sequence read. In other embodiments sequence molecules of unique identity are detected based on sequence information at the beginning (start) and end (stop) regions of the sequence read, the length of the sequence read and attachment of a barcode.

[0020] In some embodiments, amplification comprises selective amplification, non-selective amplification, suppression amplification or subtractive enrichment.

[0021] In some embodiments, the methods of the disclosure comprise removing a subset of the reads from further analysis prior to quantifying or enumerating reads.

[0022] In some embodiments, the method may comprise filtering out reads with an accuracy or quality score of less than a threshold, e.g., 90%, 99%, 99.9%, or 99.99% and/or mapping score less than a threshold, e.g., 90%, 99%, 99.9% or 99.99%. In other embodiments, methods of the disclosure comprise filtering reads with a quality score lower than a set threshold.

[0023] In some embodiments, predefined regions are uniform or substantially uniform in size, about 10kb, 20kb, 30kb 40kb, 50kb, 60kb, 70kb, 80kb, 90kb, or 100kb in size. In some embodiments, at least 50, 100, 200, 500, 1000, 2000, 5000, 10,000, 20,000, or 50,000 regions are analyzed.

IDT Exhibit 1020.4009

[0024] In some embodiments, a genetic variant, rare mutation or copy number variation occurs in a region of the genome selected from the group consisting of gene fusions, gene duplications, gene deletions, gene translocations, microsatellite regions, gene fragments or combination thereof. In other embodiments a genetic variant, rare mutation, or copy number variation occurs in a region of the genome selected from the group consisting of genes, oncogenes, tumor suppressor genes, promoters, regulatory sequence elements, or combination thereof. In some embodiments the variant is a nucleotide variant, single base substitution, or small indel, transversion, translocation, inversion, deletion, truncation or gene truncation about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 or 20 nucleotides in length.

[0025] In some embodiments, the method comprises correcting/normalizing/adjusting the quantity of mapped reads using the barcodes or unique properties of individual reads.

[0026] In some embodiments, enumerating the reads is performed through enumeration of unique barcodes in each of the predefined regions and normalizing those numbers across at least a subset of predefined regions that were sequenced. In some embodiments, samples at succeeding time intervals from the same subject are analyzed and compared to previous sample results. The method of the disclosure may further comprise determining partial copy number variation frequency, loss of heterozygosity, gene expression analysis, epigenetic analysis and hypermethylation analysis after amplifying the barcode-attached extracellular polynucleotides.

[0027] In some embodiments, copy number variation and rare mutation analysis is determined in a cell-free or substantially cell free sample obtained from a subject using multiplex sequencing, comprising performing over 10,000 sequencing reactions; simultaneously sequencing at least 10,000 different reads; or performing data analysis on at least 10,000 different reads across the genome. The method may comprise multiplex sequencing comprising performing data analysis on at least 10,000 different reads across the genome. The method may further comprise enumerating sequenced reads that are uniquely identifiable.

[0028] In some embodiments, the methods of the disclosure comprise normalizing and detection is performed using one or more of hidden markov, dynamic programming, support vector machine, Bayesian network, trellis decoding, Viterbi decoding, expectation maximization, Kalman filtering, or neural network methodologies.

[0029] In some embodiments the methods of the disclosure comprise monitoring disease progression, monitoring residual disease, monitoring therapy, diagnosing a condition, prognosing a condition, or selecting a therapy based on discovered variants.

IDT Exhibit 1020.4010

[0030] In some embodiments, a therapy is modified based on the most recent sample analysis. Further, the methods of the disclosure comprise inferring the genetic profile of a tumor, infection or other tissue abnormality. In some embodiments growth, remission or evolution of a tumor, infection or other tissue abnormality is monitored. In some embodiments the subject's immune system are analyzed and monitored at single instances or over time.

[0031] In some embodiments, the methods of the disclosure comprise identification of a variant that is followed up through an imaging test (e.g., CT, PET-CT, MRI, X-ray, ultrasound) for localization of the tissue abnormality suspected of causing the identified variant.

[0032] In some embodiments, the methods of the disclosure comprise use of genetic data obtained from a tissue or tumor biopsy from the same patient. In some embodiments, whereby the phylogenetics of a tumor, infection or other tissue abnormality is inferred.

[0033] In some embodiments, the methods of the disclosure comprise performing population-based no-calling and identification of low-confidence regions. In some embodiments, obtaining the measurement data for the sequence coverage comprises measuring sequence coverage depth at every position of the genome. In some embodiments correcting the measurement data for the sequence coverage bias comprises calculating window-averaged coverage. In some embodiments correcting the measurement data for the sequence coverage bias comprises performing adjustments to account for GC bias in the library construction and sequencing process. In some embodiments correcting the measurement data for the sequence coverage bias comprises performing adjustments based on additional weighting factor associated with individual mappings to compensate for bias.

[0034] In some embodiments, the methods of the disclosure comprise extracellular polynucleotide derived from a diseased cell origin. In some embodiments, the extracellular polynucleotide is derived from a healthy cell origin.

[0035] The disclosure also provides for a system comprising a computer readable medium for performing the following steps: selecting predefined regions in a genome; enumerating number of sequence reads in the predefined regions; normalizing the number of sequence reads across the predefined regions; and determining percent of copy number variation in the predefined regions. In some embodiments, the entirety of the genome or at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, or 90% of the genome is analyzed. In some embodiments, computer readable medium provides data on percent cancer DNA or RNA in plasma or serum to the end user.

IDT Exhibit 1020.4011

[0036] In some embodiments, the amount of genetic variation, such as polymorphisms or causal variants is analyzed. In some embodiments, the presence or absence of genetic alterations is detected.

