

METHODS OF BASE CALLING NUCLEOBASES

BACKGROUND

5 Field

The disclosed technology relates to the field of nucleic acid sequencing. More particularly, the disclosed technology relates to using next generation sequencing to determine the nucleotide sequences of two or more polynucleotide sequence portions in
10 a single sequencing run.

Description of the Related Art

In some types of next-generation sequencing (NGS) technologies, a nucleic acid cluster
15 is created on a flow cell by amplifying an original template nucleic acid strand. Sequencing cycles may be performed as complementary strands of the template nucleic acids are being synthesized, i.e., using sequencing-by-synthesis (SBS) processes.

In each sequencing cycle, deoxyribonucleic acid analogs conjugated to fluorescent
20 labels are hybridized to the template nucleic acids, and excitation light sources are used to excite the fluorescent labels on the deoxyribonucleic acid analogs. Detectors capture fluorescent emissions from the fluorescent labels and identify the deoxyribonucleic acid analogs. As a result, the sequence of the template nucleic acids may be determined by repeatedly performing such sequencing cycles.

25 NGS allows for the sequencing of a number of different template nucleic acids simultaneously, significantly reducing the cost of sequencing in the last twenty years, however there remains a desire for further increases in sequencing throughput.

30 SUMMARY

According to a first aspect of the present invention, there is provided a method of base calling nucleobases of two or more polynucleotide sequence portions, the method comprising:

35 (a) obtaining first intensity data comprising a combined intensity of a first signal obtained based upon a respective first nucleobase of at least one first

polynucleotide sequence portion and a second signal obtained based upon a respective second nucleobase of at least one second polynucleotide sequence portion;

(b) obtaining second intensity data comprising a combined intensity of a third signal obtained based upon the respective first nucleobase of the at least one first polynucleotide sequence portion and a fourth signal obtained based upon the respective second nucleobase of the at least one second polynucleotide sequence portion;

(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,

wherein said polynucleotide sequence portions have been selectively processed such that an intensity of the signals obtained based upon the respective first nucleobase is greater than an intensity of the signals obtained based upon the respective second nucleobase.

In embodiments, the first and second signals and/or the third and fourth signals may be obtained substantially simultaneously.

In embodiments, 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 signals and the combined intensity of the third and fourth signals.

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

In embodiments, the polynucleotide sequence portions may have been selectively processed such that, during sequencing, a greater number of the first polynucleotide sequence portions are capable of generating a signal than a number of the second polynucleotide sequence portions that are capable of generating a signal.

In embodiments, a ratio between the number of the first polynucleotide sequence portions capable of generating a signal and the number of the second polynucleotide sequence portions capable of generating a signal may be between 1.25:1 to 5:1, between 1.5:1 to 3:1, or about 2:1.

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

In embodiments, the obtained signals are generated by:

- 5 contacting a plurality of polynucleotide molecules comprising the first and second polynucleotide sequence portions with first primers for sequencing the first polynucleotide sequence portions and second primers for sequencing the second polynucleotide sequence portions;
- 10 extending the first primers and the second primers by contacting the polynucleotide molecules with labeled nucleobases to form first labeled primers and second labeled primers;
- stimulating the light emissions from the first and second labeled primers; and
- detecting the light emissions at a sensor.

15 In embodiments:

 the first and second signals may be based on light emissions detected in a first range of optical frequencies;

 the third and fourth signals may be based on light emissions detected in a second range of optical frequencies; and

20 the first range of optical frequencies and the second range of optical frequencies may be not identical.

For example, the first range of optical frequencies may correspond to the color red, e.g., 400–484 THz (or equivalently, 620–750 nm in terms of wavelength), and the second range of optical frequencies may correspond to the color green, e.g., 526–606 THz (or

25 equivalently, 495–570 nm in terms of wavelength).

In embodiments, the plurality of polynucleotide molecules may be attached to a substrate, and the light emissions from the first labeled primers and the light emissions from the second labeled primers may be emitted from the same region or substantially

30 overlapping regions of the substrate.

In embodiments, the light emissions detected at the sensor may be spatially unresolved.

In embodiments, the sensor may be configured to provide a single output based upon

35 the first and second signals.

In embodiments, the sensor may comprise a single sensing element.

In embodiments, each polynucleotide molecule may comprise one or more copies of the first polynucleotide sequence portion and one or more copies of the second polynucleotide sequence portion.

5

In embodiments, the first and second polynucleotide sequence portions may be respective portions of different polynucleotide molecules.

10

In embodiments, the polynucleotide sequence portions may have been selectively processed by contacting the plurality of polynucleotide molecules with unblocked first primers and a predetermined fraction of second primers which have a blocked 3' end.

15

In embodiments, selectively processing may comprise preparing for selective sequencing or conducting selective sequencing. For example, selective sequencing may be achieved using a mixture of unblocked and blocked sequencing primers.

20

In embodiments, the polynucleotide sequence portions may have been selectively processed to provide a greater total number of the first polynucleotide sequence portions than a total number of the second polynucleotide sequence portions.

25

In embodiments, selectively processing may comprise conducting selective amplification.

In embodiments, the at least one first polynucleotide sequence portion and the at least one second polynucleotide sequence portion may be present in a cluster.

30

In embodiments, the one of the plurality of classifications may be selected based on the first and the second intensity data using a Gaussian mixture model.

In embodiments, the method may further comprise, based on the selected classification, determining that the second polynucleotide sequence portion is modified relative to the first polynucleotide sequence portion, at a location associated with the first and second nucleobases.

35

Said modification may have been made to any of the first polynucleotide sequence portion, the second polynucleotide sequence portion, or any sequence from which either

of the first and second portions are derived, provided that it results in the modification of the sequences of the first and second portions relative to one another.

5 In embodiments, the second polynucleotide sequence portion may be modified relative to the first polynucleotide sequence portion resulting from a library preparation and/or sequencing error. The method may, therefore, further comprise: based on the selected classification, determining that a library preparation and/or sequencing errors has occurred.

10 In embodiments, the second polynucleotide sequence portion may be modified relative to the first polynucleotide sequence portion resulting from conversion of a modified cytosine to thymine or a nucleobase which is read as thymine/uracil, and/or of an unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil. The method may, therefore, further comprise: based on the selected classification,
15 determining the identity and the methylation status of a nucleobase at a corresponding position of a polynucleotide molecule from which the first and second polynucleotide sequence portions are derived.

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

According to a second aspect of the present invention, there is provided a method of base calling nucleobases of n polynucleotide sequence portions, the method comprising:

25 (a) obtaining first intensity data comprising a combined intensity of respective first signal components generated by each of the n th portions obtained based upon respective n th nucleobases in each of the n portions, wherein the respective first signal components are obtained simultaneously;

30 (b) obtaining second intensity data comprising a combined intensity of respective second signal components generated by each of the n th portions obtained based upon respective n th nucleobases in each of the n portions, wherein the respective second 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 n th nucleobases; and

35 (d) based on the selected classification, base calling the respective n th nucleobases for all n portions,

wherein said polynucleotide sequence portions have been selectively processed such that an intensity of the signals obtained based upon the respective first nucleobase is greater than an intensity of the signals obtained based upon the respective second nucleobase.

5

In embodiments, selecting the classification based on the first and second intensity data may comprise selecting the classification based on the combined intensity of respective first signal components and second signal components.

10

In embodiments, the plurality of classifications may comprise 4^n classifications, each classification representing one of 4^n unique combinations of n th nucleobases.

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

15

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

In embodiments, the data processing device may be a polynucleotide sequencer.

20

According to a fourth 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 above.

25

According to a fifth 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 above.

30

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

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

35

In one embodiment, the disclosed technology provides systems and methods for determining the nucleobase sequences of two or more polynucleotide sequence portions

in parallel (i.e., substantially simultaneously, using the same sequencing run), while the two sequence portions are co-localized within the same nucleic acid cluster. Thus, in some embodiments, the sequencing yield per flow cell area may be doubled as compared to prior systems which performed the sequencing serially.

5

In one embodiment, the disclosed method can include providing a substrate comprising a plurality of single or double stranded polynucleotide molecules in a cluster. The disclosed method can further include contacting the plurality of polynucleotide molecules with first primers for sequencing a first polynucleotide sequence portion and second primers for sequencing a second polynucleotide sequence portion. The first and second polynucleotide sequence portions may be present as respective portions of the same polynucleotide molecules. Alternatively, the first and second polynucleotide sequence portions may be present in different polynucleotide molecules of the plurality of polynucleotide molecules. The disclosed method can further include extending the first primers and the second primers by contacting the cluster with labeled nucleobases to form first labeled primers and second labeled primers. The disclosed method can further include stimulating light emissions from the first and second labelled primers, wherein an amplitude of the signal generated by the first labeled primers is greater than an amplitude of the signal generated by the second labeled primers (or vice versa). The disclosed method can further include identifying the labeled nucleobases added to the first primers and the second primers based on the amplitude of the signal generated by the labelled nucleobases. In some embodiments, the first primers are index primers that hybridize to a site adjacent to a barcode index portion associated with the first sequence portion. In some embodiments, the second primers are index primers that hybridize to a site adjacent to a barcode index portion associated with the second sequence portion. In some embodiments, the first primers are index primers that hybridize to a site adjacent to a barcode index portion associated with the first sequence portion, and the second primers are index primers that hybridize to a site adjacent to a barcode index portion associated with the second sequence portion.

30

In some embodiments, identifying the labeled nucleobases added to the first primers and identifying the labeled nucleobases added to the second primers are performed substantially simultaneously. In some embodiments, the signal generated by the first labelled primers and the signal generated by the second labeled primers are emitted from the same region or substantially overlapping regions of the substrate. In some embodiments, each polynucleotide molecule is attached to the substrate. In some embodiments, the plurality of polynucleotide molecules in the cluster are generated by a

35

bridge amplification process, an exclusion amplification process, a rolling circle amplification process, or any other suitable amplification process. In some embodiments, the substrate comprises a plurality of clusters of nucleic acids, the clusters being randomly distributed on the substrate. In alternative embodiments, the clusters are
5 arranged in a patterned array.

In some embodiments, the amplitude of the signal generated by the first labeled primers corresponds with a first quantity of the first labeled primers in the cluster, and the amplitude of the signal generated by the second labeled primers corresponds with a
10 second quantity of the second labeled primers in the cluster. In some embodiments, contacting the plurality of polynucleotide molecules with first primers for sequencing the first sequence portion and second primers for sequencing the second sequence portion comprises contacting the molecules with unblocked first primers and a predetermined fraction of second primers which have a blocked 3'-end. The blocked 3'-end may be
15 formed by any way of blocking the ability of a primer to extend a nucleic acid strand, for example by modifications in the sugar or nucleobase. In some embodiments, the blocked 3'-end comprises a hairpin loop, a deoxynucleotide, a phosphate group, a propyl spacer, a modification blocking the 3'-hydroxyl group, or an inverted nucleobase. In some embodiments, the first primers are formed of a locked nucleic acid or a peptide nucleic
20 acid. In some embodiments, the second primers are formed of a locked nucleic acid or a peptide nucleic acid.

In some embodiments, the disclosed method further includes: detecting the signal generated by the first labeled primers in a first range of optical frequencies and a second
25 range of optical frequencies; and detecting the signal generated by the second labeled primers in the first range of optical frequencies and the second range of optical frequencies, wherein the first range of optical frequencies and the second range of optical frequencies are not identical. For example, the first range of optical frequencies may correspond to the color red, e.g., 400–484 THz (or equivalently, 620–750 nm in terms of wavelength), and the second range of optical frequencies may correspond to
30 the color green, e.g., 526–606 THz (or equivalently, 495–570 nm in terms of wavelength).

In some embodiments, the disclosed method further includes: acquiring a first fluorescent image of the cluster in a first range of optical frequencies; acquiring a second
35 fluorescent image of the cluster in a second range of optical frequencies, wherein the first range of optical frequencies and the second range of optical frequencies are not identical; and obtaining the signals generated by the first and second labeled primers by

extracting fluorescence intensities from the first and second fluorescent images of the cluster. In some examples, the first range of optical frequencies and the second range of optical frequencies may partially overlap. For example, the first range of optical frequencies may be 500–580 THz, and the second range of optical frequencies may be 540–620 THz.

In some embodiments, the disclosed method further includes extracting fluorescence intensities from the first and second fluorescent images of the same region or substantially overlapping regions of the substrate. In some embodiments, identifying the labeled nucleobases added to the first primers and the second primers is based on a combination of the extracted fluorescence intensities from the first and second fluorescent images. In some embodiments, a combination of identities of the labeled nucleobases added to the first primers and the second primers is classified as one of sixteen combinations of types of nucleobases, based on the combination of the extracted fluorescence intensities and predetermined fluorescence intensity distributions for the sixteen combinations of types of nucleobases. In some embodiments, the disclosed method further includes: normalizing the extracted fluorescence intensities; and classifying a combination of identities of the labeled nucleobases added to the first primers and the second primers as one of sixteen combinations of types of nucleobases, based on a combination of the normalized extracted fluorescence intensities and predetermined normalized fluorescence intensity distributions for the sixteen combinations of types of nucleobases.

In some embodiments, the disclosed method further includes stimulating fluorescent emissions from the first labeled primers and second labeled primers in the cluster with light at a predetermined optical frequency. In some embodiments, the disclosed method further includes stimulating fluorescent emissions from the first labeled primers and second labeled primers in the cluster with light at two predetermined optical frequencies. In some embodiments, the disclosed method further includes identifying whether the labeled nucleobases are associated with the first sequence portion or the second sequence portion based on the amplitude of the signal generated by the labeled nucleobases.

The systems, devices, kits, and methods disclosed herein each have several aspects, no single one of which is solely responsible for their desirable attributes. Numerous other embodiments are also contemplated, including embodiments that have fewer, additional, and/or different components, steps, features, objects, benefits, and advantages. The

components, aspects, and steps may also be arranged and ordered differently. After considering this discussion, and particularly after reading the section entitled “Detailed Description”, one will understand how the features of the devices and methods disclosed herein provide advantages over other known devices and methods.

5

It is to be understood that any features of the systems disclosed herein may be combined together in any desirable manner and/or configuration. Further, it is to be understood that any features of the methods disclosed herein may be combined together in any desirable manner. Moreover, it is to be understood that any combination of features of the methods and/or the systems may be used together, and/or may be combined with any of the examples disclosed herein. It should be appreciated that all combinations of the foregoing concepts and additional concepts discussed in greater detail below are contemplated as being part of the inventive subject matter disclosed herein and may be used to achieve the benefits and advantages described herein.

10
15

BRIEF 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.

20

Figure 1 shows a block diagram which schematically illustrates an example sequencing system that may be used to perform the disclosed methods.

25

Figure 2 shows a block diagram which schematically illustrates an example imaging system that may be used in conjunction with the example sequencing system of Figure 1.

30

Figure 3 shows a functional block diagram of an example computer system that may be used in the example sequencing system of Figure 1.

35

Figure 4A and Figure 4B schematically illustrate nucleic acid clusters comprising two or more polynucleotide sequence portions for sequencing by the present methods.

Figure 5A and Figure 5B are charts which shows example dye labeling schemes that may be used in conjunction with the present methods.

5 Figure 6 is a plot showing graphical representations of sixteen distributions of signals from a nucleic acid cluster according to one embodiment of the disclosed technology.

Figure 7 schematically illustrates how library preparation errors can obscure true variants in NGS methods.

10 Figure 8 schematically illustrates the use of unique molecular indices (UMIs) for eliminating library preparation errors.

Figure 9 shows schematically how library preparation errors and true variants can be associated with different distributions of combined signal intensities.

15 Figure 10 shows the effect of methylcytosine (^mC) to thymine conversion treatment of a top strand of a tandem polynucleotide molecule, and a scatter plot showing the resulting distributions of signals from a nucleic acid cluster generated from the treated molecule.

20 Figure 11 shows the effect of cytosine to uracil conversion treatment of top and bottom strands of a tandem polynucleotide molecule, and a scatter plot showing the resulting distributions of signals from a nucleic acid cluster generated from the treated molecule.

25 Figure 12 is a diagram showing a method for classifying combined signal intensities into one of a plurality of distributions.

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

30 Figure 14 is a flow diagram showing a method for generating signals for use in the method of base calling shown in Figure 13.

Figure 15 shows a forward strand, reverse strand, forward complement strand, and reverse complement strand of a polynucleotide molecule.

35 Figure 16 shows an example of a polynucleotide sequence (or insert) with 5' and 3' adaptor sequences.

Figure 17 shows a typical polynucleotide with 5' and 3' adaptor sequences.

Figure 18 shows an example of PCR stitching. Here, two sequences – a strand of a human library and a strand of a phiX library are joined together to create a single polynucleotide strand comprising both a first portion (comprising the strand of the human sequence) and a second portion (comprising the strand of the phiX sequence), as well as terminal and internal adaptor sequences.

Figure 19 shows the preparation of a concatenated polynucleotide sequence comprising a first portion and a second portion using a tandem insert method.

Figure 20A shows the effect of pre-treatment of library strands using C to U conversion on bases in template strands. Figure 20B shows the effect of pre-treatment of library strands using mC to T conversion on bases in template strands.

Figure 21 shows the preparation of a concatenated polynucleotide sequence comprising a first portion and a second portion using a loop fork method.

Figure 22 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 23 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 24 shows a typical solid support.

Figure 25 shows the stages of bridge amplification and the generation of an amplified cluster.

Figure 26 shows the stages of bridge amplification for concatenated polynucleotide templates and the generation of an amplified cluster.

Figure 27 shows a method of selective sequencing.

Figure 28 shows a method of selective amplification.

Figure 29 shows a method of selective amplification.

Figure 30 shows a method of selective amplification.

Figure 31 shows (A) that by plotting relative intensities of light signals obtained from a first channel (ch1) and a second channel (ch2), a constellation of 16 clouds is obtained;

and (B) alignment of R1 and R2 (minor and major reads respectively) with the known human and PhiX sequence.

Figure 32 shows that by plotting relative intensities of light signals obtained from a first channel (ch1) and a second channel (ch2), a constellation of 16 clouds is obtained.

5 Figure 33 shows (A) that by plotting relative intensities of light signals obtained from a first channel (ch1) and a second channel (ch2), a constellation of 16 clouds is obtained for R1 and R2 concurrently and R3 and R4 concurrently; (B) alignment of R1, R2, R3 and R4 with the known sequence; and (C) annotation of where R1, R2, R3 and R4 appear on the known sequence.

10

DETAILED DESCRIPTION

All patents, patent applications, and other publications, including all sequences disclosed within these references, referred to herein 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.

20

Introduction

In one aspect, the disclosed technology provides systems and methods that can dramatically shorten the total sequencing time and reduce the number of reagents used in next generation sequencing workflows. In addition, sequencing yield per flow cell area may be increased. In some embodiments, the disclosed method enables simultaneous sequencing of two or more polynucleotide sequence portions without the need for the signals generated from the different portions to be separately detectable given the configuration of the portions and the sequencing equipment. It is therefore possible to simultaneously sequence multiple polynucleotide sequence portions that are not possible to spatially resolve, for example sequencing two polynucleotide sequence portions from a signal detected at a single sensing region (for example a single pixel of an imaging sensor) and/or from a signal obtained from a single cluster (i.e. a single contiguous cluster containing both of the two or more sequence portions), thus

35

increasing the efficiency of the sequencing workflow. Typically, sequencing data from clusters comprising more than one polynucleotide sequence portion of interest ("polyclonal clusters") are filtered out and are excluded from the sequencing output. Therefore, the present methods may also allow for an increase in the number of usable clusters in a given area of a substrate.

In some embodiments, the primer for sequencing a first sequence portion and the primer for sequencing a second sequence portion are annealed/hybridized to the molecules in the same reaction step to reduce chemical reaction steps, thus saving time and increasing the efficiency of sequencing-by-synthesis (SBS) workflows. Then, both sequence portions may be read-out through SBS chemistry cycles in the same reaction run.

In some embodiments, in order to separate the signals received from the dye-labeled nucleobases hybridized to each sequence portion, the signal from one of the portions is diminished, e.g., by 50%, in comparison to the signal generated by the other portion. In one example, the difference in signal intensity may be achieved by blocking the addition of labeled nucleobases to some of the primers. For example, half of the primers which bind to a first portion may be blocked so that no fluorescent nucleotides can be added during the sequencing reactions. Thus, the overall intensity of the nucleobases added to the first portion will be 50% lower than the intensity of the nucleobases added to a second portion in this example. By reviewing not only the wavelength of light emitted from the dyes from each nucleic acid cluster on the flow cell, but also the intensity of that light, the labeled nucleobase hybridized to the first portion can be distinguished from the labeled nucleobase hybridized to the second portion. This will be discussed more completely in the sections below. In another example, the difference in signal intensity may be achieved by selectively controlling the number of copies of a first sequence portion relative to the number of copies of a second sequence portion. For example, the number of copies of the second sequence portion may be lower, e.g. by 50%, in comparison to the number of copies of the first sequence portion.

In some embodiments, the disclosed technology comprises obtaining sequence information using Illumina's sequencing-by-synthesis and reversible terminator-based sequencing chemistry with removable fluorescent dyes (e.g., as described in Bentley et al., Nature 6:53-59 [2009]). Short sequence reads of about tens to a few hundred base pairs may be aligned against a reference genome and unique mapping of the short sequence reads to the reference genome may be identified. Further details regarding

the sequencing-by-synthesis and dye labeling methods which can be used by the disclosed technology are described in U.S. Patent Application Publication Numbers 2007/0166705, 2006/0188901, 2006/0240439, 2006/0281109, 2005/0100900, 2013/0079232, U.S. Patent Number 7,057,026, PCT Application Publication Numbers
5 WO 2005/065814, WO 2006/064199, WO 2007/010251, and WO 2018/165099, U.S. Patent Application Number 17/338590, U.S. Patent No. 7,601,499, U.S. Patent No. 9,267,173, and U.S. Patent Publication No. 2012/0053063, the disclosures of which are incorporated herein by reference in their entireties.

10 Example Sequencer

Referring to Figure 1, a diagrammatical representation of an example sequencing system 10 is illustrated as including a sequencer 12 designed to determine sequences of genetic material of a sample 14. The sequencer may function in a variety of manners, and based
15 upon a variety of techniques, including sequencing by primer extension using labeled nucleotides, as in a presently contemplated embodiment, as well as other sequencing techniques such as sequencing by ligation or pyrosequencing. In some embodiments, the sequencer 12 progressively moves samples through reaction cycles and imaging cycles to progressively build oligonucleotides by binding nucleotides to templates at
20 individual sites on the sample. In some embodiments, the sample may be prepared by a sample preparation system 16. This process may include amplification of fragments of DNA or RNA on a support to create a multitude of sites of DNA or RNA fragments the sequence of which are determined by the sequencing process. Exemplary methods for producing sites of amplified nucleic acids suitable for sequencing include, but are not
25 limited to, rolling circle amplification (RCA) (Lizardi et al., Nat. Genet. 19:225-232 (1998)), bridge PCR (Adams and Kron, Method for Performing Amplification of Nucleic Acid with Two Primers Bound to a Single Solid Support, Mosaic Technologies, Inc. (Winter Hill, Mass.); Whitehead Institute for Biomedical Research, Cambridge, Mass., (1997); Adessi et al., Nucl. Acids Res. 28:E87 (2000); Pemov et al., Nucl. Acids Res. 33:e11 (2005); or U.S. Pat. No. 5,641,658), polony generation (Mitra et al., Proc. Natl. Acad. Sci. USA 100:5926-5931 (2003); Mitra et al., Anal. Biochem. 320:55-65 (2003)),
30 or clonal amplification on beads using emulsions (Dressman et al., Proc. Natl. Acad. Sci. USA 100:8817-8822 (2003)) or ligation to bead-based adapter libraries (Brenner et al., Nat. Biotechnol. 18:630-634 (2000); Brenner et al., Proc. Natl. Acad. Sci. USA 97:1665-1670 (2000)); Reinartz, et al., Brief Funct. Genomic Proteomic 1:95-104 (2002)), each of the aforementioned publications is incorporated herein by reference. The sample

preparation system 16 may dispose the sample, which may be in the form of an array of sites, in a sample container for processing and imaging.

