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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 December 2024 (19.12.2024)

WIPO | PCT

(10) International Publication Number
WO 2024/256581 A1

(51) International Patent Classification:
C12Q 1/6806 (2018.01)

(21) International Application Number:
PCT/EP2024/066447

(22) International Filing Date:
13 June 2024 (13.06.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/508,166 14 June 2023 (14.06.2023) US

(71) Applicant: **ILLUMINA, INC.** [US/US]; 5200 Illumina Way, San Diego, California 92122 (US).

(72) Inventors: **BROWN, Colin**; c/o Illumina, Inc., 5200 Illumina Way, San Diego, California 92122 (US). **SHULTZ-ABERGER, Sarah**; c/o Illumina, Inc., 5200 Illumina Way, San Diego, California 92122 (US). **KARUNAKARAN, Aathavan**; c/o Illumina, Inc., 5200 Illumina Way, San Diego, California 92122 (US).

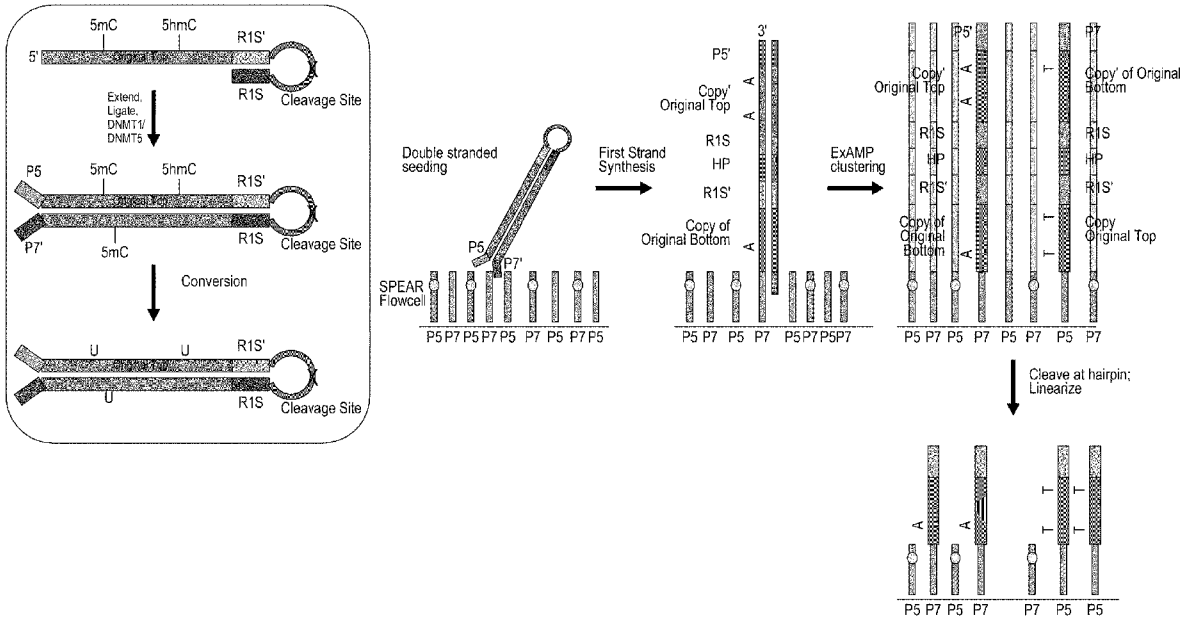
(74) Agent: **MARKS & CLERK LLP**; 15 Fetter Lane, London EC4A 1BW (GB).

(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, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, 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, ME, 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).

(54) Title: DETERMINATION OF MODIFIED CYTOSINES

Figure 18



(57) Abstract: Aspects relate to methods of distinguishing between different types of modified cytosines in nucleic acid sequences.

[Continued on next page]

WO 2024/256581 A1

Published:

— *with international search report (Art. 21(3))*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*
— *with sequence listing part of description (Rule 5.2(a))*
— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

DETERMINATION OF MODIFIED CYTOSINES

Field of the Invention

The invention relates to methods of distinguishing between different types of modified cytosines in nucleic acid sequences.

Background of the Invention

Modified cytosines, including 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC), are well-studied epigenetic modifications that play fundamental roles in human development and disease. Its genome-wide distribution differs between tissue types, and between healthy and diseased states.

As a result, there has been an intense focus on developing methods for mapping modified cytosines at single base resolution, with minimal loss of sample polynucleotide quantity, quality, and complexity.

Füllgrabe et al. in "Accurate simultaneous sequencing of genetic and epigenetic bases in DNA", BioRxiv, 2022.07.08.499285 (https://doi.org/10.1101/2022.07.08.499285) describe a single-base-resolution sequencing methodology (6-letter sequencing) to distinguish between 5-methylcytosine and 5-hydroxymethylcytosine.

However, there remains a need to develop new methods for detecting modified cytosines, and in particular methods that enable quick and accurate distinction between different types of modified cytosines.

Summary of the Invention

According to an aspect of the present invention, there is provided a method of preparing polynucleotide templates for distinguishing between modified cytosines, comprising:

(a) providing a polynucleotide library hairpin strand comprising:

a double-stranded polynucleotide comprising a forward library strand and a reverse library strand,

a hairpin loop adaptor ligated to an end of the double-stranded polynucleotide, wherein the hairpin loop adaptor comprises a cleavable site,

wherein the polynucleotide library hairpin strand has been generated from a precursor polynucleotide library hairpin strand such that any CpG dyads in the precursor polynucleotide library hairpin comprising only unmodified cytosine are converted to a first dyad in the polynucleotide library hairpin strand, any CpG dyads in the precursor polynucleotide library hairpin comprising 5-methylcytosine are converted to a second dyad in the polynucleotide library hairpin strand, and any CpG dyads in the precursor polynucleotide library hairpin comprising 5-hydroxymethylcytosine are converted to a third dyad in the polynucleotide library hairpin strand,

wherein the first dyad, second dyad and third dyad are different to each other when read; and

(b) synthesising at least one template strand by generating a complement of the polynucleotide library hairpin strand, each of the template strands comprising a forward template strand complementary to the forward library strand, a spacer strand complementary to the hairpin loop adaptor, and a reverse template strand complementary to the reverse library strand, wherein the spacer strand comprises a first cleavable site.

In one aspect, the method further comprises a step of:

(c) synthesising at least one template complement strand by generating a complement of the template strand, each of the template complement strands comprising a forward complement template strand, a spacer complement strand, and a reverse complement template strand, wherein the spacer complement strand comprises a second cleavable site.

In one aspect, the method further comprises a step of:

(d) cleaving the first cleavable site on the at least one template strand to generate at least one first polynucleotide sequence each comprising a first portion and cleaving the second cleavable site on the at least one template complement strand to generate at least one second polynucleotide sequence each comprising a second portion,

wherein the first portion corresponds with the forward template strand and the second portion corresponds with the reverse complement template strand, or wherein the first portion corresponds with the reverse template strand and the second portion corresponds with the forward complement template strand.

In one aspect, the first portion is at least 25 base pairs and the second portion is at least 25 base pairs.

In one aspect, the first cleavable site is a first restriction site for an endonuclease.

In one aspect, the second cleavable site is a second restriction site for an endonuclease.

In one aspect, the at least one first polynucleotide sequence each comprise a first sequencing primer binding site.

In one aspect, the first sequencing primer binding site is located after a 3'-end of the first portion.

In one aspect, the at least one second polynucleotide sequence each comprise a second sequencing primer binding site.

In one aspect, the second sequencing primer binding site is located after a 3'-end of the second portion.

In one aspect, where CpG dyads comprising only unmodified cytosine were present in the precursor polynucleotide library hairpin, then a double C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; where CpG dyads comprising 5-methylcytosine were present in the precursor polynucleotide library hairpin, then a double mismatch is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; and where CpG dyads comprising 5-hydroxymethylcytosine were present in the precursor polynucleotide library hairpin, then a single mismatch and single C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad.

In one aspect, where CpG dyads comprising only unmodified cytosine were present in the precursor polynucleotide library hairpin, then a double mismatch is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; where CpG dyads comprising 5-methylcytosine were present in the precursor polynucleotide library hairpin, then a double C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; and where CpG dyads comprising 5-hydroxymethylcytosine were present in the precursor polynucleotide library hairpin, then a single mismatch and single C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad.

In one aspect, the method further comprises a step of preparing the first portion and the second portion for concurrent sequencing.

In one aspect, the method comprises simultaneously contacting first sequencing primer binding sites located after a 3'-end of the first portions with first primers and second sequencing primer binding sites located after a 3'-end of the second portions with second primers.

In one aspect, the method further comprises a step of processing the at least one first polynucleotide sequence comprising a first portion and the at least one second polynucleotide sequence comprising a second portion, such that a proportion of first portions are capable of generating a first signal and a proportion of second portions are capable of generating a second signal.

In one aspect, the processing involves selective processing to cause an intensity of the first signal to be greater than an intensity of the second signal.

In one aspect, a concentration of the first portions capable of generating the first signal is greater than a concentration of the second portions capable of generating the second signal.

In one aspect, a ratio between the concentration of the first portions capable of generating the first signal and the concentration of the second portions capable of generating the second signal is between 1.25:1 to 5:1.

In one aspect, the ratio is between 1.5:1 to 3:1.

In one aspect, the ratio is about 2:1.

In one aspect, selective processing comprises preparing for selective sequencing or conducting selective sequencing.

In one aspect, selectively processing comprises conducting selective amplification.

In one aspect, selectively processing comprises contacting first sequencing primer binding sites located after a 3'-end of the first portions with first primers and contacting second sequencing primer binding sites located after a 3'-end of the second portions with second primers, wherein the second primers comprises a mixture of blocked second primers and unblocked second primers.

In one aspect, the blocked second primer comprises a blocking group at a 3' end of the blocked second primer.

In one aspect, the blocking group is selected from the group consisting of: a hairpin loop, a deoxynucleotide, a deoxyribonucleotide, a hydrogen atom instead of a 3'-OH group, a phosphate group, a phosphorothioate group, a propyl spacer, a modification blocking the 3'-hydroxyl group, or an inverted nucleobase.

In one aspect, the selective processing comprises selectively removing some or substantially all of second immobilised primers that are not yet extended, and conducting a further amplification cycle in order to selectively amplify the first polynucleotide sequence(s) relative to the second polynucleotide sequence(s).

In one aspect, selectively processing comprises selectively blocking some or substantially all of second immobilised primers that are not yet extended using a primer blocking agent, wherein the primer blocking agent is configured to limit or prevent synthesis of a strand extending from the second immobilised primer, and conducting a further amplification cycle in order to selectively amplify the first polynucleotide sequence(s) relative to the second polynucleotide sequence(s).

In one aspect, the primer blocking agent is added whilst first polynucleotide sequence(s) are hybridised to the second immobilised primers.

In one aspect, the method comprises contacting some or substantially all of the second immobilised primers with an extended primer sequence, wherein the extended primer sequence is substantially complementary to the second immobilised primer and further comprises a 5' additional nucleotide; and adding the primer blocking agent, wherein the primer blocking agent is complementary to the 5' additional nucleotide.

In one aspect, the primer blocking agent is a blocked nucleotide.

In one aspect, the blocked nucleotide comprises a blocking group at a 3' end of the blocked nucleotide.

In one aspect, the blocking group is selected from the group consisting of: a hairpin loop, a deoxynucleotide, a deoxyribonucleotide, a hydrogen atom instead of a 3'-OH group, a phosphate

group, a phosphorothioate group, a propyl spacer, a modification blocking the 3'-hydroxyl group, or an inverted nucleobase.

In one aspect, the blocked nucleotide is A or G.

In one aspect, the first signal and the second signal are spatially resolved.

In one aspect, the first signal and the second signal are spatially unresolved.

In one aspect, the at least one first polynucleotide sequence comprising the first portion and the at least one second polynucleotide sequence comprising the second portion are attached to a solid support.

In one aspect, the solid support is a flow cell.

In one aspect, the at least one first polynucleotide sequence comprising the first portion and the at least one second polynucleotide sequence comprising the second portion forms a cluster on the solid support.

In one aspect, the cluster is formed by bridge amplification.

In one aspect, the at least one first polynucleotide sequence comprising the first portion and the at least one second polynucleotide sequence comprising the second portion form a duoclonal cluster.

In one aspect, the solid support comprises at least one first immobilised primer and at least one second immobilised primer.

In one aspect, the first immobilised primer comprises a sequence as defined in SEQ ID NO. 1 or 5, or a variant or fragment thereof; and the second immobilised primer comprises a sequence as defined in SEQ ID NO. 2, or a variant or fragment thereof.

In one aspect, each first polynucleotide sequence is attached to a first immobilised primer, and wherein each second polynucleotide sequence is attached to a second immobilised primer.

In one aspect, each first polynucleotide sequence comprises a second adaptor sequence and wherein each second polynucleotide sequence comprises a first adaptor sequence, wherein the

second adaptor sequence is substantially complementary to the second immobilised primer and wherein the first adaptor sequence is substantially complementary to the first immobilised primer.

According to another aspect of the present invention, there is provided a method of sequencing polynucleotide sequences to distinguish between modified cytosines, comprising:

preparing polynucleotide templates for distinguishing between modified cytosines using a method as described herein;

sequencing nucleobases in the first portion and the second portion; and

identifying the presence of 5-methylcytosine or 5-hydroxymethylcytosine by detecting differences when comparing a sequence output from the first portion with a sequence output from the second portion.

In one aspect, the step of sequencing nucleobases in the first portion and the second portion involves concurrent sequencing of nucleobases in the first portion and the second portion.

In one aspect, the step of sequencing nucleobases comprises performing sequencing-by-synthesis.

In one aspect, the method further comprises a step of conducting paired-end reads.

In one aspect, the step of concurrently sequencing nucleobases comprises:

(a) obtaining first intensity data comprising a combined intensity of a first signal component obtained based upon a respective first nucleobase at the first portion and a second signal component obtained based upon a respective second nucleobase at the second portion, wherein the first and second signal components are obtained simultaneously;

(b) obtaining second intensity data comprising a combined intensity of a third signal component obtained based upon the respective first nucleobase at the first portion and a fourth signal component obtained based upon the respective second nucleobase at the second portion, wherein the third and fourth signal components are obtained simultaneously;

(c) selecting one of a plurality of classifications based on the first and the second intensity data, wherein each classification represents a possible combination of respective first and second nucleobases; and

(d) based on the selected classification, base calling the respective first and second nucleobases.

In one aspect, selecting the classification based on the first and second intensity data comprises selecting the classification based on the combined intensity of the first and second signal components and the combined intensity of the third and fourth signal components.

In one aspect, the plurality of classifications comprises sixteen classifications, each classification representing one of sixteen unique combinations of first and second nucleobases.

In one aspect, the first signal component, second signal component, third signal component and fourth signal component are generated based on light emissions associated with the respective nucleobase.

In one aspect, the light emissions are detected by a sensor, wherein the sensor is configured to provide a single output based upon the first and second signals.

In one aspect, the sensor comprises a single sensing element.

In one aspect, the method further comprises repeating steps (a) to (d) for each of a plurality of base calling cycles.

According to another aspect of the present invention, there is provided a kit comprising instructions for preparing polynucleotide templates for distinguishing between modified cytosines as described herein, and/or for sequencing polynucleotide sequences to distinguish between modified cytosines as described herein.

According to another aspect of the present invention, there is provided a data processing device comprising means for carrying out a method as described herein.

In one aspect, the data processing device is a polynucleotide sequencer.

According to another aspect of the present invention, there is provided a computer program product comprising instructions which, when the program is executed by a processor, cause the processor to carry out a method as described herein.

According to another aspect of the present invention, there is provided a computer-readable storage medium comprising instructions which, when executed by a processor, cause the processor to carry out a method as described herein.

According to another aspect of the present invention, there is provided a computer-readable data carrier having stored thereon a computer program product as described herein.

According to another aspect of the present invention, there is provided a data carrier signal carrying a computer program product as described herein.

According to another aspect of the present invention, there is provided a method of preparing a polynucleotide library hairpin strand, comprising:

(a) providing a double-stranded polynucleotide comprising a precursor forward library strand and a precursor reverse library strand; and

(b) ligating a hairpin loop adaptor to an end of the double-stranded polynucleotide to generate a first hairpin polynucleotide, wherein the hairpin loop adaptor comprises a cleavable site.

In one aspect, the hairpin loop adaptor comprises a base-paired stem and a non-base-paired loop.

In one aspect, the cleavable site is located in the non-base-paired loop.

In one aspect, the hairpin loop adaptor connects a 3'-end of the precursor forward library strand with a 5'-end of the precursor reverse library strand; or wherein the hairpin loop adaptor connects a 3'-end of the precursor reverse library strand with a 5'-end of the precursor forward library strand.

In one aspect, the cleavable site is a restriction site for an endonuclease.

In one aspect, the method further comprises a step of:

(c) removing the precursor reverse library strand from the first hairpin polynucleotide to generate a second hairpin polynucleotide comprising the precursor forward library strand and the hairpin loop adaptor, wherein the hairpin loop adaptor comprises the cleavable site.

In one aspect, the method further comprises a step of:

(d) forming a resynthesised reverse library strand from the second hairpin polynucleotide to generate a third hairpin polynucleotide, wherein when any cytosine bases are present in the resynthesised reverse library strand, then all such cytosine bases are unmodified cytosine.

In one aspect, the method further comprises a step of:

(e) exposing the third hairpin polynucleotide to an enzyme configured to convert hemimethylated 5-methylcytosine CpG dyads to fully methylated 5-methylcytosine CpG dyads,

but not convert hemimethylated 5-hydroxymethylcytosine dyads, in order to generate a fourth hairpin polynucleotide.

In one aspect, the enzyme configured to convert hemimethylated 5-methylcytosine CpG dyads to fully methylated 5-methylcytosine CpG dyads, but not convert hemimethylated 5-hydroxymethylcytosine dyads is a DNA methyltransferase.

In one aspect, the enzyme configured to convert hemimethylated 5-methylcytosine CpG dyads to fully methylated 5-methylcytosine CpG dyads, but not convert hemimethylated 5-hydroxymethylcytosine dyads is a member of the DNA methyltransferase 1 (DNMT1) family or the DNA methyltransferase 5 (DNMT5) family.

In one aspect, the method further comprises a step of:

(f) exposing the fourth hairpin polynucleotide to a conversion agent configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase which is read as thymine/uracil, or to a conversion agent configured to convert an unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil, in order to generate a fifth hairpin polynucleotide.

In one aspect, the conversion agent is configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase which is read as thymine/uracil.

In one aspect, the conversion agent is configured to convert unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil.

In one aspect, the conversion agent comprises a chemical agent and/or an enzyme.

In one aspect, the conversion agent comprises a boron-based reducing agent and a ten-eleven translocation (TET) methylcytosine dioxygenase.

In one aspect, the boron-based reducing agent is an amine-borane compound or an azine-borane compound.

In one aspect, the boron-based reducing agent is selected from the group consisting of pyridine borane, 2-picoline borane, t-butylamine borane, ammonia borane, ethylenediamine borane and dimethylamine borane.

In one aspect, the TET methylcytosine dioxygenase is a member of the TET1 subfamily, the TET2 subfamily, or the TET3 subfamily.

In one aspect, the conversion agent comprises sulfite.

In one aspect, the sulfite is bisulfite.

In one aspect, the bisulfite is sodium bisulfite.

In one aspect, the conversion agent comprises a cytidine deaminase.

In one aspect, the cytidine deaminase is a wild-type cytidine deaminase or a mutant cytidine deaminase.

In one aspect, the cytidine deaminase is a member of the AID subfamily, the APOBEC1 subfamily, the APOBEC2 subfamily, the APOBEC3A subfamily, the APOBEC3B subfamily, the APOBEC3C subfamily, the APOBEC3D subfamily, the APOBEC3F subfamily, the APOBEC3G subfamily, the APOBEC3H subfamily, or the APOBEC4 subfamily.

In one aspect, the cytidine deaminase is a member of the APOBEC3A subfamily.

In one aspect, the cytidine deaminase comprises amino acid substitution mutations at positions functionally equivalent to (Tyr/Phe)130 and Tyr132 in a wild-type APOBEC3A protein.

In one aspect, the (Tyr/Phe)130 is Tyr130, and the wild-type APOBEC3A protein is SEQ ID NO. 16.

In one aspect, the substitution mutation at the position functionally equivalent to Tyr130 comprises Ala, Val or Trp.

In one aspect, the substitution mutation at the position functionally equivalent to Tyr132 comprises a mutation to His, Arg, Gln or Lys.

In one aspect, the mutant cytidine deaminase comprises a ZDD motif H-[P/A/V]-E-$X_{[23-28]}$-P-C-$X_{[2-4]}$-C (SEQ ID NO. 51).

In one aspect, the mutant cytidine deaminase is a member of the APOBEC3A subfamily and comprises a ZDD motif $HXEX_{24}SW(S/T)PCX_{[2-4]}CX_6FX_8LX_5R(L/I)YX_{[8-11]}LX_2LX_{[10]}M$ (SEQ ID NO. 52).

In one aspect, the mutant cytidine deaminase converts 5-methylcytosine to thymine by deamination at a greater rate than a conversion rate of cytosine to uracil by deamination.

In one aspect, the rate is at least 100-fold greater.

In one aspect, the conversion agent further comprises a glycosyltransferase.

In one aspect, the glycosyltransferase is a β-glucosyltransferase.

In one aspect, the method further comprises a step of ligating a flanking adaptor to an end of the double-stranded polynucleotide away from the hairpin loop adaptor to the third hairpin polynucleotide, the fourth hairpin polynucleotide or the fifth hairpin polynucleotide, wherein the flanking adaptor comprises a primer-binding sequence and a primer-binding complement sequence.

In one aspect, the flanking adaptor is a forked adaptor comprising a base-paired stem, a first arm and a second arm.

In one aspect, the primer-binding sequence is located on the first arm, and the primer-binding complement sequence is located on the second arm.

According to another aspect of the present invention, there is provided a polynucleotide library hairpin strand prepared according to a method of preparing a polynucleotide library hairpin strand as described herein.

Description of the Drawings

Figure 1 shows a forward strand, reverse strand, forward complement strand, and reverse complement strand of a polynucleotide molecule.

Figure 2 shows the steps involved in a loop fork method.

Figure 3 shows an example of a polynucleotide sequence prepared using a loop fork method.

Figure 4 shows an example of a polynucleotide sequence prepared using a loop fork method.

Figure 5 shows a typical solid support.

Figure 6 shows the stages of bridge amplification for polynucleotide templates prepared using a loop fork method and the generation of an amplified cluster, comprising (Panel A) a concatenated library strand hybridising to a immobilised primer; (Panel B) generation of a template strand from the library strand; (Panel C) dehybridisation and washing away the library strand; (Panel D) generation of a template complement strand from the template strand via bridge amplification and dehybridisation of the sequence bridge; and (Panel E) further amplification to provide a plurality of template and template complement strands.

Figure 7 shows the detection of nucleobases using 4-channel, 2-channel and 1-channel chemistry.

Figure 8 shows a method of selective sequencing.

Figure 9 shows a method of selective amplification comprising (Panel A) starting from a plurality of template and template complement strands; (Panel B) selective cleavage of one type of immobilised primer from the support; (Panel C) only template (or template complement) strands complementary to the free immobilised primer anneal and undergo bridge amplification, (Panel D) producing different proportions of template and template complement strands; (Panel E) subsequent standard (non-selective) sequencing occurs in different proportions enabling signal differentiation.

Figure 10 shows a method of selective amplification comprising (Panel A) template and template complement strands annealing to immobilised primers; (Panel B) addition of a primer-blocking agent that binds only to one type of immobilised primer, preventing the extension from that one type of immobilised primer, preventing the extension from one type of immobilised primer; (Panel C) producing different proportions of template and template complement strands; (Panel D) subsequent standard (non-selective) sequencing occurs in different proportions enabling signal differentiation.

Figure 11 shows a method of selective amplification comprising (Panel A) flowing a (or a plurality of) extended primer sequence(s) containing at least one additional 5' nucleotide across the surface of the solid support; (Panel B) addition of a primer-blocking agent that binds only to

one type of immobilised primer and is complementary to the additional 5' nucleotide of the extended primer sequence, preventing the extension from one type of immobilised primer.

Figure 12 is a plot showing graphical representations of sixteen distributions of signals generated by polynucleotide sequences according to one embodiment.

Figure 13 is a flow diagram showing a method for base calling according to one embodiment.

Figure 14 shows a prior art method for detecting 5-hydroxymethylcytosine and 5-methylcytosine. This involves conducting a sequencing run to determine the presence of both 5-hydroxymethylcytosine and 5-methylcytosine (left), then another separate sequencing run to determine the presence of only 5-methylcytosine (right). The presence of 5-hydroxymethylcytosine is obtained by comparing the two separate runs.

Figure 15 shows a prior art method (Füllgrabe et al.) for detecting 5-hydroxymethylcytosine and 5-methylcytosine using hairpin polynucleotides. In the prior art method, the hairpin loop adaptor does not comprise a cleavable site.

Figure 16 shows an example workflow of preparing a polynucleotide library hairpin strand according to a method as described herein, then a subsequent example workflow for preparing polynucleotide templates for distinguishing between modified cytosines according to a method as described herein. 5-methylcytosine is represented as 5m in bubbles, 5-hydroxymethylcytosine is represented as 5hm in bubbles, U represents uracil, thymine or a nucleobase which is read as thymine/uracil, and X represents A or T (the exact nature of A or T is not material and is not shown for clarity – XX base pairs are AT/TA base pairs which remain unchanged during the preparation process).

Figure 17 shows another example workflow of preparing a polynucleotide library hairpin strand according to a method as described herein, then a subsequent example workflow for preparing polynucleotide templates for distinguishing between modified cytosines according to a method as described herein. 5-methylcytosine is represented as 5m in bubbles, 5-hydroxymethylcytosine is represented as 5hm in bubbles, U represents uracil, thymine or a nucleobase which is read as thymine/uracil, and X represents A or T (the exact nature of A or T is not material and is not shown for clarity – XX base pairs are AT/TA base pairs which remain unchanged during the preparation process).

Figure 18 shows a further example workflow conducted according to the method shown in Figure 17.

Detailed Description

All patents, patent applications, and other publications referred to herein, including all sequences disclosed within these references, are expressly incorporated herein by reference, to the same extent as if each individual publication, patent or patent application was specifically and individually indicated to be incorporated by reference. All documents cited are, in relevant part, incorporated herein by reference in their entireties for the purposes indicated by the context of their citation herein. However, the citation of any document is not to be construed as an admission that it is prior art with respect to the present disclosure.

The present invention can be used in sequencing, in particular concurrent sequencing. Methodologies applicable to the present invention have been described in WO 08/041002, WO 07/052006, WO 98/44151, WO 00/18957, WO 02/06456, WO 07/107710, WO05/068656, US 13/661,524 and US 2012/0316086, the contents of which are herein incorporated by reference. Further information can be found in US 20060024681, US 20060292611, WO 06/110855, WO 06/135342, WO 03/074734, WO07/010252, WO 07/091077, WO 00/179553, WO 98/44152 and WO 2022/087150, the contents of which are herein incorporated by reference.

As used herein, the term "variant" refers to a variant polypeptide sequence or part of the polypeptide sequence that retains desired function of the full non-variant sequence. For example, a desired function of the immobilised primer retains the ability to bind (i.e. hybridise) to a target sequence.

As used in any aspect described herein, a "variant" has at least 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%, or at least 99% overall sequence identity to the non-variant nucleic acid sequence. The sequence identity of a variant can be determined using any number of sequence alignment programs known in the art. As an example, Emboss Stretcher from the EMBL-EBI may be used as found on the Internet at: ebi.ac.uk/Tools/psa/emboss_stretcher/ (using default parameters: pair output format, Matrix = BLOSUM62, Gap open = 1, Gap extend = 1 for proteins; pair output format, Matrix = DNAfull, Gap open = 16, Gap extend = 4 for nucleotides).

As used herein, the term "fragment" refers to a functionally active series of consecutive nucleic acids from a longer nucleic acid sequence. The fragment may be at least 99%, at least 95%, at least 90%, at least 80%, at least 70%, at least 60%, at least 50%, at least 40% or at least 30% the length of the longer nucleic acid sequence. A fragment as used herein may also retain the ability to bind (i.e. hybridise) to a target sequence.

Sequencing generally comprises four fundamental steps: 1) library preparation to form a plurality of target polynucleotides for identification; 2) cluster generation to form an array of amplified template polynucleotides; 3) sequencing the cluster array of amplified template polynucleotides; and 4) data analysis to identify characteristics of the target polynucleotides from the amplified template polynucleotide sequences. These steps are described in greater detail below.

Library strands and template terminology

As shown in Figure 1, for a given double-stranded polynucleotide sequence 100 to be identified, the polynucleotide sequence 100 comprises a forward strand of the sequence 101 and a reverse strand of the sequence 102.

When the polynucleotide sequence 100 is replicated (e.g. using a DNA/RNA polymerase), complementary versions of the forward strand 101 of the sequence 100 and the reverse strand 102 of the sequence 100 are generated. Thus, replication of the polynucleotide sequence 100 provides a double-stranded polynucleotide sequence 100a that comprises a forward strand of the sequence 101 and a forward complement strand of the sequence 101', and a double-stranded polynucleotide sequence 100b that comprises a reverse strand of the sequence 102 and a reverse complement strand of the sequence 102'.

The term "template" may be used to describe a complementary version of the double-stranded polynucleotide sequence 100. As such, the "template" comprises a forward complement strand of the sequence 101' and a reverse complement strand of the sequence 102'. Thus, by using the forward complement strand of the sequence 101' as a template for complementary base pairing, a sequencing process (e.g. a sequencing-by-synthesis or a sequencing-by-ligation process) reproduces information that was present in the original forward strand of the sequence 101. Similarly, by using the reverse complement strand of the sequence 102' as a template for complementary base pairing, a sequencing process (e.g. a sequencing-by-synthesis or a sequencing-by-ligation process) reproduces information that was present in the original reverse strand of the sequence 102.

The two strands in the template may also be referred to as a forward strand of the template 101' and a reverse strand of the template 102'. The complement of the forward strand of the template 101' is termed the forward complement strand of the template 101, whilst the complement of the reverse strand of the template 102' is termed the reverse complement strand of the template 102.

Language for original polynucleotide sequence 100	**Corresponding language for the "template"**
Forward strand of the sequence 101	Forward complement strand of the template 101 (sometimes referred to herein as forward complement strand 101)
Reverse strand of the sequence 102	Reverse complement strand of the template 102 (sometimes referred to herein as reverse complement strand 102)
Forward complement strand of the sequence 101'	Forward strand of the template 101' (sometimes referred to herein as forward strand 101')
Reverse complement strand of the sequence 102'	Reverse strand of the template 102' (sometimes referred to herein as reverse strand 102')

Library preparation

Library preparation is the first step in any high-throughput sequencing platform. These libraries allow templates to be generated via complementary base pairing that can subsequently be clustered and amplified. During library preparation, nucleic acid sequences, for example genomic DNA sample, or cDNA or RNA sample, is converted into a sequencing library, which can then be sequenced. By way of example with a DNA sample, the first step in library preparation is random fragmentation of the DNA sample. Sample DNA is first fragmented and the fragments of a specific size (typically 200–500 bp, but can be larger) are ligated, sub-cloned or "inserted" in-between two oligo adaptors (adaptor sequences). The original sample DNA fragments are referred to as "inserts". The target polynucleotides may advantageously also be size-fractionated prior to modification with the adaptor sequences.

As described herein, typically the templates to be generated from the libraries may include separate polynucleotide sequences, in particular a first polynucleotide sequence comprising a first portion and a second polynucleotide sequence comprising a second portion. Generating these

templates from particular libraries may be performed according to methods known to persons of skill in the art. However, some example approaches of preparing libraries suitable for generation of such templates are described below.

In some embodiments, the library may be prepared using a loop fork method, which is described below. This procedure may be used, for example, for preparing templates including a first polynucleotide sequence comprising a first portion and a second polynucleotide sequence comprising a second portion, wherein the first portion is a forward strand of the template, and the second portion is a reverse complement strand of the template (or alternatively, wherein the first portion is a reverse strand of the template, and the second portion is a forward complement strand of the template). A representative process for conducting a loop fork method is shown in Figure 2.

Starting from a double-stranded polynucleotide sequence comprising a forward strand of the sequence and a reverse strand of the sequence, adaptors may be ligated to a first end of the sequence (e.g. using processes as described in more detail in e.g. WO 07/052006, or "tagmentation" methods as described above). A second end of the sequence (different from the first end) may be ligated to a loop, which connects the forward strand of the sequence and the reverse strand of the sequence, thus generating a loop fork ligated polynucleotide sequence. Between these principal steps of ligating the loop and the adaptors during the generation of a loop fork ligated polynucleotide sequence, additional steps (e.g. removal of the reverse library strand, resynthesis of the reverse library strand, exposure to an enzyme configured to convert hemimethylated 5-methylcytosine CpG dyads to fully methylated 5-methylcytosine CpG dyads, but not convert hemimethylated 5-hydroxymethylcytosine dyads, and exposure to a conversion agent) may be conducted – these additional steps are explained in greater detail herein. Once the loop fork ligated polynucleotide sequence has been generated and appropriately treated, the library is now ready for seeding, clustering and amplification.

As will be described later, during clustering and amplification, further processes may be used to generate templates including a first polynucleotide sequence comprising a first portion and a second polynucleotide sequence comprising a second portion, wherein the first portion is a forward strand of the template, and the second portion is a reverse complement strand of the template (or alternatively, wherein the first portion is a reverse strand of the template, and the second portion is a forward complement strand of the template).

The processes described above in relation to loop fork methods generate libraries that have self-tandem insert polynucleotides.

Thus, one strand of a polynucleotide within a polynucleotide library may comprise, in a 5' to 3' direction, a second primer-binding complement sequence 302 (e.g. P7), an optional first terminal sequencing primer binding site complement 303', a first insert sequence 401 (A and B), a loop sequence 403 (L) (also referred to herein as a hairpin loop adaptor), a second insert sequence 402 (B' and A'), an optional second terminal sequencing primer binding site 304, and a first primer-binding sequence 301' (e.g. P5') (Figures 3 and 4).

Alternatively, or in addition, one or more sequencing primer binding sites (or complements) may be provided within the loop sequence 403 (L) (or hairpin loop adaptor).

Although not shown in Figures 3 and 4, the strand may further comprise one or more index sequences. As such, a first index sequence (e.g. i7) may be provided between the second primer-binding complement sequence 302 (e.g. P7) and the optional first terminal sequencing primer binding site complement 303'. Separately, or in addition, a second index complement sequence (e.g. i5') may be provided between the optional second terminal sequencing primer binding site 304 and the first primer-binding sequence 301' (e.g. P5'). Thus, in some embodiments, one strand of a polynucleotide within a polynucleotide library may comprise, in a 5' to 3' direction, a second primer-binding complement sequence 302 (e.g. P7), a first index sequence (e.g. i7), an optional first terminal sequencing primer binding site complement 303', a first insert sequence 401 (A and B), a loop sequence 403 (L) (or hairpin loop adaptor), a second insert sequence 402 (B' and A'), an optional second terminal sequencing primer binding site 304, a second index complement sequence (e.g. i5'), and a first primer-binding sequence 301' (e.g. P5').

Alternatively, or in addition, one or more index sequences (or complements) may be provided within the loop sequence 403 (L) (or hairpin loop adaptor).

It should be noted that the arrangement of the loop sequence 403 (L) (or hairpin loop adaptor) is shown in Figures 3 and 4 as being ligated on the right hand side of the double-stranded fragment, and the P5'/P7 adaptor being ligated on the left hand side of the double-stranded fragment – in other words, the loop sequence 403 (L) (or hairpin loop adaptor) connects a 3'-end of the forward library strand with a 5'-end of the reverse library strand. However, the positions of the hairpin loop adaptor and the P5'/P7 adaptor may be reversed, such that the loop sequence 403 (L) (or hairpin loop adaptor) is ligated on the left hand side of the double-stranded fragment, and the P5'/P7 adaptor being ligated on the right hand side of the double-stranded fragment – in other words, the loop sequence 403 (L) (or hairpin loop adaptor) connects a 3'-end of the reverse library

strand with a 5'-end of the forward library strand. Both variations are encompassed by the present disclosure.

In addition, it should be noted that a P5'/P7 adaptor is used in Figures 3 and 4. However, a P7'/P5 adaptor may equally be used, where P7' is used instead of P5' and P5 is used instead of P7. Both variations are encompassed by the present disclosure.

In one embodiment, the first insert sequence 401 may comprise a forward strand of the sequence 101, and the second insert complement sequence 402' may comprise a reverse complement strand of the sequence 102' (or the first insert sequence 401 may comprise a reverse strand of the sequence 102, and the second insert complement sequence 402' may comprise a forward complement strand of the sequence 101'), for example where the library is prepared using a loop fork method.

Although Figure 3 shows the presence of a first terminal sequencing primer binding site complement 303', a second terminal sequencing primer binding site 304, these are optional as mentioned above (in addition, their complements, a second terminal sequencing primer binding site complement 304', and a first terminal sequencing primer binding site 303). Accordingly, these sections may be omitted from the library.

As will be understood by the skilled person, a double-stranded nucleic acid will typically be formed from two complementary polynucleotide strands comprised of deoxyribonucleotides or ribonucleotides joined by phosphodiester bonds, but may additionally include one or more ribonucleotides and/or non-nucleotide chemical moieties and/or non-naturally occurring nucleotides and/or non-naturally occurring backbone linkages. In particular, the double-stranded nucleic acid may include non-nucleotide chemical moieties, e.g. linkers or spacers, at the 5' end of one or both strands. By way of non-limiting example, the double-stranded nucleic acid may include methylated nucleotides, uracil bases, phosphorothioate groups, peptide conjugates etc. Such non-DNA or non-natural modifications may be included in order to confer some desirable property to the nucleic acid, for example to enable covalent, non-covalent or metal-coordination attachment to a solid support, or to act as spacers to position the site of cleavage an optimal distance from the solid support. A single stranded nucleic acid consists of one such polynucleotide strand. Where a polynucleotide strand is only partially hybridised to a complementary strand – for example, a long polynucleotide strand hybridised to a short nucleotide primer – it may still be referred to herein as a single stranded nucleic acid.

A sequence comprising at least a primer-binding sequence (e.g. a primer-binding sequence and a sequencing primer binding site, or a combination of a primer-binding sequence, an index sequence and a sequencing primer binding site) may be referred to herein as an adaptor sequence, and an insert is flanked by a 5' adaptor sequence and a 3' adaptor sequence. The primer-binding sequence may also comprise a sequencing primer for the index read.

As used herein, an "adaptor" refers to a sequence that comprises a short sequence-specific oligonucleotide that is ligated to the 5′ and 3′ ends of each DNA (or RNA) fragment in a sequencing library as part of library preparation. The adaptor sequence may further comprise non-peptide linkers.

In a further embodiment, the P5' and P7' primer-binding sequences are complementary to short primer sequences (or lawn primers) present on the surface of a flow cell. Binding of P5' and P7' to their complements (P5 and P7) on – for example – the surface of the flow cell, permits nucleic acid amplification. As used herein "'" denotes the complementary strand.

The primer-binding sequences in the adaptor which permit hybridisation to amplification primers (e.g. lawn primers) will typically be around 20-40 nucleotides in length, although the invention is not limited to sequences of this length. The precise identity of the amplification primers (e.g. lawn primers), and hence the cognate sequences in the adaptors, are generally not material to the invention, as long as the primer-binding sequences are able to interact with the amplification primers in order to direct PCR amplification. The sequence of the amplification primers may be specific for a particular target nucleic acid that it is desired to amplify, but in other embodiments these sequences may be "universal" primer sequences which enable amplification of any target nucleic acid of known or unknown sequence which has been modified to enable amplification with the universal primers. The criteria for design of PCR primers are generally well known to those of ordinary skill in the art.

The index sequences (also known as a barcode or tag sequence) are unique short DNA (or RNA) sequences that are added to each DNA (or RNA) fragment during library preparation. The unique sequences allow many libraries to be pooled together and sequenced simultaneously. Sequencing reads from pooled libraries are identified and sorted computationally, based on their barcodes, before final data analysis. Library multiplexing is also a useful technique when working with small genomes or targeting genomic regions of interest. Multiplexing with barcodes can exponentially increase the number of samples analysed in a single run, without drastically increasing run cost or run time. Examples of tag sequences are found in WO05/068656, whose contents are incorporated herein by reference in their entirety. The tag can be read at the end of

the first read, or equally at the end of the second read, for example using a sequencing primer complementary to the strand marked P7. The invention is not limited by the number of reads per cluster, for example two reads per cluster: three or more reads per cluster are obtainable simply by dehybridising a first extended sequencing primer, and rehybridising a second primer before or after a cluster repopulation/strand resynthesis step. Methods of preparing suitable samples for indexing are described in, for example WO 2008/093098, which is incorporated herein by reference. Single or dual indexing may also be used. With single indexing, up to 48 unique 6-base indexes can be used to generate up to 48 uniquely tagged libraries. With dual indexing, up to 24 unique 8-base Index 1 sequences and up to 16 unique 8-base Index 2 sequences can be used in combination to generate up to 384 uniquely tagged libraries. Pairs of indexes can also be used such that every i5 index and every i7 index are used only one time. With these unique dual indexes, it is possible to identify and filter indexed hopped reads, providing even higher confidence in multiplexed samples.

The sequencing primer binding sites are sequencing and/or index primer binding sites and indicate the starting point of the sequencing read. During the sequencing process, a sequencing primer anneals (i.e. hybridises) to at least a portion of the sequencing primer binding site on the template strand. The polymerase enzyme binds to this site and incorporates complementary nucleotides base by base into the growing opposite strand.

Cluster generation and amplification

Once a double stranded nucleic acid library is formed, typically, the library has previously been subjected to denaturing conditions to provide single stranded nucleic acids. Suitable denaturing conditions will be apparent to the skilled reader with reference to standard molecular biology protocols (Sambrook et al., 2001, Molecular Cloning, A Laboratory Manual, 4th Ed, Cold Spring Harbor Laboratory Press, Cold Spring Harbor Laboratory Press, NY; Current Protocols, eds Ausubel et al). In one embodiment, chemical denaturation may be used.

Following denaturation, a single-stranded library may be contacted in free solution onto a solid support comprising surface capture moieties (for example P5 and P7 lawn primers).

Thus, embodiments of the present invention may be performed on a solid support 200, such as a flowcell. However, in alternative embodiments, seeding and clustering can be conducted off-flowcell using other types of solid support.

The solid support 200 may comprise a substrate 204. See Figure 5. The substrate 204 comprises at least one well 203 (e.g. a nanowell), and typically comprises a plurality of wells 203 (e.g. a plurality of nanowells).

In one embodiment, the solid support comprises at least one first immobilised primer and at least one second immobilised primer.

Thus, each well 203 may comprise at least one first immobilised primer 201, and typically may comprise a plurality of first immobilised primers 201. In addition, each well 203 may comprise at least one second immobilised primer 202, and typically may comprise a plurality of second immobilised primers 202. Thus, each well 203 may comprise at least one first immobilised primer 201 and at least one second immobilised primer 202, and typically may comprise a plurality of first immobilised primers 201 and a plurality of second immobilised primers 202.

The first immobilised primer 201 may be attached via a 5'-end of its polynucleotide chain to the solid support 200. When extension occurs from first immobilised primer 201, the extension may be in a direction away from the solid support 200.

The second immobilised primer 202 may be attached via a 5'-end of its polynucleotide chain to the solid support 200. When extension occurs from second immobilised primer 202, the extension may be in a direction away from the solid support 200.

The first immobilised primer 201 may be different to the second immobilised primer 202 and/or a complement of the second immobilised primer 202. The second immobilised primer 202 may be different to the first immobilised primer 201 and/or a complement of the first immobilised primer 201.

The (or each of the) first immobilised primer(s) 201 may comprise a sequence as defined in SEQ ID NO. 1 or 5, or a variant or fragment thereof. The second immobilised primer(s) 202 may comprise a sequence as defined in SEQ ID NO. 2, or a variant or fragment thereof. Whilst first immobilised primer(s) 201 are shown here to correspond to P5 and second immobilised primer(s) 202 are shown here to correspond to P7, the definitions of these may be swapped – in other words, first immobilised primer(s) 201 may correspond instead to P7, and second immobilised primer(s) 202 may correspond to P5.

In some embodiments, the first immobilised primer(s) 201 and the second immobilised primer(s) 202 within a well 203 may be spatially separated from each other. In other words, the first

immobilised primer(s) 201 may occupy a first region, and the second immobilised primer(s) 202 may occupy a second region, wherein the first region and the second region do not overlap with each other. This means that any signals generated (e.g. a first signal and a second signal as referred to herein) are spatially resolved.