[0037] The disclosure also provides for a method for detecting a rare mutation in a cell-free or a substantially cell free sample obtained from a subject comprising: a) sequencing extracellular polynucleotides from a bodily sample from a subject, wherein each of the extracellular polynucleotides generate a plurality of sequencing reads; b) filtering out reads that fail to meet a set threshold; c) mapping sequence reads derived from the sequencing onto a reference sequence; d) identifying a subset of mapped sequence reads that align with a variant of the reference sequence at each mappable base position; e) for each mappable base position, calculating a ratio of (a) a number of mapped sequence reads that include a variant as compared to the reference sequence, to (b) a number of total sequence reads for each mappable base position; f) normalizing the ratios or frequency of variance for each mappable base position and determining potential rare variant(s) or other genetic alteration(s); and g) comparing the resulting number for each of the regions

[0038] This disclosure also provides for a method comprising: a. providing at least one set of tagged parent polynucleotides, and for each set of tagged parent polynucleotides; b. amplifying the tagged parent polynucleotides in the set to produce a corresponding set of amplified progeny polynucleotides; c. sequencing a subset (including a proper subset) of the set of amplified progeny polynucleotides, to produce a set of sequencing reads; and d. collapsing the set of sequencing reads to generate a set of consensus sequences, each consensus sequence corresponding to a unique polynucleotide among the set of tagged parent polynucleotides. In certain embodiments the method further comprises: e. analyzing the set of consensus sequences for each set of tagged parent molecules.

[0039] In some embodiments each polynucleotide in a set is mappable to a reference sequence.

[0040] In some embodiments the method comprises providing a plurality of sets of tagged parent polynucleotides, wherein each set is mappable to a different reference sequence.

[0041] In some embodiments the method further comprises converting initial starting genetic material into the tagged parent polynucleotides.

[0042] In some embodiments the initial starting genetic material comprises no more than 100 ng of polynucleotides.

IDT Exhibit 1020.4012

[0043] In some embodiments the method comprises bottlenecking the initial starting genetic material prior to converting.

[0044] In some embodiments the method comprises converting the initial starting genetic material into tagged parent polynucleotides with a conversion efficiency of at least 10%, at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 80% or at least 90%.

[0045] In some embodiments converting comprises any of blunt-end ligation, sticky end ligation, molecular inversion probes, PCR, ligation-based PCR, single strand ligation and single strand circularization.

[0046] In some embodiments the initial starting genetic material is cell-free nucleic acid.

[0047] In some embodiments a plurality of the reference sequences are from the same genome.

[0048] In some embodiments each tagged parent polynucleotide in the set is uniquely tagged.

[0049] In some embodiments the tags are non-unique.

[0050] In some embodiments the generation of consensus sequences is based on information from the tag and/or at least one of sequence information at the beginning (start) region of the sequence read, the end (stop) regions of the sequence read and the length of the sequence read.

[0051] In some embodiments the method comprises sequencing a subset of the set of amplified progeny polynucleotides sufficient to produce sequence reads for at least one progeny from of each of at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90% at least 95%, at least 98%, at least 99%, at least 99.9% or at least 99.99% of unique polynucleotides in the set of tagged parent polynucleotides.

[0052] In some embodiments the at least one progeny is a plurality of progeny, e.g., at least 2, at least 5 or at least 10 progeny.

[0053] In some embodiments the number of sequence reads in the set of sequence reads is greater than the number of unique tagged parent polynucleotides in the set of tagged parent polynucleotides.

[0054] In some embodiments the subset of the set of amplified progeny polynucleotides sequenced is of sufficient size so that any nucleotide sequence represented in the set of tagged parent polynucleotides at a percentage that is the same as the percentage per-base sequencing error rate of the sequencing platform used, has at least a 50%, at least a 60%, at least a 70%, at least a 80%, at least a 90% at least a 95%, at least a 98%, at least a 99%, at least a 99.9% or at least a 99.99% chance of being represented among the set of consensus sequences.

IDT Exhibit 1020.4013

[0055] In some embodiments the method comprises enriching the set of amplified progeny polynucleotides for polynucleotides mapping to one or more selected reference sequences by: (i) selective amplification of sequences from initial starting genetic material converted to tagged parent polynucleotides; (ii) selective amplification of tagged parent polynucleotides; (iii) selective sequence capture of amplified progeny polynucleotides; or (iv) selective sequence capture of initial starting genetic material.

[0056] In some embodiments analyzing comprises normalizing a measure (e.g., number) taken from a set of consensus sequences against a measure taken from a set of consensus sequences from a control sample.

[0057] In some embodiments analyzing comprises detecting mutations, rare mutations, single nucleotide variants, indels, copy number variations, transversions, translocations, inversion, deletions, aneuploidy, partial aneuploidy, polyploidy, chromosomal instability, chromosomal structure alterations, gene fusions, chromosome fusions, gene truncations, gene amplification, gene duplications, chromosomal lesions, DNA lesions, abnormal changes in nucleic acid chemical modifications, abnormal changes in epigenetic patterns, abnormal changes in nucleic acid methylation infection or cancer.

[0058] In some embodiments the polynucleotides comprise DNA, RNA, a combination of the two or DNA plus RNA-derived cDNA.

[0059] In some embodiments a certain subset of polynucleotides is selected for or is enriched based on polynucleotide length in base-pairs from the initial set of polynucleotides or from the amplified polynucleotides.

[0060] In some embodiments analysis further comprises detection and monitoring of an abnormality or disease within an individual, such as, infection and/or cancer.

[0061] In some embodiments the method is performed in combination with immune repertoire profiling.

[0062] In some embodiments the polynucleotides are extract from the group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, and tears.

[0063] In some embodiments collapsing comprising detecting and/or correcting errors, nicks or lesions present in the sense or anti-sense strand of the tagged parent polynucleotides or amplified progeny polynucleotides.

[0064] This disclosure also provides for a method comprising detecting genetic variation in initial starting genetic material with a sensitivity of at least 5%, at least 1%, at least 0.5%, at least 0.1%

IDT Exhibit 1020.4014

or at least 0.05%. In some embodiments the initial starting genetic material is provided in an amount less than 100 ng of nucleic acid, the genetic variation is copy number/heterozygosity variation and detecting is performed with sub-chromosomal resolution; e.g., at least 100 megabase resolution, at least 10 megabase resolution, at least 1 megabase resolution, at least 100 kilobase resolution, at least 10 kilobase resolution or at least 1 kilobase resolution. In another embodiment the method comprises providing a plurality of sets of tagged parent polynucleotides, wherein each set is mappable to a different reference sequence. In another embodiment the reference sequence is the locus of a tumor marker, and analyzing comprises detecting the tumor marker in the set of consensus sequences. In another embodiment the tumor marker is present in the set of consensus sequences at a frequency less than the error rate introduced at the amplifying step. In another embodiment the at least one set is a plurality of sets, and the reference sequences comprise a plurality of reference sequences, each of which is the locus of a tumor marker. In another embodiment analyzing comprises detecting copy number variation of consensus sequences between at least two sets of parent polynucleotides. In another embodiment analyzing comprises detecting the presence of sequence variations compared with the reference sequences. In another embodiment analyzing comprises detecting the presence of sequence variations compared with the reference sequences and detecting copy number variation of consensus sequences between at least two sets of parent polynucleotides. In another embodiment collapsing comprises: i. grouping sequences reads sequenced from amplified progeny polynucleotides into families, each family amplified from the same tagged parent polynucleotide; and ii. determining a consensus sequence based on sequence reads in a family.

