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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 September 2023 (21.09.2023)

WIPO | PCT

(10) International Publication Number
WO 2023/175041 A1

(51) International Patent Classification:
C12Q 1/6869 (2018.01) *C12Q 1/6874* (2018.01)

(21) International Application Number:
PCT/EP2023/056669

(22) International Filing Date:
15 March 2023 (15.03.2023)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

63/269,383	15 March 2022 (15.03.2022)	US
63/439,415	17 January 2023 (17.01.2023)	US
63/439,417	17 January 2023 (17.01.2023)	US
63/439,438	17 January 2023 (17.01.2023)	US
63/439,443	17 January 2023 (17.01.2023)	US
63/439,466	17 January 2023 (17.01.2023)	US
63/439,491	17 January 2023 (17.01.2023)	US
63/439,501	17 January 2023 (17.01.2023)	US
63/439,519	17 January 2023 (17.01.2023)	US
63/439,522	17 January 2023 (17.01.2023)	US

(71) Applicants: **ILLUMINA, INC.** [US/US]; 5200 Illumina Way, San Diego, California 92122 (US). **ILLUMINA CAMBRIDGE LIMITED** [GB/GB]; Illumina Centre, 19 Granta Park, Great Abington, Cambridge Cambridgeshire CB21 6DF (GB). **ILLUMINA SOFTWARE, INC.** [US/US]; 5200 Illumina Way, San Diego, California 92122 (US).

(72) Inventors: **BOUTELL, Jonathan**; c/o Illumina Cambridge Limited, Illumina Centre, 19 Granta Park, Great Abington, Cambridge Cambridgeshire CB21 6DF (GB). **GORMLEY, Niall**; c/o Illumina Cambridge Limited, Illumina Centre, 19 Granta Park, Great Abington, Cambridge Cambridgeshire CB21 6DF (GB). **VESSERE, Gery**; c/o Illumina Software, Inc., 5200 Illumina Way, San Diego, California 92122 (US). **KARUNAKARAN, Aathavan**; c/o Illumina, Inc., 5200 Illumina Way, San Diego, California 92122 (US). **SRIDHARAN, Shagesh**; c/o Illumina, Inc., 5200 Illumina Way, San Diego, California 92122 (US). **SARAF, Nileshi**; 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,

(54) Title: CONCURRENT SEQUENCING OF FORWARD AND REVERSE COMPLEMENT STRANDS ON CONCATENATED POLYNUCLEOTIDES

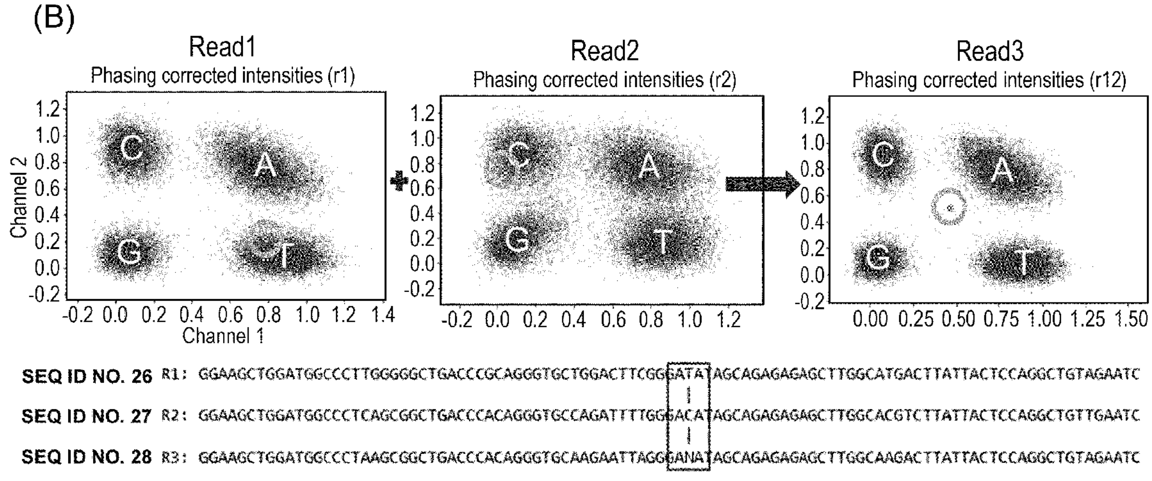


FIG. 15

(57) Abstract: The invention relates to methods of detecting mismatched base pairs in nucleic acid sequences.

[Continued on next page]

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).

Published:

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

— *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*

Concurrent sequencing of forward and reverse complement strands on concatenated polynucleotides

Field of the Invention

The invention relates to methods of detecting mismatched base pairs in nucleic acid sequences.

Background of the Invention

The common expectation is that the complementary sequences of a double-stranded DNA molecule should carry exactly similar information, and as such, sequencing one strand of the molecule should be sufficient. In practice, however, this notion is not accurate.

The most common occasion where the symmetry of information between complementary strands may break is due to DNA damage. Different bases of DNA have different susceptibilities to different forms of damage. For instance, G is very sensitive to oxidative damage leading to the formation of oxo-G, the formation of which is one of the main reasons of library prep dependent sequencing errors, as DNA polymerases often unfaithfully pair oxo-G with A, leading to high quality C>A sequencing errors. This results in the creation of mismatched base pairs.

At the same time, there remains a need to develop more accurate nucleic acid sequencing methods. Identifying such mismatched base pairs would allow such sequencing errors to be identified.

Summary of the Invention

According to an aspect of the present invention, there is provided a method of preparing at least one polynucleotide sequence for detection of mismatched base pairs, comprising:

synthesising at least one polynucleotide sequence comprising a first portion and a second portion,

wherein the at least one polynucleotide sequence comprises portions of a double-stranded nucleic acid template, and the first portion comprises a forward strand of the template, and the second portion comprises a reverse complement strand of the template; or wherein the first portion comprises a reverse strand of the template, and the second portion comprises a forward complement strand of the template.

In one embodiment, the forward strand of the template is not identical to the reverse complement strand of the template.

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

In one embodiment, 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 embodiment, a proportion of first portions is capable of generating a first signal and a proportion of second portions is 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 one embodiment, the method further 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.

In one embodiment, 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 embodiment, 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, or between 1.5:1 to 3:1, or about 2:1.

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

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

In one embodiment, 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 embodiment, the blocked second primer comprises a sequence as defined in SEQ ID NO. 11 to 16 or a variant or fragment thereof and/or the unblocked second primer comprises a sequence as defined in SEQ ID NO. 11 to 14 or a variant or fragment thereof.

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

In one embodiment, the at least one polynucleotide sequence comprising the first portion and the second portion is/are attached to a solid support, wherein the solid support may be a flow cell.

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

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

In one embodiment, the at least one polynucleotide sequence comprising the first portion and the second portion forms a monoclonal cluster.

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

In one embodiment, 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 embodiment, each polynucleotide sequence comprising the first portion and the second portion is attached to a first immobilised primer.

In one embodiment, each polynucleotide sequence comprising the first portion and the second portion further comprises a second adaptor sequence, wherein the second adaptor sequence is substantially complementary to the second immobilised primer.

In one embodiment, the step of synthesising the at least one polynucleotide sequence comprising a first portion and a second portion comprises:

synthesising a first precursor polynucleotide fragment comprising a complement of the first portion and a hybridisation complement sequence,

synthesising a second precursor polynucleotide fragment comprising a second portion and a hybridisation sequence,

annealing the hybridisation complement sequence of the first precursor polynucleotide fragment with the hybridisation sequence on the second precursor polynucleotide fragment to form a hybridised adduct,

synthesising a first precursor polynucleotide sequence by extending the first precursor polynucleotide fragment to form a complement of the second portion, and

synthesising the at least one polynucleotide sequence by forming a complement of the first precursor polynucleotide sequence.

In one embodiment, the first precursor polynucleotide fragment comprises a first sequencing primer binding site complement.

In one embodiment, the first sequencing primer binding site complement is located before a 5’-end of the complement of the first portion, such as immediately before the 5’-end of the complement of the first portion.

In one embodiment, the first precursor polynucleotide fragment comprises a second adaptor complement sequence.

In one embodiment, the second adaptor complement sequence is located before a 5’-end of the complement of the first portion.

In one embodiment, the first precursor polynucleotide fragment comprises a first sequencing primer binding site complement and a second adaptor complement sequence.

In one embodiment, the first sequencing primer binding site complement is located before a 5’-end of the complement of the first portion, and wherein the second adaptor complement sequence is located before a 5’-end of the first sequencing primer binding site complement.

In one embodiment, the first precursor polynucleotide fragment comprises a second sequencing primer binding site complement.

In one embodiment, the hybridisation sequence complement comprises the second sequencing primer binding site complement.

In one embodiment, the second precursor polynucleotide fragment comprises a first adaptor complement sequence.

In one embodiment, the method further comprises concurrently sequencing nucleobases in the first portion and the second portion.

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

According to another aspect of the present invention, there is provided a method of sequencing at least one polynucleotide sequence to detect mismatched base pairs, comprising:

preparing at least one polynucleotide sequence for detection of mismatched base pairs using a method as described herein;

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

identifying mismatched base pairs 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 concurrently sequencing nucleobases comprises performing sequencing-by-synthesis or sequencing-by-ligation.

In one embodiment, the step of preparing the at least one polynucleotide sequence comprises using a method as described herein; and wherein the step of concurrent sequencing nucleobases in the first portion and the second portion is based on the intensity of the first signal and the intensity of the second signal.

In one embodiment, the mismatched base pair comprises an oxo-G to A base pair.

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

In one embodiment, 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 embodiment, 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 embodiment, the plurality of classifications comprises 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 are generated based on light emissions associated with the respective nucleobase.

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

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

In one embodiment, 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 of the plurality of classifications represents one or more possible combinations of respective first and second nucleobases, and wherein at least one classification of the plurality of classifications represents more than one possible combination of respective first and second nucleobases; and

(d) based on the selected classification, determining sequence information from the first portion and the second portion.

In one embodiment, 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 embodiment, when based on a nucleobase of the same identity, an intensity of the first signal component is substantially the same as an intensity of the second signal component and an intensity of the third signal component is substantially the same as an intensity of the fourth signal component.

In one embodiment, the plurality of classifications consists of a predetermined number of classifications.

In one embodiment, the plurality of classifications comprises:

one or more classifications representing matching first and second nucleobases; and

one or more classifications representing mismatching first and second nucleobases, and

wherein determining sequence information of the first portion and second portion comprises:

in response to selecting a classification representing matching first and second nucleobases, determining a match between the first and second nucleobases; or

in response to selecting a classification representing mismatching first and second nucleobases, determining a mismatch between the first and second nucleobases.

In one embodiment, determining sequence information of the first portion and the second portion comprises, in response to selecting a classification representing a match between the first and second nucleobases, base calling the first and second nucleobases.

In one embodiment, determining sequence information of the first portion and the second portion comprises, based on the selected classification, determining that the second portion is modified relative to the first portion at a location associated with the first and second nucleobases.

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

In one embodiment, 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 at least one polynucleotide sequence for detection of mismatched base pairs as described herein, and/or for sequencing at least one polynucleotide sequence to detect mismatched base pairs 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 embodiment, 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.

Description of the Drawings

Features of examples of the present disclosure will become apparent by reference to the following detailed description and drawings, in which like reference numerals correspond to similar, though perhaps not identical, components. For the sake of brevity, reference numerals or features having a previously described function may or may not be described in connection with other drawings in which they appear.

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

Figure 2 shows the preparation of a concatenated polynucleotide sequence comprising a first portion and a second portion using a tandem insert method, comprising (A) preparation of a desired first (forked) adaptor and second (forked) adaptor from three oligos; (B) different types of first (forked) adaptors and second (forked) adaptors that do not anneal to each other due to the presence of a third oligo on at least one of the first (forked) adaptor and/or the second (forked) adaptor; (C) ligation of the template polynucleotide strand and adaptors generates three products, with the desired product containing both types of adaptor being produced at a proportion of 50%; (D) synthesis of concatenated strands from the desired product; and (E) completion of the synthesis of the concatenated strands from the desired product.

Figure 3 shows an example of a concatenated polynucleotide sequence comprising a first portion and a second portion, as well as terminal and internal adaptor sequences.

Figure 4 shows an example of a concatenated polynucleotide sequence comprising a first portion and a second portion, as well as terminal and internal adaptor sequences.

Figure 5 shows a typical solid support.

Figure 6 shows the stages of bridge amplification for concatenated polynucleotide sequences and the generation of an amplified cluster, comprising (A) a concatenated library strand hybridising to a immobilised primer; (B) generation of a template strand from the library strand; (C) dehybridisation and washing away the library strand; (D) generation of a template complement strand from the template strand via bridge amplification and dehybridisation of the sequence bridge; (E) further amplification to provide a plurality of template and template complement strands; and (F) cleavage of one set of the 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 is a plot showing graphical representations of sixteen distributions of signals generated by polynucleotide sequences according to one embodiment.

Figure 10 is a flow diagram showing a method for base calling according to one embodiment.

Figure 11 is a plot showing graphical representations of nine distributions of signals generated by polynucleotide sequences according to one embodiment.

Figure 12 is a plot showing graphical representations of nine distributions of signals generated by polynucleotide sequences according to one embodiment, highlighting distributions that may be associated with library preparation errors.

Figure 13 is a flow diagram showing a method for determining sequence information according to one embodiment.

Figure 14A shows 9 QaM analysis conducted on the signals obtained from the custom second hyb run of Example 1. The x-axis shows signal intensity from a “red” wavelength channel, whilst the y-axis shows signal intensity from a “green” wavelength channel. G is not associated with any dyes and as such appears contributes no intensity for both “red” and “green” channels. C is associated with a “red” dye and as such contributes intensity to the “red” channel, but not the “green” channel. T is associated with a “green” dye and as such contributes intensity to the “green” channel, but not the “red channel. A is associated with both a “red” dye and a “green” dye, and as such contributes intensity to both the “red” channel and “green” channel. Since the template comprises forward and reverse complement strands that are sequenced simultaneously, most of the readout will generate (G,G) read (bottom left corner), (C,C) read (bottom right corner), (T,T) read (top left corner), and (A,A) read (top right corner) clouds. However, any mismatched base pairs will appear in regions other than the four corner clouds. A central cloud corresponding to (C,T) or (T,C) reads corresponds with the presence of modified cytosines; in addition, side clouds located at the top middle, bottom middle, centre left and centre right sections corresponds with the presence of other mismatched base pairs. Figure 14B shows sequence data generated from two different primers used (HYB2’-ME and HP10) in the custom second hyb run of Example 1. Mismatches between the two sequences allow identification of modified cytosines. For example, 5-mC present in the original forward strand of the target polynucleotide is read as T in the HP10 read, whereas C present in the original reverse complement strand of the target polynucleotide (corresponding to the same position as 5-mC in the original forward strand of the target polynucleotide) is read as C in the HYB2’-ME read.

Figure 15A shows the sequencing primer binding modes used in Example 2 – Read 1 (control) is conducted using only a single sequencing primer type (HP21 mix), Read 2 (control) is conducted using a single sequencing primer type (HYB2’-ME), and Read 3 is conducted using two sequencing primer types (HP10 mix and HYB2’-ME) to enable concurrent sequencing to generate a 9 QaM signal. Figure 15B shows the results from the Read 1, Read 2 and Read 3 runs in Example 2. The plot is arranged so that G is disposed on the bottom left corner, C is disposed on the top left corner, T is disposed on the bottom right corner, and A is disposed on the top right corner. The Read 1 plot has a T base call for one of the reads (highlighted as a circled point). The Read 2 plot has a C base call for the read corresponding to the same position (highlighted as a circled point). The Read 3 plot contains (G,G) reads at the bottom left corner, (C,C) reads at the top left corner, (T,T) reads at the bottom right corner, and (A,A) reads at the top right corner. An mismatched base pair error was detected due to the presence of a (C,T) read in the central middle portion of the plot.

Detailed Description of the Invention

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: https://www.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

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. See Figure 1.

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.

Generally, where forward strand, reverse strand, forward complement strand, and reverse complement strand are used herein without qualifying whether they are with respect to the original polynucleotide sequence 100 or with respect to the “template”, these terms may be interpreted as referring to the “template”.