In other embodiments, the first immobilised primer(s) 201 and the second immobilised primer(s) 202 within a well 203 may not be spatially separated from each other. In other words, the first immobilised primer(s) 201 may occupy a first region, and the second immobilised primer(s) 202 may occupy a second region, wherein the first region and the second region may correspond to the same region or may be substantially overlapping. This means that any signals generated (e.g. a first signal and a second signal as referred to herein) are spatially unresolved.

By way of brief example, following attachment of the P5 and P7 primers to the solid support, the solid support may be contacted with the template to be amplified under conditions which permit hybridisation (or annealing – such terms may be used interchangeably) between the template and the immobilised primers. The template is usually added in free solution under suitable hybridisation conditions, which will be apparent to the skilled reader. Typically, hybridisation conditions are, for example, 5xSSC at 40°C. However, other temperatures may be used during hybridisation, for example about 50°C to about 75°C, about 55°C to about 70°C, or about 60°C to about 65°C. Solid-phase amplification can then proceed. The first step of the amplification is a primer extension step in which nucleotides are added to the 3' end of the immobilised primer using the template to produce a fully extended complementary strand. The template is then typically washed off the solid support. The complementary strand will include at its 3' end a primer-binding sequence (i.e. either P5' or P7') which is capable of bridging to the second primer molecule immobilised on the solid support and binding. Further rounds of amplification (analogous to a standard PCR reaction) leads to the formation of clusters or colonies of template molecules bound to the solid support. This is called clustering.

Thus, solid-phase amplification by either a method analogous to that of WO 98/44151 or that of WO 00/18957 (the contents of which are incorporated herein in their entirety by reference) will result in production of a clustered array comprised of colonies of "bridged" amplification products. This process is known as bridge amplification. Both strands of the amplification products will be immobilised on the solid support at or near the 5' end, this attachment being derived from the original attachment of the amplification primers. Typically, the amplification products within each colony will be derived from amplification of a single template molecule. Other amplification procedures may be used, and will be known to the skilled person. For example, amplification may be isothermal amplification using a strand displacement polymerase;

or may be exclusion amplification as described in WO 2013/188582. Further information on amplification can be found in WO 02/06456 and WO 07/107710, the contents of which are incorporated herein in their entirety by reference.

Through such approaches, a cluster of template molecules is formed, comprising copies of a template strand and copies of the complement of the template strand.

The steps of cluster generation and amplification for templates including a first polynucleotide sequence comprising a first portion and a second polynucleotide sequence comprising a second portion are illustrated below and in Figure 6.

In cases where (separate) polynucleotide strands are used, each first polynucleotide sequence may be attached (via the 5'-end of the first polynucleotide sequence) to a first immobilised primer, and wherein each second polynucleotide sequence is attached (via the 5'-end of the second polynucleotide sequence) to a second immobilised primer. Each first polynucleotide sequence may comprise a second adaptor sequence, wherein the second adaptor sequence comprises a portion which is substantially complementary to the second immobilised primer (or is substantially complementary to the second immobilised primer). The second adaptor sequence may be at a 3'-end of the first polynucleotide sequence. Each second polynucleotide sequence may comprise a first adaptor sequence, wherein the first adaptor sequence comprises a portion which is substantially complementary to the first immobilised primer (or is substantially complementary to the first immobilised primer). The first adaptor sequence may be at a 3'-end of the second polynucleotide sequence.

In an embodiment, a solution comprising a polynucleotide library prepared by a loop fork method as described above may be flowed across a flowcell.

A particular polynucleotide strand from the polynucleotide library to be sequenced comprising, in a 5' to 3' direction, a second primer-binding complement sequence 302 (e.g. P7), an optional first terminal sequencing primer binding site complement 303', a first insert sequence 401 (A and B), a loop sequence 403 (L) (or hairpin loop adaptor), a second insert sequence 402 (B' and A'), an optional second terminal sequencing primer binding site 304, and a first primer-binding sequence 301' (e.g. P5'), may anneal (via the first primer-binding sequence 301') to the first immobilised primer 201 (e.g. P5 lawn primer) located within a particular well 203 (Figure 6A).

The polynucleotide library may comprise other polynucleotide strands with different first insert sequences 401 and second insert sequences 402. Such other polynucleotide strands may anneal to

corresponding first immobilised primers 201 (e.g. P5 lawn primers) in different wells 203, thus enabling parallel processing of the various different strands within the polynucleotide library.

A new polynucleotide strand may then be synthesised, extending from the first immobilised primer 201 (e.g. P5 lawn primer) in a direction away from the substrate 204. By using complementary base-pairing, this generates a template strand comprising, in a 5' to 3' direction, the first immobilised primer 201 (e.g. P5 lawn primer) which is attached to the solid support 200, an optional second terminal sequencing primer binding site complement 304', a second insert complement sequence 402' (A' copy and B' copy), a loop complement sequence 403' (L') (also referred to herein as a spacer strand, that is complementary to the hairpin loop adaptor), a first insert complement sequence 401' (B copy and A copy), an optional first terminal sequencing primer binding site 303, and a second primer-binding sequence 302' (e.g. P7') (Figure 6B). Such a process may utilise a polymerase, such as a DNA or RNA polymerase.

If the polynucleotides in the library comprise index sequences, then corresponding index sequences are also produced in the template.

The polynucleotide strand from the polynucleotide library may then be dehybridised and washed away, leaving a template strand attached to the first immobilised primer 201 (e.g. P5 lawn primer) (Figure 6C).

The second primer-binding sequence 302' (e.g. P7') on the template strand may then anneal to a second immobilised primer 202 (e.g. P7 lawn primer) located within the well 203. This forms a "bridge".

A new polynucleotide strand may then be synthesised by bridge amplification, extending from the second immobilised primer 202 (e.g. P7 lawn primer) (initially) in a direction away from the substrate 204. By using complementary base-pairing, this generates another template strand (also referred herein as a template complement strand) comprising, in a 5' to 3' direction, the second immobilised primer 202 (e.g. P7 lawn primer) which is attached to the solid support 200, an optional first terminal sequencing primer binding site complement 303', a first insert sequence 401 (A and B), a loop sequence 403 (L) (also referred to herein as a spacer complement strand), a second insert sequence 402 (B' and A'), an optional second terminal sequencing primer binding site 304, and a first primer-binding sequence 301' (e.g. P5'). Again, such a process may utilise a polymerase, such as a DNA or RNA polymerase.

The strand attached to the second immobilised primer 202 (e.g. P7 lawn primer) may then be dehybridised from the strand attached to the first immobilised primer 201 (e.g. P5 lawn primer) (Figure 6D).

A subsequent bridge amplification cycle can then lead to amplification of the strand attached to the first immobilised primer 201 (e.g. P5 lawn primer) and the strand attached to the second immobilised primer 202 (e.g. P7 lawn primer). The second primer-binding sequence 302' (e.g. P7') on the template strand attached to the first immobilised primer 201 (e.g. P5 lawn primer) may then anneal to another second immobilised primer 202 (e.g. P7 lawn primer) located within the well 203. In a similar fashion, the first primer-binding sequence 301' (e.g. P5') on the template strand attached to the second immobilised primer 202 (e.g. P7 lawn primer) may then anneal to another first immobilised primer 201 (e.g. P5 lawn primer) located within the well 203.

Completion of bridge amplification and dehybridisation may then provide an amplified cluster, thus providing a plurality of polynucleotide sequences comprising a first insert complement sequence 401' and a second insert complement sequence 402', as well as a plurality of polynucleotide sequences comprising a first insert sequence 401 and a second insert sequence 402 (Figure 6E).

If desired, further bridge amplification cycles may be conducted to increase the number of polynucleotide sequences within the well 203.

Once again, although Figure 6 shows the presence of a first terminal sequencing primer binding site complement 303', a second terminal sequencing primer binding site 304, a second terminal sequencing primer binding site complement 304', and a first terminal sequencing primer binding site 303, these are optional as mentioned above. Accordingly, these sections may be omitted from the template and template complement strands.

The methods for clustering and amplification described above generally relate to conducting non-selective amplification. However, methods of the present invention relating to selective processing may comprise conducting selective amplification, which is described in further detail below under selective processing.

Sequencing

As described herein, the template provides information (e.g. identification of the genetic sequence, identification of epigenetic modifications) on the original target polynucleotide

sequence. For example, a sequencing process (e.g. a sequencing-by-synthesis or sequencing-by-ligation process) may reproduce information that was present in the original target polynucleotide sequence, by using complementary base pairing.

In one embodiment, sequencing may be carried out using any suitable "sequencing-by-synthesis" technique, wherein nucleotides are added successively in cycles to the free 3' hydroxyl group, resulting in synthesis of a polynucleotide chain in the 5' to 3' direction. The nature of the nucleotide added may be determined after each addition. One particular sequencing method relies on the use of modified nucleotides that can act as reversible chain terminators. Such reversible chain terminators comprise removable 3' blocking groups. Once such a modified nucleotide has been incorporated into the growing polynucleotide chain complementary to the region of the template being sequenced there is no free 3'-OH group available to direct further sequence extension and therefore the polymerase cannot add further nucleotides. Once the nature of the base incorporated into the growing chain has been determined, the 3' block may be removed to allow addition of the next successive nucleotide. By ordering the products derived using these modified nucleotides it is possible to deduce the DNA sequence of the DNA template. Such reactions can be done in a single experiment if each of the modified nucleotides has attached thereto a different label, known to correspond to the particular base, to facilitate discrimination between the bases added at each incorporation step. Suitable labels are described in PCT application PCT/GB2007/001770, the contents of which are incorporated herein by reference in their entirety. Alternatively, a separate reaction may be carried out containing each of the modified nucleotides added individually.

The modified nucleotides may carry a label to facilitate their detection. Such a label may be configured to emit a signal, such as an electromagnetic signal, or a (visible) light signal.

In a particular embodiment, the label is a fluorescent label (e.g. a dye). Thus, such a label may be configured to emit an electromagnetic signal, or a (visible) light signal. One method for detecting the fluorescently labelled nucleotides comprises using laser light of a wavelength specific for the labelled nucleotides, or the use of other suitable sources of illumination. The fluorescence from the label on an incorporated nucleotide may be detected by a CCD camera or other suitable detection means. Suitable detection means are described in PCT/US2007/007991, the contents of which are incorporated herein by reference in their entirety.

However, the detectable label need not be a fluorescent label. Any label can be used which allows the detection of the incorporation of the nucleotide into the DNA sequence.

Each cycle may involve simultaneous delivery of four different nucleotide types to the array of template molecules. Alternatively, different nucleotide types can be added sequentially and an image of the array of template molecules can be obtained between each addition step.

In some embodiments, each nucleotide type may have a (spectrally) distinct label. In other words, four channels may be used to detect four nucleobases (also known as 4-channel chemistry) (Figure 7 – left). For example, a first nucleotide type (e.g. A) may include a first label (e.g. configured to emit a first wavelength, such as red light), a second nucleotide type (e.g. G) may include a second label (e.g. configured to emit a second wavelength, such as blue light), a third nucleotide type (e.g. T) may include a third label (e.g. configured to emit a third wavelength, such as green light), and a fourth nucleotide type (e.g. C) may include a fourth label (e.g. configured to emit a fourth wavelength, such as yellow light). Four images can then be obtained, each using a detection channel that is selective for one of the four different labels. For example, the first nucleotide type (e.g. A) may be detected in a first channel (e.g. configured to detect the first wavelength, such as red light), the second nucleotide type (e.g. G) may be detected in a second channel (e.g. configured to detect the second wavelength, such as blue light), the third nucleotide type (e.g. T) may be detected in a third channel (e.g. configured to detect the third wavelength, such as green light), and the fourth nucleotide type (e.g. C) may be detected in a fourth channel (e.g. configured to detect the fourth wavelength, such as yellow light). Although specific pairings of bases to signal types (e.g. wavelengths) are described above, different signal types (e.g. wavelengths) and/or permutations may also be used.

In some embodiments, detection of each nucleotide type may be conducted using fewer than four different labels. For example, sequencing-by-synthesis may be performed using methods and systems described in US 2013/0079232, which is incorporated herein by reference.

Thus, in some embodiments, two channels may be used to detect four nucleobases (also known as 2-channel chemistry) (Figure 7 – middle). For example, a first nucleotide type (e.g. A) may include a first label (e.g. configured to emit a first wavelength, such as green light) and a second label (e.g. configured to emit a second wavelength, such as red light), a second nucleotide type (e.g. G) may not include the first label and may not include the second label, a third nucleotide type (e.g. T) may include the first label (e.g. configured to emit the first wavelength, such as green light) and may not include the second label, and a fourth nucleotide type (e.g. C) may not include the first label and may include the second label (e.g. configured to emit the second wavelength, such as red light). Two images can then be obtained, using detection channels for the first label and the second label. For example, the first nucleotide type (e.g. A) may be detected in both a first channel (e.g. configured to detect the first wavelength, such as red light) and a second channel

(e.g. configured to detect the second wavelength, such as green light), the second nucleotide type (e.g. G) may not be detected in the first channel and may not be detected in the second channel, the third nucleotide type (e.g. T) may be detected in the first channel (e.g. configured to detect the first wavelength, such as red light) and may not be detected in the second channel, and the fourth nucleotide type (e.g. C) may not be detected in the first channel and may be detected in the second channel (e.g. configured to detect the second wavelength, such as green light). Although specific pairings of bases to signal types (e.g. wavelengths) and/or combinations of channels are described above, different signal types (e.g. wavelengths) and/or permutations may also be used.

In some embodiments, one channel may be used to detect four nucleobases (also known as 1-channel chemistry) (Figure 7 – right). For example, a first nucleotide type (e.g. A) may include a cleavable label (e.g. configured to emit a wavelength, such as green light), a second nucleotide type (e.g. G) may not include a label, a third nucleotide type (e.g. T) may include a non-cleavable label (e.g. configured to emit the wavelength, such as green light), and a fourth nucleotide type (e.g. C) may include a label-accepting site which does not include the label. A first image can then be obtained, and a subsequent treatment carried out to cleave the label attached to the first nucleotide type, and to attach the label to the label-accepting site on the fourth nucleotide type. A second image may then be obtained. For example, the first nucleotide type (e.g. A) may be detected in a channel (e.g. configured to detect the wavelength, such as green light) in the first image and not detected in the channel in the second image, the second nucleotide type (e.g. G) may not be detected in the channel in the first image and may not be detected in the channel in the second image, the third nucleotide type (e.g. T) may be detected in the channel (e.g. configured to detect the wavelength, such as green light) in the first image and may be detected in the channel (e.g. configured to detect the wavelength, such as green light) in the second image, and the fourth nucleotide type (e.g. C) may not be detected in the channel in the first image and may be detected in the channel in the second image (e.g. configured to detect the wavelength, such as green light). Although specific pairings of bases to signal types (e.g. wavelengths) and/or combinations of images are described above, different signal types (e.g. wavelengths), images and/or permutations may also be used.

In one embodiment, the sequencing process comprises a first sequencing read and second sequencing read. In some embodiments, the first sequencing read is conducted separately from the second sequencing read. In other embodiments, the first sequencing read and the second sequencing read may be conducted concurrently. In other words, the first sequencing read and the second sequencing read may be conducted at the same time.

The first sequencing read may comprise the binding of a first sequencing primer (also known as a read 1.1 sequencing primer) to the first sequencing primer binding site (e.g. within loop complement sequence 403'). The second sequencing read may comprise the binding of a second sequencing primer (also known as a read 1.2 sequencing primer) to the second sequencing primer binding site (e.g. within loop sequence 403).

This leads to sequencing of the first portion (e.g. second insert complement sequence 402') and the second portion (e.g. first insert sequence 401).

The methods for sequencing described above generally relate to conducting non-selective sequencing. However, methods of the present invention relating to selective processing may comprise conducting selective sequencing, which is described in further detail below under selective processing.

In particular, where concurrent sequencing is conducted, the signals generated may be spatially resolved or spatially unresolved. In the case where the signals generated are spatially resolved, the signals generated by the first portion and the second portion may be parsed by interpreting these signals separately in view of the spatial separation, and non-selective processing methods (such as non-selective amplification and non-selective sequencing) may be used. However, where the signals generated by the first portion and the second portion are spatially unresolved, other methods may be required to parse the information generated – as such, spatially unresolved signals may involve selective processing methods (such as selective amplification and/or selective sequencing).

Selective processing methods

In some embodiments, selective processing methods may be used to generate signals of different intensities. Accordingly, in some embodiments, the method may comprise selectively processing the at least one first polynucleotide sequence comprising a first portion and the at least one second polynucleotide sequence comprising a second portion, such that a proportion of first portions are capable of generating a first signal and a proportion of second portions are capable of generating a second signal, wherein the selective processing causes an intensity of the first signal to be greater than an intensity of the second signal.

The method may comprise selectively processing a plurality of first polynucleotide sequences each comprising a first portion and a plurality of second polynucleotide sequences each comprising a second portion, such that a proportion of first portions are capable of generating a

first signal and a proportion of second portions are capable of generating a second signal, wherein the selective processing causes an intensity of the first signal to be greater than an intensity of the second signal.

By "selective processing" is meant here performing an action that changes relative properties of the first portion and the second portion in the at least one first polynucleotide sequence comprising a first portion and at least one second polynucleotide sequence comprising a second portion (or the plurality of first polynucleotide sequences each comprising a first portion and the plurality of second polynucleotide sequences each comprising a second portion), so that the intensity of the first signal is greater than the intensity of the second signal. The property may be, for example, a concentration of first portions capable of generating the first signal relative to a concentration of second portions capable of generating the second signal. The action may include, for example, conducting selective amplification, conducting selective sequencing, or preparing for selective sequencing.

In one embodiment, the selective processing results in the concentration of the first portions capable of generating the first signal being greater than the concentration of the second portions capable of generating the second signal. In other words, the method of the invention results in an altered ratio of R1:R2 molecules, such as within a single cluster or a single well.

In one embodiment, the ratio may be between 1.25:1 to 5:1. In a further embodiment, the ratio may be between 1.5:1 to 3:1. In an even further embodiment, the ratio may be about 2:1.

Selective processing may refer to conducting selective sequencing. Alternatively, selective processing may refer to preparing for selective sequencing. As shown in Figure 8, in one example, selective sequencing may be achieved using a mixture of unblocked and blocked sequencing primers.

Where the method of the invention involves (separate) polynucleotide strands, with a first polynucleotide strand with a first portion, and a second polynucleotide strand with a second portion, the first polynucleotide strand may comprise a first sequencing primer binding site, and the second polynucleotide strand may comprise a second sequencing primer binding site, where the first sequencing primer binding site and second sequencing primer binding site are of a different sequence to each other and bind different sequencing primers.

In one embodiment, binding of first sequencing primers to the first sequencing primer site generates a first signal and binding of second sequencing primers to the second sequencing primer

site generates a second signal, where the intensity of the first signal is greater than the intensity of the second signal. This may be applied to embodiments where the first polynucleotide strand comprises a first sequencing primer binding site, and the second polynucleotide strand comprises a second sequencing primer binding site. This is achieved using a mixed population of blocked and unblocked second sequencing primers that bind the second sequencing primer site. Any ratio of blocked:unblocked second primers can be used that generates a second signal that is of a lower intensity than the first signal, for example, the ratio of blocked:unblocked primers may be 20:80 to 80:20. In a further embodiment, the ratio may be 1:2 to 2:1.

In an even further embodiment, a ratio of 50:50 of blocked:unblocked second primers is used, which in turn generates a second signal that is around 50% of the intensity of the first signal.

The first and second sequencing primers may be added to the flow cell at the same time, or separately but sequentially.

By "blocked" is meant that the sequencing primer comprises a blocking group at a 3' end of the sequencing primer. Suitable blocking groups include a hairpin loop (e.g. a polynucleotide attached to the 3'-end, comprising in a 5' to 3' direction, a cleavable site such as a nucleotide comprising uracil, a loop portion, and a complement portion, wherein the complement portion is substantially complementary to all or a portion of the immobilised primer), a deoxynucleotide, a deoxyribonucleotide, a hydrogen atom instead of a 3'-OH group, a phosphate group, a phosphorothioate group, a propyl spacer (e.g. -O-$(CH_2)_3$-OH instead of a 3'-OH group), a modification blocking the 3'-hydroxyl group (e.g. hydroxyl protecting groups, such as silyl ether groups (e.g. trimethylsilyl, triethylsilyl, triisopropylsilyl, t-butyl(dimethyl)silyl, t-butyl(diphenyl)silyl), ether groups (e.g. benzyl, allyl, t-butyl, methoxymethyl (MOM), 2-methoxyethoxymethyl (MEM), tetrahydropyranyl), or acyl groups (e.g. acetyl, benzoyl)), or an inverted nucleobase. However, the blocking group may be any modification that prevents extension (i.e. elongation) of the primer by a polymerase.

The sequence of the sequencing primers and the sequence primer binding sites are not material to the methods of the invention, as long as the sequencing primers are able to bind to the sequence primer binding site to enable amplification and sequencing of the regions to be identified.

In one aspect, the unblocked and blocked second sequencing primers are present in the sequencing composition in equal concentrations. That is, the ratio of blocked:unblocked second sequencing primers is around 50:50. The sequencing composition may further comprise at least one additional (first) sequencing primer. In one example, the sequencing composition comprises blocked second

sequencing primers, unblocked second sequencing primers and at least one first sequencing primer.

As shown in Figure 8, selective sequencing may be conducted on the amplified (duoclonal) cluster shown in Figure 6E, after restriction sites in the loop complement sequence 403' and the loop sequence 403 are cleaved by an endonuclease, as described in further detail below. A plurality of first sequencing primers 501 are added. These sequencing primers 501 anneal to a sequencing primer binding site present in the loop complement sequence 403'. A plurality of second unblocked sequencing primers 502a and a plurality of second blocked sequencing primers 502b are added, either at the same time as the first sequencing primers 501, or sequentially (e.g. prior to or after addition of first sequencing primers 501). These second unblocked sequencing primers 502a and second blocked sequencing primers 502b anneal to a sequencing primer binding site present in the loop sequence 403. This then allows the second insert complement sequences 402' (i.e. "first portions") to be sequenced and the first insert sequences 401 (i.e. "second portions") to be sequenced, wherein a greater proportion of second insert complement sequences 402' are sequenced (black arrow) compared to a proportion of first insert sequences 401 (grey arrow).

In other embodiments, the positioning of first sequencing primers and second sequencing primers may be swapped. In other words, the first sequencing binding primers may anneal instead to the loop sequence 403, and the second sequencing binding primers may anneal instead to the loop complement sequence 403'.

Alternatively, or in addition, selective processing may refer to selective amplification. That is, selectively amplifying one portion (e.g. the first or second portion) on a first or second polynucleotide strand.

In one example, selective processing comprises selectively removing some or substantially all of second immobilised primers that have not yet been extended (extended to form a second polynucleotide strand), and conducting at least one further amplification cycle in order to selectively amplify the first polynucleotide sequence(s) relative to the second polynucleotide sequence(s). Immobilised primers that have not yet been extended may be referred to herein as free or un-extended second immobilised primers.

Accordingly, in this example, selective removal of some or substantially all free second immobilised primers is carried out before at least one further round of bridge amplification and before any sequencing of the target regions. As a consequence, the ratio of first polynucleotide capable of generating a first signal to the second polynucleotide that is capable of generating a

second signal is altered, which in turn leads to two signals of different intensities, permitting concurrent sequencing of both sequences (or the target regions within those sequences).

By "some or substantially all" is meant that at least 75%, at least 80%, at least 90% or between 95% and 100% of free second immobilised primers are removed.

The selective removal of all or substantially all free second immobilised primers may be carried out using a reagent capable of cleaving the immobilised primer from the solid support. This reagent may be added following at least 5, at least 10, at least 15 or following at least 20 to 24 rounds of bridge amplification. The reagent may be added separately or together with the amplification reagents for performing the at least one further round of amplification.

As described above, and described in further detail in WO 2008/041002, the first and second immobilised primers may be attached to the surface of a solid support though a linker. The linker may be different for the first and second immobilised primers. The linker may be any cleavable linker; that is the linker may comprise one or more moieties, such as modified nucleotides, that enable selective cleavage of the immobilised primer from the surface of the solid support. By way of non-limiting example, the linker may comprise uracil bases, phosphorothioate groups, ribonucleotides, diol linkages, disulphide linkages, peptides etc. which may be included, not only to allow covalent attachment to a solid support, but also to allow selective cleavage of the linker.

In one example, the first immobilised primer is attached to a solid support through a first linker, where the linker comprises uracil, or 2-deoxyuridine. In this example, free first immobilised primers (that is, primers that are not extended) can be removed using uracil glycosylase. In one embodiment, free first immobilised primers can be removed using a USER enzyme mix (which is a cocktail of uracil glycosylase and endonuclease VIII).

In one example, the sequence of the first immobilised primer comprises the following sequence or a variant of fragment thereof:

5'-PS-TTTTTTTTTTAATGATACGGCGACCACCGAUCTACAC-3' where U = 2-deoxyuridine (SEQ ID NO. 53).

In another example, the second immobilised primer is attached to a solid support through a second linker, where the linker comprises 8-oxoguanine. In this example, free second immobilised primers (that is, primers that are not extended) can be removed using a FPG glycosylase.

In one example, the sequence of the second immobilised primer comprises the following sequence or a variant of fragment thereof:

5'-PS-TTTTTTTTTTCAAGCAGAAGACGGCATACGA[G^{oxo}]AT-3', where [G^{oxo}] = 8-oxoguanine (SEQ ID NO. 54).

One example of this method is shown in Figure 9. Selective amplification may be conducted on the amplified (duoclonal) cluster as shown in Figure 6E. The solid support 200 comprises free first immobilised primers 201 and free second immobilised primers 202 (Figure 9A). For simplicity, strand 1001' represents second insert complement sequence 402', loop complement sequence 403' and first insert complement sequence 401', whilst strand 1001 represents first insert sequence 401, loop sequence 403 and second insert sequence 402. Free second immobilised primers 202 are cleaved from the solid support 200, thus leaving behind free first immobilised primers 201 (Figure 9B).

The first primer-binding sequence 301' (e.g. P5') on one set of template strands may then anneal to the free first immobilised primers 201 (e.g. P5 lawn primer) located within the well 203. By contrast, since free second immobilised primers 202 (e.g. P7 lawn primer) have been removed, second primer-binding sequences 302' (e.g. P7') are not able to anneal (Figure 9C).

After conducting a cycle of bridge amplification, this leads to selective amplification of the strand 1001', relative to the strand 1001 (Figure 9D).

Conducting standard (non-selective) sequencing then allows strands 1001' and strands 1001 to be sequenced, wherein a greater proportion of strands 1001' are sequenced (grey arrow) compared to a proportion of strands 1001 (black arrow) (Figure 9E).

In another example, selectively processing comprises selectively blocking the extension of some or substantially all of the second immobilised primers that have not yet been extended (extended to form a second polynucleotide strand). Again, these primers may be referred to herein as free or un-extended second immobilised primers. The method may involve using a primer-blocking agent, wherein the primer-blocking agent is configured to limit or prevent synthesis of a strand (i.e. a polynucleotide strand) extending from the second immobilised primer. The method may further involve conducting at least one further amplification cycle. As the free second immobilised primers are blocked from being extended by the primer-blocking agent, only the first immobilised primers can be extended. This leads to amplification of only the first polynucleotide

strand (i.e. not the second polynucleotide strand), and as a consequence, an increase in the amount of first polynucleotide sequences relative to the second polynucleotide sequences.

By "some or substantially all" is meant that at least 75%, at least 80%, at least 90% or between 95% and 100% of free second immobilised primers are blocked.

The primer-blocking agent may be flowed across the solid support following bridge amplification. In one embodiment, the primer-blocking agent is flowed across the solid support following at least 1, 2, 3, 4, 5, 6, 7, 8, 9 or 10 cycles, following at least 15, following at least 20 or following at least 25 rounds of bridge amplification.

In one example, the primer-blocking agent is added whilst first polynucleotide sequence(s) are hybridised to the second immobilised primers. That is, the primer-blocking agent is added during amplification and following extension of at least the first polynucleotide strand. At this stage the extended first polynucleotide strand bends (bridges) and hybridises at its 5' end to the second immobilised primer. Addition of the primer-blocking agent at this stage prevents extension of the second immobilised primer, which would normally occur using the first polynucleotide strand as its template.

In one embodiment, the primer-blocking agent is a blocked nucleotide. In one example, the blocked nucleotide may be A, C, T or G, but may be selected from A or G.

Again, by "blocked" is meant that the sequencing primer comprises a blocking group at a 3' end of the sequencing primer. Suitable blocking groups include a hairpin loop (e.g. a polynucleotide attached to the 3'-end, comprising in a 5' to 3' direction, a cleavable site such as a nucleotide comprising uracil, a loop portion, and a complement portion, wherein the complement portion is substantially complementary to all or a portion of the immobilised primer), a deoxynucleotide, a deoxyribonucleotide, a hydrogen atom instead of a 3'-OH group, a phosphate group, a phosphorothioate group, a propyl spacer (e.g. -O-$(CH_2)_3$-OH instead of a 3'-OH group)), a modification blocking the 3'-hydroxyl group (e.g. hydroxyl protecting groups, such as silyl ether groups (e.g. trimethylsilyl, triethylsilyl, triisopropylsilyl, t-butyl(dimethyl)silyl, t-butyl(diphenyl)silyl), ether groups (e.g. benzyl, allyl, t-butyl, methoxymethyl (MOM), 2-methoxyethoxymethyl (MEM), tetrahydropyranyl), or acyl groups (e.g. acetyl, benzoyl)), or an inverted nucleobase. However, the blocking group may be any modification that prevents extension (i.e. elongation) of the primer by a polymerase. The block may be reversible or irreversible.

The blocked nucleotide may be added as part of a mixture comprising both blocked and unblocked nucleotides. Alternatively, the blocked nucleotide may be added to the flow cell separately and either before or after unblocked nucleotides are added. Following addition of the blocked nucleotide, at least one more round of bridge amplification is performed.

One example of this method is shown in Figure 10. Selective amplification may be conducted on the amplified (duoclonal) cluster as shown in Figure 9A. The first primer-binding sequence 301' (e.g. P5') on one set of template strands may anneal to first immobilised primers 201 (e.g. P5 lawn primer), and the second primer-binding sequence 302' (e.g. P7') on another set of template strands may anneal to second immobilised primers 202 (e.g. P7 lawn primer) (Figure 10A).

Whilst the second primer-binding sequence 302' (e.g. P7') is annealed to the second immobilised primer 202, a primer-blocking agent 601 is selectively installed onto a 3'-end of the second immobilised primer 202, whilst no installation occurs to the 3'-end of the first immobilised primer 201 (Figure 10B).

After conducting a cycle of bridge amplification, this leads to selective amplification of the strands 1001', relative to the strands 1001. The primer-blocking agent 601 prevents extension from the second immobilised primer 202 (Figure 10C).

Conducting standard (non-selective) sequencing then allows strands 1001' and strands 1001 to be sequenced, wherein a greater proportion of strands 1001' are sequenced (grey arrow) compared to a proportion of strands 1001 (black arrow) (Figure 10D).

In an alternative example, the method comprises flowing at least one, or a plurality of, extended primer sequence(s) across the surface of the solid support (e.g. a flow cell), wherein such sequences can bind (e.g. hybridise) free immobilised primers (e.g. P5 or P7) and wherein the extended primer sequences further comprise at least one 5' additional nucleotide; and (b) adding the primer blocking agent, where the primer blocking agent is complementary to the 5' additional nucleotide.

In one embodiment, the extended primer sequences are substantially complementary to the first or second immobilised primers (e.g. P5 or P7), or substantially complementary to a portion of the first or second immobilised primer.

The 5' additional nucleotide may be selected from A, T, C or G, but may be T (or U) or C. In one embodiment, the 5' additional nucleotide is not a complement of the 3' nucleotide of the second

immobilised primer (where the extended primer sequence binds the first immobilised primer) or is not a complement of the 3' nucleotide of the first immobilised primer (where the extended primer sequence binds the second immobilised primer). For example, where the first immobilised primer is P5 (for example as defined in SEQ ID NO. 1) and the second immobilised primer is P7 for example as defined in SEQ ID NO. 2), and where the extended primer sequence binds the first immobilised primer, the 5' additional nucleotide is not A. Similarly, where the extended primer sequence binds the second immobilised primer, the 5' additional nucleotide is not G.

In one embodiment, the primer-blocking agent is a blocked nucleotide, for example, as described above. In one embodiment, the blocked nucleotide may be A, C, T or G, but may be selected from A or G. Accordingly, where the 5' additional nucleotide is T or U, the primer-blocking agent is A, and where the 5' additional nucleotide is C, the primer-blocking agent is G.

Again, the extended primer sequence(s) and primer-blocking agent may be flowed across the solid support following bridge amplification. In one embodiment, the primer-blocking agent is flowed across the solid support following at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 15, at least 20 or following at least 25 rounds of bridge amplification.

In one embodiment, the extended primer sequence is selected from SEQ ID NO. 55 to 66 or a variant or fragment thereof.

One example of this method is shown in Figure 11. Selective amplification may be conducted on the amplified (duoclonal) cluster as shown in Figure 9A; as such following a number of rounds of amplification, a cluster is formed comprising both extended first (e.g. P5) and second (e.g. P7) immobilised polynucleotide strands. Before the next round of amplification, a (or a plurality of) extended primer sequence(s) is flowed across the surface of the solid support 200. The extended primer sequence 701 is substantially complementary to at least a portion, if not all of the immobilised primer (e.g. either P5 or P7) and binds to the immobilised primer (e.g. P5 or P7) as shown in Figure 11A. As also shown in Figure 11A, the extended primer sequence 701 comprises at least one additional 5' nucleotide.

Following addition of the extended primer sequence 701, a primer blocking agent 601 is added and flowed across the surface of the solid support (e.g. flow cell). As the primer-blocking agent 601 is complementary to the 5' additional nucleotide of the extended primer sequence 701 the primer-blocking agent 601 binds to the 3'-end of the immobilised strands that are hybridised to the extended primer sequence 701, as shown in Figure 16B. As a consequence, addition of the

primer-blocking agent 601 prevents not only extension of the immobilised strand (e.g. P5 or P7) but renders the immobilised primer (P5 or P7) unavailable for hybridisation and subsequent bridge amplification for other extended strands (e.g. 101') (see Figure 11B).

Performing at least one more cycle of bridge amplification, leads to selective amplification of strands 1001' (in a 2:1 ratio of 1001' to 1001). Again, similar to Figure 10D, conducting standard (non-selective) sequencing then allows strands 1001' and strands 1001 to be sequenced, wherein a greater proportion of strands 1001' are sequenced (grey arrow) compared to a proportion of strands 1001 (black arrow) (Figure 10D).

The extended primer sequences may be added as part of the amplification mixture described above. Alternatively, the blocked immobilised primer-binding sequence may be added to the flow cell separately and may be before the amplification mixture is added. Following addition of the blocked immobilised primer-binding sequence, at least one more round of bridge amplification is performed.

Data analysis using 16 QaM

Figure 12 is a scatter plot showing an example of sixteen distributions of signals generated by polynucleotide sequences disclosed herein.

The scatter plot of Figure 12 shows sixteen distributions (or bins) of intensity values from the combination of a brighter signal (i.e. a first signal as described herein) and a dimmer signal (i.e. a second signal as described herein); the two signals may be co-localized and may not be optically resolved as described above. The intensity values shown in Figure 12 may be up to a scale or normalisation factor; the units of the intensity values may be arbitrary or relative (i.e., representing the ratio of the actual intensity to a reference intensity). The sum of the brighter signal generated by the first portions and the dimmer signal generated by the second portions results in a combined signal. The combined signal may be captured by a first optical channel and a second optical channel. Since the brighter signal may be A, T, C or G, and the dimmer signal may be A, T, C or G, there are sixteen possibilities for the combined signal, corresponding to sixteen distinguishable patterns when optically captured. That is, each of the sixteen possibilities corresponds to a bin shown in Figure 12. The computer system can map the combined signal generated into one of the sixteen bins, and thus determine the added nucleobase at the first portion and the added nucleobase at the second portion, respectively.

For example, when the combined signal is mapped to bin 1612 for a base calling cycle, the computer processor base calls both the added nucleobase at the first portion and the added nucleobase at the second portion as C. When the combined signal is mapped to bin 1614 for the base calling cycle, the processor base calls the added nucleobase at the first portion as C and the added nucleobase at the second portion as T. When the combined signal is mapped to bin 1616 for the base calling cycle, the processor base calls the added nucleobase at the first portion as C and the added nucleobase at the second portion as G. When the combined signal is mapped to bin 1618 for the base calling cycle, the processor base calls the added nucleobase at the first portion as C and the added nucleobase at the second portion as A.

When the combined signal is mapped to bin 1622 for the base calling cycle, the processor base calls the added nucleobase at the first portion as T and the added nucleobase at the second portion as C. When the combined signal is mapped to bin 1624 for the base calling cycle, the processor base calls both the added nucleobase at the first portion and the added nucleobase at the second portion as T. When the combined signal is mapped to bin 1626 for the base calling cycle, the processor base calls the added nucleobase at the first portion as T and the added nucleobase at the second portion as G. When the combined signal is mapped to bin 1628 for the base calling cycle, the processor base calls the added nucleobase at the first portion as T and the added nucleobase at the second portion as A.

When the combined signal is mapped to bin 1632 for the base calling cycle, the processor base calls the added nucleobase at the first portion as G and the added nucleobase at the second portion as C. When the combined signal is mapped to bin 1634 for the base calling cycle, the processor base calls the added nucleobase at the first portion as G and the added nucleobase at the second portion as T. When the combined signal is mapped to bin 1636 for the base calling cycle, the processor base calls both the added nucleobase at the first portion and the added nucleobase at the second portion as G. When the combined signal is mapped to bin 1638 for the base calling cycle, the processor base calls the added nucleobase at the first portion as G and the added nucleobase at the second portion as A.

When the combined signal is mapped to bin 1642 for the base calling cycle, the processor base calls the added nucleobase at the first portion as A and the added nucleobase at the second portion as C. When the combined signal is mapped to bin 1644 for the base calling cycle, the processor base calls the added nucleobase at the first portion as A and the added nucleobase at the second portion as T. When the combined signal is mapped to bin 1646 for the base calling cycle, the processor base calls the added nucleobase at the first portion as A and the added nucleobase at the second portion as G. When the combined signal is mapped to bin 1648 for the base calling

cycle, the processor base calls both the added nucleobase at the first portion and the added nucleobase at the second portion as A.

In this particular example, T is configured to emit a signal in both the IMAGE 1 channel and the IMAGE 2 channel, A is configured to emit a signal in the IMAGE 1 channel only, C is configured to emit a signal in the IMAGE 2 channel only, and G does not emit a signal in either channel. However, different permutations of nucleobases can be used to achieve the same effect by performing dye swaps. For example, A may be configured to emit a signal in both the IMAGE 1 channel and the IMAGE 2 channel, T may be configured to emit a signal in the IMAGE 1 channel only, C may be configured to emit a signal in the IMAGE 2 channel only, and G may be configured to not emit a signal in either channel.

Further details regarding performing base-calling based on a scatter plot having sixteen bins may be found in U.S. Patent Application Publication No. 2019/0212294, the disclosure of which is incorporated herein by reference.

Figure 13 is a flow diagram showing a method 1700 of base calling according to the present disclosure. The described method allows for simultaneous sequencing of two (or more) portions (e.g. the first portion and the second portion) in a single sequencing run from a single combined signal obtained from the first portion and the second portion, thus requiring less sequencing reagent consumption and faster generation of data from both the first portion and the second portion. Further, the simplified method may reduce the number of workflow steps while producing the same yield as compared to existing next-generation sequencing methods. Thus, the simplified method may result in reduced sequencing runtime.

As shown in Figure 13, the disclosed method 1700 may start from block 1701. The method may then move to block 1710.

At block 1710, intensity data is obtained. The intensity data includes first intensity data and second intensity data. The first intensity data comprises a combined intensity of a first signal component obtained based upon a respective first nucleobase of the first portion and a second signal component obtained based upon a respective second nucleobase of the second portion. Similarly, the second intensity data comprises a combined intensity of a third signal component obtained based upon the respective first nucleobase of the first portion and a fourth signal component obtained based upon the respective second nucleobase of the second portion.

As such, the first portion is capable of generating a first signal comprising a first signal component and a third signal component. The second portion is capable of generating a second signal comprising a second signal component and a fourth signal component.

As described above, the first portion and the second portion may be arranged on the solid support such that signals from the first portion and the second portion are detected by a single sensing portion and/or may comprise a single cluster such that first signals and second signals from each of the respective first portions and second portions cannot be spatially resolved.

In one example, obtaining the intensity data comprises selecting intensity data that corresponds to two (or more) different portions (e.g. the first portion and the second portion). In one example, intensity data is selected based upon a chastity score. A chastity score may be calculated as the ratio of the brightest base intensity divided by the sum of the brightest and second brightest base intensities. The desired chastity score may be different depending upon the expected intensity ratio of the light emissions associated with the different portions. As described above, it may be desired to produce clusters comprising the first portion and the second portion, which give rise to signals in a ratio of 2:1. In one example, high-quality data corresponding to two portions with an intensity ratio of 2:1 may have a chastity score of around 0.8 to 0.9.

After the intensity data has been obtained, the method may proceed to block 1720. In this step, one of a plurality of classifications is selected based on the intensity data. Each classification represents a possible combination of respective first and second nucleobases. In one example, the plurality of classifications comprises sixteen classifications as shown in Figure 12, each representing a unique combination of first and second nucleobases. Where there are two portions, there are sixteen possible combinations of first and second nucleobases. Selecting the classification based on the first and second intensity data comprises selecting the classification based on the combined intensity of the first and second signal components and the combined intensity of the third and fourth signal components.

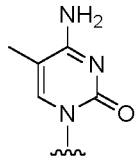
The method may then proceed to block 1730, where the respective first and second nucleobases are base called based on the classification selected in block 1720. The signals generated during a cycle of a sequencing are indicative of the identity of the nucleobase(s) added during sequencing (e.g. using sequencing-by-synthesis). It will be appreciated that there is a direct correspondence between the identity of the nucleobases that are incorporated and the identity of the complementary base at the corresponding position of the template sequence bound to the solid support. Therefore, any references herein to the base calling of respective nucleobases at the two portions encompasses the base calling of nucleobases hybridised to the template sequences and,

alternatively or additionally, the identification of the corresponding nucleobases of the template sequences. The method may then end at block 1740.

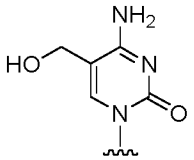
Sequencing of modified cytosines

The polynucleotide library hairpin strand (in particular, a fourth hairpin polynucleotide as described herein) may be exposed to a conversion agent configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase which is read as thymine/uracil, or to a conversion agent configured to convert an unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil (in particular, in order to generate a fifth hairpin polynucleotide as described herein).

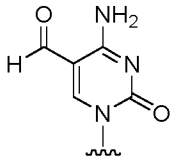
As used herein, the term "modified cytosine" may refer to any one or more of 5-methylcytosine (5-mC), 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine (5-fC) and 5-carboxylcytosine (5-caC):



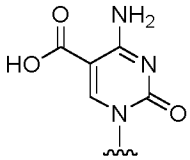
5-methylcytosine
(5-mC)



5-hydroxymethylcytosine
(5-hmC)



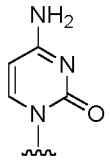
5-formylcytosine
(5-fC)



5-carboxylcytosine
(5-caC)

wherein the wavy line indicates an attachment point of the modified cytosine to the polynucleotide.

As used herein, the term "unmodified cytosine" refers to cytosine (C):



cytosine
(C)

wherein the wavy line indicates an attachment point of the unmodified cytosine to the polynucleotide.

As used herein, the term "conversion agent configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase which is read as thymine/uracil" may refer to a reagent which converts 5-methylcytosine and 5-hydroxymethylcytosine to thymine (i.e. would base pair with adenine), or to an equivalent nucleobase which would base pair with adenine. The

conversion may comprise a deamination reaction converting 5-methylcytosine or 5-hydroxymethylcytosine to thymine or nucleobase which is read as thymine/uracil.

As used herein, the term "conversion agent configured to convert an unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil, in order to generate a fifth hairpin polynucleotide" may refer to a reagent which converts one or more unmodified cytosines to uracil (i.e. would base pair with adenine), or to an equivalent nucleobase which would base pair with adenine. The conversion may comprise a deamination reaction converting the unmodified cytosine to uracil or nucleobase which is read as thymine/uracil.