[0065] This disclosure also provides for a system comprising a computer readable medium for performing the following steps: a. providing at least one set of tagged parent polynucleotides, and for each set of tagged parent polynucleotides; b. amplifying the tagged parent polynucleotides in the set to produce a corresponding set of amplified progeny polynucleotides; c. sequencing a subset (including a proper subset) of the set of amplified progeny polynucleotides, to produce a set of sequencing reads; and d. collapsing the set of sequencing reads to generate a set of consensus sequences, each consensus sequence corresponding to a unique polynucleotide among the set of tagged parent polynucleotides and, optionally, e. analyzing the set of consensus sequences for each set of tagged parent molecules.

[0066] This disclosure also provides a method comprising: a. providing at least one set of tagged parent polynucleotides, and for each set of tagged parent polynucleotides; b. amplifying the

IDT Exhibit 1020.4015

tagged parent polynucleotides in the set to produce a corresponding set of amplified progeny polynucleotides; c. sequencing a subset (including a proper subset) of the set of amplified progeny polynucleotides, to produce a set of sequencing reads; d. collapsing the set of sequencing reads to generate a set of consensus sequences, each consensus sequence corresponding to a unique polynucleotide among the set of tagged parent polynucleotides; and e. filtering out from among the consensus sequences those that fail to meet a quality threshold. In one embodiment the quality threshold considers a number of sequence reads from amplified progeny polynucleotides collapsed into a consensus sequence. In another embodiment the quality threshold considers a number of sequence reads from amplified progeny polynucleotides collapsed into a consensus sequence. This disclosure also provides a system comprising a computer readable medium for performing the aforesaid method.

[0067] This disclosure also provides a method comprising: a. providing at least one set of tagged parent polynucleotides, wherein each set maps to a different reference sequence in one or more genomes, and, for each set of tagged parent polynucleotides; i. amplifying the first polynucleotides to produce a set of amplified polynucleotides; ii. sequencing a subset of the set of amplified polynucleotides, to produce a set of sequencing reads; and iii. collapsing the sequence reads by: 1. grouping sequences reads sequenced from amplified progeny polynucleotides into families, each family amplified from the same tagged parent polynucleotide. In one embodiment collapsing further comprises: 2. determining a quantitative measure of sequence reads in each family. In another embodiment the method further comprises (including a) including a): b. determining a quantitative measure of unique families; and c. based on (1) the quantitative measure of unique families and (2) the quantitative measure of sequence reads in each group, inferring a measure of unique tagged parent polynucleotides in the set. In another embodiment inferring is performed using statistical or probabilistic models. In another embodiment wherein the at least one set is a plurality of sets. In another embodiment the method further comprises correcting for amplification or representational bias between the two sets. In another embodiment the method further comprises using a control or set of control samples to correct for amplification or representational biases between the two sets. In another embodiment the method further comprises determining copy number variation between the sets. In another embodiment the method further comprises (including a, b, c): d. determining a quantitative measure of polymorphic forms among the families; and e. based on the determined quantitative measure of polymorphic forms, inferring a quantitative measure of polymorphic forms in the number of

IDT Exhibit 1020.4016

inferred unique tagged parent polynucleotides. In another embodiment wherein polymorphic forms include but are not limited to: substitutions, insertions, deletions, inversions, microsatellite changes, transversions, translocations, fusions, methylation, hypermethylation, hyrdroxymethylation, acetylation, epigenetic variants, regulatory-associated variants or protein binding sites. In another embodiment wherein the sets derive from a common sample, the method further comprising: a. inferring copy number variation for the plurality of sets based on a comparison of the inferred number of tagged parent polynucleotides in each set mapping to each of a plurality of reference sequences. In another embodiment the original number of polynucleotides in each set is further inferred. This disclosure also provides a system comprising a computer readable medium for performing the aforesaid methods.

[0068] This disclosure also provides a method of determining copy number variation in a sample that includes polynucleotides, the method comprising: a. providing at least two sets of first polynucleotides, wherein each set maps to a different reference sequence in a genome, and, for each set of first polynucleotides; i. amplifying the polynucleotides to produce a set of amplified polynucleotides; ii. sequencing a subset of the set of amplified polynucleotides, to produce a set of sequencing reads; iii. grouping sequences reads sequenced from amplified polynucleotides into families, each family amplified from the same first polynucleotide in the set; iv. inferring a quantitative measure of families in the set; b. determining copy number variation by comparing the quantitative measure of families in each set. This disclosure also provides a system comprising a computer readable medium for performing the aforesaid methods.

[0069] This disclosure also provides a method of inferring frequency of sequence calls in a sample of polynucleotides comprising: a. providing at least one set of first polynucleotides, wherein each set maps to a different reference sequence in one or more genomes, and, for each set of first polynucleotides; i. amplifying the first polynucleotides to produce a set of amplified polynucleotides; ii. sequencing a subset of the set of amplified polynucleotides, to produce a set of sequencing reads; iii. grouping the sequence reads into families, each family comprising sequence reads of amplified polynucleotides amplified from the same first polynucleotide; b. inferring, for each set of first polynucleotides, a call frequency for one or more bases in the set of first polynucleotides, wherein inferring comprises: i. assigning, for each family, confidence score for each of a plurality of calls, the confidence score taking into consideration a frequency of the call among members of the family; and ii. estimating a frequency of the one or more calls taking into consideration the confidence scores of the one or more calls assigned to each family. This

IDT Exhibit 1020.4017

disclosure also provides a system comprising a computer readable medium for performing the aforesaid methods.

