METHODS FOR MAINTAINING THE INTEGRITY AND IDENTIFICATION OF A NUCLEIC ACID TEMPLATE IN A MULTIPLEX SEQUENCING REACTION

Field of the Invention

The invention generally relates to methods for maintaining the integrity and identification of a nucleic acid template in a multiplex sequencing reaction.

Background

Sequencing-by-synthesis involves template-dependent addition of nucleotides to a template/primer duplex. Nucleotide addition is mediated by a polymerase enzyme and added nucleotides may be labeled in order to facilitate their detection. Single molecule sequencing has been used to obtain high-throughput sequence information on individual DNA or RNA. The ability to multiplex samples, i.e., pool different patient samples, is important for decreasing costs and increasing the through-put of sequencing-by-synthesis platforms.

One issue that presents itself in a multiplex sequencing reaction is maintaining accurate identification of a sample throughout the sequencing process and. Unfortunately, the sample preparation phase frequently introduces errors into the subsequent sequencing reaction. For example, template molecules may physically overlap on a substrate; and in some cases result in a detectable erroneous signal. In any sequencing process that requires high accuracy, such errors can have a significant impact on results. Next generation sequencing typically involves the generation of an in vitro library; arraying of nucleic acid templates into physically distinct locations on a solid support (and optional clonal amplification of the templates into spatially localized clusters); and the sequencing-by-synthesis reaction itself. Errors can occur during generation of the library, arraying and amplification steps that result in molecular overlap of different patient samples. For example, heteroduplexes, i.e., improper pairing of template strands from different samples, may form during the amplification of the multiplexed samples. Molecular overlap during any of the first three steps in the process may lead to assigning sequencing data to the wrong patient sample. Other errors can be introduced due to incomplete amplification, stray labeled primers, and other sources that result in erroneous signal detection. In a single molecule sequencing protocol, such errors can have a significant affect on the results. Accordingly, methods for reducing or eliminating errors in such procedures are needed.

Summary

The invention generally provides methods for validating the results of a molecular detection assay and for enabling the detection of errors introduced in sample preparation. The invention is applicable to nucleic acid sequencing, protein detection, and other methods that involve precise measurement of the presence and/or amount of an analyte. The invention utilizes two or more identifiers that are uniquely associated with an analyte of interest, such that the identifiers will only be present when the analyte of interest is detected. In this way, false positive results are avoided by requiring concomitant detection of two or more independent markers associated with the analyte to be detected.

The invention is, especially useful in multiplex next-generation sequencing applications in which errors can have a significant impact on results, but is also applicable across a broad range of detection assays, including protein detection assays. The invention is based upon using two or more identifiers, such as nucleic acid barcodes that are uniquely associated with an analyte to be detected. For example, in sequencing applications, a pair of barcodes are placed such that they flank the sequence of interest. The identifiers are unique to the sample, so that valid sequence data are confirmed by the presence of the pair of flanking identifiers. In contrast, template containing only one identifier or an incorrect pair (i.e., a pair or more of identifiers not associated with the same template) will be excluded from analysis. Thus, methods of the invention make it possible to determine whether the integrity of a sample was maintained during a multiplex sequencing reaction, and prevent the assigning of sequence data to the wrong sample.

In other applications, two or more nucleic acid tags are associated with a protein or other analyte of interest. Identification of the sequence tags confirms the accuracy of the detection of the protein. In this embodiment of the invention, a protein is capture by a binding agent, such as an antibody, that is coupled to at least two oligonucleotides of known sequence. The identification of the captured protein is confirmed via sequencing of the tags that are uniquely associated with the binding agent. The presence of both sequence tags validates the capture of the target protein.

In protein detection, identifiers can also include detectable labels, antibodies, other proteins, including glycoproteins, vitamins, steroids, viral particles, and the like that bind

specifically to a protein of interest. Detection of the identifiers validates that the protein of interest has been found.