In some embodiments, the conversion agent configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase which is read as thymine/uracil may further be configured to be selective for converting 5-methylcytosine and/or 5-hydroxymethylcytosine over converting unmodified cytosine. The selectivity may be measured by comparing reaction parameters (e.g. deamination reaction parameters) of the conversion of 5-methylcytosine and/or 5-hydroxymethylcytosine to thymine or equivalent nucleobase which is read as thymine/uracil, with corresponding reaction parameters (e.g. deamination reaction parameters) of the conversion of unmodified cytosine to uracil or nucleobase which is read as thymine/uracil. For example, reaction parameters such as rate of reaction or yield may be compared. In the case of rate of reaction, a rate of a reaction (e.g. deamination) of 5-methylcytosine and/or 5-hydroxymethylcytosine to thymine or nucleobase which is read as thymine/uracil may be greater (e.g. at least 2 times greater, at least 5 times greater, at least 10 times greater, at least 20 times greater, at least 50 times greater, or at least 100 times greater) than a corresponding rate of a reaction (e.g. deamination) of the unmodified cytosine to uracil or nucleobase which is read as thymine/uracil. In the case of yield, a yield of a reaction (e.g. deamination) of 5-methylcytosine and/or 5-hydroxymethylcytosine to thymine or nucleobase which is read as thymine/uracil may be greater (e.g. at least 2 times greater, at least 5 times greater, at least 10 times greater, at least 20 times greater, at least 50 times greater, or at least 100 times greater) than a corresponding yield of a reaction (e.g. deamination) of the unmodified cytosine to uracil or nucleobase which is read as thymine/uracil.

In some embodiments, the conversion agent configured to convert an unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil may further be configured to be selective for converting unmodified cytosine over converting 5-methylcytosine and/or 5-hydroxymethylcytosine. The selectivity may be measured by comparing reaction parameters (e.g. deamination reaction parameters) of the conversion of unmodified cytosine to uracil or nucleobase which is read as thymine/uracil, with corresponding reaction parameters (e.g.

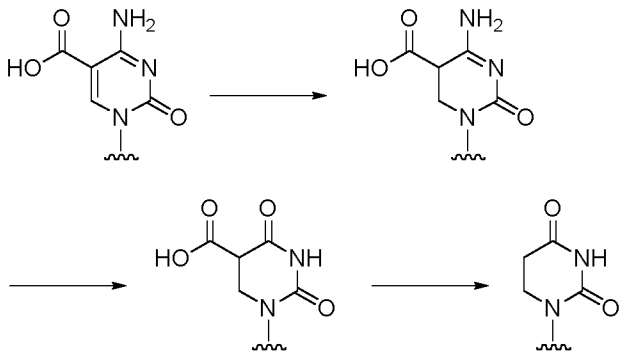
deamination reaction parameters) of the conversion of 5-methylcytosine and/or 5-hydroxymethylcytosine to thymine or nucleobase which is read as thymine/uracil. For example, reaction parameters such as rate of reaction or yield may be compared. In the case of rate of reaction, a rate of a reaction (e.g. deamination) of the unmodified cytosine to uracil or nucleobase which is read as thymine/uracil may be greater (e.g. at least 2 times greater, at least 5 times greater, at least 10 times greater, at least 20 times greater, at least 50 times greater, or at least 100 times greater) than a rate of a reaction (e.g. deamination) of 5-methylcytosine and/or 5-hydroxymethylcytosine to uracil or the nucleobase which is read as thymine/uracil. In the case of yield, a yield of a reaction (e.g. deamination) the unmodified cytosine to uracil or nucleobase which is read as thymine/uracil may be greater (e.g. at least 2 times greater, at least 5 times greater, at least 10 times greater, at least 20 times greater, at least 50 times greater, or at least 100 times greater) than a corresponding yield of a reaction (e.g. deamination) of 5-methylcytosine and/or 5-hydroxymethylcytosine to uracil or the nucleobase which is read as thymine/uracil.

In some embodiments, the conversion agent may comprise a chemical agent and/or an enzyme.

In some embodiments, the conversion agent may comprise a boron-based reducing agent and a ten-eleven translocation (TET) methylcytosine dioxygenase.

In a further embodiment, the boron-based reducing agent is an amine-borane compound or an azine-borane compound (wherein the term “azine” refers to a nitrogenous heterocyclic compound comprising a 6-membered aromatic ring). Non-limiting examples of amine-borane compounds include compounds such as t-butylamine borane, ammonia borane, ethylenediamine borane and dimethylamine borane. Non-limiting examples of azine-borane compounds include compounds such as pyridine borane and 2-picoline borane.

In general, boron-based reducing agents are able to convert 5-formylcytosine and 5-carboxylcytosine to dihydrouracil (i.e. a nucleobase which is read as thymine/uracil). The reaction proceeds by reduction of the internal C=C bond of 5-formylcytosine or 5-carboxylcytosine, deamination, and then decarboxylation to form dihydrouracil (illustrated below using 5-carboxylcytosine):



This process is selective for a particular type of modified cytosine (5-carboxylcytosine) and does not convert unmodified cytosine. When treatment of 5-methylcytosine and 5-hydroxymethylcytosine is required, treatment with additional reagents may be included to convert 5-methylcytosine and 5-hydroxymethylcytosine to 5-formylcytosine and/or 5-carboxylcytosine. In particular, boron-based reducing agents may be combined with ten-eleven translocation (TET) methylcytosine dioxygenases as described herein.

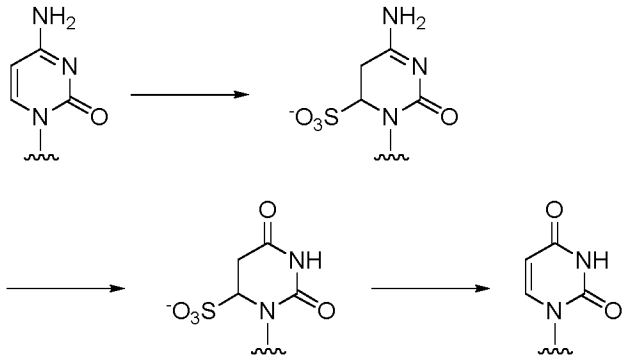
In some embodiments, the TET methylcytosine dioxygenase may be a member of the TET1 subfamily, the TET2 subfamily, or the TET3 subfamily. The enzyme may be configured to convert 5-methylcytosine to 5-hydroxymethylcytosine, 5-hydroxymethylcytosine to 5-formylcytosine, and 5-formylcytosine to 5-carboxylcytosine. Non-limiting examples of the TET methylcytosine dioxygenase include:

TET protein	**Non-limiting examples**
TET1	UniProt: Q8NFU7 (SEQ ID NO. 43) UniProt: Q3URK3 (SEQ ID NO. 44)
TET2	UniProt: Q6N021 (SEQ ID NO. 45) UniProt: Q4JK59 (SEQ ID NO. 46)
TET3	UniProt: O43151 (SEQ ID NO. 47) UniProt: Q8BG87 (SEQ ID NO. 48)

In some embodiments, the conversion agent (e.g. chemical agent) may comprise sulfite. The sulfite may be present in a partially acid/salt form (e.g. as bisulfite ions), or be present in a salt form (e.g. as sulfite ions). In cases where the sulfite is present in a salt form, the sulfite may comprise a cation (not including H^+). For example, the cation may be selected from "metal cations" or "non-metal cations". Metal cations may include alkali metal ions (e.g. lithium, sodium, potassium, rubidium or caesium ions). Non-metal cations may include ammonium salts (e.g. alkylammonium salts) or phosphonium salts (e.g. alkylphosphonium salts). The term "sulfite" also encompasses "metabisulfite", which dissolves in aqueous solution to form bisulfite. In a

further embodiment, the sulfite is bisulfite. In an even further embodiment, the bisulfite is sodium bisulfite.

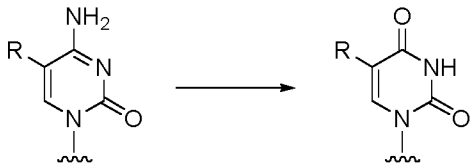
In general, sulfite (e.g. bisulfite) is able to convert unmodified cytosine to uracil. The reaction proceeds via conjugate addition of sulfite to the internal C=C of unmodified cytosine, deamination, and then elimination of sulfite to reform the internal C=C bond to form uracil:



This process is selective for unmodified cytosine over 5-methylcytosine and 5-hydroxymethylcytosine.

In one embodiment, the conversion agent (e.g. enzyme) may comprise a cytidine deaminase.

As used herein, the term "cytidine deaminase" may refer to an enzyme which is able to catalyse the following reaction:



wherein R is hydrogen, methyl, hydroxymethyl, formyl or carboxyl, and wherein the wavy line indicates an attachment point to a polynucleotide.

In one embodiment, the cytidine deaminase is a wild-type cytidine deaminase or a mutant cytidine deaminase. In a further embodiment, the cytidine deaminase is a mutant cytidine deaminase.

In some embodiments, the cytidine deaminase is a member of the APOBEC protein family. In a further embodiment, the cytidine deaminase is a member of the AID subfamily, the APOBEC1 subfamily, the APOBEC2 subfamily, the APOBEC3 subfamily (e.g. the APOBEC3A subfamily, the APOBEC3B subfamily, the APOBEC3C subfamily, the APOBEC3D subfamily, the APOBEC3F subfamily, the APOBEC3G subfamily, or the APOBEC3H subfamily), or the

APOBEC4 subfamily. In an even further embodiment, the cytidine deaminase is a member of the APOBEC3A subfamily.

In general, cytidine deaminases are able to catalyse the deamination of 5-methylcytosine and 5-hydroxymethylcytosine to their equivalent deaminated versions (i.e. nucleobases which are read as thymine/uracil), as well as catalysing the deamination of unmodified cytosines to uracil. Nevertheless, rates of reaction may differ depending on the type of modified cytosine; for example, wild-type APOBEC3A catalyses the deamination of unmodified cytosine and 5-methylcytosine relatively efficiently, whereas deamination of 5-hydroxymethylcytosine is ~5000-fold slower relative to unmodified cytosine. However, particular cytidine deaminases (e.g. mutant cytidine deaminases) may be chosen which have higher affinities for 5-methylcytosine and/or 5-hydroxymethylcytosine as substrates over unmodified cytosines, or vice versa.

The APOBEC protein family is a member of the large cytidine deaminase superfamily that contains a canonical zinc-dependent deaminase (ZDD) signature motif embedded within a core cytidine deaminase fold. This fold includes a five-stranded mixed beta (b)-sheet surrounded by six alpha (a)-helices with the order a1-b1-b2-a2-b3-a3-b4-a4-b5-a5-a6 (Salter et al., Trends Biochem Sci. 2016 41(7):578–594. doi:10.1016/j.tibs.2016.05.001; Salter et al., Trends Biochem. Sci. 2018, 43(8):606-622 doi.org/10.1016/j.tibs.2018.04.013). Each cytidine deaminase domain core structure of APOBEC proteins contains a highly conserved spatial arrangement of the catalytic centre residues of a zinc-binding motif H-[P/A/V]-E-$X_{[23-28]}$-P-C-$X_{[2-4]}$-C (SEQ ID NO. 51) (referred to herein as the ZDD motif, where X is any amino acid, and the subscript range of numbers after X refers to the number of amino acids) (Salter et al., Trends Biochem Sci. 2016 41(7):578–594. doi:10.1016/j.tibs.2016.05.001). Without intending to be limited by theory, the H and two C residues coordinate a Zn atom, and the E residue polarises a water molecule near the Zn-atom for catalysis (Chen et al., 2021, Viruses, 13:497, doi.org/10.3390/v13030497).

Some members of the APOBEC protein family, e.g., the AID subfamily, the APOBEC1 subfamily, the APOBEC2 subfamily, the APOBEC3A subfamily, the APOBEC3C subfamily, the APOBEC3H subfamily, and the APOBEC4 subfamily, include one copy of the ZDD motif. Other members of the APOBEC protein family, e.g., the APOBEC3B subfamily, the APOBEC3D subfamily, the APOBEC3F subfamily, and the APOBEC3G subfamily, include two copies of the ZDD motif, but often only the C-terminal copy is active (Salter et al., Trends Biochem Sci. 2016 41(7):578–594. doi:10.1016/j.tibs.2016.05.001). Thus, a mutant cytidine deaminase disclosed herein includes one or two ZDD motifs. In one embodiment, a mutant cytidine deaminase based on a member of the APOBEC3A subfamily includes the following ZDD motif: $HXEX_{24}SW(S/T)PCX_{[2-4]}CX_6FX_8LX_5R(L/I)YX_{[8-11]}LX_2LX_{[10]}M$ (SEQ ID NO. 52) (where X is

any amino acid, and the subscript number or range of numbers after X refers to the number of amino acids) (Salter et al., Trends Biochem Sci. 2016 41(7):578–594. doi:10.1016/j.tibs.2016.05.001).

Non-limiting examples of wild-type cytidine deaminases in the APOBEC protein family are shown in the table below (from UniProt, database of protein sequence and functional information, available at uniprot.org; or GenBank, collection of nucleotide sequences and their protein translations, available at ncbi.nlm.nih.gov/protein/):

APOBEC protein	**Non-limiting examples**
AID	UniProt: Q9GZX7 (SEQ ID NO. 7); UniProt: G3QLD2 (SEQ ID NO. 8); Uniprot Q9WVE0 (SEQ ID NO. 9)
APOBEC1	UniProt: P41238 (SEQ ID NO. 10); NCBI XP_030856728.1 (SEQ ID NO. 11); Uniprot P51908 (SEQ ID NO. 12)
APOBEC2	UniProt: Q9Y235 (SEQ ID NO. 13); Uniprot G3SGN8 (SEQ ID NO. 14); Uniprot Q9WV35 (SEQ ID NO. 15)
APOBEC3A	UniProt: P31941(SEQ ID NO. 16); GenBank: XP_045219544.1 (SEQ ID NO. 17) GenBank: AER45717.1 (SEQ ID NO. 18); GenBank: XP_003264816.1 (SEQ ID NO. 19); GenBank: PNI48846.1 (SEQ ID NO. 20); GenBank: ADO85886.1 (SEQ ID NO. 21)
APOBEC3B	UniProt: Q9UH17 (SEQ ID NO. 22); Uniprot G3QV16 (SEQ ID NO. 23); Uniprot F6M3K5 (SEQ ID NO. 24)
APOBEC3C	UniProt: Q9NRW3 (SEQ ID NO. 25); Uniprot Q694B5 (SEQ ID NO. 26); Uniprot B0LW74 (SEQ ID NO. 27)
APOBEC3D	UniProt: Q96AK3 (SEQ ID NO. 28); NCBI NP_001332895.1 (SEQ ID NO. 29); NCBI NP_001332931.1 (SEQ ID NO. 30)
APOBEC3F	UniProt: Q8IUX4 (SEQ ID NO. 31); Uniprot G3RD21 (SEQ ID NO. 32);

	Uniprot Q1G0Z6 (SEQ ID NO. 33)
APOBEC3G	UniProt: Q9HC16 (SEQ ID NO. 34); Uniprot Q694C1 (SEQ ID NO. 35); Uniprot U5NDB3 (SEQ ID NO. 36)
APOBEC3H	UniProt: Q6NTF7 (SEQ ID NO. 37); Uniprot B7T0U7 (SEQ ID NO. 38); Uniprot Q19Q52 (SEQ ID NO. 39)
APOBEC4	UniProt: Q8WW27(SEQ ID NO. 40); NCBI XP_004028087.1 (SEQ ID NO. 41); Uniprot Q497M3 (SEQ ID NO. 42)

In one embodiment, the mutant cytidine deaminase may comprise amino acid substitution mutations at positions functionally equivalent to (Tyr/Phe)130 and Tyr132 in a wild-type APOBEC3A protein. Such mutant cytidine deaminases are described in further detail in US Provisional Application 63/328,444, which is incorporated herein by reference. By "functionally equivalent" it is meant that the mutant cytidine deaminase has the amino acid substitution at the amino acid position in a reference (wild-type) cytidine deaminase that has the same functional role in both the reference (wild-type) cytidine deaminase and the mutant cytidine deaminase.

In one embodiment, the (Tyr/Phe)130 may be Tyr130, and the wild-type APOBEC3A protein may be SEQ ID NO. 16.

In some embodiments, the mutant cytidine deaminase may convert 5-methylcytosine to thymine by deamination at a greater rate than conversion rate of cytosine to uracil by deamination. In a further embodiment, the rate may be at least 100-fold greater.

In some embodiments, the mutant cytidine deaminase may comprise both a 5-methylcytosine specific deaminase and 5-hydroxymethylcytosine specific deaminase.

In one embodiment, the substitution mutation at the position functionally equivalent to Tyr130 may comprise Ala, Val or Trp.

In one embodiment, the substitution mutation at the position functionally equivalent to Tyr132 may comprise a mutation to His, Arg, Gln or Lys.

In one embodiment, the mutant cytidine deaminase may comprise a ZDD motif H-[P/A/V]-E-$X_{[23-28]}$-P-C-$X_{[2-4]}$-C (SEQ ID NO. 51).

In one embodiment, the mutant cytidine deaminase may be a member of the APOBEC3A subfamily and may comprise a ZDD motif $HXEX_{24}SW(S/T)PCX_{[2-4]}CX_6FX_8LX_5R(L/I)YX_{[8-11]}LX_2LX_{[10]}M$ (SEQ ID NO. 52).

In some embodiments, the conversion agent may further comprise a glycosyltransferase (e.g. α-glucosyltransferase or β-glucosyltransferase). In a further embodiment, the glycosyltransferase may be a β-glucosyltransferase. Such a glycosyltransferase may be configured to convert 5-hydroxymethylcytosine to a 5-hydroxymethylcytosine analogue bearing a hydroxyl protecting group, wherein the hydroxyl protecting group is glycosyl. A non-limiting example of the enzyme includes T4-βGT, for example as supplied by New England BioLabs (catalog # M0357S, M0357L) or by ThermoFisher Scientific (catalog # EO0831); further non-limiting examples of glycosyltransferases include:

Glucosyltransferase	**Non-limiting examples**
α-glucosyltransferase	UniProt: P04519 (SEQ ID NO. 49)
β-glucosyltransferase	UniProt: P04547 (SEQ ID NO. 50)

Specific methods of modified cytosine sequencing using conversion agents are further illustrated below. However, the type of conversion agents are not limited thereto.

BS-seq

Bisulfite sequencing (BS-seq) involves using bisulfite as the conversion agent. This process is described in Frommer et al. (Proc. Natl. Acad. Sci. U.S.A., 1992, 89, pp. 1827-1831), which is incorporated herein by reference. This process converts unmodified cytosines in the target polynucleotide to uracil to deaminated analogues, but does not convert 5-methylcytosine and 5-hydroxymethylcytosine. Accordingly, BS-seq allows identification of the modified cytosines 5-mC and 5-hmC by reading them as C; whereas unmodified C is converted to nucleobases which are read as T/U.

EM-seq

Enzymatic Methyl sequencing (EM-seq) involves using T4 bacteriophage β-glucosyltransferase and a TET2 enzyme as the further agents and APOBEC3A as the conversion agent. This process is described in Vaisvila et al. (Genome Res. 2021, 31, pp. 1280-1289), US 10,619,200 B2 and US 9,121,061 B2, which are incorporated herein by reference. The T4 bacteriophage β-glucosyltransferase converts 5-hydroxymethylcytosine in the target polynucleotide to β-glucosyl-

5-hydroxymethylcytosine, which prevents oxidation. The TET2 enzyme causes oxidation of 5-methylcytosine in the target polynucleotide to 5-hydroxymethylcytosine, which in turn is converted to β-glucosyl-5-hydroxymethylcytosine by the T4 bacteriophage β-glucosyltransferase. Subsequent treatment with APOBEC3A converts unmodified cytosines in the target polynucleotide to uracil. Accordingly, EM-seq allows identification of the modified cytosines 5-mC and 5-hmC (as protected glycosyl residues) by reading them as C; whereas unmodified C is converted to U.

Modified APOBEC

Modified APOBEC sequencing involves using a mutant APOBEC3A enzyme as the conversion agent, which is described in more detail in the Reference Examples 1 to 4 below. This process is described in US Provisional Application 63/328,444, which is incorporated herein by reference.

TAPS

TET-assisted pyridine borane sequencing (TAPS) involves using a TET1 enzyme as the further agent and pyridine borane as the conversion agent. This process is described in Liu et al. (Nature Biotechnology, 2019, 37, pp. 424-429), which is incorporated herein by reference. The TET1 enzyme causes oxidation of 5-methylcytosine, 5-hydroxymethylcytosine and 5-formylcytosine in the target polynucleotide to 5-carboxylcytosine. Subsequent treatment with pyridine borane converts 5-carboxylcytosine (including residues that used to be 5-methylcytosine, 5-hydroxymethylcytosine and 5-formylcytosine) to dihydrouracil, but does not convert unmodified cytosine. Accordingly, TAPS allows identification of the modified cytosines 5-mC and 5-hmC by reading them as T/U; whereas unmodified cytosine is read as C.

Methods of preparing polynucleotide templates

According to an embodiment of the present invention, a method of preparing polynucleotide templates for distinguishing between modified cytosines is described, comprising:

(a) providing a polynucleotide library hairpin strand comprising:

a double-stranded polynucleotide comprising a forward library strand and a reverse library strand,

a hairpin loop adaptor ligated to an end of the double-stranded polynucleotide, wherein the hairpin loop adaptor comprises a cleavable site,

wherein the polynucleotide library hairpin strand has been generated from a precursor polynucleotide library hairpin strand such that any CpG dyads in the precursor polynucleotide library hairpin comprising only unmodified cytosine are converted to a first dyad in the polynucleotide library hairpin strand, any CpG dyads in the precursor polynucleotide library hairpin comprising 5-methylcytosine are converted to a second dyad in the polynucleotide library hairpin strand, and any CpG dyads in the precursor polynucleotide library hairpin comprising 5-hydroxymethylcytosine are converted to a third dyad in the polynucleotide library hairpin strand,

wherein the first dyad, second dyad and third dyad are different to each other; and

(b) synthesising at least one template strand by generating a complement of the polynucleotide library hairpin strand, each of the template strands comprising a forward template strand complementary to the forward library strand, a spacer strand complementary to the hairpin loop adaptor, and a reverse template strand complementary to the reverse library strand, wherein the spacer strand comprises a first cleavable site.

Advantageously, the hairpin loop adaptor used in the polynucleotide library hairpin strand contains a cleavable site. In addition, the template strand that is generated from the polynucleotide library hairpin strand contains a first cleavable site in the spacer strand. This differs from previous methods, which do not contain a cleavable site within the hairpin loop adaptor, or in the resulting template strand generated from the library. A risk associated with templates generated from these types of polynucleotide library hairpin strands is that the template may rehybridise to itself – in other words, the forward template strand (complementary to the forward library strand) may self-hybridise to the reverse template strand (complementary to the reverse library strand). The formation of such self-hybridised structures can interfere with the sequencing process since nucleotides that are capable of emitting a signal (e.g. fluorescent nucleotides) are now unable to base-pair with the forward template strand (or reverse template strand). By installing the first cleavable site within the spacer strand, the forward template strand (or the reverse template strand) may be removed at a later stage, removing the risk of hairpins forming within the template. A further advantage is that longer reads are enabled.

As mentioned herein, the hairpin loop adaptor may be ligated to one end of the double-stranded polynucleotide such that the hairpin loop adaptor connects a 3'-end of the forward library strand with a 5'-end of the reverse library strand. In such a case, the polynucleotide library hairpin strand may comprise, in a 5' to 3' direction, the forward library strand, the hairpin loop adaptor (which

comprises the cleavable site), and the reverse library strand. However, the hairpin loop adaptor may instead be ligated to another end of the of the double-stranded polynucleotide such that the hairpin loop adaptor connects a 3'-end of the reverse library strand with a 5'-end of the forward library strand. In such a case, the polynucleotide library hairpin strand may comprise, in a 5' to 3' direction, the reverse library strand, the hairpin loop adaptor (which comprises the cleavable site), and the forward library strand.

As used herein, the term "dyad" refers to a portion of a double-stranded polynucleotide containing two adjacent nucleotides on one strand, and two adjacent polynucleotides at corresponding positions on the other strand. Both base pairs within the dyad may be complementary (e.g. in DNA, a CG/GC base pair or an AT/TA base pair; in RNA, a CG/GC base pair or an AU/UA base pair); however, in some cases, one or both base pairs within the dyad may not be complementary. For example, where certain types of dyad have been treated with conversion agents, this may cause one base pair within the dyad to be complementary, and one base pair in the dyad to not be complementary; in other cases, both base pairs within the dyad may not be complementary.

As used herein, the term "CpG dyad" refers to a double-stranded polynucleotide containing the following motif:

(5′)-C̲G-(3′)

(3′)-GC̲-(5′)

In the "CpG dyad" motif, C̲ may represent either unmodified cytosine, 5-methylcytosine or 5-hydroxymethylcytosine.

In particular, where the CpG dyad contains only unmodified cytosine, then both C's in the motif are unmodified cytosine.

Where the CpG dyad comprises 5-methylcytosine, then one or both C's may be 5-methylcytosine. In particular, where the library is prepared using a method of preparing a polynucleotide library hairpin strand as described in further detail below, the C in the forward library strand may be 5-methylcytosine and the C in the reverse library strand may be unmodified cytosine (e.g. in a third hairpin polynucleotide as described herein) – such a CpG dyad may be described as a "hemimethylated 5-methylcytosine CpG dyad". In other cases, both the C in the forward library strand may be 5-methylcytosine and the C in the reverse library strand may also be 5-methylcytosine (e.g. in a fourth hairpin polynucleotide as described herein) – such a CpG dyad may be described as a "fully methylated 5-methylcytosine CpG dyad".

Where the CpG dyad comprises 5-hydroxymethylcytosine, then one or both C's may be 5-hydroxymethylcytosine. In particular, where the library is prepared using a method of preparing a polynucleotide library hairpin strand as described in further detail below, the C in the forward library strand may be 5-hydroxymethylcytosine and the C in the reverse library strand may be unmodified cytosine (e.g. in a third hairpin polynucleotide and/or fourth hairpin polynucleotide as described herein) – such a CpG dyad may be described as a "hemimethylated 5-hydroxymethylcytosine CpG dyad".

As mentioned herein, the first dyad (produced from the CpG dyads comprising only unmodified cytosine), the second dyad (produced from the CpG dyads comprising 5-methylcytosine) and the third dyad (produced from the CpG dyads comprising 5-hydroxymethylcytosine) are different to each other when read. As used herein for the terms "first dyad", "second dyad" and "third dyad", these are different to each other when read if the complement of respective dyads are different. In other words, when a "first dyad" is sequenced, the sequence output is different compared to a sequence output of the "second dyad" and the "third dyad", since the complement of the "first dyad" is different to that of the complement of the "second dyad" and the complement of the "third dyad"; similarly, when a "second dyad" is sequenced, the sequence output is different compared to a sequence output of the "first dyad" and the "third dyad", since the complement of the "second dyad" is different to that of the complement of the "first dyad" and the complement of the "third dyad"; and similarly, when a "third dyad" is sequenced, the sequence output is different compared to a sequence output of the "first dyad" and the "second dyad", since the complement of the "third dyad" is different to that of the complement of the "first dyad" and the complement of the "second dyad".

In some embodiments (e.g. for fifth hairpin polynucleotides as described herein which have been prepared by exposing fourth hairpin polynucleotides as described herein to a conversion agent configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase which is read as thymine/uracil), the "first dyad" (produced from the CpG dyads comprising only unmodified cytosine) may contain the following motif:

(5′)-CG-(3′)

(3′)-GC-(5′);

the "second dyad" (produced from the CpG dyads comprising 5-methylcytosine) may contain the following motif:

(5′)-UG-(3′)

(3′)-GU-(5′);

and the "third dyad" (produced from the CpG dyads comprising 5-hydroxymethylcytosine) may contain the following motif:

(5′)-UG-(3′)

(3′)-GC-(5′);

where C represents unmodified cytosine, and U represents thymine or a nucleobase which is read as thymine/uracil. Since the "first dyad" is read as (5')-CG-(3') in the forward library strand and (3')-GC-(5') in the reverse library strand, the "second dyad" is read as (5')-TG-(3') in the forward library strand and (3')-GT-(5') in the reverse library strand, and the "third dyad" is read as (5')-TG-(3') in the forward library strand and (3')-GC-(5') in the reverse library strand (where all modified and unmodified cytosines are read only as cytosine during sequencing, and U is read only as thymine during sequencing), any CpG dyads comprising only unmodified cytosine, any CpG dyads comprising 5-methylcytosine, and any CpG dyads comprising 5-hydroxymethylcytosine can be distinguished from each other.

In some alternative embodiments (e.g. for fifth hairpin polynucleotides as described herein which have been prepared by exposing fourth hairpin polynucleotides as described herein to a conversion agent configured to convert an unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil), the "first dyad" (produced from the CpG dyads comprising only unmodified cytosine) may contain the following motif:

(5′)-UG-(3′)

(3′)-GU-(5′);

the "second dyad" (produced from the CpG dyads comprising 5-methylcytosine) may contain the following motif:

(5′)-C̲G-(3′)

(3′)-GC̲-(5′);

and the "third dyad" (produced from the CpG dyads comprising 5-hydroxymethylcytosine) may contain the following motif:

(5′)-C̲G-(3′)

(3′)-GU-(5′);

where C̲ represents modified cytosine (e.g. 5-methylcytosine, 5-hydroxymethylcytosine, or derivatives thereof such as protected 5-hydroxymethylcytosine, for example β-glucosyl-5-hydroxymethylcytosine), and U represents uracil or a nucleobase which is read as thymine/uracil. Since the "first dyad" is read as (5')-TG-(3') in the forward library strand and (3')-GT-(5') in the reverse library strand, the "second dyad" is read as (5')-CG-(3') in the forward library strand and (3')-GC-(5') in the reverse library strand, and the "third dyad" is read as (5')-CG-(3') in the forward library strand and (3')-GT-(5') in the reverse library strand (where all modified and unmodified cytosines are read only as cytosine during sequencing, and U is read only as thymine during sequencing), any CpG dyads comprising only unmodified cytosine, any CpG dyads

comprising 5-methylcytosine, and any CpG dyads comprising 5-hydroxymethylcytosine can again be distinguished from each other.

Where the polynucleotide library hairpin strand comprises, in a 5' to 3' direction, the forward library strand, the hairpin loop adaptor (which comprises the cleavable site), and the reverse library strand, the corresponding template strand may comprise, in a 5' to 3' direction, the reverse template strand, the spacer strand (which comprises the first cleavable site), and the forward template strand. Where the polynucleotide library hairpin strand comprises, in a 5' to 3' direction, the reverse library strand, the hairpin loop adaptor (which comprises the cleavable site), and the forward library strand, the corresponding template strand may comprise, in a 5' to 3' direction, the forward template strand, the spacer strand (which comprises the first cleavable site), and the reverse template strand.

In one embodiment, step (b) may comprise synthesising a plurality of template strands. Accordingly, the method of preparing polynucleotide templates for distinguishing between different modified cytosines may instead comprise a step of:

(b) synthesising a plurality of template strands by generating a complement of the polynucleotide library hairpin strand, each of the template strands comprising a forward template strand complementary to the forward library strand, a spacer strand complementary to the hairpin loop adaptor, and a reverse template strand complementary to the reverse library strand, wherein the spacer strand comprises a first cleavable site.

In one embodiment, the method may further comprise a step of:

(c) synthesising at least one template complement strand by generating a complement of the template strand, each of the template complement strands comprising a forward complement template strand, a spacer complement strand, and a reverse complement template strand, wherein the spacer complement strand comprises a second cleavable site.

Where the template strand comprises, in a 5' to 3' direction, the reverse template strand, the spacer strand (which comprises the first cleavable site), and the forward template strand, the corresponding template complement strand may comprise, in a 5' to 3' direction, the forward complement strand, the spacer complement strand (which comprises the second cleavable site), and the reverse complement template strand. Where the template strand comprises, in a 5' to 3' direction, the forward template strand, the spacer strand (which comprises the first cleavable site),

and the reverse template strand, the corresponding template complement strand may comprise, in a 5' to 3' direction, the reverse complement strand, the spacer complement strand (which comprises the second cleavable site), and the forward complement template strand.

In one embodiment, step (c) may comprise synthesising a plurality of template complement strands. Accordingly, the method of preparing polynucleotide templates for distinguishing between different modified cytosines may instead comprise a step of:

(c) synthesising a plurality of template complement strands by generating a complement of the template strand, each of the template complement strands comprising a forward complement template strand, a spacer complement strand, and a reverse complement template strand, wherein the spacer complement strand comprises a second cleavable site.

In one embodiment, the method may further comprise a step of:

(d) cleaving the first cleavable site on the at least one template strand to generate at least one first polynucleotide sequence each comprising a first portion and cleaving the second cleavable site on the at least one template complement strand to generate at least one second polynucleotide sequence each comprising a second portion,

wherein the first portion corresponds with the forward template strand and the second portion corresponds with the reverse complement template strand, or wherein the first portion corresponds with the reverse template strand and the second portion corresponds with the forward complement template strand.

Where the first portion corresponds with the forward template strand and the second portion corresponds with the reverse complement template strand, then it is the reverse template strand and the forward complement template strand that have been removed. Where the first portion corresponds with the reverse template strand and the second portion corresponds with the forward complement template strand, then it is the forward template strand and the reverse complement template strand that have been removed. Since portions of the original template and template complement strands that would self-hybridise have been removed, this reduces the risk of unwanted hairpins forming within the first polynucleotide sequence and the second polynucleotide sequence, as mentioned above.

As mentioned herein, step (b) may comprise synthesising a plurality of template strands, and step (c) may comprise synthesising a plurality of template complement strands. Accordingly, the

method of preparing polynucleotide templates for distinguishing between different modified cytosines may instead comprise a step of:

(d) cleaving the first cleavable site on the plurality of template strands to generate a plurality of first polynucleotide sequences each comprising a first portion and cleaving the second cleavable site on the plurality of template complement strands to generate a plurality of second polynucleotide sequences each comprising a second portion,

wherein the first portion corresponds with the forward template strand and the second portion corresponds with the reverse complement template strand, or wherein the first portion corresponds with the reverse template strand and the second portion corresponds with the forward complement template strand.

By "cleavable site" is meant any moiety that allows the hairpin loop adaptor, spacer strand, or spacer complement strand to be separated into two strands from a single strand. In some embodiments, the cleavable site is a restriction site.

By "restriction site" is meant a sequence of nucleotides recognised by an endonuclease. In particular, the endonuclease may be a double strand restriction endonuclease or restriction enzyme. By either of these terms is meant an enzyme that can hydrolyze both strands of a double-stranded polynucleotide (duplex), to produce polynucleotide molecules that are cleaved on both strands. In a further embodiment, the restriction enzyme may be a type II restriction enzyme. In an even further embodiment, the restriction enzyme may be a type IIP restriction enzyme, a type IIS restriction enzyme, a type IIC restriction enzyme, or a type IIT restriction enzyme.

In one example, the type II restriction enzyme may be EcoRI and the restriction enzyme is G/AATTC wherein EcoRI catalyzes a double stranded break within the recognition site. In another example, the type II restriction enzyme may be BglII and the restriction site is A/GATCT, wherein BglII catalyzes a double stranded break within the recognition site. In a further example, the type II restriction enzyme may be NotI and the restriction site is GC/GGCCGC, wherein NotI catalyses a double stranded break within the recognition site. In a yet further example, the type II restriction enzyme may be FokI and the restriction site is GGATG(9/13). Other suitable endonucleases are available from commercial sources, including New England Biolabs and Fisher Scientific.

In an alternative embodiment, the endonuclease is a CRISPR enzyme.

CRISPR-Cas mechanisms are currently classified into two classes (classes 1 and 2) and six types (types I to VI). In one embodiment, the CRISPR enzyme is a class I enzyme, and is selected from type I, III and IV. In one embodiment, the CRISPR enzyme is Cas6. In another embodiment, the CRISPR enzyme is a class 2 enzyme, and is selected from type II, IV, V, and VI. In one embodiment, the CRISPR enzyme is selected from Cas9, Cpf1 (Cas12a), Mad7, CasC2c1, C2C2 and C2c3. In another embodiment, the CRISPR enzyme is a type VI CRISPR enzyme. In one embodiment, the enzyme is selected from CasC2c1, C2C2 and C2c3.

CRISPR enzymes may be naturally occurring, for example Cas9 may be obtained from any one of *Staphylococcus aureus, Neisseria meningitides, Streptococcus thermophiles, Treponema denticola* or *Campylobacter jejuni*. Cpf1 enzymes may be selected from the *Acidaminococcus sp* (AsCpf1) or *Lachnospiraceae bacterium* (LbCpf1).

In an alternative embodiment, the CRISPR enzyme is a Cas9 paired nickase. Examples of a Cas9 paired nickase include Cas9 D10A and Cas9 H840A. For example, in one embodiment, the Cas9 protein may comprise the D10A or H840A amino acid substitutions. These nickases cleave only the DNA strand that is complementary to and recognised by a gRNA.

In one embodiment, the restriction site may be or may comprise a PAM (protospacer adjacent motif) sequence. Examples of suitable PAM sequences include NGG, NGAG, NGCG, NGN, NG, GAA, GAT, NNG, NGN, NRN, YG, NNGRRT, NNNRRT, NNAGAA, NNNNGATT and NNNNCRAA and complements thereof.

In a further embodiment, the Cas9 protein may alternatively or additionally comprise the N863A or N854A amino acid substitutions.

In a further embodiment, the Cas9 protein has been modified to improve activity. For example, in one embodiment, the Cas9 protein may additionally comprise a D1135E substitution. Alternatively, the Cas9 protein may also be the VQR variant.

Thus, in one embodiment, the first cleavable site may be a first restriction site for an endonuclease. In a further embodiment, the endonuclease may be a type II restriction enzyme as defined herein, or a CRISPR enzyme as defined herein (e.g. a Cas9 paired nickase as defined herein). In an even further embodiment, the restriction enzyme is a type IIP restriction enzyme, a type IIS restriction enzyme, a type IIC restriction enzyme, or a type IIT restriction enzyme. Non-limiting examples of the type II restriction enzyme may include EcoRI, BglII, NotI and FokI; non-limiting examples of the CRISPR enzyme include Cas6, Cas9, Cpf1 (Cas12a), Mad7, CasC2c1, C2C2 and C2c3, and Cas9 paired nickases such as Cas9 D10A and Cas9 H840A.

In another embodiment, the second cleavable site may be a second restriction site for an endonuclease. In a further embodiment, the endonuclease may be a type II restriction enzyme as defined herein, or a CRISPR enzyme as defined herein (e.g. a Cas9 paired nickase as defined herein). In an even further embodiment, the restriction enzyme is a type IIP restriction enzyme, a type IIS restriction enzyme, a type IIC restriction enzyme, or a type IIT restriction enzyme. Non-limiting examples of the type II restriction enzyme may include EcoRI, BglII, NotI and FokI; non-limiting examples of the CRISPR enzyme include Cas6, Cas9, Cpf1 (Cas12a), Mad7, CasC2c1, C2C2 and C2c3, and Cas9 paired nickases such as Cas9 D10A and Cas9 H840A.

In some embodiments, the first cleavable site and the second cleavable site may be cleaved under the same reaction conditions. For example, the first cleavable site and the second cleavable site may be a restriction site recognised by the same endonuclease. In a further embodiment, the endonuclease may be a type II restriction enzyme as defined herein that recognises both the first cleavable site and the second cleavable site, or a CRISPR enzyme as defined herein (e.g. a Cas9 paired nickase as defined herein) that recognises both the first cleavable site and the second cleavable site. In an even further embodiment, the restriction enzyme is a type IIP restriction enzyme, a type IIS restriction enzyme, a type IIC restriction enzyme, or a type IIT restriction enzyme, that recognises both the first cleavable site and the second cleavable site. Again, non-limiting examples of the type II restriction enzyme may include EcoRI, BglII, NotI and FokI, recognising both the first cleavable site and the second cleavable site; non-limiting examples of the CRISPR enzyme include Cas6, Cas9, Cpf1 (Cas12a), Mad7, CasC2c1, C2C2 and C2c3, and Cas9 paired nickases such as Cas9 D10A and Cas9 H840A, recognising both the first cleavable site and the second cleavable site.

In some embodiments, the at least one first polynucleotide sequence (or plurality of first polynucleotide sequences) may each comprise a first sequencing primer binding site. In a further embodiment, the first sequencing primer binding site may be located after a 3'-end of the first portion. The first sequencing primer binding site may, for instance, have formed part of the spacer strand.

In some embodiments, the at least one second polynucleotide sequence (or plurality second polynucleotide sequences) may each comprise a second sequencing primer binding site. In a further embodiment, the second sequencing primer binding site may be located after a 3'-end of the second portion. The second sequencing primer binding site may, for instance, have formed part of the spacer complement strand.

As shown in Figures 16 and 17, the first dyad (produced from the CpG dyads comprising only unmodified cytosine), the second dyad (produced from the CpG dyads comprising 5-methylcytosine) and the third dyad (produced from the CpG dyads comprising 5-hydroxymethylcytosine) generate different patterns when comparing appropriate bases in the forward template strand and the reverse complement template strand, or when comparing appropriate bases in the reverse template strand and the forward complement template strand.

For example, in some embodiments (e.g. for fifth hairpin polynucleotides as described herein which have been prepared by exposing fourth hairpin polynucleotides as described herein to a conversion agent configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase which is read as thymine/uracil), Figure 16 shows how the first dyad, second dyad and the third dyad form certain patterns that allow each of these to be distinguished from each other.

In such an example, the first dyad (produced from the CpG dyads comprising only unmodified cytosine) produces a (5')-CG-(3') sequence in the forward template strand at positions corresponding to the original CpG dyad, and a (5')-CG-(3') sequence in the reverse complement template strand at positions corresponding to the original CpG dyad. Similarly, the first dyad (produced from the CpG dyads comprising only unmodified cytosine) produces a (5')-CG-(3') sequence in the reverse template strand at positions corresponding to the original CpG dyad, and a (5')-CG-(3') sequence in the forward complement template strand at positions corresponding to the original CpG dyad. Accordingly, a double C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad (e.g. for the first base in each sequence, both are C and match, and for the second base in each sequence, both are G and match). This allows easy identification of the original CpG dyad comprising only unmodified cytosine (as compared to CpG dyads comprising 5-methylcytosine and CpG dyads comprising 5-hydroxymethylcytosine).

In addition, the second dyad (produced from the CpG dyads comprising 5-methylcytosine) produces a (5')-CA-(3') sequence in the forward template strand at positions corresponding to the original CpG dyad, and a (5')-TG-(3') sequence in the reverse complement template strand at positions corresponding to the original CpG dyad. Similarly, the second dyad (produced from the CpG dyads comprising 5-methylcytosine) produces a (5')-CA-(3') sequence in the reverse template strand at positions corresponding to the original CpG dyad, and a (5')-TG-(3') sequence in the forward complement template strand at positions corresponding to the original CpG dyad. Accordingly, a double mismatch is present when comparing corresponding positions in the at

least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad (e.g. for the first base in each sequence, one is C, and the other is T, showing a mismatch, and for the second base in each sequence, one is A, and the other is G, showing another mismatch). This allows easy identification of the original CpG dyad comprising 5-methylcytosine (as compared to CpG dyads comprising only unmodified cytosine and CpG dyads comprising 5-hydroxymethylcytosine).

Furthermore, the third dyad (produced from the CpG dyads comprising 5-hydroxymethylcytosine) produces a (5')-CA-(3') sequence in the forward template strand at positions corresponding to the original CpG dyad, and a (5')-CG-(3') sequence in the reverse complement template strand at positions corresponding to the original CpG dyad. Similarly, the third dyad (produced from the CpG dyads comprising 5-hydroxymethylcytosine) produces a (5')-CG-(3') sequence in the reverse template strand at positions corresponding to the original CpG dyad, and a (5')-TG-(3') sequence in the forward complement template strand at positions corresponding to the original CpG dyad. Accordingly, a single mismatch and single C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad (e.g. for the first base in each sequence, both are C and match, and for the second base in each sequence, one is A, and the other is G, showing a mismatch; alternatively, for the first base in each sequence, one is C, and the other is T, showing a mismatch, and for the second base in each sequence, both are G and match). This allows easy identification of the original CpG dyad comprising 5-hydroxymethylcytosine (as compared to CpG dyads comprising only unmodified cytosine and CpG dyads comprising 5-methylcytosine).

Accordingly, in one embodiment, where CpG dyads comprising only unmodified cytosine were present in the precursor polynucleotide library hairpin, then a double C-C/G-G match may be present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; where CpG dyads comprising 5-methylcytosine were present in the precursor polynucleotide library hairpin, then a double mismatch may be present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; and where CpG dyads comprising 5-hydroxymethylcytosine were present in the precursor polynucleotide library hairpin, then a single mismatch and single C-C/G-G match may be present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad.

In some alternative embodiments (e.g. for fifth hairpin polynucleotides as described herein which have been prepared by exposing fourth hairpin polynucleotides as described herein to a conversion agent configured to convert an unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil), Figure 17 shows how the first dyad, second dyad and the third dyad form certain patterns that allow each of these to be distinguished from each other.