[0070] This disclosure also provides a method of communicating sequence information about at least one individual polynucleotide molecule comprising: a. providing at least one individual polynucleotide molecule; b. encoding sequence information in the at least one individual polynucleotide molecule to produce a signal; c. passing at least part of the signal through a channel to produce a received signal comprising nucleotide sequence information about the at least one individual polynucleotide molecule, wherein the received signal comprises noise and/or distortion; d. decoding the received signal to produce a message comprising sequence information about the at least one individual polynucleotide molecule, wherein decoding reduces noise and/or distortion in the message; and e. providing the message to a recipient. In one embodiment the noise comprises incorrect nucleotide calls. In another embodiment distortion comprises uneven amplification of the individual polynucleotide molecule compared with other individual polynucleotide molecules. In another embodiment distortion results from amplification or sequencing bias. In another embodiment the at least one individual polynucleotide molecule is a plurality of individual polynucleotide molecules, and decoding produces a message about each molecule in the plurality. In another embodiment encoding comprises amplifying the at least individual polynucleotide molecule which has optionally been tagged, wherein the signal comprises a collection of amplified molecules. In another embodiment the channel comprises a polynucleotide sequencer and the received signal comprises sequence reads of a plurality of polynucleotides amplified from the at least one individual polynucleotide molecule. In another embodiment decoding comprises grouping sequence reads of amplified molecules amplified from each of the at least one individual polynucleotide molecules. In another embodiment the decoding consists of a probabilistic or statistical method of filtering the generated sequence signal. This disclosure also provides a system comprising a computer readable medium for performing the aforesaid methods.

[0071] In another embodiment the polynucleotides are derived from tumor genomic DNA or RNA. In another embodiment the polynucleotides are derived from cell-free polynucleotides, exosomal polynucleotides, bacterial polynucleotides or viral polynucleotides. In another embodiment further comprising the detection and/or association of affected molecular pathways. In another embodiment further comprising serial monitoring of the health or disease state of an individual. In another embodiment whereby the phylogeny of a genome associated with a disease

IDT Exhibit 1020.4018

within an individual is inferred. In another embodiment further comprising diagnosis, monitoring or treatment of a disease. In another embodiment the treatment regimen is selected or modified based on detected polymorphic forms or CNVs or associated pathways. In another embodiment the treatment comprises of a combination therapy.

[0072] This disclosure also provides a computer readable medium in non-transitory, tangible form comprising executable code configured to perform the following steps: selecting predefined regions in a genome; accessing sequence reads and enumerating number of sequence reads in the predefined regions; normalizing the number of sequence reads across the predefined regions; and determining percent of copy number variation in the predefined regions.

[0073] This disclosure also provides a computer readable medium in non-transitory, tangible form comprising executable code configured to perform the following steps: a. accessing a data file comprising a plurality of sequencing reads; b. filtering out reads that fail to meet a set threshold; c. mapping sequence reads derived from the sequencing onto a reference sequence; d. identifying a subset of mapped sequence reads that align with a variant of the reference sequence at each mappable base position; e. for each mappable base position, calculating a ratio of (a) a number of mapped sequence reads that include a variant as compared to the reference sequence, to (b) a number of total sequence reads for each mappable base position; f. normalizing the ratios or frequency of variance for each mappable base position and determining potential rare variant(s) or other genetic alteration(s); and g. comparing the resulting number for each of the regions with potential rare variant(s) or mutation(s) to similarly derived numbers from a reference sample.

[0074] This disclosure also provides a computer readable medium in non-transitory, tangible form comprising executable code configured to perform the following steps: a. accessing a data file comprising a plurality of sequencing reads, wherein the sequence reads derive from a set of progeny polynucleotides amplified from at least one set of tagged parent polynucleotides; b. collapsing the set of sequencing reads to generate a set of consensus sequences, each consensus sequence corresponding to a unique polynucleotide among the set of tagged parent polynucleotides.

[0075] This disclosure also provides a computer readable medium in non-transitory, tangible form comprising executable code configured to perform the following steps: a. accessing a data file comprising a plurality of sequencing reads, wherein the sequence reads derive from a set of progeny polynucleotides amplified from at least one set of tagged parent polynucleotides; b. collapsing the set of sequencing reads to generate a set of consensus sequences, each consensus

IDT Exhibit 1020.4019

sequence corresponding to a unique polynucleotide among the set of tagged parent polynucleotides; c. filtering out from among the consensus sequences those that fail to meet a quality threshold.

[0076] This disclosure also provides a computer readable medium in non-transitory, tangible form comprising executable code configured to perform the following steps: a. accessing a data file comprising a plurality of sequencing reads, wherein the sequence reads derive from a set of progeny polynucleotides amplified from at least one set of tagged parent polynucleotides; and i. collapsing the sequence reads by: 1. grouping sequences reads sequenced from amplified progeny polynucleotides into families, each family amplified from the same tagged parent polynucleotide and, optionally, 2. determining a quantitative measure of sequence reads in each family. In certain embodiments, the executable code further performs the steps of: b. determining a quantitative measure of unique families; c. based on (1) the quantitative measure of unique families and (2) the quantitative measure of sequence reads in each group, inferring a measure of unique tagged parent polynucleotides in the set. In certain embodiments, the executable code further performs the steps of: d. determining a quantitative measure of polymorphic forms among the families; and e. based on the determined quantitative measure of polymorphic forms, inferring a quantitative measure of polymorphic forms in the number of inferred unique tagged parent polynucleotides.

[0077] This disclosure also provides a computer readable medium in non-transitory, tangible form comprising executable code configured to perform the following steps: a. accessing a data file comprising a plurality of sequencing reads, wherein the sequence reads derive from a set of progeny polynucleotides amplified from at least one set of tagged parent polynucleotides grouping sequences reads sequenced from amplified polynucleotides into families, each family amplified from the same first polynucleotide in the set; b. inferring a quantitative measure of families in the set; c. determining copy number variation by comparing the quantitative measure of families in each set.