5 In some embodiments, the sequencer 12 includes a fluidics control/delivery system 18 and a detection system 20. The fluidics control/delivery system 18 may receive a plurality of process fluids as indicated by reference numeral 22, for circulation through the sample containers of the samples in process, designated by reference numeral 24. As will be appreciated by those skilled in the art, the process fluids may vary depending upon the particular stage of sequencing. For example, in sequencing-by-synthesis (SBS) using
10 labeled nucleotides, the process fluids introduced to the sample may include a polymerase and tagged nucleotides of the four common DNA types, each nucleotide having a unique fluorescent tag and a blocking agent linked to it. The fluorescent tag allows the detection system 20 to detect which nucleotides were last added to primers hybridized to template nucleic acids at individual sites in the array, and the blocking agent
15 prevents addition of more than one nucleotide per cycle at each site.

At other phases of the sequencing cycles, the process fluids 22 may include other fluids and reagents, such as reagents for removing extension blocks from nucleotides or cleaving nucleotide linkers to release a newly extendable primer terminus. For example,
20 once reactions have taken place at individual sites in the array of the samples, the initial process fluid containing the tagged nucleotides may be washed from the sample in one or more flushing operations. The sample may then undergo detection, such as by the optical imaging at the detection system 20. Subsequently, reagents may be added by the fluidics control/delivery system 18 to de-block the last added nucleotide and remove
25 the fluorescent tag from each. The fluidics control/delivery system 18 may then again wash the sample, which is then prepared for a subsequent cycle of sequencing. Exemplary fluidic and detection configurations that can be used in the methods and devices set forth herein are described in WO 07/123744, which is incorporated herein by reference. In some embodiments, such sequencing may continue until the quality of data
30 derived from sequencing degrades due to cumulative loss of yield or until a predetermined number of cycles have been completed.

In some embodiments, the quality of samples 24 in process as well as the quality of the data derived by the system, and the various parameters used for processing the samples
35 is controlled by a quality/process control system 26. The quality/process control system 26 may include one or more programmed processors, or general purpose or application-specific computers which communicate with sensors and other processing systems

within the fluidics control/delivery system 18 and the detection system 20. A number of process parameters may be used for sophisticated quality and process control, for example, as part of a feedback loop that can change instrument operation parameters during the course of a sequencing run.

5

In some embodiments, the sequencer 12 also communicates with a system control/operator interface 28 and ultimately with a post-processing system 30. The system control/operator interface 28 may include a general purpose or application-specific computer designed to monitor process parameters, acquired data, system settings, and so forth. The operator interface may be generated by a program executed locally or by programs executed within the sequencer 12. In some embodiments, these may provide visual indications of the health of the systems or subsystems of the sequencer, the quality of the data acquired, and so forth. The system control/operator interface 28 may also permit human operators to interface with the system to regulate operation, initiate and interrupt sequencing, and any other interactions that may be desired with the system hardware or software. For instance, the system control/operator interface 28 may automatically undertake and/or modify steps to be performed in a sequencing procedure, without input from a human operator. Alternatively or additionally, the system control/operator interface 28 may generate recommendations regarding steps to be performed in a sequencing procedure and display these recommendations to the human operator. This mode may allow for input from the human operator before undertaking and/or modifying steps in the sequencing procedure. In addition, the system control/operator interface 28 may provide an option to the human operator allowing the human operator to select certain steps in a sequencing procedure to be automatically performed by the sequencer 12 while requiring input from the human operator before undertaking and/or modifying other steps. In any event, allowing both automated and operator interactive modes may provide increased flexibility in performing the sequencing procedure. In addition, the combination of automation and human-controlled interaction may further allow for a system capable of creating and modifying new sequencing procedures and algorithms through adaptive machine learning based on the inputs gathered from human operators.

The post-processing system 30 may further include one or more programmed computers that receive detected information, which may be in the form of pixilated image data and derive sequence data from the image data. The post-processing system 30 may include image recognition algorithms which distinguish between colors of dyes (e.g., fluorescent emission spectra of dyes) attached to nucleotides that bind at individual sites as

35

sequencing progresses (e.g., by analysis of the image data encoding specific colors and/or intensities), and logs the sequence of the nucleotides at the individual site locations. Progressively, then, the post-processing system 30 may build sequence lists for the individual sites of the sample array which can be further processed to establish
5 genetic information for extended lengths of material by various bioinformatics algorithms.

The sequencing system 10 may be configured to handle individual samples or may be designed for higher throughput in a manner in which multiple stations are provided for the delivery of reagents and other fluids, and for detection of progressively building
10 sequences of nucleotides. Further details can be found in U.S. Patent No. 9797012, which is incorporated herein by reference.

Samples may be removed from processing, reprocessed, and scheduling of such processing may be altered in real time, particularly where the fluidics control system 18
15 or the quality/process control system 26 detect that one or more operations were not performed in an optimal or desired manner. In embodiments wherein a sample is removed from the process or experiences a pause in processing that is of a substantial duration, the sample can be placed in a storage state. Placing the sample in a storage state can include altering the environment of the sample or the composition of the sample
20 to stabilize biomolecule reagents, biopolymers or other components of the sample. Exemplary methods for altering the sample environment include, but are not limited to, reducing temperature to stabilize sample constituents, addition of an inert gas to reduce oxidation of sample constituents, and removing from a light source to reduce photobleaching or photodegradation of sample constituents. Exemplary methods of
25 altering sample composition include, without limitation, adding stabilizing solvents such as antioxidants, glycerol and the like, altering pH to a level that stabilizes enzymes, or removing constituents that degrade or alter other constituents. In addition, certain steps in the sequencing procedure may be performed before removing the sample from processing. For instance, if it is determined that the sample should be removed from
30 processing, the sample may be directed to the fluidics control/delivery system 18 so that the sample may be washed before storage. These steps may be taken to ensure that no information from the sample is lost.

Moreover, sequencing operations may be interrupted by the sequencer 12 at any time
35 upon the occurrence of certain predetermined events. These events may include, without limitation, unacceptable environmental factors such as undesirable temperature, humidity, vibrations or stray light; inadequate reagent delivery or hybridization;

unacceptable changes in sample temperature; unacceptable sample site number/quality/distribution; decayed signal-to-noise ratio; insufficient image data; and so forth. It should be noted that the occurrence of such events need not require interruption of sequencing operations. Rather, such events may be factors weighed by the quality/process control system 26 in determining whether sequencing operations should continue. For example, if an image of a particular cycle is analyzed in real time and shows a low signal for that optical channel, the image can be re-exposed using a longer exposure time, or have a particular chemical treatment repeated. If the image shows a bubble in a flow cell, the instrument can automatically flush more reagent to remove the bubble, then re-record the image. If the image shows low signal for a particular optical channel in one cycle due to a fluidics problem, the instrument can automatically halt scanning and reagent delivery for that particular optical channel, thus saving on analysis time and reagent consumption.

Although the system has been exemplified above with regard to a system in which a sample interfaces with different stations by physical movement of the sample, it will be understood that the principles set forth herein are also applicable to a system in which the steps occurring at each station are achieved by other means not requiring movement of the sample. For example, reagents present at the stations can be delivered to a sample by means of a fluidic system connected to reservoirs containing the various reagents. Similarly, an optics system can be configured to detect a sample that is in fluid communication with one or more reagent stations. Thus, detection steps can be carried out before, during or after delivery of any particular reagent described herein. Accordingly, samples can be effectively removed from processing by discontinuing one or more processing steps, be it fluid delivery or optical detection, without necessarily physically removing the sample from its location in the device.

Disclosed systems can be used to continuously sequence nucleic acids in a plurality of different samples. Disclosed systems can be configured to include an arrangement of samples and an arrangement of stations for carrying out sequencing steps. The samples in the arrangement of samples can be placed in a fixed order and at fixed intervals relative to each other. For example, an arrangement of nucleic acid arrays can be placed along the outer edge of a circular table. Similarly, the stations can be placed in a fixed order and at fixed intervals relative to each other. For example, the stations can be placed in a circular arrangement having a perimeter that corresponds to the layout for the arrangement of sample arrays. Each of the stations can be configured to carry out a different manipulation in a sequencing protocol. The arrangements of sample arrays and

stations can be moved relative to each other such that the stations carry out desired steps of a reaction scheme at each reaction site. The relative locations of the stations and the schedule for the relative movement can correlate with the order and duration of reaction steps in the sequencing reaction scheme such that once a sample array has completed a cycle of interacting with the full set of stations, then a single sequencing reaction cycle is complete. For example, primers that are hybridized to nucleic acid targets on an array can each be extended by addition of a single nucleotide, detected and de-blocked if the order of the stations, spacing between the stations, and rate of passage for the array corresponds to the order of reagent delivery and reaction time for a complete sequencing reaction cycle.

In accordance with the configuration set forth above, each lap (or full revolution in embodiments where a circular table is used) completed by an individual sample array can correspond to determination of a single nucleotide for each of the target nucleic acids on the array (e.g., including the steps of incorporation, imaging, cleavage and de-blocking carried out in each cycle of a sequencing run). Furthermore, several sample arrays present in the system (for example, on the circular table) concurrently move along similar, repeated laps through the system, thereby resulting in continuous sequencing by the system. Using the disclosed systems or methods, reagents can be actively delivered or removed from a first sample array in accordance with a first reaction step of a sequencing cycle while incubation, or some other reaction step in the cycle, occurs for a second sample array. Thus, a set of stations can be configured in a spatial and temporal relationship with an arrangement of sample arrays such that reactions occur at multiple sample arrays concurrently even as the sample arrays are subjected to different steps of the sequencing cycle at any given time, thereby allowing continuous and simultaneous sequencing to be performed. Such a circular system may be used when the chemistry and imaging times are disproportionate. For small flow cells that only take a short time to scan, the system may have a number of flow cells running in parallel in order to optimize the time the instrument spends acquiring data. When the imaging time and chemistry time are equal, a system that is sequencing a sample on a single flow cell spends half the time performing a chemistry cycle rather than an imaging cycle, and therefore a system that can process two flow cells could have one on the chemistry cycle and one on the imaging cycle. When the imaging time is ten-fold less than the chemistry time, the system can have ten flow cells at various stages of the chemistry process whilst continually acquiring data.

In some embodiments, the disclosed system is configured to allow replacement of a first sample array with a second sample array while the system continuously sequences nucleic acids of a third sample array. Thus, a first sample array can be individually added or removed from the system without interrupting sequencing reactions occurring at another sample array, thereby allowing continuous sequencing for the set of sample arrays. Moreover, sequencing runs of different lengths can be performed continuously and simultaneously in the system because individual sample arrays can complete a different number of laps through the system and the sample arrays can be removed or added to the system in an independent fashion such that reactions occurring at other sites are not perturbed.

Figure 2 illustrates an exemplary detection station 38 which can detect nucleotides added at sites of an array and can be used in conjunction with the example sequencing system of Figure 1. As set forth above, a sample can be moved to two or more stations of the device that are located in physically different locations or alternatively one or more steps can be carried out on a sample that is in communication with the one or more stations without necessarily being moved to different locations. Accordingly, the description herein with regard to particular stations is understood to relate to stations in a variety of configurations whether or not the sample moves between stations, the stations move to the sample, or the stations and sample are static with respect to each other. In the embodiment illustrated in Figure 2, one or more light sources 46 provide light beams that are directed to conditioning optics 48. The light sources 46 may include one or more lasers, with multiple lasers being used for detecting dyes that fluoresce at different corresponding wavelengths. The light sources may direct beams to the conditioning optics 48 for filtering and shaping of the beams in the conditioning optics. For example, in a presently contemplated embodiment, the conditioning optics 48 combine beams from multiple lasers and generate a substantially linear beam of radiation that is conveyed to focusing optics 50. The laser modules can additionally include a measuring component that records the power of each laser. The measurement of power may be used as a feedback mechanism to control the length of time an image is recorded in order to obtain a uniform exposure energy, and therefore signal, for each image. If the measuring component detects a failure of the laser module, then the instrument can flush the sample with a "holding buffer" to preserve the sample until the error in the laser can be corrected.

The sample 24 is positioned on a sample positioning system 52 that may appropriately position the sample in three dimensions, and may displace the sample for progressive

imaging of sites on the sample array. In a presently contemplated embodiment, the focusing optics 50 confocally direct radiation to one or more surfaces of the array at which individual sites are located that are to be sequenced. Depending upon the wavelengths of light in the focused beam, a retrobeam of radiation is returned from the sample due to fluorescence of dyes bound to the nucleotides at each site.

The retrobeam is then returned through retrobeam optics 54 which may filter the beam, such as to separate different wavelengths in the beam, and direct these separated beams to one or more cameras 56. The cameras 56 may be based upon any suitable technology, such as including charge coupled devices that generate pixilated image data based upon photons impacting locations in the devices. In some embodiments, the cameras 56 may include CMOS sensors. In some embodiments, the cameras 56 may include one or more point-and-shoot cameras. In some embodiments, the cameras 56 may include one or more time delay and integration (TDI) cameras. The cameras generate image data that is then forwarded to image processing circuitry 58. In some embodiments, the processing circuitry 58 may perform various operations, such as analog-to-digital conversion, scaling, filtering, and association of the data in multiple frames to appropriately and accurately image multiple sites at specific locations on the sample. The image processing circuitry 58 may store the image data, and may ultimately forward the image data to the post-processing system 30 where sequence data can be derived from the image data. Example detection devices that can be used at a detection station include, for example, those described in US 2007/0114362 (U.S. patent application Ser. No. 11/286,309) and WO 07/123744, each of which is incorporated herein by reference.

A computer system 106 as illustrated in Figure 3 may be used to implement the system control/operator interface 28 and the post-processing system 30 of the example sequencing system 10 in Figure 1. As shown in Figure 3, the computer system 106 can include functionalities for controlling optics/fluidics systems and determining nucleobase sequences of polynucleotides.

In one embodiment, the computer system 106 includes a processor 202 that is in electrical communication with a memory 204, a storage 206, and a communication interface 208. The processor 202 can be configured to execute instructions that cause the fluidics system 104 to supply reagents to the flow cell 114 during sequencing reactions. The processor 202 can execute instructions that control the light source 120 of the optics system 102 to generate light at around a predetermined wavelength. The

processor 202 can execute instructions that control the detector 126 of the optics system 102 and receive data from the detector 126. The processor 202 can execute instructions to process data, for example fluorescent images, received from the detector 126 and to determine the nucleotide sequences of polynucleotides based on the data received from the detector 126. The memory 204 can be configured to store instructions for configuring the processor 202 to perform the functions of the computer system 106 when the sequencing system 100 is powered on. When the sequencing system 100 is powered off, the storage 206 can store the instructions for configuring the processor 202 to perform the functions of the computer system 106. The communication interface 208 can be configured to facilitate the communications between the computer system 106, the optics system 102, and the fluidics system 104.

The computer system 106 can include a user interface 210 configured to communicate with a display device (not shown) for displaying the sequencing results of the sequencing system 100. The user interface 210 can be configured to receive inputs from users of the sequencing system 100. An optics system interface 212 and a fluidics system interface 214 of the computer system 106 can be configured to control the optics system 102 and the fluidics system 104 through communication links (not shown). For example, the optics system interface 212 can communicate with the computer interface 110 of the optics system 102 through a communication link 108a.

The computer system 106 can include a nucleic base determiner 216 configured to determine the nucleotide sequence of polynucleotides using the data received from the detector 126. The nucleic base determiner 216 can include one or more of: a template generator 218, a location registrator 220, an intensity extractor 222, an intensity corrector 224, a base caller 226, and a quality score determiner 228. The template generator 218 can be configured to generate a template of the locations of polynucleotide clusters in the flow cell 114 using the fluorescent images captured by the detector 126. The location registrator 220 can be configured to register the locations of polynucleotide clusters in the flow cell 114 in the fluorescent images captured by the detector 126 based on the location template generated by the template generator 218. The intensity extractor 222 can be configured to extract intensities of the fluorescent emissions from the fluorescent images to generate extracted intensities. For example, the peak intensity value found in a diffraction-limited spot of a DNA cluster may be extracted from the image and used to represent the signal of the DNA cluster. For another example, the total intensity included within a diffraction-limited spot of a DNA cluster may be extracted from the image and

used to represent the signal of the DNA cluster. Alternatively, the intensity estimate can be made through the use of equalization and channel estimation.

The intensity corrector 224 can be configured to reduce or eliminate noise or aberration inherent in the sequencing reaction or optical system. For example, intensity may be influenced by laser intensity fluctuation, DNA cluster shape/size variation, uneven illumination, optical distortions or aberrations, and/or phasing/pre-phasing that occur in the DNA clusters. In some embodiments, the intensity corrector 224 can phase correct or pre-phase correct extracted intensities. In some embodiments, the intensity corrector 224 can normalize extracted fluorescence intensities to reduce or eliminate the effect of DNA cluster size variation. For example, each DNA template may contain the same calibration oligonucleotide. Thus, the extracted fluorescence intensity of a cluster obtained from sequencing a known nucleotide in the calibration oligonucleotide can be used as a normalization factor for that cluster. The intensity corrector 224 can divide the extracted fluorescence intensities of that cluster obtained from sequencing nucleotides in other regions of the DNA template by the normalization factor to obtain the normalized extracted fluorescence intensities. The base caller 226 can be configured to determine the nucleobases of a polynucleotide from the corrected intensities. The bases of a polynucleotide determined by the base caller 226 can be associated with quality scores determined by the quality score determiner 228. Quality scoring refers to the process of assigning a quality score to each base call. To evaluate the quality of a base call from a sequencing read, example processes can include calculating a set of predictor values for the base call and using the predictor values to look up a quality score in a quality table. The quality score can be presented in any suitable format that allows a user to determine the probability of error of any given base call. In some embodiments, the quality score is presented as a numerical value. For example, the quality score can be quoted as QXX where the XX is the score and it means that that particular call has a probability of error of $10^{-XX/10}$. Thus, as an example, Q30 equates to an error rate of 1 in 1000, or 0.1% and Q40 equates to an error rate of 1 in 10,000 or 0.01%. The error rate can be calculated using a control nucleic acid. Additionally, some metrics displays can include the error rate on a per-cycle basis. In some embodiments, the quality table is generated using on a calibration data set, the calibration set being representative of run and sequence variability. Further details of the computations that can be performed by the nucleic base determiner, calculation of error rate and quality score may be found in U.S. Patent Number 8392126, U.S. Patent Application Publication Numbers 2020/0080142 and 2012/0020537, each of which is incorporated by reference herein in its entirety. While nucleic base determiner 216 is shown as part of computer system 106

in Figure 3, it will be appreciated that nucleic base determiner 216 may be a separate computing device from the other components shown in Figure 3 such that nucleic base determiner 216 may receive and process image data in a computing device that is different to a computing device that provides optics and fluidics control.

5

Clusters

Figures 4A-4B each illustrate a respective plurality of polynucleotide molecules 400 comprising multiple copies of two polynucleotide sequence portions of interest 401a, 401b for base calling simultaneously based upon a single combined signal obtained from the two portions according to the present methods. For example, the plurality of polynucleotide molecules 400 illustrated in Figures 4A and 4B may be configured on a substrate 410 such that light emissions from the plurality of polynucleotide molecules are detected by a single sensing portion (for example a single pixel of an imaging sensor 420). Additionally or alternatively, the plurality of polynucleotide molecules 400 may comprise a single cluster (i.e. a single contiguous cluster containing both of the two or more sequence portions 401a, 401b) such that light emissions from each of the two respective portions cannot be spatially resolved. The substrate 410 may be a flow cell, which may be patterned or unpatterned. In one example, the substrate 410 may be a patterned flow cell comprising a number of discrete nanowells 411, with each well containing polynucleotide molecules comprising two or more polynucleotide sequence portions for sequencing and each well having a single respective sensor associated with the well. Because each a single sensor is associated with the well, signals from the two or more portions of interest cannot be resolved, irrespective of whether the different portions (or respective clusters) are spatially resolved within the well. Two or more polynucleotide sequence portions of interest contained within a single well in this way is sometimes referred to herein as a "cluster" irrespective of whether the different portions are spatially resolved in the well given that light emissions from such a well form a single combined signal.

30

In one example, as shown in Figure 4A, the first and second sequence portions 401a, 401b are present in different polynucleotide molecules 400. In the example shown in Figure 4B, the first and second sequence portions 401a, 401b are present as respective portions of the same molecules 400. As described in detail below, by diminishing the signal from one of the portions 401b relative to the other of the portions 401a it is possible to separate the signals received from the dye-labeled nucleobases hybridized to each portion. Figures 4A and 4B each illustrate a respective way in which the signal intensity

35

from the first portion 401a may be modified relative to the signal intensity from the second portion 401b (i.e. in which the signal intensity of one portion may be diminished relative to the signal intensity of the other portion). In the example illustrated in Figure 4A, the number of first portions 401a relative to the number of second portions 401b is uneven, with one second portion for each two first portions). In the example illustrated in Figure 4B, the number of first portions 401a and second portions 401b is the same, however some of the primers 402b used to sequence the second portions are blocked such that during a sequencing run, an uneven number of first portions relative to second portions emit light. While blocking is illustrated with respect to Figure 4B, in which first and second sequence portions 401a, 401b are present as respective portions of the same polynucleotide molecules 400, it will be appreciated that blocking can also be used to diminish the signal from one sequence portion relative to another sequence portion where first and second portions are present in different polynucleotide molecules. Various ways of providing a plurality of polynucleotide molecules 400 comprising multiple copies of two polynucleotide sequence portions of interest 401a, 401b in which a signal intensity of one polynucleotide sequence portion is greater relative to the other polynucleotide sequence portion are outlined below.

Differential signal intensity

Both the first and second sequence portions 401a, 401b in the cluster can be sequenced simultaneously using first primers 402a specific to the first portion 401a, or to a region 403a adjacent to the first portion, and second primers 402b specific to the second portion 401b, or to a region 403b adjacent to the second portion, in the same reaction run. For example, the first and second sequence portions 401a, 401b may be flanked at one or both ends by respective primer binding sites 403a, 403b having a known sequence. Sequencing primers 402a, 402b specific to the different primer binding sites 403a, 403b can therefore be designed and used for simultaneous sequencing of the two sequence portions 401a, 401b.

As described above, a single combined signal may be obtained from the two polynucleotide sequence portions of interest 401a, 401b according to the present methods. For example, the plurality of polynucleotide molecules 400 may be configured on the flow cell 410 such that light emissions from the plurality of polynucleotide molecules are detected by a single sensing portion 420. Alternatively or additionally, the plurality of template polynucleotides may comprise a single cluster such that light

emissions from each of the respective polynucleotide sequence portions cannot be spatially resolved.

5 Since the fluorescent signal associated with the extended first portion sequencing primers 402a and the fluorescent signal associated with the extended second portion sequencing primers 402b is combined, the signals may not be optically resolved. Therefore, methods for determining whether a fluorescent signal is associated with the extended first portion sequencing primers 402a or the extended second portion sequencing primers 402b are needed, at least when the dye-labeled nucleotide analogs at the extended first portion sequencing primers are not the same as the dye-labeled nucleotide analogs at the extended second portion sequencing primers (e.g., when "A"s are added at the first sequence portion 401a and "C"s are added at the second sequence portion 401b), in order to correctly determine the nucleic acid sequences of both the first and second portions.