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 a concatenated polynucleotide sequence comprising a first portion and 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 by using a tandem insert method described in more detail in e.g. WO 2022/087150, which is incorporated herein by reference. This procedure may be used, for example, for preparing templates comprising concatenated polynucleotide sequences comprising a first portion and 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). Such libraries may also be referred to as cross-tandem inserts. A representative process for conducting a tandem insert method is shown in Figure 2A to 2E.

The processes described above in relation to tandem insert methods generate libraries that have concatenated polynucleotides.

Thus, one strand of a concatenated 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 terminal sequencing primer binding site complement 303’ (e.g. B15-ME; or if ME is not present, then B15), a first insert sequence 401, a hybridisation complement sequence 403 (e.g. ME’-HYB2-ME; or if ME’ and ME are not present, then HYB2), a second insert sequence 402, a second terminal sequencing primer binding site 304 (e.g. ME'-A14'; or if ME' is not present, then A14'), and a first primer-binding sequence 301' (e.g. P5') (Figures 3 and 4 – bottom strand).

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 first terminal sequencing primer binding site complement 303' (e.g. B15-ME; or if ME is not present, then B15). Separately, or in addition, a second index complement sequence (e.g. i5') may be provided between the second terminal sequencing primer binding site 304 (e.g. ME'-A14') 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), a first terminal sequencing primer binding site complement 303' (e.g. B15-ME; or if ME is not present, then B15), a first insert sequence 401, a hybridisation complement sequence 403 (e.g. ME'-HYB2-ME; or if ME' and ME are not present, then HYB2), a second insert sequence 402, a second terminal sequencing primer binding site 304 (e.g. ME'-A14'; or if ME' is not present, then A14'), a second index complement sequence (e.g. i5'), and a first primer-binding sequence 301' (e.g. P5')

Another strand of a concatenated polynucleotide within a polynucleotide library may comprise, in a 5' to 3' direction, a first primer-binding complement sequence 301 (e.g. P5), a second terminal sequencing primer binding site complement 304' (e.g. A14-ME; or if ME is not present, then A14), a second insert complement sequence 402', a hybridisation sequence 403' (e.g. ME'-HYB2'-ME; or if ME' and ME are not present, then HYB2'), a first insert complement sequence 401', a first terminal sequencing primer binding site 303 (e.g. ME'-B15'; or if ME' is not present, then B15'), and a second primer-binding sequence 302' (e.g. P7') (Figures 3 and 4 – top strand).

Although not shown in Figures 3 and 4, the another strand may further comprise one or more index sequences. As such, a second index sequence (e.g. i5) may be provided between the first primer-binding complement sequence 301 (e.g. P5) and the second terminal sequencing primer binding site complement 304' (e.g. A14-ME; or if ME is not present, then A14). Separately, or in addition, a first index complement sequence (e.g. i7') may be provided between the first terminal sequencing primer binding site 303 (e.g.

ME’-B15’; or if ME’ is not present, then B15’) and the second primer-binding sequence 302’ (e.g. P7’). Thus, in some embodiments, another strand of a polynucleotide within a polynucleotide library may comprise, in a 5’ to 3’ direction, a first primer-binding complement sequence 301 (e.g. P5), a second index sequence (e.g. i5), a second terminal sequencing primer binding site complement 304’ (e.g. A14-ME; or if ME is not present, then A14).), a second insert complement sequence 402’, a hybridisation sequence 403’ (e.g. ME’-HYB2’-ME; or if ME’ and ME are not present, then HYB2’), a first insert complement sequence 401’, a first terminal sequencing primer binding site 303 (e.g. ME’-B15’; or if ME’ is not present, then B15’), a first index complement sequence (e.g. i7’), and a second primer-binding sequence 302’ (e.g. P7’).

As described herein, the first insert sequence 401 and the second insert sequence 402 may comprise different types of library sequences.

In one embodiment, the first insert sequence 401 may comprise a forward strand of the sequence 101, and the second insert sequence 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 sequence 402 may comprise a forward complement strand of the sequence 101’), for example where the library is prepared using a tandem insert method.

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 (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 (or inserts in concatenated strands) 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.

In concatenated strands, the hybridisation sequence (or the hybridisation sequence complement) may comprise an internal sequencing primer binding site. In other words, an internal sequencing primer binding site may form part of the hybridisation sequence. For example, ME’-HYB2 (or ME’-HYB2’) may act as an internal sequencing primer binding site to which a sequencing primer can bind. Alternatively, the hybridisation sequence may be an internal sequencing primer binding site. For example, HYB2 (or HYB2’) may act as an internal sequencing primer binding site to which a sequencing primer can bind. Accordingly, we may refer to the hybridisation site herein as comprising a second sequencing primer binding site, or as a second sequencing primer binding site.

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.

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.

In some cases, to facilitate sequencing, one set of strands (either the original template strands or the complement strands thereof) may be removed from the solid support leaving either the original template strands or the complement strands. Suitable methods for removing such strands are described in more detail in application number WO 07/010251, the contents of which are incorporated herein by reference in their entirety.

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

In cases where single (concatenated) polynucleotide strands are used, each polynucleotide sequence may be attached (via the 5'-end of the (concatenated) polynucleotide sequence) to a first immobilised primer. Each polynucleotide sequence may comprise a second adaptor sequence, wherein the second adaptor 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 (concatenated) polynucleotide sequence.

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

A particular concatenated 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), a first terminal sequencing primer binding site complement 303' (e.g. B15-ME), a first insert sequence 401, a hybridisation complement sequence 403 (e.g. ME'-HYB2-ME), a second insert sequence 402, a second terminal sequencing primer binding site 304 (e.g. ME'-A14'), 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 concatenated 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 concatenated 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, a second terminal sequencing primer binding site complement 304' (e.g. A14-ME; or if ME is not present, then A14), a second insert complement sequence 402' (which represents a type of "second portion"), a hybridisation sequence 403' (which comprises a type of "second sequencing primer binding site") (e.g. ME'-HYB2'-ME; or if ME' and ME are not present, then HYB2'), a first insert complement sequence 401' (which represents a type of "first portion"), a first terminal sequencing primer binding site 303 (which represents a type of "first sequencing primer binding site") (e.g. ME'-B15'; or if ME' is not present, then B15'), 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 concatenated 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 a template 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, a first terminal sequencing primer binding site complement 303’ (e.g. B15-ME; or if ME is not present, then B15), a first insert sequence 401, a hybridisation complement sequence 403 (e.g. ME’-HYB2-ME; or if ME’ and ME are not present, then HYB2), a second insert sequence 402, a second terminal sequencing primer binding site 304 (e.g. ME’-A14’; or if ME’ is not present, then A14’), 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 concatenated polynucleotide sequences comprising a first insert complement sequence 401’ (i.e. “first portions”) and a second insert complement sequence 402’ (i.e. second portions”), as well as a plurality of concatenated 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.

In one aspect, before sequencing, one group of strands (either the group of template polynucleotides, or the group of template complement polynucleotides thereof) is removed from the solid support to form a (monoclonal) cluster, leaving either the templates or the template complements (Figure 6F).

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. 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 sequencing primer) to the first sequencing primer binding site (e.g. first terminal sequencing primer binding site 303 in templates including a concatenated polynucleotide sequence comprising a first portion and a second portion). The second sequencing read may comprise the binding of a second sequencing primer (also known as a read 2 sequencing primer) to the second sequencing primer binding site (e.g. a portion of hybridisation sequence 403' in templates including a concatenated polynucleotide sequence comprising a first portion and a second portion).

This leads to sequencing of the first portion (e.g. first insert complement sequence 401' in templates including a concatenated polynucleotide sequence comprising a first portion and a second portion) and the second portion (e.g. second insert complement sequence 402' in templates including a concatenated polynucleotide sequence comprising a first portion and a second portion).

Alternative methods of sequencing include sequencing by ligation, for example as described in US 6,306,597 or WO 06/084132, the contents of which are incorporated herein by reference.

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.

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 at least one polynucleotide sequence comprising a first portion and 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 polynucleotide sequences each comprising a first portion and 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 polynucleotide sequence comprising a first portion and a second portion (or the plurality of polynucleotide sequences each comprising a first portion and 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 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, or between 1.5:1 to 3:1, or 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 a single (concatenated) polynucleotide strand with a first and second portion, the single (concatenated) polynucleotide strand may comprise a first sequencing primer binding site and 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 single (concatenated) polynucleotide strand comprises a first sequencing primer binding site and 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, or 1:2 to 2:1.

In one 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 embodiment, the first sequencing primer binding site may be selected from ME’-A14’ (as defined in SEQ ID NO. 17 or a variant or fragment thereof), A14’ (as defined in SEQ ID NO. 18 or a variant or fragment thereof), ME’-B15’ (as defined in SEQ ID NO. 19 or a variant or fragment thereof) and B15’ (as defined in SEQ ID NO. 20 or a variant or fragment thereof); and the second sequencing primer binding site may be selected from ME'-HYB2 (as defined in SEQ ID NO. 21 or a variant or fragment thereof), HYB2 (as defined in SEQ ID NO. 11 or a variant or fragment thereof), ME'-HYB2' (as defined in SEQ ID NO. 22 or a variant or fragment thereof) and HYB2' (as defined in SEQ ID NO. 13 or a variant or fragment thereof).

In another embodiment, the first sequencing primer binding site is ME'-B15' (as defined in SEQ ID NO. 19 or a variant or fragment thereof), and the second sequencing primer binding site is ME'-HYB2' (as defined in SEQ ID NO. 22 or a variant or fragment thereof). Alternatively, the first sequencing primer binding site is B15' (as defined in SEQ ID NO. 20 or a variant or fragment thereof), and the second sequencing primer binding site is HYB2' (as defined in SEQ ID NO. 13 or a variant or fragment thereof). The first and second sequencing primer sites may be located after (e.g. immediately after) a 3'-end of the first and second portions to be identified.

In another embodiment, the first sequencing primer binding site is ME'-A14' (as defined in SEQ ID NO. 17 or a variant or fragment thereof), and the second sequencing primer binding site is ME'-HYB2 (as defined in SEQ ID NO. 21 or a variant or fragment thereof). Alternatively, the first sequencing primer binding site may be A14' (as defined in SEQ ID NO. 18 or a variant or fragment thereof) and the second sequencing primer binding site may be HYB2 (as defined in SEQ ID NO. 11 or a variant or fragment thereof). The first and second sequencing primer sites may be located after (e.g. immediately after) a 3'-end of the first and second portions to be identified.

In one example, the sequencing primer (which may be referred to herein as the second sequencing primer) comprises or consists of a sequence as defined in SEQ ID NO. 11 to 16, or a variant or fragment thereof. The sequencing primer may further comprise a 3' blocking group as described above to create a blocked sequencing primer. Alternatively, the primer comprises a 3'-OH group. Such a primer is unblocked and can be elongated with a polymerase.

In one embodiment, 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. This additional sequencing primer may be selected from A14-ME (as defined in SEQ ID NO.

9 or a variant or fragment thereof), A14 (as defined in SEQ ID NO. 7 or a variant or fragment thereof), B15-ME (as defined in SEQ ID NO. 10 or a variant or fragment thereof) and B15 (as defined in SEQ ID NO. 8 or a variant or fragment thereof). In one embodiment, the sequencing composition comprises blocked second sequencing primers, unblocked second sequencing primers and at least one first sequencing primer, wherein the first sequencing primer is A14, or B15, or is both A14 and B15.

As shown in Figure 8, selective sequencing may be conducted on the amplified (monoclonal) cluster shown in Figure 6F. A plurality of first sequencing primers 501 are added. These first sequencing primers 501 (e.g. B15-ME; or if ME is not present, then B15) anneal to the first terminal sequencing primer binding site 303 (which represents a type of “first sequencing primer binding site”) (e.g. ME’-B15’; or if ME’ is not present, then B15’). 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 (e.g. HYB2-ME; or if ME is not present, then HYB2) and second blocked sequencing primers 502b (e.g. blocked HYB2-ME; or if ME is not present, then blocked HYB2) anneal to an internal sequencing primer binding site in the hybridisation sequence 403’ (which represents a type of “second sequencing primer binding site”) (e.g. ME’-HYB2’; or if ME’ is not present, then HYB2’). This then allows the first insert complement sequences 401’ (i.e. “first portions”) to be sequenced and the second insert complement sequences 402’ (i.e. “second portions”) to be sequenced, wherein a greater proportion of first insert complement sequences 401’ are sequenced (grey arrow) compared to a proportion of second insert complement sequences 402’ (black arrow).

Although Figure 8 shows selective sequencing being conducted on a template strand attached to first immobilised primer 201, in some embodiments the (monoclonal) cluster may instead have template strands attached to second immobilised primer 202. In such a case, the first sequencing primers may instead correspond to A14-ME (or if ME is not present, then A14), and the second unblocked sequencing primers may instead correspond to HYB2’-ME (or if ME is not present, then HYB2’) and second blocked sequencing primers may instead correspond to blocked HYB2’-ME (or if ME is not present, then blocked HYB2’).

In yet 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 internal sequencing primer binding site, and the second sequencing binding primers may anneal instead to the terminal sequencing primer binding site.

Figure 8 shows concurrent sequencing of a concatenated strand according to the above method. As shown in Figure 8, a polynucleotide strand with a first portion (insert) and second portion (insert) can be accurately and simultaneously sequenced by a selective sequencing method that uses a mixture of unblocked and blocked sequencing primers as described above.

<u>Data analysis using 16 QaM</u>

Figure 9 is a scatter plot showing an example of sixteen distributions of signals generated by polynucleotide sequences disclosed herein.

The scatter plot of Figure 9 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 9 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 9. 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 10 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 10, 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 9, 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.

Data analysis using 9 QaM

For two portions of polynucleotide sequences (e.g. a first portion and a second portion as described herein), there are sixteen possible combinations of nucleobases at any given position (i.e., an A in the first portion and an A in the second portion, an A in the first portion and a T in the second portion, and so on). When the same nucleobase is present at a given position in both portions, the light emissions associated with each target sequence during the relevant base calling cycle will be characteristic of the same nucleobase. In effect, the two portions behave as a single portion, and the identity of the bases at that position are uniquely callable.

However, when a nucleobase of the first portion is different from a nucleobase at a corresponding position of the second portion, the signals associated with each portion in the relevant base calling cycle will be characteristic of different nucleobases. In one embodiment, the first signal coming from the first portion have substantially the same intensity as the second signal coming from the second portion. The two signals may also be co-localised, and may not be spatially and/or optically resolved. Therefore, when different nucleobases are present at corresponding positions of the two portions, the identity of the nucleobases cannot be uniquely called from the combined signal alone. However, useful sequencing information can still be determined from these signals.

The scatter plot of Figure 11 shows nine distributions (or bins) of intensity values from the combination of two co-localised signals of substantially equal intensity.

The intensity values shown in Figure 11 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 first signal generated from the first portion and the second signal generated from the second portion results in a combined signal. The combined signal may be captured by a first optical channel and a second optical channel. The computer system can map the combined signal generated into one of the nine bins, and thus determine sequence information relating to the added nucleobase at the first portion and the added nucleobase at the second portion.

Bins are selected based upon the combined intensity of the signals originating from each target sequence during the base calling cycle. For example, bin 1803 may be selected following the detection of a high-intensity (or “on/on”) signal in the first channel and a high-intensity signal in the second channel. Bin 1806 may be selected following the detection of a high-intensity signal in the first channel and an intermediate-intensity (“on/off” or “off/on”) signal in the second channel. Bin 1809 may be selected following the detection of a high-intensity signal in the first channel and a low-intensity or zero-intensity (“off/off”) signal in the second channel. Bin 1802 may be selected following the detection of an intermediate-intensity signal in the first channel and a high-intensity signal in the second channel. Bin 1805 may be selected following the detection of an intermediate-intensity signal in the first channel and an intermediate-intensity signal in the second channel. Bin 1808 may be selected following the detection of an intermediate-intensity signal in the first channel and a low-intensity or zero-intensity signal in the second channel. Bin 1801 may be selected following the detection of a low-intensity signal in the first channel and a high-intensity signal in the second channel. Bin 1804 may be selected following the detection of a low-intensity or zero-intensity signal in the first channel and an intermediate-intensity signal in the second channel. Bin 1807 may be selected following the detection of a low-intensity or zero-intensity signal in the first channel and a low-intensity signal in the second channel.