[0078] This disclosure also provides a computer readable medium in non-transitory, tangible form comprising executable code configured to perform the following steps: a. accessing a data file comprising a plurality of sequencing reads, wherein the sequence reads derive from a set of progeny polynucleotides amplified from at least one set of tagged parent polynucleotides grouping the sequence reads into families, each family comprising sequence reads of amplified polynucleotides amplified from the same first polynucleotide; b. inferring, for each set of first

IDT Exhibit 1020.4020

polynucleotides, a call frequency for one or more bases in the set of first polynucleotides, wherein inferring comprises: c. assigning, for each family, confidence score for each of a plurality of calls, the confidence score taking into consideration a frequency of the call among members of the family; and d. estimating a frequency of the one or more calls taking into consideration the confidence scores of the one or more calls assigned to each family.

[0079] This disclosure also provides a computer readable medium in non-transitory, tangible form comprising executable code configured to perform the following steps: a. accessing a data accessing a data file comprising a received signal that comprises endoded sequence information from at least one individual polynucleotide molecule wherein the received signal comprises noise and/or distortion; b. decoding the received signal to produce a message comprising sequence information about the at least one individual polynucleotide molecule, wherein decoding reduces noise and/or distortion about each individual polynucleotide in the message; and c. writing the message comprising sequence information about the at least one individual polynucleotide molecule to a computer file.

[0080] This disclosure also provides a computer readable medium in non-transitory, tangible form comprising executable code configured to perform the following steps: a. accessing a data file comprising a plurality of sequencing reads, wherein the sequence reads derive from a set of progeny polynucleotides amplified from at least one set of tagged parent polynucleotides; b. collapsing the set of sequencing reads to generate a set of consensus sequences, each consensus sequence corresponding to a unique polynucleotide among the set of tagged parent polynucleotides; c. filtering out from among the consensus sequences those that fail to meet a quality threshold.

[0081] This disclosure also provides a computer readable medium in non-transitory, tangible form comprising executable code configured to perform the following steps: a. accessing a data file comprising a plurality of sequencing reads, wherein the sequence reads derive from a set of progeny polynucleotides amplified from at least one set of tagged parent polynucleotides; and b. collapsing the sequence reads by: i. grouping sequences reads sequenced from amplified progeny polynucleotides into families, each family amplified from the same tagged parent polynucleotide; and ii. optionally, determining a quantitative measure of sequence reads in each family. In certain embodiments, the executable code further performs the steps of: c. determining a quantitative measure of unique families; d. based on (1) the quantitative measure of unique families and (2) the quantitative measure of sequence reads in each group, inferring a measure of unique tagged

IDT Exhibit 1020.4021

parent polynucleotides in the set. In certain embodiments, the executable code further performs the steps of: e. determining a quantitative measure of polymorphic forms among the families; and f. based on the determined quantitative measure of polymorphic forms, inferring a quantitative measure of polymorphic forms in the number of inferred unique tagged parent polynucleotides. In certain embodiments, the executable code further performs the steps of: e. inferring copy number variation for the plurality of sets based on a comparison of the inferred number of tagged parent polynucleotides in each set mapping to each of a plurality of reference sequences.

[0082] This disclosure also provides a computer readable medium in non-transitory, tangible form comprising executable code configured to perform the following steps: a. accessing a data file comprising a plurality of sequencing reads, wherein the sequence reads derive from a set of progeny polynucleotides amplified from at least one set of tagged parent polynucleotides; b. grouping sequences reads sequenced from amplified polynucleotides into families, each family amplified from the same first polynucleotide in the set; c. inferring a quantitative measure of families in the set; d. determining copy number variation by comparing the quantitative measure of families in each set.

[0083] This disclosure also provides a computer readable medium in non-transitory, tangible form comprising executable code configured to perform the following steps: a. accessing a data file comprising a plurality of sequencing reads, wherein the sequence reads derive from a set of progeny polynucleotides amplified from at least one set of tagged parent polynucleotides grouping the sequence reads into families, each family comprising sequence reads of amplified polynucleotides amplified from the same first polynucleotide; and b. inferring, for each set of first polynucleotides, a call frequency for one or more bases in the set of first polynucleotides, wherein inferring comprises: i. assigning, for each family, confidence score for each of a plurality of calls, the confidence score taking into consideration a frequency of the call among members of the family; and ii. estimating a frequency of the one or more calls taking into consideration the confidence scores of the one or more calls assigned to each family.

[0084] This disclosure also provides a method comprising: a. providing a sample comprising between 100 and 100,000 haploid human genome equivalents of cell free DNA (cfDNA) polynucleotides; and b. tagging the polynucleotides with between 2 and 1,000,000 unique identifiers. In certain embodiments, the number of unique identifiers is at least 3, at least 5, at least 10, at least 15 or at least 25 and at most 100, at most 1000 or at most 10,000. In certain

IDT Exhibit 1020.4022

embodiments, the number of unique identifiers is at most 100, at most 1000, at most 10,000, at most 100,000.

[0085] This disclosure also provides a method comprising: a. providing a sample comprising a plurality of human haploid genome equivalents of fragmented polynucleotides; b. determining z, wherein z is a measure of central tendency (e.g., mean, median or mode) of expected number of duplicate polynucleotides starting at any position in the genome, wherein duplicate polynucleotides have the same start and stop positions; and c. tagging polynucleotides in sample with n unique identifiers, wherein n is between 2 and 100,000*z, 2 and 10,000*z, 2 and 1,000*z or 2 and 100*z.

[0086] This disclosure also provides a method comprising: a. providing at least one set of tagged parent polynucleotides, and for each set of tagged parent polynucleotides; b. producing a plurality of sequence reads for each tagged parent polynucleotide in the set to produce a set of sequencing reads; and c. collapsing the set of sequencing reads to generate a set of consensus sequences, each consensus sequence corresponding to a unique polynucleotide among the set of tagged parent polynucleotides.

[0087] The disclosure provides for a method for detecting copy number variation comprising: a) sequencing extracellular polynucleotides from a bodily sample from a subject, wherein each of the extracellular polynucleotide generate a plurality of sequencing reads; b) filtering out reads that fail to meet a set threshold; c) mapping the sequence reads obtained from step (a), after reads are filtered out, to a reference sequence; d) quantifying or enumerating mapped reads in two or more predefined regions of the reference sequence; and e) determining copy number variation in one or more of the predefined regions by: (ii) normalizing number of reads in the predefined regions to each other and/or the number of unique sequence reads in the predefined regions to one another; (ii) comparing the normalized numbers obtained in step (i) to normalized numbers obtained from a control sample.