In nucleic acid sequencing, preferred identifiers are nucleic acid barcodes. The barcodes can be of any appropriate length (e.g., from about 2 to about 50 nucleotides) and any number of barcode sequences can be used. Preferably, the barcodes are present in pairs that flank the template to be sequenced, or portions thereof. While the pair of barcodes associated with any given template is unique, individual barcode sequences can be used with multiple templates as long as no pair is used across to different samples. While it is possible, and may be preferred in some circumstances, to use more than two barcoded sequences or other identifiers, the invention will be exemplified using pairs of identifier molecules. In a preferred configuration, each template from a particular sample is associated with the same unique pair of barcode sequence. The pairs ideally flank the region of the template that is to be sequenced. Thus, the pairs can be contiguous with a region of interest or can be separated from the region of interest by spacers (which include intervening sequence in the template or inserted spacer sequence) or by naturally-occurring sequence. Other identifiers, such as nucleic acid binding proteins, optical labels, nucleotide analogs, and others known in the art can be used as identifiers according to the invention.

In certain aspects, methods of the invention involve obtaining a template nucleic acid, incorporating a pair of sequence identifiers into the template, and sequencing the template. Methods of the invention may further include directly or indirectly attaching the template to a substrate. The template may be amplified prior to attachment to the substrate or may be amplified on the substrate after it has been attached to the substrate. In particular embodiments, the template is amplified after it has been attached to the substrate.

Identifiers can be incorporated anywhere on a template. For example, a first identifier can be incorporated into a 5' end of the template and a second identifier can incorporated into a 3' end of the template. The first and second identifiers are designed such that each set of identifiers is correlated to a particular sample (e.g., material obtained from a particular patient or group of patients), allowing samples to be distinguished and/or validated. The first and second identifiers may be the same or different, so long as the combination is unique with respect to the sample or sub-sample that one seeks to distinguish and/or validate.

After identifiers have been incorporated into the template, the template is sequenced. Sequencing may be by any method known in the art. Sequencing-by-synthesis is a common technique used in next generation procedures and works well with the instant invention. However, other sequencing methods can be used, including sequence-by-ligation, sequencing-by-hybridization; gel-based techniques and others. In general, sequencing involves hybridizing a primer to a template to form a template/primer duplex, contacting the duplex with a polymerase in the presence of a detectably-labeled nucleotides under conditions that permit the polymerase to add nucleotides to the primer in a template-dependent manner. Signal from the detectable label is then used as to identify the incorporated base and the steps are sequentially repeated in order to determine the linear order of nucleotides in the template. Exemplary detectable labels include radiolabels, florescent labels, enzymatic labels, etc. In particular embodiments, the detectable label may be an optically detectable label, such as a fluorescent label. Exemplary fluorescent labels include cyanine, rhodamine, fluorescien, coumarin, BODIPY, alexa, or conjugated multi-dyes.

Numerous techniques are known for detecting sequence and for identifying barcodes and some are exemplified below. However, the exact means for detecting and compiling sequence data does not affect the function of the invention described herein.

If the analyte to be detected is not a nucleic acid, the identifiers can still be nucleic acid barcodes that are associated with the analyte binding moiety. Sequencing of the barcodes then validates detection of the analyte (i.e., the presence of two or more barcodes in association with the binder must be detected in order to validate the detection). In protein detection, the idenfiers can also be part of the primary structure of the protein to be detected. In this embodiment, the amino acid sequence of the protein is determined (e.g., by mass spectrometry or Edman degradation). The embedded amino acid sequence identifiers would have to be present in the amino acid sequence in order to validate the identity of the protein.

Brief Description of the Drawings

Figure 1 is a set of drawings showing how molecular overlap can lead to apparent high quality, but incorrect sequence data.

Figure 2 is a set of drawings depicting how methods of the invention allow one to detect, during analysis, whether the integrity of a patient's sample was maintained during a multiplex sequencing reaction.