Four of the nine bins represent matches between respective nucleobases of the two portions sensed during the cycle (bins 1801, 1803, 1807, and 1809). In response to mapping the combined signal to a bin representing a match, the computer processor may detect a match between the first portion and the second portion at the sensed position. In response to mapping the combined signal to a bin representing a match, the computer processor may base call the respective nucleobases. For example, when the combined signal is mapped to bin 1801 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 T. When the combined signal is mapped to bin 1803 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. When the combined signal is mapped to bin 1807 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 1809 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 C.

The remaining five bins are “ambiguous”. That is to say that these bins each represent more than one possible combination of first and second nucleobases. Bins 1802, 1804, 1806, and 1808 each represent two possible combinations of first and second nucleobases. Bin 1805, meanwhile, represents four possible combinations. Nevertheless, mapping the combined signal to an ambiguous bin may still allow for sequencing information to be determined. For example, bins 1802, 1804, 1805, 1806, and 1808 represent mismatches between respective nucleobases of the two portions sensed during the cycle. Therefore, in response to mapping the combined signal to a bin representing a mismatch, the computer processor may detect a mismatch between the first portion and the second portion at the sensed position.

In this particular example, A is configured to emit a signal in both the first channel and the second channel, C is configured to emit a signal in the first channel only, T is configured to emit a signal in the second 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 first channel and the second channel, T may be configured to emit a signal in the first channel only, C may be configured to emit a signal in the second channel only, and G may be configured to not emit a signal in either channel.

The number of classifications which may be selected based upon the combined signal intensities may be predetermined, for example based on the number of portions expected to be present in the nucleic acid cluster. Whilst Figure 11 shows a set of nine possible classifications, the number of classifications may be greater or smaller.

In addition to identifying matches and mismatches, the mapping of the combined signal to each of the different bins (e.g. in combination with additional knowledge, such as the library preparation methods used) can provide additional information about the first portion and the second portion, or about sequences from which the first portion and the second portion were derived. For example, given the nucleic acid material input and the processing methods used to generate the nucleic acid clusters, the first portion and the second portion may be expected to be identical at a given position. In this case, the mapping of the combined signal to a bin representing a mismatch may be indicative of an error introduced during library preparation.

Errors arise during NGS library preparation, for example due to PCR artefacts or DNA damage. The error rate is determined by the library preparation method used, for example the number of cycles of PCR amplification carried out, and a typical error rate may be of the order of 0.1%. This limits the sensitivity of diagnostic assays based on the sequencing method, and may obscure true variants. The present methods allow for the identification of library preparation errors from fewer sequencing reads.

In the absence of any library preparation/sequencing errors, the signals produced by sequencing the two portions (e.g. using sequencing-by-synthesis) will match. The combined signal may therefore be mapped to one of the four “corner” clouds shown in Figure 11 and Figure 12, and the identity of the nucleobase at the corresponding position of the original library polynucleotide can be determined. Should the identity of the nucleobase at that position suggest a rare, or even unknown, variant, it can be determined with a high level of confidence that the base call represents a true variant, as opposed to a library preparation error. If, on the other hand, the combined signal is mapped to any of the other clouds, this indicates that the sequences of the first portion and the second portion do not match, and that an error has occurred in library preparation. Therefore, in response to mapping the combined signal to a classification representing a mismatch between the two nucleobases, a library preparation error may be identified.

Figure 13 is a flow diagram showing a method 1900 of determining sequence information according to the present disclosure. The described method allows for the determination of sequence information from 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.

As shown in Figure 13, the disclosed method 1900 may start from block 1901. The method may then move to block 1910.

At block 1910, 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, for example 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. In one example, high-quality data corresponding to two portions with a substantially equal intensity ratio may have a chastity score of around 0.8 to 0.9, for example 0.89-0.9.

After the intensity data has been obtained, the method may proceed to block 1920. In this step, one of a plurality of classifications is selected based on the intensity data. Each classification represents one or more possible combinations of respective first and second nucleobases, and at least one classification of the plurality of classifications represents more than one possible combination of respective first and second nucleobases. In one example, the plurality of classifications comprises nine classifications as shown in Figure 11. 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 1930, where sequence information of the respective first and second nucleobases is determined based on the classification selected in block 1920. 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). For example, it may be determined that there is a match or a mismatch between the respective first and second nucleobases. Where it is determined that there is a match between the first and second respective nucleobases, the nucleobases may be base called. Whether there is a match or a mismatch, additional or alternative information may be obtained, as described above. 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 1940.

Detection of mismatched base pairs

The present invention is directed to a method of preparing at least one polynucleotide sequence for detection of mismatched base pairs, comprising:

synthesising at least one polynucleotide sequence comprising a first portion and a second portion,

wherein the at least one polynucleotide sequence comprises portions of a double-stranded nucleic acid template, and the first portion comprises a forward strand of the template, and the second portion comprises a reverse complement strand of the template; or wherein the first portion comprises a reverse strand of the template, and the second portion comprises a forward complement strand of the template.

Advantageously, by synthesising at least one polynucleotide sequence comprising the first portion and the second portion, wherein the first portion comprises a forward strand of the template (or reverse strand of the template), and the second portion comprises a reverse complement strand of the template (or forward complement strand of the template), mismatched base pairs can be detected quickly and reliably, which in turn allows errors in the sequencing output to be corrected. The odds of an error appearing from a typical library preparation method are usually in the order of 1 in 10^3. However, with the present invention, the odds that two identical library preparation errors occur in both the forward strand of the template and the reverse complement strand of the template (or the reverse strand of the template and the forward complement strand of the template) is in the order of 1 in 10^7. Thus, sequencing output and accuracy can be increased drastically with the methods of the present invention.

In some embodiments, selective processing methods may be used when preparing the templates. This leads to further advantages, as it also becomes possible to attribute specific nucleobases of the mismatched base pair to particular strands of the original library, thus leading to more precise error detection, whilst maintaining reductions in time taken to detect mismatched base pairs.

The first portion may comprise (or be) the forward strand of a polynucleotide sequence (e.g. forward strand of a template), and the second portion may comprise (or be) the reverse complement strand of the polynucleotide sequence (e.g. reverse complement strand of the template) (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), and the second portion may comprise (or be) the forward complement strand of the polynucleotide sequence (e.g. forward complement strand of the template) (in effect, a forward complement may be considered a “copy” of the reverse strand). In some embodiments, the first portion may be derived from a forward strand of a target polynucleotide to be sequenced, and the second portion may be derived from a reverse complement strand of the target polynucleotide to be sequenced; or the first portion may be derived from a reverse strand of a target polynucleotide to be sequenced, and the second portion may be derived from a forward complement strand of the target polynucleotide to be sequenced. In these particular embodiments, concurrent sequencing of both the forward and reverse complement strands (or the reverse and forward complement strands) allows mismatched base pairs and/or epigenetic modification to be detected.

Where mismatched base pairs are detected, the forward strand of the template may not be identical to the reverse complement strand of the template. Alternatively, the reverse strand of the template may not be identical to the forward complement strand of the template.

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, 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 an intensity of the first signal is substantially the same as 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 sequencing or preparing for selective sequencing) have already been described herein and apply to the methods of preparing at least one polynucleotide sequence for detection of mismatched base pairs as described herein.

The first portion may be referred to herein as read 1 (R1). The second portion may be referred to herein as read 2 (R2).

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 single (concatenated) polynucleotide strand may be attached to a solid support. In one embodiment, this solid support is a flow cell. In one embodiment, the polynucleotide strand is attached to the solid support in a single well of the solid support.

The polynucleotide strand or 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 1400 copies, about 900 to 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 single polynucleotide strand with a first and second portion, before sequencing one group of strands (either the group of template polynucleotides, or the group of template complement polynucleotides thereof) may be removed from the solid support, leaving either the templates or the template complements, as explained above. Such a cluster may be considered to be a "monoclonal" cluster.

By "monoclonal" cluster is meant that the population of polynucleotide sequences that are then sequenced (as the next step) are substantially the same – i.e. copies of the same sequence. As such, a "monoclonal" cluster may refer to the population of single polynucleotide molecules within a well that are then sequenced. A "monoclonal" cluster may contain a sufficient number of copies of a single template polynucleotide (or copies of a single template complement polynucleotide) 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 “monoclonal” 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 1400 copies, about 900 to 1200 copies, or about 1000 copies of a single template polynucleotide (or copies of a single template complement polynucleotide). The copies of the single template polynucleotide (and/or single template complement polynucleotides) 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 monoclonal “cluster”.

The at least one polynucleotide sequence comprising a first portion and a second portion may be prepared using a tandem insert method as described herein. Accordingly, in one embodiment, the step of synthesising the at least one polynucleotide sequence comprising a first portion and a second portion may comprise:

> synthesising a first precursor polynucleotide fragment comprising a complement of the first portion and a hybridisation complement sequence,
>
> synthesising a second precursor polynucleotide fragment comprising a second portion and a hybridisation sequence,
>
> annealing the hybridisation complement sequence of the first precursor polynucleotide fragment with the hybridisation sequence on the second precursor polynucleotide fragment to form a hybridised adduct,
>
> synthesising a first precursor polynucleotide sequence by extending the first precursor polynucleotide fragment to form a complement of the second portion, and
>
> synthesising the at least one polynucleotide sequence by forming a complement of the first precursor polynucleotide sequence.

In one embodiment, the first precursor polynucleotide fragment may comprise a first sequencing primer binding site complement.

In one embodiment, the first sequencing primer binding site complement may be located before a 5’-end of the complement of the first portion, such as immediately before the 5’-end of the complement of the first portion.

In one aspect, the first precursor polynucleotide fragment may comprise a second adaptor complement sequence.

In one example, the second adaptor complement sequence may be located before a 5’-end of the complement of the first portion.

In another embodiment, the first precursor polynucleotide fragment may comprise a first sequencing primer binding site complement and a second adaptor complement sequence.

In one embodiment, the first sequencing primer binding site complement may be located before a 5’-end of the complement of the first portion, and wherein the second adaptor complement sequence may be located before a 5’-end of the first sequencing primer binding site complement.

In one aspect, the first precursor polynucleotide fragment may comprise a second sequencing primer binding site complement.

In one embodiment, the hybridisation sequence complement may comprise the second sequencing primer binding site complement.

In one embodiment, the second precursor polynucleotide fragment may comprise a first adaptor complement sequence.

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

Methods of sequencing

Also described herein is a method of sequencing polynucleotide sequences to detect mismatched base pairs, comprising:

preparing polynucleotide sequences for detection of mismatched base pairs using a method as described herein;

concurrently sequencing nucleobases in the first portion and the second portion; and identifying mismatched base pairs by detecting differences when comparing a sequence output from the first portion with a sequence output from the second portion.

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

In one embodiment, the step of preparing the polynucleotide sequences comprises using a selective processing method as described herein; and wherein the step of concurrent sequencing nucleobases in the first portion and the second portion is based on the intensity of the first signal and the intensity of the second signal.

In one embodiment, the mismatched base pair comprises an oxo-G to A base pair.

In one aspect, 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.

In some embodiments, where a proportion of first portions is capable of generating a first signal and a proportion of second portions is capable of generating a second signal, wherein an intensity of the first signal is substantially the same as an intensity of the second signal, the data may be analysed using 9 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 of the plurality of classifications represents one or more possible combinations of respective first and second nucleobases, and wherein at least one classification of the plurality of classifications represents more than one possible combination of respective first and second nucleobases; and

(d) based on the selected classification, determining sequence information from the first portion and the second portion.

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, when based on a nucleobase of the same identity, an intensity of the first signal component may be substantially the same as an intensity of the second signal component and an intensity of the third signal component is substantially the same as an intensity of the fourth signal component.

In one embodiment, the plurality of classifications may consist of a predetermined number of classifications.

In one embodiment, the plurality of classifications may comprise:

one or more classifications representing matching first and second nucleobases; and

one or more classifications representing mismatching first and second nucleobases, and

wherein determining sequence information of the first portion and second portion comprises:

in response to selecting a classification representing matching first and second nucleobases, determining a match between the first and second nucleobases; or

in response to selecting a classification representing mismatching first and second nucleobases, determining a mismatch between the first and second nucleobases.

In one embodiment, determining sequence information of the first portion and the second portion may comprise, in response to selecting a classification representing a match between the first and second nucleobases, base calling the first and second nucleobases.

In one embodiment, determining sequence information of the first portion and the second portion may comprise, based on the selected classification, determining that the second portion is modified relative to the first portion at a location associated with the 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 sequences for detection of mismatched base pairs 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 sequences for detection of mismatched base pairs as described herein, and/or for sequencing polynucleotide sequences to detect mismatched base pairs as described herein.

In one embodiment, the kit may further comprise a sequencing primer comprising or consisting of a sequence selected from SEQ ID NO. 7 to 16 or a variant or fragment thereof.

In one embodiment, the kit may comprise a sequencing composition comprising a sequencing primer selected from SEQ ID NO. 7 to 10 or a variant or fragment thereof, and a sequencing primer selected from SEQ ID NO. 11 to 16 or a variant or fragment thereof.

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 sequences for detection of mismatched base pairs 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, such as 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.

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.

The present invention will now be described by way of the following non-limiting examples.

Examples

Example 1 – Mismatched base pair analysis and methylation analysis on methylated pUC19 sample using 9 QaM

Oligo sequences:

For transposon annealing (underline indicates ME’ or ME):

ME’-HYB2 (SEQ ID NO. 21)

`/5Phos/CTGTCTCTTATACACATCTGAGTAAGTGGAAGAGATAGGAAGG`

ME’-HYB2’ (SEQ ID NO. 22)

`/5Phos/CTGTCTCTTATACACATCTCCTTCCTATCTCTTCCACTTACTC`

Biotin-A14-ME (SEQ ID NO. 9)

`Biotin-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG`

Biotin-B15-ME (SEQ ID NO. 10)

`Biotin-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG`

Sequencing oligos (underline indicates ME):

HYB2-ME (SEQ ID NO. 12)

`GAGTAAGTGGAAGAGATAGGAAGGAGATGTGTATAAGAGACAG`

HYB2’-ME (SEQ ID NO. 14)

`CCTTCCTATCTCTTCCACTTACTCAGATGTGTATAAGAGACAG`

Preparation of forked adaptors:

1. 5µl of 200µM stock of biotin-A14-ME oligo was combined with 10µl of 100µM stock of ME’-HYB2 oligo. 2µl of 10x TEN Annealing buffer (Illumina) and 3µl of IDTE buffer (Illumina) was added (“A14” transposome mixture).
2. Separately, 5µl of 200µM stock of biotin-B15-ME oligo was combined with 10µl of 100µM stock of ME’-HYB2’ oligo. 2µl of 10x TEN Annealing buffer (Illumina) and 3µl of IDTE buffer (Illumina) was added (“B15” transposome mixture) with 10x TEN and IDTE buffers.
3. Each mixture was heated to 95C for 30s followed by a slow cool (0.1C/s ramp rate) to 10C.
4. 2µl of each annealed mixture was combined with 46µl of Standard Storage Buffer (contains 50% glycerol, Illumina) and 2µl of Tn5 transposase (~90µM stock).

5. Each mixture was mixed and incubated overnight at 37C. Following the incubation step, the two separately prepared transposome complexes were combined together by adding 50µl of each to another 100µl of Standard Storage Buffer to give 200µl of 1µM transposome mix.

Loading of forked adaptors onto beads:

1. 200µl of MyOne T1 Streptavidin beads (Thermofisher) were washed twice with 200µl Tagmentation Wash buffer (TWB, Illumina).
2. Beads were resuspended in 960µl of TWB and 40µl of 1µM transposome mix from step 5 of “Preparation of forked adaptors” was added.
3. Beads were mixed on a rotator for 30mins to 1hr at room temperature.
4. Beads were put on a magnet and beads were washed twice with TWB.
5. Beads were resuspended in original volume (200µl) of BLT Storage Buffer (Illumina). The BLTs were stored at 4C until needed.