In such an example, the first dyad (produced from the CpG dyads comprising only unmodified cytosine) produces a (5')-CA-(3') sequence in the forward template strand at positions corresponding to the original CpG dyad, and a (5')-TG-(3') sequence in the reverse complement template strand at positions corresponding to the original CpG dyad. Similarly, the first dyad (produced from the CpG dyads comprising only unmodified cytosine) produces a (5')-CA-(3') sequence in the reverse template strand at positions corresponding to the original CpG dyad, and a (5')-TG-(3') sequence in the forward complement template strand at positions corresponding to the original CpG dyad. Accordingly, a double mismatch is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad (e.g. for the first base in each sequence, one is C, and the other is T, showing a mismatch, and for the second base in each sequence, one is A, and the other is G, showing another mismatch). This allows easy identification of the original CpG dyad comprising only unmodified cytosine (as compared to CpG dyads comprising 5-methylcytosine and CpG dyads comprising 5-hydroxymethylcytosine).

In addition, the second dyad (produced from the CpG dyads comprising 5-methylcytosine) produces a (5')-CG-(3') sequence in the forward template strand at positions corresponding to the original CpG dyad, and a (5')-CG-(3') sequence in the reverse complement template strand at positions corresponding to the original CpG dyad. Similarly, the second dyad (produced from the CpG dyads comprising 5-methylcytosine) produces a (5')-CG-(3') sequence in the reverse template strand at positions corresponding to the original CpG dyad, and a (5')-CG-(3') sequence in the forward complement template strand at positions corresponding to the original CpG dyad. Accordingly, a double C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad (e.g. for the first base in each sequence, both are C and match, and for the second base in each sequence, both are G and match). This allows easy identification of the original CpG dyad comprising 5-methylcytosine (as compared to CpG dyads comprising only unmodified cytosine and CpG dyads comprising 5-hydroxymethylcytosine).

Furthermore, the third dyad (produced from the CpG dyads comprising 5-hydroxymethylcytosine) produces a (5')-CG-(3') sequence in the forward template strand at

positions corresponding to the original CpG dyad, and a (5')-TG-(3') sequence in the reverse complement template strand at positions corresponding to the original CpG dyad. Similarly, the third dyad (produced from the CpG dyads comprising 5-hydroxymethylcytosine) produces a (5')-CA-(3') sequence in the reverse template strand at positions corresponding to the original CpG dyad, and a (5')-CG-(3') sequence in the forward complement template strand at positions corresponding to the original CpG dyad. Accordingly, a single mismatch and single C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad (e.g. for the first base in each sequence, one is C, and the other is T, showing a mismatch, and for the second base in each sequence, both are G and match; alternatively, for the first base in each sequence, both are C and match, and for the second base in each sequence, one is A, and the other is G, showing a mismatch). This allows easy identification of the original CpG dyad comprising 5-hydroxymethylcytosine (as compared to CpG dyads comprising only unmodified cytosine and CpG dyads comprising 5-methylcytosine).

Accordingly, in another embodiment, wherein where CpG dyads comprising only unmodified cytosine were present in the precursor polynucleotide library hairpin, then a double mismatch may be present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; where CpG dyads comprising 5-methylcytosine were present in the precursor polynucleotide library hairpin, then a double C-C/G-G match may be present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; and where CpG dyads comprising 5-hydroxymethylcytosine were present in the precursor polynucleotide library hairpin, then a single mismatch and single C-C/G-G match may be present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad.

Whilst Figures 16 and 17 show that the reverse template strand and the forward complement template strand are retained as part of the at least one first polynucleotide sequence and the at least one second polynucleotide sequence after cleavage, it should be appreciated that in other cases, it may be the forward template strand and the reverse complement template strand that are retained instead. This can be achieved, for example, by attaching the hairpin loop adaptor at the other end to that shown in Figures 16 and 17, as explained above.

Furthermore, whilst Figures 16 and 17 show that the template strand is generated extending from the second immobilised primer (e.g. P7), and that the template complement strand is generated

extending from the first immobilised primer (e.g. P5), it should be appreciated that in other cases, the template strand may instead extend from the first immobilised primer (e.g. P5), and that the template complement strand may instead extend from the second immobilised primer (e.g. P7). This can be achieved, for example, by using a P5'/P7 adaptor instead of a P7'/P5 adaptor, as explained above.

As described herein, the polynucleotide sequences each comprise portions of a double-stranded nucleic acid template, and the first portion may comprise (or be) the forward strand of a polynucleotide sequence (e.g. forward strand of a template, or forward template strand), and the second portion may comprise (or be) the reverse complement strand of the polynucleotide sequence (e.g. reverse complement strand of the template, or reverse complement template strand) (in effect, a reverse complement strand may be considered a "copy" of the forward strand). Alternatively, the first portion may comprise (or be) the reverse strand of a polynucleotide sequence (e.g. reverse strand of a template, or reverse template strand), and the second portion may comprise (or be) the forward complement strand of the polynucleotide sequence (e.g. forward complement strand of the template, or forward complement template strand) (in effect, a forward complement may be considered a "copy" of the reverse strand).

The first portion may be derived from a forward strand of a target polynucleotide to be sequenced (also referred to herein as a forward library strand), and the second portion may be derived from a reverse complement strand of the target polynucleotide to be sequenced (also may be considered as a reverse complement library strand, a complement of a reverse library strand); or the first portion may be derived from a reverse strand of a target polynucleotide to be sequenced (also referred to herein as a reverse library strand), and the second portion may be derived from a forward complement strand of the target polynucleotide to be sequenced (also may be considered as a forward complement library strand, a complement of a forward library strand).

The template is generated from a (double-stranded) target polynucleotide to be sequenced via complementary base pairing. The (double-stranded) target polynucleotide may be one (double-stranded) polynucleotide present in a polynucleotide library to be sequenced. As such, the template allows sequence information to be obtained for that particular polynucleotide.

The method may further comprise a step of preparing the first portion and the second portion for concurrent sequencing.

For example, the method may comprise simultaneously contacting first sequencing primer binding sites located after a 3'-end of the first portions with first primers and second sequencing

primer binding sites located after a 3'-end of the second portions with second primers. Thus, the first portions and second portions are primed for concurrent sequencing.

In some embodiments, the method may comprises a step of processing the at least one first polynucleotide sequence comprising a first portion and the at least one second polynucleotide sequence comprising a second portion, such that a proportion of first portions are capable of generating a first signal and a proportion of second portions are capable of generating a second signal.

In some embodiments, the first signal and the second signal may be spatially resolved. In other embodiments, the first signal and the second signal may be spatially unresolved.

In some embodiments (e.g. where spatially resolvable signals are used), a proportion of first portions may be capable of generating a first signal and a proportion of second portions may be capable of generating a second signal, wherein an intensity of the first signal is substantially the same as an intensity of the second signal.

In other embodiments (e.g. where selective processing methods are used as described herein), a proportion of first portions may be capable of generating a first signal and a proportion of second portions may be capable of generating a second signal, wherein the selective processing causes an intensity of the first signal to be greater than an intensity of the second signal. The first signal and the second signal may be spatially unresolved (e.g. generated from the same region or substantially overlapping regions).

Further aspects relating to selective processing methods (e.g. conducting selective amplification, conducting selective sequencing or preparing for selective sequencing) have already been described herein and apply to the methods of preparing polynucleotide templates for distinguishing between modified cytosines as described herein.

The first portion may be referred to herein as read 1.1 (R1.1). The second portion may be referred to herein as read 1.2 (R1.2).

In one embodiment, the first portion is at least 25 or at least 50 base pairs and the second portion is at least 25 base pairs or at least 50 base pairs.

The first and second strand may be separately attached to a solid support. In a further embodiment, this solid support may be a flow cell. In a further embodiment, each of the first and second strands are attached to the solid support (e.g. flow cell) in a single well of the solid support.

The polynucleotide strands may form or be part of a cluster on the solid support.

As used herein, the term "cluster" may refer to a clonal group of template polynucleotides (e.g. DNA or RNA) bound within a single well of a solid support (e.g. flow cell). As such, a cluster may refer to the population of polynucleotide molecules within a well that are then sequenced. A "cluster" may contain a sufficient number of copies of template polynucleotides such that the cluster is able to output a signal (e.g. a light signal) that allows sequencing reads to be performed on the cluster. A "cluster" may comprise, for example, about 500 to about 2000 copies, about 600 to about 1800 copies, about 700 to about 1600 copies, about 800 to about 1400 copies, about 900 to about 1200 copies, or about 1000 copies of template polynucleotides.

A cluster may be formed by bridge amplification, as described above.

Where the method of the invention involves a first polynucleotide strand and a second polynucleotide strand, the cluster formed may be a duoclonal cluster.

By "duoclonal" cluster is meant that the population of polynucleotide sequences that are then sequenced (as the next step) are substantially of two types – e.g. a first sequence and a second sequence. As such, a "duoclonal" cluster may refer to the population of single first sequences and single second sequences within a well that are then sequenced. A "duoclonal" cluster may contain a sufficient number of copies of a single first sequence and copies of a single second sequence such that the cluster is able to output a signal (e.g. a light signal) that allows sequencing reads to be performed on the "monoclonal" cluster. A "duoclonal" cluster may comprise, for example, about 500 to about 2000 combined copies, about 600 to about 1800 combined copies, about 700 to about 1600 combined copies, about 800 to about 1400 combined copies, about 900 to about 1200 combined copies, or about 1000 combined copies of single first sequences and single second sequences. The copies of single first sequences and single second sequences together may comprise at least about 50%, at least about 60%, at least about 70%, at least about 80%, at least about 90%, or about 95%, 98%, 99% or 100% of all polynucleotides within a single well of the flow cell, and thus providing a substantially duoclonal "cluster".

The at least one first polynucleotide sequence comprising a first portion and at least one second polynucleotide sequence may be prepared using a loop fork method as described herein (see

Figure 4). For example, the polynucleotide library hairpin strand may be prepared using methods of preparing a polynucleotide library hairpin strand as described herein.

In some embodiments, the method may further comprise a step of concurrently sequencing nucleobases in the first portion and the second portion.

Figure 16 shows an example workflow for preparing a polynucleotide library hairpin strand according to a method as described herein (using a conversion agent configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase which is read as thymine/uracil), then a subsequent example workflow for preparing polynucleotide templates for distinguishing between modified cytosines according to a method as described herein.

For library preparation, starting from a double-stranded polynucleotide comprising a precursor forward library strand 2001 and a precursor reverse library strand (not shown), a hairpin loop adaptor 2002 is ligated, where the hairpin loop adaptor comprises a cleavable site 2003, to generate a first hairpin polynucleotide (not shown) (steps (a) and (b) for methods of preparing a polynucleotide library hairpin strand as described herein). Removing the precursor reverse library strand then generates second hairpin polynucleotide 2200, which comprises the precursor forward library strand 2001 and the hairpin loop adaptor 2002, wherein the hairpin loop adaptor 2002 comprises the cleavable site 2003 (step (c) for methods of preparing a polynucleotide library hairpin strand as described herein). As shown in Figure 16, the precursor forward library strand 2001 comprises a CpG sequence comprising 5-methylcytosine, a CpG sequence comprising 5-hydroxymethylcytosine, and a CpG sequence comprising unmodified cytosine.

By extending from the 3'-end of the hairpin loop adaptor 2002 in a 5' to 3' direction, a resynthesised reverse library strand 2004 can be generated (e.g. using a DNA polymerase), thus forming third hairpin polynucleotide 2300 (step (d) for methods of preparing a polynucleotide library hairpin strand as described herein). Therefore, third hairpin polynucleotide 2300 comprises the precursor forward library strand 2001, the hairpin loop adaptor 2002 (comprising the cleavable site 2003), and resynthesised reverse library strand 2004. This process may be conducted using unmodified cytosines, thus meaning that the resynthesised reverse library strand 2004 is devoid of modified cytosines such as 5-methylcytosine and 5-hydroxymethylcytosine. This generates a CpG dyad 2005 comprising only unmodified cytosine, a CpG dyad 2006 comprising 5-methylcytosine (a hemimethylated 5-methylcytosine CpG dyad), and a CpG dyad 2007 comprising 5-hydroxymethylcytosine (a hemimethylated 5-hydroxymethylcytosine CpG dyad).

Ligation of a flanking adaptor 2005 (a P7'/P5 fork adaptor) to third hairpin polynucleotide 2300 at an end away from the hairpin loop adaptor 2002 then generates hairpin polynucleotide 2300-1.

Treatment of hairpin polynucleotide 2300-1 with DNA methyltransferase 1 enzyme transforms CpG dyad 2006 comprising 5-methylcytosine (a hemimethylated 5-methylcytosine CpG dyad) to a fully methylated 5-methylcytosine CpG dyad. However, CpG dyad 2007 comprising 5-hydroxymethylcytosine (a hemimethylated 5-hydroxymethylcytosine CpG dyad) is unaffected, as well as CpG dyad 2005 comprising only unmodified cytosine. This generates fourth hairpin polynucleotide 2400 (step (e) for methods of preparing a polynucleotide library hairpin strand as described herein). Therefore, fourth hairpin polynucleotide 2400 comprises the precursor forward library strand 2001, the hairpin loop adaptor 2002 (comprising the cleavable site 2003), and partially methylated reverse library strand 2008.

Finally, exposing the fourth hairpin polynucleotide 2400 to a conversion agent configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase which is read as thymine/uracil generates fifth hairpin polynucleotide 2500 (step (f) for methods of preparing a polynucleotide library hairpin strand as described herein). The precursor forward library strand 2001 is converted to forward library strand 2009. The partially methylated reverse library strand 2008 is converted to reverse library strand 2010. The CpG dyad 2005 comprising only unmodified cytosine is unaffected, which can now be considered as first dyad 2011. The CpG dyad 2006 comprising 5-methylcytosine (a hemimethylated 5-methylcytosine CpG dyad), which was transformed to a fully methylated 5-methylcytosine CpG dyad, is converted to second dyad 2012. The CpG dyad 2007 comprising 5-hydroxymethylcytosine (a hemimethylated 5-hydroxymethylcytosine CpG dyad) is converted to third dyad 2013.

For template generation, fifth hairpin polynucleotide 2500 anneals (via P7') to immobilised primer P7 on a solid support. Extending from the 3'-end of immobilised primer P7 in a 5' to 3' direction generates a template strand (e.g. using a DNA polymerase). The template strand comprises, in a 5' to 3' direction, a reverse template strand 2050 which is complementary to reverse library strand 2010, a spacer strand 2051 (comprising a first cleavable site 2053) which is complementary to hairpin loop adaptor 2002, and a forward template strand 2052 which is complementary to forward library strand 2009.

After conducting amplification, a template complement strand is generated which extends from the 3'-end of immobilised primer P5. The template complement strand comprises, in a 5' to 3' direction, a forward complement template strand 2052', a spacer complement strand 2051' (comprising a second cleavable site 2054), and a reverse complement template strand 2050'.

Cleavage of the first cleavable site 2053 and second cleavable site 2054 causes removal of forward template strand 2052 and reverse complement template strand 2050', thus leaving behind a first polynucleotide sequence comprising a first portion (reverse template strand 2050) and a second polynucleotide sequence comprising a second portion (forward complement template strand 2052'). The templates are now ready for sequencing to determine the presence of 5-methylcytosine and 5-hydroxymethylcytosine.

Figure 17 shows another example workflow for preparing a polynucleotide library hairpin strand according to a method as described herein (using a conversion agent configured to convert an unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil), then a subsequent example workflow for preparing polynucleotide templates for distinguishing between modified cytosines according to a method as described herein.

For library preparation, again starting from a double-stranded polynucleotide comprising a precursor forward library strand 2001 and a precursor reverse library strand (not shown), a hairpin loop adaptor 2002 is ligated, where the hairpin loop adaptor comprises a cleavable site 2003, to generate a first hairpin polynucleotide (not shown) (steps (a) and (b) for methods of preparing a polynucleotide library hairpin strand as described herein). Removing the precursor reverse library strand then generates second hairpin polynucleotide 2200, which comprises the precursor forward library strand 2001 and the hairpin loop adaptor 2002, wherein the hairpin loop adaptor 2002 comprises the cleavable site 2003 (step (c) for methods of preparing a polynucleotide library hairpin strand as described herein). As shown in Figure 17, the precursor forward library strand 2001 comprises a CpG sequence comprising 5-methylcytosine, a CpG sequence comprising 5-hydroxymethylcytosine, and a CpG sequence comprising unmodified cytosine.

By extending from the 3'-end of the hairpin loop adaptor 2002 in a 5' to 3' direction, a resynthesised reverse library strand 2004 can be generated (e.g. using a DNA polymerase), thus forming third hairpin polynucleotide 2300 (step (d) for methods of preparing a polynucleotide library hairpin strand as described herein). Therefore, third hairpin polynucleotide 2300 comprises the precursor forward library strand 2001, the hairpin loop adaptor 2002 (comprising the cleavable site 2003), and resynthesised reverse library strand 2004. This process may be conducted using unmodified cytosines, thus meaning that the resynthesised reverse library strand 2004 is devoid of modified cytosines such as 5-methylcytosine and 5-hydroxymethylcytosine. This generates a CpG dyad 2005 comprising only unmodified cytosine, a CpG dyad 2006 comprising 5-methylcytosine (a hemimethylated 5-methylcytosine CpG dyad), and a CpG dyad

2007 comprising 5-hydroxymethylcytosine (a hemimethylated 5-hydroxymethylcytosine CpG dyad).

Ligation of a flanking adaptor 2005 (a P7'/P5 fork adaptor) to third hairpin polynucleotide 2300 at an end away from the hairpin loop adaptor 2002 then generates hairpin polynucleotide 2300-1.

Treatment of hairpin polynucleotide 2300-1 with DNA methyltransferase 1 enzyme transforms CpG dyad 2006 comprising 5-methylcytosine (a hemimethylated 5-methylcytosine CpG dyad) to a fully methylated 5-methylcytosine CpG dyad. However, CpG dyad 2007 comprising 5-hydroxymethylcytosine (a hemimethylated 5-hydroxymethylcytosine CpG dyad) is unaffected, as well as CpG dyad 2005 comprising only unmodified cytosine. This generates fourth hairpin polynucleotide 2400 (step (e) for methods of preparing a polynucleotide library hairpin strand as described herein). Therefore, fourth hairpin polynucleotide 2400 comprises the precursor forward library strand 2001, the hairpin loop adaptor 2002 (comprising the cleavable site 2003), and partially methylated reverse library strand 2008.

Finally, exposing the fourth hairpin polynucleotide 2400 to a conversion agent configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase which is read as thymine/uracil generates fifth hairpin polynucleotide 2500 (step (f) for methods of preparing a polynucleotide library hairpin strand as described herein). The precursor forward library strand 2001 is converted to forward library strand 2009. The partially methylated reverse library strand 2008 is converted to reverse library strand 2010. The CpG dyad 2005 comprising only unmodified cytosine is converted to first dyad 2011. The CpG dyad 2006 comprising 5-methylcytosine (a hemimethylated 5-methylcytosine CpG dyad), which was transformed to a fully methylated 5-methylcytosine CpG dyad, is unchanged in this step and can be considered as second dyad 2012. The CpG dyad 2007 comprising 5-hydroxymethylcytosine (a hemimethylated 5-hydroxymethylcytosine CpG dyad) is converted to third dyad 2013.

For template generation, fifth hairpin polynucleotide 2500 anneals (via P7') to immobilised primer P7 on a solid support. Extending from the 3'-end of immobilised primer P7 in a 5' to 3' direction generates a template strand (e.g. using a DNA polymerase). The template strand comprises, in a 5' to 3' direction, a reverse template strand 2050 which is complementary to reverse library strand 2010, a spacer strand 2051 (comprising a first cleavable site 2053) which is complementary to hairpin loop adaptor 2002, and a forward template strand 2052 which is complementary to forward library strand 2009.

After conducting amplification, a template complement strand is generated which extends from the 3'-end of immobilised primer P5. The template complement strand comprises, in a 5' to 3' direction, a forward complement template strand 2052', a spacer complement strand 2051' (comprising a second cleavable site 2054), and a reverse complement template strand 2050'.

Cleavage of the first cleavable site 2053 and second cleavable site 2054 causes removal of forward template strand 2052 and reverse complement template strand 2050', thus leaving behind a first polynucleotide sequence comprising a first portion (reverse template strand 2050) and a second polynucleotide sequence comprising a second portion (forward complement template strand 2052'). The templates are now ready for sequencing to determine the presence of 5-methylcytosine and 5-hydroxymethylcytosine.

Methods of sequencing

Also described herein is a method of sequencing polynucleotide sequences to distinguish between modified cytosines, comprising:

preparing polynucleotide templates for distinguishing between modified cytosines using a method as described herein;

sequencing nucleobases in the first portion and the second portion; and

identifying the presence of 5-methylcytosine or 5-hydroxymethylcytosine by detecting differences when comparing a sequence output from the first portion with a sequence output from the second portion.

In one embodiment, the step of sequencing nucleobases in the first portion and the second portion may involve concurrent sequencing of nucleobases in the first portion and the second portion.

In one embodiment, sequencing is performed by sequencing-by-synthesis.

In one embodiment, the method may further comprise a step of conducting paired-end reads.

In some embodiments, where the method comprises a step of selectively processing the at least one polynucleotide sequence comprising the first portion and the second portion, such that a proportion of first portions are capable of generating a first signal and a proportion of second portions are capable of generating a second signal, wherein the selective processing causes an intensity of the first signal to be greater than an intensity of the second signal, the data may be analysed using 16 QAM as mentioned herein.

Accordingly, the step of concurrently sequencing nucleobases may comprise:

(a) obtaining first intensity data comprising a combined intensity of a first signal component obtained based upon a respective first nucleobase at the first portion and a second signal component obtained based upon a respective second nucleobase at the second portion, wherein the first and second signal components are obtained simultaneously;

(b) obtaining second intensity data comprising a combined intensity of a third signal component obtained based upon the respective first nucleobase at the first portion and a fourth signal component obtained based upon the respective second nucleobase at the second portion, wherein the third and fourth signal components are obtained simultaneously;

(c) selecting one of a plurality of classifications based on the first and the second intensity data, wherein each classification represents a possible combination of respective first and second nucleobases; and

(d) based on the selected classification, base calling the respective first and second nucleobases.

In one embodiment, selecting the classification based on the first and second intensity data may comprise selecting the classification based on the combined intensity of the first and second signal components and the combined intensity of the third and fourth signal components.

In one embodiment, the plurality of classifications may comprise sixteen classifications, each classification representing one of sixteen unique combinations of first and second nucleobases.

In one embodiment, the first signal component, second signal component, third signal component and fourth signal component may be generated based on light emissions associated with the respective nucleobase.

In one embodiment, the light emissions may be detected by a sensor, wherein the sensor is configured to provide a single output based upon the first and second signals.

In one embodiment, the sensor may comprise a single sensing element.

In one embodiment, the method may further comprise repeating steps (a) to (d) for each of a plurality of base calling cycles.

Kits

Methods as described herein may be performed by a user physically. In other words, a user may themselves conduct the methods of preparing polynucleotide templates for distinguishing between modified cytosines as described herein, and as such the methods as described herein may not need to be computer-implemented.

In another aspect of the invention, there is provided a kit comprising instructions for preparing polynucleotide templates for distinguishing between modified cytosines as described herein, and/or for sequencing polynucleotide sequences to distinguish between modified cytosines as described herein.

Computer programs and products

In other embodiments, methods as described herein may be performed by a computer. In other words, a computer may contain instructions to conduct the methods of preparing polynucleotide templates for distinguishing between modified cytosines as described herein, and as such the methods as described herein may be computer-implemented.

Accordingly, in another aspect of the invention, there is provided a data processing device comprising means for carrying out the methods as described herein.

The data processing device may be a polynucleotide sequencer.

The data processing device may comprise reagents used for synthesis methods as described herein.

The data processing device may comprise a solid support. In a further embodiment, the solid support may be a flow cell.

In another aspect of the invention, there is provided a computer program product comprising instructions which, when the program is executed by a processor, cause the processor to carry out the methods as described herein.

In another aspect of the invention, there is provided a computer-readable storage medium comprising instructions which, when executed by a processor, cause the processor to carry out the methods as described herein.

In another aspect of the invention, there is provided a computer-readable data carrier having stored thereon the computer program product as described herein.

In another aspect of the invention, there is provided a data carrier signal carrying the computer program product as described herein.

The various illustrative imaging or data processing techniques described in connection with the embodiments disclosed herein can be implemented as electronic hardware, computer software, or combinations of both. To illustrate this interchangeability of hardware and software, various illustrative components, blocks, modules, and steps have been described above generally in terms of their functionality. Whether such functionality is implemented as hardware or software depends upon the particular application and design constraints imposed on the overall system. The described functionality can be implemented in varying ways for each particular application, but such implementation decisions should not be interpreted as causing a departure from the scope of the disclosure.

The various illustrative detection systems described in connection with the embodiments disclosed herein can be implemented or performed by a machine, such as a processor configured with specific instructions, a digital signal processor (DSP), an application specific integrated circuit (ASIC), a field programmable gate array (FPGA) or other programmable logic device, discrete gate or transistor logic, discrete hardware components, or any combination thereof designed to perform the functions described herein. A processor can be a microprocessor, but in the alternative, the processor can be a controller, microcontroller, or state machine, combinations of the same, or the like. A processor can also be implemented as a combination of computing devices, e.g., a combination of a DSP and a microprocessor, a plurality of microprocessors, one or more microprocessors in conjunction with a DSP core, or any other such configuration. For example, systems described herein may be implemented using a discrete memory chip, a portion of memory in a microprocessor, flash, EPROM, or other types of memory.

The elements of a method, process, or algorithm described in connection with the embodiments disclosed herein can be embodied directly in hardware, in a software module executed by a processor, or in a combination of the two. A software module can reside in RAM memory, flash memory, ROM memory, EPROM memory, EEPROM memory, registers, hard disk, a removable disk, a CD-ROM, or any other form of computer-readable storage medium known in the art. An exemplary storage medium can be coupled to the processor such that the processor can read information from, and write information to, the storage medium. In the alternative, the storage

medium can be integral to the processor. The processor and the storage medium can reside in an ASIC. A software module can comprise computer-executable instructions which cause a hardware processor to execute the computer-executable instructions.

Computer-executable instructions may be stored in a (transitory or non-transitory) computer readable storage medium (e.g., memory, storage system, etc.) storing code, or computer readable instructions.

Methods of preparing library hairpin strands

Also described herein is a method of preparing a polynucleotide library hairpin strand, comprising:

(a) providing a double-stranded polynucleotide comprising a precursor forward library strand and a precursor reverse library strand; and

(b) ligating a hairpin loop adaptor to an end of the double-stranded polynucleotide to generate a first hairpin polynucleotide, wherein the hairpin loop adaptor comprises a cleavable site.

Advantageously, since the polynucleotide library hairpin strand comprises a cleavable site within the hairpin loop adaptor, this allows template strands (and template complement strands) to be generated that themselves have first cleavable sites and second cleavable sites as described herein. Accordingly, these polynucleotide library hairpin strand allow the generation of template strands (and template complement strands) that have a reduced risk of forming hairpins during sequencing.

The hairpin loop adaptor may be an oligonucleotide of any structure or any sequence that allows the (precursor) forward library strand and the (precursor) reverse library strand to be connected via a loop.

In one embodiment, the hairpin loop adaptor may connect a 3'-end of the precursor forward library strand with a 5'-end of the precursor reverse library strand. In another embodiment, the hairpin loop adaptor may connect a 3'-end of the precursor reverse library strand with a 5'-end of the precursor forward library strand.

In one embodiment, the hairpin loop adaptor may comprise a base-paired stem and a non-base-paired loop (e.g. a loop structure with unpaired or non-Watson-Crick paired nucleotides) and connects the 3' end of the (precursor) forward library strand with the 5' end of the (precursor) reverse library strand, or the 5' end of the (precursor) forward library strand with the 3' end of the (precursor) reverse library strand.

In one embodiment, the cleavable site may be located in the non-base-paired loop.

In one embodiment, the cleavable site may be a restriction site for an endonuclease. In a further embodiment, the endonuclease may be a type II restriction enzyme as defined herein, or a CRISPR enzyme as defined herein (e.g. a Cas9 paired nickase as defined herein). In an even further embodiment, the restriction enzyme is a type IIP restriction enzyme, a type IIS restriction enzyme, a type IIC restriction enzyme, or a type IIT restriction enzyme. Non-limiting examples of the type II restriction enzyme may include EcoRI, BglII, NotI and FokI; non-limiting examples of the CRISPR enzyme include Cas6, Cas9, Cpf1 (Cas12a), Mad7, CasC2c1, C2C2 and C2c3, and Cas9 paired nickases such as Cas9 D10A and Cas9 H840A.

In one embodiment, the method may further comprise a step of:

(c) removing the precursor reverse library strand from the first hairpin polynucleotide to generate a second hairpin polynucleotide comprising the precursor forward library strand and the hairpin loop adaptor, wherein the hairpin loop adaptor comprises the cleavable site.

Accordingly, in the second hairpin polynucleotide, the precursor reverse library strand may not be present. Since the native library may additionally contain modified cytosines on the precursor reverse library strand, it is advantageous to remove these as these may interfere with the generation of particular dyad patterns depending on the methylation status (e.g. 5-methylcytosine or 5-hydroxymethylcytosine) that is present on the precursor forward library strand.

In one embodiment, the method may further comprise a step of:

(d) forming a resynthesised reverse library strand from the second hairpin polynucleotide to generate a third hairpin polynucleotide, wherein when any cytosine bases are present in the resynthesised reverse library strand, then all such cytosine bases are unmodified cytosine.

Accordingly, in the third hairpin polynucleotide, only the precursor forward library strand contains modified cytosines (e.g. 5-methylcytosine or 5-hydroxymethylcytosine). Since the resynthesised reverse library strand is produced in a way such that all cytosine bases are unmodified cytosine, the "methylation" status of all cytosine bases in the resynthesised reverse library strand is known (i.e. all demethylated).

In one embodiment, the method may further comprise a step of:

(e) exposing the third hairpin polynucleotide to an enzyme configured to convert hemimethylated 5-methylcytosine CpG dyads to fully methylated 5-methylcytosine CpG dyads, but not convert hemimethylated 5-hydroxymethylcytosine dyads, in order to generate a fourth hairpin polynucleotide.

Accordingly, in the fourth hairpin polynucleotide, the methylation status of all CpG dyads is controlled. The resynthesised reverse library strand is converted to a partially methylated reverse library strand in step (e). In particular, where a CpG sequence is present in the precursor forward library strand which contains unmodified cytosine, the corresponding CpG dyad also contains only unmodified cytosine. Where a CpG sequence is present in the precursor forward library strand which contains 5-methylcytosine, the corresponding CpG dyad contains 5-methylcytosine on both the precursor forward library stand and the partially methylated reverse library strand. Where a CpG sequence is present in the precursor forward library strand which contains 5-hydroxymethylcytosine, the corresponding CpG dyad contains 5-hydroxymethylcytosine on the precursor forward library strand, but unmodified cytosine in the partially methylated reverse library strand.

In one example, the enzyme configured to convert hemimethylated 5-methylcytosine CpG dyads to fully methylated 5-methylcytosine CpG dyads, but not convert hemimethylated 5-hydroxymethylcytosine dyads may be a DNA methyltransferase. In a further embodiment, the DNA methyltransferase may be a member of the DNA methyltransferase 1 (DNMT1) family or the DNA methyltransferase 5 (DNMT5) family. Non-limiting examples of the DNA methyltransferase 1 or DNA methyltransferase 5 enzyme include:

DNMT protein	**Non-limiting examples**
DNMT1	UniProt: Q24K09 (SEQ ID NO. 67) UniProt: Q9Z330 (SEQ ID NO. 68) UniProt: P13864 (SEQ ID NO. 69) UniProt: Q92072 (SEQ ID NO. 70)

	UniProt: P26358 (SEQ ID NO. 71) UniProt: Q27746 (SEQ ID NO. 72)
DNMT5	UniProt: J9VI03 (SEQ ID NO. 73)

In one embodiment, the method may further comprise a step of:

(f) exposing the fourth hairpin polynucleotide to a conversion agent configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase which is read as thymine/uracil, or to a conversion agent configured to convert an unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil, in order to generate a fifth hairpin polynucleotide.

Suitable conversion agents have been described herein (under "Sequencing of modified cytosines"). In particular, the fifth hairpin polynucleotide may comprise first dyads, second dyads and third dyads as described herein.

In one embodiment, the method may further comprise a step of ligating a flanking adaptor to an end of the double-stranded polynucleotide away from the hairpin loop adaptor to the third hairpin polynucleotide, the fourth hairpin polynucleotide or the fifth hairpin polynucleotide, wherein the flanking adaptor comprises a primer-binding sequence and a primer-binding complement sequence.

Although Figures 16 and 17 show ligation of the flanking adaptor (P7'/P5 fork adaptor) to the third hairpin polynucleotide (in other words, immediately after step (d) as described herein), the flanking adaptor need not be ligated at this stage. For example, the flanking adaptor may instead be ligated onto the fourth hairpin polynucleotide (in other words, immediately after step (e) as described herein), or the fifth hairpin polynucleotide (in other words, immediately after step (f) as described herein). In a further embodiment, the flanking adaptor may instead be ligated onto the fourth hairpin polynucleotide (in other words, immediately after step (e) as described herein).

Furthermore, whilst Figures 16 and 17 show ligation of the flanking adaptor (P7'/P5 fork adaptor) on the left hand side, the positions of the hairpin loop adaptor and the flanking adaptor may be swapped instead. In other words, the flanking adaptor may instead be ligated onto the right hand side, whilst the hairpin loop adaptor may be ligated to the left hand side.

In addition, whilst Figures 16 and 17 show ligation of a P7'/P5 type fork adaptor, A P5'/P7 type fork adaptor may be used instead, as mentioned above.

In one embodiment, the flanking adaptor may comprise a first and second strand, wherein the first and second strands are base-paired for a portion of their sequence (forming the base-paired stem) and are non-complementary for the remainder of their sequence, for example, P5' and P7 or P7' and P5, which subsequently forms a fork structure, wherein a first arm of the fork structure comprises a primer-binding sequence and the second arm of the fork structure comprises a primer-binding complement sequence. Thus, the flanking adaptor may be a forked adaptor comprising a base-paired stem, a first arm and a second arm.

In one embodiment, the primer-binding sequence may be located on the first arm, and the primer-binding complement sequence may be located on the second arm.

The primer-binding sequence may be capable of binding to a lawn or immobilised primer that is immobilised on the surface of a solid support. For example, the primer-binding sequence may be either P5' (for example, SEQ ID NO. 3 or 6 or a variant or fragment thereof) or P7' (for example, SEQ ID NO. 4 or a variant or fragment thereof). Similarly, the primer-binding complement sequence may be either P5 (for example, SEQ ID NO. 1 or 5 or a variant or fragment thereof) or P7 (for example, SEQ ID NO. 2 or a variant or fragment thereof). If the primer-binding sequence is P5', the primer-binding complement sequence is P7. If the primer-binding sequence is P7', the primer-binding complement sequence is P5.

The hairpin loop adaptor may comprise one or more sequencing primer binding sites (or sequencing primer binding site complements). The sequencing primer binding sites and the sequencing primer binding site complements may allow binding of a sequencing primer.

In the hairpin loop adaptor, the sequencing primer-binding sites may be in the non-base-paired loop or in the base-paired stem.

In one embodiment, the base-paired stem may comprise at least one sequencing primer binding site. In a further embodiment, the sequencing primer-binding site may be in the base-paired stem, and a sequencing primer-binding site complement may be also be in the base-paired stem. In an even further embodiment, the sequencing primer-binding site and sequencing primer-binding site complement may be in the base-paired stem, and the cleavable site may be in the non-base-paired loop.

In another embodiment, the non-base-paired loop may comprise two sequencing primer binding sites. In a further embodiment, non-base-paired loop may comprise two sequencing primer-binding sites, wherein the sequencing primer-binding sites are either side of the cleavable site.

The sequencing primer binding sites are sequencing primer binding sites and indicate the starting point of the sequencing read. During the sequencing process, a sequencing primer anneals (i.e. hybridises) to at least a portion of the sequencing primer binding site on the template strand. The polymerase enzyme binds to this site and incorporates complementary nucleotides base by base into the growing opposite strand.

The sequence of the sequencing primers and the sequence primer binding sites are not material to the methods of the invention, as long as the sequencing primers are able to bind to the sequence primer binding site (or sequencing binding site complement) to enable amplification and sequencing of the regions to be identified.

Also described herein is a polynucleotide library hairpin strand prepared according to a method of preparing a polynucleotide library hairpin strand as described herein.

Any of the methods of preparing polynucleotide library hairpin strands as described herein may be utilised in methods of preparing polynucleotide templates for distinguishing between modified cytosines as described herein.

Additional Notes

The embodiments described herein are exemplary. Modifications, rearrangements, substitute processes, etc. may be made to these embodiments and still be encompassed within the teachings set forth herein. One or more of the steps, processes, or methods described herein may be carried out by one or more processing and/or digital devices, suitably programmed.

Conditional language used herein, such as, among others, “can,” “might,” “may,” “e.g.,” and the like, unless specifically stated otherwise, or otherwise understood within the context as used, is generally intended to convey that certain embodiments include, while other embodiments do not include, certain features, elements and/or states. Thus, such conditional language is not generally intended to imply that features, elements and/or states are in any way required for one or more embodiments or that one or more embodiments necessarily include logic for deciding, with or without author input or prompting, whether these features, elements and/or states are included or are to be performed in any particular embodiment. The terms “comprising,” “including,”

"having," "involving," and the like are synonymous and are used inclusively, in an open-ended fashion, and do not exclude additional elements, features, acts, operations, and so forth. Also, the term "or" is used in its inclusive sense (and not in its exclusive sense) so that when used, for example, to connect a list of elements, the term "or" means one, some, or all of the elements in the list. The term "comprising" may be considered to encompass "consisting".

Disjunctive language such as the phrase "at least one of X, Y or Z," unless specifically stated otherwise, is otherwise understood with the context as used in general to present that an item, term, etc., may be either X, Y or Z, or any combination thereof (e.g., X, Y and/or Z). Thus, such disjunctive language is not generally intended to, and should not, imply that certain embodiments require at least one of X, at least one of Y or at least one of Z to each be present.

The terms "about" or "approximate" and the like are synonymous and are used to indicate that the value modified by the term has an understood range associated with it, where the range can be ±20%, ±15%, ±10%, ±5%, or ±1%. The term "substantially" is used to indicate that a result (e.g., measurement value) is close to a targeted value, where close can mean, for example, the result is within 80% of the value, within 90% of the value, within 95% of the value, or within 99% of the value. The term "partially" is used to indicate that an effect is only in part or to a limited extent.

Unless otherwise explicitly stated, articles such as "a" or "an" should generally be interpreted to include one or more described items. Accordingly, phrases such as "a device configured to" or "a device to" are intended to include one or more recited devices. Such one or more recited devices can also be collectively configured to carry out the stated recitations. For example, "a processor to carry out recitations A, B and C" can include a first processor configured to carry out recitation A working in conjunction with a second processor configured to carry out recitations B and C.

While the above detailed description has shown, described, and pointed out novel features as applied to illustrative embodiments, it will be understood that various omissions, substitutions, and changes in the form and details of the devices or algorithms illustrated can be made without departing from the spirit of the disclosure. As will be recognized, certain embodiments described herein can be embodied within a form that does not provide all of the features and benefits set forth herein, as some features can be used or practiced separately from others. All changes which come within the meaning and range of equivalency of the claims are to be embraced within their scope.

It should be appreciated that all combinations of the foregoing concepts (provided such concepts are not mutually inconsistent) are contemplated as being part of the inventive subject matter disclosed herein. In particular, all combinations of claimed subject matter appearing at the end of this disclosure are contemplated as being part of the inventive subject matter disclosed herein.

5

Reference Examples

Reference Examples 1 to 4 – Purification and deaminase activity of mutant APOBEC3A proteins

*Swa*I assay method

This assay was adapted and modified from Schutsky et al., Nucleic Acid Research, 45, 7655-7665, 2017. doi: 10.1093/nar/gkx345. Modifications to Schutsky et al. included the following. Instead of performing DNA precipitation and redissolving the DNA substrate into SwaI compatible buffer, 1 µL of the altered cytidine deaminase APOBEC3A(Y130A) deamination reaction mixture was aliquoted into 9 µL *Swa*I compatible buffer for restriction enzyme digestion for our gel assay. Appropriate controls were performed to determine the *Swa*I restriction enzyme digestion efficiency was not compromised by the APOBEC reaction buffer. Instead of introducing 1.5-fold excess complementary strand prior to overnight *Swa*I restriction enzyme digestion, 3-fold excess complementary strand was introduced. Instead of running the pre-heated 20% acrylamide/Tris-Borate-EDTA (TBE)/urea gel reported by Schutsky et al., the gel run was performed at room temperature with a 15% acrylamide/Tris-Borate-EDTA (TBE)/urea gel and observed good resolution between cut (deaminated) and uncut (unreacted) oligo substrates.

Reference Example 1 – Purification of APOBEC3A(Y130X) mutant proteins

The impact of all possible amino acid substitutions at position 130 of APOBEC3A on the deaminase activity of this enzyme was systematically assessed. To this end, 19 different His-tagged APOBEC3A constructs were cloned, each encoding a different amino acid at position 130 relative to the wild type tyrosine. The corresponding proteins were expressed in BL21(DE3) cells, purified using Ni-NTA agarose beads, and desalted/concentrated using spin columns to storage buffer (50mM Tris pH 7.5, 200mM NaCl, 5%(v/v) glycerol, 0.01% (v/v) Tween-20, 0.5mM DTT). This yielded APOBEC3A(Y130X) mutant protein preparations with 80-85% purity, as judged by SDS-PAGE analysis.

Reference Example 2 – DNA deaminase activity of APOBEC3A(Y130X) mutant proteins

The deaminase activity of all purified APOBEC3A(Y130X) proteins was then analysed using the *Swa*I assay, with a 37°C/2 hour reaction time and NEB APOBEC3A as positive control. 10-20 µM final concentration of Y130X recombinant enzymes were incubated with oLB1609 (C oligo, top panel) and oLB1612 (5mC oligo, bottom panel) at 37°C for 2 hours. NEB APOBEC3A enzyme was purchased from NEBNext® Enzymatic Methyl-seq Kit (Catalog # E7120). Wild type

APOBEC3A deaminated 5mC and C substrates to completion, consistent with previous literature. Different mutants exhibited a wide range of reactivities towards 5mC and C substrates, with some showing preference towards either substrate. Remarkably, APOBEC3A(Y130A) (first box) deaminated 5mC substrates almost completely (94.2%), while it deaminated the corresponding C substrate to a minor extent (29.4%). Other mutants, such as APOBEC3A(Y130P) and APOBEC3A(Y130T), also exhibited more complete deamination of the 5mC than C substrate, albeit to a lesser extent than APOBEC3A(Y130A). In contrast, APOBEC3A(Y130L) (second box) deaminated approximately half of the C substrate (56%), but almost none of the 5mC substrate (6.8%). The deaminase activity of all APOBEC3A(Y130X) mutants is quantified and summarised in the table below:

Protein	% C deamination	% 5mC deamination
NEB APOBEC	92	94.9
Y130A	**29.4**	**94.2**
Y130G	3.7	19.1
Y130L	**56**	**6.8**
Y130F	84.6	94.3
Y130I	8.1	8.1
Y130H	83.4	95.8
Y130Q	1.6	14.6
Y130M	37	55.4
Y130N	3.6	9.1
Y130K	0.3	0.9

Protein	% C deamination	% 5mC deamination
NEB APOBEC	99.3	95.3
Y130V	11.6	15
Y130D	0.8	2.3
Y130E	0	2.9
Y130S	53.5	95.4
Y130C	88.2	96.2
Y130W	97.6	71.6
Y130P	0.2	22.3
Y130R	0.6	0.8
Y130T	8	28.1

Because these *Swa*I assays were performed as a single endpoint measurement (2 hour), it could be possible that the respective deamination reactions had already saturated. A time course analysis of APOBEC3A(Y130A) deaminase activity was therefore performed. The extent of C and 5mC deamination was monitored at 0, 5, 10, 30, 60 and 120 minutes by incubation of ~10-20 μM of APOBEC(Y130A) with 500nM C and 5mC oligonucleotide substrate. A greater difference in the extent of 5mC versus C deamination was observed at $t \leq 30$ min.

The kinetics of deamination by wild type APOBEC3A and mutant APOBEC3A(Y130A) were quantitatively compared. The initial deamination reaction velocity was measured at a range of DNA substrate concentrations and used to construct Michaelis-Menten curves for 5mC and C substrates, respectively. The resulting Km and Kcat values were then derived from these data. The catalytic efficiency of APOBEC3A(Y130A) was ~100-fold higher on 5mC than C substrates corroborating the endpoint *Swa*I assays shown above.

Reference Example 3 – Purification of APOBEC3A(Y130A-Y132H) double mutant protein

APOBEC3A(Y130A-Y132H) protein was expressed in BL21(DE3) cells, purified using Ni-NTA agarose beads, and desalted/concentrated using spin columns to storage buffer (50mM Tris pH 7.5, 200mM NaCl, 5%(v/v) glycerol, 0.01% (v/v) Tween-20, 0.5mM DTT). This yielded APOBEC3A(Y130A-Y132H) mutant protein preparations with 90-95% purity, as judged by SDS-PAGE analysis.