Detailed Description

Methods of the invention relate to validating and maintaining the integrity of a nucleic acid template in a multiplex sequencing reaction. Methods of the invention involve obtaining a template nucleic acid, incorporating a pair of sequence identifiers into the template, and sequencing the template.

Reference is now made to Figure 1, which shows a set of drawings depicting one example of how molecular overlap can lead to apparent high quality, but incorrect sequence data. Panel 1 shows the pooling of nucleic acid libraries from two patient samples, A and B. Each template has a gene fragment from that patient (e.g., GA), and a molecule barcode specific to that patient (e.g., A). As shown in panel 1, the pool includes patient A's sample which is nucleic acid templates have barcode A and gene fragment GA, and also includes patient B's sample which is nucleic acid templates have barcode B and gene fragment GB.

Panel 2 shows that after pooling, the samples are then hybridized to primers at discrete locations on a flow cell. In this figure, the nucleic acid from patient A and the nucleic acid from patient B land within a cluster radius on the solid support. When the amplification process begins, primers attached to the solid support extend on each template. In one case (patient A's nucleic acid template), the primer extends across a full length of the template. In the other case (patient B's nucleic acid template), the primer partially extends, incorporating the sequence of barcode B and common priming sequence, but not the gene fragment sequence GB from patient B's nucleic acid template.

Panel 3 shows an example of improper molecular overlap. After extension, the duplexes are denatured and re-annealed for another amplification cycle. This example shows improper molecular overlap via formation of a heteroduplex between patient A's nucleic acid template and patient B's nucleic acid template. In this example, strand *i* hybridizes to strand *ii*, forming a bridge. Barcode regions of each strand are not complementary, but the strands still hybridize because of the amount of common priming sequence on both sides of the barcode.

Consequently, a polymerase will extend strand *ii*, copying patient A's gene fragment onto a strand containing patient B's barcode. A chimeric molecule has been formed.

Panel 4 shows the results of this molecular overlap. Strand *ii* contains the barcode for patient B but the gene fragment for patient A. During the sequencing process, patient A's sequence data will be improperly and incorrectly associated with patient B. If the chimeric molecule is formed early in the amplification process, it can dominate the cluster of amplification products, leading to a homogenous clonal chimeric cluster. During sequencing, this would give a high-quality signal, but the data will be assigned to the wrong sample because it has become associated with the wrong barcode.

Reference is now made to Figure 2, which shows a set of drawings depicting how methods of the invention allow one to detect, during analysis, whether the integrity of a patient's sample was maintained during a multiplex sequencing reaction. Panel 1 shows the pooling of nucleic acid libraries from two patient samples, A and B. Each template has a gene fragment from that patient (e.g., GA), and a pair of sequence identifiers (first and second) specific to that patient (e.g., A1 and A2).

The first and second identifiers are designed such that each set of identifiers is correlated to a particular patient, allowing patient samples to be distinguished. The first and second identifiers may be the same. Alternatively, the first and second identifiers may be different. In certain embodiments, the pair of identifiers are first and second barcode sequences that are attached to the template. The barcode sequences may be attached to the template such that a first barcode sequence is attached to a 5' end of the template and a second barcode sequence is attached to a 3' end of the template. The first and second barcode sequences may be the same, or they may be different.

In this figure, the sequence identifiers are pairs of barcode sequences. This figure also shows that the pair of barcode sequences are attached to the nucleic acid template at the 5' and the 3' ends of the template, such that the barcode sequences flank the gene fragment of each template. As shown in panel 1, the pool includes patient A's sample which is nucleic acid templates having barcode sequences A1 and A2 that flank gene fragment GA, and also includes patient B's sample which is nucleic acid templates having barcode sequences B1 and B2 and gene fragment GB.