15 In some embodiments, whether a fluorescent signal is associated with the first sequence portion 401a or the second sequence portion 401b can be determined by using distinguishable levels of signal intensity. In particular, the polynucleotide molecules 400 may be selectively processed such that an intensity of the light emissions associated with respective nucleobases in each of the different sequence portions of interest is different.

25 It will be appreciated that for dye labeling schemes which include an unlabeled or "dark" base (e.g., G), the signal intensity will be zero for both portions. Similarly, the signal associated with a nucleobase may be zero for an image captured in one particular channel of a base calling cycle. Accordingly, it will be appreciated that the polynucleotide molecules may be selectively processed such that, for signals of non-zero intensity, an intensity of the signals obtained based upon respective nucleobases of the different sequence portions is different.

30 Sequencing

35 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.

5 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

10 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

15 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

20 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

25 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, such as a (visible) light signal. One method for detecting the fluorescently labelled nucleotides comprises using laser

30 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.

35 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
5 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 5A – left). For example, a first nucleotide type (e.g. A) may
10 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
15 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
20 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.

25 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.

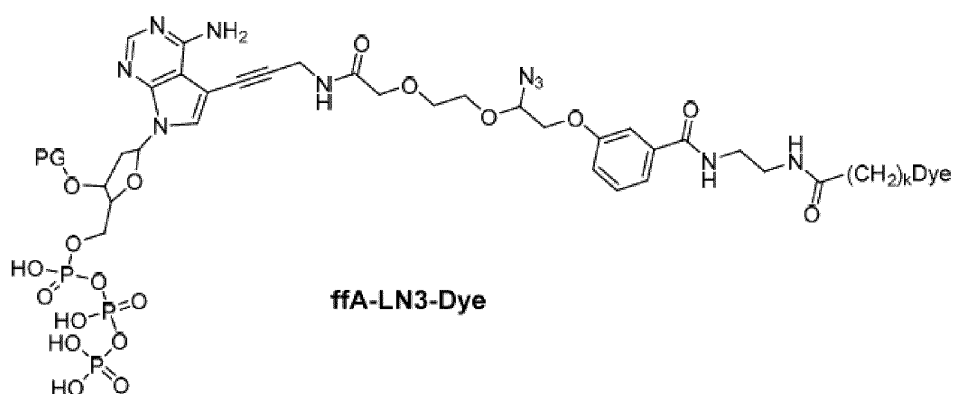
30 Thus, in some embodiments, two channels may be used to detect four nucleobases (also known as 2-channel chemistry) (Figure 5A – middle). For example, a first nucleotide type (e.g. A) may include a first label (e.g. configured to emit a first wavelength, such as green light) and a second label (e.g. configured to emit a second wavelength, such as red light),
35 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,

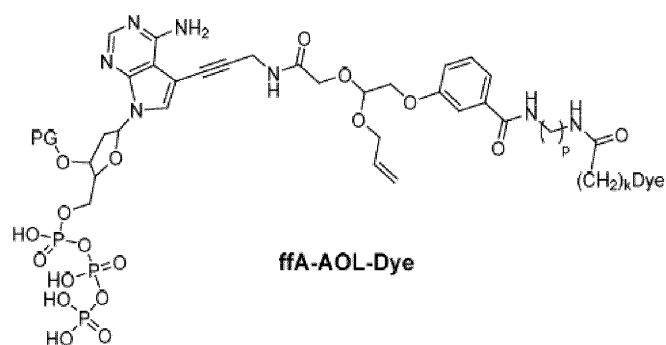
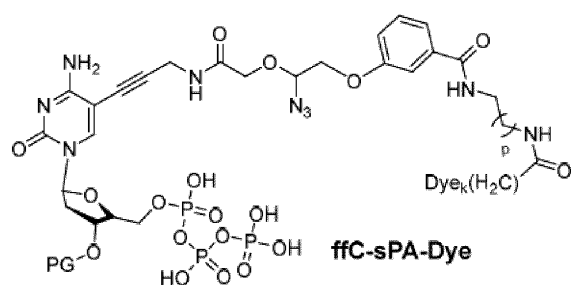
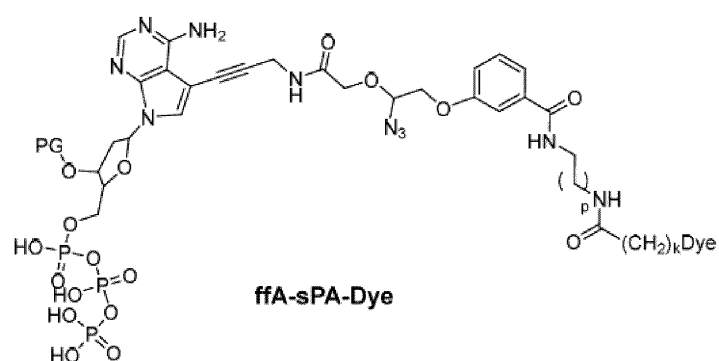
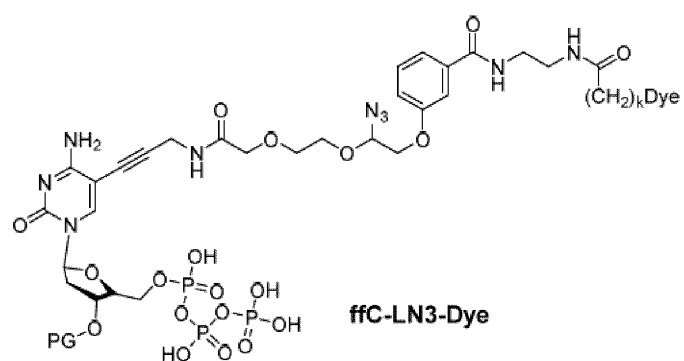
and a fourth nucleotide type (e.g. C) may not include the first label and may include the second label (e.g. configured to emit the second wavelength, such as red light). Two images can then be obtained, using detection channels for the first label and the second label. For example, the first nucleotide type (e.g. A) may be detected in both a first
5 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
10 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
15 be used.

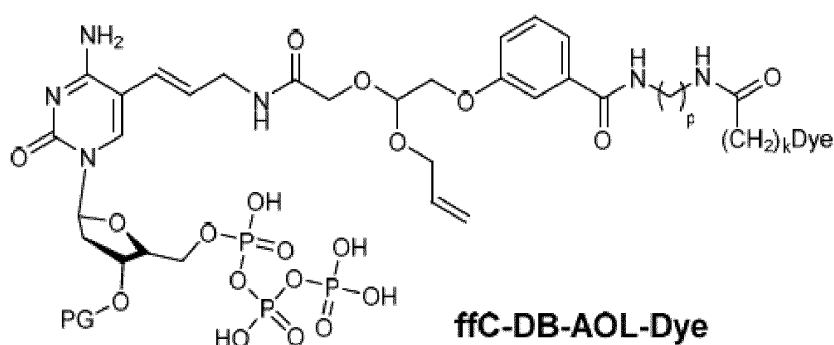
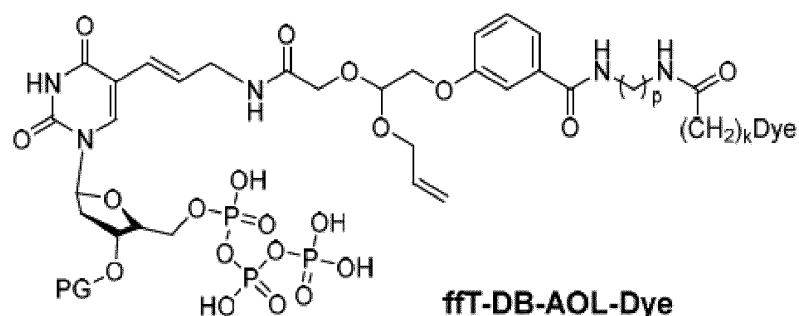
In some embodiments, one channel may be used to detect four nucleobases (also known as 1-channel chemistry) (Figure 5A – right). For example, a first nucleotide type (e.g. A) may include a cleavable label (e.g. configured to emit a wavelength, such as green light),
20 a second nucleotide type (e.g. G) 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
25 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
30 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
35 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 some embodiments, the fluorescent labels are selected from the group consisting of polymethine derivatives, coumarin derivatives, benzopyran derivatives, chromenoquinoline derivatives, compounds containing bis-boron heterocycles such as BOPPY and BOPYPY. In some embodiments, the fluorescent label is attached to the nucleotide through a cleavable linker. In some further embodiments, the labeled nucleotide may have the fluorescent label attached to the C5 position of a pyrimidine base or the C7 position of a 7-deaza purine base, optionally through a cleavable linker moiety. For example, the nucleobase may be 7-deaza adenine and the dye is attached to the 7-deaza adenine at the C7 position, optionally through a cleavable linker. The nucleobase may be 7-deaza guanine and the dye is attached to the 7-deaza guanine at the C7 position, optionally through a cleavable linker. The nucleobase may be cytosine and the dye is attached to the cytosine at the C5 position, optionally through a cleavable linker. As another example, the nucleobase may be thymine or uracil and the dye is attached to the thymine or uracil at the C5 position, optionally through a cleavable linker. In some further embodiments, the cleavable linker may comprise similar or the same chemical moiety as the reversible terminator 3' hydroxy blocking group such that the 3' hydroxy blocking group and the cleavable linker may be removed under the same reaction condition or in a single chemical reaction. Non-limiting example of the cleavable linker include the LN3 linker, the sPA linker, and the AOL linker, each of which is exemplified below.







In some embodiments, the nucleotides are selected from the group consisting of an analog of dGTP, an analog of dTTP, an analog of dUTP, an analog of dCTP, and an analog of dATP. In some embodiments, the first nucleotide is a first reversibly blocked nucleotide triphosphate (rbNTP), the second nucleotide is a second rbNTP, the third nucleotide is a third rbNTP, and the fourth nucleotide is a fourth rbNTP, wherein each of the first nucleotide, second nucleotide, third nucleotide and fourth nucleotide is a different type of nucleotide from the other. In some embodiments, the four rbNTPs are selected from the group consisting of rbATP, rbTTP, rbUTP, rbCTP, and rbGTP. In some

5 nucleotides, each of the four rbNTPs includes a modified base and a reversible terminator 3' blocking group. Non-limiting example of the 3' blocking group include azidomethyl (*-CH₂N₃), substituted azidomethyl (e.g., *-CH(CHF₂)N₃ or *-CH(CH₂F)N₃) and *-CH₂-O-CH₂-CH=CH₂, where the asterisk * indicates the point attachment to the 3' oxygen of the ribose or deoxyribose ring of the nucleotide.

15

Further details about the dyes and the fully functionalized nucleotides can be found in U.S. Patent Application Publication Numbers 2018/0094140 and 2020/0277670, International Patent Application Publication Number 2017/051201, and U.S. Provisional Patent Application Numbers 63/057758 and 63/127061, the disclosures of which are

20 incorporated herein by reference in their entireties.

Signal processing

Figure 6 is a scatter plot showing an example of sixteen distributions of signals from a nucleic acid cluster as illustrated in Figures 4A-4B, which may be implemented with the dye labeling scheme shown in Figure 5B in one example. As explained in connection with Figures 4A-4B, in one embodiment, the fluorescent signal coming from the collection of extended first portion sequencing primers 402a will be brighter than the fluorescent signal coming from the collection of extended second portion sequencing primers 402b in the same cluster. The scatter plot of Figure 6 shows sixteen distributions (or "bins"/"classifications") of intensity values from the combination of a brighter signal and a dimmer signal; the two signals may be co-localized and may not be optically resolved as described above. The intensity values shown in Figure 6 may be up to a scale or normalization 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 from the extended first portion primers 402a and the dimmer signal from the extended second portion primers 402b results in a combined signal. The combined signal may be captured by the first optical channel and the second optical channel (e.g., the "IMAGE 1" channel and the "IMAGE 2" channel in Figure 5B). 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 according to the embodiment shown in connection with Figure 5B. That is, each of the sixteen possibilities corresponds to a bin shown in Figure 6. The computer system can map the combined signal from a cluster into one of the sixteen bins, and thus determine the added nucleobase at the extended first portion primers 402a and the added nucleobase at the extended second portion primers 402b, respectively.

For example, when the combined signal is mapped to bin 612 for a base calling cycle, the computer processor base calls both the added nucleobase at the extended first portion primers 402a and the added nucleobase at the extended second portion primers 402b as C. When the combined signal is mapped to bin 614 for the base calling cycle, the processor base calls the added nucleobase at the extended first portion primers 402a as C and the added nucleobase at the extended second portion primers 402b as T. When the combined signal is mapped to bin 616 for the base calling cycle, the processor base calls the added nucleobase at the extended first portion primers 402a as C and the added nucleobase at the extended second portion primers 402b as G. When the combined

signal is mapped to bin 618 for the base calling cycle, the processor base calls the added nucleobase at the extended first portion primers 402a as C and the added nucleobase at the extended second portion primers 402b as A.

5 When the combined signal is mapped to bin 622 for the base calling cycle, the processor base calls the added nucleobase at the extended first portion primers 402a as T and the added nucleobase at the extended second portion primers 402b as C. When the combined signal is mapped to bin 624 for the base calling cycle, the processor base calls both the added nucleobase at the extended first portion primers 402a and the added
10 nucleobase at the extended second portion primers 402b as T. When the combined signal is mapped to bin 626 for the base calling cycle, the processor base calls the added nucleobase at the extended first portion primers 402a as T and the added nucleobase at the extended second portion primers 402b as G. When the combined signal is mapped to bin 628 for the base calling cycle, the processor base calls the added nucleobase at
15 the extended first portion primers 402a as T and the added nucleobase at the extended second portion primers 402b as A.

When the combined signal is mapped to bin 632 for the base calling cycle, the processor base calls the added nucleobase at the extended first portion primers 402a as G and the
20 added nucleobase at the extended second portion primers 402b as C. When the combined signal is mapped to bin 634 for the base calling cycle, the processor base calls the added nucleobase at the extended first portion primers 402a as G and the added nucleobase at the extended second portion primers 402b as T. When the combined signal is mapped to bin 636 for the base calling cycle, the processor base calls both the
25 added nucleobase at the extended first portion primers 402a and the added nucleobase at the extended second portion primers 402b as G. When the combined signal is mapped to bin 638 for the base calling cycle, the processor base calls the added nucleobase at the extended first portion primers 402a as G and the added nucleobase at the extended first portion primers 402b as A.

30 When the combined signal is mapped to bin 642 for the base calling cycle, the processor base calls the added nucleobase at the extended first portion primers 402a as A and the added nucleobase at the extended second portion primers 402b as C. When the combined signal is mapped to bin 644 for the base calling cycle, the processor base calls
35 the added nucleobase at the extended first portion primers 402a as A and the added nucleobase at the extended second portion primers 402b as T. When the combined signal is mapped to bin 646 for the base calling cycle, the processor base calls the added

nucleobase at the extended first portion primers 402a as A and the added nucleobase at the extended second portion primers 402b as G. When the combined signal is mapped to bin 648 for the base calling cycle, the processor base calls both the added nucleobase at the extended first portion primers 402a and the added nucleobase at the extended
5 second portion primers 402b as A. 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.

10 It will be appreciated that the number of distributions may be different, depending upon the number of sequence portions of interest which are concurrently sequenced. For example, the plurality of classifications may comprise 4^n classifications, each classification representing one of 4^n unique combinations of nth nucleobases in each of n sequence portions.

15 Accordingly, in one aspect of the present disclosure, there is provided a method of base calling nucleobases of n polynucleotide sequence portions, the method comprising:

(a) obtaining first intensity data comprising a combined intensity of respective first signal components generated by each of the nth portions obtained based upon
20 respective nth nucleobases in each of the n portions, wherein the respective first signal components are obtained simultaneously;

(b) obtaining second intensity data comprising a combined intensity of respective second signal components generated by each of the nth portions obtained based upon respective nth nucleobases in each of the n portions, wherein the respective
25 second 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 nth nucleobases; and

(d) based on the selected classification, base calling the respective nth
30 nucleobases for all n portions,

wherein said polynucleotide sequence portions have been selectively processed such that an intensity of the signals obtained based upon the respective first nucleobase is greater than an intensity of the signals obtained based upon the respective second nucleobase.

35 Selecting the classification based on the first and second intensity data may comprise selecting the classification based on the combined intensity of respective first signal

components and second signal components. The plurality of classifications may comprise 4^n classifications, each classification representing one of 4^n unique combinations of n th nucleobases.

5 In addition to base calling the first and second polynucleotide sequence portions, 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 and second polynucleotide sequence portions, or about sequences from which the first and second portions were derived. For example,
10 given the nucleic acid material input and the processing methods used to generate the nucleic acid clusters, the first and second sequence portions 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. Alternatively, the first and second sequence portions may be expected to
15 be different, for example due to deliberate sequence modifications introduced during library preparation.

Detecting library preparation errors

20 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 (as shown in Figure 7). One solution
25 to this problem is to use unique molecular identifiers or indices (UMIs) to distinguish true (e.g. rare) mutations from mutations arising due to library preparation errors (as shown in Figure 8). The present methods, however, allow for the identification of library preparation errors from fewer sequencing reads.

30 In one example, a plurality of clusters are generated, each comprising at least one first polynucleotide sequence portion and at least one second polynucleotide sequence portion, wherein the first and second portions will be the same in the absence of any library preparation or sequencing errors. For example, the nucleic acid clusters may be processed such that each of the one or more first portions corresponds to the sequence
35 of a forward strand (or a reverse strand) of a double-stranded template molecule, and the one or more second portions corresponds to the complement of the reverse strand (or the forward strand) of the template. Examples of suitable methods for generating

such clusters are described below. In the absence of any library preparation/sequencing errors, the signals produced by subjecting the two sequence portions to sequencing-by-synthesis will match. The combined signal may therefore be mapped to one of the four “corner” clouds shown in Figure 9. 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 and second sequence portions 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.

Depending upon the library preparation methods used, it is not necessarily the case, however, that a match between the first and second nucleobases added to the first and second polynucleotide sequence portions is indicative of the absence of a library preparation error. Similarly, it is not necessarily the case that a mismatch between the first and second nucleobases is indicative of the presence of a library preparation error. For example, one or more sequence modifications may be intentionally introduced during library preparation as described in the example below.

Detecting sequence modifications

The present technology also allows for the detection of sequence modifications (e.g. deliberate sequence modifications) made during library preparation. In particular, *a priori* information of a modification performed to obtain a sequence may be used to obtain information associated with the modification.

Of the many possible DNA modifications, the methylation of cytosines is the most frequently observed in relation to gene regulation. In order to determine the methylation profile of a nucleic acid sequence, it is known to treat the nucleic acid sequence (e.g. chemically or enzymatically) to convert either methylated or unmethylated cytosine to a different base. For example, bisulfite treatment may be used to convert unmethylated cytosine to uracil. Alternatively, borane treatment, or enzymatic conversion (e.g. using an APOBEC or activation-induced cytidine deaminase (AID) enzyme) may be used to convert 5-methylcytosine to thymine. According to prior methods, inferring methylation status requires conversion, sequencing, and comparison either to a database reference, to an unconverted reference, or to an opposite strand in a consensus pileup. The present

methods, however, allow for the methylation status of a sequence to be determined in real-time, from a single sequencing run, and without the need for alignment.

Figure 10 shows an example of a method for the identification of methylated bases according to the present disclosure. Concatenated polynucleotide sequences comprising a first portion and a second portion are prepared using a tandem insert method as described below. A first portion will have the methylation fingerprint of the original molecule, while the second portion will have lost the methylation fingerprint but will retain the original base information. Following conversion treatment of methylcytosines to thymine (e.g. using Borane or APOBEC treatment), cytosines will be converted to the modified base, while the second portion retains the original base information. The sequence of the original molecule, including methylation status, can then be determined based upon the distribution to which the combined signal intensities from the two portions are mapped.

Alternatively, as shown in Figure 11, unmodified cytosines may be converted to uracil, for example using bisulfite treatment. Here, the complementary strand may also be used in order to extract the full original sequence information.

Alternative library preparation and cluster generation methods may also be used for the purpose of determining methylation status. For example, nucleic acid clusters may be prepared, each comprising one or more first polynucleotide sequence portions corresponding to the sequence of a forward strand (or a reverse strand) of a treated molecule, and one or more second polynucleotide sequence portions corresponding to the complement of the reverse strand (or the forward strand) of the treated molecule. By virtue of the methylation conversion happening prior to the copying process, methylation information of both strands of the original molecule can be determined in a single sequencing run.

30 Signal mapping

In one example, the combined signals may be mapped to the distributions by using a Gaussian Mixture Model (GMM). For example, as shown in Figure 12, raw intensities for each cluster may first be separately normalized. A GMM with four sources may then be fitted and used to predict the brighter or "major" read (bottom left-hand panel). The intensities with the same major read may then be used to train a GMM with four sources for the dimmer or "minor" read (top right-hand panel), and the trained GMM may be used

to predict the minor read for each major read, resulting in 16 cluster centers (bottom right-hand panel). A GMM with 16 sources may then be initialized with these 16 cluster centers, and trained with all cluster intensities for each cycle. The two reads may then be predicted using GMM cluster assignment and correlation with the expected cloud constellation.

Simplified Sequencing Workflow

Figure 13 is a flow diagram showing a method 700 of base calling according to the present disclosure. The described method allows for simultaneous sequencing of two or more sequence portions in a single sequencing run from a single combined signal obtained from the two or more portions, thus requiring less sequencing reagent consumption and faster generation of data from both portions. Further, the simplified method may reduce the number of workflow steps while producing the same yield as compared to existing next-generation sequencing methods. Thus, the simplified method may result in reduced sequencing runtime.

As shown in Figure 13, the disclosed method 700 may start from block 701. The method may then move to block 710.

At block 710, 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 obtained based upon a respective first nucleobase of at least one first polynucleotide sequence portion and a second signal obtained based upon a respective second nucleobase of at least one second polynucleotide sequence portion. Similarly, the second intensity data comprises a combined intensity of a third signal obtained based upon the respective first nucleobase of the at least one first polynucleotide sequence portion and a fourth signal obtained based upon the respective second nucleobase of the at least one second polynucleotide sequence portion.

As described above, polynucleotide molecules comprising the at least one first polynucleotide sequence portion and the at least one second polynucleotide sequence portion may be arranged on the flow cell such that light emissions from the first and second portions are detected by a single sensing portion and/or may comprise a single cluster such that light emissions from each of the respective two polynucleotide sequence portions cannot be spatially resolved.

In one example, the signals may be generated according to the method shown in Figure 14.

5 In one example, obtaining the intensity data comprises selecting intensity data that corresponds to two or more different sequence portions. 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
10 described above, it may be desired to produce clusters comprising two different sequence portions of interest, which give rise to signals in a ratio of 2:1. In one example, high-quality data corresponding to two sequence portions with an intensity ratio of 2:1 may have a chastity score of around 0.8 to 0.9. In one example, clusters may be identified as containing one or more than one polynucleotide sequence portion of interest (e.g. by
15 chastity score) and processed accordingly. For example, clusters containing more than one sequence portion of interest may be base called according to the present methods, whereas clusters containing a single sequence portion of interest may be base called according to known methods.

20 After the intensity data has been obtained, the method may proceed to block 720. 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 6, each representing a unique combination of first and
25 second nucleobases. Where there are two polynucleotide sequence regions of interest, 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 signals and the combined intensity of the third and fourth signals, for example using GMMs as described
30 above.