Reference Example 4 – DNA deaminase activity of APOBEC3A(Y130A-Y132H) double mutant protein

The deaminase activity of purified APOBEC3A(Y130A-Y132H) double mutant protein was then analyzed using the *Swa*I assay, with a 37°C/ 2 hour reaction time and NEB APOBEC3A as positive control. The conditions used were the same as described in Reference Example 2 with the exception that the *Swa*I assay used reaction conditions of 40 mM sodium acetate pH 5.2, 37°C for 1 hour to 16 hours. The DNA substrates are shown below:

5'GAGGTGTATGGTTGTACTAAT/5mC/ACT/5mC/CTGGA/5mC/GAATCTTAA/5mC/ACAA/5mC/GTGCAG/5mC/CAAA/5mC/GCTT/5mC/GC/5mC/ACGG/5mC/AACGTG/5mC/GGACT/5mC/GTCG/5mC/CTTA/5mC/AATCG/5mC/GCAGGT/5mC/ACGTTGAAGATGAGGATG-3' (SEQ ID NO: 74)

GAGGTGTATGGTTGTAG/5mC/GCAAATCGTAAAA/5mC/GCAAAGCGAAAAC/5mC/GCAAACCGTAAAC/5mC/GAAAAGCGCTTGAAGATGAGGATG (SEQ ID NO: 75)

GAGGTGTATGGTTGTAG/5mC/GGAAAACGGAAAT/5mC/GGAAAACGTAAAG/5mC/GTAAATCGGAAAG/5mC/GAAAAGCGGTTGAAGATGAGGATG (SEQ ID NO: 76)

GAGGTGTATGGTTGTAA/5mC/GTAAACCGCAAAC/5mC/GGAAAACGAAAAT/5mC/GCAAACCGAAAAC/5mC/GTAAAACGCTTGAAGATGAGGATG (SEQ ID NO: 77)

GAGGTGTATGGTTGTAA/5mC/GAAAACCGGAAAT/5mC/GAAAAGCGTAAAT/5mC/GTAAATCGCAAAA/5mC/GGAAATCGATTGAAGATGAGGATG (SEQ ID NO: 78)

After the deaminase reaction the deaminated oligo substrates were PCR-amplified, sequenced, and the number of C and 5mC deamination events per read were counted. APOBEC3A(Y130A-Y132H) exhibited higher levels of deamination at all methylated sites compared to unmethylated sites. This was consistent across both CpG and non-CpG contexts, and was robust to variation in reaction time. The difference in deamination level between methylated and unmethylated sites

was markedly higher for APOBEC3A(Y130A-Y132H) than APOBEC3A(Y130A), indicating that APOBEC3A(Y130A-Y132H) achieves better discrimination of methylated sites than APOBEC3A(Y130A). In addition, APOBEC3A(Y130A-Y132H) deaminated methylated sites more efficiently than unmethylated sites across all xCpGx motifs.

Embodiments are set out in the following clauses:

Clause 1. A method of preparing polynucleotide templates for distinguishing between modified cytosines, comprising:

(a) providing a polynucleotide library hairpin strand comprising:

a double-stranded polynucleotide comprising a forward library strand and a reverse library strand,

a hairpin loop adaptor ligated to an end of the double-stranded polynucleotide, wherein the hairpin loop adaptor comprises a cleavable site,

wherein the polynucleotide library hairpin strand has been generated from a precursor polynucleotide library hairpin strand such that any CpG dyads in the precursor polynucleotide library hairpin comprising only unmodified cytosine are converted to a first dyad in the polynucleotide library hairpin strand, any CpG dyads in the precursor polynucleotide library hairpin comprising 5-methylcytosine are converted to a second dyad in the polynucleotide library hairpin strand, and any CpG dyads in the precursor polynucleotide library hairpin comprising 5-hydroxymethylcytosine are converted to a third dyad in the polynucleotide library hairpin strand,

wherein the first dyad, second dyad and third dyad are different to each other when read; and

(b) synthesising at least one template strand by generating a complement of the polynucleotide library hairpin strand, each of the template strands comprising a forward template strand complementary to the forward library strand, a spacer strand complementary to the hairpin loop adaptor, and a reverse template strand complementary to the reverse library strand, wherein the spacer strand comprises a first cleavable site.

Clause 2. A method according to clause 1, wherein the method further comprises a step of:

(c) synthesising at least one template complement strand by generating a complement of the template strand, each of the template complement strands comprising

a forward complement template strand, a spacer complement strand, and a reverse complement template strand, wherein the spacer complement strand comprises a second cleavable site.

Clause 3. A method according to clause 2, wherein the method further comprises a step of:

(d) cleaving the first cleavable site on the at least one template strand to generate at least one first polynucleotide sequence each comprising a first portion and cleaving the second cleavable site on the at least one template complement strand to generate at least one second polynucleotide sequence each comprising a second portion,

wherein the first portion corresponds with the forward template strand and the second portion corresponds with the reverse complement template strand, or wherein the first portion corresponds with the reverse template strand and the second portion corresponds with the forward complement template strand.

Clause 4. A method according to any one of clauses 1 to 3, wherein the first cleavable site is a first restriction site for an endonuclease.

Clause 5. A method according to any one of clauses 2 to 4, wherein the second cleavable site is a second restriction site for an endonuclease.

Clause 6. A method according to any one of clauses 3 to 5, wherein the at least one first polynucleotide sequence each comprise a first sequencing primer binding site.

Clause 7. A method according to clause 6, wherein the first sequencing primer binding site is located after a 3'-end of the first portion.

Clause 8. A method according to any one of clauses 3 to 7, wherein the at least one second polynucleotide sequence each comprise a second sequencing primer binding site.

Clause 9. A method according to clause 8, wherein the second sequencing primer binding site is located after a 3'-end of the second portion.

Clause 10. A method according to any one of clauses 3 to 9, wherein where CpG dyads comprising only unmodified cytosine were present in the precursor polynucleotide library hairpin, then a double C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; where CpG dyads comprising 5-methylcytosine were present in the precursor

polynucleotide library hairpin, then a double mismatch is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; and where CpG dyads comprising 5-hydroxymethylcytosine were present in the precursor polynucleotide library hairpin, then a single mismatch and single C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad.

Clause 11. A method according to any one of clauses 3 to 9, wherein where CpG dyads comprising only unmodified cytosine were present in the precursor polynucleotide library hairpin, then a double mismatch is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; where CpG dyads comprising 5-methylcytosine were present in the precursor polynucleotide library hairpin, then a double C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; and where CpG dyads comprising 5-hydroxymethylcytosine were present in the precursor polynucleotide library hairpin, then a single mismatch and single C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad.

Clause 12. A method according to any one of clauses 3 to 11, wherein the method further comprises a step of preparing the first portion and the second portion for concurrent sequencing.

Clause 13. A method according to clause 12, wherein the method comprises simultaneously contacting first sequencing primer binding sites located after a 3'-end of the first portions with first primers and second sequencing primer binding sites located after a 3'-end of the second portions with second primers.

Clause 14. A method according to any one of clauses 3 to 13, wherein the method further comprises a step of processing the at least one first polynucleotide sequence comprising a first portion and the at least one second polynucleotide sequence comprising a second portion, such that a proportion of first portions are capable of generating a first signal and a proportion of second portions are capable of generating a second signal.

Clause 15. A method according to clause 14, wherein the processing involves selective processing to cause an intensity of the first signal to be greater than an intensity of the second signal.

Clause 16. A method according to clause 15, wherein a concentration of the first portions capable of generating the first signal is greater than a concentration of the second portions capable of generating the second signal.

Clause 17. A method according to clause 16, wherein a ratio between the concentration of the first portions capable of generating the first signal and the concentration of the second portions capable of generating the second signal is between 1.25:1 to 5:1.

Clause 18. A method according to clause 17, wherein the ratio is between 1.5:1 to 3:1.

Clause 19. A method according to clause 18, wherein the ratio is about 2:1.

Clause 20. A method according to any one of clauses 15 to 19, wherein selective processing comprises preparing for selective sequencing or conducting selective sequencing.

Clause 21. A method according to any one of clauses 15 to 19, wherein selectively processing comprises conducting selective amplification.

Clause 22. A method according to any one of clauses 15 to 20, wherein selectively processing comprises contacting first sequencing primer binding sites located after a 3'-end of the first portions with first primers and contacting second sequencing primer binding sites located after a 3'-end of the second portions with second primers, wherein the second primers comprises a mixture of blocked second primers and unblocked second primers.

Clause 23. A method according to clause 22, wherein the blocked second primer comprises a blocking group at a 3' end of the blocked second primer.

Clause 24. A method according to clause 23, wherein the blocking group is selected from the group consisting of: a hairpin loop, a deoxynucleotide, a deoxyribonucleotide, a hydrogen atom instead of a 3'-OH group, a phosphate group, a phosphorothioate group, a propyl spacer, a modification blocking the 3'-hydroxyl group, or an inverted nucleobase.

Clause 25. A method according to any one of clauses 15 to 19 or 21, wherein the selective processing comprises selectively removing some or substantially all of second immobilised primers that are not yet extended, and conducting a further amplification cycle in order to

selectively amplify the first polynucleotide sequence(s) relative to the second polynucleotide sequence(s).

Clause 26. A method according to any one of clauses 15 to 19 or 21, wherein selectively processing comprises selectively blocking some or substantially all of second immobilised primers that are not yet extended using a primer blocking agent, wherein the primer blocking agent is configured to limit or prevent synthesis of a strand extending from the second immobilised primer, and conducting a further amplification cycle in order to selectively amplify the first polynucleotide sequence(s) relative to the second polynucleotide sequence(s).

Clause 27. A method according to clause 26, wherein the primer blocking agent is added whilst first polynucleotide sequence(s) are hybridised to the second immobilised primers.

Clause 28. A method according to clause 26, wherein the method comprises contacting some or substantially all of the second immobilised primers with an extended primer sequence, wherein the extended primer sequence is substantially complementary to the second immobilised primer and further comprises a 5' additional nucleotide; and adding the primer blocking agent, wherein the primer blocking agent is complementary to the 5' additional nucleotide.

Clause 29. A method according to any one of clauses 26 to 28, wherein the primer blocking agent is a blocked nucleotide.

Clause 30. A method according to clause 29, wherein the blocked nucleotide comprises a blocking group at a 3' end of the blocked nucleotide.

Clause 31. A method according to clause 30, wherein the blocking group is selected from the group consisting of: a hairpin loop, a deoxynucleotide, a deoxyribonucleotide, a hydrogen atom instead of a 3'-OH group, a phosphate group, a phosphorothioate group, a propyl spacer, a modification blocking the 3'-hydroxyl group, or an inverted nucleobase.

Clause 32. A method according to any one of clauses 29 to 31, wherein the blocked nucleotide is A or G.

Clause 33. A method according to any one of clauses 14 to 32, wherein the first signal and the second signal are spatially resolved.

Clause 34. A method according to any one of clauses 14 to 32, wherein the first signal and the second signal are spatially unresolved.

Clause 35. A method according to any one of clauses 3 to 34, wherein the at least one first polynucleotide sequence comprising the first portion and the at least one second polynucleotide sequence comprising the second portion are attached to a solid support.

Clause 36. A method according to clause 35, wherein the solid support is a flow cell.

Clause 37. A method according to clause 35 or clause 36, wherein the at least one first polynucleotide sequence comprising the first portion and the at least one second polynucleotide sequence comprising the second portion forms a cluster on the solid support.

Clause 38. A method according to clause 37, wherein the cluster is formed by bridge amplification.

Clause 39. A method according to any one of clauses 35 to 38, wherein the at least one first polynucleotide sequence comprising the first portion and the at least one second polynucleotide sequence comprising the second portion form a duoclonal cluster.

Clause 40. A method according to any one of clauses 35 to 39, wherein the solid support comprises at least one first immobilised primer and at least one second immobilised primer.

Clause 41. A method according to clause 40, wherein the first immobilised primer comprises a sequence as defined in SEQ ID NO. 1 or 5, or a variant or fragment thereof; and the second immobilised primer comprises a sequence as defined in SEQ ID NO. 2, or a variant or fragment thereof.

Clause 42. A method according to clause 40 or clause 41, wherein each first polynucleotide sequence is attached to a first immobilised primer, and wherein each second polynucleotide sequence is attached to a second immobilised primer.

Clause 43. A method according to any one of clauses 40 to 42, wherein each first polynucleotide sequence comprises a second adaptor sequence and wherein each second polynucleotide sequence comprises a first adaptor sequence, wherein the second adaptor sequence is substantially complementary to the second immobilised primer and wherein the first adaptor sequence is substantially complementary to the first immobilised primer.

Clause 44. A method of sequencing polynucleotide sequences to distinguish between modified cytosines, comprising:

preparing polynucleotide templates for distinguishing between modified cytosines using a method according to any one of clauses 3 to 43;

sequencing nucleobases in the first portion and the second portion; and

identifying the presence of 5-methylcytosine or 5-hydroxymethylcytosine by detecting differences when comparing a sequence output from the first portion with a sequence output from the second portion.

Clause 45. A method according to clause 44, wherein the step of sequencing nucleobases in the first portion and the second portion involves concurrent sequencing of nucleobases in the first portion and the second portion.

Clause 46. A method according to clause 44 or clause 45, wherein the step of sequencing nucleobases comprises performing sequencing-by-synthesis.

Clause 47. A method according to any one of clauses 44 to 46, wherein the method further comprises a step of conducting paired-end reads.

Clause 48. A kit comprising instructions for preparing polynucleotide templates for distinguishing between modified cytosines according to any one of clauses 1 to 43, and/or for sequencing polynucleotide sequences to distinguish between modified cytosines according to any one of clauses 44 to 47.

Clause 49. A data processing device comprising means for carrying out a method according to any one of clauses 1 to 47.

Clause 50. A data processing device according to clause 49, wherein the data processing device is a polynucleotide sequencer.

Clause 51. A computer program product comprising instructions which, when the program is executed by a processor, cause the processor to carry out a method according to any one of clauses 1 to 47.

Clause 52. A computer-readable storage medium comprising instructions which, when executed by a processor, cause the processor to carry out a method according to any one of clauses 1 to 47.

Clause 53. A computer-readable data carrier having stored thereon a computer program product according to clause 51.

Clause 54. A data carrier signal carrying a computer program product according to clause 51.

Clause 55. A method of preparing a polynucleotide library hairpin strand, comprising:

(a) providing a double-stranded polynucleotide comprising a precursor forward library strand and a precursor reverse library strand; and

(b) ligating a hairpin loop adaptor to an end of the double-stranded polynucleotide to generate a first hairpin polynucleotide, wherein the hairpin loop adaptor comprises a cleavable site.

Clause 56. A method according to clause 55, wherein the hairpin loop adaptor comprises a base-paired stem and a non-base-paired loop.

Clause 57. A method according to clause 55 or clause 56, wherein the cleavable site is located in the non-base-paired loop.

Clause 58. A method according to any one of clauses 55 to 57, wherein the hairpin loop adaptor connects a 3'-end of the precursor forward library strand with a 5'-end of the precursor reverse library strand; or wherein the hairpin loop adaptor connects a 3'-end of the precursor reverse library strand with a 5'-end of the precursor forward library strand.

Clause 59. A method according to any one of clauses 55 to 58, wherein the cleavable site is a restriction site for an endonuclease.

Clause 60. A method according to any one of clauses 55 to 59, wherein the method further comprises a step of:

(c) removing the precursor reverse library strand from the first hairpin polynucleotide to generate a second hairpin polynucleotide comprising the precursor forward library strand and the hairpin loop adaptor, wherein the hairpin loop adaptor comprises the cleavable site.

Clause 61. A method according to clause 60, wherein the method further comprises a step of:

(d) forming a resynthesised reverse library strand from the second hairpin polynucleotide to generate a third hairpin polynucleotide, wherein when any cytosine bases are present in the resynthesised reverse library strand, then all such cytosine bases are unmodified cytosine.

Clause 62. A method according to clause 61, wherein the method further comprises a step of:

(e) exposing the third hairpin polynucleotide to an enzyme configured to convert hemimethylated 5-methylcytosine CpG dyads to fully methylated 5-methylcytosine CpG dyads, but not convert hemimethylated 5-hydroxymethylcytosine dyads, in order to generate a fourth hairpin polynucleotide.

Clause 63. A method according to clause 62, wherein the enzyme configured to convert hemimethylated 5-methylcytosine CpG dyads to fully methylated 5-methylcytosine CpG dyads, but not convert hemimethylated 5-hydroxymethylcytosine dyads is a DNA methyltransferase.

Clause 64. A method according to clause 63, wherein the enzyme configured to convert hemimethylated 5-methylcytosine CpG dyads to fully methylated 5-methylcytosine CpG dyads, but not convert hemimethylated 5-hydroxymethylcytosine dyads is a member of the DNA methyltransferase 1 (DNMT1) family or the DNA methyltransferase 5 (DNMT5) family.

Clause 65. A method according to any one of clauses 62 to 64, wherein the method further comprises a step of:

(f) exposing the fourth hairpin polynucleotide to a conversion agent configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase which is read as thymine/uracil, or to a conversion agent configured to convert an unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil, in order to generate a fifth hairpin polynucleotide.

Clause 66. A method according to clause 65, wherein the conversion agent is configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase which is read as thymine/uracil.

Clause 67. A method according to clause 65, wherein the conversion agent is configured to convert unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil.

Clause 68. A method according to any one of clauses 65 to 67, wherein the conversion agent comprises a chemical agent and/or an enzyme.

Clause 69. A method according to clause 68, wherein the conversion agent comprises a boron-based reducing agent and a ten-eleven translocation (TET) methylcytosine dioxygenase.

Clause 70. A method according to clause 69, wherein the boron-based reducing agent is an amine-borane compound or an azine-borane compound.

Clause 71. A method according to clause 69 or clause 70, wherein the boron-based reducing agent is selected from the group consisting of pyridine borane, 2-picoline borane, t-butylamine borane, ammonia borane, ethylenediamine borane and dimethylamine borane.

Clause 72. A method according to any one of clauses 69 to 71, wherein the TET methylcytosine dioxygenase is a member of the TET1 subfamily, the TET2 subfamily, or the TET3 subfamily.

Clause 73. A method according to clause 68, wherein the conversion agent comprises sulfite.

Clause 74. A method according to clause 73, wherein the sulfite is bisulfite.

Clause 75. A method according to clause 74, wherein the bisulfite is sodium bisulfite.

Clause 76. A method according to clause 68, wherein the conversion agent comprises a cytidine deaminase.

Clause 77. A method according to clause 76, wherein the cytidine deaminase is a wild-type cytidine deaminase or a mutant cytidine deaminase.

Clause 78. A method according to clause 76 or clause 77, wherein the cytidine deaminase is a member of the AID subfamily, the APOBEC1 subfamily, the APOBEC2 subfamily, the APOBEC3A subfamily, the APOBEC3B subfamily, the APOBEC3C subfamily, the APOBEC3D subfamily, the APOBEC3F subfamily, the APOBEC3G subfamily, the APOBEC3H subfamily, or the APOBEC4 subfamily.

Clause 79. A method according to clause 78, wherein the cytidine deaminase is a member of the APOBEC3A subfamily.

Clause 80. A method according to any one of clauses 76 to 79, wherein the cytidine deaminase comprises amino acid substitution mutations at positions functionally equivalent to (Tyr/Phe)130 and Tyr132 in a wild-type APOBEC3A protein.

Clause 81. A method according to clause 80, wherein the (Tyr/Phe)130 is Tyr130, and the wild-type APOBEC3A protein is SEQ ID NO. 16.

Clause 82. A method according to clause 80 or clause 81, wherein the substitution mutation at the position functionally equivalent to Tyr130 comprises Ala, Val or Trp.

Clause 83. A method according to clause 80, wherein the substitution mutation at the position functionally equivalent to Tyr132 comprises a mutation to His, Arg, Gln or Lys.

Clause 84. A method according to any one of clauses 77 to 83, wherein the mutant cytidine deaminase comprises a ZDD motif H-[P/A/V]-E-$X_{[23-28]}$-P-C-$X_{[2-4]}$-C (SEQ ID NO. 51).

Clause 85. A method according to any one of clauses 77 to 83, wherein the mutant cytidine deaminase is a member of the APOBEC3A subfamily and comprises a ZDD motif $HXEX_{24}SW(S/T)PCX_{[2-4]}CX_6FX_8LX_5R(L/I)YX_{[8-11]}LX_2LX_{[10]}M$ (SEQ ID NO. 52).

Clause 86. A method according to any one of clauses 77 to 85, wherein the mutant cytidine deaminase converts 5-methylcytosine to thymine by deamination at a greater rate than conversion rate of cytosine to uracil by deamination.

Clause 87. A method according to clause 86, wherein the rate is at least 100-fold greater.

Clause 88. A method according to any one of clauses 65 to 87, wherein the conversion agent further comprises a glycosyltransferase.

Clause 89. A method according to clause 88, wherein the glycosyltransferase is a β-glucosyltransferase.

Clause 90. A method according to any one of clauses 61 to 89, wherein the method further comprises a step of ligating a flanking adaptor to an end of the double-stranded polynucleotide away from the hairpin loop adaptor to the third hairpin polynucleotide, the fourth hairpin polynucleotide or the fifth hairpin polynucleotide, wherein the flanking adaptor comprises a primer-binding sequence and a primer-binding complement sequence.

Clause 91. A method according to clause 90, wherein the flanking adaptor is a forked adaptor comprising a base-paired stem, a first arm and a second arm.

Clause 92. A method according to clause 90 or clause 91, wherein the primer-binding sequence is located on the first arm, and the primer-binding complement sequence is located on the second arm.

Clause 93. A polynucleotide library hairpin strand prepared according to any one of clauses 55 to 92.

SEQUENCE LISTING

SEQ ID NO. 1: P5 sequence

AATGATACGGCGACCACCGAGATCTACAC

SEQ ID NO. 2: P7 sequence

CAAGCAGAAGACGGCATACGAGAT

SEQ ID NO. 3: P5' sequence (complementary to P5)

GTGTAGATCTCGGTGGTCGCCGTATCATT

SEQ ID NO. 4: P7' sequence (complementary to P7)

ATCTCGTATGCCGTCTTCTGCTTG

SEQ ID NO. 5: Alternative P5 sequence

AATGATACGGCGACCGA

SEQ ID NO. 6: Alternative P5' sequence (complementary to alternative P5 sequence)

TCGGTCGCCGTATCATT

SEQ ID NO. 7: UniProt Q9GZX7

MDSLLMNRRKFLYQFKNVRWAKGRRETYLCYVVKRRDSATSFSLDFGYLRNKNGCHVELLFLRY
ISDWDLDPGRCYRVTWFTSWSPCYDCARHVADFLRGNPNLSLRIFTARLYFCEDRKAEPEGLRR
LHRAGVQIAIMTFKDYFYCWNTFVENHERTFKAWEGLHENSVRLSRQLRRILLPLYEVDDLRDA
FRTLGL

SEQ ID NO. 8: UniProt G3QLD2

MDSLLMNRRKFLYQFKNVRWAKGRRETYLCYVVKRRDSATSFSLDFGYLRNKNGCHVELLFLRY
ISDWDLDPGRCYRVTWFTSWSPCYDCARHVADFLRGNPNLSLRIFTARLYFCEDRKAEPEGLRR
LHRAGVQIAIMTFKENHERTFKAWEGLHENSVRLSRQLRRILLPLYEVDDLRDAFRTLGL

SEQ ID NO. 9: Uniprot Q9WVE0

MDSLLMKQKKFLYHFKNVRWAKGRHETYLCYVVKRRDSATSCSLDFGHLRNKSGCHVELLFLRY
ISDWDLDPGRCYRVTWFTSWSPCYDCARHVAEFLRWNPNLSLRIFTARLYFCEDRKAEPEGLRR
LHRAGVQIGIMTFKDYFYCWNTFVENRERTFKAWEGLHENSVRLTRQLRRILLPLYEVDDLRDA
FRMLGF

SEQ ID NO. 10: UniProt P41238

MTSEKGPSTGDPTLRRRIEPWEFDVFYDPRELRKEACLLYEIKWGMSRKIWRSSGKNTTNHVEV
NFIKKFTSERDFHPSMSCSITWFLSWSPCWECSQAIREFLSRHPGVTLVIYVARLFWHMDQQNR
QGLRDLVNSGVTIQIMRASEYYHCWRNFVNYPPGDEAHWPQYPPLWMMLYALELHCIILSLPPC
LKISRRWQNHLTFFRLHLQNCHYQTIPPHILLATGLIHPSVAWR

SEQ ID NO. 11: NCBI XP_030856728.1

MSRKIWRSSGKNTTNHVEVNFIKKFTSERHFHPSISCSITWFLSWSPCWECSQAIREFLSQHPG
VTLVIYVARLFWHMDQQNRQGLRDLVNSGVTIQIMRASEYYHCWRNFVNYPPGDEAHWPQYPPL
WMMLYALELHCIILSLPPCLKISRRWQNHLTFFRLHLQNCHYQTIPPHILLATGLIHPSVAWR

SEQ ID NO. 12: Uniprot P51908

MSSETGPVAVDPTLRRRIEPHEFEVFFDPRELRKETCLLYEINWGGRHSVWRHTSQNTSNHVEV
NFLEKFTTERYFRPNTRCSITWFLSWSPCGECSRAITEFLSRHPYVTLFIYIARLYHHTDQRNR
QGLRDLISSGVTIQIMTEQEYCYCWRNFVNYPPSNEAYWPRYPHLWVKLYVLELYCIILGLPPC
LKILRRKQPQLTFFTITLQTCHYQRIPPHLLWATGLK

SEQ ID NO. 13: UniProt Q9Y235

MAQKEEAAVATEAASQNGEDLENLDDPEKLKELIELPPFEIVTGERLPANFFKFQFRNVEYSSG
RNKTFLCYVVEAQGKGGQVQASRGYLEDEHAAAHAEEAFFNTILPAFDPALRYNVTWYVSSSPC
AACADRIIKTLSKTKNLRLLILVGRLFMWEEPEIQAALKKLKEAGCKLRIMKPQDFEYVWQNFV
EQEEGESKAFQPWEDIQENFLYYEEKLADILK

SEQ ID NO. 14: Uniprot G3SGN8

MAQKEEAAAATEAAAATEAASQNGEDLENLDDPEKLKELIELPPFEIVTGERLPANFFKFQFRN
VEYSSGRNKTFLCYVVEAQGKGGQVQASRGYLEDEHAAAHAEEAFFNTILPAFDPALRYNVTWY
VSSSPCAACADRIIKTLSKTKNLRLLILVGRLFMWEEPEIQAALKKLKEAGCKLRIMKPQDFEY
VWQNFVEQEEGESKAFQPWEDIQENFLYYEEKLADILK

SEQ ID NO. 15: Uniprot Q9WV35

MAQKEEAAEAAAPASQNGDDLENLEDPEKLKELIDLPPFEIVTGVRLPVNFFKFQFRNVEYSSG
RNKTFLCYVVEVQSKGGQAQATQGYLEDEHAGAHAEEAFFNTILPAFDPALKYNVTWYVSSSPC
AACADRILKTLSKTKNLRLLILVSRLFMWEEPEVQAALKKLKEAGCKLRIMKPQDFEYIWQNFV
EQEEGESKAFEPWEDIQENFLYYEEKLADILK

SEQ ID NO. 16: UniProt P31941

MEASPASGPRHLMDPHIFTSNFNNGIGRHKTYLCYEVERLDNGTSVKMDQHRGFLHNQAKNLLC
GFYGRHAELRFLDLVPSLQLDPAQIYRVTWFISWSPCFSWGCAGEVRAFLQENTHVRLRIFAAR
IYDYDPLYKEALQMLRDAGAQVSIMTYDEFKHCWDTFVDHQGCPFQPWDGLDEHSQALSGRLRA
ILQNQGN

SEQ ID NO. 17: GenBank XP_045219544.1

MDGSPASRPRHLMDPNTFTFNFNNDLSVRGRHQTYLCYEVERLDNGTWVPMDERRGFLHNKAKN
VPCGDYGCHVELRFLCEVPSWQLDPAQTYRVTWFISWSPCFRKGCAGQVRAFLQENKHVRLRIF
AARIYDYDPRYQEALRTLRDAGAQVSIMTYEEFKHCWDTFVDRQGRPFQPWDGLDEHSQALSGR
LRAILQNQGN

SEQ ID NO. 18: GenBank AER45717.1

MEASPASGPRHLMDPCVFTSNFNNGIRWHKTYLCYEVERLDNGTWVKMDQHRGFLHNQARNPLY
GLDGRHAELRFLGLLPYWQLDPAQIYRVTWFISWSPCFSWGCARQVRAFLQENTHVRLRIFAAR
IYDYDPLYKEALQMLRDAGAQVSIMTYDEFEYCWNTFVDHQGCPFQPWDGLEEHSQALSGKLQA
ILLNQGN

SEQ ID NO. 19: GenBank XP_003264816.1

MEASPASRPGHLMDPQVFTSNFNNGIRWHKTYLCYEVERLDNGTWVKMDQHRGFLHNQAKNLFC
GFYGRHAELCFLDLVPSLQLDPAQTYRVTWFISWSPCFSWGCAEQVRAFLQENTHVRLRLFAAR
IYDYDPLYKEALQMLRGAGAQVSIMTYHEFKHCWDTFVDHQGRPFQPWDGLEEHSQALSGRLQA
ILQNQGN

SEQ ID NO. 20: GenBank PNI48846.1

TEASPASGPRHLMDPHIFTSNFNNGIGRRKTYLCYEVERLDNGTSVKMDQHRGFLHNQAKNLLC
GFYGRHAELCFLDLVPSLQLDPAQIYRVTWFISWSPCFSWGCAGQVRAFLQENTHVRLRIFAAR
IYDYDPLYKEALQMLRDAGAQVSIMTYDEFKHCWDTFVDHQGCPFQPWDGLEEHSQALSGRLRA
ILQNQGN

SEQ ID NO. 21: GenBank ADO85886.1

VEASPASGPRHLMDPHIFTSNFNNVIGRHKTYLCYEVERLDNGTWVKMDQHRGFLHNQAKNLLC
GFYGRHAELRFLDLVPSLQLDPAQIYRVTWFISWSPCFSWGCAGQVRAFLQENTHVRLHIFAAR
IYDYDPLYKEALQMLRDAGAQVSIMTYDEFKHCWDTFVDHQGCPFQPWDGLEEHSQALSGRLRA
ILQNQGN

SEQ ID NO. 22: UniProt Q9UH17

MNPQIRNPMERMYRDTFYDNFENEPILYGRSYTWLCYEVKIKRGRSNLLWDTGVFRGQVYFKPQ
YHAEMCFLSWFCGNQLPAYKCFQITWFVSWTPCPDCVAKLAEFLSEHPNVTLTISAARLYYYWE
RDYRRALCRLSQAGARVTIMDYEEFAYCWENFVYNEGQQFMPWYKFDENYAFLHRTLKEILRYL
MDPDTFTFNFNNDPLVLRRRQTYLCYEVERLDNGTWVLMDQHMGFLCNEAKNLLCGFYGRHAEL
RFLDLVPSLQLDPAQIYRVTWFISWSPCFSWGCAGEVRAFLQENTHVRLRIFAARIYDYDPLYK
EALQMLRDAGAQVSIMTYDEFEYCWDTFVYRQGCPFQPWDGLEEHSQALSGRLRAILQNQGN

SEQ ID NO. 23: Uniprot G3QV16

MNPQIRNPMERMYRGTFYNNFENEPILYGRSYNWLCYEVKIKRGRSNLLWNTGVFRGQMYSQPE
HHAEMCFLSWFCGNQLPAYKCFQITWFVSWTPCPDCVAKLAEFLAEYPNVTLTISTARLYYYWE
RDYRRALCRLSQAGARMKIMDYEECAYCWENFVYKEGQQFMPWYKFDENYAFLHHTLKEILRHL
MDPDTFTFNFNNDPLVLRRHQTYLCYEVERLDNGTWVLMDRHMGFLCNEAKNLLCGFYGRHAEL
RFLDLVPSLQLDPAQIYRVTWFISWSPCFSWGCAGQVCEFLQENTHMRLRIFAARIYDYDPLYK
KALQMLRDAGAQVSIMTYDEFKHCWDTFVYRQGCPFQPWDGLEEHSQALSGRLQAILQNQGN

SEQ ID NO. 24: Uniprot F6M3K5

MNPQIRNPMERMYRRTFNYNFENEPILYGRSYTWLCYEVKIRKDPSKLPWDTGVFRGQMYSKPE
HHAEMCFLSWFCGNQLPAHKRFQITWFVSWTPCPDCVAKVAEFLAEYPNVTLTISAARLYYYWE
TDYRRALCRLRQAGARVKIMDYEEFAYCWENFVYNEDQSFMPWYKFDDNYAFLHHKLKEILRHL
MDPDTFTSNFNNDLSVLGRHQTYLCYEVERLDNGTWVPMDQHWGFLCNQAKNVPRGDYGCHAEL
CFLDQVSSWQLDPAQTYRVTWFISWSPCFSWGCADQVYAFLQENTHVRLRIFAARIYDYNPLYQ
EALRTLRDAGAQVSIMTYDEFEYCWDTFVDRQGRPFQPWDGLDEHSQALSGRLRAILQNQGN

SEQ ID NO. 25: UniProt Q9NRW3

MNPQIRNPMKAMYPGTFYFQFKNLWEANDRNETWLCFTVEGIKRRSVVSWKTGVFRNQVDSETH
CHAERCFLSWFCDDILSPNTKYQVTWYTSWSPCPDCAGEVAEFLARHSNVNLTIFTARLYYFQY
PCYQEGLRSLSQEGVAVEIMDYEDFKYCWENFVYNDNEPFKPWKGLKTNFRLLKRRLRESLQ

SEQ ID NO. 26: Uniprot Q694B5

MNPQIRNPMKAMYPGTFYFQFKNLWEANDRNETWLCFTVEGIKRRSVVSWKTGVFRNQVDSETH
CHAERCFLSWFCDDILSPNTNYQVTWYTSWSPCPECAGEVAEFLARHSNVNLTIFTARLYYFQD
TDYQEGLRSLSQEGVAVKIMDYKDFKYCWENFVYNDDEPFKPWKGLKYNFRFLKRRLQEILE

SEQ ID NO. 27: Uniprot B0LW74

MNPQIRNPMKAMDPGTFYFQFKNLWEANNRNETWLCFTVEVIKQHSTVSWETGVFRNQVDLETH
CHAERCFLSWFCEDILSPNTDYQVTWYTSWSPCLDCAGEVAKFLARHNNVMLTIYTARLYYSQY
PNYQQGLRSLSEKGVSVKIMDYEDFKYCWEKFVYDDGEPFKPWKGLKTSFRFLKRRLREILQ

SEQ ID NO. 28: UniProt Q96AK3

MNPQIRNPMERMYRDTFYDNFENEPILYGRSYTWLCYEVKIKRGRSNLLWDTGVFRGPVLPKRQ
SNHRQEVYFRFENHAEMCFLSWFCGNRLPANRRFQITWFVSWNPCLPCVVKVTKFLAEHPNVTL
TISAARLYYYRDRDWRWVLLRLHKAGARVKIMDYEDFAYCWENFVCNEGQPFMPWYKFDDNYAS
LHRTLKEILRNPMEAMYPHIFYFHFKNLLKACGRNESWLCFTMEVTKHHSAVFRKRGVFRNQVD
PETHCHAERCFLSWFCDDILSPNTNYEVTWYTSWSPCPECAGEVAEFLARHSNVNLTIFTARLC
YFWDTDYQEGLCSLSQEGASVKIMGYKDFVSCWKNFVYSDDEPFKPWKGLQTNFRLLKRRLREI
LQ

SEQ ID NO. 29: NCBI NP_001332895.1

MNPQIRNPMERMYRRTFYNHFENEPILYGRSYTWLCYEVKIKRGCSNLIWDTGVFRGPVLPKLQ
SNHRQEVYFQFENHAEMCFFSWFCGNRLPANRRFQITWFVSWNPCLPCVVKVTKFLAEHPNVTL
TISAARLYYYQDREWRRVLRRLHKAGARVKIMDYKDFAHCWENFVYNEGQQFMPWYKFDDNYAS
LHRTLKEILRNPMEAMYPHVFYFHFKNLLKACGRNESWLCFTVDVTEHHPPVSWKRGVFRNPVD
PETHCHAERCFLSWFCDDILSPNTNYQVTWYTSWSPCPECAREVAEFLARHSNVKLTIFTARLY
HFWNTDYQEGLCSLSQEGASVKIMSYKDFVSCWKNFVYSDDEPFKPWKGLKTNFRLLKTMLREI
LQ

SEQ ID NO. 30: NCBI NP_001332931.1

MNPQIRNPMERMYRRTFNYNFENEPILYGRSYTWLCYEVKIRKDPSKLPWDTGVFRGQVYFQPQ
YHAEMCFLSWFCGNQLPAYKRFQITWFVSWNPCPDCVAKVTEFLAEHPNVTLTISVARLYYYRG
KDWRRALCRLHQAGARVKIMDYEEFAYCWENFVYNEGQSFMPWDKFDDNYAFLHHKLKEILRNP
MKAMYPHTFYFHFENLQKAYGRNETWLCFAVEIIKQHSTVPWKTGVFRNQVDPETHCHAERCFL
SWFCDNTLSPKKNYQVTWYISWSPCPECAGEVAEFLATHSNVKLTIYTARLYYFWDTDYQEGLR
SLSEEGASMEIMGYEDFKYCWENFVYNDGEPFKPWKGINTNFRFLERRLWKILQ

SEQ ID NO. 31: UniProt Q8IUX4

MKPHFRNTVERMYRDTFSYNFYNRPILSRRNTVWLCYEVKTKGPSRPRLDAKIFRGQVYSQPEH
HAEMCFLSWFCGNQLPAYKCFQITWFVSWTPCPDCVAKLAEFLAEHPNVTLTISAARLYYYWER
DYRRALCRLSQAGARVKIMDDEEFAYCWENFVYSEGQPFMPWYKFDDNYAFLHRTLKEILRNPM
EAMYPHIFYFHFKNLRKAYGRNESWLCFTMEVVKHHSPVSWKRGVFRNQVDPETHCHAERCFLS
WFCDDILSPNTNYEVTWYTSWSPCPECAGEVAEFLARHSNVNLTIFTARLYYFWDTDYQEGLRS
LSQEGASVEIMGYKDFKYCWENFVYNDDEPFKPWKGLKYNFLFLDSKLQEILE

SEQ ID NO. 32: Uniprot G3RD21

MKPQFRNTVERMYRGTFSYNFNNRPILSRRNTVWLCYEVKTKGPSRPPLDAKIFRGQVYFQPQY
HAEMCFLSWFCGNQLPAYKCFQITCFVSWTPCPDCVAKLAEFLAEHPNVTLTISAARLYYYWER
DYRRALRRLRQAGARVKIMDDEEFAYCWENFVYSEGQPFMPWHKFDDNYAFLHRTLKEILRNPM
EAMYPHIFYFHFKNLLKAYGRNESWLCFTMEVIKHHSPVSWKRGVFRNQVDSETHCHAERCFLS

WFCDDILSPNTNYQVTWYTSWSPCPECAGEVAEFLARHSNVNLTIFTARLYYFWDTDYQEGLRS
LNQEGASVKIMGYKDFKYCWENFVYNDDEPFKPWKGLKYNFLFLDSKLQEILE

SEQ ID NO. 33: Uniprot Q1G0Z6

MQPQYRNTVERMYRGTFFYNFNNRPILSRRNTVWLCYEVKTRGPSMPTWDTKIFRGQVYSKPEH
HAEMCFLSRFCGNQLPAYKRFQITWFVSWTPCPDCVAKVAEFLAEHPNVTLTISAARLYYYWET
DYRRALCRLRQAGARVKIMDYEEFAYCWENFVYNEGQSFMPWDKFDDNYAFLHHKLKEILRNPM
EATYPHIFYFHFKNLRKAYGRNETWLCFTMEIIKQHSTVSWETGVFRNQVDPESRCHAERCFLS
WFCEDILSPNTDYQVTWYTSWSPCLDCAGEVAEFLARHSNVKLAIFAARLYYFWDTHYQQGLRS
LSEKGASVEIMGYKDFKYCWENFVYNGDEPFKPWKGLKYNFLFLDSKLQEILE

SEQ ID NO. 34: UniProt Q9HC16

MKPHFRNTVERMYRDTFSYNFYNRPILSRRNTVWLCYEVKTKGPSRPPLDAKIFRGQVYSELKY
HPEMRFFHWFSKWRKLHRDQEYEVTWYISWSPCTKCTRDMATFLAEDPKVTLTIFVARLYYFWD
PDYQEALRSLCQKRDGPRATMKIMNYDEFQHCWSKFVYSQRELFEPWNNLPKYYILLHIMLGEI
LRHSMDPPTFTFNFNNEPWVRGRHETYLCYEVERMHNDTWVLLNQRRGFLCNQAPHKHGFLEGR
HAELCFLDVIPFWKLDLDQDYRVTCFTSWSPCFSCAQEMAKFISKNKHVSLCIFTARIYDDQGR
CQEGLRTLAEAGAKISIMTYSEFKHCWDTFVDHQGCPFQPWDGLDEHSQDLSGRLRAILQNQEN

SEQ ID NO. 35: Uniprot Q694C1

MTPQFRNTVERMYRDTFSYNFNNRPILSRRNTVWLCYEVKTKDPSRPPLDAKIFRGQVYSELKY
HPEMRFFHWFSKWRKLHRDQEYEVTWYISWSPCTKCTRNVATFLAEDPKVTLTIFVARLYYFWD
QDYQEALRSLCQKRDGPRATMKIMNYDEFQHCWSKFVYSQRELFEPWNNLPKYYMLLHIMLGEI
LRHSMDPPTFTSNFNNEHWVRGRHETYLCYEVERLHNDTWVLLNQRRGFLCNQAPHKHGFLEGR
HAELCFLDVIPFWKLDLHQDYRVTCFTSWSPCFSCAQEMAKFISNKKHVSLCIFAARIYDDQGR
CQEGLRTLAEAGAKISIMTYSEFKHCWDTFVYHQGCPFQPWDGLEEHSQALSGRLQAILQNQGN

SEQ ID NO. 36: Uniprot U5NDB3

MNPQIRNMVEPMDPRTFVSNFNNRPILSGLNTVWLCCEVKTKDPSGPPLDAKIFQGKVLRSKAK
YHPEMRFLQWFREWRQLHHDQEYKVTWYVSWSPCTRCANSVATFLAKDPKVTLTIFVARLYYFW
KPNYQQALRILCQKRDGPHATMKIMNYNEFQDCWNKFVDGRGKPFKPWNNLPKHYTLLQATLGE
LLRHLMDPGTFTSNFNNKPWVSGQHETYLCYKVERLHNDTWVPLNQHRGFLRNQAPNIHGFPKG
RHAELCFLDLIPFWKLDGQQYRVTCFTSWSPCFSCAQEMAKFISNNEHVSLCIFAARIYDDQGR
YQEGLRTLHRDGAKIAMMNYSEFEYCWDTFVDCQGCPFQPWDGLDEHSQALSERLRAILQNQGN

SEQ ID NO. 37: UniProt Q6NTF7

MALLTAETFRLQFNNKRRLRRPYYPRKALLCYQLTPQNGSTPTRGYFENKKKCHAEICFINEIK
SMGLDETQCYQVTCYLTWSPCSSCAWELVDFIKAHDHLNLGIFASRLYYHWCKPQQKGLRLLCG
SQVPVEVMGFPEFADCWENFVDHEKPLSFNPYKMLEELDKNSRAIKRRLERIKIPGVRAQGRYM
DILCDAEV

SEQ ID NO. 38: Uniprot B7T0U7

MALLTAETFRLQFNNKLRLRRPYYRRKTLLCYQLTPQNGSMPTRGYFKNKKKCHAEICFINEIK
SMGLDETQCYQVTCYLTWSPCSSCAWKLVDFIKAHDHLNLRIFASRLYYHWCKRQQEGLRLLCG
SQVPVEVMGFPEFADCWENFVDHEKPLSFDPSKMLEELDKNSQAIKRRLERIKSRSVDVLENGL
RSLQLGPVTPSSSRSNSR

SEQ ID NO. 39: Uniprot Q19Q52

MALLTAKTFSLQFNNKRRVNKPYYPRKALLCYQLTPQNGSTPTRGHLKNKKKDHAEIRFINKIK
SMGLDETQCYQVTCYLTWSPCPSCAGELVDFIKAHRHLNLRIFASRLYYHWRPNYQEGLLLLCG
SQVPVEVMGLPEFTDCWENFVDHKEPPSFNPSEKLEELDKNSQAIKRRLERIKSRSVDVLENGL
RSLQLGPVTPSSSIRNSR