Panel 2 shows that after pooling, the samples are then hybridized to primers at discrete locations on a flow cell. In this figure, the nucleic acid from patient A and the nucleic acid from patient B land within a cluster radius on the solid support. When the amplification process begins, primers attached to the solid support extend on each template. In one case (patient A's nucleic acid template), the primer extends across a full length of the template. In the other case (patient B's nucleic acid template), the primer partially extends, incorporating the sequence of barcode B2 and common priming sequence, but not the gene fragment sequence GB from patient B's nucleic acid template or the sequence of barcode B1.

Panel 3 shows an example of improper molecular overlap. After extension, the duplexes are denatured and re-annealed for another amplification cycle. This example shows improper molecular overlap via formation of a heteroduplex between patient A's nucleic acid template and patient B's nucleic acid template. In this example, strand *i* hybridizes to strand *ii*, forming a bridge. Barcode regions A2 and B2 of each strand are not complementary, but the strands still hybridize because of the amount of common priming sequence on both sides of the barcode. Consequently, a polymerase will extend strand *ii*, copying patient A's gene fragment. However, in the case of two barcodes, only one of patient B's barcodes (B2) will be contained in the copied strand. The other barcode will be one of patient A's barcode sequences (A1). Unlike the example in Figure 1, the chimeric molecule formed in this example includes one barcode sequence from patient A (A1) and one barcode sequence from patient B (B2).

Panel 4 shows the results of this molecular overlap. While strand *ii* contains the gene fragment for patient A, it contains only one of patient B's barcodes (B2). The other barcode will be one of patient A's barcode sequences (A1). While each barcode sequence on its own is a valid barcode sequence, taken together, the two barcode sequences do not form a valid barcode sequence pair. During the sequencing process, this would yield a high-quality signal, but during data analysis the molecules would be recognized as invalid because the data contains an invalid barcode sequence pair, i.e., it is known that barcode sequence A1 should not be seen together with barcode sequence B2 and hence a crossover has occurred. Thus during analysis, it is determined that the integrity of the patient's sample was not maintained and the data can then be excluded from final analysis.

The following sections discuss general considerations for barcode sequences, attaching barcode sequences to nucleic acid templates, and nucleic acid sequencing, for example, template

considerations, polymerases useful in sequencing-by-synthesis, choice of surfaces, reaction conditions, signal detection and analysis.

Nucleic Acid Templates

Nucleic acid templates include deoxyribonucleic acid (DNA) and/or ribonucleic acid (RNA). Nucleic acid templates can be synthetic or derived from naturally occurring sources, or may include both synthetic and natural sequence; and may include PCR product. In one embodiment, nucleic acid template molecules are isolated from a biological sample containing a variety of other components, such as proteins, lipids and non-template nucleic acids. Nucleic acid template molecules can be obtained from any cellular material, obtained from an animal, plant, bacterium, fungus, or any other cellular organism. Biological samples for use in the present invention include viral particles or preparations. Nucleic acid template molecules can be obtained directly from an organism or from a biological sample obtained from an organism, e.g., from blood, urine, cerebrospinal fluid, seminal fluid, saliva, sputum, stool and tissue. Any tissue or body fluid specimen may be used as a source for nucleic acid for use in the invention. Nucleic acid template molecules can also be isolated from cultured cells, such as a primary cell culture or a cell line. The cells or tissues from which template nucleic acids are obtained can be infected with a virus or other intracellular pathogen. A sample can also be total RNA extracted from a biological specimen, a cDNA library, viral, or genomic DNA.

Nucleic acid obtained from biological samples typically is fragmented to produce suitable fragments for analysis. In one embodiment, nucleic acid from a biological sample is fragmented by sonication. Nucleic acid template molecules can be obtained as described in U.S. Patent Application Publication Number US2002/0190663 A1, published Oct. 9, 2003. Generally, nucleic acid can be extracted from a biological sample by a variety of techniques such as those described by Maniatis, et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor, N.Y., pp. 280-281 (1982). Generally, individual nucleic acid template molecules can be from about 1 base to about 20 kb. Nucleic acid molecules may be single-stranded, double-stranded, or double-stranded with single-stranded regions (for example, stem- and loop-structures).