Tagmentation:

1. 10µl BLT (bead linked transposomes) from step 5 of “Loading of forked adaptors onto beads” were combined with 100ng DNA in 30µl (pUC19 methylated control DNA) and 10µl of TB1 (5x Tag buffer, Illumina).
2. The combination was mixed and incubated at 55C for 5min, followed by a hold step at 10C.
3. 10µl ST2 Stop buffer was added and mixed.
4. The mixture was incubated at room temp for 5mins.
5. The tubes were transferred to a magnet.
6. The beads were washed twice with 100µl Tagmentation Wash buffer (TWB, Illumina).
7. The beads were resuspended in 50µl of ELM (Extension Ligation Mix, Illumina).
8. The mixture was incubated at 37C for 5mins, then 50C for 5mins, followed by a hold step at 10C.

Hybridisation and extension on beads:

1. The tubes from step 8 of “Tagmentation” were placed on a magnet until the BLT beads pelleted.
2. The beads were washed once with 200µl of Tagmentation Wash Buffer (TWB, Illumina).

3. The beads were washed once with 200µl of 0.1N NaOH – the beads were left to sit in 0.1N NaOH for 30s during this wash step.
4. Beads were washed once with 200µl of TWB.
5. Bears were resuspended in 100µl of HT1 (Hybridisation Buffer, Illumina).
6. Beads were heated in HT1 to 70C for 30s followed by a slow cool (0.1C/s) down to 10C.
7. Beads were washed twice with 200µl of TWB.
8. Beads were resuspended in 100µl of PAM (Patterned Amplification Mix, Illumina) supplemented with 50mM KCl.
9. Beads were heated in PAM to 50C for 5mins, then 60C for 5mins.
10. Beads were washed twice with 200µl of TWB.
11. Beads were resuspended in 50µl of RSB (Resuspension Buffer, Illumina).

Methylation analysis conversion method:

(N.B. For the purposes of detecting mismatched base pairs in the library, the methylation analysis conversion method is not strictly necessary. As such, this step may be skipped if the end goal is to identify only mismatched base pairs, rather than both mismatched base pairs and methylation status.)

1. The following TET master mix (TET MM) was prepared and kept on ice:

	1x (µl)	**4.5x (µl)**
Water	9.00	40.50
Reconstituted TET2 Reaction Buffer (NEB EM-seq kit)	10	45
Oxidation Supplement (NEB EM-seq kit)	1	4.5
DTT (NEB EM-seq kit)	1	4.5
TET2 (NEB EM-seq kit)	4	18
Total	**25**	**112.5**

2. On ice, 25µl of TET MM was added to 20µl of adaptor-ligated DNA in the form of BLTs in RSB (from step 11 of “Hybridisation and extension on beads”).
3. The mixture was vortexed and centrifuged briefly.
4. 500mM of Fe(II) solution (NEB EM-seq kit) was freshly prepared and diluted by adding 1µl to 1249µl of water.
5. 5µl of the diluted Fe(II) solution was added to the 45 µl of adaptor-ligated DNA with TET MM prepared in step 2.

6. The mixture was vortexed (or pipette mixed 10x), centrifuged briefly, incubated for 1hr at 37C, then put on ice.
7. 1µl of Stop reagent was added, vortexed (or pipette mixed 10x), and incubated at 37C for 30 mins.
8. The beads were washed once with 100µl Wash buffer, and then resuspended in 35µl water.
9. In a PCR tube, the 35µl of TET-oxidised DNA from step 8 was combined with 10µl of sodium acetate / acetic acid buffer (pH 4.3) and 5 µl of 1 M pyridine borane. The mixture was incubated overnight at 40C.
10. The beads were washed twice with 100µl Wash buffer, then resuspended in 20 µl of RSB.
11. The 20ul of beads+DNA in RSB from step 10 was combined with 25µl of Q5U Mastermix (NEB) and 5µl of UDI primers (Unique Dual Index primers, Illumina).
12. The mixture was amplified by PCR: cycling procedure – 98C for 30s followed by 3 cycles of (98C 10s, 62C 30s, 65C 3min), then 6 cycles of (98C 10s, 62C 30s, 65C 30s), 65C for 5 mins and then hold at 4C.
13. PCR products were analysed by TapeStation D1000 (Agilent), and then subjected to a further SPRI clean-up before quantification using a Qubit Broad Range dsDNA assay kit (Thermofisher).

Sequencing:

Sequencing was conducted on the MiniSeq. Standard clustering on the MiniSeq and a standard first hyb was conducted for the 1st 36 cycles of sequencing.

A custom second hyb was used from the “Cust3” position of the reagent cartridge. This primer hyb maintains a higher temperature (60C) than normal during the post-hyb wash (which usually drops to 40C). This higher temperature was to ensure that the right sequencing primers hyb to the right places on the cluster strands.

The primer mix for this custom hyb was HP10 R1 primer mix (Illumina) spiked with 0.5µM each of HYB2’-ME and HYB2-ME primers. These primers are all unblocked and allow concurrent sequencing of both the first portion and the second portion, and so generate the 9 QaM signal during sequencing. The converted library was loaded onto the MiniSeq cartridge at 1pM final concentration. The MiniSeq was set up to save 3 tiles of images per cycle, for later off-line analysis. The 9 QaM results are shown in Figure 14A, where modified cytosines can be identified by a characteristic central cloud in the plot (indicated by circled region). Of course, the (5-mC)-G base pair (or a G-(5-mC) base pair), which is subsequently converted to a mismatched T-G base pair (or a G-T base pair) by TAPS, represents a type of mismatched base pair. Other mismatched base pairs can be identified by side clouds (top middle, bottom middle, centre left, centre right – indicated by boxed regions) The actual genetic sequences are shown in Figure 14B, where modified cytosines can be assigned to cases where a C-T mismatch is observed between the HYB2'-ME read and the HP10 read.

Overall, these results (in particular the custom second hyb results) show that analysis can be conducted on polynucleotide sequences to find mismatched base pairs. In addition, methylation analysis can be conducted on polynucleotide sequences to identify modified cytosines – however, this is not strictly necessary for the purposes of the present invention if the methylation analysis conversion method is skipped. In particular, by enabling concurrent sequencing of the forward and reverse complement strands of the template (or reverse and forward complement strands of the template), mismatched base pairs can be identified quickly and accurately.

Example 2 – Mismatched base pair analysis on human DNA sample using 9 QaM

A similar experiment to Example 1 was conducted except that the DNA during the "Tagmentation" section was replaced with a Promega human blend DNA spiked with 5% PhiX (as control). In addition, the steps from "Methylation analysis conversion method" were not conducted – thus, any errors would be indicative of mismatched base pairs, for example, as a result of errors resulting from library preparation.

Sequencing was conducted on the NextSeq 2000. A custom hyb was conducted where the usual primer mix was replaced with HP10 primer mix (Illumina) spiked with HYB2'-ME primer (0.3 µM each). These primers are all unblocked and allow concurrent sequencing of both the first portion and the second portion, and so generate the 9 QaM signal during sequencing. The library was loaded onto the NextSeq 2000 at 650pM final concentration. These results are presented in Figure 15B (Read 3 – combined Read 1 and Read 2), where mismatched base pairs can be identified by characteristic off-corner clouds in the plot (indicated by point in circled region). In this case, a C-T mismatch (a middle cloud) was detected, leading to an "N" readout in the Read 3 sequence.

Control experiments were also conducted where individual reads were done separately using only one sequencing primer type (Read 1 and Read 2 separately). One of the reads on the tandem insert corresponds to a readout for the forward strand, whilst the other read on the tandem insert corresponds to a readout for the reverse complement strand. In the Read 1 case, using a HP21 primer mix (Illumina), one of the bases is detected as T (indicated by point in circled region); in the Read 2 case, using a HYB2'-ME primer, one of the bases is detected as C (indicated by point in circled region). The control experiment confirms that the detection of the C-T mismatch in the Read 3 case was correct, using only one read run.

Overall, these results show that analysis can be conducted on polynucleotide sequences to find mismatched base pairs. Again, by enabling concurrent sequencing of the forward and reverse complement strands of the template (or reverse and forward complement strands of the template), mismatched base pairs can be identified quickly and accurately.

SEQUENCE LISTING

(Underlined sequences are ME or ME’ sequences)

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: A14

TCGTCGGCAGCGTC

SEQ ID NO. 8: B15

GTCTCGTGGGCTCGG

SEQ ID NO. 9: A14-ME

TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG

SEQ ID NO. 10: B15-ME

GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

SEQ ID NO. 11: HYB2

GAGTAAGTGGAAGAGATAGGAAGG

SEQ ID NO. 12: HYB2-ME

GAGTAAGTGGAAGAGATAGGAAGGAGATGTGTATAAGAGACAG

SEQ ID NO. 13: HYB2’

CCTTCCTATCTCTTCCACTTACTC

SEQ ID NO. 14: HYB2'-ME

CCTTCCTATCTCTTCCACTTACTCAGATGTGTATAAGAGACAG

SEQ ID NO. 15: HYB2'-block

CCTTCCTATCTCTTCCACTTACT-3'propanol

SEQ ID NO. 16: HYB2'-ME-block

CCTTCCTATCTCTTCCACTTACTCAGATGTGTATAAGAGACAG-3'propanol

SEQ ID NO. 17: ME'-A14'

CTGTCTCTTATACACATCTGACGCTGCCGACGA

SEQ ID NO. 18: A14'

GACGCTGCCGACGA

SEQ ID NO. 19: ME'-B15'

CTGTCTCTTATACACATCTCCGAGCCCACGAGAC

SEQ ID NO. 20: B15'

CCGAGCCCACGAGAC

SEQ ID NO. 21: ME'-HYB2

CTGTCTCTTATACACATCTGAGTAAGTGGAAGAGATAGGAAGG

SEQ ID NO. 22: ME'-HYB2'

CTGTCTCTTATACACATCTCCTTCCTATCTCTTCCACTTACTC

CLAIMS:

1. A method of preparing at least one polynucleotide sequence for detection of mismatched base pairs, comprising:

   synthesising at least one polynucleotide sequence comprising a first portion and a second portion,

   wherein the at least one polynucleotide sequence comprises portions of a double-stranded nucleic acid template, and the first portion comprises a forward strand of the template, and the second portion comprises a reverse complement strand of the template; or wherein the first portion comprises a reverse strand of the template, and the second portion comprises a forward complement strand of the template.

2. A method according to claim 1, wherein the forward strand of the template is not identical to the reverse complement strand of the template.

3. A method according to claim 1 or claim 2, wherein the method further comprises a step of preparing the first portion and the second portion for concurrent sequencing.

4. A method according to claim 3, 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.

5. A method according to any one of claims 1 to 4, wherein a proportion of first portions is capable of generating a first signal and a proportion of second portions is capable of generating a second signal, wherein an intensity of the first signal is substantially the same as an intensity of the second signal.

6. A method according to any one of claims 1 to 4, wherein the method further 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.

7. A method according to claim 6, 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.

8. A method according to claim 7, 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, preferably between 1.5:1 to 3:1, more preferably about 2:1.

9. A method according to any one of claims 6 to 8, wherein selective processing comprises preparing for selective sequencing or conducting selective sequencing.

10. A method according to any one of claims 6 to 9, 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.

11. A method according to claim 10, wherein the blocked second primer comprises a blocking group at a 3’ end of the blocked second primer.

12. A method according to claim 11, 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.

13. A method according to any one of claims 10 to 12, wherein the blocked second primer comprises a sequence as defined in SEQ ID NO. 11 to 16 or a variant or fragment thereof and/or the unblocked second primer comprises a sequence as defined in SEQ ID NO. 11 to 14 or a variant or fragment thereof.

14. A method according to any one of claims 5 to 13, wherein the first signal and the second signal are spatially unresolved.

15. A method according to any one of claims 1 to 14, wherein the at least one polynucleotide sequence comprising the first portion and the second portion is/are attached to a solid support, preferably wherein the solid support is a flow cell.

16. A method according to claim 15, wherein the at least one polynucleotide sequence comprising the first portion and the second portion forms a cluster on the solid support.

17. A method according to claim 16, wherein the cluster is formed by bridge amplification.

18. A method according to any one of claims 15 to 17, wherein the at least one polynucleotide sequence comprising the first portion and the second portion forms a monoclonal cluster.

19. A method according to any one of claims 15 to 18, wherein the solid support comprises at least one first immobilised primer and at least one second immobilised primer.

20. A method according to claim 19, 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.

21. A method according to claim 19 or claim 20, wherein each polynucleotide sequence comprising the first portion and the second portion is attached to a first immobilised primer.

22. A method according to any one of claims 19 to 21, wherein each polynucleotide sequence comprising the first portion and the second portion further comprises a

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

23. A method according to any one of claims 1 to 22, wherein the step of synthesising the at least one polynucleotide sequence comprising a first portion and a second portion comprises:

synthesising a first precursor polynucleotide fragment comprising a complement of the first portion and a hybridisation complement sequence,

synthesising a second precursor polynucleotide fragment comprising a second portion and a hybridisation sequence,

annealing the hybridisation complement sequence of the first precursor polynucleotide fragment with the hybridisation sequence on the second precursor polynucleotide fragment to form a hybridised adduct,

synthesising a first precursor polynucleotide sequence by extending the first precursor polynucleotide fragment to form a complement of the second portion, and

synthesising the at least one polynucleotide sequence by forming a complement of the first precursor polynucleotide sequence.

24. A method according to claim 23, wherein the first precursor polynucleotide fragment comprises a first sequencing primer binding site complement.

25. A method according to claim 24, wherein the first sequencing primer binding site complement is located before a 5'-end of the complement of the first portion, preferably immediately before the 5'-end of the complement of the first portion.

26. A method according to any one of claims 23 to 25, wherein the first precursor polynucleotide fragment comprises a second adaptor complement sequence.

27. A method according to claim 26, wherein the second adaptor complement sequence is located before a 5'-end of the complement of the first portion.

28. A method according to any one of claims 23 to 27, wherein the first precursor polynucleotide fragment comprises a first sequencing primer binding site complement and a second adaptor complement sequence.

29. A method according to claim 28, wherein the first sequencing primer binding site complement is located before a 5'-end of the complement of the first portion, and wherein the second adaptor complement sequence is located before a 5'-end of the first sequencing primer binding site complement.

30. A method according to any one of claims 23 to 29, wherein the first precursor polynucleotide fragment comprises a second sequencing primer binding site complement.

31. A method according to claim 30, wherein the hybridisation sequence complement comprises the second sequencing primer binding site complement.

32. A method according to any one of claims 23 to 31, wherein the second precursor polynucleotide fragment comprises a first adaptor complement sequence.

33. A method according to any one of claims 1 to 32, wherein the method further comprises concurrently sequencing nucleobases in the first portion and the second portion.

34. A method of sequencing at least one polynucleotide sequence to detect mismatched base pairs, comprising:

    preparing at least one polynucleotide sequence for detection of mismatched base pairs using a method according to any one of claims 1 to 32;

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

    identifying mismatched base pairs by detecting differences when comparing a sequence output from the first portion with a sequence output from the second portion.

35. A method according to claim 34, wherein the step of concurrently sequencing nucleobases comprises performing sequencing-by-synthesis or sequencing-by-ligation.

36. A method according to claim 34 or claim 35, wherein the step of preparing the at least one polynucleotide sequence comprises using a method according to any one of claims 5 to 14; and wherein the step of concurrent sequencing

nucleobases in the first portion and the second portion is based on the intensity of the first signal and the intensity of the second signal.

37. A method according to any one of claims 34 to 36, wherein the mismatched base pair comprises an oxo-G to A base pair.

38. A method according to any one of claims 34 to 37, wherein the method further comprises a step of conducting paired-end reads.

39. A kit comprising instructions for preparing at least one polynucleotide sequence for detection of mismatched base pairs according to any one of claims 1 to 33, and/or for sequencing at least one polynucleotide sequence to detect mismatched base pairs according to any one of claims 34 to 38.

40. A data processing device comprising means for carrying out a method according to any one of claims 1 to 38.

41. A data processing device according to claim 40, wherein the data processing device is a polynucleotide sequencer.

42. 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 38.