[0088] The disclosure also provides for a method for detecting a rare mutation in a cell-free or substantially cell free sample obtained from a subject comprising: a) sequencing extracellular polynucleotides from a bodily sample from a subject, wherein each of the extracellular polynucleotide generate a plurality of sequencing reads; b) performing multiplex sequencing on regions or whole-genome sequencing if enrichment is not performed; c) filtering out reads that fail to meet a set threshold; d) mapping sequence reads derived from the sequencing onto a reference sequence; e) identifying a subset of mapped sequence reads that align with a variant of

IDT Exhibit 1020.4023

the reference sequence at each mappable base position; f) for each mappable base position, calculating a ratio of (a) a number of mapped sequence reads that include a variant as compared to the reference sequence, to (b) a number of total sequence reads for each mappable base position; g) normalizing the ratios or frequency of variance for each mappable base position and determining potential rare variant(s) or mutation(s); and h)and comparing the resulting number for each of the regions with potential rare variant(s) or mutation(s) to similarly derived numbers from a reference sample.

[0089] The disclosure also provides for a method of characterizing the heterogeneity of an abnormal condition in a subject, the method comprising generating a genetic profile of extracellular polynucleotides in the subject, wherein the genetic profile comprises a plurality of data resulting from copy number variation and rare mutation analyses.

[0090] In some embodiments, the prevalence/concentration of each rare variant identified in the subject is reported and quantified simultaneously. In some embodiments, a confidence score, regarding the prevalence/concentrations of rare variants in the subject, is reported.

[0091] In some embodiments, the extracellular polynucleotides comprise DNA. In some embodiments, the extracellular polynucleotides comprise RNA.

[0092] In some embodiments, the methods further comprise isolating extracellular polynucleotides from the bodily sample. In some embodiments, the isolating comprises a method for circulating nucleic acid isolation and extraction. In some embodiments, the methods further comprise fragmenting said isolated extracellular polynucleotides. In some embodiments, the bodily sample is selected from the group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool and tears.

[0093] In some embodiments, the methods further comprise the step of determining the percent of sequences having copy number variation or rare mutation or variant in said bodily sample. In some embodiments, the determining comprises calculating the percentage of predefined regions with an amount of polynucleotides above or below a predetermined threshold.

[0094] In some embodiments, the subject is suspected of having an abnormal condition. In some embodiments, the abnormal condition is selected from the group consisting of mutations, rare mutations, indels, copy number variations, transversions, translocations, inversion, deletions, aneuploidy, partial aneuploidy, polyploidy, chromosomal instability, chromosomal structure alterations, gene fusions, chromosome fusions, gene truncations, gene amplification, gene duplications, chromosomal lesions, DNA lesions, abnormal changes in nucleic acid chemical

IDT Exhibit 1020.4024

modifications, abnormal changes in epigenetic patterns, abnormal changes in nucleic acid methylation infection and cancer.

[0095] In some embodiments, the subject is a pregnant female. In some embodiments, the copy number variation or rare mutation or genetic variant is indicative of a fetal abnormality. In some embodiments, the fetal abnormality is selected from the group consisting of mutations, rare mutations, indels, copy number variations, transversions, translocations, inversion, deletions, aneuploidy, partial aneuploidy, polyploidy, chromosomal instability, chromosomal structure alterations, gene fusions, chromosome fusions, gene truncations, gene amplification, gene duplications, chromosomal lesions, DNA lesions, abnormal changes in nucleic acid chemical modifications, abnormal changes in epigenetic patterns, abnormal changes in nucleic acid methylation infection and cancer.

[0096] In some embodiments, the methods further comprise attaching one or more barcodes to the extracellular polynucleotides or fragments thereof prior to sequencing. In some embodiments, each barcode attached to extracellular polynucleotides or fragments thereof prior to sequencing is unique. In some embodiments, each barcode attached to extracellular polynucleotides or fragments thereof prior to sequencing is not unique.

[0097] In some embodiments, the methods further comprise selectively enriching regions from the subject's genome or transcriptome prior to sequencing. In some embodiments, the methods further comprise non-selectively enriching regions from the subject's genome or transcriptome prior to sequencing.

[0098] In some embodiments, the methods further comprise attaching one or more barcodes to the extracellular polynucleotides or fragments thereof prior to any amplification or enrichment step. In some embodiments, the barcode is a polynucleotide. In some embodiments, the barcode comprises random sequence. In some embodiments, the barcode comprises a fixed or semi-random set of oligonucleotides that in combination with the diversity of molecules sequenced from a select region enables identification of unique molecules. In some embodiments, the barcodes comprise oligonucleotides is at least a 3, 5, 10, 15, 20 25, 30, 35, 40, 45, or 50mer base pairs in length.

[0099] In some embodiments, the methods further comprise amplifying the extracellular polynucleotides or fragments thereof. In some embodiments, the amplification comprises global amplification or whole genome amplification. In some embodiments, the amplification comprises selective amplification. In some embodiments, the amplification comprises non-selective

IDT Exhibit 1020.4025

amplification. In some embodiments, suppression amplification or subtractive enrichment is performed.

[00100] In some embodiments, sequence reads of unique identity are detected based on sequence information at the beginning (start) and end (stop) regions of the sequence read and the length of the sequence read. In some embodiments, sequence molecules of unique identity are detected based on sequence information at the beginning (start) and end (stop) regions of the sequence read, the length of the sequence read and attachment of a barcode.

[00101] In some embodiments, the methods further comprise removing a subset of the reads from further analysis prior to quantifying or enumerating reads. In some embodiments, removing comprises filtering out reads with an accuracy or quality score of less than a threshold, e.g., 90%, 99%, 99.9%, or 99.99% and/or mapping score less than a threshold, e.g., 90%, 99%, 99.9% or 99.99%. In some embodiments, the methods further comprise filtering reads with a quality score lower than a set threshold.

[00102] In some embodiments, the predefined regions are uniform or substantially uniform in size. In some embodiments, the predefined regions are at least about 10kb, 20kb, 30kb 40kb, 50kb, 60kb, 70kb, 80kb, 90kb, or 100kb in size.