The method may then proceed to block 730, where the respective first and second nucleobases are base called based on the classification selected in block 720. The light emissions generated during a cycle of a sequencing-by-synthesis method are indicative
35 of the identity of the nucleobase(s) added to the sequencing primers undergoing extension. It will be appreciated that there is a direct correspondence between the identity of the nucleobases incorporated into the sequencing primers and the identity of

the complementary base at the corresponding position of the sequence portion bound to the flow cell. Therefore, any references herein to the base calling of respective nucleobases of polynucleotide sequence portions encompasses the base calling of nucleobases hybridized to the polynucleotide sequence portions and, alternatively or
5 additionally, the identification of the corresponding nucleobases of the portions. The method may then end at block 740.

Figure 14 is a flow diagram showing a method 800 by which the signals discussed in relation to block 710 of Figure 13 may be generated. The method may start from block
10 801.

The method may then move to block 810, default oligo grafting, which may include the attachment of oligonucleotide anchors/graft sequences to a planar, optically transparent surface of the flow cell. The method may then move to block 820, generating DNA
15 libraries from a sample, where template polynucleotides in a sample may be end-repaired to generate 5'-phosphorylated blunt ends, and the polymerase activity of Klenow fragment may be used to add a single A base to the 3' end of the blunt phosphorylated nucleic acid fragments. This addition prepares the nucleic acid fragments for ligation to oligonucleotide adapters, which have an overhang of a single T
20 base at their 3' end to increase ligation efficiency. The adapter oligonucleotides are complementary to the flow cell anchor oligos.

After DNA library generation, the method may then move to block 830, denaturing the double stranded DNA libraries to generate single stranded template polynucleotides for
25 seeding on the flow cell. The method may then move to block 840, clustering from the single stranded template polynucleotides. Under limiting-dilution conditions, adapter-modified, single-stranded template polynucleotides are added to the flow cell and immobilized by hybridization to the anchor oligos. Attached nucleic acid fragments are extended and bridge amplified to create an ultra-high density sequencing flow cell with
30 hundreds of millions of clusters, each containing about 1,000 copies of the same template. Details regarding enrichment of nucleic acids using cluster amplification may be found in Kozarewa et al., Nature Methods 6:291-295 (2009), which is incorporated herein by reference.

35 After cluster generation, the method may directly move to block 850, hybridizing/annealing first and second primers 402a, 402b simultaneously to both the first and second polynucleotide sequence portions 401a, 401b on the flow cell 410. Next,

the method may move to block 860 of signal generation. Signal generation proceeds by simultaneously extending the hybridized primers 402a, 402b. With each cycle, fluorescently tagged nucleotides compete for addition to the growing chains of extended primers. Only one is incorporated at a primer location based on the sequence of the
5 template strand. After the addition of nucleotides, the cluster is excited by a light source, and characteristic fluorescent signals are emitted. The emission spectra and the signal intensities uniquely determine the base call. Hundreds of millions of nucleic acid clusters, or thousands to tens of thousands of millions of clusters, may be sequenced in a massively parallel manner. After sequencing the polynucleotide sequence portions 401a,
10 401b on the flow cell 410, the method may end at block 870.

Methods of library preparation, cluster generation and amplification, sequencing, and selective processing which are suitable for use with the present base calling methods will now be described in further detail.

15

Library strands and template terminology

For a given double-stranded polynucleotide sequence 1100 to be identified, the polynucleotide sequence 1100 comprises a forward strand of the sequence 1101 and a
20 reverse strand of the sequence 1102. See Figure 15.

When the polynucleotide sequence 1100 is replicated (e.g. using a DNA/RNA polymerase), complementary versions of the forward strand 1101 of the sequence 1100 and the reverse strand 1102 of the sequence 1100 are generated. Thus, replication of
25 the polynucleotide sequence 1100 provides a double-stranded polynucleotide sequence 1100a that comprises a forward strand of the sequence 1101 and a forward complement strand of the sequence 1101', and a double-stranded polynucleotide sequence 1100b that comprises a reverse strand of the sequence 1102 and a reverse complement strand of the sequence 1102'.

30

The term "template" may be used to describe a complementary version of the double-stranded polynucleotide sequence 1100. As such, the "template" comprises a forward complement strand of the sequence 1101' and a reverse complement strand of the sequence 1102'. Thus, by using the forward complement strand of the sequence 1101'
35 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 1101. Similarly, by using the

reverse complement strand of the sequence 1102' 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 1102.

5

The two strands in the template may also be referred to as a forward strand of the template 1101' and a reverse strand of the template 1102'. The complement of the forward strand of the template 1101' is termed the forward complement strand of the template 1101, whilst the complement of the reverse strand of the template 1102' is termed the reverse complement strand of the template 1102.

10

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 1100 or with respect to the "template", these terms may be interpreted as referring to the "template".

15

Language for original polynucleotide sequence 1100	Corresponding language for the "template"
Forward strand of the sequence 1101	Forward complement strand of the template 1101 (sometimes referred to herein as forward complement strand 1101)
Reverse strand of the sequence 1102	Reverse complement strand of the template 1102 (sometimes referred to herein as reverse complement strand 1102)
Forward complement strand of the sequence 1101'	Forward strand of the template 1101' (sometimes referred to herein as forward strand 1101')
Reverse complement strand of the sequence 1102'	Reverse strand of the template 1102' (sometimes referred to herein as reverse strand 1102')

Library preparation

20

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.

5 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.

10

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. Alternatively, the templates to be generated typically include separate polynucleotide sequences, in particular a first polynucleotide sequence comprising a first
15 portion and a second polynucleotide sequence comprising a second portion. Generating these templates from particular libraries may be performed according to methods known to persons of skill in the art. However, some example approaches of preparing libraries suitable for generation of such templates are described below.

20

In some embodiments, the library may be prepared by ligating adaptor sequences to double-stranded polynucleotide sequences, each comprising a forward strand of the sequence and a reverse strand of the sequence, as described in more detail in e.g. WO 07/052006, which is incorporated herein by reference. In some cases, “tagmentation” can be used to attach the sample DNA to the adaptors, as described in more detail in
25 e.g. WO 10/048605, US 2012/0301925, US 2013/0143774 and WO 2016/189331, each of which are incorporated herein by reference. In tagmentation, double-stranded DNA is simultaneously fragmented and tagged with adaptor sequences and PCR primer binding sites. The combined reaction eliminates the need for a separate mechanical shearing step during library preparation. These procedures may be used, for example, for
30 preparing templates including a first polynucleotide sequence comprising a first portion and a second polynucleotide sequence comprising a second portion, wherein the first portion is a forward strand of the template, and the second portion is a forward complement strand of the template – i.e. a copy of the forward strand (or alternatively, wherein the first portion is a reverse strand of the template, and the second portion is a
35 reverse complement strand of the template).

Where features herein are described in relation to the “forward” strand, it should be considered that these features could equally be applied to the “reverse strand”.

Where libraries are prepared by ligating adaptor sequences to double-stranded polynucleotide sequences as described above, library preparation may comprise ligating a first primer-binding sequence 1301' (e.g. P5', such as SEQ ID NO: 3) and a second terminal sequencing primer binding site 1304 (e.g. SBS3', for example, SEQ ID NO: 38) to a 3'-end of a forward strand of a sequence 1101. See Figure 16. The library preparation may be arranged such that the second terminal sequencing primer binding site 1304 is attached (e.g. directly attached) to the 3'-end of the forward strand of the sequence 1101, and such that the first primer-binding sequence 1301' is attached (e.g. directly attached) to the 3'-end of the second terminal sequencing primer binding site 1304.

The library preparation may further comprise ligating a complement of first terminal sequencing primer binding site 1303' (e.g. SBS12, such as SEQ ID NO: 39) (also referred to herein as a first terminal sequencing primer binding site complement 1303') and a complement of a second primer-binding sequence 1302 (also referred to herein as a second primer-binding complement sequence 1302) (e.g. P7, such as SEQ ID NO: 2) to a 5'-end of the forward strand of the sequence 1101. The library preparation may be arranged such that first terminal sequencing primer binding site complement 1303' is attached (e.g. directly attached) to the 5'-end of the forward strand of the sequence 1101, and such that second primer-binding complement sequence 1302 is attached (e.g. directly attached) to the 5'-end of first terminal sequencing primer binding site complement 1303'.

Thus, one strand of a polynucleotide within a polynucleotide library may comprise, in a 5' to 3' direction, a second primer-binding complement sequence 1302 (e.g. P7), a first terminal sequencing primer binding site complement 1303' (e.g. SBS12), a forward strand of the sequence 1101, a second terminal sequencing primer binding site 1304 (e.g. SBS3'), and a first primer-binding sequence 1301' (e.g. P5') (Figure 16 – bottom strand).

Although not shown in Figure 16, 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 1302 (e.g. P7) and the first terminal sequencing primer binding site complement 1303' (e.g. SBS12). Separately, or in

addition, a second index complement sequence (e.g. i5') may be provided between the second terminal sequencing primer binding site 1304 (e.g. SBS3') and the first primer-binding sequence 1301' (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 1302 (e.g. P7), a first index sequence (e.g. i7), a first terminal sequencing primer binding site complement 1303' (e.g. SBS12), a forward strand of the sequence 1101, a second terminal sequencing primer binding site 1304 (e.g. SBS3'), a second index complement sequence (e.g. i5'), and a first primer-binding sequence 1301' (e.g. P5'). A typical polynucleotide is shown in Figure 17 (bottom strand).

When a double-stranded sequence 1100 is used, the library preparation may also comprise ligating a second primer-binding sequence 1302' (e.g. P7') and a first terminal sequencing primer binding site 1303 (e.g. SBS12') to a 3'-end of a reverse strand of a sequence 1102. The library preparation may be arranged such that first terminal sequencing primer binding site 1303 is attached (e.g. directly attached) to the 3'-end of the reverse strand of the sequence 1102, and such that the second primer-binding sequence 1302' is attached (e.g. directly attached) to the 3'-end of first terminal sequencing primer binding site 1303.

The library preparation may further comprise ligating a complement of a second terminal sequencing primer binding site 1304' (e.g. SBS3) (also referred to herein as a second terminal sequencing primer binding site complement 1304') and a complement of a first primer-binding sequence 1301 (also referred to herein as a first primer-binding complement sequence 1301) (e.g. P5) to a 5'-end of the reverse strand of the sequence 1102. The library preparation may be arranged such that the second terminal sequencing primer binding site complement 1304' is attached (e.g. directly attached) to the 5'-end of the reverse strand of the sequence 1102, and such that the first primer-binding complement sequence 1301 is attached (e.g. directly attached) to the 5'-end of the second terminal sequencing primer binding site complement 1304'.

Thus, another strand of a polynucleotide within a polynucleotide library may comprise, in a 5' to 3' direction, a first primer-binding complement sequence 1301 (e.g. P5), a second terminal sequencing primer binding site complement 1304' (e.g. SBS3), a reverse strand of the sequence 1102, a first terminal sequencing primer binding site 1303 (e.g. SBS12'), and a second primer-binding sequence 1302' (e.g. P7') (Figure 16 – top strand).

Although not shown in Figure 16, 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 1301 (e.g. P5) and the second terminal sequencing primer binding site complement 1304' (e.g. SBS3). Separately, or in addition, a first index complement sequence (e.g. i7') may be provided between the first terminal sequencing primer binding site 1303 (e.g. SBS12') and the second primer-binding sequence 1302' (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 1301 (e.g. P5), a second index sequence (e.g. i5), a second terminal sequencing primer binding site complement 1304' (e.g. SBS3), a reverse strand of the sequence 1102, a first terminal sequencing primer binding site 1303 (e.g. SBS12'), a first index complement sequence (e.g. i7'), and a second primer-binding sequence 1302' (e.g. P7'). A typical polynucleotide is shown in Figure 17 (top strand).

In some embodiments, the library may be prepared using PCR stitching methods, such as (splicing by) overlap extension PCR (also known as OE-PCR or SOE-PCR), as described in more detail in e.g. Higuchi et al. (Nucleic Acids Res., 1988, vol. 16, pp. 7351-7367), which is incorporated herein by reference. This procedure may be used, for example, for preparing templates including concatenated polynucleotide sequences comprising a first portion and a second portion, wherein the first portion and the second portion are different polynucleotide sequences (e.g. genetically unrelated, and/or obtained from different sources). A representative process for conducting PCR stitching for a human and PhiX library is shown in Figure 18.

As used herein, the term "genetically unrelated" refers to portions which are not related in the sense of being any two of the group consisting of: forward strands, reverse strands, forward complement strands, and reverse complement strands. However, the "genetically unrelated" sequences could be different fragment sequences which are derived from the same source, but are different fragments from that source (e.g. from the same fragmented library preparation process). This includes sequences that can be overlapping in sequence (but not identical in sequence).

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 19(A) to 19(E).

Figure 20A shows the effect of optional pre-treatment of library strands using mC to T conversion on bases in template strands. Figure 20B shows the effect of optional pre-treatment of library strands using C to U conversion on bases in template strands.

In some embodiments, the library may be prepared using a loop fork method, which is described below. This procedure may be used, for example, for preparing templates including a first polynucleotide sequence comprising a first portion and a second polynucleotide sequence comprising a second portion, wherein the first portion is a forward strand of the template, and the second portion is a reverse complement strand of the template (or alternatively, wherein the first portion is a reverse strand of the template, and the second portion is a forward complement strand of the template). This procedure may also 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 strand of the template. Such libraries may also be referred to as self-tandem inserts. A representative process for conducting a loop fork method is shown in Figure 21.

Starting from a double-stranded polynucleotide sequence comprising a forward strand of the sequence and a reverse strand of the sequence, adaptors may be ligated to a first end of the sequence (e.g. using processes as described in more detail in e.g. WO 07/052006, or "tagmentation" methods as described above). A second end of the sequence (different from the first end) may be ligated to a loop, which connects the forward strand of the sequence and the reverse strand of the sequence, thus generating a loop fork ligated polynucleotide sequence. Conducting PCR on the loop fork ligated polynucleotide sequence produces a new double-stranded polynucleotide sequence, one strand comprising the forward strand of the sequence and the reverse strand of the sequence, and the other strand comprising a forward complement strand of the sequence and a reverse complement strand of the sequence. The library is now ready for seeding, clustering and amplification.

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

The processes described above in relation to PCR stitching, tandem insert methods, and loop fork 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 1302 (e.g. P7), a first terminal sequencing primer binding site complement 1303' (e.g. B15-ME; or if ME is not present, then B15), a first insert sequence 1401, a hybridisation complement sequence 1403 (e.g. ME'-HYB2-ME; or if ME' and ME are not present, then HYB2), a second insert sequence 1402, a second terminal sequencing primer binding site 1304 (e.g. ME'-A14'; or if ME' is not present, then A14'), and a first primer-binding sequence 1301' (e.g. P5') (Figures 16 and 17 – bottom strand).

Although not shown in Figures 16 and 17, 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 1302 (e.g. P7) and the first terminal sequencing primer binding site complement 1303' (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 1304 (e.g. ME'-A14') and the first primer-binding sequence 1301' (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 1302 (e.g. P7), a first index sequence (e.g. i7), a first terminal sequencing primer binding site complement 1303' (e.g. B15-ME; or if ME is not present, then B15), a first insert sequence 1401, a hybridisation complement sequence 1403 (e.g. ME'-HYB2-ME; or if ME' and ME are not present, then HYB2), a second insert sequence 1402, a second terminal sequencing primer binding site 1304 (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 1301' (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 1301 (e.g. P5), a second terminal sequencing primer binding site complement 1304' (e.g. A14-ME; or if ME is not present, then A14), a second insert complement sequence 1402', a hybridisation sequence 1403' (e.g. ME'-HYB2'-ME; or if ME' and ME are not present, then HYB2'), a first insert complement sequence 1401', a first terminal sequencing primer binding site 1303 (e.g. ME'-B15'; or if ME' is not present, then B15'), and a second primer-binding sequence 1302' (e.g. P7') (Figures 16 and 17 – top strand).

Although not shown in Figures 16 and 17, 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 1301 (e.g. P5) and the second terminal sequencing primer binding site complement 1304' (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 1303 (e.g. ME'-B15'; or if ME' is not present, then B15') and the second primer-binding sequence 1302' (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 1301 (e.g. P5), a second index sequence (e.g. i5), a second terminal sequencing primer binding site complement 1304' (e.g. A14-ME; or if ME is not present, then A14).), a second insert complement sequence 1402', a hybridisation sequence 1403' (e.g. ME'-HYB2'-ME; or if ME' and ME are not present, then HYB2'), a first insert complement sequence 1401', a first terminal sequencing primer binding site 1303 (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 1302' (e.g. P7').

As described herein, the first insert sequence 1401 and the second insert sequence 1402 may comprise different types of library sequences.

In one embodiment, the first insert sequence 1401 may be different to the second insert sequence 1402 (e.g. genetically unrelated, and/or obtained from different sources), for example where the library is prepared using PCR stitching.

In another embodiment, the first insert sequence 1401 may comprise a forward strand of the sequence 1101, and the second insert sequence may comprise a reverse complement strand of the sequence 1102' (or the first insert sequence 1401 may comprise a reverse strand of the sequence 1102, and the second insert sequence 1402

may comprise a forward complement strand of the sequence 1101'), for example where the library is prepared using a tandem insert method.

In another embodiment, the first insert sequence 1401 may comprise a forward strand of the sequence 1101, and the second insert sequence 1402 may comprise a reverse strand of the sequence 1102 (or the first insert sequence 1401 may comprise a forward complement strand of the sequence 1101', and the second insert sequence 1402 may comprise a reverse complement strand of the sequence 1102'), for example where the library is prepared using a loop fork 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.

5 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.

10 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
15 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
20 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
25 (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
30 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
35 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).

5 Thus, embodiments of the present invention may be performed on a solid support 1200, such as a flowcell. However, in alternative embodiments, seeding and clustering can be conducted off-flowcell using other types of solid support.

10 The solid support 1200 may comprise a substrate 1204. See Figure 24. The substrate 1204 comprises at least one well 1203 (e.g. a nanowell), and typically comprises a plurality of wells 1203 (e.g. a plurality of nanowells).

The solid support may comprise at least one first immobilised primer and at least one second immobilised primer.

15 Thus, each well 1203 may comprise at least one first immobilised primer 1201, and typically may comprise a plurality of first immobilised primers 1201. In addition, each well 1203 may comprise at least one second immobilised primer 1202, and typically may comprise a plurality of second immobilised primers 1202. Thus, each well 1203 may
20 comprise at least one first immobilised primer 1201 and at least one second immobilised primer 1202, and typically may comprise a plurality of first immobilised primers 1201 and a plurality of second immobilised primers 1202.

25 The first immobilised primer 1201 may be attached via a 5'-end of its polynucleotide chain to the solid support 1200. When extension occurs from first immobilised primer 1201, the extension may be in a direction away from the solid support 1200.

30 The second immobilised primer 1202 may be attached via a 5'-end of its polynucleotide chain to the solid support 1200. When extension occurs from second immobilised primer 1202, the extension may be in a direction away from the solid support 1200.

35 The first immobilised primer 1201 may be different to the second immobilised primer 1202 and/or a complement of the second immobilised primer 1202. The second immobilised primer 1202 may be different to the first immobilised primer 1201 and/or a complement of the first immobilised primer 1201.

The (or each of the) first immobilised primer(s) 1201 may comprise a sequence as defined in SEQ ID NO: 1 or 5, or a variant or fragment thereof. The second immobilised primer(s) 1202 may comprise a sequence as defined in SEQ ID NO: 2, or a variant or fragment thereof.

5

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.

5 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.

10 The steps of cluster generation and amplification for templates including a concatenated polynucleotide sequence comprising a first portion and a second portion, as well as templates including a first polynucleotide sequence comprising a first portion and a second polynucleotide sequence comprising a second portion, are illustrated below and in Figure 25.

15 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.

20

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

30

35

In an embodiment, a solution comprising a polynucleotide library prepared by ligating adaptor sequences to double-stranded polynucleotide sequences as described above may be flown across a flowcell.

- 5 A particular polynucleotide strand from the polynucleotide library to be sequenced comprising, in a 5' to 3' direction, a second primer-binding complement sequence 1302 (e.g. P7), a first terminal binding site complement 1303' (e.g. SBS12), a forward strand of the sequence 1101, a second terminal sequencing primer binding site 1304 (e.g. SBS3') and a first primer-binding sequence 1301' (e.g. P5'), may anneal (via the first
10 primer-binding sequence 1301') to the first immobilised primer 1201 (e.g. P5 lawn primer) located within a particular well 1203 (Figure 25A).

- The polynucleotide library may comprise other polynucleotide strands with different forward strands of the sequence 1101. Such other polynucleotide strands may anneal to
15 corresponding first immobilised primers 1201 (e.g. P5 lawn primers) in different wells 1203, thus enabling parallel processing of the various different strands within the polynucleotide library.

- A new polynucleotide strand may then be synthesised, extending from the first
20 immobilised primer 1201 (e.g. P5 lawn primer) in a direction away from the substrate 1204. By using complementary base-pairing, this generates a template strand comprising, in a 5' to 3' direction, the first immobilised primer 1201 (e.g. P5 lawn primer) which is attached to the solid support 1200, a second terminal sequencing primer binding site complement 1304' (e.g. SBS3), a forward strand of the template 1101' (which
25 represents a type of "first portion"), a first terminal sequencing primer binding site 1303 (which represents a type of "first sequencing primer binding site") (e.g. SBS12'), and a second primer-binding sequence 1302' (e.g. P7') (Figure 25B). Such a process may utilise an appropriate polymerase, such as a DNA or RNA polymerase.

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

- The polynucleotide strand from the polynucleotide library may then be dehybridised and washed away, leaving a template strand attached to the first immobilised primer 1201
35 (e.g. P5 lawn primer) (Figure 25C).

The second primer-binding sequence 1302' (e.g. P7') on the template strand may then anneal to a second immobilised primer 1202 (e.g. P7 lawn primer) located within the well 1203. This forms a "bridge" (Figure 25D).

5 A new polynucleotide strand may then be synthesised by bridge amplification, extending from the second immobilised primer 1202 (e.g. P7 lawn primer) (initially) in a direction away from the substrate 1204. By using complementary base-pairing, this generates a template strand comprising, in a 5' to 3' direction, the second immobilised primer 1202 (e.g. P7 lawn primer) which is attached to the solid support 1200, a first terminal sequencing primer binding site complement 1303' (e.g. SBS12), a forward complement strand of the template 1101 (which represents a type of "second portion"), a second terminal sequencing primer binding site 1304 (which represents a type of "second sequencing primer binding site") (e.g. SBS3'), and a first primer-binding sequence 1301' (e.g. P5') (Figure 25E). Again, such a process may utilise a suitable polymerase, such as a DNA or RNA polymerase.

The strand attached to the second immobilised primer 1202 (e.g. P7 lawn primer) may then be dehybridised from the strand attached to the first immobilised primer 1201 (e.g. P5 lawn primer) (Figure 25F).

20 A subsequent bridge amplification cycle can then lead to amplification of the strand attached to the first immobilised primer 1201 (e.g. P5 lawn primer) and the strand attached to the second immobilised primer 1202 (e.g. P7 lawn primer). Similar to Figure 25D, the second primer-binding sequence 1302' (e.g. P7') on the template strand attached to the first immobilised primer 1201 (e.g. P5 lawn primer) may then anneal to another second immobilised primer 1202 (e.g. P7 lawn primer) located within the well 1203. In a similar fashion, the first primer-binding sequence 1301' (e.g. P5') on the template strand attached to the second immobilised primer 1202 (e.g. P7 lawn primer) may then anneal to another first immobilised primer 1201 (e.g. P5 lawn primer) located within the well 1203 (Figure 25G).

35 Completion of bridge amplification and dehybridisation may then provide an amplified (duoclonal) cluster, thus providing a plurality of first polynucleotide sequences comprising the forward strand of the template 1101' (i.e. "first portions"), and a plurality of second polynucleotide sequences comprising the forward complement strand of the template 1101 (i.e. "second portions") (Figure 25H).