SEQ ID NO. 40: UniProt Q8WW27

MEPIYEEYLANHGTIVKPYYWLSFSLDCSNCPYHIRTGEEARVSLTEFCQIFGFPYGTTFPQTK
HLTFYELKTSSGSLVQKGHASSCTGNYIHPESMLFEMNGYLDSAIYNNDSIRHIILYSNNSPCN
EANHCCISKMYNFLITYPGITLSIYFSQLYHTEMDFPASAWNREALRSLASLWPRVVLSPISGG
IWHSVLHSFISGVSGSHVFQPILTGRALADRHNAYEINAITGVKPYFTDVLLQTKRNPNTKAQE
ALESYPLNNAFPGQFFQMPSGQLQPNLPPDLRAPVVFVLVPLRDLPPMHMGQNPNKPRNIVRHL
NMPQMSFQETKDLGRLPTGRSVEIVEITEQFASSKEADEKKKKKGKK

SEQ ID NO. 41: NCBI XP_004028087.1

MEPIYEEYLANHGTIVKPYYWLSFSLDCSNCPYHIRTGEEARVSLTEFCQIFGFPYGTTFPQTK
HLTFYELKTSSGSLVQKGHASSCTGNYIHPESMLFEMNGYLDSAIYNNDSIRHIILYSNNSPCN
EANHCCISKMYNFLITYPGITLSIYFSQLYHTEMDFPASAWNREALRSLASLWPRVVLSPISGG
IWHSVLHSFISGVSGSHVFQPILTGRALADRHNAYEINAITGVKPYFTDVLLQTKRNPNTKAQE
ALESYPLNNAFPGQSFQMPSGQLQPNLPPDVRAPVVFVLVPLRDLPPMHMGQNPNKPRNIVRHL
NMPQMSFQETEDLGRLPTGRSVEIVEITERFASSKEADEKKKKKKGKK

SEQ ID NO. 42: Uniprot Q497M3

MEPLYEEILTQGGTIVKPYYWLSLSLGCTNCPYHIRTGEEARVPYTEFHQTFGFPWSTYPQTKH
LTFYELRSSSKNLIQKGLASNCTGSHNHPEAMLFEKNGYLDAVIFHNSNIRHIILYSNNSPCNE
AKHCCISKMYNFLMNYPEVTLSVFFSQLYHTEKQFPTSAWNRKALQSLASLWPQVTLSPICGGL
WHAILEKFVSNISGSTVPQPFIAGRILADRYNTYEINSIIAAKPYFTDGLLSRQKENQNREAWA
AFEKHPLGSAAPAQRQPTRGQDPRTPAVLMLVSNRDLPPIHVGSTPQKPRTVVRHLNMLQLSSF
KVKDVKKPPSGRPVEEVEVMKESARSQKANKKNRSQWKKQTLVIKPRICRLLER

SEQ ID NO. 43: UniProt Q8NFU7

MSRSRHARPSRLVRKEDVNKKKKNSQLRKTTKGANKNVASVKTLSPGKLKQLIQERDVKKKTEP
KPPVPVRSLLTRAGAARMNLDRTEVLFQNPESLTCNGFTMALRSTSLSRRLSQPPLVVAKSKKV
PLSKGLEKQHDCDYKILPALGVKHSENDSVPMQDTQVLPDIETLIGVQNPSLLKGKSQETTQFW
SQRVEDSKINIPTHSGPAAEILPGPLEGTRCGEGLFSEETLNDTSGSPKMFAQDTVCAPFPQRA
TPKVTSQGNPSIQLEELGSRVESLKLSDSYLDPIKSEHDCYPTSSLNKVIPDLNLRNCLALGGS
TSPTSVIKFLLAGSKQATLGAKPDHQEAFEATANQQEVSDTTSFLGQAFGAIPHQWELPGADPV
HGEALGETPDLPEIPGAIPVQGEVFGTILDQQETLGMSGSVVPDLPVFLPVPPNPIATFNAPSK
WPEPQSTVSYGLAVQGAIQILPLGSGHTPQSSSNSEKNSLPPVMAISNVENEKQVHISFLPANT
QGFPLAPERGLFHASLGIAQLSQAGPSKSDRGSSQVSVTSTVHVVNTTVVTMPVPMVSTSSSSY
TTLLPTLEKKKRKRCGVCEPCQQKTNCGECTYCKNRKNSHQICKKRKCEELKKKPSVVVPLEVI
KENKRPQREKKPKVLKADFDNKPVNGPKSESMDYSRCGHGEEQKLELNPHTVENVTKNEDSMTG
IEVEKWTQNKKSQLTDHVKGDFSANVPEAEKSKNSEVDKKRTKSPKLFVQTVRNGIKHVHCLPA
ETNVSFKKFNIEEFGKTLENNSYKFLKDTANHKNAMSSVATDMSCDHLKGRSNVLVFQQPGFNC
SSIPHSSHSIINHHASIHNEGDQPKTPENIPSKEPKDGSPVQPSLLSLMKDRRLTLEQVVAIEA
LTQLSEAPSENSSPSKSEKDEESEQRTASLLNSCKAILYTVRKDLQDPNLQGEPPKLNHCPSLE
KQSSCNTVVFNGQTTTLSNSHINSATNQASTKSHEYSKVTNSLSLFIPKSNSSKIDTNKSIAQG
IITLDNCSNDLHQLPPRNNEVEYCNQLLDSSKKLDSDDLSCQDATHTQIEEDVATQLTQLASII
KINYIKPEDKKVESTPTSLVTCNVQQKYNQEKGTIQQKPPSSVHNNHGSSLTKQKNPTQKKTKS
TPSRDRRKKKPTVVSYQENDRQKWEKLSYMYGTICDIWIASKFQNFGQFCPHDFPTVFGKISSS
TKIWKPLAQTRSIMQPKTVFPPLTQIKLQRYPESAEEKVKVEPLDSLSLFHLKTESNGKAFTDK
AYNSQVQLTVNANQKAHPLTQPSSPPNQCANVMAGDDQIRFQQVVKEQLMHQRLPTLPGISHET
PLPESALTLRNVNVVCSGGITVVSTKSEEEVCSSSFGTSEFSTVDSAQKNFNDYAMNFFTNPTK

NLVSITKDSELPTCSCLDRVIQKDKGPYYTHLGAGPSVAAVREIMENRYGQKGNAIRIEIVVYT
GKEGKSSHGCPIAKWVLRRSSDEEKVLCLVRQRTGHHCPTAVMVVLIMVWDGIPLPMADRLYTE
LTENLKSYNGHPTDRRCTLNENRTCTCQGIDPETCGASFSFGCSWSMYFNGCKFGRSPSPRRFR
IDPSSPLHEKNLEDNLQSLATRLAPIYKQYAPVAYQNQVEYENVARECRLGSKEGRPFSGVTAC
LDFCAHPHRDIHNMNNGSTVVCTLTREDNRSLGVIPQDEQLHVLPLYKLSDTDEFGSKEGMEAK
IKSGAIEVLAPRRKKRTCFTQPVPRSGKKRAAMMTEVLAHKIRAVEKKPIPRIKRKNNSTTTNN
SKPSSLPTLGSNTETVQPEVKSETEPHFILKSSDNTKTYSLMPSAPHPVKEASPGFSWSPKTAS
ATPAPLKNDATASCGFSERSSTPHCTMPSGRLSGANAAAADGPGISQLGEVAPLPTLSAPVMEP
LINSEPSTGVTEPLTPHQPNHQPSFLTSPQDLASSPMEEDEQHSEADEPPSDEPLSDDPLSPAE
EKLPHIDEYWSDSEHIFLDANIGGVAIAPAHGSVLIECARRELHATTPVEHPNRNHPTRLSLVF
YQHKNLNKPQHGFELNKIKFEAKEAKNKKMKASEQKDQAANEGPEQSSEVNELNQIPSHKALTL
THDNVVTVSPYALTHVAGPYNHWV

SEQ ID NO. 44: UniProt Q3URK3

MSRSRPAKPSKSVKTKLQKKKDIQMKTKTSKQAVRHGASAKAVNPGKPKQLIKRRDGKKETEDK
TPTPAPSFLTRAGAARMNRDRNQVLFQNPDSLTCNGFTMALRRTSLSWRLSQRPVVTPKPKKVP
PSKKQCTHNIQDEPGVKHSENDSVPSQHATVSPGTENGEQNRCLVEGESQEITQSCPVFEERIE
DTQSCISASGNLEAEISWPLEGTHCEELLSHQTSDNECTSPQECAPLPQRSTSEVTSQKNTSNQ
LADLSSQVESIKLSDPSPNPTGSDHNGFPDSSFRIVPELDLKTCMPLDESVYPTALIRFILAGS
QPDVFDTKPQEKTLITTPEQVGSHPNQVLDATSVLGQAFSTLPLQWGFSGANLVQVEALGKGSD
SPEDLGAITMLNQQETVAMDMDRNATPDLPIFLPKPPNTVATYSSPLLGPEPHSSTSCGLEVQG
ATPILTLDSGHTPQLPPNPESSSVPLVIAANGTRAEKQFGTSLFPAVPQGFTVAAENEVQHAPL
DLTQGSQAAPSKLEGEISRVSITGSADVKATAMSMPVTQASTSSPPCNSTPPMVERRKRKACGV
CEPCQQKANCGECTYCKNRKNSHQICKKRKCEVLKKKPEATSQAQVTKENKRPQREKKPKVLKT
DFNNKPVNGPKSESMDCSRRGHGEEEQRLDLITHPLENVRKNAGGMTGIEVEKWAPNKKSHLAE
GQVKGSCDANLTGVENPQPSEDDKQQTNPSPTFAQTIRNGMKNVHCLPTDTHLPLNKLNHEEFS
KALGNNSSKLLTDPSNCKDAMSVTTSGGECDHLKGPRNTLLFQKPGLNCRSGAEPTIFNNHPNT
HSAGSRPHPPEKVPNKEPKDGSPVQPSLLSLMKDRRLTLEQVVAIEALTQLSEAPSESSSPSKP
EKDEEAHQKTASLLNSCKAILHSVRKDLQDPNVQGKGLHHDTVVFNGQNRTFKSPDSFATNQAL
IKSQGYPSSPTAEKKGAAGGRAPFDGFENSHPLPIESHNLENCSQVLSCDQNLSSHDPSCQDAP
YSQIEEDVAAQLTQLASTINHINAEVRNAESTPESLVAKNTKQKHSQEKRMVHQKPPSSTQTKP
SVPSAKPKKAQKKARATPHANKRKKKPPARSSQENDQKKQEQLAIEYSKMHDIWMSSKFQRFGQ
SSPRSFPVLLRNIPVFNQILKPVTQSKTPSQHNELFPPINQIKFTRNPELAKEKVKVEPSDSLP
TCQFKTESGGQTFAEPADNSQGQPMVSVNQEAHPLPQSPPSNQCANIMAGAAQTQFHLGAQENL
VHQIPPPTLPGTSPDTLLPDPASILRKGKVLHFDGITVVTEKREAQTSSNGPLGPTTDSAQSEF
KESIMDLLSKPAKNLIAGLKEQEAAPCDCDGGTQKEKGPYYTHLGAGPSVAAVRELMETRFGQK
GKAIRIEKIVFTGKEGKSSQGCPVAKWVIRRSGPEEKLICLVRERVDHHCSTAVIVVLILLWEG
IPRLMADRLYKELTENLRSYSGHPTDRRCTLNKKRTCTCQGIDPKTCGASFSFGCSWSMYFNGC
KFGRSENPRKFRLAPNYPLHEKQLEKNLQELATVLAPLYKQMAPVAYQNQVEYEEVAGDCRLGN
EEGRPFSGVTCCMDFCAHSHKDIHNMHNGSTVVCTLIRADGRDTNCPEDEQLHVLPLYRLADTD
EFGSVEGMKAKIKSGAIQVNGPTRKRRLRFTEPVPRCGKRAKMKQNHNKSGSHNTKSFSSASST
SHLVKDESTDFCPLQASSAETSTCTYSKTASGGFAETSSILHCTMPSGAHSGANAAAGECTGTV
QPAEVAAHPHQSLPTADSPVHAEPLTSPSEQLTSNQSNQQLPLLSNSQKLASCQVEDERHPEAD
EPQHPEDDNLPQLDEFWSDSEEIYADPSFGGVAIAPIHGSVLIECARKELHATTSLRSPKRGVP
FRVSLVFYQHKSLNKPNHGFDINKIKCKCKKVTKKKPADRECPDVSPEANLSHQIPSRVASTLT
RDNVVTVSPYSLTHVAGPYNRWV

SEQ ID NO. 45: UniProt Q6N021

MEQDRTNHVEGNRLSPFLIPSPPICQTEPLATKLQNGSPLPERAHPEVNGDTKWHSFKSYYGIP
CMKGSQNSRVSPDFTQESRGYSKCLQNGGIKRTVSEPSLSGLLQIKKLKQDQKANGERRNFGVS
QERNPGESSQPNVSDLSDKKESVSSVAQENAVKDFTSFSTHNCSGPENPELQILNEQEGKSANY
HDKNIVLLKNKAVLMPNGATVSASSVEHTHGELLEKTLSQYYPDCVSIAVQKTTSHINAINSQA
TNELSCEITHPSHTSGQINSAQTSNSELPPKPAAVVSEACDADDADNASKLAAMLNTCSFQKPE
QLQQQKSVFEICPSPAENNIQGTTKLASGEEFCSGSSSNLQAPGGSSERYLKQNEMNGAYFKQS

SVFTKDSFSATTTPPPPSQLLLSPPPPLPQVPQLPSEGKSTLNGGVLEEHHHYPNQSNTTLLRE
VKIEGKPEAPPSQSPNPSTHVCSPSPMLSERPQNNCVNRNDIQTAGTMTVPLCSEKTRPMSEHL
KHNPPIFGSSGELQDNCQQLMRNKEQEILKGRDKEQTRDLVPPTQHYLKPGWIELKAPRFHQAE
SHLKRNEASLPSILQYQPNLSNQMTSKQYTGNSNMPGGLPRQAYTQKTTQLEHKSQMYQVEMNQ
GQSQGTVDQHLQFQKPSHQVHFSKTDHLPKAHVQSLCGTRFHFQQRADSQTEKLMSPVLKQHLN
QQASETEPFSNSHLLQHKPHKQAAQTQPSQSSHLPQNQQQQQKLQIKNKEEILQTFPHPQSNND
QQREGSFFGQTKVEECFHGENQYSKSSEFETHNVQMGLEEVQNINRRNSPYSQTMKSSACKIQV
SCSNNTHLVSENKEQTTHPELFAGNKTQNLHHMQYFPNNVIPKQDLLHRCFQEQEQKSQQASVL
QGYKNRNQDMSGQQAAQLAQQRYLIHNHANVFPVPDQGGSHTQTPPQKDTQKHAALRWHLLQKQ
EQQQTQQPQTESCHSQMHRPIKVEPGCKPHACMHTAPPENKTWKKVTKQENPPASCDNVQQKSI
IETMEQHLKQFHAKSLFDHKALTLKSQKQVKVEMSGPVTVLTRQTTAAELDSHTPALEQQTTSS
EKTPTKRTAASVLNNFIESPSKLLDTPIKNLLDTPVKTQYDFPSCRCVEQIIEKDEGPFYTHLG
AGPNVAAIREIMEERFGQKGKAIRIERVIYTGKEGKSSQGCPIAKWVVRRSSSEEKLLCLVRER
AGHTCEAAVIVILILVWEGIPLSLADKLYSELTETLRKYGTLTNRRCALNEERTCACQGLDPET
CGASFSFGCSWSMYYNGCKFARSKIPRKFKLLGDDPKEEEKLESHLQNLSTLMAPTYKKLAPDA
YNNQIEYEHRAPECRLGLKEGRPFSGVTACLDFCAHAHRDLHNMQNGSTLVCTLTREDNREFGG
KPEDEQLHVLPLYKVSDVDEFGSVEAQEEKKRSGAIQVLSSFRRKVRMLAEPVKTCRQRKLEAK
KAAAEKLSSLENSSNKNEKEKSAPSRTKQTENASQAKQLAELLRLSGPVMQQSQQPQPLQKQPP
QPQQQQRPQQQQPHHPQTESVNSYSASGSTNPYMRRPNPVSPYPNSSHTSDIYGSTSPMNFYST
SSQAAGSYLNSSNPMNPYPGLLNQNTQYPSYQCNGNLSVDNCSPYLGSYSPQSQPMDLYRYPSQ
DPLSKLSLPPIHTLYQPRFGNSQSFTSKYLGYGNQNMQGDGFSSCTIRPNVHHVGKLPPYPTHE
MDGHFMGATSRLPPNLSNPNMDYKNGEHHSPSHIIHNYSAAPGMFNSSLHALHLQNKENDMLSH
TANGLSKMLPALNHDRTACVQGGLHKLSDANGQEKQPLALVQGVASGAEDNDEVWSDSEQSFLD
PDIGGVAVAPTHGSILIECAKRELHATTPLKNPNRNHPTRISLVFYQHKSMNEPKHGLALWEAK
MAEKAREKEEECEKYGPDYVPQKSHGKKVKREPAEPHETSEPTYLRFIKSLAERTMSVTTDSTV
TTSPYAFTRVTGPYNRYI

SEQ ID NO. 46: UniProt Q4JK59

MEQDRTTHAEGTRLSPFLIAPPSPISHTEPLAVKLQNGSPLAERPHPEVNGDTKWQSSQSCYGI
SHMKGSQSSHESPHEDRGYSRCLQNGGIKRTVSEPSLSGLHPNKILKLDQKAKGESNIFEESQE
RNHGKSSRQPNVSGLSDNGEPVTSTTQESSGADAFPTRNYNGVEIQVLNEQEGEKGRSVTLLKN
KIVLMPNGATVSAHSEENTRGELLEKTQCYPDCVSIAVQSTASHVNTPSSQAAIELSHEIPQPS
LTSAQINFSQTSSLQLPPEPAAMVTKACDADNASKPAIVPGTCPFQKAEHQQKSALDIGPSRAE
NKTIQGSMELFAEEYYPSSDRNLQASHGSSEQYSKQKETNGAYFRQSSKFPKDSISPTTVTPPS
QSLLAPRLVLQPPLEGKGALNDVALEEHHDYPNRSNRTLLREGKIDHQPKTSSSQSLNPSVHTP
NPPLMLPEQHQNDCGSPSPEKSRKMSEYLMYYLPNHGHSGGLQEHSQYLMGHREQEIPKDANGK
QTQGSVQAAPGWIELKAPNLHEALHQTKRKDISLHSVLHSQTGPVNQMSSKQSTGNVNMPGGFQ
RLPYLQKTAQPEQKAQMYQVQVNQGPSPGMGDQHLQFQKALYQECIPRTDPSSEAHPQAPSVPQ
YHFQQRVNPSSDKHLSQQATETQRLSGFLQHTPQTQASQTPASQNSNFPQICQQQQQQQLQRKN
KEQMPQTFSHLQGSNDKQREGSCFGQIKVEESFCVGNQYSKSSNFQTHNNTQGGLEQVQNINKN
FPYSKILTPNSSNLQILPSNDTHPACEREQALHPVGSKTSNLQNMQYFPNNVTPNQDVHRCFQE
QAQKPQQASSLQGLKDRSQGESPAPPAEAAQQRYLVHNEAKALPVPEQGGSQTQTPPQKDTQKH
AALRWLLLQKQEQQQTQQSQPGHNQMLRPIKTEPVSKPSSYRYPLSPPQENMSSRIKQEISSPS
RDNGQPKSIIETMEQHLKQFQLKSLCDYKALTLKSQKHVKVPTDIQAAESENHARAAEPQATKS
TDCSVLDDVSESDTPGEQSQNGKCEGCNPDKDEAPYYTHLGAGPDVAAIRTLMEERYGEKGKAI
RIEKVIYTGKEGKSSQGCPIAKWVYRRSSEEEKLLCLVRVRPNHTCETAVMVIAIMLWDGIPKL
LASELYSELTDILGKCGICTNRRCSQNETRNCCCQGENPETCGASFSFGCSWSMYYNGCKFARS
KKPRKFRLHGAEPKEEERLGSHLQNLATVIAPIYKKLAPDAYNNQVEFEHQAPDCCLGLKEGRP
FSGVTACLDFSAHSHRDQQNMPNGSTVVVTLNREDNREVGAKPEDEQFHVLPMYIIAPEDEFGS
TEGQEKKIRMGSIEVLQSFRRRRVIRIGELPKSCKKKAEPKKAKTKKAARKRSSLENCSSRTEK
GKSSSHTKLMENASHMKQMTAQPQLSGPVIRQPPTLQRHLQQGQRPQQPQPPQPQPQTTPQPQP
QPQHIMPGNSQSVGSHCSGSTSVYTRQPTPHSPYPSSAHTSDIYGDTNHVNFYPTSSHASGSYL
NPSNYMNPYLGLLNQNNQYAPFPYNGSVPVDNGSPFLGSYSPQAQSRDLHRYPNQDHLTNQNLP
PIHTLHQQTFGDSPSKYLSYGNQNMQRDAFTTNSTLKPNVHHLATFSPYPTPKMDSHFMGAASR
SPYSHPHTDYKTSEHHLPSHTIYSYTAAASGSSSSHAFHNKENDNIANGLSRVLPGFNHDRTAS

AQELLYSLTGSSQEKQPEVSGQDAAAVQEIEYWSDSEHNFQDPCIGGVAIAPTHGSILIECAKC
EVHATTKVNDPDRNHPTRISLVLYRHKNLFLPKHCLALWEAKMAEKARKEEECGKNGSDHVSQK
NHGKQEKREPTGPQEPSYLRFIQSLAENTGSVTTDSTVTTSPYAFTQVTGPYNTFV

SEQ ID NO. 47: UniProt O43151

MSQFQVPLAVQPDLPGLYDFPQRQVMVGSFPGSGLSMAGSESQLRGGGDGRKKRKRCGTCEPCR
RLENCGACTSCTNRRTHQICKLRKCEVLKKKVGLLKEVEIKAGEGAGPWGQGAAVKTGSELSPV
DGPVPGQMDSGPVYHGDSRQLSASGVPVNGAREPAGPSLLGTGGPWRVDQKPDWEAAPGPAHTA
RLEDAHDLVAFSAVAEAVSSYGALSTRLYETFNREMSREAGNNSRGPRPGPEGCSAGSEDLDTL
QTALALARHGMKPPNCNCDGPECPDYLEWLEGKIKSVVMEGGEERPRLPGPLPPGEAGLPAPST
RPLLSSEVPQISPQEGLPLSQSALSIAKEKNISLQTAIAIEALTQLSSALPQPSHSTPQASCPL
PEALSPPAPFRSPQSYLRAPSWPVVPPEEHSSFAPDSSAFPPATPRTEFPEAWGTDTPPATPRS
SWPMPRPSPDPMAELEQLLGSASDYIQSVFKRPEALPTKPKVKVEAPSSSPAPAPSPVLQREAP
TPSSEPDTHQKAQTALQQHLHHKRSLFLEQVHDTSFPAPSEPSAPGWWPPPSSPVPRLPDRPPK
EKKKKLPTPAGGPVGTEKAAPGIKPSVRKPIQIKKSRPREAQPLFPPVRQIVLEGLRSPASQEV
QAHPPAPLPASQGSAVPLPPEPSLALFAPSPSRDSLLPPTQEMRSPSPMTALQPGSTGPLPPAD
DKLEELIRQFEAEFGDSFGLPGPPSVPIQDPENQQTCLPAPESPFATRSPKQIKIESSGAVTVL
STTCFHSEEGGQEATPTKAENPLTPTLSGFLESPLKYLDTPTKSLLDTPAKRAQAEFPTCDCVE
QIVEKDEGPYYTHLGSGPTVASIRELMEERYGEKGKAIRIEKVIYTGKEGKSSRGCPIAKWVIR
RHTLEEKLLCLVRHRAGHHCQNAVIVILILAWEGIPRSLGDTLYQELTDTLRKYGNPTSRRCGL
NDDRTCACQGKDPNTCGASFSFGCSWSMYFNGCKYARSKTPRKFRLAGDNPKEEEVLRKSFQDL
ATEVAPLYKRLAPQAYQNQVTNEEIAIDCRLGLKEGRPFAGVTACMDFCAHAHKDQHNLYNGCT
VVCTLTKEDNRCVGKIPEDEQLHVLPLYKMANTDEFGSEENQNAKVGSGAIQVLTAFPREVRRL
PEPAKSCRQRQLEARKAAAEKKKIQKEKLSTPEKIKQEALELAGITSDPGLSLKGGLSQQGLKP
SLKVEPQNHFSSFKYSGNAVVESYSVLGNCRPSDPYSMNSVYSYHSYYAQPSLTSVNGFHSKYA
LPSFSYYGFPSSNPVFPSQFLGPGAWGHSGSSGSFEKKPDLHALHNSLSPAYGGAEFAELPSQA
VPTDAHHPTPHHQQPAYPGPKEYLLPKAPLLHSVSRDPSPFAQSSNCYNRSIKQEPVDPLTQAE
PVPRDAGKMGKTPLSEVSQNGGPSHLWGQYSGGPSMSPKRTNGVGGSWGVFSSGESPAIVPDKL
SSFGASCLAPSHFTDGQWGLFPGEGQQAASHSGGRLRGKPWSPCKFGNSTSALAGPSLTEKPWA
LGAGDFNSALKGSPGFQDKLWNPMKGEEGRIPAAGASQLDRAWQSFGLPLGSSEKLFGALKSEE
KLWDPFSLEEGPAEEPPSKGAVKEEKGGGGAEEEEEELWSDSEHNFLDENIGGVAVAPAHGSIL
IECARRELHATTPLKKPNRCHPTRISLVFYQHKNLNQPNHGLALWEAKMKQLAERARARQEEAA
RLGLGQQEAKLYGKKRKWGGTVVAEPQQKEKKGVVPTRQALAVPTDSAVTVSSYAYTKVTGPYS
RWI

SEQ ID NO. 48: UniProt Q8BG87

MSQFQVPLAVQPDLSGLYDFPQGQVMVGGFQGPGLPMAGSETQLRGGGDGRKKRKRCGTCDPCR
RLENCGSCTSCTNRRTHQICKLRKCEVLKKKAGLLKEVEINAREGTGPWAQGATVKTGSELSPV
DGPVPGQMDSGPVYHGDSRQLSTSGAPVNGAREPAGPGLLGAAGPWRVDQKPDWEAASGPTHAA
RLEDAHDLVAFSAVAEAVSSYGALSTRLYETFNREMSREAGSNGRGPRPESCSEGSEDLDTLQT
ALALARHGMKPPNCTCDGPECPDFLEWLEGKIKSMAMEGGQGRPRLPGALPPSEAGLPAPSTRP
PLLSSEVPQVPPLEGLPLSQSALSIAKEKNISLQTAIAIEALTQLSSALPQPSHSTSQASCPLP
EALSPSAPFRSPQSYLRAPSWPVVPPEEHPSFAPDSPAFPPATPRPEFSEAWGTDTPPATPRNS
WPVPRPSPDPMAELEQLLGSASDYIQSVFKRPEALPTKPKVKVEAPSSSPAPVPSPISQREAPL
LSSEPDTHQKAQTALQQHLHHKRNLFLEQAQDASFPTSTEPQAPGWWAPPGSPAPRPPDKPPKE
KKKKPPTPAGGPVGAEKTTPGIKTSVRKPIQIKKSRSRDMQPLFLPVRQIVLEGLKPQASEGQA
PLPAQLSVPPPASQGAASQSCATPLTPEPSLALFAPSPSGDSLLPPTQEMRSPSPMVALQSGST
GGPLPPADDKLEELIRQFEAEFGDSFGLPGPPSVPIQEPENQSTCLPAPESPFATRSPKKIKIE
SSGAVTVLSTTCFHSEEGGQEATPTKAENPLTPTLSGFLESPLKYLDTPTKSLLDTPAKKAQSE
FPTCDCVEQIVEKDEGPYYTHLGSGPTVASIRELMEDRYGEKGKAIRIEKVIYTGKEGKSSRGC
PIAKWVIRRHTLEEKLLCLVRHRAGHHCQNAVIVILILAWEGIPRSLGDTLYQELTDTLRKYGN
PTSRRCGLNDDRTCACQGKDPNTCGASFSFGCSWSMYFNGCKYARSKTPRKFRLTGDNPKEEEV
LRNSFQDLATEVAPLYKRLAPQAYQNQVTNEDVAIDCRLGLKEGRPFSGVTACMDFCAHAHKDQ
HNLYNGCTVVCTLTKEDNRCVGQIPEDEQLHVLPLYKMASTDEFGSEENQNAKVSSGAIQVLTA

FPREVRRLPEPAKSCRQRQLEARKAAAEKKKLQKEKLSTPEKIKQEALELAGVTTDPGLSLKGG
LSQQSLKPSLKVEPQNHFSSFKYSGNAVVESYSVLGSCRPSDPYSMSSVYSYHSRYAQPGLASV
NGFHSKYTLPSFGYYGFPSSNPVFPSQFLGPSAWGHGGSGGSFEKKPDLHALHNSLNPAYGGAE
FAELPGQAVATDNHHPIPHHQQPAYPGPKEYLLPKVPQLHPASRDPSPFAQSSSCYNRSIKQEP
IDPLTQAESIPRDSAKMSRTPLPEASQNGGPSHLWGQYSGGPSMSPKRTNSVGGNWGVFPPGES
PTIVPDKLNSFGASCLTPSHFPESQWGLFTGEGQQSAPHAGARLRGKPWSPCKFGNGTSALTGP
SLTEKPWGMGTGDFNPALKGGPGFQDKLWNPVKVEEGRIPTPGANPLDKAWQAFGMPLSSNEKL
FGALKSEEKLWDPFSLEEGTAEEPPSKGVVKEEKSGPTVEEDEEELWSDSEHNFLDENIGGVAV
APAHCSILIECARRELHATTPLKKPNRCHPTRISLVFYQHKNLNQPNHGLALWEAKMKQLAERA
RQRQEEAARLGLGQQEAKLYGKKRKWGGAMVAEPQHKEKKGAIPTRQALAMPTDSAVTVSSYAY
TKVTGPYSRWI

SEQ ID NO. 49: UniProt P04519

MRICIFMARGLEGCGVTKFSLEQRDWFIKNGHEVTLVYAKDKSFTRTSSHDHKSFSIPVILAKE
YDKALKLVNDCDILIINSVPATSVQEATINNYKKLLDNIKPSIRVVVYQHDHSVLSLRRNLGLE
ETVRRADVIFSHSDNGDFNKVLMKEWYPETVSLFDDIEEAPTVYNFQPPMDIVKVRSTYWKDVS
EINMNINRWIGRTTTWKGFYQMFDFHEKFLKPAGKSTVMEGLERSPAFIAIKEKGIPYEYYGNR
EIDKMNLAPNQPAQILDCYINSEMLERMSKSGFGYQLSKLNQKYLQRSLEYTHLELGACGTIPV
FWKSTGENLKFRVDNTPLTSHDSGIIWFDENDMESTFERIKELSSDRALYDREREKAYEFLYQH
QDSSFCFKEQFDIITK

SEQ ID NO. 50: UniProt P04547

MKIAIINMGNNVINFKTVPSSETIYLFKVISEMGLNVDIISLKNGVYTKSFDEVDVNDYDRLIV
VNSSINFFGGKPNLAILSAQKFMAKYKSKIYYLFTDIRLPFSQSWPNVKNRPWAYLYTEEELLI
KSPIKVISQGINLDIAKAAHKKVDNVIEFEYFPIEQYKIHMNDFQLSKPTKKTLDVIYGGSFRS
GQRESKMVEFLFDTGLNIEFFGNAREKQFKNPKYPWTKAPVFTGKIPMNMVSEKNSQAIAALII
GDKNYNDNFITLRVWETMASDAVMLIDEEFDTKHRIINDARFYVNNRAELIDRVNELKHSDVLR
KEMLSIQHDILNKTRAKKAEWQDAFKKAIDL

SEQ ID NO. 51: ZDD motif

H-[P/A/V]-E-$X_{[23-28]}$-P-C-$X_{[2-4]}$-C

SEQ ID NO. 52: ZDD motif

HXEX_{24}SW(S/T)PC$X_{[2-4]}CX_6FX_8LX_5$R(L/I)Y$X_{[8-11]}LX_2LX_{[10]}$M

SEQ ID NO. 53: Removable P5 sequence

TTTTTTTTTTAATGATACGGCGACCACCGAUCTACAC
(where U = 2-deoxyuridine)

SEQ ID NO. 54: Removable P7 sequence

TTTTTTTTTTCAAGCAGAAGACGGCATACGA[G^{oxo}]AT
(where [G^{oxo}] = 8-oxoguanine)

SEQ ID NO. 55: Extended primer sequence with A as 5' additional nucleotide and P5' sequence (complementary to P5)

AGTGTAGATCTCGGTGGTCGCCGTATCATT

SEQ ID NO. 56: Extended primer sequence with T as 5' additional nucleotide and P5' sequence (complementary to P5)

TGTGTAGATCTCGGTGGTCGCCGTATCATT

SEQ ID NO. 57: Extended primer sequence with C as 5' additional nucleotide and P5' sequence (complementary to P5)

CGTGTAGATCTCGGTGGTCGCCGTATCATT

SEQ ID NO. 58: Extended primer sequence with G as 5' additional nucleotide and P5' sequence (complementary to P5)

GGTGTAGATCTCGGTGGTCGCCGTATCATT

SEQ ID NO. 59: Extended primer sequence with A as 5' additional nucleotide and P7' sequence (complementary to P7)

AATCTCGTATGCCGTCTTCTGCTTG

SEQ ID NO. 60: Extended primer sequence with T as 5' additional nucleotide and P7' sequence (complementary to P7)

TATCTCGTATGCCGTCTTCTGCTTG

SEQ ID NO. 61: Extended primer sequence with C as 5' additional nucleotide and P7' sequence (complementary to P7)

CATCTCGTATGCCGTCTTCTGCTTG

SEQ ID NO. 62: Extended primer sequence with G as 5' additional nucleotide and P7' sequence (complementary to P7)

GATCTCGTATGCCGTCTTCTGCTTG

SEQ ID NO. 63: Extended primer sequence with A as 5' additional nucleotide and alternative P5' sequence (complementary to alternative P5)

ATCGGTCGCCGTATCATT

SEQ ID NO. 64: Extended primer sequence with T as 5' additional nucleotide and alternative P5' sequence (complementary to alternative P5)

TTCGGTCGCCGTATCATT

SEQ ID NO. 65: Extended primer sequence with C as 5' additional nucleotide and alternative P5' sequence (complementary to alternative P5)

CTCGGTCGCCGTATCATT

SEQ ID NO. 66: Extended primer sequence with G as 5' additional nucleotide and alternative P5' sequence (complementary to alternative P5)

GTCGGTCGCCGTATCATT

SEQ ID NO. 67: UniProt Q24K09

MPARTAPARVPALASRAFSLPDDVRRRLKDLERDSLTEKECVKEKLNLLHEFLRTEIKNQLCDL
ETKLHKEELSEEGYLAKVKSLLNKDLSLENGAHAFSREANGCLENGSQTSGEDCRVVMAEKGKP
PKPVSRLYTPRRSKSDGETKSEVSSSPRITRKTTRQTTITSHFPRGPAKRKPEEEPEKVKSDDS
VDEEKDQEEKRRRVTSRERVAGLLPAEEPGRVRPGTHMEEEGRDDKEEKRLRSQTKEPTPKHKA
KEEPDRDVRPGGAQAEMNEGEDKDEKRHRSQPKDLASKRRPEEKEPERVKPQVSDEKDEDEKEE
KRRRTTYRELTEKKMTRTKIAVVSKTNPPKCTECLQYLDDPELRYEQHPPDAVEEIQILTNERL
SIFDANESGFESYEDLPQHKLTCFSVYCKRGHLCPIDTGLIEKDVELLFSGSAKPIYEDDPSPE
GGINGKNFGPINEWWIAGFDGGEKALLGFSTSFAEYILMDPSPEYAPLFSVMQEKIYISKIVVE
FLQSNPDSTYEDLINKIETTVPPCMLNLNRFTEDSLLRHAQFVVEQVESYDRAGDSDEQPIFLS
PCMRDLIKLAGVTLGKRRAERRQTIRQPAKEKDKGPTKATTTKLVYQIFDTFFAEQIEKDDKED
KENAFKRRRCGVCEICQQPECGKCKACKDMVKFGGSGRSKQACQKRRCPNMAMKEADDDEEVDD
NIPEMPSPKKMHQGKKKKQNKNRISWVGDAVKTDGKKSYYKKVCIDSETLEVGDCVSVIPDDSS
KPLYLARVTALWEDSSNGQMFHAHWFCAGTDTVLGATSDPLELFLVDECEDMQLSYIHSKVQVI
YKAPSENWAMEGGVDPEALMSEDDGKTYFYQLWYDQDYARFESPPKTQPTEDNKYKFCASCARL
AEMRQKEIPRVVEQLQDLEGRVLYSLATKNGVQYRVGDGVYLPPEAFTFNIKLSSPVKRPRKEP
VDEALYPEHYRKYSDYIKGSNLDAPEPYRIGRIKEIFCSKKSNGRPNETDIKIRVNKFYRPENT
HKSTPASYHADINLLYWSDEEAVVDFKAVQGRCTVEYGEDLPQCLQDFSAGGPDRFYFLEAYNA
KSKSFEDPPNHARSTGNKGKGKGKGKNRTKSQTCEPSELETEIKLPKLRTLDVFSGCGGLSEGF
HQAGISETLWAIEMWDPAAQAFRLNNPGSTVFTEDCNVLLKLVMAGEVTNSRGQKLPQKGDVEM
LCGGPPCQGFSGMNRFNSRTYSKFKNSLVVSFLSYCDYYRPRYFLLENVRNFVSFKRSMVLKLT
LRCLVRMGYQCTFGVLQAGQYGVAQTRRRAIILAAAPGEPLPLFPEPLHVFAPRACQLSVVVDD
KKFVSNITRLSSGPFRTITVRDTMSDLPEIRNGASALEISYNGEPQSWFQRQLRGSQYQPILRD
HICKDMSALVAARMRHIPLAPGSDWRDLPNIEVRLSDGTLARKLRYNYHDKKNGCSSSGALRGV
CSCVEGKPCEPAARQFNTLIPWCLPHTGNRHNHWAGLYGRLEWDGFFSTTVTNPEPMGKQGRVL
HPEQHRVVSVRECARSQGFPDTYRLFGNILDKHRQVGNAVPPPLAKAIGLEIKRCMLAKARESA
SAKIKEEAAKD

SEQ ID NO. 68: UniProt Q9Z330

MPARTAPARVPALASPAGSLPDHVRRRLKDLERDGLTEKECVKEKLNLLHEFLQTEIKSQLCDL
ETKLHKEELSEEGYLAKVKTLLNKDLCLENGTLSLTQKANGCPANGSRPTWKAEMADSNRSPRS
RPKPRGPRRSKSDSETMIEASSSSVATRRTTRQTTITSHFKGPAKRKPKEDSEKGNANESAAEE
RDQDKKRRVAGTESRASRAGESVEKPERVRPGTQLCQEEQGEQEDDRRPRRQTRELASRRKSRE
DPDREARPGTHLDVDDDDEKDKRSSRPRSQPRDLATKRRPKEEVEQITPEPPEGKDEDEREEKR
RKTTRKKPEPLSIPVQSRVERKASQGKASAIPKLNPPQCPECGQYLDDPDLKYQQHPVDAVDEP
QMLTNEALSVFDSNSSWFETYDSSPMHKFTFFSVYCSRGHLCPVDTGLIEKNVELYFSGVAKAI
HEENPSVEGGVNGKNLGPINQWWISGFDGGEKALIGFSTAFAEYFLMEPSPEYAPIFGLMQEKI
YISKIVVEFLQSNPDAVYEDLINKIETTVPPSAINVNRFTEDSLLRHAQFVVSQVESYDDAKDD
DETPIFLSPCMRSLIHLAGVSLGQRRATRRTVINSAKVKRKGPTKATTTKLVYQIFDTFFSEQI
EKDDKEDKENTMKRRRCGVCEVCQQPECGKCKACKDMVKFGGTGRSKQACLKRRCPNLAVKEAD
EDEEADDDIPELPSPKKLHQGKKKKQNKDRISWLGEPVKIEENRTYYWKVSIDEETLEVGDCVS
VIPDDPSKPLYLARVTALWEDKNGQMFHAHWFCAGTDTVLGATSDPLELFLVGECENMQLSYIH
SKVKVIYRGPSPNWAMEGGMDPEAMLPGAEDGKTYFYQFWYSQDYARFESPPKTQPAEDNKHKF
CLSCIRLAELRQKEMPKVLEQLEEVDGRVYCSSITKNGVVYRLGDSVYLPPEAFTFNIKMASPM
KRSKRDPVNENPVPRDTYRKYSDYIKGSNLDAPEPYRIGRIKEIYCGKKKGGKVNEADIKIRLY
KFYRPENTHKSIQATYHADINLLYWSDEEAVVDFSDVQGRCTVEYGEDLLESIQDYSQGGPDRF
YFLEAYNSKTKSFEDPPNHARSPGNKGKGKGKGKGKGKPQVSEPKEPEAAIKLPKLRTLDVFSG
CGGLTEGFHQAGISETLWAIEMWEPAAQAFRLNNPGTTVFTEDCNVLLKLVMAGEVTNSLGQRL
PQKGDVEMLCGGPPCQGFSGMNRFNSRTYSKFKNSLVVSFLSYCDYYRPRFFLLENVRNFVSFR
RSMVLKLTLRCLVRMGYQCTFGVLQAGQYGVAQTRRRAIILAAAPGEKLPLFPEPLHVFAPRAC
QLSVVVDDKKFVSNITRLSSGPFRTITMRDTMSDLPEIQNGASAPEISYKWRATVLVPEAAARV
ALPAHPQGPYPQVHERAGGCRMRHIPLSPGSDWRDLPNIQVRLRDGVITNKLRYTFHDTKNGCS
STGALRGVCSCAEGKTCDPASRQFNTLIPWCLPHTGNRHNHWAGLYGRLEWDGFFSTTVTNPEP

MGKQGRVLHPEQHRVVSVRECARSQGFPDTYRLFGNILDRHRQVGNAVPPPLAKAIGLEIKLCL
LASAQESASAAVKGKEETTTED

SEQ ID NO. 69: UniProt P13864

MPARTAPARVPALASPAGSLPDHVRRRLKDLERDGLTEKECVREKLNLLHEFLQTEIKSQLCDL
ETKLHKEELSEEGYLAKVKSLLNKDLSLENGTHTLTQKANGCPANGSRPTWRAEMADSNRSPRS
RPKPRGPRRSKSDSDTLSVETSPSSVATRRTTRQTTITAHFTKGPTKRKPKEESEEGNSAESAA
EERDQDKKRRVVDTESGAAAAVEKLEEVTAGTQLGPEEPCEQEDDNRSLRRHTRELSLRRKSKE
DPDREARPETHLDEDEDGKKDKRSSRPRSQPRDPAAKRRPKEAEPEQVAPETPEDRDEDEREEK
RRKTTRKKLESHTVPVQSRSERKAAQSKSVIPKINSPKCPECGQHLDDPNLKYQQHPEDAVDEP
QMLTSEKLSIYDSTSTWFDTYEDSPMHRFTSFSVYCSRGHLCPVDTGLIEKNVELYFSGCAKAI
HDENPSMEGGINGKNLGPINQWWLSGFDGGEKVLIGFSTAFAEYILMEPSKEYEPIFGLMQEKI
YISKIVVEFLQNNPDAVYEDLINKIETTVPPSTINVNRFTEDSLLRHAQFVVSQVESYDEAKDD
DETPIFLSPCMRALIHLAGVSLGQRRATRRVMGATKEKDKAPTKATTTKLVYQIFDTFFSEQIE
KYDKEDKENAMKRRRCGVCEVCQQPECGKCKACKDMVKFGGTGRSKQACLKRRCPNLAVKEADD
DEEADDDVSEMPSPKKLHQGKKKKQNKDRISWLGQPMKIEENRTYYQKVSIDEEMLEVGDCVSV
IPDDSSKPLYLARVTALWEDKNGQMMFHAHWFCAGTDTVLGATSDPLELFLVGECENMQLSYIH
SKVKVIYKAPSENWAMEGGTDPETTLPGAEDGKTYFFQLWYNQEYARFESPPKTQPTEDNKHKF
CLSCIRLAELRQKEMPKVLEQIEEVDGRVYCSSITKNGVVYRLGDSVYLPPEAFTFNIKVASPV
KRPKKDPVNETLYPEHYRKYSDYIKGSNLDAPEPYRIGRIKEIHCGKKKGKVNEADIKLRLYKF
YRPENTHRSYNGSYHTDINMLYWSDEEAVVNFSDVQGRCTVEYGEDLLESIQDYSQGGPDRFYF
LEAYNSKTKNFEDPPNHARSPGNKGKGKGKGKGKGKHQVSEPKEPEAAIKLPKLRTLDVFSGCG
GLSEGFHQAGISETLWAIEMWDPAAQAFRLNNPGTTVFTEDCNVLLKLVMAGEVTNSLGQRLPQ
KGDVEMLCGGPPCQGFSGMNRFNSRTYSKFKNSLVVSFLSYCDYYRPRFFLLENVRNFVSYRRS
MVLKLTLRCLVRMGYQCTFGVLQAGQYGVAQTRRRAIILAAAPGEKLPLFPEPLHVFAPRACQL
SVVVDDKKFVSNITRLSSGPFRTITVRDTMSDLPEIQNGASNSEIPYNGEPLSWFQRQLRGSHY
QPILRDHICKDMSPLVAARMRHIPLFPGSDWRDLPNIQVRLGDGVIAHKLQYTFHDVKNGYSST
GALRGVCSCAEGKACDPESRQFSTLIPWCLPHTGNRHNHWAGLYGRLEWDGFFSTTVTNPEPMG
KQGRVLHPEQHRVVSVRECARSQGFPDSYRFFGNILDRHRQVGNAVPPPLAKAIGLEIKLCLLS
SARESASAAVKAKEEAATKD