[00103] In some embodiments, at least 50, 100, 200, 500, 1000, 2000, 5000, 10,000, 20,000, or 50,000 regions are analyzed.

[00104] In some embodiments, the variant occurs in a region of the genome selected from the group consisting of gene fusions, gene duplications, gene deletions, gene translocations, microsatellite regions, gene fragments or combination thereof. In some embodiments, the variant occurs in a region of the genome selected from the group consisting of genes, oncogenes, tumor suppressor genes, promoters, regulatory sequence elements, or combination thereof. In some embodiments, the variant is a nucleotide variant, single base substitution, small indel, transversion, translocation, inversion, deletion, truncation or gene truncation of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 or 20 nucleotides in length.

[00105] In some embodiments, the methods further comprise correcting/normalizing/adjusting the quantity of mapped reads using the barcodes or unique properties of individual reads. In some embodiments, enumerating the reads is performed through enumeration of unique barcodes in each of the predefined regions and normalizing those numbers across at least a subset of predefined regions that were sequenced.

IDT Exhibit 1020.4026

[00106] In some embodiments, samples at succeeding time intervals from the same subject are analyzed and compared to previous sample results. In some embodiments, the method further comprises amplifying the barcode-attached extracellular polynucleotides. In some embodiments, the methods further comprise determining partial copy number variation frequency, determining loss of heterozygosity, performing gene expression analysis, performing epigenetic analysis and/or performing hypermethylation analysis.

[00107] The disclosure also provides for a method comprising determining copy number variation or performing rare mutation analysis in a cell-free or substantially cell free sample obtained from a subject using multiplex sequencing.

[00108] In some embodiments, the multiplex sequencing comprises performing over 10,000 sequencing reactions. In some embodiments, the multiplex sequencing comprises simultaneously sequencing at least 10,000 different reads. In some embodiments, the multiplex sequencing comprising performing data analysis on at least 10,000 different reads across the genome. In some embodiments, the normalizing and detection is performed using one or more of hidden markov, dynamic programming, support vector machine, Bayesian or probabilistic modeling, trellis decoding, Viterbi decoding, expectation maximization, Kalman filtering, or neural network methodologies. In some embodiments, the methods further comprise monitoring disease progression, monitoring residual disease, monitoring therapy, diagnosing a condition, prognosing a condition, or selecting a therapy based on discovered variants for the subject. In some embodiments, a therapy is modified based on the most recent sample analysis. In some embodiments, the genetic profile of a tumor, infection or other tissue abnormality is inferred.

[00109] In some embodiments, the growth, remission or evolution of a tumor, infection or other tissue abnormality is monitored. In some embodiments, sequences related to the subject's immune system are analyzed and monitored at single instances or over time. In some embodiments, identification of a variant is followed up through an imaging test (e.g., CT, PET-CT, MRI, X-ray, ultrasound) for localization of the tissue abnormality suspected of causing the identified variant. In some embodiments, the analysis further comprises use of genetic data obtained from a tissue or tumor biopsy from the same patient. In some embodiments, the phylogenetics of a tumor, infection or other tissue abnormality is inferred. In some embodiments, the method further comprises performing population-based no-calling and identification of low-confidence regions. In some embodiments, obtaining the measurement data for the sequence coverage comprises measuring sequence coverage depth at every position of the genome. In some

IDT Exhibit 1020.4027

embodiments, correcting the measurement data for the sequence coverage bias comprises calculating window-averaged coverage. In some embodiments, correcting the measurement data for the sequence coverage bias comprises performing adjustments to account for GC bias in the library construction and sequencing process. In some embodiments, correcting the measurement data for the sequence coverage bias comprises performing adjustments based on additional weighting factor associated with individual mappings to compensate for bias.

[00110] In some embodiments, extracellular polynucleotide is derived from a diseased cell origin. In some embodiments, extracellular polynucleotide is derived from a healthy cell origin.

[00111] The disclosure also provides for a system comprising a computer readable medium for performing the following steps: selecting predefined regions in a genome; enumerating number of sequence reads in the predefined regions; normalizing the number of sequence reads across the predefined regions; and determining percent of copy number variation in the predefined regions.

[00112] In some embodiments, the entirety of the genome or at least 85% of the genome is analyzed. In some embodiments, the computer readable medium provides data on percent cancer DNA or RNA in plasma or serum to the end user. In some embodiments, the copy number variants identified are fractional (i.e., non-integer levels) due to heterogeneity in the sample. In some embodiments, enrichment of selected regions is performed. In some embodiments, copy number variation information is simultaneously extracted based on the methods described herein. In some embodiments, the methods comprise an initial step of polynucleotide bottlenecking to limit the number of starting initial copies or diversity of polynucleotides in the sample.

[00113] The disclosure also provides for a method for detecting a rare mutation in a cell-free or a substantially cell free sample obtained from a subject comprising: a) sequencing extracellular polynucleotides from a bodily sample of a subject, wherein each of the extracellular polynucleotides generate a plurality of sequencing reads; b) filtering out reads that fail to meet a set quality threshold; c) mapping sequence reads derived from the sequencing onto a reference sequence; d) identifying a subset of mapped sequence reads that align with a variant of the reference sequence at each mappable base position; e) for each mappable base position, calculating a ratio of (a) a number of mapped sequence reads that include a variant as compared to the reference sequence, to (b) a number of total sequence reads for each mappable base position; f) normalizing the ratios or frequency of variance for each mappable base position and determining potential rare variant(s) or other genetic alteration(s); and g) comparing the resulting

IDT Exhibit 1020.4028

number for each of the regions with potential rare variant(s) or mutation(s) to similarly derived numbers from a reference sample.

[00114] The disclosure also provides for a method comprising: a) providing at least one set of tagged parent polynucleotides, and for each set of tagged parent polynucleotides; b) amplifying the tagged parent polynucleotides in the set to produce a corresponding set of amplified progeny polynucleotides; c) sequencing a subset (including a proper subset) of the set of amplified progeny polynucleotides, to produce a set of sequencing reads; and d) collapsing the set of sequencing reads to generate a set of consensus sequences, each consensus sequence corresponding to a unique polynucleotide among the set of tagged parent polynucleotides.