If desired, further bridge amplification cycles may be conducted to increase the number of first polynucleotide sequences and second polynucleotide sequences within the well 1203.

- 5 In this particular example, the “first portion” corresponds with the forward strand of the template 1101’, and the “second portion” corresponds with the forward complement strand of the template 1101.

10 However, other set-ups may be obtained by changing the library used. For example, by using a loop fork method to prepare a library, a portion at or close to the loop (or the loop complement) may be cleaved (e.g. by nicking). In these cases, the loop may comprise a cleavage site (e.g. a restriction recognition site, a cleavable linker, a modified nucleotide, or the like). By conducting cleavage at the loop, it is possible to produce a well 1203, where the “first portion” corresponds with a forward strand of the template, and the
15 “second portion” corresponds with a reverse complement strand of the template. As such, different types of strands for the “first portions” and “second portions” may be prepared for templates including a first polynucleotide sequence comprising a first portion and a second polynucleotide sequence comprising a second portion.

- 20 In an embodiment, a solution comprising a polynucleotide library prepared by PCR stitching, a tandem insert method or a loop fork method as described above may be flowed across a flowcell.

25 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 1302 (e.g. P7), a first terminal sequencing primer binding site complement 1303’ (e.g. B15-ME), a first insert sequence 1401, a hybridisation complement sequence 1403 (e.g. ME’-HYB2-ME), a second insert sequence 1402, a second terminal sequencing primer binding site 1304 (e.g. ME’-A14’), and a first primer-binding sequence
30 1301’ (e.g. P5’), may anneal (via the first primer-binding sequence 1301’) to the first immobilised primer 1201 (e.g. P5 lawn primer) located within a particular well 1203 (Figure 26A).

- 35 The polynucleotide library may comprise other concatenated polynucleotide strands with different first insert sequences 1401 and second insert sequences 1402. Such other polynucleotide strands may anneal to corresponding first immobilised primers 1201 (e.g.

P5 lawn primers) in different wells 1203, thus enabling parallel processing of the various different concatenated strands within the polynucleotide library.

5 A new polynucleotide strand may then be synthesised, extending from the first immobilised primer 1201 (e.g. P5 lawn primer) in a direction away from the substrate 1204. By using complementary base-pairing, this generates a template strand comprising, in a 5' to 3' direction, the first immobilised primer 1201 (e.g. P5 lawn primer) which is attached to the solid support 1200, a second terminal sequencing primer binding site complement 1304' (e.g. A14-ME; or if ME is not present, then A14), a second insert complement sequence 1402' (which represents a type of "second portion"), a hybridisation sequence 1403' (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 1401' (which represents a type of "first portion"), a first terminal sequencing primer binding site 1303 (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 1302' (e.g. P7') (Figure 26B). Such a process may utilise a polymerase, such as a DNA or RNA polymerase.

20 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 1201 (e.g. P5 lawn primer) (Figure 26C).

25 The second primer-binding sequence 1302' (e.g. P7') on the template strand may then anneal to a second immobilised primer 1202 (e.g. P7 lawn primer) located within the well 1203. This forms a "bridge" (not shown, but using a similar process as shown in Figure 25D).

30 A new polynucleotide strand may then be synthesised by bridge amplification, extending from the second immobilised primer 1202 (e.g. P7 lawn primer) (initially) in a direction away from the substrate 1204. By using complementary base-pairing, this generates a template strand comprising, in a 5' to 3' direction, the second immobilised primer 1202 (e.g. P7 lawn primer) which is attached to the solid support 1200, a first terminal sequencing primer binding site complement 1303' (e.g. B15-ME; or if ME is not present, then B15), a first insert sequence 1401, a hybridisation complement sequence 1403 (e.g.

ME'-HYB2-ME; or if ME' and ME are not present, then HYB2), a second insert sequence 1402, a second terminal sequencing primer binding site 1304 (e.g. ME'-A14'; or if ME' is not present, then A14'), and a first primer-binding sequence 1301' (e.g. P5') (not shown, but using a similar process as shown in Figure 25E). Again, such a process may utilise
5 a polymerase, such as a DNA or RNA polymerase.

The strand attached to the second immobilised primer 1202 (e.g. P7 lawn primer) may then be dehybridised from the strand attached to the first immobilised primer 1201 (e.g. P5 lawn primer) (Figure 26D).
10

A subsequent bridge amplification cycle can then lead to amplification of the strand attached to the first immobilised primer 1201 (e.g. P5 lawn primer) and the strand attached to the second immobilised primer 1202 (e.g. P7 lawn primer). Similar to Figure 25D, the second primer-binding sequence 1302' (e.g. P7') on the template strand
15 attached to the first immobilised primer 1201 (e.g. P5 lawn primer) may then anneal to another second immobilised primer 1202 (e.g. P7 lawn primer) located within the well 1203. In a similar fashion, the first primer-binding sequence 1301' (e.g. P5') on the template strand attached to the second immobilised primer 1202 (e.g. P7 lawn primer) may then anneal to another first immobilised primer 1201 (e.g. P5 lawn primer) located
20 within the well 1203 (not shown, but using a similar process as shown in Figure 25G).

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 1401' (i.e. "first portions") and a second insert complement sequence 1402' (i.e. second portions"), as well as a plurality of concatenated polynucleotide sequences comprising a first insert sequence 1401 and a second insert sequence 1402 (Figure 26E).
25

If desired, further bridge amplification cycles may be conducted to increase the number of first polynucleotide sequences and second polynucleotide sequences within the well 1203.
30

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
35 templates or the template complements (Figure 26F).

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

5

Sequencing

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 1303 in templates including a first polynucleotide sequence comprising a first portion and a second polynucleotide sequence comprising a second portion, or 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. second terminal sequencing primer binding site 1304 in templates including a first polynucleotide sequence comprising a first portion and a second polynucleotide sequence comprising a second portion, or a portion of hybridisation sequence 1403' 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. forward strand of the template 1101' in templates including a first polynucleotide sequence comprising a first portion and a second polynucleotide sequence comprising a second portion, or first insert complement sequence 1401' in templates including a concatenated polynucleotide sequence comprising a first portion and a second portion) and the second portion (e.g. forward complement strand of the template 1101 in templates including a first polynucleotide sequence comprising a first portion and a second polynucleotide sequence comprising a second portion, or second insert complement sequence 1402' 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.

- 5 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.

10 Selective processing methods

As described herein, the present methods may be used to concurrently sequence a polynucleotide strand or strands comprising two portions (in other words, a concatenated polynucleotide sequence comprising a first portion and a second portion) to be identified, or where separate strands each comprise a portion to be identified (in other words, a first polynucleotide sequence comprising a first portion and a second polynucleotide sequence comprising a second portion), such portions can be identified concurrently. This may be achieved by altering the ratio of the different portions which are capable of emitting a signal, which in turn means that during sequencing the signal from the first portion will be greater than the signal from the second portion. It is this difference in the intensity of the first and second signals that allows for the two portions, either on the same or different polynucleotide strands, to be identified simultaneously. It is of course desirable to be able to maximise the throughput and decrease the run time of a sequencing reaction. Concurrent sequencing, achieved by the methods of the present invention, enables at least a doubling of the throughput of a sequencing reaction (i.e. increased sequencing efficiency) as well as a decrease in the time taken to sequence a target polynucleotide strand(s).

Accordingly, we describe a method of preparing at least one polynucleotide sequence (or strand, such terms may be used interchangeably herein) for identification, where the method comprises selectively processing at least one polynucleotide sequence comprising a first portion and a second portion, or at least one first polynucleotide sequence comprising a first portion and at least one second polynucleotide sequence comprising a second portion, such that a proportion of first portions are capable of generating a first signal and a proportion of second portions are capable of generating a second signal, wherein the selective processing causes an intensity of the first signal to be greater than an intensity of the second signal.

The at least one polynucleotide sequence comprising the first portion and a second portion may be a plurality of polynucleotide sequences each comprising a first portion and a second portion.

5

The at least one first polynucleotide sequence comprising a first portion and the at least one second polynucleotide sequence comprising a second portion may be a plurality of first polynucleotide sequences each comprising a first portion, and a plurality of second polynucleotide sequences each comprising a second portion.

10

Accordingly, the method may comprise selectively processing a plurality of polynucleotide sequences each comprising a first portion and a second portion, or a plurality of first polynucleotide sequences each comprising a first portion and a plurality of second polynucleotide sequences each comprising a second portion, such that a proportion of first portions are capable of generating a first signal and a proportion of second portions are capable of generating a second signal, wherein the selective processing causes an intensity of the first signal to be greater than an intensity of the second signal.

15

20

By “identification” is meant here obtaining genetic information from the polynucleotide strand or polynucleotide strands. This may include identification of the genetic sequence of the polynucleotide strand or polynucleotide strands (i.e. sequencing). Furthermore, this may instead, or additionally, include identification of mismatched base pairs. In addition, this may instead, or additionally, include identification of any epigenetic modifications, for example methylation. Accordingly, “identification” may mean identification of the genetic sequence of the polynucleotide strand or polynucleotide strands, mismatched base pairs, and/or identification of any epigenetic modifications.

25

30

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), or the at least one first polynucleotide sequence comprising a first portion and at least one second polynucleotide sequence comprising a second portion (or the plurality of first polynucleotide sequences each comprising a first portion and the plurality of second polynucleotide sequences each comprising a second portion), so that the intensity of the first signal is greater than the intensity of the second signal. The property may be, for

35

example, a concentration of first portions capable of generating the first signal relative to a concentration of second portions capable of generating the second signal. The action may include, for example, conducting selective amplification, conducting selective sequencing, or preparing for selective sequencing.

5

The present invention may be applied to a single (concatenated) polynucleotide strand that comprises, on the same strand, a first portion and a second portion to be identified. As explained above, such a strand can be produced using known techniques in the art, such as PCR stitching, tandem insert methods or loop fork methods.

10

The first portions and second portions may be different polynucleotide sequences. That is, the sequences may be genetically unrelated and/or derived from different sources.

Alternatively, the first portions and second portions may be genetically related.

15

For example, 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 strand of the polynucleotide sequence (e.g. reverse strand of the template) or the forward complement strand of the polynucleotide sequence (e.g. forward complement strand of the template). As a further alternative, 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 strand of the polynucleotide sequence (e.g. forward strand of the template) or the reverse complement strand of the polynucleotide sequence (e.g. reverse complement strand of the template).

20

Alternatively, 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). As a further alternative, 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

25

30

35

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.

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

Accordingly, the method may comprise selectively processing at least one polynucleotide sequence comprising a first portion and a second portion, wherein each polynucleotide sequence is attached to a first immobilised primer. The method may comprise selectively processing a plurality of polynucleotide sequences each comprising a first portion and a second portion, wherein each polynucleotide sequence is attached to a first immobilised primer.

Alternatively, the present invention can be applied to (separate) polynucleotide strands where a first strand comprises a first portion to be identified and a second strand comprises a second portion to be identified.

The first portions and second portions may be different polynucleotide sequences. That is, the sequences may be genetically unrelated and/or derived from different sources.

Alternatively, the first portions and second portions may be genetically related.

For example, the (separate) polynucleotide strands may comprise a first strand that comprises a first portion that may comprise (or be) the forward strand of a polynucleotide sequence (e.g. forward strand of a template), and a second strand that comprises a second portion that may comprise (or be) the reverse strand of the polynucleotide sequence (e.g. reverse strand of the template) or the forward complement strand of the polynucleotide sequence (e.g. forward complement strand of the template). As a further

alternative, the (separate) polynucleotide strands may comprise a first strand that comprises a first portion that may comprise (or be) the reverse strand of a polynucleotide sequence (e.g. reverse strand of a template), and a second strand that comprises a second portion that may comprise (or be) the forward strand of the polynucleotide sequence (e.g. forward strand of the template) or the reverse complement strand of the polynucleotide sequence (e.g. reverse complement strand of the template).

Alternatively, the (separate) polynucleotide strands may comprise a first strand that comprises a first portion that may comprise (or be) the forward strand of a polynucleotide sequence (e.g. forward strand of a template), and a second strand that comprises a second portion that 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). As a further alternative, the (separate) polynucleotide strands may comprise a first strand that comprises a first portion that may comprise (or be) the reverse strand of a polynucleotide sequence (e.g. reverse strand of a template), and a second strand that comprises a second portion that 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 strand 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.

Again, 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 some embodiments relating to (separate) polynucleotide strands, the first and second strand may be separately attached to a solid support. This solid support may be a flow cell. For example, each of the first and second strands may be attached to the solid support (e.g. flow cell) in a single well of the solid support.

Accordingly, the method may comprise selectively processing at least one first polynucleotide sequence comprising a first portion and at least one second polynucleotide sequence comprising a second portion, wherein each first polynucleotide sequence is attached to a first immobilised primer, and each second polynucleotide sequence is attached to a second immobilised primer. The method may comprise selectively processing a plurality of first polynucleotide sequences each comprising a first portion and a plurality of at least one second polynucleotide sequences each comprising a second portion, wherein each first polynucleotide sequence is attached to a first immobilised primer, and each second polynucleotide sequence is attached to a second immobilised primer.

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”.

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

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

The selective processing may result 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. It is this altered ratio that primes the first portions and second portions to be ready for concurrent sequencing.

The ratio may be between 1.25:1 to 5:1, or between 1.5:1 to 3:1, or about 2:1.

5

The first signal and the second signal may be spatially unresolved (e.g. generated from the same region or substantially overlapping regions). A first region may be occupied by the at least one first polynucleotide sequence comprising the first portion within the duoclonal cluster is the same as, or substantially overlapping with, a second region occupied by the at least one second polynucleotide sequence comprising the second portion within the duoclonal cluster.

10

Selective processing may refer to conducting selective sequencing. Alternatively, selective processing may refer to preparing for selective sequencing. As shown in Figure 27, in one example, selective sequencing may be achieved using a mixture of unblocked and blocked sequencing primers.

15

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.

20

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

25

30

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

35

primer binding site and a second sequencing primer binding site, or to embodiments where the first polynucleotide strand comprises a first sequencing primer binding site, and the second polynucleotide strand comprises a second sequencing primer binding site. In other embodiments, the binding of first sequencing primers and second sequencing primers may not be applied to cases where the first polynucleotide strand comprises a first sequencing primer binding site, and the second polynucleotide strand comprises a second sequencing primer binding site. This is achieved using a mixed population of blocked and unblocked second sequencing primers that bind the second sequencing primer site. Any ratio of blocked:unblocked second primers can be used that generates a second signal that is of a lower intensity than the first signal, for example, the ratio of blocked:unblocked primers may be: 20:80 to 80:20, such as 1:2 to 2:1.

For example, a ratio of 50:50 of blocked:unblocked second primers may be 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-(CH₂)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.

Accordingly, in an aspect of the invention, there is provided a sequencing primer comprising or consisting of a sequence selected from SEQ ID NO: 11 to 16 or a variant or fragment thereof.

5 In another aspect of the invention there is provided a sequencing composition (also referred to herein as a sequencing mix), comprising a blocked second sequencing primer selected from SEQ ID NO: 15 and 16 or a variant or fragment thereof, and an unblocked second sequencing primer selected from SEQ ID NO: 13 and 14, or a variant or fragment thereof. In one embodiment, the sequencing composition comprises a blocked
10 sequencing primer selected from SEQ ID NO: 15 or a variant or fragment thereof, and an unblocked sequencing primer selected from SEQ ID NO: 13 or a variant or fragment thereof. In another embodiment, the sequencing composition comprises a blocked sequencing primer selected from SEQ ID NO: 16 or a variant or fragment thereof, and an unblocked sequencing primer selected from SEQ ID NO: 14, or a variant or fragment
15 thereof.

In some embo, 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
20 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 aspect,
25 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.

In another aspect of the invention, there is provided the use of a blocked sequencing
30 primer, such as a blocked sequencing primer comprising SEQ ID NO: 11 to 16 or a variant or fragment thereof in preparing at least one polynucleotide sequence, for example, a plurality of polynucleotide sequences, for identification.

As shown in Figure 27, selective sequencing may be conducted on the amplified
35 (monoclonal) cluster shown in Figure 26F. A plurality of first sequencing primers 1501 are added. These first sequencing primers 1501 (e.g. B15-ME; or if ME is not present, then B15) anneal to the first terminal sequencing primer binding site 1303 (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 1502a and a plurality of second blocked sequencing primers 1502b are added, either at the same time as the first sequencing primers 1501, or sequentially (e.g. prior to or after addition of first sequencing primers 1501). These second unblocked sequencing primers 1502a (e.g. HYB2-ME; or if ME is not present, then HYB2) and second blocked sequencing primers 1502b (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 1403' (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 1401' (i.e. "first portions") to be sequenced and the second insert complement sequences 1402' (i.e. "second portions") to be sequenced, wherein a greater proportion of first insert complement sequences 1401' are sequenced (grey arrow) compared to a proportion of second insert complement sequences 1402' (black arrow).

Although Figure 27 shows selective sequencing being conducted on a template strand attached to first immobilised primer 1201, in some embodiments the (monoclonal) cluster may instead have template strands attached to second immobilised primer 1202. 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 27 shows concurrent sequencing of a concatenated strand according to the above method. As shown in Figure 27, 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.

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

5

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

15

Accordingly, in this example, selective removal of some or substantially all free second immobilised primers is carried out before at least one further round of bridge amplification and before any sequencing of the target regions. As a consequence, the ratio of first polynucleotide capable of generating a first signal to the second polynucleotide that is capable of generating a second signal is altered, which in turn leads to two signals of different intensities, permitting concurrent sequencing of both sequences (or the target regions within those sequences).

20

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

25

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

30

As described above, and described in further detail in WO 2008/041002, the first and second immobilised primers may be attached to the surface of a solid support through a linker. The linker may be different for the first and second immobilised primers. The linker may be any cleavable linker; that is the linker may comprise one or more moieties, such as modified nucleotides, that enable selective cleavage of the immobilised primer from the surface of the solid support. By way of non-limiting example, the linker may comprise uracil bases, phosphorothioate groups, ribonucleotides, diol linkages, disulphide

35

linkages, peptides etc. which may be included, not only to allow covalent attachment to a solid support, but also to allow selective cleavage of the linker.

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

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

5'-PS-TTTTTTTTTTAAATGATACGGCGACCACCGAUCTACAC-3' where U = 2-deoxyuridine (SEQ ID NO: 23).

15 In another example, the second immobilised primer is attached to a solid support through a second linker, where the linker comprises uracil or 2-deoxyuridine. In this example, free second immobilised primers (that is, primers that are not extended) can be removed using uracil glycosylase. In some embodiments, free second immobilised primers can be removed using a USER enzyme mix (which is a cocktail of uracil glycosylase and endonuclease VIII). In one example, the sequence of the second immobilised primer
20 comprises the following sequence or a variant of fragment thereof:

5'-PS-TTTTTTTTTTCAAGCAGAAGACGGCATACGA[G^{oxo}]AT-3', where [G^{oxo}] = 8-oxoguanine (SEQ ID NO: 24).

25 Accordingly, in a further aspect of the invention, there is provided an amplification mixture comprising a recombinase, a DNA polymerase, a single-stranded DNA binding protein (SSB) and a glycosylase, wherein the glycosylase is either FPG glycosylase or uracil glycosylase or the USER enzyme mix.

30 One example of this method is shown in Figure 28. Selective amplification may be conducted on the amplified (duoclonal) cluster as shown in Figure 25H. The solid support 1200 comprises free first immobilised primers 1201 and free second immobilised primers 1202. Free second immobilised primers 1202 are cleaved from the solid support 1200, thus leaving behind free first immobilised primers 1201 (Figure 28A).

35 The first primer-binding sequence 1301' (e.g. P5') on one set of template strands may then anneal to the free first immobilised primers 1201 (e.g. P5 lawn primer) located within

the well 1203. By contrast, since free second immobilised primers 1202 (e.g. P7 lawn primer) have been removed, second primer-binding sequences 1302' (e.g. P7') are not able to anneal (Figure 28B).

5 After conducting a cycle of bridge amplification, this leads to selective amplification of the template strands comprising the forward strand of the template 1101' and the first terminal sequencing primer binding site 1303, relative to the template strands comprising the forward complement strand of the template 1101 and the second terminal sequencing primer binding site 1304 (Figure 28C).

10

Conducting standard (non-selective) sequencing then allows the forward strands of the template 1101' (i.e. "first portions") to be sequenced and the forward complement strands of the template 1101 (i.e. "second portions") to be sequenced, wherein a greater proportion of forward strands of the template 1101' are sequenced (grey arrow) compared to a proportion of forward complement strands of the template 1101 (black arrow) (Figure 28D).

15

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

20

25

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

30

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

35

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

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

Again, by "blocked" is meant that the sequencing primer comprises a blocking group at a 3' end of the sequencing primer. Suitable blocking groups include a hairpin loop (e.g. a polynucleotide attached to the 3'-end, comprising in a 5' to 3' direction, a cleavable site such as a nucleotide comprising uracil, a loop portion, and a complement portion, wherein the complement portion is substantially complementary to all or a portion of the immobilised primer), a deoxynucleotide, a deoxyribonucleotide, a hydrogen atom instead of a 3'-OH group, a phosphate group, a phosphorothioate group, a propyl spacer (e.g. -O-(CH₂)₃-OH instead of a 3'-OH group)), a modification blocking the 3'-hydroxyl group (e.g. hydroxyl protecting groups, such as silyl ether groups (e.g. trimethylsilyl, triethylsilyl, triisopropylsilyl, t-butyl(dimethyl)silyl, t-butyl(diphenyl)silyl), ether groups (e.g. benzyl, allyl, t-butyl, methoxymethyl (MOM), 2-methoxyethoxymethyl (MEM), tetrahydropyranyl), or acyl groups (e.g. acetyl, benzoyl)), or an inverted nucleobase. However, the blocking group may be any modification that prevents extension (i.e. elongation) of the primer by a polymerase. The block may be reversible or irreversible.

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

One example of this method is shown in Figure 29. Selective amplification may be conducted on the amplified (duoclonal) cluster as shown in Figure 25H. The first primer-binding sequence 1301' (e.g. P5') on one set of template strands may anneal to first immobilised primers 1201 (e.g. P5 lawn primer), and the second primer-binding

sequence 1302' (e.g. P7') on another set of template strands may anneal to second immobilised primers 1202 (e.g. P7 lawn primer) (Figure 29A).

5 Whilst the second primer-binding sequence 1302' (e.g. P7') is annealed to the second immobilised primer 1202, a primer-blocking agent 1601 is selectively installed onto a 3'-end of the second immobilised primer 1202, whilst no installation occurs to the 3'-end of the first immobilised primer 1201 (Figure 29B).

10 After conducting a cycle of bridge amplification, this leads to selective amplification of the template strands comprising the forward strand of the template 1101' and the first terminal sequencing primer binding site 1303, relative to the template strands comprising the forward complement strand of the template 1101 and the second terminal sequencing primer binding site 1304. The primer-blocking agent 1601 prevents extension from the second immobilised primer 1202. (Figure 29C).

15 Conducting standard (non-selective) sequencing then allows the forward strands of the template 1101' (i.e. "first portions") to be sequenced and the forward complement strands of the template 1101 (i.e. "second portions") to be sequenced, wherein a greater proportion of forward strands of the template 1101' are sequenced (grey arrow) compared to a proportion of forward complement strands of the template 1101 (black arrow) (Figure 29D).

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

30 The extended primer sequences may be substantially complementary to the first or second immobilised primers (e.g. P5 or P7), or substantially complementary to a portion of the first or second immobilised primer.

35 The 5' additional nucleotide may be selected from A, T, C or G, but may be T (or U) or C. In some aspects, the 5' additional nucleotide is not a complement of the 3' nucleotide of the second immobilised primer (where the extended primer sequence binds the first immobilised primer) or is not a complement of the 3' nucleotide of the first immobilised

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

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

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

In one embodiment, the extended primer sequence is selected from SEQ ID NO: 25 to 36 or a variant or fragment thereof.