SEQ ID NO. 70: UniProt Q92072

MPARSAPPPPALPPALRRRLRDLERDEDSLSEKETLQEKLRLTRGFLRAEVQRRLSALDADVRC
RELSEERYLAKVKALLRRELAAENGDAAKLFSRASNGCAGNGEEEWERGGRGEDGAMEVEEAAA
SSSSSSSSSSSSSSSSSSSSLLPAPRARKARRSRSNGESKKSPASSRVTRSSGRQPTILSVFS
KGSTKRKSEEVNGAVKPEVSAEKDEEEEEELEEKEQDEKRIKIETKEGSEIKDEITQVKTSTPA
KTTPPKCVDCRQYLDDPDLKFFQGDPDDALEEPEMLTDERLSIFDANEDGFESYEDLPQHKVTS
FSVYDKRGHLCPFDTGLIERNIELYFSGAVKPIYDDNPCLDGGVRAKKLGPINAWWITGFDGGE
KALIGFTTAFADYILMEPSEEYAPIFALMQEKIYMSKIVVEFLQNNRDVSYEDLLNKIETTVPP
VGLNFNRFTEDSLLRHAQFVVEQVESYDEAGDSDEPPVLITPCMRDLIKLAGVTLGKRRAVRRQ
AIRHPTRIDKDKGPTKATTTKLVYLIFDTFFSEQIEKDEREDDKENAMKRRRCGVCEVCQQPEC
GKCKACQNMVKFGGSGRSKQACLQRRCPNLAVREADEDEEVDDNIPEMPSPKKMLQGRKKKQNK
SRISWVGEPIKSDGKKDFYQRVCIDSETLEVGDCVSVSPDDPTKPLYLARVTAMWEDSSGQMFH
AHWFCPGSDTVLGATSDPLELFLVDECEDMQLSYIHGKVNVIYKPPSENWAMEGGLDMEIKMVE
DDGRTYFYQMWYDQEYARFETPPRAQPMEDNKYKFCLSCARLDEVRHKEIPKVAEPLDEGDGKM
FYAMATKNGVQYRVGDSVYLLPEAFSFSMKPASPAKRPKKEAVDEDLYPEHYRKYSEYIKGSNL
DAPDPYRVGRIKEIFCHIRTNGKPNEADIKLRIWKFYRPENTHKSMKATYHADINLLYWSDEET
TVDFCAVQGRCTVVYGEDLTESIQDYSAGGLDRFYFLEAYNAKTKSFEDPPNHARSSGNKGKGK
GKGKGKGKGKSSTTCEQSEPEPTELKLPKLRTLDVFSGCGGLSEGFHQAGVSETLWAIEMWEPA
AQAFRLNNPGTTVFTEDCNVLLKLVMSGEKTNSLGQKLPQKGDVEMLCGGPPCQGFSGMNRFNS
RTYSKFKNSLVVSFLSYCDYYRPRFFLLENVRNFVSFKRSMVLKLTLRCLVRMGYQCTFGVLQA
GQYGVAQTRRRAIVLAAAPGEKLPMFPEPLHVFAPRACQLSVVVDDKKFVSNITRTYSGPFRTI
TVRDTMSDLPEIRNGASALEISYNGEPQSWFQRQIRGSQYQPILRDHICKDMSALVAARMRHIP
LAPGSDWRDLPNIEVRLSDGTSTRKLRYTHHEKKNGRSSSGALRGVCSCAEGKPCDPADRQFNT

LIPWCLPHTGNRHNHWAGLYGRLEWDGFFSTTVTNPEPMGKQGRVLHPEQHRVVSVRECARSQG
FPDTYRLFGNILDKHRQVGNAVPPPLAKAIGLEIRACVGARMREESGAAVAPPAPEKMEMTAAA
D

SEQ ID NO. 71: UniProt P26358

MPARTAPARVPTLAVPAISLPDDVRRRLKDLERDSLTEKECVKEKLNLLHEFLQTEIKNQLCDL
ETKLRKEELSEEGYLAKVKSLLNKDLSLENGAHAYNREVNGRLENGNQARSEARRVGMADANSP
PKPLSKPRTPRRSKSDGEAKPEPSPSPRITRKSTRQTTITSHFAKGPAKRKPQEESERAKSDES
IKEEDKDQDEKRRRVTSRERVARPLPAEEPERAKSGTRTEKEEERDEKEEKRLRSQTKEPTPKQ
KLKEEPDREARAGVQADEDEDGDEKDEKKHRSQPKDLAAKRRPEEKEPEKVNPQISDEKDEDEK
EEKRRKTTPKEPTEKKMARAKTVMNSKTHPPKCIQCGQYLDDPDLKYGQHPPDAVDEPQMLTNE
KLSIFDANESGFESYEALPQHKLTCFSVYCKHGHLCPIDTGLIEKNIELFFSGSAKPIYDDDPS
LEGGVNGKNLGPINEWWITGFDGGEKALIGFSTSFAEYILMDPSPEYAPIFGLMQEKIYISKIV
VEFLQSNSDSTYEDLINKIETTVPPSGLNLNRFTEDSLLRHAQFVVEQVESYDEAGDSDEQPIF
LTPCMRDLIKLAGVTLGQRRAQARRQTIRHSTREKDRGPTKATTTKLVYQIFDTFFAEQIEKDD
REDKENAFKRRRCGVCEVCQQPECGKCKACKDMVKFGGSGRSKQACQERRCPNMAMKEADDDEE
VDDNIPEMPSPKKMHQGKKKKQNKNRISWVGEAVKTDGKKSYYKKVCIDAETLEVGDCVSVIPD
DSSKPLYLARVTALWEDSSNGQMFHAHWFCAGTDTVLGATSDPLELFLVDECEDMQLSYIHSKV
KVIYKAPSENWAMEGGMDPESLLEGDDGKTYFYQLWYDQDYARFESPPKTQPTEDNKFKFCVSC
ARLAEMRQKEIPRVLEQLEDLDSRVLYYSATKNGILYRVGDGVYLPPEAFTFNIKLSSPVKRPR
KEPVDEDLYPEHYRKYSDYIKGSNLDAPEPYRIGRIKEIFCPKKSNGRPNETDIKIRVNKFYRP
ENTHKSTPASYHADINLLYWSDEEAVVDFKAVQGRCTVEYGEDLPECVQVYSMGGPNRFYFLEA
YNAKSKSFEDPPNHARSPGNKGKGKGKGKGKPKSQACEPSEPEIEIKLPKLRTLDVFSGCGGLS
EGFHQAGISDTLWAIEMWDPAAQAFRLNNPGSTVFTEDCNILLKLVMAGETTNSRGQRLPQKGD
VEMLCGGPPCQGFSGMNRFNSRTYSKFKNSLVVSFLSYCDYYRPRFFLLENVRNFVSFKRSMVL
KLTLRCLVRMGYQCTFGVLQAGQYGVAQTRRRAIILAAAPGEKLPLFPEPLHVFAPRACQLSVV
VDDKKFVSNITRLSSGPFRTITVRDTMSDLPEVRNGASALEISYNGEPQSWFQRQLRGAQYQPI
LRDHICKDMSALVAARMRHIPLAPGSDWRDLPNIEVRLSDGTMARKLRYTHHDRKNGRSSSGAL
RGVCSCVEAGKACDPAARQFNTLIPWCLPHTGNRHNHWAGLYGRLEWDGFFSTTVTNPEPMGKQ
GRVLHPEQHRVVSVRECARSQGFPDTYRLFGNILDKHRQVGNAVPPPLAKAIGLEIKLCMLAKA
RESASAKIKEEEAAKD

SEQ ID NO. 72: UniProt Q27746

MPSKTICDQVIPPNVRDRVQELDGDLNDGLITEKGYVKKKSKILFEHLSPDIQTKLKGLEDELK
DEELTEKGYLNKVQSILAKFIETCSPVNGDTKEEASSNGKDDEKAESTVANGTTSNGSTTNGSS
GSSKANGHTNGGYVQSSSQEETGTSQSEEEMDMDTPTSGKGGSKKKKKSKGSGGGDAGKGRKRK
VLGDDERDGVEKKEGEKKDVEGEEGEEAKEESATPDEKTLRTSKRKRSPKADAKQPSIMSMFTK
KPAKKEEEKMEESSSMEVDKKEMENGDNGKKEEEEPSGPGGKRIKKEEEEEEKAKVEPMSPSRD
LRHKANHETAESKQPPLRCKECRQLLDDPDLKIFPGDPEDAREEYITLTDPRLSLLTGDEGDAM
SYDERLQHKITNFCVYDKSTHICAFDRGMIEKNKELYFSGYVKPIYDDNPSTEGGIPTKRIGPI
NEWYTTGFDGGHKALIGFSTAFAEYIVMSPSEEYKPFWTAVQEKIYMSKILIEFLQNNVDPVYE
DLLTQIETTVPPEGCNRFTEDSLLRHAQFVVEQVESYDDAADRDEVLLITMPCMRDLIKLAGVT
LGKRRAARKAAAVKKDKKPVFTMATVTPLVSHIFDAIFKDQIADEMKAAASERKKRCGVCEICQ
APDCGKCTACKDMIKFGGSGKAKQACKDRRCPNMAVQEADENDIDEMDNSSNKENKDEKKAKKG
RKLETPLKKKKRAKVTWLDEPTEVTEERAYYKAAMLDDEKIEIGDCVLIHPDDPTKPLFMARVI
YMWQESQGEMMFHAQWFVYGSETVLGETSDPLEVFPIDECQDTYLGSVNAKCTVIYKAPPNDWS
MIGGIDDPETDHVIKEDDGKTFFYQKWYDPELARFEDYEVLMAPDDIPAHRFCSCCLKNERAQE
KETARPGAKLEDQDDSSKVLYSSWHYKGNEFQIGDGVYLLPEVFSFNIKQKVVTKKPVSKKDVD
EDLYPENYRKSSEYVKGSNLECPEPFRIGKIISIYTTKSNSTVRLRVNKMYRPEDTHKGRTAAY
QADLNVLYWSEEEAVTELEVVQGKCSVVCAEDLNVSTDEYSAGGPHKFYFREAYDSERKCFEDP
PSKSRSTRMKGKGKGKGKGKAKGKIAVEKEEEKESTETPFNKLKCLDVFAGCGGLSEGFHQAGI
CESSWAIEKEEPAAQAYRLNNPGSTVFSDDCNELLRLVMQGEKTSRTGQKLPQKGDVELLCGGP
PCQGFSGMNRFNSREYSKFKNSLISSYLSYCDYYRPRFFLLENVRNFVSYKKNMVLKLALRCLI

RMGYQCTFGILQAGQYGVPQTRRRAIILAAAPGEKLPFYPEPLHVFSSRACSLSVMIGEKKIES
NNQWCLSAPYRTITVRDTMSDLPTINNGAQKLEISYDGEPQSDFQKKIRGNQYQPILRDHICKD
MSSLVAARMKHIPLAPGSDWRDLPNIPVTLKDGTTCRKLRYTHKDKKNGKSSTGALRGVCSCAE
GDACDPSDRQFSTLIPWCLPHTGNRHNNWAGLYGRLEWDGFFSTTVTNPEPMGKQGRVLHPEQH
RVVSVRECARSQGFPDTYRFFGSILDKHRQIGNAVPPPMAAAIGMEIKVCLQTKTKRDQERAAL
EPVKEETEESMD

SEQ ID NO. 73: UniProt J9VI03

MTTALTFGGGLFKDNTKFDIDMRGTADGAVNGGNIPNSQSQKRKRASPSPEIESEEDGDDWYEI
DYIADSRVIRRKGRQILQYLIHWAGYAVHERTWEDEDGIGGEDCALVQEFYRKNPGKPRLSPSS
VRKEVKLARMVEVVITTRRIDGKSRAASSTDQPSPHRLGITSPQANNIGGEDPNPSLTRRPVRS
TVSEIAKRPTSKKVHPNKKCKASSDDESDFVFEEGEWDEDEDDDNDVDFRSSEDDEDDEQERSA
EEPESDEEIIKPAKKTKSSLPKAKLRPKPANLGGFVTGVRPLNQGLDIKAAVRNMSDDLPPISD
IEAMFDHLVSRIPDIVELVRQLNGRKLRVATMCSGTESPLLALNMIAKAIKAQHGLTLAFEHVF
SCEIEPFKQAYIERNFTPPILFRDVTELGKKRAHTAYGSMVDVPGDVDILIAGTSCVDYSNLNN
VQQDIDANGESGRTFRGMLQWVKKHQPPIVILENVCNAPWDKVVEYFGQIDYDAQYTRLDTKEF
YIPHTRTRVYLFATPSSSESDDLPEKWAQTVKDLRRPWSSPFEAFLLHTDDPNIHRARLELASA
RAQTDGTSRKTTDWNRCESRHQRARQDEALGLLRPLTSWQEAGVCKGLDWTWNDWLLAQTERVV
DLLEISTLRMAKDGIDSGFKACIWNVSQNVDRQTGSSKTALAPCLTPNMIPWVTIRGGPVTGRE
ALALQGIPVRELLLTSENEDQLADLAGNAMTTTVVGSAMIAALKVACHKITEGANPEKEAALIL
EKEAVDDEQVANRIIGEDYLEHHDLDLAKVTKSNLSEILDLACRSSRHCQCEGQSGTAPNILEC
QECSYRACKSCGGRPEHVYAPCANQRVEPAEFEKRFKGLLPMRVRIAGLTDQCLNAVRKAAEKS
NKGSVNDNDWQLWSTALLEGIHDAEFRFRYLKRQSTWTAVYEARRAMLSLVLRNQIPEWRLTIK
APASEPNNSQLRALLLHPVARLQIDIAGQDVLCGPWELCIPSMKTIDIEITGKGELLPSWQASL
GLQGPFANTTRWSEVEISLQAEDENTLDRKLSGTYQLLPRCGQAMSSLHKKRPDLSDDGLPQLY
FFLDPTRCGESREDRYVFSTSTERLDYGTERPVIARLDSHWREGNEKQRKVKLDVSGAWVKCPE
AHLTAIGGDDIAVVANDAAANEIHRDRATFAIPSSASAISASLTTEGCSHAMALLSCRVPLDPT
HSESMWRRGAWAEIDLSHQGNTTFANLAWITERLPPLDGLKNWAHIADDVSEHVCERCAPRPPK
IHWIKREGKANKKGNKTKSTIIAFEDKLEAGQYEHALKHRPSPFVVQLRLDQDIGSFRIGLNIV
SLAHRALSRLPPTTSEHKISLSWRLTPGHVTESPQPRRVFILPSNKQDPENSQPEAFKLPLRKE
QLRSLWWMLEQEKATGKTHTFVEEEISESLLPAVGWRAEGKAERPVMVRGGVIADQVGYGKTVI
SIALVAQTLSLPAPEPATPGLIDLKATLIVVPGHLSKQWPNEIARFTGSMFKVIVIQGMKDLQE
KTIAELGKADIIVMASEIFESDVYWSRLEYLSAQPREWLHDTQGGRFFCDRLDAAMESLVSQTK
ILKEKGSEAAMRAMEDKKKSLVDNVGSKKEVHTAVNFGKRMKGQAYRDKHDSDSKAKPITKEEL
ERWEASEDEDDDENSKTYIPIPKFHSFTGSESIFSASVKKDYKLLPNPVLHMFRFRRVIADEFT
YLQKKSLAAVLRLSSSYRWILSGTPPVSDFAAIRSIATFMGIHLGVEDDGEGDVQYQKARAKDQ
TQAEKFHAFREVHSRAWHNRRDELAQEFLNVFVRQNIAEIEDIPTVEHIHTFKLPASEGAVYLE
LEHHLQALEMQARKETKFKNVSQGDRNARLEEALSDSKTAEEALLKRCCHFTLDLSDKTQDAKS
AQEACDHITSARARQLLACQEDLSRSVNQAIALHGWIKKKGGFSKNDDERQPFAEWIAFSSNIS
KHQGDIEAARILLKVIEKCGVKDGNIPPSPSDKQSPSIASGAKMDDVKWQLREQTHLLRKLVKE
LVARVRSLRFFEVVRKIQKGKSDAQIVLESSECGHKPSTNPDIEMAILSCCGHVACHKCMRKAA
ASQRCVKSGECQAAVRPTNIVKVSSLGVEGELSSGRYGAKLEHLVNLIHSIPKNERVLVFLQWE
DLAGKVSEALSAGRIPHVTLSGSAKSRANTLDRFQSTNADSARVLLLKMNDASAAGSNLTTANH
AVFLGPLFTNSLFNYRAVETQAIGRVRRYGQQKKVHIHRLLALDTIDMTIFNARRTELKEKTDW
EEIPQEEYKGRGSSISMTNEKRTPTLTVKSNPFKRSSSWALASSFRSKKRSMEARDAEGVSDDD
ENSELSDII

CLAIMS:

1. A method of preparing polynucleotide templates for distinguishing between modified cytosines, comprising:

(a) providing a polynucleotide library hairpin strand comprising:

a double-stranded polynucleotide comprising a forward library strand and a reverse library strand,

a hairpin loop adaptor ligated to an end of the double-stranded polynucleotide, wherein the hairpin loop adaptor comprises a cleavable site,

wherein the polynucleotide library hairpin strand has been generated from a precursor polynucleotide library hairpin strand such that any CpG dyads in the precursor polynucleotide library hairpin comprising only unmodified cytosine are converted to a first dyad in the polynucleotide library hairpin strand, any CpG dyads in the precursor polynucleotide library hairpin comprising 5-methylcytosine are converted to a second dyad in the polynucleotide library hairpin strand, and any CpG dyads in the precursor polynucleotide library hairpin comprising 5-hydroxymethylcytosine are converted to a third dyad in the polynucleotide library hairpin strand,

wherein the first dyad, second dyad and third dyad are different to each other when read; and

(b) synthesising at least one template strand by generating a complement of the polynucleotide library hairpin strand, each of the template strands comprising a forward template strand complementary to the forward library strand, a spacer strand complementary to the hairpin loop adaptor, and a reverse template strand complementary to the reverse library strand, wherein the spacer strand comprises a first cleavable site.

2. A method according to claim 1, wherein the method further comprises a step of:

(c) synthesising at least one template complement strand by generating a complement of the template strand, each of the template complement strands comprising a forward complement template strand, a spacer complement strand, and a reverse complement template strand, wherein the spacer complement strand comprises a second cleavable site.

3. A method according to claim 2, wherein the method further comprises a step of:

(d) cleaving the first cleavable site on the at least one template strand to generate at least one first polynucleotide sequence each comprising a first portion and cleaving the second cleavable site on the at least one template complement strand to generate at least one second polynucleotide sequence each comprising a second portion,

wherein the first portion corresponds with the forward template strand and the second portion corresponds with the reverse complement template strand, or wherein the first portion corresponds with the reverse template strand and the second portion corresponds with the forward complement template strand.

4. A method according to any one of claims 1 to 3, wherein the first cleavable site is a first restriction site for an endonuclease.

5. A method according to any one of claims 2 to 4, wherein the second cleavable site is a second restriction site for an endonuclease.

6. A method according to any one of claims 3 to 5, wherein the at least one first polynucleotide sequence each comprise a first sequencing primer binding site.

7. A method according to claim 6, wherein the first sequencing primer binding site is located after a 3'-end of the first portion.

8. A method according to any one of claims 3 to 7, wherein the at least one second polynucleotide sequence each comprise a second sequencing primer binding site; optionally wherein the second sequencing primer binding site is located after a 3'-end of the second portion.

9. A method according to any one of claims 3 to 8, wherein where CpG dyads comprising only unmodified cytosine were present in the precursor polynucleotide library hairpin, then a double C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; where CpG dyads comprising 5-methylcytosine were present in the precursor polynucleotide library hairpin, then a double mismatch is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; and where CpG dyads comprising 5-hydroxymethylcytosine were present in the precursor polynucleotide library hairpin, then a single mismatch and single C-C/G-G match is present when comparing corresponding

positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad.

10. A method according to any one of claims 3 to 8, wherein where CpG dyads comprising only unmodified cytosine were present in the precursor polynucleotide library hairpin, then a double mismatch is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; where CpG dyads comprising 5-methylcytosine were present in the precursor polynucleotide library hairpin, then a double C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad; and where CpG dyads comprising 5-hydroxymethylcytosine were present in the precursor polynucleotide library hairpin, then a single mismatch and single C-C/G-G match is present when comparing corresponding positions in the at least one first polynucleotide sequence and the at least one second polynucleotide sequence corresponding to the CpG dyad.

11. A method according to any one of claims 3 to 10, wherein the method further comprises a step of preparing the first portion and the second portion for concurrent sequencing.

12. A method according to claim 11, wherein the method comprises simultaneously contacting first sequencing primer binding sites located after a 3'-end of the first portions with first primers and second sequencing primer binding sites located after a 3'-end of the second portions with second primers.

13. A method according to any one of claims 3 to 12, wherein the method further comprises a step of processing the at least one first polynucleotide sequence comprising a first portion and the at least one second polynucleotide sequence comprising a second portion, such that a proportion of first portions are capable of generating a first signal and a proportion of second portions are capable of generating a second signal.

14. A method according to claim 13, wherein the processing involves selective processing to cause an intensity of the first signal to be greater than an intensity of the second signal.

15. A method according to claim 14, wherein a concentration of the first portions capable of generating the first signal is greater than a concentration of the second portions capable of generating the second signal.

16. A method according to claim 15, wherein a ratio between the concentration of the first portions capable of generating the first signal and the concentration of the second portions capable of generating the second signal is between 1.25:1 to 5:1; optionally wherein the ratio is between 1.5:1 to 3:1 or 2:1.

17. A method according to any one of claims 14 to 16, wherein selective processing comprises preparing for selective sequencing or conducting selective sequencing.

18. A method according to any one of claims 14 to 16, wherein selectively processing comprises conducting selective amplification.

19. A method according to any one of claims 14 to 17, wherein selectively processing comprises contacting first sequencing primer binding sites located after a 3'-end of the first portions with first primers and contacting second sequencing primer binding sites located after a 3'-end of the second portions with second primers, wherein the second primers comprises a mixture of blocked second primers and unblocked second primers.

20. A method according to any one of claims 14 to 16 or 18, wherein the selective processing comprises selectively removing some or substantially all of second immobilised primers that are not yet extended, and conducting a further amplification cycle in order to selectively amplify the first polynucleotide sequence(s) relative to the second polynucleotide sequence(s).

21. A method according to any one of claims 14 to 16 or 18, wherein selectively processing comprises selectively blocking some or substantially all of second immobilised primers that are not yet extended using a primer blocking agent, wherein the primer blocking agent is configured to limit or prevent synthesis of a strand extending from the second immobilised primer, and conducting a further amplification cycle in order to selectively amplify the first polynucleotide sequence(s) relative to the second polynucleotide sequence(s).

22. A method according to claim 21, wherein the primer blocking agent is added whilst first polynucleotide sequence(s) are hybridised to the second immobilised primers.

23. A method according to claim 21, wherein the method comprises contacting some or substantially all of the second immobilised primers with an extended primer sequence,

wherein the extended primer sequence is substantially complementary to the second immobilised primer and further comprises a 5' additional nucleotide; and adding the primer blocking agent, wherein the primer blocking agent is complementary to the 5' additional nucleotide.

24. A method according to any one of claims 13 to 23, wherein the first signal and the second signal are spatially resolved; or wherein the first signal and the second signal are spatially unresolved.

25. A method according to any one of claims 3 to 24, wherein the at least one first polynucleotide sequence comprising the first portion and the at least one second polynucleotide sequence comprising the second portion are attached to a solid support; optionally wherein the solid support is a flow cell.

26. A method according to claim 25, wherein the at least one first polynucleotide sequence comprising the first portion and the at least one second polynucleotide sequence comprising the second portion forms a cluster on the solid support.

27. A method according to claim 25 or claim 26, wherein the at least one first polynucleotide sequence comprising the first portion and the at least one second polynucleotide sequence comprising the second portion form a duoclonal cluster.

28. A method of sequencing polynucleotide sequences to distinguish between modified cytosines, comprising:

preparing polynucleotide templates for distinguishing between modified cytosines using a method according to any one of claims 3 to 27;

sequencing nucleobases in the first portion and the second portion; and

identifying the presence of 5-methylcytosine or 5-hydroxymethylcytosine by detecting differences when comparing a sequence output from the first portion with a sequence output from the second portion.

29. A method according to claim 28, wherein the step of sequencing nucleobases in the first portion and the second portion involves concurrent sequencing of nucleobases in the first portion and the second portion.

30. A kit comprising instructions for preparing polynucleotide templates for distinguishing between modified cytosines according to any one of claims 1 to 27, and/or for sequencing

polynucleotide sequences to distinguish between modified cytosines according to claim 28 or claim 29.

31. A data processing device comprising means for carrying out a method according to any one of claims 1 to 29; optionally wherein the data processing device is a polynucleotide sequencer.

32. A computer program product comprising instructions which, when the program is executed by a processor, cause the processor to carry out a method according to any one of claims 1 to 29.

33. A computer-readable storage medium comprising instructions which, when executed by a processor, cause the processor to carry out a method according to any one of claims 1 to 29.

34. A computer-readable data carrier having stored thereon a computer program product according to claim 32.

35. A data carrier signal carrying a computer program product according to claim 32.

36. A method of preparing a polynucleotide library hairpin strand, comprising:

(a) providing a double-stranded polynucleotide comprising a precursor forward library strand and a precursor reverse library strand; and

(b) ligating a hairpin loop adaptor to an end of the double-stranded polynucleotide to generate a first hairpin polynucleotide, wherein the hairpin loop adaptor comprises a cleavable site.

37. A method according to claim 36, wherein the hairpin loop adaptor comprises a base-paired stem and a non-base-paired loop.

38. A method according to claim 36 or claim 37, wherein the cleavable site is located in the non-base-paired loop.

39. A method according to any one of claims 36 to 38, wherein the hairpin loop adaptor connects a 3'-end of the precursor forward library strand with a 5'-end of the precursor

reverse library strand; or wherein the hairpin loop adaptor connects a 3'-end of the precursor reverse library strand with a 5'-end of the precursor forward library strand.

40. A method according to any one of claims 36 to 39, wherein the cleavable site is a restriction site for an endonuclease.

41. A method according to any one of claims 36 to 40, wherein the method further comprises a step of:

(c) removing the precursor reverse library strand from the first hairpin polynucleotide to generate a second hairpin polynucleotide comprising the precursor forward library strand and the hairpin loop adaptor, wherein the hairpin loop adaptor comprises the cleavable site.

42. A method according to claim 41, wherein the method further comprises a step of:

(d) forming a resynthesised reverse library strand from the second hairpin polynucleotide to generate a third hairpin polynucleotide, wherein when any cytosine bases are present in the resynthesised reverse library strand, then all such cytosine bases are unmodified cytosine.

43. A method according to claim 42, wherein the method further comprises a step of:

(e) exposing the third hairpin polynucleotide to an enzyme configured to convert hemimethylated 5-methylcytosine CpG dyads to fully methylated 5-methylcytosine CpG dyads, but not convert hemimethylated 5-hydroxymethylcytosine dyads, in order to generate a fourth hairpin polynucleotide.

44. A method according to claim 43, wherein the enzyme configured to convert hemimethylated 5-methylcytosine CpG dyads to fully methylated 5-methylcytosine CpG dyads, but not convert hemimethylated 5-hydroxymethylcytosine dyads is a DNA methyltransferase; optionally wherein the enzyme configured to convert hemimethylated 5-methylcytosine CpG dyads to fully methylated 5-methylcytosine CpG dyads, but not convert hemimethylated 5-hydroxymethylcytosine dyads is a member of the DNA methyltransferase 1 (DNMT1) family or the DNA methyltransferase 5 (DNMT5) family.

45. A method according to claim 43 or claim 44, wherein the method further comprises a step of:

(f) exposing the fourth hairpin polynucleotide to a conversion agent configured to convert 5-methylcytosine and 5-hydroxymethylcytosine to thymine or a nucleobase

which is read as thymine/uracil, or to a conversion agent configured to convert an unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil, in order to generate a fifth hairpin polynucleotide.

46. A method according to claim 45, wherein the conversion agent comprises a chemical agent and/or an enzyme.

47. A method according to any one of claims 42 to 46, wherein the method further comprises a step of ligating a flanking adaptor to an end of the double-stranded polynucleotide away from the hairpin loop adaptor to the third hairpin polynucleotide, the fourth hairpin polynucleotide or the fifth hairpin polynucleotide, wherein the flanking adaptor comprises a primer-binding sequence and a primer-binding complement sequence.

48. A method according to claim 47, wherein the flanking adaptor is a forked adaptor comprising a base-paired stem, a first arm and a second arm.

49. A method according to claim 47 or claim 48, wherein the primer-binding sequence is located on the first arm, and the primer-binding complement sequence is located on the second arm.

50. A polynucleotide library hairpin strand prepared according to any one of claims 36 to 49.

Figure 1

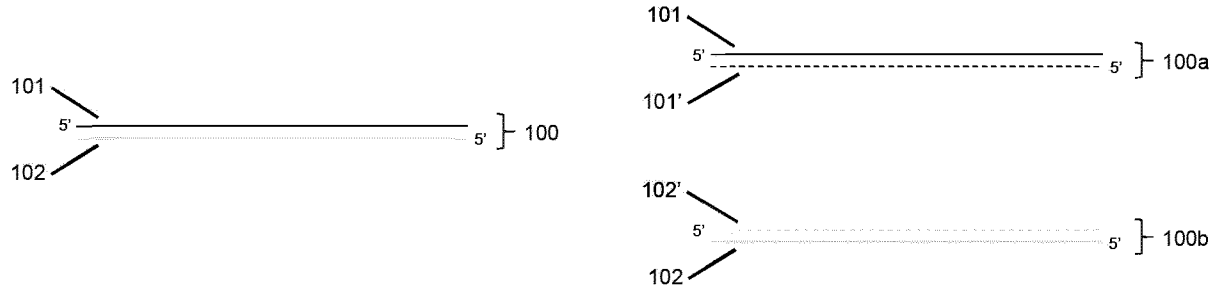


Figure 2

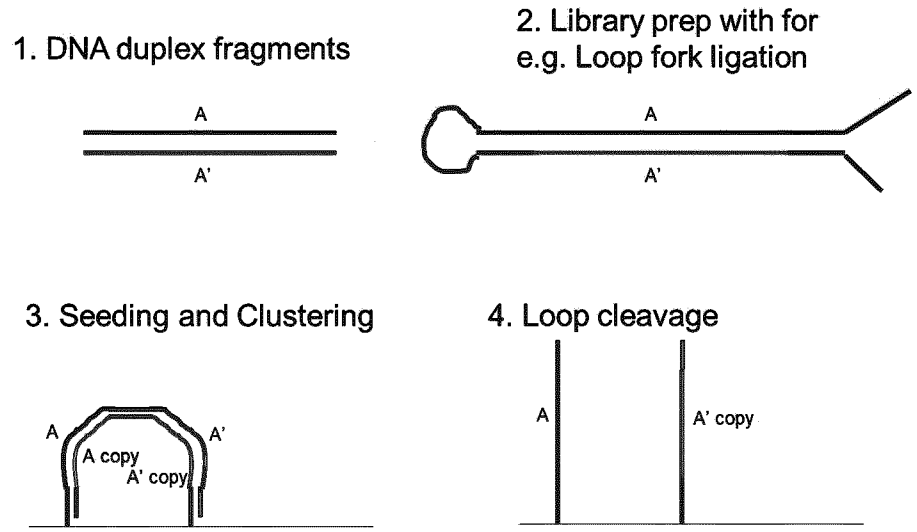


Figure 3

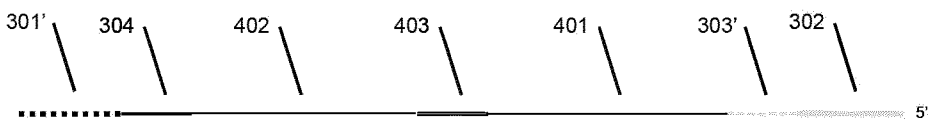


Figure 4

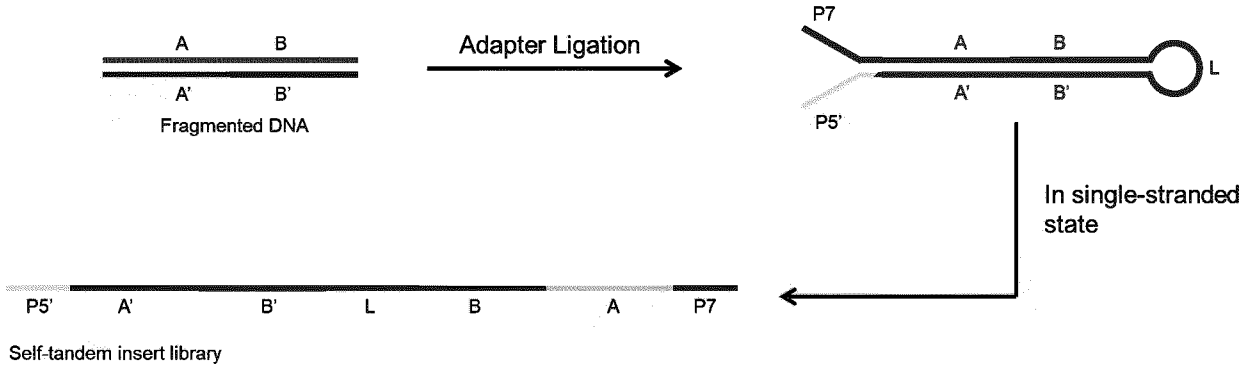


Figure 5

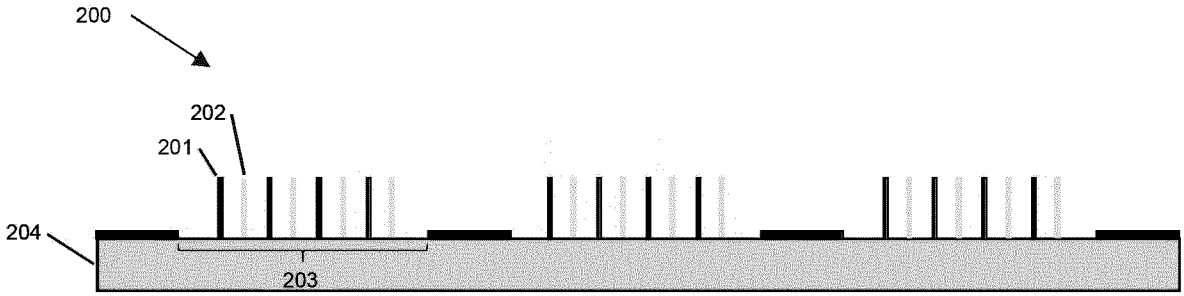
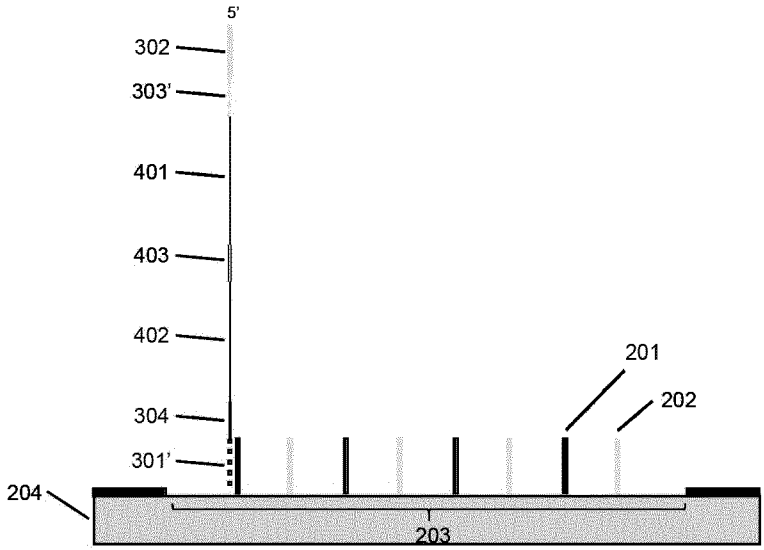


Figure 6

(A)


5'
302
303'
401
403
402
304
301'
204
201
202
203

(B)

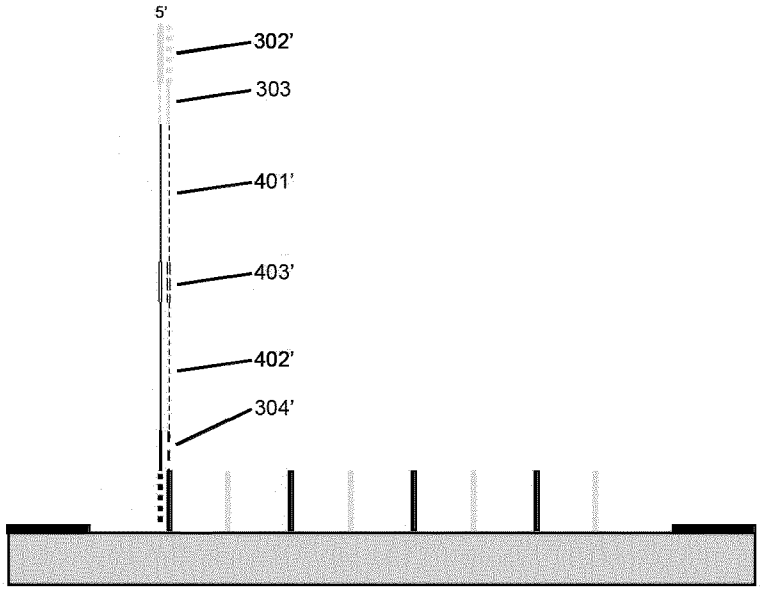
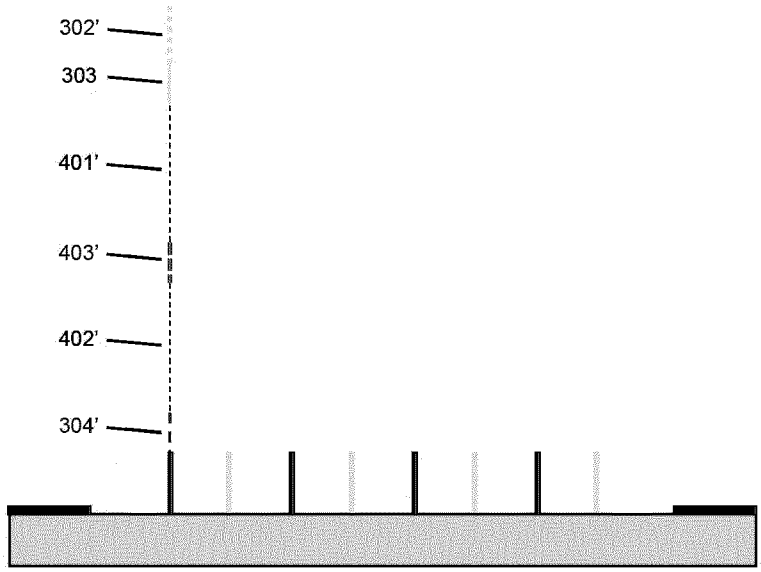

5'
302'
303
401'
403'
402'
304'

Figure 6 (cont.)

(C)



(D)

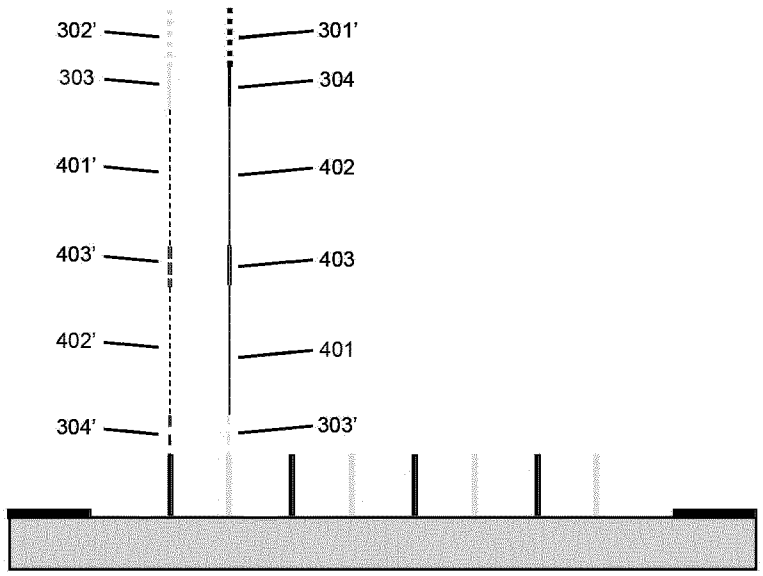


Figure 6 (cont.)