[00115] In some embodiments, each polynucleotide in a set is mappable to a reference sequence. In some embodiments, the methods comprise providing a plurality of sets of tagged parent polynucleotides, wherein each set is mappable to a different mappable position in the reference sequence. In some embodiments, the method further comprises: e) analyzing the set of consensus sequences for each set of tagged parent molecules separately or in combination. In some embodiments, the method further comprises converting initial starting genetic material into the tagged parent polynucleotides. In some embodiments, the initial starting genetic material comprises no more than 100 ng of polynucleotides. In some embodiments, the method comprises bottlenecking the initial starting genetic material prior to converting. In some embodiments, the method comprises converting the initial starting genetic material into tagged parent polynucleotides with a conversion efficiency of at least 10%, at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 80% or at least 90%. In some embodiments, the converting comprises any of blunt-end ligation, sticky end ligation, molecular inversion probes, PCR, ligation-based PCR, single strand ligation and single strand circularization. In some embodiments, the initial starting genetic material is cell-free nucleic acid. In some embodiments, a plurality of the sets map to different mappable positions in a reference sequence from the same genome.

[00116] In some embodiments, each tagged parent polynucleotide in the set is uniquely tagged. In some embodiments, each set of parent polynucleotides is mappable to a position in a reference sequence, and the polynucleotides in each set are not uniquely tagged. In some embodiments, the generation of consensus sequences is based on information from the tag and/or at least one of (i) sequence information at the beginning (start) region of the sequence read, (ii) the end (stop) regions of the sequence read and (iii) the length of the sequence read.

IDT Exhibit 1020.4029

[00117] In some embodiments, the method comprises sequencing a subset of the set of amplified progeny polynucleotides sufficient to produce sequence reads for at least one progeny from of each of at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90% at least 95%, at least 98%, at least 99%, at least 99.9% or at least 99.99% of unique polynucleotides in the set of tagged parent polynucleotides. In some embodiments, the at least one progeny is a plurality of progeny, e.g., at least 2, at least 5 or at least 10 progeny. In some embodiments, the number of sequence reads in the set of sequence reads is greater than the number of unique tagged parent polynucleotides in the set of tagged parent polynucleotides. In some embodiments, the subset of the set of amplified progeny polynucleotides sequenced is of sufficient size so that any nucleotide sequence represented in the set of tagged parent polynucleotides at a percentage that is the same as the percentage per-base sequencing error rate of the sequencing platform used, has at least a 50%, at least a 60%, at least a 70%, at least a 80%, at least a 90% at least a 95%, at least a 98%, at least a 99%, at least a 99.9% or at least a 99.99% chance of being represented among the set of consensus sequences.

[00118] In some embodiments, the method comprises enriching the set of amplified progeny polynucleotides for polynucleotides mapping to one or more selected mappable positions in a reference sequence by: (i) selective amplification of sequences from initial starting genetic material converted to tagged parent polynucleotides; (ii) selective amplification of tagged parent polynucleotides; (iii) selective sequence capture of amplified progeny polynucleotides; or (iv) selective sequence capture of initial starting genetic material.

[00119] In some embodiments, analyzing comprises normalizing a measure (e.g., number) taken from a set of consensus sequences against a measure taken from a set of consensus sequences from a control sample. In some embodiments, analyzing comprises detecting mutations, rare mutations, indels, copy number variations, transversions, translocations, inversion, deletions, aneuploidy, partial aneuploidy, polyploidy, chromosomal instability, chromosomal structure alterations, gene fusions, chromosome fusions, gene truncations, gene amplification, gene duplications, chromosomal lesions, DNA lesions, abnormal changes in nucleic acid chemical modifications, abnormal changes in epigenetic patterns, abnormal changes in nucleic acid methylation infection or cancer.

[00120] In some embodiments, the polynucleotides comprise DNA, RNA, a combination of the two, or DNA plus RNA-derived cDNA. In some embodiments, a certain subset of polynucleotides is selected for, or is enriched based on, polynucleotide length in base-pairs from

IDT Exhibit 1020.4030

the initial set of polynucleotides or from the amplified polynucleotides. In some embodiments, analysis further comprises detection and monitoring of an abnormality or disease within an individual, such as, infection and/or cancer. In some embodiments, the method is performed in combination with immune repertoire profiling. In some embodiments, the polynucleotides are extracted from a sample selected from the group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, and tears. In some embodiments, collapsing comprises detecting and/or correcting errors, nicks or lesions present in the sense or anti-sense strand of the tagged parent polynucleotides or amplified progeny polynucleotides.

[00121] The disclosure also provides for a method comprising detecting genetic variation in non-uniquely tagged initial starting genetic material with a sensitivity of at least 5%, at least 1%, at least 0.5%, at least 0.1% or at least 0.05%.

[00122] In some embodiments, the initial starting genetic material is provided in an amount less than 100 ng of nucleic acid, the genetic variation is copy number/heterozygosity variation and detecting is performed with sub-chromosomal resolution; e.g., at least 100 megabase resolution, at least 10 megabase resolution, at least 1 megabase resolution, at least 100 kilobase resolution, at least 10 kilobase resolution or at least 1 kilobase resolution. In some embodiments, the method comprises providing a plurality of sets of tagged parent polynucleotides, wherein each set is mappable to a different mappable position in a reference sequence. In some embodiments, the mappable position in the reference sequence is the locus of a tumor marker and analyzing comprises detecting the tumor marker in the set of consensus sequences.

[00123] In some embodiments, the tumor marker is present in the set of consensus sequences at a frequency less than the error rate introduced at the amplifying step. In some embodiments, the at least one set is a plurality of sets, and the mappable position of the reference sequence comprise a plurality of mappable positions in the reference sequence, each of which mappable position is the locus of a tumor marker. In some embodiments, analyzing comprises detecting copy number variation of consensus sequences between at least two sets of parent polynucleotides. In some embodiments, analyzing comprises detecting the presence of sequence variations compared with the reference sequences.

[00124] In some embodiments, analyzing comprises detecting the presence of sequence variations compared with the reference sequences and detecting copy number variation of consensus sequences between at least two sets of parent polynucleotides. In some embodiments, collapsing comprises: (i) grouping sequences reads sequenced from amplified progeny

IDT Exhibit 1020.4031