One example of this method is shown in Figure 30. Selective amplification may be conducted on the amplified (duoclonal) cluster as shown in Figure 25H; as such following a number of rounds of amplification, a cluster is formed comprising both extended first (e.g. P5) and second (e.g. P7) immobilised polynucleotide strands. Before the next round of amplification, a (or a plurality of) extended primer sequence(s) is flowed across the surface of the solid support 1200. The extended primer sequence 1701 is substantially complementary to at least a portion, if not all of the immobilised primer (e.g. either P5 or P7) and binds to the immobilised primer (e.g. P5 or P7) as shown in Figure 30A. As also shown in Figure 30A, the extended primer sequence 1701 comprises at least one additional 5' nucleotide.

Following addition of the extended primer sequence 1701, a primer blocking agent 1601 is added and flowed across the surface of the solid support (e.g. flow cell). As the primer-blocking agent 1601 is complementary to the 5' additional nucleotide of the extended primer sequence 1701 the primer-blocking agent 1601 binds to the 3'-end of the immobilised strands that are hybridised to the extended primer sequence 1701, as

shown in Figure 30B. As a consequence, addition of the primer-blocking agent 1601 prevents not only extension of the immobilised strand (e.g. P5 or P7) but renders the immobilised primer (P5 or P7) unavailable for hybridisation and subsequent bridge amplification for other extended strands (e.g. 1101') (see Figure 30B).

5

Performing at least one more cycle of bridge amplification, leads to selective amplification of the template strands comprising the forward strand of the template 1101' (in a 2:1 ratio of 1101' to 1101). Again, similar to Figure 25D, conducting standard (non-selective) sequencing then allows the forward strands of the template 1101' (i.e. "first portions") to be sequenced and the forward complement strands of the template 1101 (i.e. "second portions") to be sequenced, wherein a greater proportion of forward strands of the template 1101' are sequenced (grey arrow) compared to a proportion of forward complement strands of the template 1101 (black arrow) (Figure 25D).

10

15

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

20

Accordingly, in a further aspect of the invention, there is provided an extended primer sequence comprising a sequence selected from SEQ ID NO: 25 to 35 or a variant or fragment thereof.

25

In a further aspect of the invention, there is provided an amplification composition comprising a recombinase, a DNA polymerase, a single-stranded DNA binding protein (SSB) and at least one blocked immobilised-primer-binding sequence.

30

By "amplification composition" is meant a composition that is suitable for the amplification of a target nucleic acid template.

In another aspect of the invention, there is provided the use of a blocked immobilised-primer-binding sequence, for example a blocked immobilised primer-binding sequence comprising a sequence selected from SEQ ID NO: 25 to 35, in preparing at least one polynucleotide sequence for identification.

35

Samples

In some embodiments, the sample comprises or consists of a purified or isolated polynucleotide derived from a tissue sample, a biological fluid sample, a cell sample, and the like. Suitable biological fluid samples include, but are not limited to blood, plasma, serum, sweat, tears, sputum, urine, sputum, ear flow, lymph, saliva, cerebrospinal fluid, rashes, bone marrow suspension, vaginal flow, trans-cervical lavage, brain fluid, ascites, milk, secretions of the respiratory, intestinal and genitourinary tracts, amniotic fluid, milk, and leukopheresis samples. In some embodiments, the sample is a sample that is easily obtainable by non-invasive procedures, e.g., blood, plasma, serum, sweat, tears, sputum, urine, sputum, ear flow, saliva or feces. In certain embodiments the sample is a peripheral blood sample, or the plasma and/or serum fractions of a peripheral blood sample. In other embodiments, the biological sample is a swab or smear, a biopsy specimen, or a cell culture. In another embodiment, the sample is a mixture of two or more biological samples, e.g., a biological sample can comprise two or more of a biological fluid sample, a tissue sample, and a cell culture sample. As used herein, the terms "blood," "plasma" and "serum" expressly encompass fractions or processed portions thereof. Similarly, where a sample is taken from a biopsy, swab, smear, etc., the "sample" expressly encompasses a processed fraction or portion derived from the biopsy, swab, smear, etc.

In certain embodiments, samples can be obtained from sources, including, but not limited to, samples from different individuals, samples from different developmental stages of the same or different individuals, samples from different diseased individuals (e.g., individuals with cancer or suspected of having a genetic disorder), normal individuals, samples obtained at different stages of a disease in an individual, samples obtained from an individual subjected to different treatments for a disease, samples from individuals subjected to different environmental factors, samples from individuals with predisposition to a pathology, samples individuals with exposure to an infectious disease agent, and the like.

In one illustrative, but non-limiting embodiment, the sample is a maternal sample that is obtained from a pregnant female, for example a pregnant woman. The maternal sample can be a tissue sample, a biological fluid sample, or a cell sample. In another illustrative, but non-limiting embodiment, the maternal sample is a mixture of two or more biological samples, e.g., the biological sample can comprise two or more of a biological fluid sample, a tissue sample, and a cell culture sample.

In certain embodiments samples can also be obtained from in vitro cultured tissues, cells, or other polynucleotide-containing sources. The cultured samples can be taken from sources including, but not limited to, cultures (e.g., tissue or cells) maintained in different media and conditions (e.g., pH, pressure, or temperature), cultures (e.g., tissue or cells) maintained for different periods of length, cultures (e.g., tissue or cells) treated with different factors or reagents (e.g., a drug candidate, or a modulator), or cultures of different types of tissue and/or cells.

In some embodiments, the use of the disclosed sequencing technology does not involve the preparation of sequencing libraries. In other embodiments, the sequencing technology contemplated herein involve the preparation of sequencing libraries. In one illustrative approach, sequencing library preparation involves the production of a random collection of adapter-modified DNA fragments (e.g., polynucleotides) that are ready to be sequenced.

Sequencing libraries of polynucleotides can be prepared from DNA or RNA, including equivalents, analogs of either DNA or cDNA, for example, DNA or cDNA that is complementary or copy DNA produced from an RNA template, by the action of reverse transcriptase. The polynucleotides may originate in double-stranded form (e.g., dsDNA such as genomic DNA fragments, cDNA, PCR amplification products, and the like) or, in certain embodiments, the polynucleotides may originated in single-stranded form (e.g., ssDNA, RNA, etc.) and have been converted to dsDNA form. By way of illustration, in certain embodiments, single stranded mRNA molecules may be copied into double-stranded cDNAs suitable for use in preparing a sequencing library. The precise sequence of the primary polynucleotide molecules is generally not material to the method of library preparation, and may be known or unknown. In one embodiment, the polynucleotide molecules are DNA molecules. More particularly, in certain embodiments, the polynucleotide molecules represent the entire genetic complement of an organism or substantially the entire genetic complement of an organism, and are genomic DNA molecules (e.g., cellular DNA, cell free DNA (cfDNA), etc.), that typically include both intron sequence and exon sequence (coding sequence), as well as non-coding regulatory sequences such as promoter and enhancer sequences. In certain embodiments, the primary polynucleotide molecules comprise human genomic DNA molecules, e.g., cfDNA molecules present in peripheral blood of a pregnant subject.

Methods of isolating nucleic acids from biological sources may differ depending upon the nature of the source. One of skill in the art can readily isolate nucleic acids from a source as needed for the method described herein. In some instances, it can be advantageous to fragment large nucleic acid molecules (e.g. cellular genomic DNA) in the nucleic acid sample to obtain polynucleotides in the desired size range. Fragmentation can be random, or it can be specific, as achieved, for example, using restriction endonuclease digestion. Methods for random fragmentation may include, for example, limited DNase digestion, alkali treatment and physical shearing. Fragmentation can also be achieved by any of a number of methods known to those of skill in the art. For example, fragmentation can be achieved by mechanical means including, but not limited to nebulization, sonication and hydroshear.

In some embodiments, sample nucleic acids are obtained from as cfDNA, which is not subjected to fragmentation. For example, cfDNA, typically exists as fragments of less than about 300 base pairs and consequently, fragmentation is not typically necessary for generating a sequencing library using cfDNA samples.

Typically, whether polynucleotides are forcibly fragmented (e.g., fragmented in vitro), or naturally exist as fragments, they are converted to blunt-ended DNA having 5'-phosphates and 3'-hydroxyl. Standard protocols, e.g., protocols for sequencing using, for example, the Illumina platform, instruct users to end-repair sample DNA, to purify the end-repaired products prior to dA-tailing, and to purify the dA-tailing products prior to the adaptor-ligating steps of the library preparation.

In various embodiments, verification of the integrity of the samples and sample tracking can be accomplished by sequencing mixtures of sample genomic nucleic acids, e.g., cfDNA, and accompanying marker nucleic acids that have been introduced into the samples, e.g., prior to processing.

30 Computing Systems

In some embodiments, the disclosed systems and methods may involve approaches for shifting or distributing certain sequence data analysis features and sequence data storage to a cloud computing environment or cloud-based network. User interaction with sequencing data, genome data, or other types of biological data may be mediated via a central hub that stores and controls access to various interactions with the data. In some embodiments, the cloud computing environment may also provide sharing of protocols,

analysis methods, libraries, sequence data as well as distributed processing for sequencing, analysis, and reporting. In some embodiments, the cloud computing environment facilitates modification or annotation of sequence data by users. In some embodiments, the systems and methods may be implemented in a computer browser,
5 on-demand or on-line.

In some embodiments, software written to perform the methods as described herein is stored in some form of computer readable medium, such as memory, CD ROM, DVD-ROM, memory stick, flash drive, hard drive, SSD hard drive, server, mainframe storage
10 system and the like.

In some embodiments, the methods may be written in any of various suitable programming languages, for example compiled languages such as C, C#, C++, Fortran, and Java. Other programming languages could be script languages, such as Perl,
15 MatLab, SAS, SPSS, Python, Ruby, Pascal, Delphi, R and PHP. In some embodiments, the methods are written in C, C#, C++, Fortran, Java, Perl, R, Java or Python. In some embodiments, the method may be an independent application with data input and data display modules. Alternatively, the method may be a computer software product and may include classes wherein distributed objects comprise applications including
20 computational methods as described herein.

In some embodiments, the methods may be incorporated into pre-existing data analysis software, such as that found on sequencing instruments. Software comprising computer implemented methods as described herein are installed either onto a computer system
25 directly, or are indirectly held on a computer readable medium and loaded as needed onto a computer system. Further, the methods may be located on computers that are remote to where the data is being produced, such as software found on servers and the like that are maintained in another location relative to where the data is being produced, such as that provided by a third party service provider.

30
An assay instrument, desktop computer, laptop computer, or server which may contain a processor in operational communication with accessible memory comprising instructions for implementation of systems and methods. In some embodiments, a desktop computer or a laptop computer is in operational communication with one or more
35 computer readable storage media or devices and/or outputting devices. An assay instrument, desktop computer and a laptop computer may operate under a number of different computer based operational languages, such as those utilized by Apple based

- computer systems or PC based computer systems. An assay instrument, desktop and/or laptop computers and/or server system may further provide a computer interface for creating or modifying experimental definitions and/or conditions, viewing data results and monitoring experimental progress. In some embodiments, an outputting device may be
- 5 a graphic user interface such as a computer monitor or a computer screen, a printer, a hand-held device such as a personal digital assistant (i.e., PDA, Blackberry, iPhone), a tablet computer (e.g., iPad), a hard drive, a server, a memory stick, a flash drive and the like.
- 10 A computer readable storage device or medium may be any device such as a server, a mainframe, a supercomputer, a magnetic tape system and the like. In some embodiments, a storage device may be located onsite in a location proximate to the assay instrument, for example adjacent to or in close proximity to, an assay instrument. For example, a storage device may be located in the same room, in the same building,
- 15 in an adjacent building, on the same floor in a building, on different floors in a building, etc. in relation to the assay instrument. In some embodiments, a storage device may be located off-site, or distal, to the assay instrument. For example, a storage device may be located in a different part of a city, in a different city, in a different state, in a different country, etc. relative to the assay instrument. In embodiments where a storage device is
- 20 located distal to the assay instrument, communication between the assay instrument and one or more of a desktop, laptop, or server is typically via Internet connection, either wireless or by a network cable through an access point. In some embodiments, a storage device may be maintained and managed by the individual or entity directly associated with an assay instrument, whereas in other embodiments a storage device may be
- 25 maintained and managed by a third party, typically at a distal location to the individual or entity associated with an assay instrument. In embodiments as described herein, an outputting device may be any device for visualizing data.
- 30 An assay instrument, desktop, laptop and/or server system may be used itself to store and/or retrieve computer implemented software programs incorporating computer code for performing and implementing computational methods as described herein, data for use in the implementation of the computational methods, and the like. One or more of an assay instrument, desktop, laptop and/or server may comprise one or more computer readable storage media for storing and/or retrieving software programs incorporating
- 35 computer code for performing and implementing computational methods as described herein, data for use in the implementation of the computational methods, and the like. Computer readable storage media may include, but is not limited to, one or more of a

hard drive, a SSD hard drive, a CD-ROM drive, a DVD-ROM drive, a floppy disk, a tape, a flash memory stick or card, and the like. Further, a network including the Internet may be the computer readable storage media. In some embodiments, computer readable storage media refers to computational resource storage accessible by a computer network via the Internet or a company network offered by a service provider rather than, for example, from a local desktop or laptop computer at a distal location to the assay instrument.

In some embodiments, computer readable storage media for storing and/or retrieving computer implemented software programs incorporating computer code for performing and implementing computational methods as described herein, data for use in the implementation of the computational methods, and the like, is operated and maintained by a service provider in operational communication with an assay instrument, desktop, laptop and/or server system via an Internet connection or network connection.

In some embodiments, a hardware platform for providing a computational environment comprises a processor (i.e., CPU) wherein processor time and memory layout such as random access memory (i.e., RAM) are systems considerations. For example, smaller computer systems offer inexpensive, fast processors and large memory and storage capabilities. In some embodiments, graphics processing units (GPUs) can be used. In some embodiments, hardware platforms for performing computational methods as described herein comprise one or more computer systems with one or more processors. In some embodiments, smaller computer are clustered together to yield a supercomputer network.

In some embodiments, computational methods as described herein are carried out on a collection of inter- or intra-connected computer systems (i.e., grid technology) which may run a variety of operating systems in a coordinated manner. For example, the CONDOR framework (University of Wisconsin-Madison) and systems available through United Devices are exemplary of the coordination of multiple stand-alone computer systems for the purpose dealing with large amounts of data. These systems may offer Perl interfaces to submit, monitor and manage large sequence analysis jobs on a cluster in serial or parallel configurations.

Definitions

Unless defined otherwise, technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which the present disclosure belongs. See, e.g. Singleton et al., Dictionary of Microbiology and Molecular Biology 2nd ed., J. Wiley & Sons (New York, NY 1994); Sambrook et al., Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Press (Cold Spring Harbor, NY 1989). For purposes of the present disclosure, the following terms are defined below.

As used herein, the term “cluster” or “clump” refers to a group of molecules, e.g., a group of DNA, or a group of signals. In some embodiments, the signals of a cluster are derived from different features. In some embodiments, a signal clump represents a physical region covered by one amplified oligonucleotide. In various examples, a physical region may be a tile, a sub-tile, a lane or a sub-lane on a flow cell, etc. Each signal clump could be ideally observed as several signals. Accordingly, duplicate signals could be detected from the same clump of signals. In some embodiments, a cluster or clump of signals can comprise one or more signals or spots that correspond to a particular feature. When used in connection with microarray devices or other molecular analytical devices, a cluster can comprise one or more signals that together occupy the physical region occupied by an amplified oligonucleotide (or other polynucleotide or polypeptide with a same or similar sequence). For example, where a feature is an amplified oligonucleotide, a cluster can be the physical region covered by one amplified oligonucleotide. In other embodiments, a cluster or clump of signals need not strictly correspond to a feature. For example, spurious noise signals may be included in a signal cluster but not necessarily be within the feature area. For example, a cluster of signals from four cycles of a sequencing reaction could comprise at least four signals.

As used herein, a “flow cell” can include a device having a lid extending over a reaction structure to form a flow channel therebetween that is in communication with a plurality of reaction sites of the reaction structure, and can include a detection device that is configured to detect designated reactions that occur at or proximate to the reaction sites. A flow cell may include a solid-state light detection or “imaging” device, such as a Charge- Coupled Device (CCD) or Complementary Metal-Oxide Semiconductor (CMOS) (light) detection device. As one specific example, a flow cell may be configured to fluidically and electrically couple to a cartridge (having an integrated pump), which may be configured to fluidically and/or electrically couple to a bioassay system. A cartridge and/or bioassay system may deliver a reaction solution to reaction sites of a flow cell

according to a predetermined protocol (e.g., sequencing-by-synthesis), and perform a plurality of imaging events. For example, a cartridge and/or bioassay system may direct one or more reaction solutions through the flow channel of the flow cell, and thereby along the reaction sites. At least one of the reaction solutions may include four types of nucleotides having the same or different fluorescent labels. The nucleotides may bind to the reaction sites of the flow cell, such as to corresponding oligonucleotides at the reaction sites. The cartridge and/or bioassay system may then illuminate the reaction sites using an excitation light source (e.g., solid-state light sources, such as light-emitting diodes (LEDs)). The excitation light may have a predetermined wavelength or wavelengths, including a range of wavelengths. The fluorescent labels excited by the incident excitation light may provide emission signals (e.g., light of a wavelength or wavelengths that differ from the excitation light and, potentially, each other) that may be detected by the light sensors of the flow cell.

Flow cells described herein may be configured to perform various biological or chemical processes. More specifically, the flow cells described herein may be used in various processes and systems where it is desired to detect an event, property, quality, or characteristic that is indicative of a designated reaction. For example, flow cells described herein may include or be integrated with light detection devices, biosensors, and their components, as well as bioassay systems that operate with biosensors. The flow cells may be configured to facilitate a plurality of designated reactions that may be detected individually or collectively. The flow cells may be configured to perform numerous cycles in which the plurality of designated reactions occurs in parallel. For example, the flow cells may be used to sequence a dense array of DNA features through iterative cycles of enzymatic manipulation and light or image detection/acquisition. As such, the flow cells may be in fluidic communication with one or more microfluidic channels that deliver reagents or other reaction components in a reaction solution to a reaction site of the flow cells. The reaction sites may be provided or spaced apart in a predetermined manner, such as in a uniform or repeating pattern. Alternatively, the reaction sites may be randomly distributed. Each of the reaction sites may be associated with one or more light guides and one or more light sensors that detect light from the associated reaction site. In one example, light guides include one or more filters for filtering certain wavelengths of light. The light guides may be, for example, an absorption filter (e.g., an organic absorption filter) such that the filter material absorbs a certain wavelength (or range of wavelengths) and allows at least one predetermined wavelength (or range of wavelengths) to pass therethrough. In some flow cells, the reaction sites

may be located in reaction recesses or chambers, which may at least partially compartmentalize the designated reactions therein.

5 As used herein, the term “spot radius” or “cluster radius” refers to a defined radius which encompasses a diffraction-limited spot or a cluster of signals. Accordingly, by defining a cluster radius as larger or smaller, a greater number of signals can fall within the radius for subsequent ordering and selection. A cluster radius can be defined by any distance measure, such as pixels, meters, millimeters, or any other useful measure of distance.

10 As used herein, a “signal” refers to a detectable event such as an emission, such as light emission, for example, in an image. Thus, in some embodiments, a signal can represent any detectable light emission that is captured in an image (i.e., a “spot”). Thus, as used herein, “signal” can refer to an actual emission from a feature of the specimen, or can refer to a spurious emission that does not correlate to an actual feature. Thus, a signal
15 could arise from noise and could be later discarded as not representative of an actual feature of a specimen.

As used herein, an “intensity” of an emitted light refers to the intensity of the light transferred per unit area, where the area is measured on the plane perpendicular to the
20 direction of propagation of the light ray, and where the intensity is the amount of energy transferred per unit time. In some embodiments, signal “strength”, “amplitude”, “magnitude” or “level” may be used synonymously with signal intensity. In some embodiments, an image taken by a detector is approximately or proportional to an intensity map integrated over some amount of time. In some embodiments, the signal of
25 a diffraction-limited spot of a DNA cluster is extracted from the image as the total intensity included in the spot, up to a factor of the integration time. For example, the signal of a DNA cluster may be defined as the intensity included within the spot radius of the DNA cluster, up to a factor of the integration time. In other embodiments, the peak intensity value found within the spot radius may be used to represent the signal of the DNA cluster,
30 up to a factor of the integration time.

As used herein, the process of aligning the template of signal positions onto a given image is referred to as “registration”, and the process for determining an intensity value or an amplitude value for each signal in the template for a given image is referred to as
35 “intensity extraction”. For registration, the methods and systems provided herein may take advantage of the random nature of signal clump positions by using image correlation to align the template to the image.

As used herein, a "nucleotide" includes a nitrogen containing heterocyclic base, a sugar, and one or more phosphate groups. Nucleotides are monomeric units of a nucleic acid sequence. Examples of nucleotides include, for example, ribonucleotides or deoxyribonucleotides. In ribonucleotides (RNA), the sugar is a ribose, and in deoxyribonucleotides (DNA), the sugar is a deoxyribose, i.e., a sugar lacking a hydroxyl group that is present at the 2' position in ribose. The nitrogen containing heterocyclic base can be a purine base or a pyrimidine base. Purine bases include adenine (A) and guanine (G), and modified derivatives or analogs thereof. Pyrimidine bases include cytosine (C), thymine (T), and uracil (U), and modified derivatives or analogs thereof. The C-1 atom of deoxyribose is bonded to N-1 of a pyrimidine or N-9 of a purine. The phosphate groups may be in the mono-, di-, or tri-phosphate form. These nucleotides may be natural nucleotides, but it is to be further understood that non-natural nucleotides, modified nucleotides or analogs of the aforementioned nucleotides can also be used.

As used herein, "nucleobase" is a heterocyclic base such as adenine, guanine, cytosine, thymine, uracil, inosine, xanthine, hypoxanthine, or a heterocyclic derivative, analog, or tautomer thereof. A nucleobase can be naturally occurring or synthetic. Non-limiting examples of nucleobases are adenine, guanine, thymine, cytosine, uracil, xanthine, hypoxanthine, 8-azapurine, purines substituted at the 8 position with methyl or bromine, 9-oxo-N6-methyladenine, 2-aminoadenine, 7-deazaxanthine, 7-deazaguanine, 7-deaza-adenine, N4-ethanocytosine, 2,6- diaminopurine, N6-ethano-2,6-diaminopurine, 5- methylcytosine, 5-(C3-C6)- alkynylcytosine, 5-fluorouracil, 5-bromouracil, thiouracil, pseudoisocytosine, 2-hydroxy-5-methyl-4-triazolopyridine, isocytosine, isoguanine, inosine, 7,8-dimethylalloxazine, 6-dihydrothymine, 5,6-dihydrouracil, 4-methyl-indole, ethenoadenine and the non-naturally occurring nucleobases described in U.S. Pat. Nos. 5,432,272 and 6,150,510 and PCT applications WO 92/002258, WO 93/10820, WO 94/22892, and WO 94/24144, and Fasman ("Practical Handbook of Biochemistry and Molecular Biology", pp. 385-394, 1989, CRC Press, Boca Raton, LO), all herein incorporated by reference in their entireties.