(E)

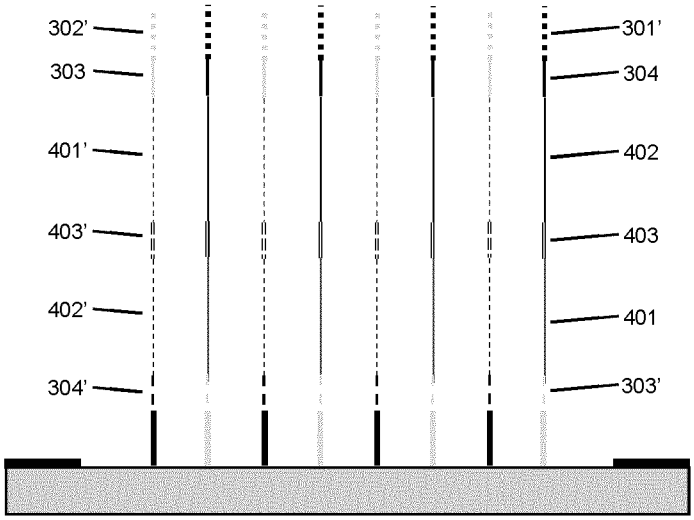


Figure 7

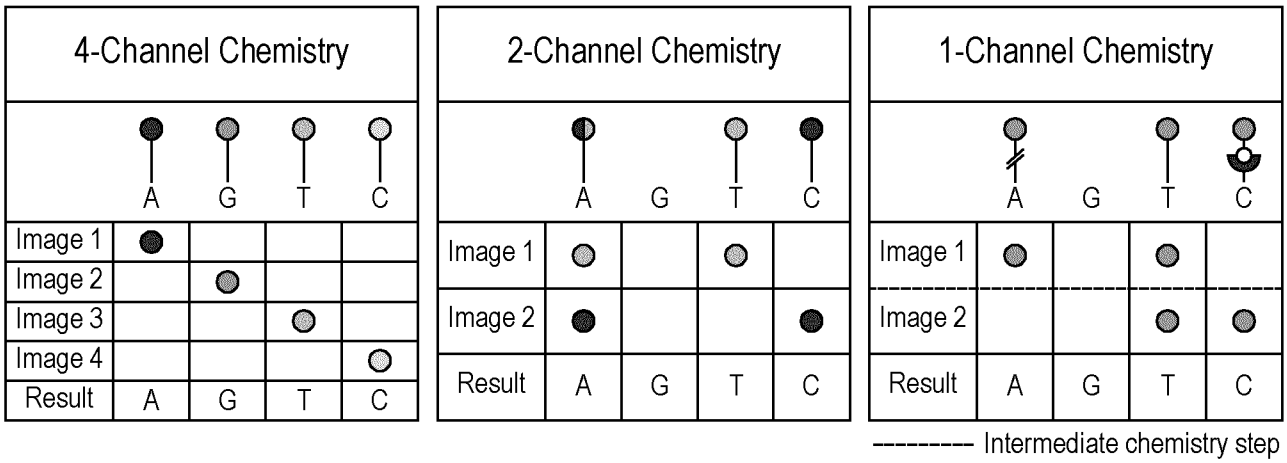


Figure 8

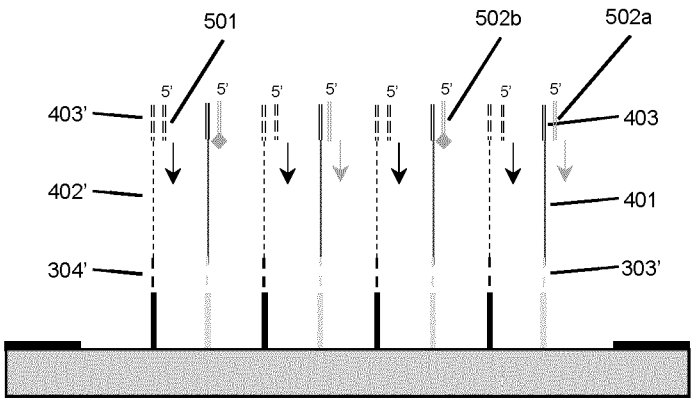
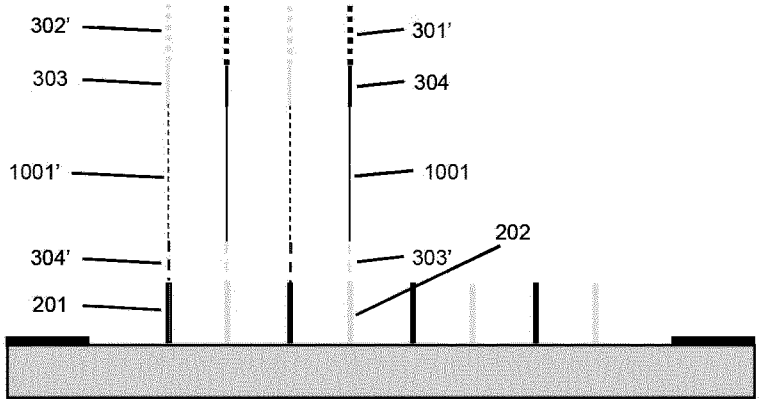
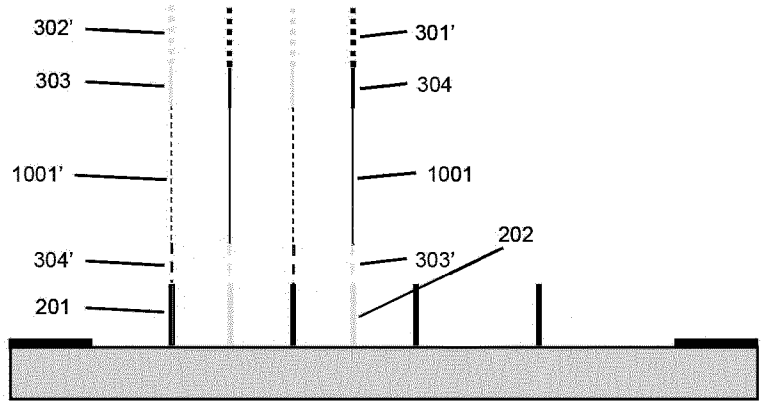


Figure 9

(A)


302'
303
1001'
304'
201
301'
304
1001
303'
202

(B)


302'
303
1001'
304'
201
301'
304
1001
303'
202

(C)

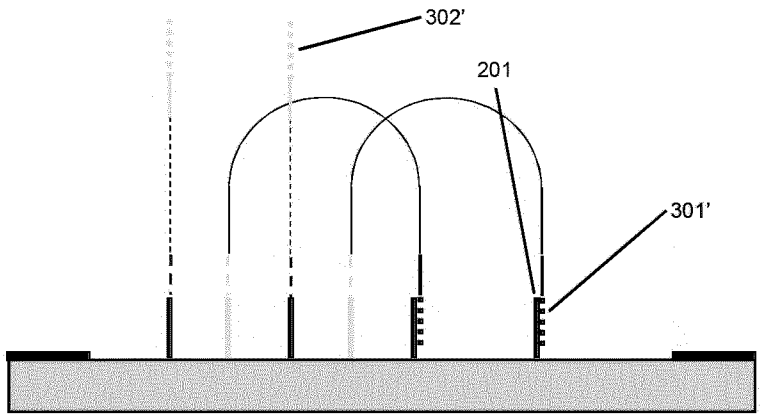
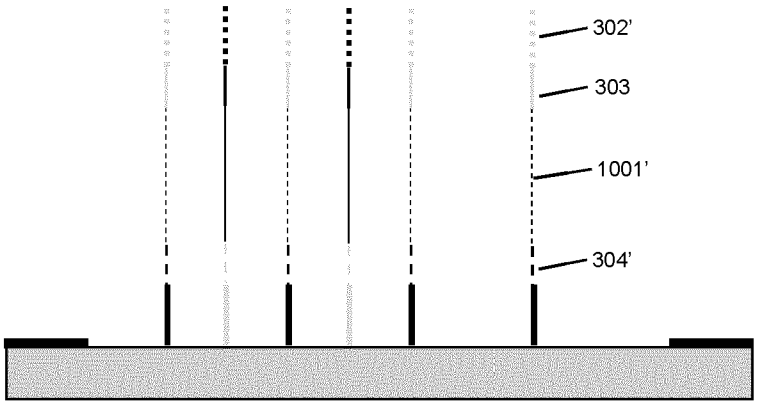

302'
201
301'

Figure 9 (cont.)

(D)



(E)

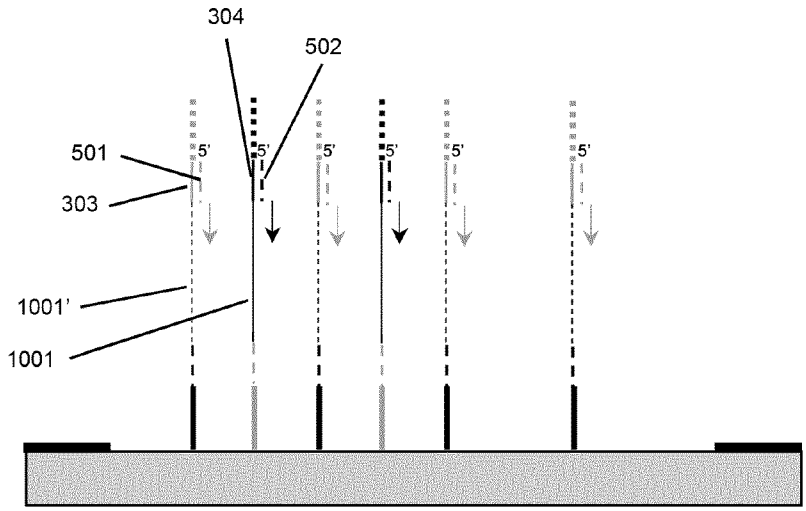
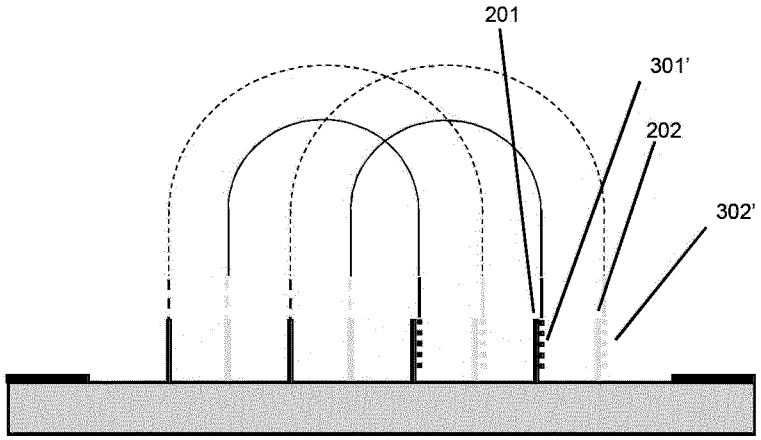
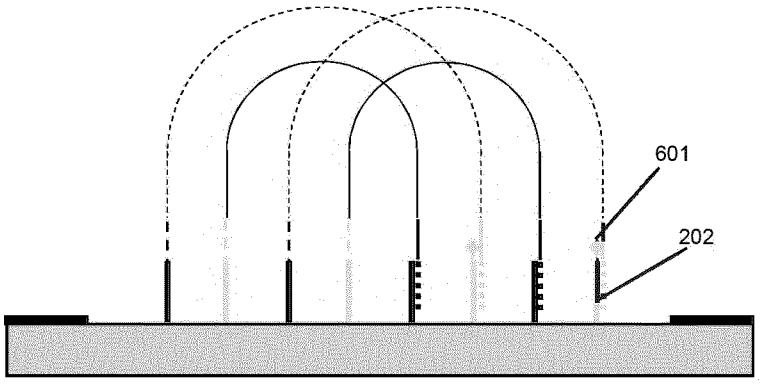


Figure 10

(A)



(B)



(C)

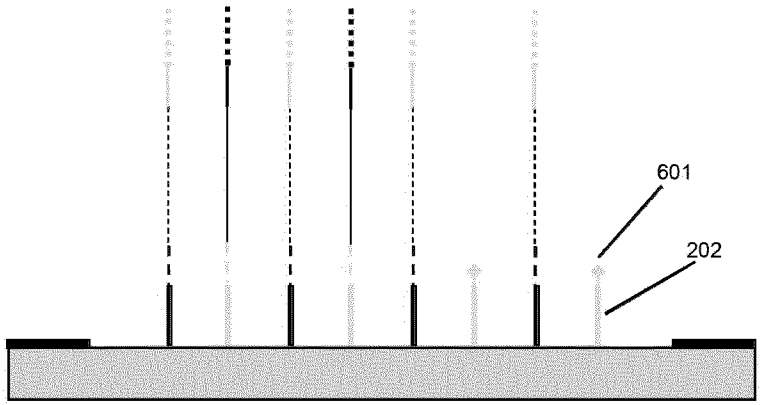


Figure 10 (cont.)

(D)

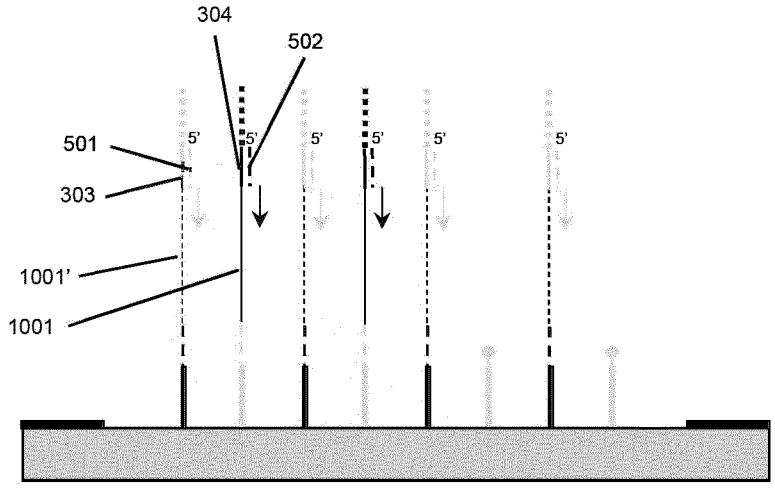
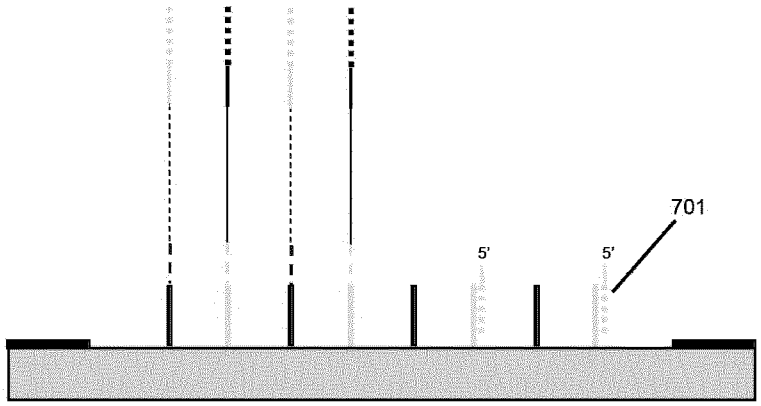


Figure 11

(A)



(B)

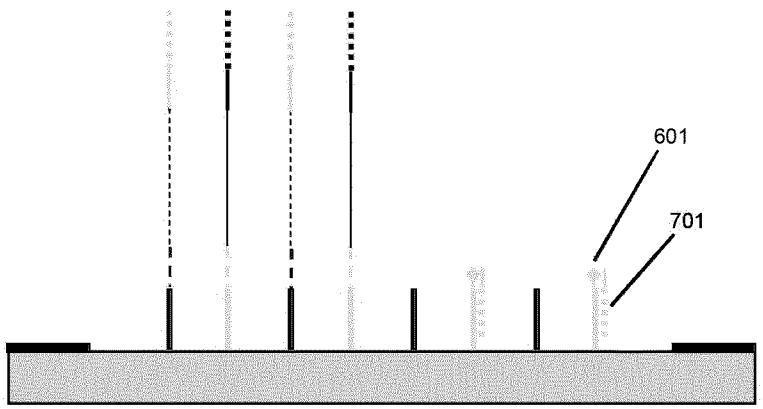


Figure 12

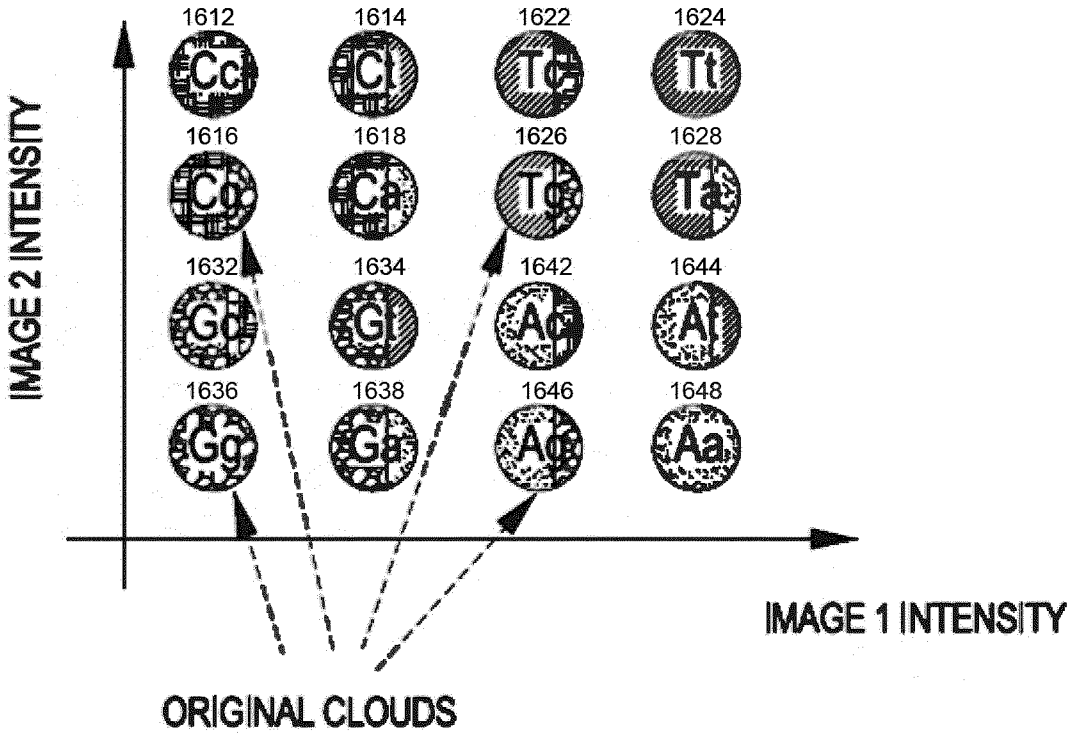

1612
1614
1622
1624
1616
1618
1626
1628
1632
1634
1642
1644
1636
1638
1646
1648
IMAGE 2 INTENSITY
IMAGE 1 INTENSITY
ORIGINAL CLOUDS

Figure 13

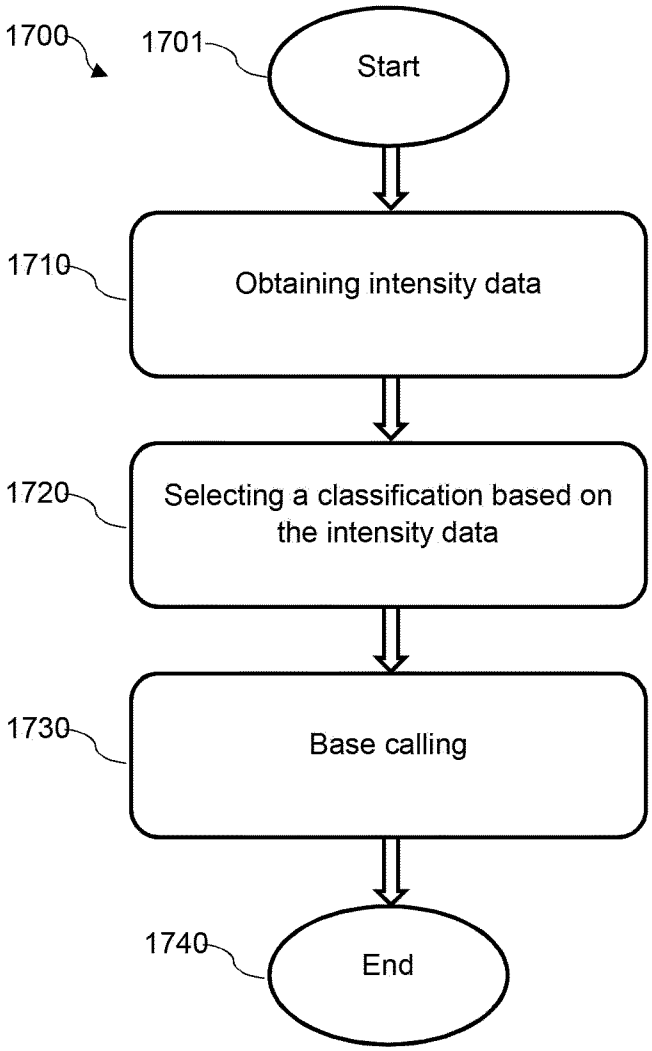

1700
1701
Start
1710
Obtaining intensity data
1720
Selecting a classification based on the intensity data
1730
Base calling
1740
End

Figure 14

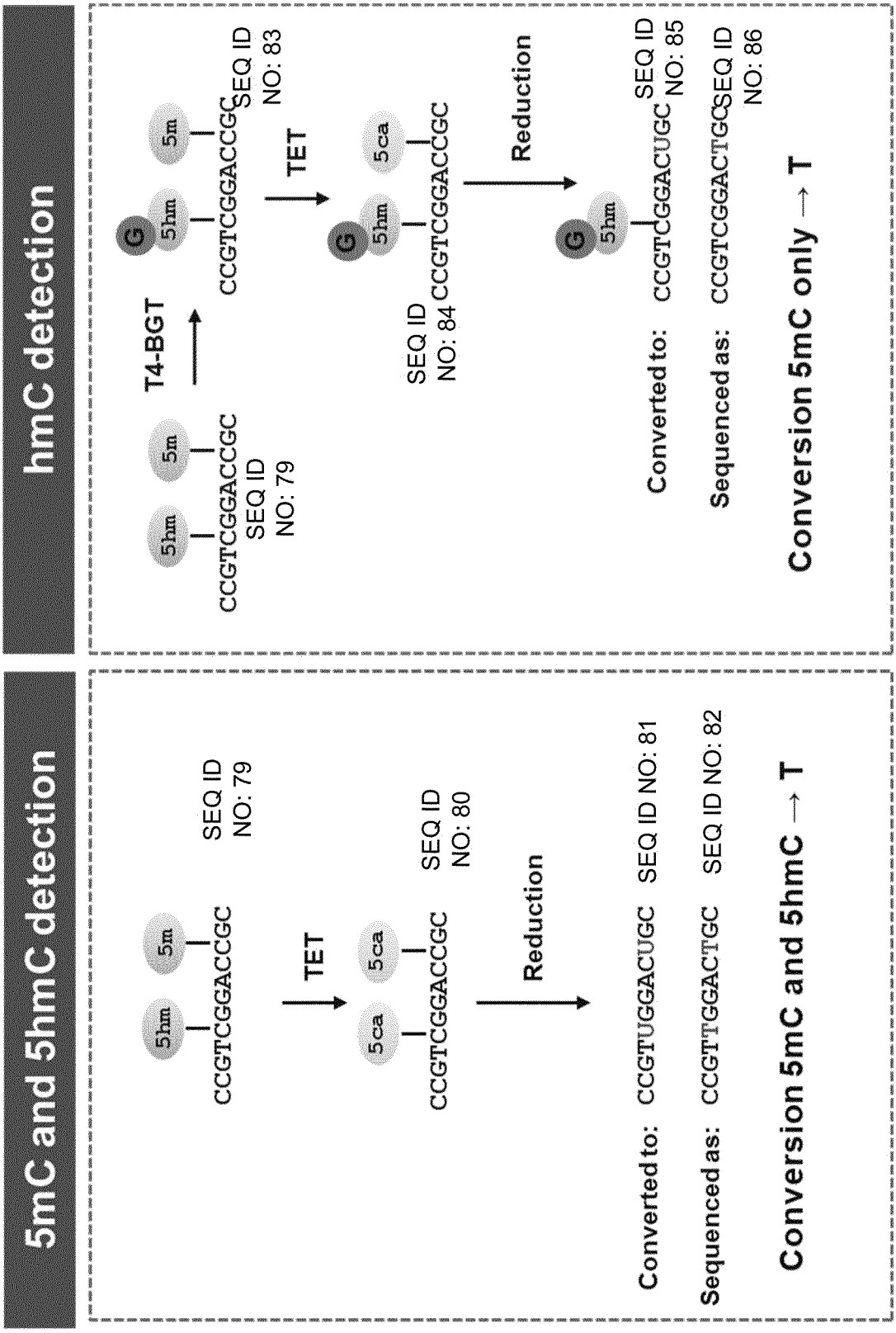

5mC and 5hmC detection
5hm
5m
CCGTCGGACCGC
SEQ ID NO: 79
TET
5ca
5ca
CCGTCGGACCGC
SEQ ID NO: 80
Reduction
Converted to: CCGTUGGACUGC SEQ ID NO: 81
Sequenced as: CCGTTGGACTGC SEQ ID NO: 82
Conversion 5mC and 5hmC → T
hmC detection
5hm
5m
CCGTCGGACCGC
SEQ ID NO: 79
T4-BGT
G
5hm
5m
CCGTCGGACCGC
SEQ ID NO: 83
TET
G
5hm
5ca
SEQ ID NO: 84 CCGTCGGACCGC
Reduction
G
5hm
Converted to: CCGTCGGACUGC SEQ ID NO: 85
Sequenced as: CCGTCGGACTGC SEQ ID NO: 86
Conversion 5mC only → T

Figure 15

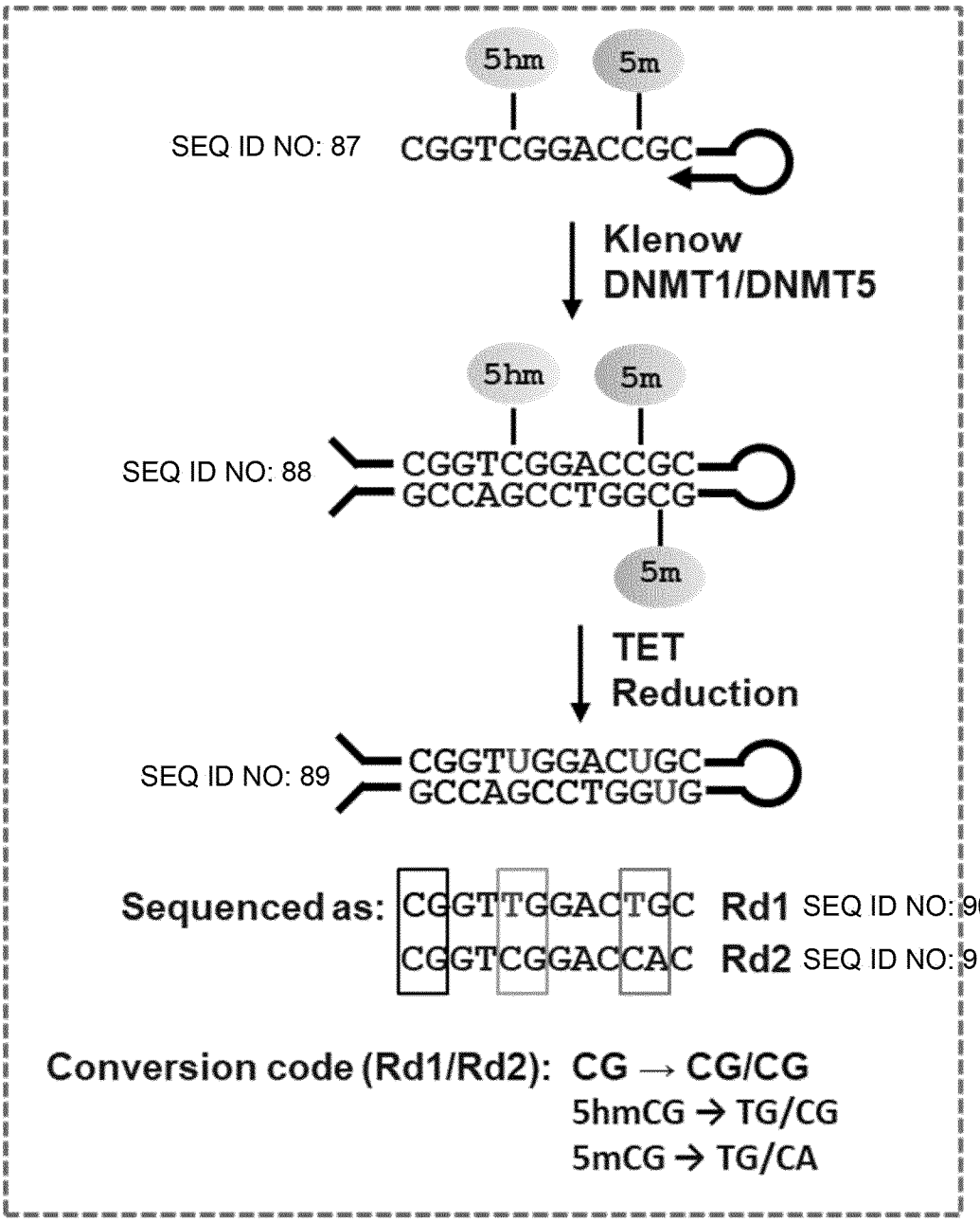

5hm
5m
SEQ ID NO: 87
CGGTCGGACCGC
Klenow
DNMT1/DNMT5
5hm
5m
SEQ ID NO: 88
CGGTCGGACCGC
GCCAGCCTGGCG
5m
TET
Reduction
SEQ ID NO: 89
CGGTUGGACUGC
GCCAGCCTGGUG
Sequenced as:
CGGTTGGACTGC Rd1 SEQ ID NO: 90
CGGTCGGACCAC Rd2 SEQ ID NO: 91
Conversion code (Rd1/Rd2): CG → CG/CG
5hmCG → TG/CG
5mCG → TG/CA

Figure 16

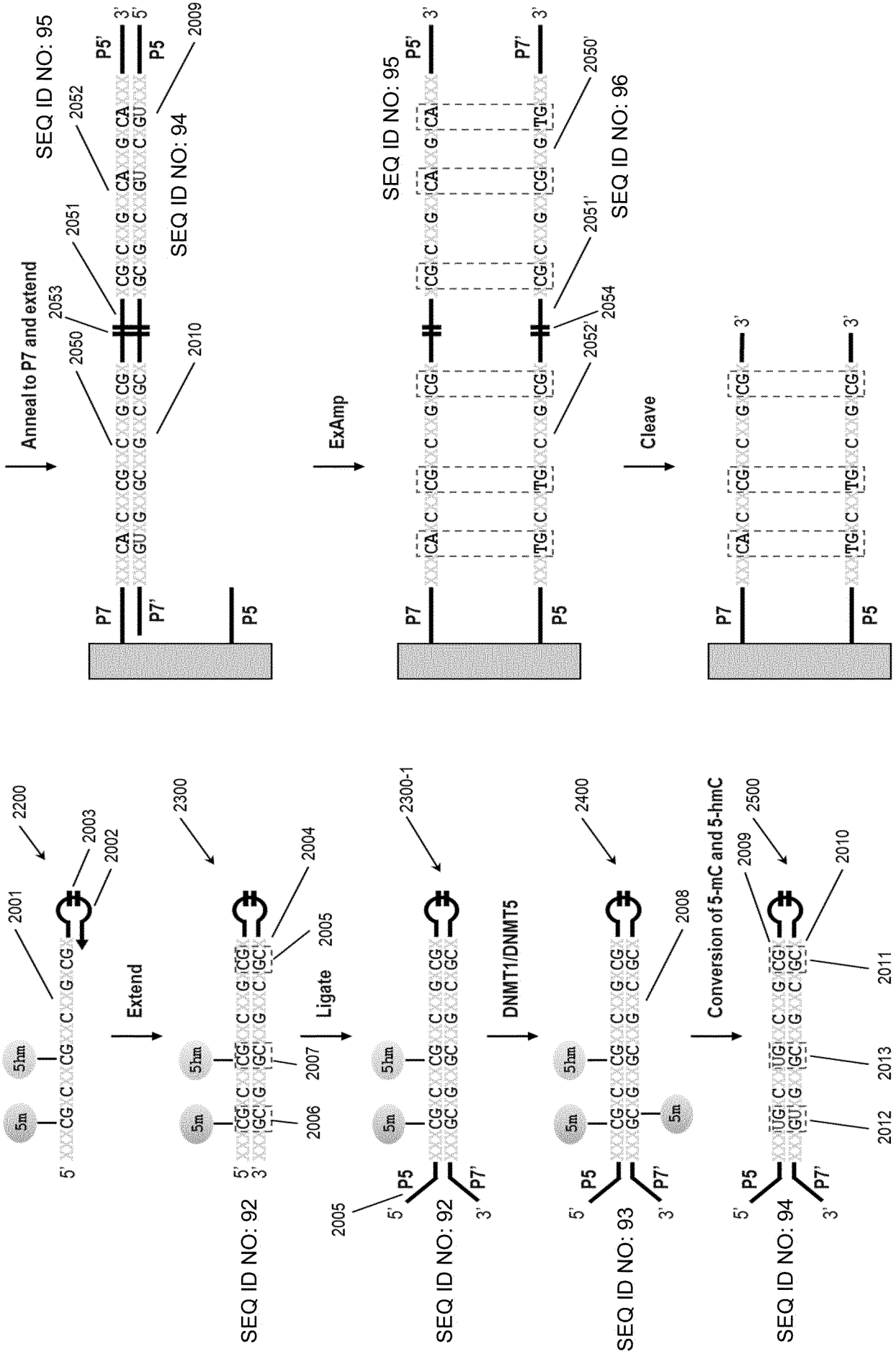

SEQ ID NO: 95
Anneal to P7 and extend
SEQ ID NO: 94
ExAmp
SEQ ID NO: 95
SEQ ID NO: 96
Cleave
Extend
Ligate
DNMT1/DNMT5
Conversion of 5-mC and 5-hmC
SEQ ID NO: 92
SEQ ID NO: 92
SEQ ID NO: 93
SEQ ID NO: 94
P5
P7
P7'
P5'
5hm
5m

Figure 17

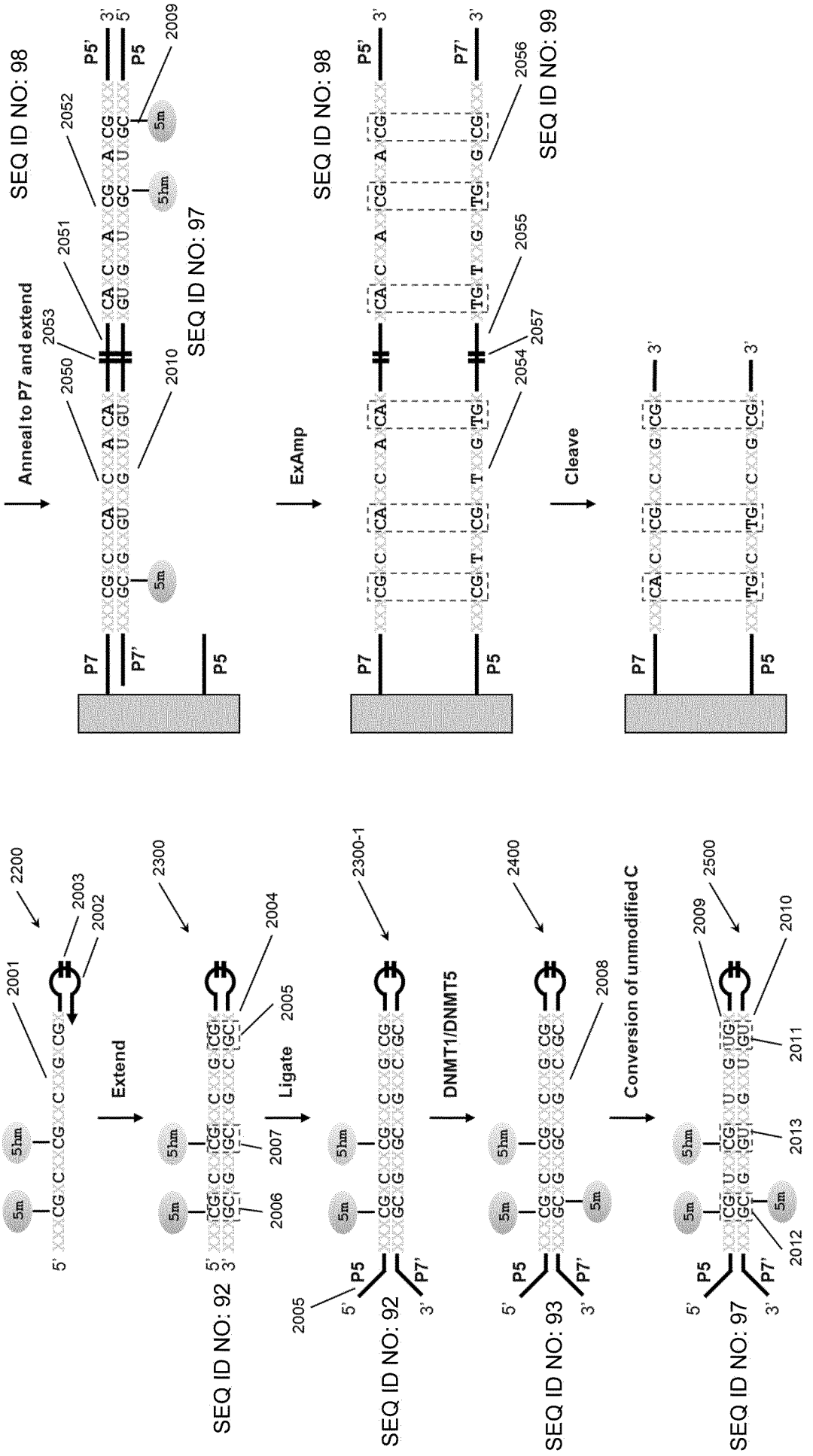
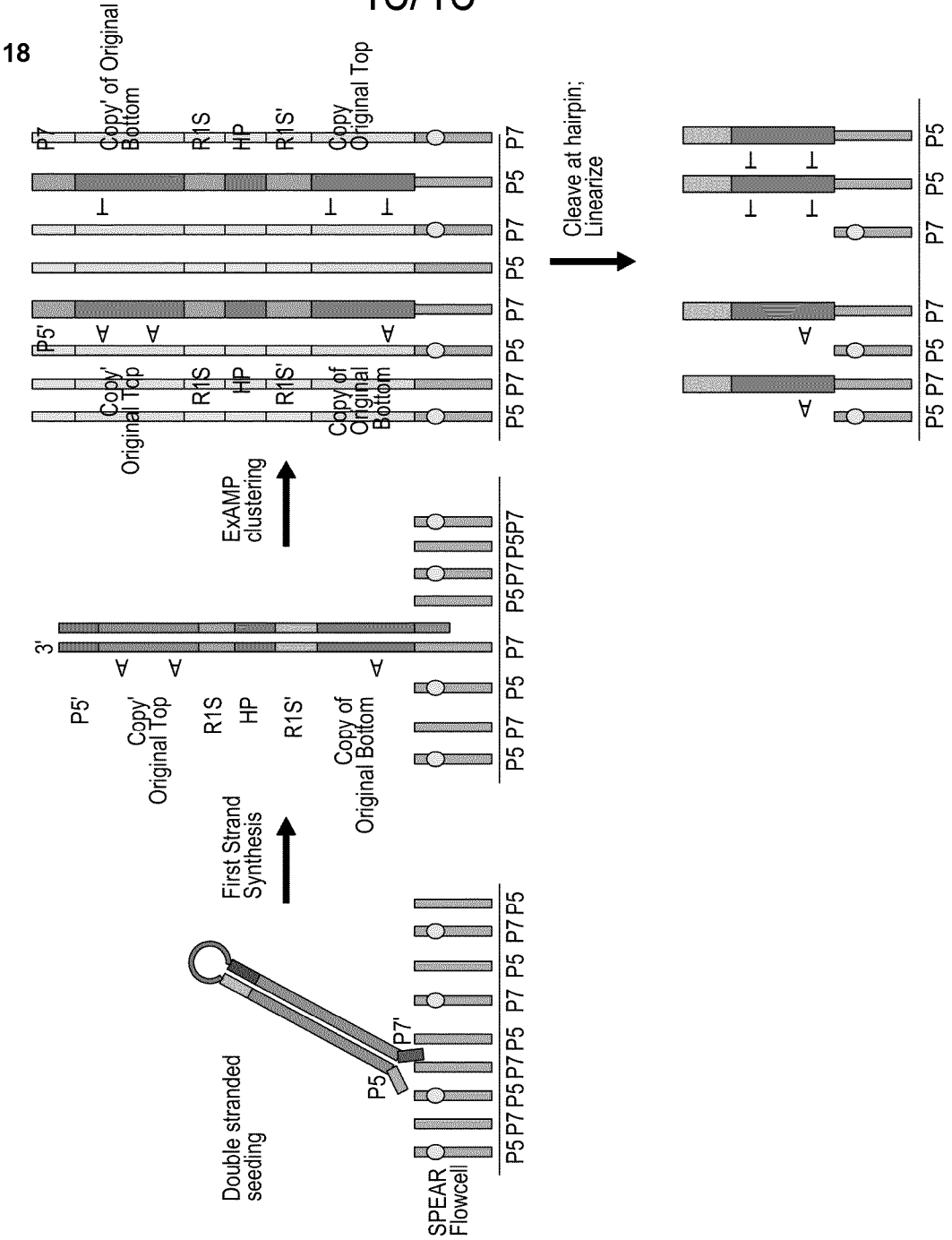
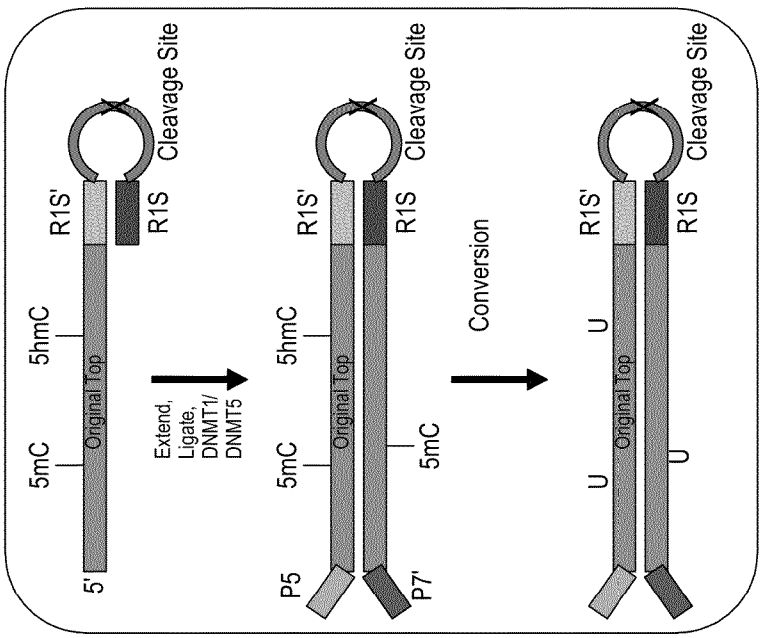

SEQ ID NO: 98
Anneal to P7 and extend
SEQ ID NO: 97
ExAmp
SEQ ID NO: 98
SEQ ID NO: 99
Cleave
Extend
SEQ ID NO: 92
Ligate
SEQ ID NO: 92
DNMT1/DNMT5
SEQ ID NO: 93
Conversion of unmodified C
SEQ ID NO: 97
P5
P7
P5'
P7'
5hm
5m

Figure 18


SPEAR Flowcell
Double stranded seeding
First Strand Synthesis
ExAMP clustering
Cleave at hairpin; Linearize
Copy' of Original Bottom
Copy Original Top
Copy' Original Top
Copy of Original Bottom
R1S
HP
R1S'
P5
P7
P5'
3'


Cleavage Site
R1S'
R1S
5hmC
5mC
Original Top
5'
Extend, Ligate, DNMT1/ DNMT5
Conversion
U
P5
P7'

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2024/066447

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/6806
ADD.

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

B. FIELDS SEARCHED

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

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	US 2022/119878 A1 (GLEZER ELI N [US] ET AL) 21 April 2022 (2022-04-21) paragraphs [0013] - [0035], [0050], [0114] - [0128], [0161] - [0168] claims 80,85-103 figures 1-19	1,2,4, 30-50 3,5-29
X Y	WO 2023/034814 A1 (SINGULAR GENOMICS SYSTEMS INC [US]) 9 March 2023 (2023-03-09) paragraphs [0003] - [0006], [0032], [0105], [0125], [0200] - [0211], [0273] - [0291]	1,2,4, 30-50 3,5-29

-/--

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

[x] See patent family annex.

* Special categories of cited documents :

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 September 2024	08/10/2024

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	Bruma, Anja

1

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

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2024/066447

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	KYRIAKOPOULOS CHARALAMPOS ET AL: "A comprehensive approach for genome-wide efficiency profiling of DNA modifying enzymes", CELL REPORTS METHODS, vol. 2, no. 3, 1 March 2022 (2022-03-01), page 100187, XP093207818, ISSN: 2667-2375, DOI: 10.1016/j.crmeth.2022.100187	31-35
Y	the whole document	1-30, 36-50

X	US 2015/011396 A1 (SCHROEDER BENJAMIN G [US] ET AL) 8 January 2015 (2015-01-08)	30-36,50
Y	paragraphs [6259] - [0062], [0071] - [0086], [0094] - [0096] claims 1-47	1-29, 37-49

X	US 2022/220543 A1 (SALK JESSE J [US]) 14 July 2022 (2022-07-14)	30,36,50
Y	claims 1-54 paragraphs [0015] - [0016], [0046], [0052], [0061] - [0076], [0080] - [0100], [0145] - [0155], [0159] - [0232]	1-29, 31-35, 37-49

Y	GIEHR PASCAL ET AL: "Two are better than one: HPoxBS - hairpin oxidative bisulfite sequencing", NUCLEIC ACIDS RESEARCH, vol. 46, no. 15, 15 June 2018 (2018-06-15) , pages e88-e88, XP093025520, GB ISSN: 0305-1048, DOI: 10.1093/nar/gky422 the whole document figures 1-4	1-50

1

INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2024/066447

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:

   a. [X] forming part of the international application as filed.

   b. [ ] furnished subsequent to the international filing date for the purposes of international search (Rule 13*ter*.1(a)).

      [ ] accompanied by a statement to the effect that the sequence listing does not go beyond the disclosure in the international application as filed.

2. [ ] With regard to any nucleotide and/or amino acid sequence disclosed in the international application, this report has been established to the extent that a meaningful search could be carried out without a WIPO Standard ST.26 compliant sequence listing.

3. Additional comments:

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2024/066447

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2022119878 A1	21-04-2022	EP 4114978 A2	11-01-2023
		US 2021277461 A1	09-09-2021
		US 2022033895 A1	03-02-2022
		US 2022119878 A1	21-04-2022
		US 2022282324 A1	08-09-2022
		US 2022325341 A1	13-10-2022
		US 2023203576 A1	29-06-2023
		US 2024150826 A1	09-05-2024
		WO 2021178893 A2	10-09-2021
WO 2023034814 A1	09-03-2023	EP 4396339 A1	10-07-2024
		US 2024271208 A1	15-08-2024
		WO 2023034814 A1	09-03-2023
US 2015011396 A1	08-01-2015	US 2015011396 A1	08-01-2015
		US 2016265042 A1	15-09-2016
		US 2021285040 A1	16-09-2021
US 2022220543 A1	14-07-2022	AU 2020321991 A1	03-03-2022
		CA 3146435 A1	04-02-2021
		CN 114502742 A	13-05-2022
		EP 4007818 A1	08-06-2022
		IL 290274 A	01-04-2022
		JP 2022543778 A	14-10-2022
		US 2022220543 A1	14-07-2022
		WO 2021022237 A1	04-02-2021