43. 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 38.

44. A computer-readable data carrier having stored thereon a computer program product according to claim 42.

45. A data carrier signal carrying a computer program product according to claim 42.

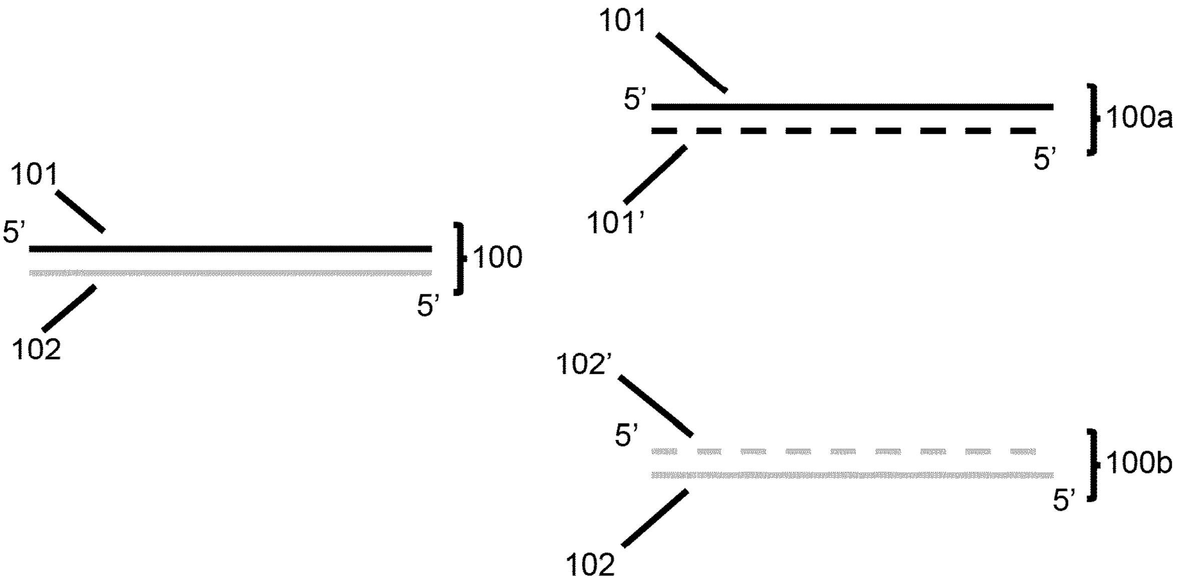


FIG. 1

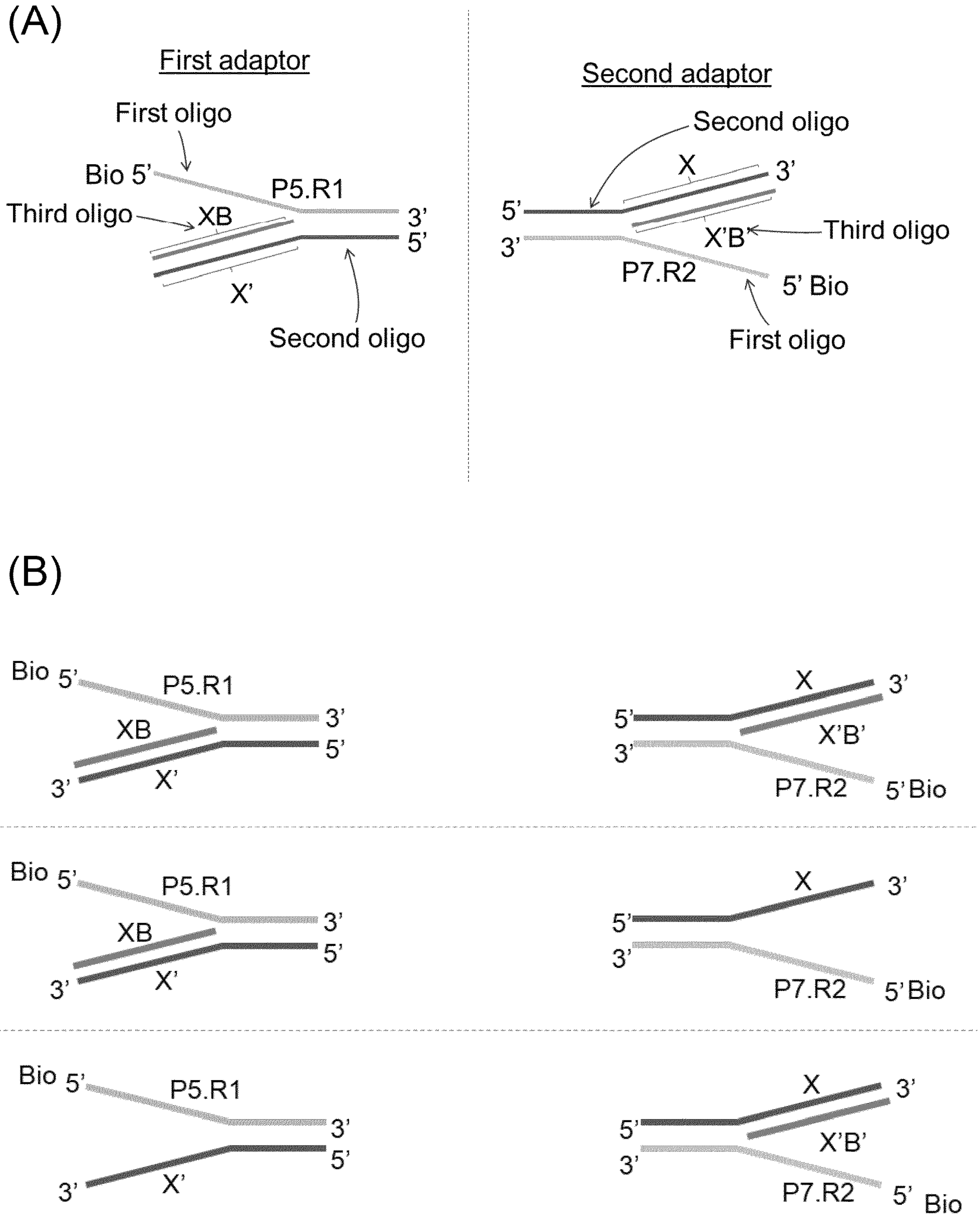

(A)
First adaptor
First oligo
Bio 5'
P5.R1
Third oligo
XB
3'
5'
X'
Second oligo
Second adaptor
Second oligo
X
3'
5'
3'
X'B'
Third oligo
P7.R2
5' Bio
First oligo
(B)
Bio 5'
P5.R1
XB
3'
5'
3'
X'
X
3'
5'
X'B'
3'
P7.R2
5' Bio
Bio 5'
P5.R1
XB
3'
5'
3'
X'
X
3'
5'
3'
P7.R2
5' Bio
Bio 5'
P5.R1
3'
5'
3'
X'
X
3'
5'
X'B'
3'
P7.R2
5' Bio

FIG. 2

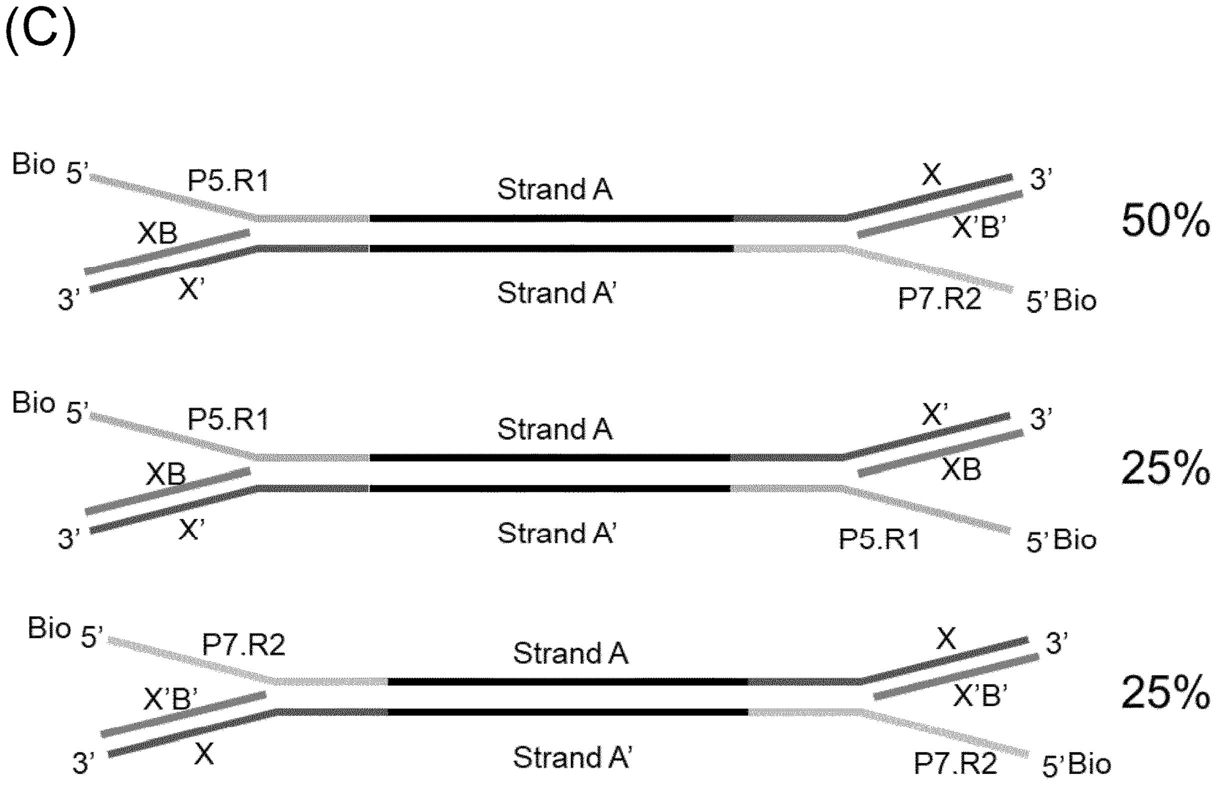


FIG. 2
(continued)

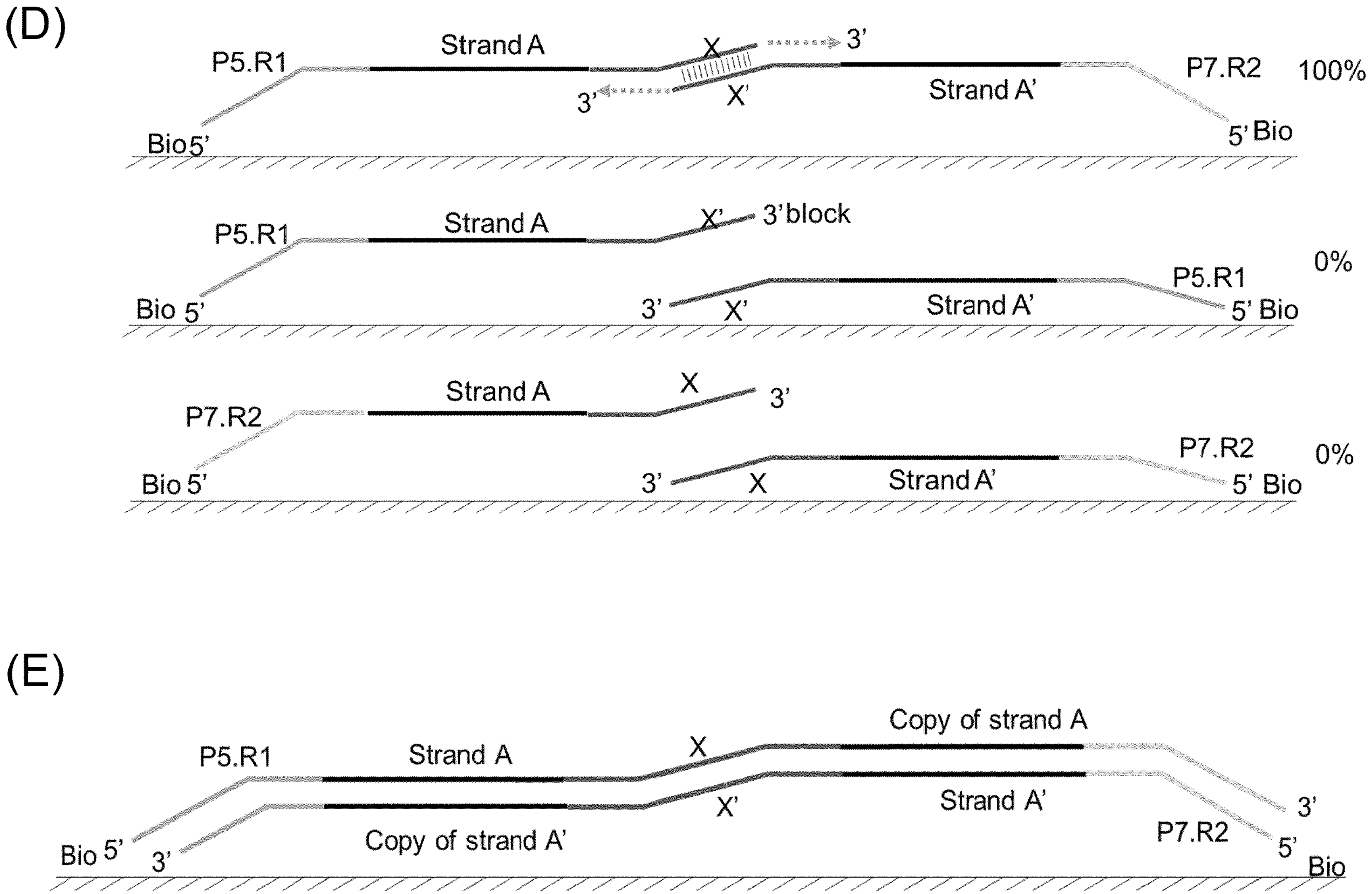


FIG. 2 (continued)