The term "nucleic acid" or "polynucleotide" refers to a deoxyribonucleotide or ribonucleotide polymer in either single- or double-stranded form, and unless otherwise limited, encompasses known analogs of natural nucleotides that hybridize to nucleic acids in manner similar to naturally occurring nucleotides, such as peptide nucleic acids (PNAs) and phosphorothioate DNA. Unless otherwise indicated, a particular nucleic acid sequence includes the complementary sequence thereof. Nucleotides include, but are

not limited to, ATP, dATP, CTP, dCTP, GTP, dGTP, UTP, TTP, dUTP, 5-methyl-CTP, 5-methyl-dCTP, ITP, dITP, 2-amino-adenosine-TP, 2-amino-deoxyadenosine-TP, 2-thiothymidine triphosphate, pyrrolo-pyrimidine triphosphate, and 2-thiocytidine, as well as the alphathiotriphosphates for all of the above, and 2'-O-methyl-ribonucleotide triphosphates for all the above bases. Modified bases include, but are not limited to, 5-Br-UTP, 5-Br-dUTP, 5-F-UTP, 5-F-dUTP, 5-propynyl dCTP, and 5-propynyl-dUTP.

The polymerase used is an enzyme generally for joining 3'-OH 5'- triphosphate nucleotides, oligomers, and their analogs. Polymerases include, but are not limited to, DNA-dependent DNA polymerases, DNA-dependent RNA polymerases, RNA-dependent DNA polymerases, RNA-dependent RNA polymerases, T7 DNA polymerase, T3 DNA polymerase, T4 DNA polymerase, T7 RNA polymerase, T3 RNA polymerase, SP6 RNA polymerase, DNA polymerase I, Klenow fragment, *Thermophilus aquaticus* DNA polymerase, Tth DNA polymerase, VentR® DNA polymerase (New England Biolabs), Deep VentR® DNA polymerase (New England Biolabs), Bst DNA Polymerase Large Fragment, Stoeffel Fragment, 90N DNA Polymerase, 90N DNA polymerase, Pfu DNA Polymerase, Tfi DNA Polymerase, Tth DNA Polymerase, RepliPhi Phi29 Polymerase, Tli DNA polymerase, eukaryotic DNA polymerase beta, telomerase, Terminator™ polymerase (New England Biolabs), KOD HiFi™ DNA polymerase (Novagen), KOD1 DNA polymerase, Q-beta replicase, terminal transferase, AMV reverse transcriptase, M-MLV reverse transcriptase, Phi6 reverse transcriptase, HIV-1 reverse transcriptase, novel polymerases discovered by bioprospecting, and polymerases cited in US 2007/0048748, US 6,329,178, US 6,602,695, and US 6,395,524 (incorporated by reference). These polymerases include wild-type, mutant isoforms, and genetically engineered variants. "Encode" or "parse" are verbs referring to transferring from one format to another, and refers to transferring the genetic information of target template base sequence into an arrangement of reporters.

Nucleosides and nucleotides may be labeled at sites on the sugar or nucleobase. A dye may be attached to any position on the nucleotide base, for example, through a linker. In particular embodiments, Watson-Crick base pairing can still be carried out for the resulting analog. Particular nucleobase labeling sites include the C5 position of a pyrimidine base or the C7 position of a 7-deaza purine base. A linker group may be used to covalently attach a dye to the nucleoside or nucleotide. As used herein, the term "covalently attached" or "covalently bonded" refers to the forming of a chemical bonding that is characterized by the sharing of pairs of electrons between atoms. For example, a covalently attached polymer coating refers to a polymer coating that forms chemical

bonds with a functionalized surface of a substrate, as compared to attachment to the surface via other means, for example, adhesion or electrostatic interaction. It will be appreciated that polymers that are attached covalently to a surface can also be bonded via means in addition to covalent attachment.

5

Various different types of linkers having different lengths and chemical properties can be used. The term "linker" encompasses any moiety that is useful to connect one or more molecules or compounds to each other, to other components of a reaction mixture, and/or to a reaction site. For example, a linker can attach a reporter molecule or "label" (e.g., a fluorescent dye) to a reaction component. In certain embodiments, the linker is a member selected from substituted or unsubstituted alkyl (e.g., a 2-5 carbon chain), substituted or unsubstituted heteroalkyl, substituted or unsubstituted aryl, substituted or unsubstituted heteroaryl, substituted or unsubstituted cycloalkyl, and substituted or unsubstituted heterocycloalkyl. In one example, the linker moiety is selected from straight- and branched carbon-chains, optionally including at least one heteroatom (e.g., at least one functional group, such as ether, thioether, amide, sulfonamide, carbonate, carbamate, urea and thiourea), and optionally including at least one aromatic, heteroaromatic or non-aromatic ring structure (e.g., cycloalkyl, phenyl). In certain embodiments, molecules that have trifunctional linkage capability are used, including, but are not limited to, cynuric chloride, mealamine, diaminopropanoic acid, aspartic acid, cysteine, glutamic acid, pyroglutamic acid, S- acetylmercaptosuccinic anhydride, carbobenzoxylysine, histine, lysine, serine, homoserine, tyrosine, piperidinyl-1,1-amino carboxylic acid, diaminobenzoic acid, etc. In certain specific embodiments, a hydrophilic PEG (polyethylene glycol) linker is used.

25

In certain embodiments, linkers are derived from molecules which comprise at least two reactive functional groups (e.g., one on each terminus), and these reactive functional groups can react with complementary reactive functional groups on the various reaction components or used to immobilize one or more reaction components at the reaction site. "Reactive functional group," as used herein refers to groups including, but not limited to, olefins, acetylenes, alcohols, phenols, ethers, oxides, halides, aldehydes, ketones, carboxylic acids, esters, amides, cyanates, isocyanates, thiocyanates, isothiocyanates, amines, hydrazines, hydrazones, hydrazides, diazo, diazonium, nitro, nitriles, mercaptans, sulfides, disulfides, sulfoxides, sulfones, sulfonic acids, sulfinic acids, acetals, ketals, anhydrides, sulfates, sulfenic acids isonitriles, amidines, imides, imidates, nitrones, hydroxylamines, oximes, hydroxamic acids thiohydroxamic acids, allenes, ortho esters, sulfites, enamines, ynamines, ureas, pseudoureas,

35

semicarbazides, carbodiimides, carbamates, imines, azides, azo compounds, azoxy compounds, and nitroso compounds. Reactive functional groups also include those used to prepare bioconjugates, e.g., N-hydroxysuccinimide esters, maleimides and the like.

5 Cleavable linkers may be, by way of non-limiting example, electrophilically cleavable linkers, nucleophilically cleavable linkers, photocleavable linkers, cleavable under reductive conditions (for example disulfide or azide containing linkers), oxidative conditions, cleavable via use of safety-catch linkers and cleavable by elimination mechanisms. The use of a cleavable linker to attach the dye compound to a substrate
10 moiety ensures that the label can, if required, be removed after detection, avoiding any interfering signal in downstream steps.

In some embodiments, one or more dye or label molecules may attach to the nucleotide base by non-covalent interactions, or by a combination of covalent and non-covalent
15 interactions via a plurality of intermediating molecules. In one example, a nucleotide or a nucleotide analog, being newly incorporated by the polymerase synthesizing from a target polynucleotide, is initially unlabeled. Then, one or more fluorescent labels may be introduced to the nucleotide or nucleotide analog by binding to labeled affinity reagents containing one or more fluorescent dyes. Uses of unlabeled nucleotides and affinity
20 reagents in sequencing by synthesis have been disclosed in U.S. Publication No. 2013/0079232, which is incorporated herein by reference. For example, one, two, three or each of the four different types of nucleotides (e.g., dATP, dCTP, dGTP and dTTP or dUTP) in the reaction mix may be initially unlabeled. Each of the four types of nucleotides (e.g., dNTPs) may have a 3' hydroxy blocking group to ensure that only a single base
25 can be added by a polymerase to the 3' end of a copy polynucleotide being synthesized from the target polynucleotide. After incorporation of an unlabeled nucleotide, an affinity reagent may be then introduced that specifically binds to the incorporated dNTP to provide a labeled extension product comprising the incorporated dNTP. The affinity reagent may be designed to specifically bind to the incorporated dNTP via antibody-
30 antigen interaction or ligand-receptor interaction, for example. The dNTP may be modified to include a specific antigen, which will pair with a specific antibody included in the corresponding affinity reagent. Thus, one, two, three or each of the four different types of nucleotides may be specifically labeled via their corresponding affinity reagents. In some embodiments, the affinity reagents may include small molecules or protein tags
35 that may bind to a hapten moiety of the nucleotide (such as streptavidin-biotin, anti-DIG and DIG, anti-DNP and DNP), antibody (including but not limited to binding fragments of antibodies, single chain antibodies, bispecific antibodies, and the like), aptamers,

knottins, affimers, or any other known agent that binds an incorporated nucleotide with a suitable specificity and affinity. In some embodiments, the hapten moiety of the unlabeled nucleotide may be attached to the nucleobase through a cleavable linker, which may be cleaved under the same reaction condition as that for removing the 3' blocking group. In some embodiments, one affinity reagent may be labeled with multiple copies of the same fluorescent dye, for example, 1, 2, 3, 4, 5, 6, 8, 10, 12, 15 copies of the same dye. In some embodiments, each affinity reagent may be labeled with a different number of copies of the same fluorescent dye. In some embodiments, a first affinity reagent may be labeled with a first number of a first fluorescent dye, a second affinity reagent may be labeled with a second number of a second fluorescent dye, a third affinity reagent may be labeled with a third number of a third fluorescent dye, and a fourth affinity reagent may be labeled with a fourth number of a fourth fluorescent dye. In some embodiments, each affinity reagent may be labeled with a distinct combination of one or more types of dye, where each type of dye has a certain copy number. In some embodiments, different affinity reagents may be labeled with different dyes that can be excited by the same light source, but each dye will have a distinguishable fluorescent intensity or a distinguishable emission spectrum. In some embodiments, different affinity reagents may be labeled with the same dye in different molar ratios to create measurable differences in their fluorescent intensities.

A nucleotide analog may be attached to or associated with one or more photo-detectable labels to provide a detectable signal. In some embodiments, a photo-detectable label may be a fluorescent compound, such as a small molecule fluorescent label. Fluorescent molecules (fluorophores) suitable as a fluorescent label include, but are not limited to: 1,5 IAEDANS; 1,8-ANS; 4-methylumbelliferone; 5-carboxy-2,7-dichlorofluorescein; 5-carboxyfluorescein (5-FAM); fluorescein amidite (FAM); 5-carboxynaphthofluorescein; tetrachloro-6-carboxyfluorescein (TET); hexachloro-6-carboxyfluorescein (HEX); 2,7-dimethoxy-4,5-dichloro-6-carboxyfluorescein (JOE); VIC®; NED™; tetramethylrhodamine (TMR); 5-carboxytetramethylrhodamine (5-TAMRA); 5-HAT (Hydroxy Tryptamine); 5-hydroxy tryptamine (HAT); 5-ROX (carboxy-X-rhodamine); 6-carboxyrhodamine 6G; 6-JOE; Light Cycler® red 610; Light Cycler® red 640; Light Cycler® red 670; Light Cycler® red 705; 7-amino-4-methylcoumarin; 7-aminoactinomycin D (7-AAD); 7-hydroxy-4-methylcoumarin; 9-amino-6-chloro-2-methoxyacridine; 6-methoxy-N-(4-aminoalkyl)quinolinium bromide hydrochloride (ABQ); Acid Fuchsin; ACMA (9-amino-6-chloro-2-methoxyacridine); Acridine Orange; Acridine Red; Acridine Yellow; Acriflavin; Acriflavin Feulgen SITSA; AFPs-AutoFluorescent Protein-(Quantum Biotechnologies); Texas Red; Texas Red-X

conjugate; Thiadicarbocyanine (DiSC3); Thiazine Red R; Thiazole Orange; Thioflavin 5; Thioflavin S; Thioflavin TCN; Thiolyte; Thiozole Orange; Tinopol CBS (Calcofluor White); TMR; TO-PRO-1; TO-PRO-3; TO-PRO-5; TOTO-1; TOTO-3; TriColor (PE-Cy5); TRITC (TetramethylRhodamine-IsoThioCyanate); True Blue; TruRed; Ultralite; Uranine B; Uvitex

5 SFC; WW 781; X-Rhodamine; X-Rhodamine-5-(and-6)-Isothiocyanate (5(6)-XRITC); Xylene Orange; Y66F; Y66H; Y66W; YO-PRO-1; YO-PRO-3; YOYO-1; interchelating dyes such as YOYO-3, Sybr Green, Thiazole orange; members of the Alexa Fluor® dye series (from Molecular Probes/Invitrogen) which cover a broad spectrum and match the principal output wavelengths of common excitation sources such as Alexa Fluor 350,

10 Alexa Fluor 405, 430, 488, 500, 514, 532, 546, 555, 568, 594, 610, 633, 635, 647, 660, 680, 700, and 750; members of the Cy Dye fluorophore series (GE Healthcare), also covering a wide spectrum such as Cy3, Cy3B, Cy3.5, Cy5, Cy5.5, Cy7; members of the Oyster® dye fluorophores (Denovo Biolabels) such as Oyster-500, -550, -556, 645, 650, 656; members of the DY-Labels series (Dyomics), for example, with maxima of

15 absorption that range from 418 nm (DY-415) to 844 nm (DY-831) such as DY-415, -495, -505, -547, -548, -549, -550, -554, -555, -556, -560, -590, -610, -615, -630, -631, -632, -633, -634, -635, -636, -647, -648, -649, -650, -651, -652, -675, -676, -677, -680, -681, -682, -700, -701, -730, -731, -732, -734, -750, -751, -752, -776, -780, -781, -782, -831, -480XL, -481XL, -485XL, -510XL, -520XL, -521XL; members of the ATTO series of

20 fluorescent labels (ATTO-TEC GmbH) such as ATTO 390, 425, 465, 488, 495, 520, 532, 550, 565, 590, 594, 610, 611X, 620, 633, 635, 637, 647, 647N, 655, 680, 700, 725, 740; members of the CAL Fluor® series or Quasar® series of dyes (Biosearch Technologies) such as CAL Fluor® Gold 540, CAL Fluor® Orange 560, Quasar® 570, CAL Fluor® Red 590, CAL Fluor® Red 610, CAL Fluor® Red 635, Quasar® 570, and Quasar® 670. In

25 some embodiments, a first photo-detectable label interacts with a second photo-detectable moiety to modify the detectable signal, e.g., via fluorescence resonance energy transfer ("FRET"; also known as Förster resonance energy transfer).

The fluorescent labels utilized by the systems and methods disclosed herein can have

30 different peak absorption wavelengths, for example, ranging from 400 nm to 800 nm. In some embodiments, the peak absorption wavelengths of the fluorescent labels can be, or be about, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 510, 520, 530, 540, 550, 560, 570, 580, 590, 600, 610, 620, 630, 640, 650, 660, 670, 680, 690, 700, 710, 720, 730, 740, 750, 760, 770, 780, 790, 800 nm, or a number or a range between any

35 two of these values. In some embodiments the peak absorption wavelengths of the fluorescent labels can be at least, or at most, 400, 410, 420, 430, 440, 450, 460, 470,

480, 490, 500, 510, 520, 530, 540, 550, 560, 570, 580, 590, 600, 610, 620, 630, 640, 650, 660, 670, 680, 690, 700, 710, 720, 730, 740, 750, 760, 770, 780, 790, or 800 nm.

5 The fluorescent labels can have different peak emission wavelength, for example, ranging from 400 nm to 800 nm. In some embodiments, the peak emission wavelengths of the fluorescent labels can be, or be about, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 510, 520, 530, 540, 550, 560, 570, 580, 590, 600, 610, 620, 630, 640, 650, 660, 670, 680, 690, 700, 710, 720, 730, 740, 750, 760, 770, 780, 790, 800 nm, or
10 a number or a range between any two of these values. In some embodiments the peak emission wavelengths of the fluorescent labels can be at least, or at most, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 510, 520, 530, 540, 550, 560, 570, 580, 590, 600, 610, 620, 630, 640, 650, 660, 670, 680, 690, 700, 710, 720, 730, 740, 750, 760, 770, 780, 790, or 800 nm.

15 The fluorescent labels can have different Stokes shift, for example, ranging from 10 nm to 200 nm. In some embodiments, the stoke shift can be, or be about, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200 nm, or a number or a range between any two of these values. In some embodiments, the stoke shift can be at least, or at most, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150,
20 160, 170, 180, 190, or 200 nm.

In some embodiments, the distance between the peak emission wavelengths of any two fluorescent labels can vary, for example, ranging from 10 nm to 200 nm. In some
25 embodiments, the distance between the peak emission wavelengths of any two fluorescent labels can be, or be about, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200 nm, or a number or a range between any two of these values. In some embodiments, the distance between the peak emission wavelengths of any two fluorescent labels can be at least, or at most, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, or 200 nm.

30 A "light source" may be any device capable of emitting energy along the electromagnetic spectrum. A light source may be a source of visible light (VIS), ultraviolet light (UV) and/or infrared light (IR). "Visible light" (VIS) generally refers to the band of electro-magnetic radiation with a wavelength from about 400 nm to about 750 nm. "Ultraviolet (UV) light"
35 generally refers to electromagnetic radiation with a wavelength shorter than that of visible light, or from about 10 nm to about 400 nm range. "Infrared light" or infrared radiation (IR) generally refers to electromagnetic radiation with a wavelength greater than the VIS

range, or from about 750 nm to about 50,000 nm. A light source may also provide full spectrum light. Light sources may output light from a selected wavelength or a range of wavelengths. In some embodiments of the invention, the light source may be configured to provide light above or below a predetermined wavelength, or may provide light within

5 a predetermined range. A light source may be used in combination with a filter, to selectively transmit or block light of a selected wavelength from the light source. A light source may be connected to a power source by one or more electrical connectors; an array of light sources may be connected to a power source in series or in parallel. A power source may be a battery, or a vehicle electrical system or a building electrical

10 system. The light source may be connected to a power source via control electronics (control circuit); control electronics may comprise one or more switches. The one or more switches may be automated, or controlled by a sensor, timer or other input, or may be controlled by a user, or a combination thereof. For example, a user may operate a switch to turn on a UV light source; the light source may be applied on a constant basis until it

15 is turned off, or it may be pulsed (repeated on/off cycles) until it is turned off. In some embodiments, the light source may be switched from a continuously-on state to a pulsed state, or vice versa. In some embodiments, the light source may be configured to be brightening or darkening over time.

20 For operation, the light source may be connected to a power source capable of providing sufficient intensity to illuminate the sample. Control electronics may be used to switch the intensity on or off based on input from a user or some other input, and can also be used to modulate the intensity to a suitable level (e.g. to control brightness of the output light). Control electronics may be configured to turn the light source on and off as desired.

25 Control electronics may include a switch for manual, automatic, or semi-automatic operation of the light sources. The one or more switches may be, for example, a transistor, a relay or an electromechanical switch. In some embodiments, the control circuit may further comprise an AC-DC and/or a DC-DC converter for converting the voltage from the voltage source to an appropriate voltage for the light source. The control

30 circuit may comprise a DC-DC regulator for regulation of the voltage. The control circuit may further comprise a timer and/or other circuitry elements for applying electric voltage to the optical filter for a fixed period of time following the receipt of input. A switch may be activated manually or automatically in response to predetermined conditions, or with a timer. For example, control electronics may process information such as user input,

35 stored instructions, or the like.

- One or more of a plurality of light sources may be provided. In some embodiments, each of the plurality of light sources may be the same. Alternatively, one or more of the light sources may vary. The light characteristics of the light emitted by the light sources may be the same or may vary. A plurality of light sources may or may not be independently
- 5 controllable. One or more characteristic of the light source may or may not be controlled, including but not limited to whether the light source is on or off, brightness of light source, wavelength of light, intensity of light, angle of illumination, position of light source, or any combination thereof.
- 10 In some embodiments, light output from a light source may be from about 350 to about 750 nm, or any amount or range therebetween, for example from about 350 nm to about 360, 370, 380, 390, 400, 410, 420, 430 or about 450 nm, or any amount or range therebetween. In other embodiments, light from a light source may be from about 550 to
- 15 about 700 nm, or any amount or range therebetween, for example from about 550 to about 560, 570, 580, 590, 600, 610, 620, 630, 640, 650, 660, 670, 680, 690 or about 700 nm, or any amount or range therebetween. In some embodiments, the wavelength of the light generated by the light source can vary, for example, ranging from 400 nm to 800 nm. In some embodiments, the wavelength of the light generated by the light source can be, or be about, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 510, 520, 530,
- 20 540, 550, 560, 570, 580, 590, 600, 610, 620, 630, 640, 650, 660, 670, 680, 690, 700, 710, 720, 730, 740, 750, 760, 770, 780, 790, 800 nm, or a number or a range between any two of these values. In some embodiments, the wavelength of the light generated by the light source can be at least, or at most, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 510, 520, 530, 540, 550, 560, 570, 580, 590, 600, 610, 620, 630, 640,
- 25 650, 660, 670, 680, 690, 700, 710, 720, 730, 740, 750, 760, 770, 780, 790, or 800 nm. The light source may be capable of emitting electromagnetic waves in any spectrum. In some embodiments, the light source may have a wavelength falling between 10 nm and 100 μm . In some embodiments, the wavelength of light may fall between 100 nm to 5000 nm, 300 nm to 1000 nm, or 400 nm to 800 nm. In some embodiments, the wavelength
- 30 of light may be less than, and/or equal to 10 nm, 100 nm, 200 nm, 300 nm, 400 nm, 500 nm, 600 nm, 700 nm, 800 nm, 900 nm, 1000 nm, 1100 nm, 1200 nm, 1300 nm, 1500 nm, 1750 nm, 2000 nm, 2500 nm, 3000 nm, 4000 nm, or 5000 nm.
- 35 In one example, a light source may be a light-emitting diode (LED) (e.g., gallium arsenide (GaAs) LED, aluminum gallium arsenide (AlGaAs) LED, gallium arsenide phosphide (GaAsP) LED, aluminum gallium indium phosphide (AlGaInP) LED, gallium(III) phosphide (GaP) LED, indium gallium nitride (InGaN)/gallium(III) nitride (GaN) LED, or

aluminum gallium phosphide (AlGaP) LED). In another example, a light source can be a laser, for example a vertical cavity surface emitting laser (VCSEL) or other suitable light emitter such as an Indium-Gallium-Aluminum-Phosphide (InGaAlP) laser, a Gallium-Arsenic Phosphide/Gallium Phosphide (GaAsP/GaP) laser, or a Gallium-Aluminum-Arsenide/Gallium-Aluminum-Arsenide (GaAlAs/GaAs) laser. Other examples of light sources may include but are not limited to electron stimulated light sources (e.g., Cathodoluminescence, Electron Stimulated Luminescence (ESL light bulbs), Cathode ray tube (CRT monitor), Nixie tube), incandescent light sources (e.g., Carbon button lamp, Conventional incandescent light bulbs, Halogen lamps, Globar, Nernst lamp), electroluminescent (EL) light sources (e.g., Light-emitting diodes—Organic light-emitting diodes, Polymer light-emitting diodes, Solid-state lighting, LED lamp, Electroluminescent sheets Electroluminescent wires), gas discharge light sources (e.g., Fluorescent lamps, Inductive lighting, Hollow cathode lamp, Neon and argon lamps, Plasma lamps, Xenon flash lamps), or high-intensity discharge light sources (e.g., Carbon arc lamps, Ceramic discharge metal halide lamps, Hydrargyrum medium-arc iodide lamps, Mercury-vapor lamps, Metal halide lamps, Sodium vapor lamps, Xenon arc lamps). Alternatively, a light source may be a bioluminescent, chemiluminescent, phosphorescent, or fluorescent light source.

As used herein, an “optical channel” is a predefined profile of optical frequencies (or equivalently, wavelengths). For example, a first optical channel may have wavelengths of 500 nm–600 nm. To take an image in the first optical channel, one may use a detector which is only responsive to 500 nm–600 nm light, or use a bandpass filter having a transmission window of 500 nm–600 nm to filter the incoming light onto a detector responsive to 300 nm–800 nm light. A second optical channel may have wavelengths of 300 nm–450 nm and 850 nm–900 nm. To take an image in the second optical channel, one may use a detector responsive to 300 nm–450 nm light and another detector responsive to 850 nm–900 nm light and then combine the detected signals of the two detectors. Alternatively, to take an image in the second optical channel, one may use a bandstop filter which rejects 451 nm–849 nm light in front of a detector responsive to 300 nm–900 nm light.

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.

5 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.

15 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.

30 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.

5 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
10 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.
15 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.

20 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.

25 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 $\pm 20\%$, $\pm 15\%$, $\pm 10\%$, $\pm 5\%$, or $\pm 1\%$. The term “substantially” is used to indicate that a result (e.g., measurement value) is close to a targeted value, where close
30 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.

35 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.

5 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
10 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.

15 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.

20 EXAMPLES

Example 1: Concurrent sequencing of a concatenated strand (different inserts, human and PhiX)

1.1 Oligo sequences for stitch PCR method:

25 HYB2-ME – SEQ ID NO: 12; HYB2'-ME – SEQ ID NO: 14

ME sequences are underlined. These were to be used with P5-UDI-A14 and P7-UDI-B15 oligos to PCR up different genomic DNA libraries, making the libraries P5-insert-HYB2' or P7-insert-HYB2. These libraries were then combined using SOE (splicing by overhang extension) PCR to combine them together. In this experiment the following two
30 oligos were used as partners as examples:

Dual-Biotin 6T-P5-nonlin

5'Dual-biotin-TTTTTTAATGATACGGCGACCGAGATCTACAC (SEQ ID NO:41)

Dual-Biotin 6T-P7-nonlin

5'Dual-biotin-TTTTTTCAAGCAGAAGACGGCATACGAGAT (SEQ ID NO:42)

The 5' dual biotin is nonetheless, irrelevant for this experiment.

1.2 Method

1. Illumina DNA Flex libraries containing human or PhiX (bacteriophage) inserts
5 were prepared following the standard Illumina protocol:

<https://emea.illumina.com/products/by-type/sequencing-kits/library-prep-kits/nextera-dna-flex.html>

2. Two initial PCRs were set up containing:

- 25ul 2x Phusion Mastermix (New England Biolabs)
- 10 • 0.25ul 100uM dual-biotin 6T-P5-nonlin
- 0.25ul 100uM HYB2-ME
- 1ul Human Flex library (~10ng)
- 23.5ul H2O

The other PCR used the dual-biotin 6T-P7-nonlin and HYB2'-ME primer pair on the PhiX
15 Flex library.

3. PCRs were cycled:

98C for 30s, followed by 10 cycles of 98C for 10s, 50C for 30s and 72C for 30s, then a
5min extension step at 72C and then held at 4C

4. After checking that material had been made in the initial PCRs via gel
20 electrophoresis, "Splice Overlap Extension" (SOE) PCRs were assembled by combining
20ul of each of the initial PCRs.

5. SOE PCRs were cycled:

98C for 30s, followed by 8 cycles of 98C for 10s, 50C for 60s and 72C for 60s, then a
5min extension step at 72C and then held at 4C.

- 25 6. SOE PCRs were cleaned up via a 1x SPRI bead clean-up and quantified using
the Qubit Broad Range dsDNA assay (Thermofisher), prior to use in sequencing
experiments.

1.3 iSeq100 sequencing details:

An iSeq100 cartridge was cracked open, and premixed HCX (90ul ECX1 + 45ul of EXC2 + 90ul HCXE3 - ExAmp mix for iSeq100) added to the HCX Mixing well. The standard HP10 read 1 primer mix was removed from its well, washed with 200ul water 5x and then replaced with 150ul of the 16QAM sequencing primer mix.

- 5 16QAM sequencing primer mix – addition of equal concentrations of HYB2'-ME and HYB2'-ME-block in the standard sequencing primer mix from Illumina. The standard sequencing primers are at 0.3uM each within HP10, and we mix the HYB2'-ME (SEQ ID NO: 14) and HYB2'-ME-block (SEQ ID NO: 16) primers into this to give 0.5uM of each of these primers. The 50:50 ratio of blocked/unblocked primers for HYB2'-ME gives us
10 the "50%" signal required at this primer site during 16QAM sequencing.

- As shown in Figure 31A, by plotting relative intensities of light signals obtained from a first channel (ch1) and a second channel (ch2), a constellation of 16 clouds is obtained. Each of these clouds allows sequence information to be identified on both the human insert and the PhiX insert, where the top left corner of four clouds corresponds with base
15 calls corresponding to C, the top right corner of four clouds corresponds with base calls corresponding to T, the bottom left corner of four clouds corresponds with base calls corresponding to G, and the bottom right corner of four clouds corresponds with base calls corresponding to A. The basecall read out (R1 and R2) of both the human insert and the PhiX insert is also shown.

- 20 As shown in Figure 31B, alignment of R1 and R2 (minor and major reads respectively) with the known human and PhiX sequence confirmed that the method accurately sequenced the inserts. In particular the sequence identity of R1 and R2 with the known sequences was 99% (150 out of 151 correct base calls for R1 and 148 out of 149 correct base calls for R2).

25

Example 2: Concurrent sequencing of separate strands (forward and forward complement, human)

P5f-Adaptor

AATGATACGGCGACCAACGAGATCTACAC*T (SEQ ID NO:43)

30 P7f-Adaptor

CAAGCAGAAGACGGCATACGAGA*T (SEQ ID NO:44)

(* indicates a phosphorothioate linkage)

2.1 Method

1. 1 ug mix human genomic DNA from Promega in 50 ul was fragmented to 400-500 bp fragment using the TruSeq-450 program on the Covaris.
- 5 2. End prep was performed using NEBNext Ultra II kit.
3. For adapters: 15 ul of F/R oligos from each P5 and P7 were mixed and 1.5 ul 10X NEBuffer 2 was added. The mix was annealed using AK_ANNEAL program (96C for 2 mins, then to 25C at -0.1/sec). 30 ul of each of the annealed oligos was then added to 140 ul of water to make 200 ul of 15 uM adapter solution (7.5 uM each side).
- 10 4. This mix was used for standard ligation using NEBNext Ultra II kit. The resulting 93 ul was mixed with 3 ul of water and 22.5 ul of iTune (SPRI-like) beads for the 1st size selection cut.
5. Supernatant was then mixed with 12.5 ul of beads for the second size selection. DNA was eluted in 20 ul.
- 15 6. 15 ul was used for 6 cycles of Q5 PCR using 10 ul of P5f/P7f primer mix (5 uM each).
7. The PCR product was purified using 0.9x iTune bead selection. It was measured to be at 23 ng/ul, or almost 68 nM.

2.2 16QAM sequencing of library

- 20 The goal is to first block 50% of P7 ends with ddNTP spiked IMX, and then nick P5 end and perform dsSBS sequencing from both ends at the same time (16QAM).
1. ShAdp Human library (with 10% PhiX) was used. After initial denaturation and neutralization of library to give a 20pM stock, 15 ul of this was added in 485 HT1 to give a 0.6pM loading concentration for the MiniSeq run.
- 25 2. 2.25 ul of 1 mM ddNTPs was added to 500 ul of MiniSeq IMX in the Cust2 position of the MiniSeq cartridge.
3. 250 ul of BMX (Blocking Mix, Illumina) was added to the "EXT" position (cartridge well to the left of the Cust positions).

As shown in Figure 32, by plotting relative intensities of light signals obtained from a first channel (x-axis) and a second channel (y-axis), a constellation of 16 clouds is obtained over multiple cycles. Again, each of these clouds allows sequence information to be identified on both the human insert and the PhiX insert.

5

Example 3: Concurrent sequencing of separate strands (different strands, separate parts of PhiX)

As shown in Figure 33A, by plotting relative intensities of light signals obtained from a first channel (ch1) and a second channel (ch2), a constellation of 16 clouds is obtained. Each of these clouds allows sequence information to be identified on both the different inserts from the PhiX genome, where the top left corner of four clouds corresponds with base calls corresponding to C, the top right corner of four clouds corresponds with base calls corresponding to A, the bottom left corner of four clouds corresponds with base calls corresponding to G, and the bottom right corner of four clouds corresponds with base calls corresponding to T. The basecall read out (R1 and R2) of both the different inserts from the PhiX genome is also shown.

10
15

A subsequent resynthesis step allows “paired end” read to be conducted. This allows a further basecall read out to be obtained (R3 and R4).

As shown in Figure 33B, alignment of R1, R2, R3 and R4 with the known sequence confirmed that the method accurately sequenced the inserts (in particular the sequence identity of R1, R2 and R3 with the known sequence was 100%).

20

SEQUENCE LISTING

(Underlined sequences are ME or ME' sequences)

SEQ ID NO: 1: P5 sequence

5

AATGATACGGCGACCACCGAGATCTACAC

SEQ ID NO: 2: P7 sequence

10

CAAGCAGAAGACGGCATACGAGAT

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

GTGTAGATCTCGGTGGTCGCCGTATCATT

15

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

ATCTCGTATGCCGTCTTCTGCTTG

20

SEQ ID NO: 5: Alternative P5 sequence

AATGATACGGCGACCGA

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

25

TCGGTCGCCGTATCATT

SEQ ID NO: 7: A14

30

TCGTCGGCAGCGTC

SEQ ID NO: 8: B15

GTCTCGTGGGCTCGG

35

SEQ ID NO: 9: A14-ME

TCGTCGGCAGCGTCCAGATGTGTATAAGAGACAG

SEQ ID NO: 10: B15-ME

5

GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

SEQ ID NO: 11: HYB2

10

GAGTAAGTGGAAGAGATAGGAAGG

SEQ ID NO: 12: HYB2-ME

GAGTAAGTGGAAGAGATAGGAAGGAGATGTGTATAAGAGACAG

15

SEQ ID NO: 13: HYB2'

CCTTCCTATCTCTTCCACTTACTC

20

SEQ ID NO: 14: HYB2'-ME

CCTTCCTATCTCTTCCACTTACTCCAGATGTGTATAAGAGACAG

SEQ ID NO: 15: HYB2'-block

25

CCTTCCTATCTCTTCCACTTACT-3'propanol

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

30

CCTTCCTATCTCTTCCACTTACTCAGATGTGTATAAGAGACAG-3'propanol

SEQ ID NO: 17: ME'-A14'

CTGTCTCTTATACACATCTGACGCTGCCGACGA

35

SEQ ID NO: 18: A14'

GACGCTGCCGACGA

SEQ ID NO: 19: ME'-B15'

5 CTGTCTCTTATACACATCTCCGAGCCCACGAGAC

SEQ ID NO: 20: B15'

CCGAGCCCACGAGAC

10

SEQ ID NO: 21: ME'-HYB2

CTGTCTCTTATACACATCTGAGTAAGTGGAAGAGATAGGAAGG

15 **SEQ ID NO: 22:** ME'-HYB2'

CTGTCTCTTATACACATCTCCTTCCTATCTCTTCCACTTACTC

SEQ ID NO: 23: Removable P5 sequence

20

TTTTTTTTTTAATGATACGGCGACCACCGAUCTACAC
(where U = 2-deoxyuridine)

SEQ ID NO: 24: Removable P7 sequence

25

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

SEQ ID NO: 25: Extended primer sequence with A as 5' additional nucleotide and P5'
sequence (complementary to P5)

30

AGTGTAGATCTCGGTGGTCGCCGTATCATT

SEQ ID NO: 26: Extended primer sequence with T as 5' additional nucleotide and P5'
sequence (complementary to P5)

35

TGTGTAGATCTCGGTGGTCGCCGTATCATT

SEQ ID NO: 27: Extended primer sequence with C as 5' additional nucleotide and P5' sequence (complementary to P5)

CGTGTAGATCTCGGTGGTCGCCGTATCATT

5

SEQ ID NO: 28: Extended primer sequence with G as 5' additional nucleotide and P5' sequence (complementary to P5)

GGTGTAGATCTCGGTGGTCGCCGTATCATT

10

SEQ ID NO: 29: Extended primer sequence with A as 5' additional nucleotide and P7' sequence (complementary to P7)

AATCTCGTATGCCGTCTTCTGCTTG

15

SEQ ID NO: 30: Extended primer sequence with T as 5' additional nucleotide and P7' sequence (complementary to P7)

TATCTCGTATGCCGTCTTCTGCTTG

20

SEQ ID NO: 31: Extended primer sequence with C as 5' additional nucleotide and P7' sequence (complementary to P7)

CATCTCGTATGCCGTCTTCTGCTTG

25

SEQ ID NO: 32: Extended primer sequence with G as 5' additional nucleotide and P7' sequence (complementary to P7)

GATCTCGTATGCCGTCTTCTGCTTG

30

SEQ ID NO: 33: Extended primer sequence with A as 5' additional nucleotide and alternative P5' sequence (complementary to alternative P5)

ATCGGTCGCCGTATCATT

35

SEQ ID NO: 34: Extended primer sequence with T as 5' additional nucleotide and alternative P5' sequence (complementary to alternative P5)

TTCGGTCGCCGTATCATT

5 **SEQ ID NO: 35:** Extended primer sequence with C as 5' additional nucleotide and alternative P5' sequence (complementary to alternative P5)

CTCGGTCGCCGTATCATT

10 **SEQ ID NO: 36:** Extended primer sequence with G as 5' additional nucleotide and alternative P5' sequence (complementary to alternative P5)

GTCGGTCGCCGTATCATT

15 **SEQ ID NO: 37 SBS3**

ACACTCTTTCCCTACACGACGCTCTTCCGATCT

SEQ ID NO: 38 SBS3'

20 AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

SEQ ID NO: 39 SBS12

25 GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

SEQ ID NO: 40 SBS12'

AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC

SEQ ID NO: 41 Dual-Biotin 6T-P5-nonlin

30 5'Dual-biotin-TTTTTTAATGATACGGCGACCAACCGAGATCTACAC

SEQ ID NO: 42 Dual-Biotin 6T-P7-nonlin

5'Dual-biotin-TTTTTTCAAGCAGAAAGACGGCATACGAGAT

SEQ ID NO: 43 P5f-Adaptor

AATGATACGGCGACCAACCGAGATCTACAC*T

35 **SEQ ID NO: 44**

CAAGCAGAAGACGGCATACGAGA*T

SEQ ID NO: 45 Human DNA sequence from clone RP11-151A10 on chromosome 10, complete sequence. GenBank Accession No: AL157888.16, Region, 137023 to 137173. 3' to 5' shown in Figure 31.

5 tctgggtcct gtgtcctgct ctgtgagctc tgcagccact cctggttgact ttgtgcctta
atgcaaatga tccctctcct tggtcacccg ctggggtttg cactgttctt gttgaccttt
agcctatgcc ttcttgggaa cctcccctgg g

SEQ ID NO: 46. Escherichia phiX174 virus strain evolved J1, complete genome.
10 GenBank Accession No: MH378443.1. Region 5039 to 5187. 3' to 5' shown in Figure 31.

Tatggaccttgctgctaaaggtctaggagctaaagaatggaacaactcactaaaaaccaagctg
tcgctacttcccaagaagctgttcagaatcagaatgagccgcaacttcgggatgaaaatgctca
caatgacaaatctgtccacgg

15 **SEQ ID NO: 47.** Escherichia phiX174 virus strain evolved J1, complete genome.
GenBank Accession No: MH378443.1. Region 3923 to 4072. 3' to 5' shown in Figure 33.

aggattgaca cctcccaat tgtatgtttt catgcctcca aatcttgag gcttttttat
20 ggttcgtcct tattaccctt ctgaatgtca cgctgattat tttgactttg agcgtatcga
ggctcttaaa cctgctattg aggcttgtgg

SEQ ID NO: 48. Escherichia phiX174 virus strain evolved J1, complete genome.
GenBank Accession No: MH378443.1. Region 5067 to 5217. 3' to 5' shown in Figure 33.

25 gctaaagaat ggaacaactc actaaaaacc aagctgtcgc tacttcccaa gaagctgttc
agaatcagaa tgagccgcaa cttcgggatg aaaatgctca caatgacaaa tctgtccacg
gagtgcctaa tccaacttac caagctgggt t

30 **SEQ ID NO: 49.** Escherichia phiX174 virus strain evolved J1, complete genome.
GenBank Accession No: MH378443.1. Region 4171 to 4320. 3' to 5' shown in Figure 33.

gcgttgagtt cgataatggt gatatgtatg ttgacggcca taaggctgct tctgacgttc
gtgatgagtt tgtatctgtt actgagaagt taatggatga attggcacia tgctacaatg
35 tgcctcccca acttgatatt aataacacta

SEQ ID NO: 50. Escherichia phiX174 virus strain evolved J1, complete genome.
GenBank Accession No: MH378443.1. Region 5016-5168. 3' to 5' shown in Figure 33.

5 atacgttaac aaaaagtcag atatggacct tgctgctaaa ggtctaggag
 ctaaagaatg gaacaactca ctaaaaacca agctgtcgct acttccaag
 aagctgttca gaatcagaat gagccgcaac ttcgggatga aaatgctcac a

10

15

20

25

30

35

40

45

50

CLAIMS

1. A method of base calling nucleobases of two or more polynucleotide sequence portions, the method comprising:
- 5 (a) obtaining first intensity data comprising a combined intensity of a first signal obtained based upon a respective first nucleobase of at least one first polynucleotide sequence portion and a second signal obtained based upon a respective second nucleobase of at least one second polynucleotide sequence portion;
- 10 (b) obtaining second intensity data comprising a combined intensity of a third signal obtained based upon the respective first nucleobase of the at least one first polynucleotide sequence portion and a fourth signal obtained based upon the respective second nucleobase of the at least one second polynucleotide sequence portion;
- 15 (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,
- wherein said polynucleotide sequence portions have been selectively processed such that an intensity of the signals obtained based upon the respective first nucleobase is greater than an intensity of the signals obtained based upon the respective second nucleobase.
- 20
2. The method of claim 1, wherein the first and second signals and/or the third and fourth signals are obtained substantially simultaneously.
- 25
3. The method of claim 1 or claim 2, wherein 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 signals and the combined intensity of the third and fourth signals.
- 30
4. The method of any one preceding claim, wherein the plurality of classifications comprises sixteen classifications, each classification representing one of sixteen unique combinations of first and second nucleobases.
- 35
5. The method of any one preceding claim, wherein the polynucleotide sequence portions have been selectively processed such that, during sequencing, a greater number of the first polynucleotide sequence portions are capable of generating a signal

than a number of the second polynucleotide sequence portions that are capable of generating a signal.

5 6. The method of claim 5, wherein a ratio between the number of the first polynucleotide sequence portions capable of generating a signal and the number of the second polynucleotide sequence portions capable of generating a signal is between 1.25:1 to 5:1, preferably between 1.5:1 to 3:1, more preferably about 2:1.

10 7. The method of any one preceding claim, wherein the first signal, second signal, third signal and fourth signal are generated based on light emissions associated with the respective nucleobase.

8. The method of claim 7, wherein the obtained signals are generated by:
contacting a plurality of polynucleotide molecules comprising the first and second
15 polynucleotide sequence portions with first primers for sequencing the first polynucleotide sequence portions and second primers for sequencing the second polynucleotide sequence portions;
extending the first primers and the second primers by contacting the
polynucleotide molecules with labeled nucleobases to form first labeled primers and
20 second labeled primers;
stimulating the light emissions from the first and second labeled primers; and
detecting the light emissions at a sensor.

9. The method of claim 8, wherein:
25 the first and second signals are based on light emissions detected in a first range of optical frequencies;
the third and fourth signals are based on light emissions detected in a second range of optical frequencies; and
wherein the first range of optical frequencies and the second range of optical
30 frequencies are not identical.

10. The method of claim 8 or claim 9, wherein the plurality of polynucleotide molecules are attached to a substrate, and wherein the light emissions from the first labeled primers and the light emissions from the second labeled primers are emitted from
35 the same region or substantially overlapping regions of the substrate.

11. The method of any one of claims 8 to 10, wherein the light emissions detected at the sensor are spatially unresolved.
12. The method of claim 11, wherein the sensor is configured to provide a single
5 output based upon the first and second signals.
13. The method of any one of claims 8 to 12, wherein the sensor comprises a single sensing element.
- 10 14. The method of any one of claims 8 to 13, wherein each polynucleotide molecule comprises one or more copies of the first polynucleotide sequence portion and one or more copies of the second polynucleotide sequence portion.
- 15 15. The method of any one of claims 1 to 13, wherein the first and second polynucleotide sequence portions are respective portions of different polynucleotide molecules.
- 20 16. The method of any one of claims 8 to 15, wherein the polynucleotide sequence portions have been selectively processed by contacting the plurality of polynucleotide molecules with unblocked first primers and a predetermined fraction of second primers which have a blocked 3' end.
- 25 17. The method of any one preceding claim, wherein selectively processing comprises preparing for selective sequencing or conducting selective sequencing.
- 30 18. The method of any one preceding claim, wherein the polynucleotide sequence portions have been selectively processed to provide a greater total number of the first polynucleotide sequence portions than a total number of the second polynucleotide sequence portions.
- 35 19. The method of claim 18, wherein selectively processing comprises conducting selective amplification.
20. The method of any one preceding claim, wherein the at least one first polynucleotide sequence portion and the at least one second polynucleotide sequence portion are present in a cluster.

21. The method of any one preceding claim, wherein the one of the plurality of classifications is selected based on the first and the second intensity data using a Gaussian mixture model.
- 5 22. The method of any one preceding claim, further comprising, based on the selected classification, determining that the second polynucleotide sequence portion is modified relative to the first polynucleotide sequence portion at a location associated with the first and second nucleobases.
- 10 23. The method of claim 22, wherein the second polynucleotide sequence portion is modified relative to the first polynucleotide sequence portion resulting from a library preparation and/or sequencing error.
- 15 24. The method of claim 22, wherein the second polynucleotide sequence portion is modified relative to the first polynucleotide sequence portion resulting from conversion of a modified cytosine to thymine or a nucleobase which is read as thymine/uracil, and/or of an unmodified cytosine to uracil or a nucleobase which is read as thymine/uracil.
- 20 25. The method of any one preceding claim, further comprising repeating steps (a) to (d) for each of a plurality of base calling cycles.
26. A method of base calling nucleobases of n polynucleotide sequence portions, the method comprising:
- 25 (a) obtaining first intensity data comprising a combined intensity of respective first signal components generated by each of the nth portions obtained based upon respective nth nucleobases in each of the n portions, wherein the respective first signal components are obtained simultaneously;
- 30 (b) obtaining second intensity data comprising a combined intensity of respective second signal components generated by each of the nth portions obtained based upon respective nth nucleobases in each of the n portions, wherein the respective second 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 nth nucleobases; and
- 35 (d) based on the selected classification, base calling the respective nth nucleobases for all n portions,

wherein said polynucleotide sequence portions have been selectively processed such that an intensity of the signals obtained based upon the respective first nucleobase is greater than an intensity of the signals obtained based upon the respective second nucleobase.

5

27. The method of claim 26, wherein selecting the classification based on the first and second intensity data comprises selecting the classification based on the combined intensity of respective first signal components and second signal components.

10

28. The method of claim 27, wherein the plurality of classifications comprises 4^n classifications, each classification representing one of 4^n unique combinations of n th nucleobases.

15

29. The method of any one of claims 26 to 28, further comprising repeating steps (a) to (d) for each of a plurality of base calling cycles.

30. A data processing device comprising means for carrying out a method according to any one preceding claim.

20

31. A data processing device according to claim 30, wherein the data processing device is a polynucleotide sequencer.

25

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

30

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

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

35

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