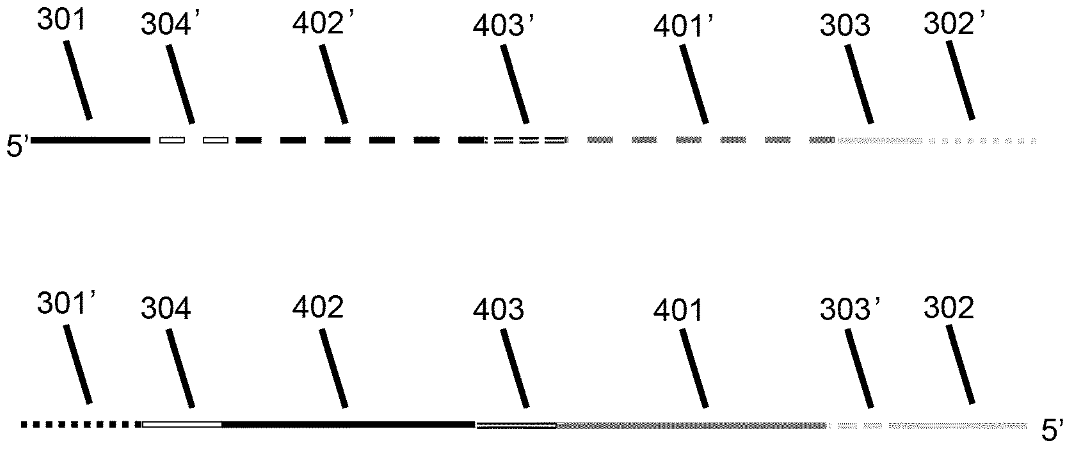


FIG. 3

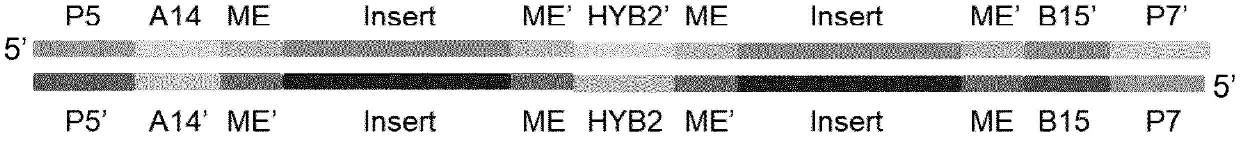


FIG. 4

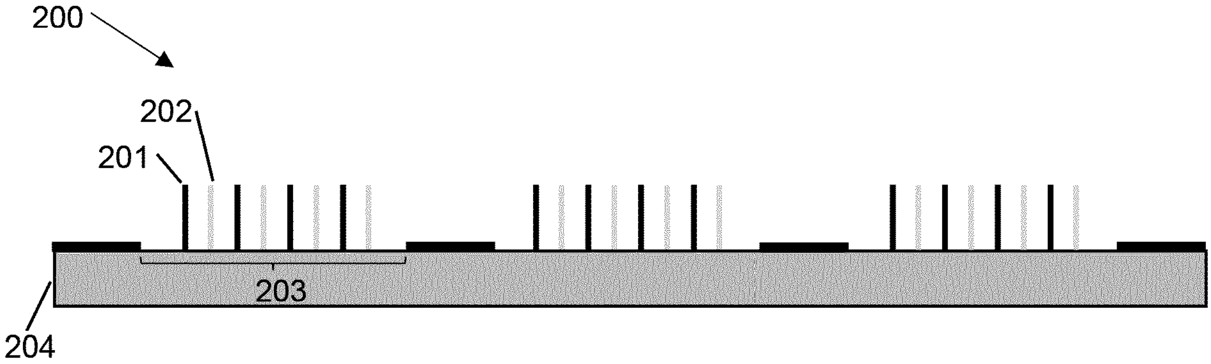


FIG. 5

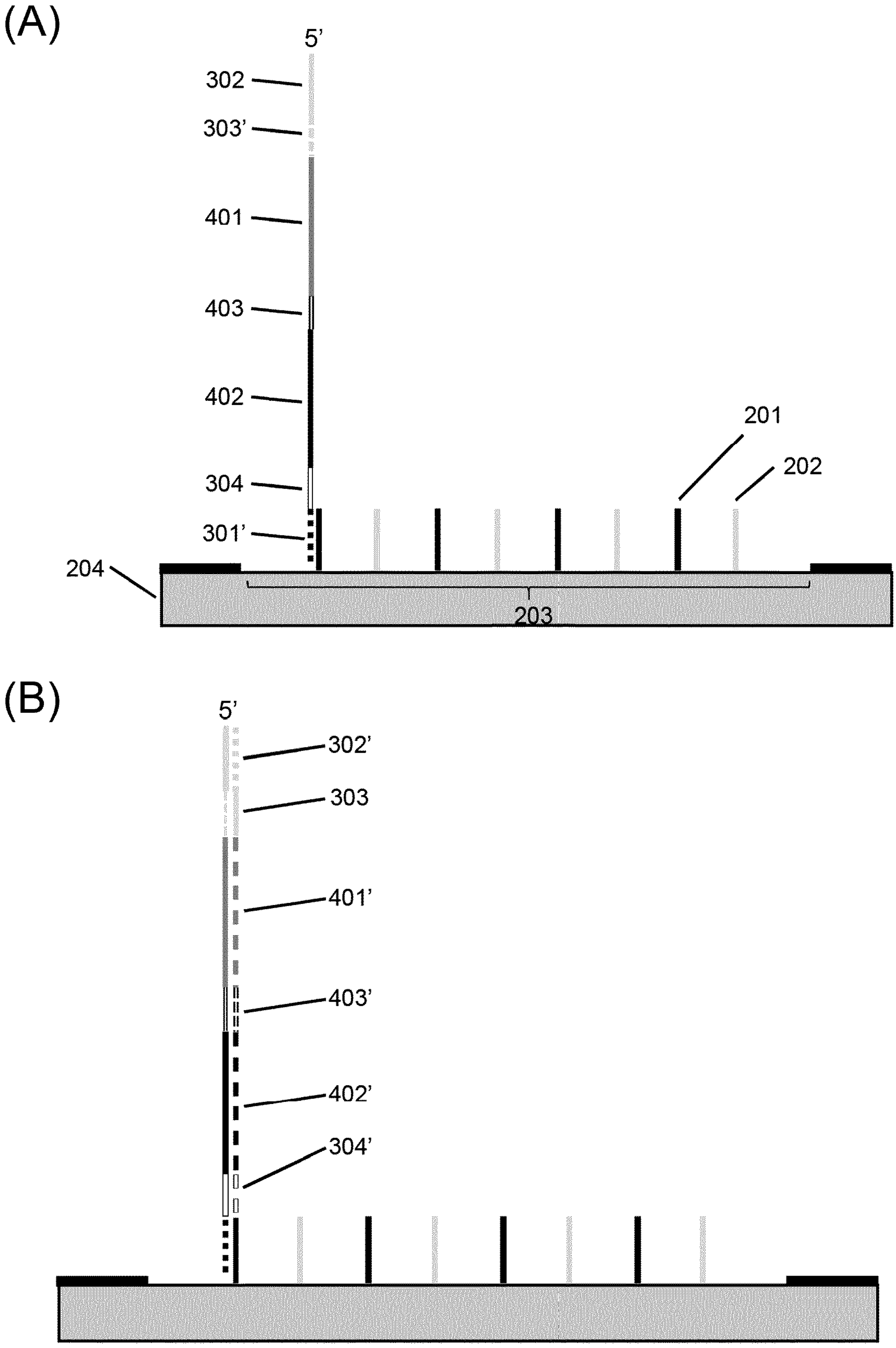

(A)
5’
302
303’
401
403
402
304
301’
201
202
204
203
(B)
5’
302’
303
401’
403’
402’
304’

FIG. 6

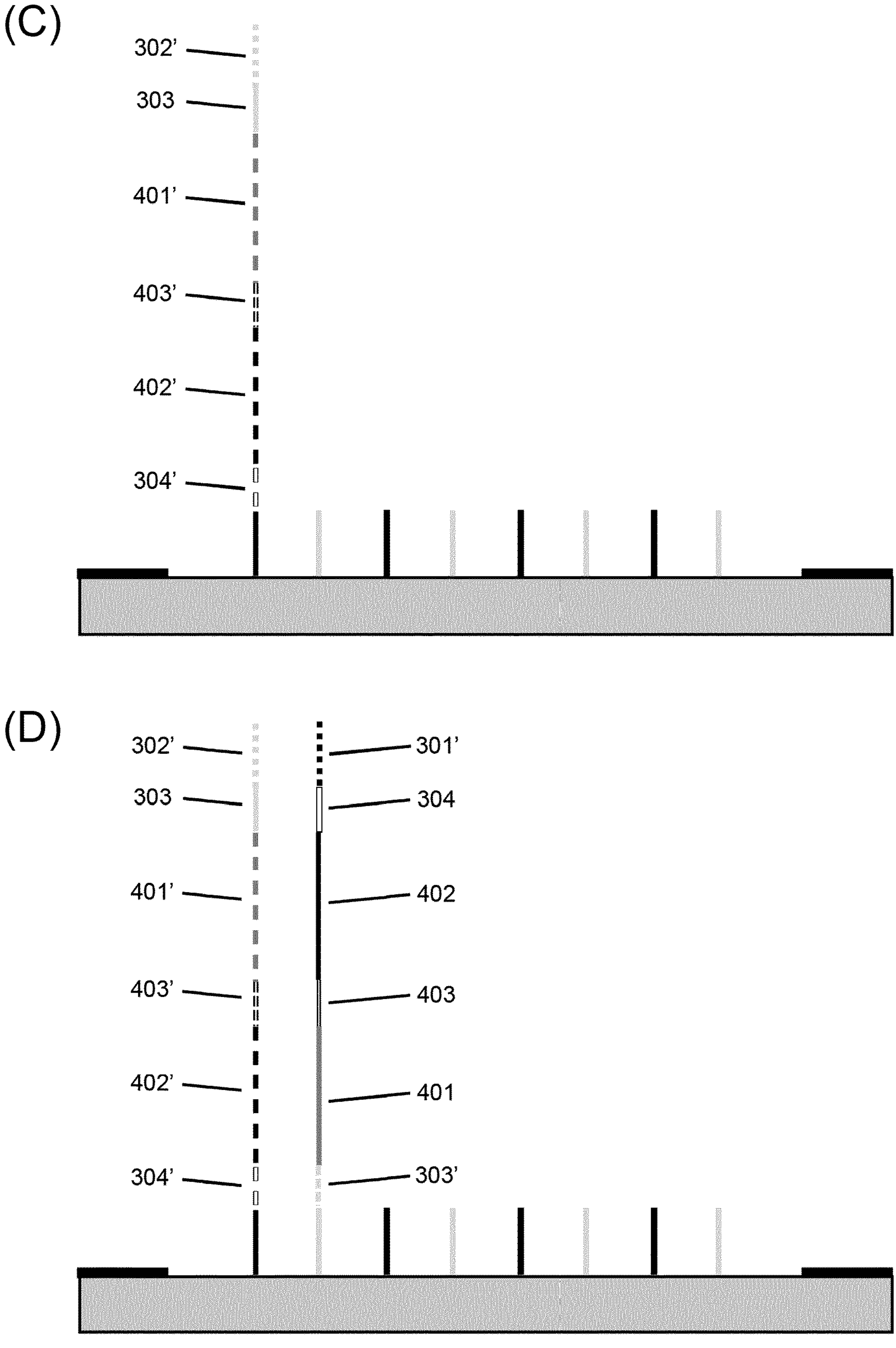


FIG. 6
(continued)

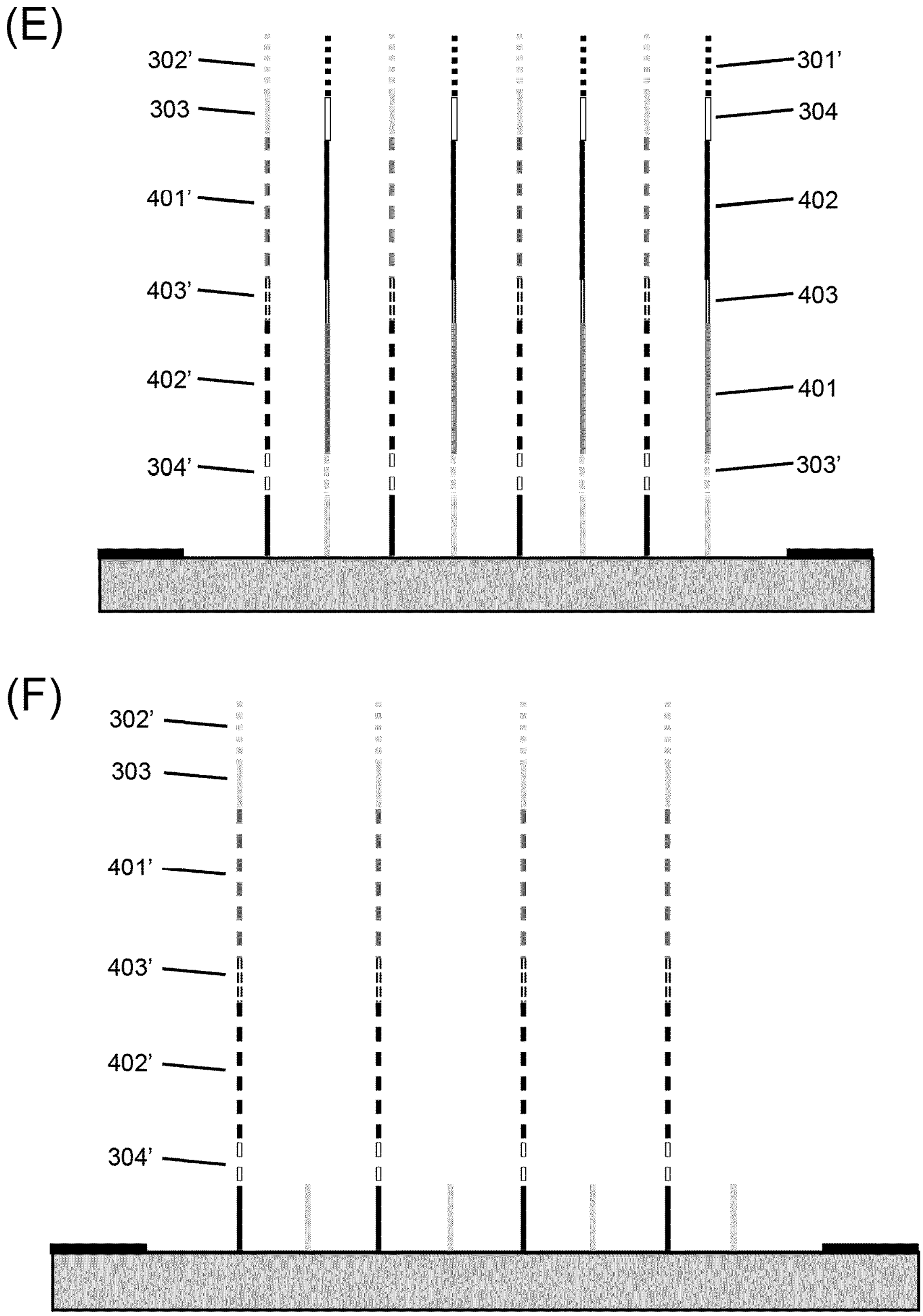
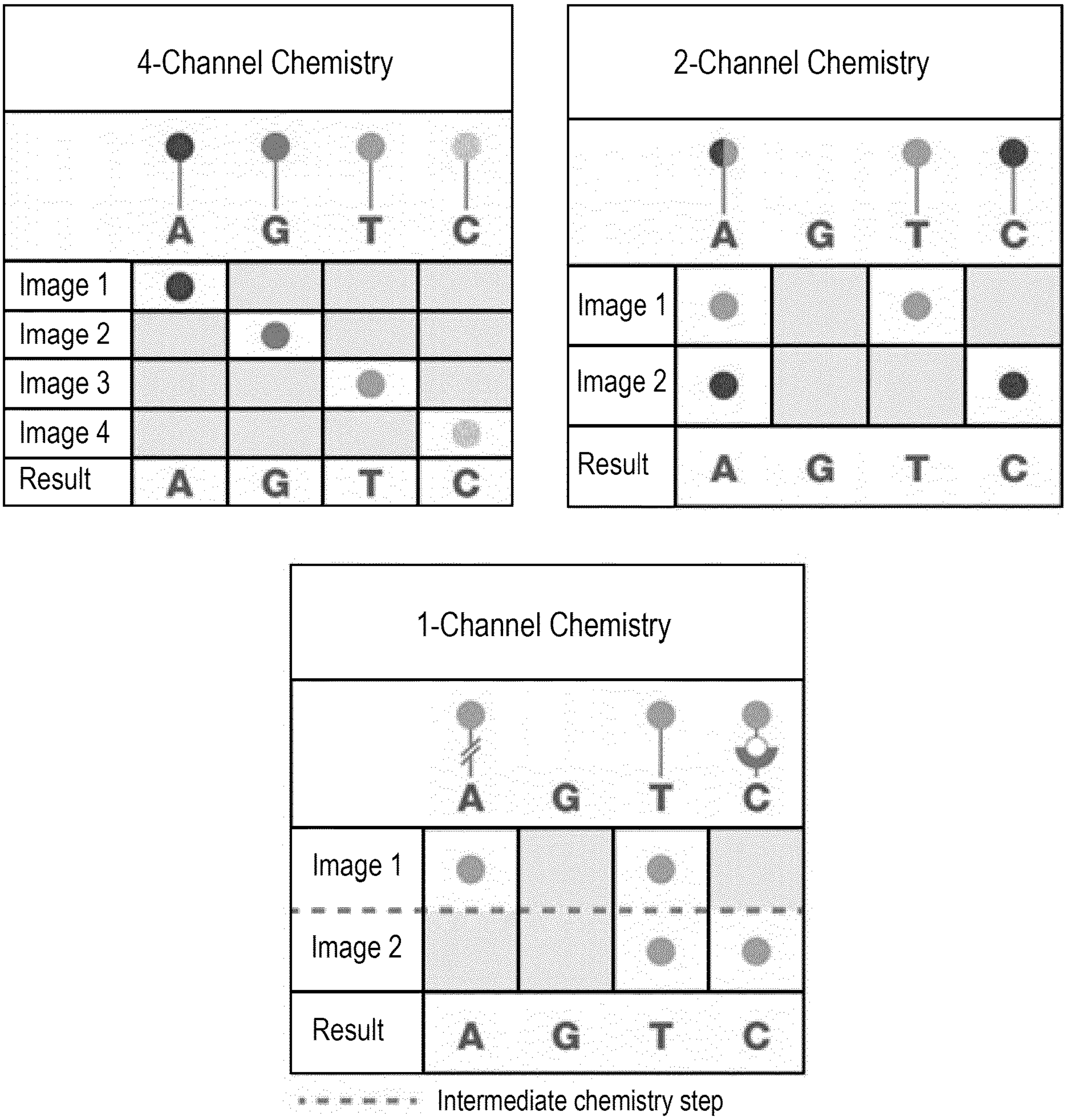


FIG. 6
(continued)



4-Channel Chemistry

	A	G	T	C
Image 1	●			
Image 2		●		
Image 3			●	
Image 4				●
Result	A	G	T	C

2-Channel Chemistry

	A	G	T	C
Image 1	●		●	
Image 2	●			●
Result	A	G	T	C

1-Channel Chemistry

	A	G	T	C
Image 1	●		●	
Image 2			●	●
Result	A	G	T	C

FIG. 7

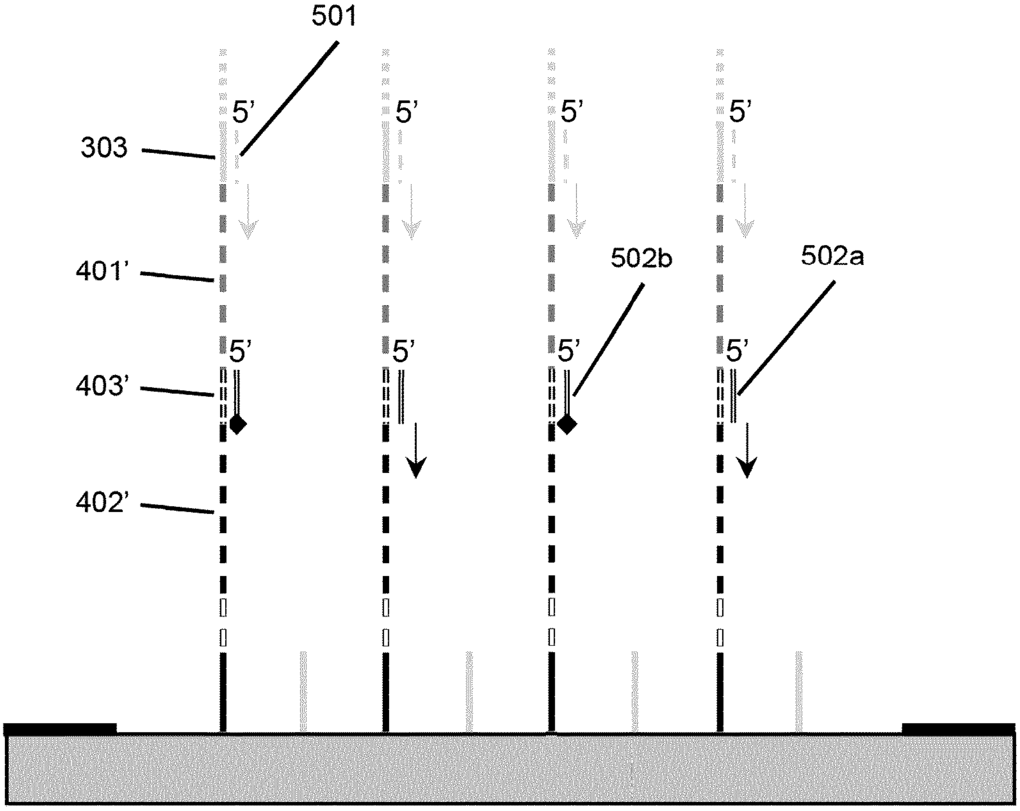


FIG. 8

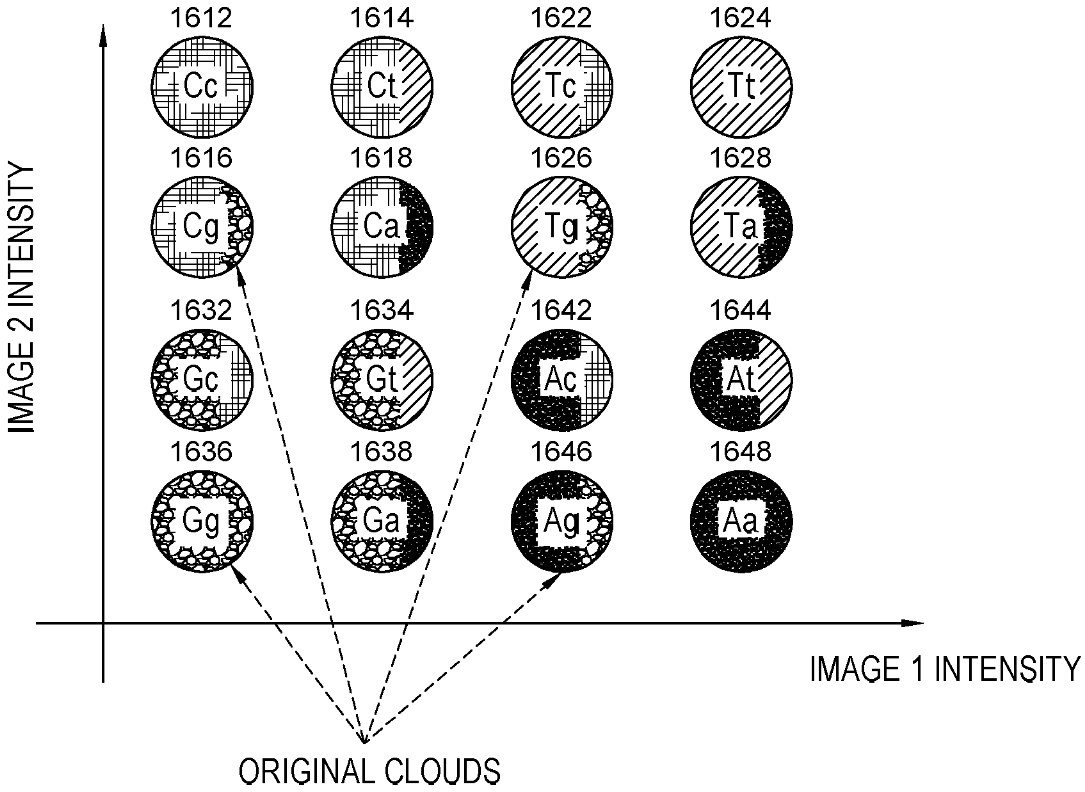

1612
Cc
1614
Ct
1622
Tc
1624
Tt
1616
Cg
1618
Ca
1626
Tg
1628
Ta
1632
Gc
1634
Gt
1642
Ac
1644
At
1636
Gg
1638
Ga
1646
Ag
1648
Aa
IMAGE 2 INTENSITY
IMAGE 1 INTENSITY
ORIGINAL CLOUDS

FIG. 9

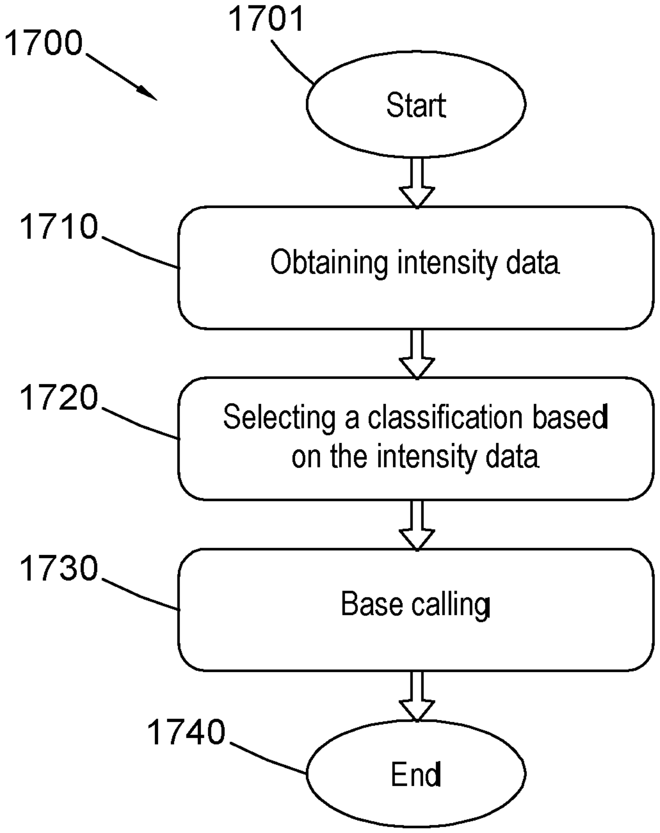

1700
1701
Start
1710
Obtaining intensity data
1720
Selecting a classification based on the intensity data
1730
Base calling
1740
End

FIG. 10

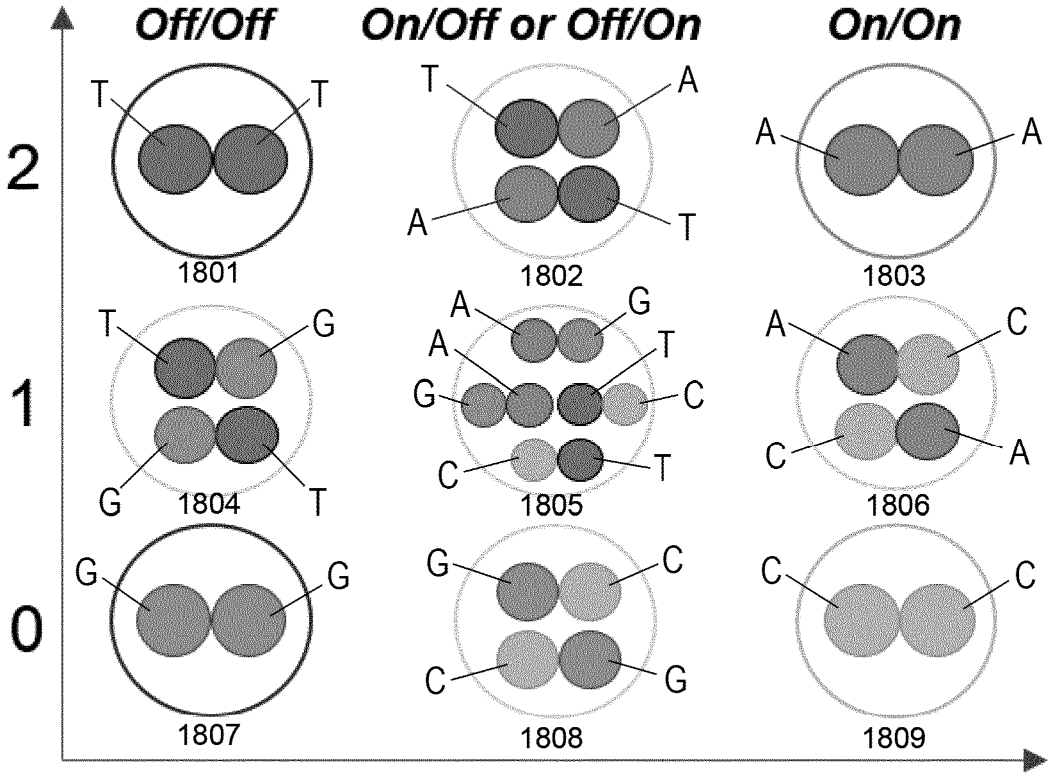

Off/Off
On/Off or Off/On
On/On
2
1
0
T
T
1801
T
A
A
T
1802
A
A
1803
T
G
G
1804
T
A
G
A
T
G
C
C
T
1805
A
C
C
A
1806
G
G
1807
G
C
C
G
1808
C
C
1809

FIG. 11

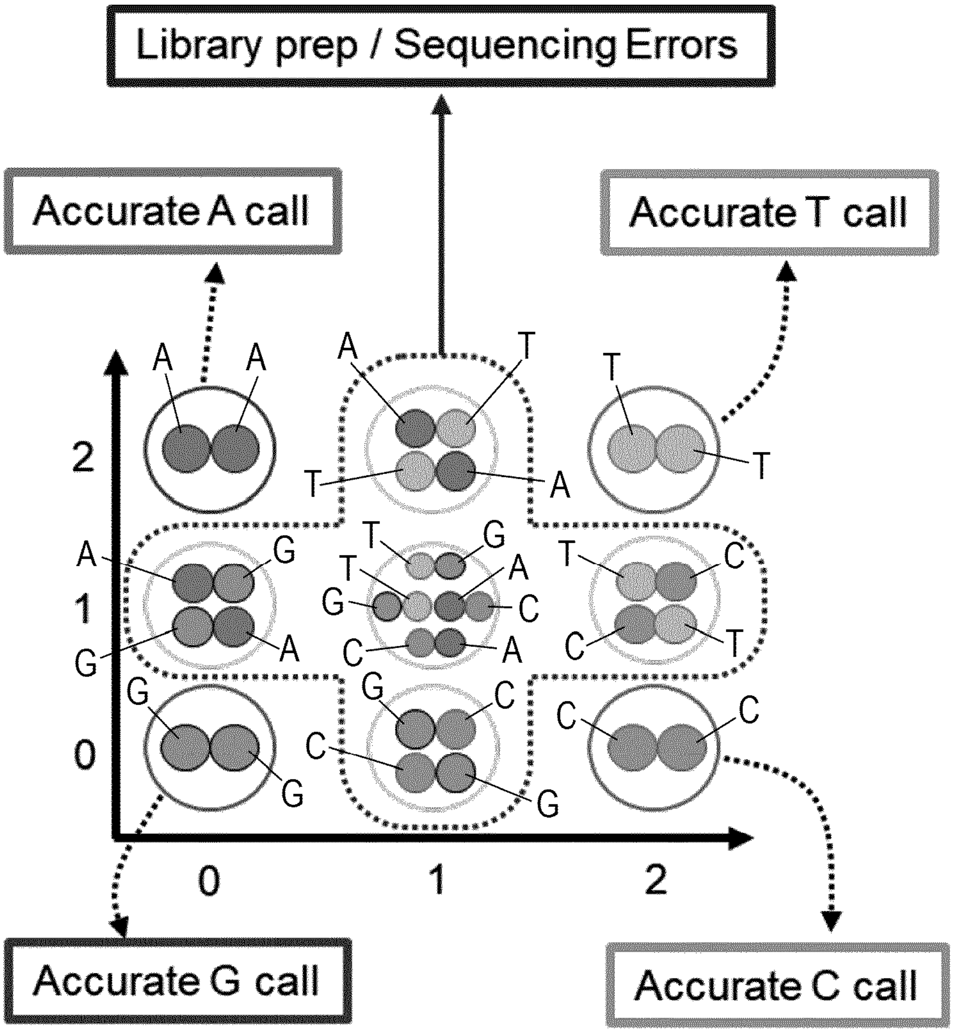

Library prep / Sequencing Errors
Accurate A call
Accurate T call
Accurate G call
Accurate C call
A
T
G
C
0
1
2

FIG. 12

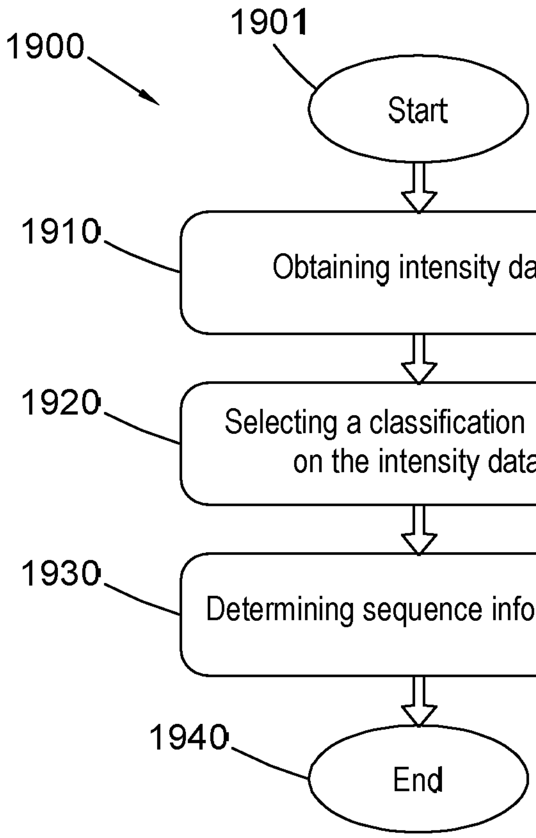

1900
1901
Start
1910
Obtaining intensity data
1920
Selecting a classification based on the intensity data
1930
Determining sequence information
1940
End

FIG. 13

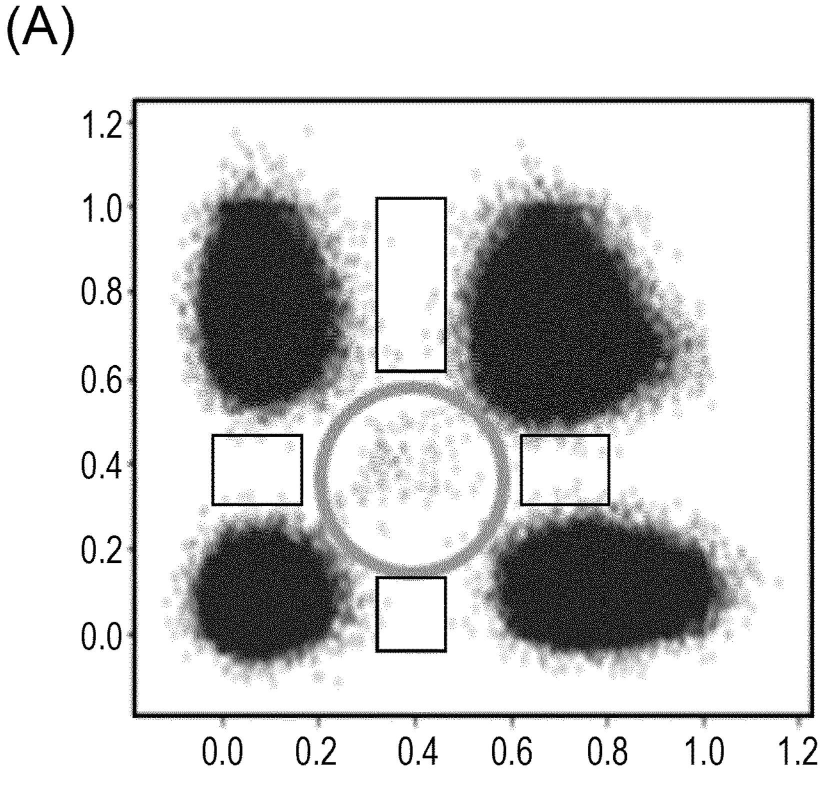
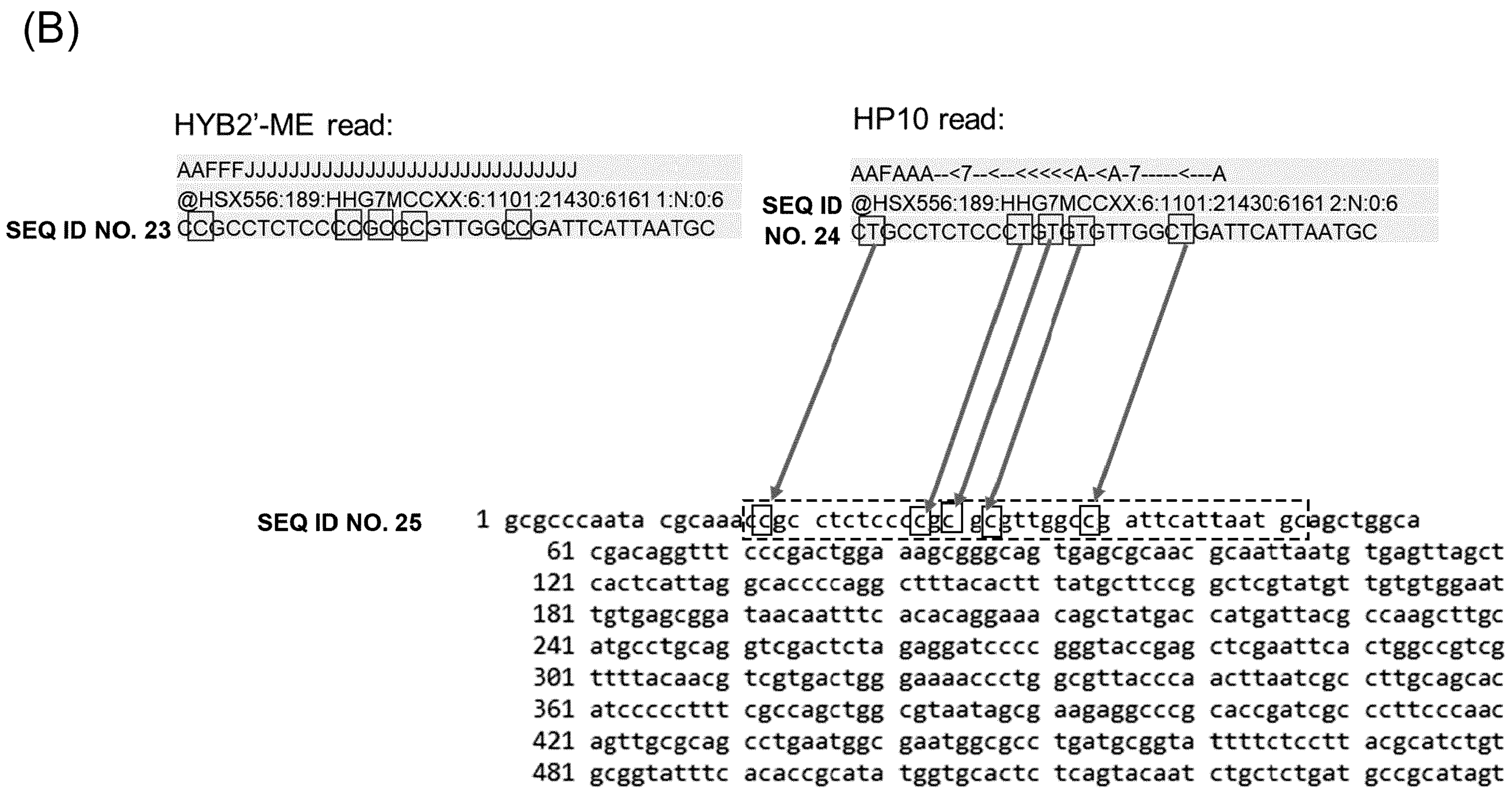

(A)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2

FIG. 14



```
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@HSX556:189:HHG7MCCXX:6:1101:21430:6161 1:N:0:6
SEQ ID NO. 23 CCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGC
```

```
AAFAAA--<7--<--<<<<<A-<A-7-----<---A
@HSX556:189:HHG7MCCXX:6:1101:21430:6161 2:N:0:6
SEQ ID NO. 24 CTGCCTCTCCCTGTGTGTTGGCTGATTCATTAATGC
```

```
SEQ ID NO. 25   1 gcgcccaata cgcaaaccgc ctctccccgc gcgttggccg attcattaat gcagctggca
               61 cgacaggttt cccgactgga aagcgggcag tgagcgcaac gcaattaatg tgagttagct
              121 cactcattag gcaccccagg ctttacactt tatgcttccg gctcgtatgt tgtgtggaat
              181 tgtgagcgga taacaatttc acacaggaaa cagctatgac catgattacg ccaagcttgc
              241 atgcctgcag gtcgactcta gaggatcccc gggtaccgag ctcgaattca ctggccgtcg
              301 ttttacaacg tcgtgactgg gaaaaccctg gcgttaccca acttaatcgc cttgcagcac
              361 atcccccttt cgccagctgg cgtaatagcg aagaggcccg caccgatcgc ccttcccaac
              421 agttgcgcag cctgaatggc gaatggcgcc tgatgcggta ttttctcctt acgcatctgt
              481 gcggtatttc acaccgcata tggtgcactc tcagtacaat ctgctctgat gccgcatagt
```

FIG. 14
(continued)

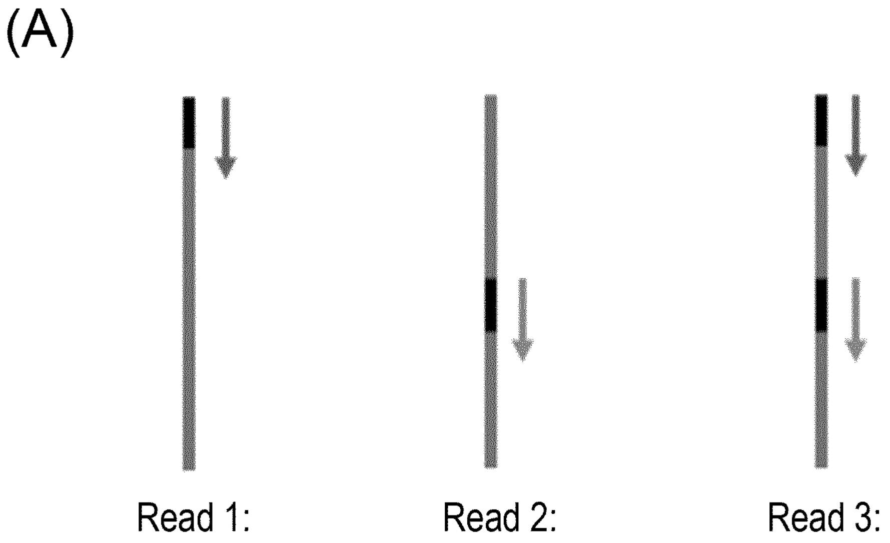
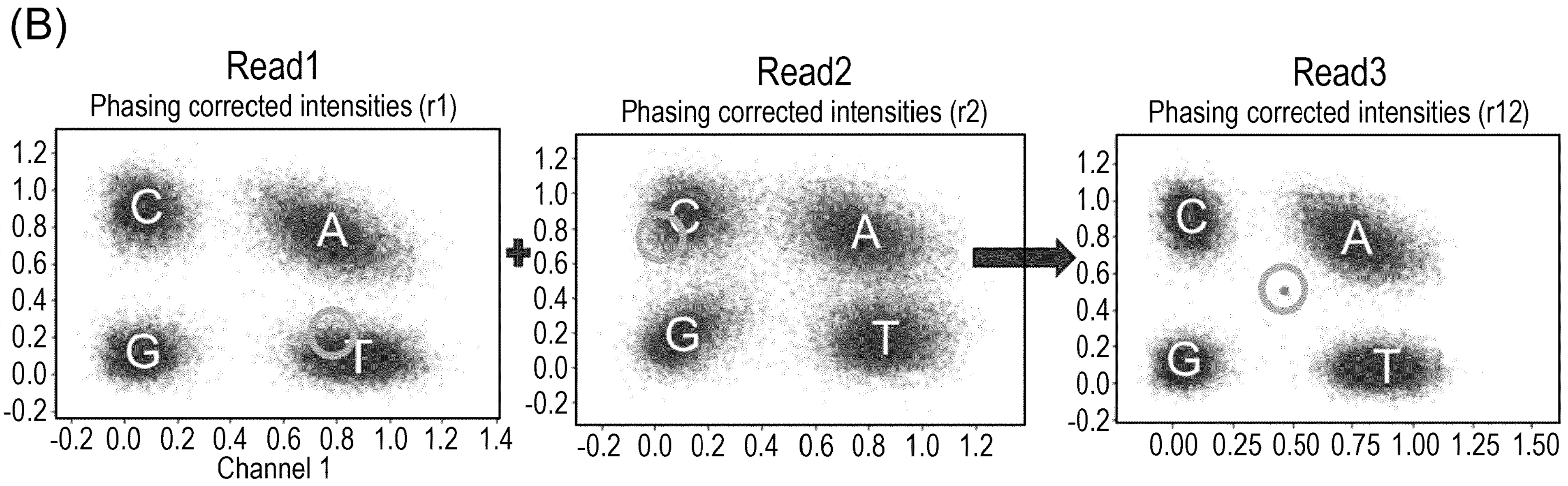


FIG. 15



```
SEQ ID NO. 26 R1:  GGAAGCTGGATGGCCCTTGGGGGCTGACCCGCAGGGTGCTGGACTTCGGGATATAGCAGAGAGAGCTTGGCATGACTTATTACTCCAGGCTGTAGAATC
                                                                     |
SEQ ID NO. 27 R2:  GGAAGCTGGATGGCCCTCAGCGGCTGACCCACAGGGTGCCAGATTTTGGGACATAGCAGAGAGAGCTTGGCACGTCTTATTACTCCAGGCTGTTGAATC
                                                                     |
SEQ ID NO. 28 R3:  GGAAGCTGGATGGCCCTAAGCGGCTGACCCACAGGGTGCAAGAATTAGGGANATAGCAGAGAGAGCTTGGCAAGACTTATTACTCCAGGCTGTAGAATC
```

FIG. 15
(continued)

INTERNATIONAL SEARCH REPORT

International application No.

PCT/EP2023/056669

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. ☒ 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

International application No
PCT/EP2023/056669

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/6869 C12Q1/6874
ADD.

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

B. FIELDS SEARCHED

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

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

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

EPO-Internal, BIOSIS, Sequence Search, EMBASE, 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	WO 2007/010263 A2 (SOLEXA LTD [GB]; SWERDLOW HAROLD PHILIP [GB]) 25 January 2007 (2007-01-25) abstract claims 12, 18, 25, 29 figures 3-6 -----	1-45
X	WO 2007/010252 A1 (SOLEXA LTD [GB]; SMITH GEOFFREY PAUL [GB]) 25 January 2007 (2007-01-25) cited in the application abstract claims 1,3,9-14 figure 1 ----- -/--	1-45

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

☒ 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 28 June 2023	Date of mailing of the international search report 05/07/2023
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Authorized officer Barz, Wolfgang

1

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2023/056669

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 98/44151 A1 (GLAXO GROUP LTD [GB]; KAWASHIMA ERIC [CH] ET AL.) 8 October 1998 (1998-10-08) cited in the application abstract claims 1, 3, 34, 58-63 figures 1,5,8,10 -----	1

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2023/056669

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2007010263 A2	25-01-2007	AT 456676 T	15-02-2010
		EP 1907573 A2	09-04-2008
		EP 2189540 A1	26-05-2010
		US 2010311597 A1	09-12-2010
		US 2017342482 A1	30-11-2017
		US 2021047685 A1	18-02-2021
		US 2023098456 A1	30-03-2023
		WO 2007010263 A2	25-01-2007
WO 2007010252 A1	25-01-2007	AT 493513 T	15-01-2011
		EP 1910560 A1	16-04-2008
		US 2009181370 A1	16-07-2009
		US 2012053074 A1	01-03-2012
		US 2013171648 A1	04-07-2013
		US 2015203911 A1	23-07-2015
		US 2016160277 A1	09-06-2016
		US 2017327885 A1	16-11-2017
		US 2020216896 A1	09-07-2020
		US 2023193384 A1	22-06-2023
		WO 2007010252 A1	25-01-2007
WO 9844151 A1	08-10-1998	AT 364718 T	15-07-2007
		AT 545710 T	15-03-2012
		AU 6846698 A	22-10-1998
		DE 69837913 T2	07-02-2008
		EP 0972081 A1	19-01-2000
		EP 1591541 A2	02-11-2005
		EP 2327797 A1	01-06-2011
		EP 3034626 A1	22-06-2016
		ES 2563643 T3	15-03-2016
		HK 1155784 A1	25-05-2012
		JP 2002503954 A	05-02-2002
		US 2005100900 A1	12-05-2005
		US 2008286795 A1	20-11-2008
		US 2011045541 A1	24-02-2011
		US 2013217586 A1	22-08-2013
		US 2013231254 A1	05-09-2013
		US 2014371100 A1	18-12-2014
		US 2015087531 A1	26-03-2015
		US 2015133320 A1	14-05-2015
		WO 9844151 A1	08-10-1998