A biological sample as described herein may be homogenized or fractionated in the presence of a detergent or surfactant. The concentration of the detergent in the buffer may be about 0.05% to about 10.0%. The concentration of the detergent can be up to an amount where

the detergent remains soluble in the solution. In a preferred embodiment, the concentration of the detergent is between 0.1% to about 2%. The detergent, particularly a mild one that is nondenaturing, can act to solubilize the sample. Detergents may be ionic or nonionic. Examples of nonionic detergents include triton, such as the Triton® X series (Triton® X-100 t-Oct-C_6H_4-$(OCH_2$-$CH_2)_x$OH, x=9-10, Triton® X-100R, Triton® X-114 x=7-8), octyl glucoside, polyoxyethylene(9)dodecyl ether, digitonin, IGEPAL® CA630 octylphenyl polyethylene glycol, n-octyl-beta-D-glucopyranoside (betaOG), n-dodecyl-beta, Tween® 20 polyethylene glycol sorbitan monolaurate, Tween® 80 polyethylene glycol sorbitan monooleate, polidocanol, n-dodecyl beta-D-maltoside (DDM), NP-40 nonylphenyl polyethylene glycol, C12E8 (octaethylene glycol n-dodecyl monoether), hexaethyleneglycol mono-n-tetradecyl ether (C14EO6), octyl-beta-thioglucopyranoside (octyl thioglucoside, OTG), Emulgen, and polyoxyethylene 10 lauryl ether (C12E10). Examples of ionic detergents (anionic or cationic) include deoxycholate, sodium dodecyl sulfate (SDS), N-lauroylsarcosine, and cetyltrimethylammoniumbromide (CTAB). A zwitterionic reagent may also be used in the purification schemes of the present invention, such as Chaps, zwitterion 3-14, and 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulf-onate. It is contemplated also that urea may be added with or without another detergent or surfactant.

Lysis or homogenization solutions may further contain other agents, such as reducing agents. Examples of such reducing agents include dithiothreitol (DTT), .beta.-mercaptoethanol, DTE, GSH, cysteine, cysteamine, tricarboxyethyl phosphine (TCEP), or salts of sulfurous acid.

Barcode Sequences

In certain embodiments, the sequence identifiers are barcode sequences that are attached to or incorporated into a nucleic acid template. The barcode sequences may be attached to the template such that a first barcode sequence is attached to a 5' end of the template and a second barcode sequence is attached to a 3' end of the template. The first and second barcode sequences may be the same, or they may be different. Barcode sequence may be incorporated into a contiguous region of a template that includes the target to be sequenced.

Exemplary methods for designing sets of barcode sequences and other methods for attaching barcode sequences are shown in U.S. paten numbers 6,138,077; 6,352,828; 5,636,400; 6,172,214; 6235,475; 7,393,665; 7,544,473; 5,846,719; 5,695,934; 5,604,097; 6,150,516;

RE39,793; 7,537,897; 6172,218; and 5,863,722, the content of each of which is incorporated by reference herein in its entirety.

The barcode sequence generally includes certain features that make the sequence useful in sequencing reactions. For example the barcode sequences are designed to have minimal or no homopolymer regions, i.e., 2 or more of the same base in a row such as AA or CCC, within the barcode sequence. The barcode sequences are also designed so that they do not overlap the target region to be sequence or contain a sequence that is identical to the target.

The first and second barcode sequences are designed such that each pair of sequences is correlated to a particular sample, allowing samples to be distinguished and validated. Methods of designing sets of barcode sequences is shown for example in Brenner et al. (U.S. patent number 6,235,475), the contents of which are incorporated by reference herein in their entirety. In certain embodiments, the barcode sequences range from about 2 nucleotides to about 50; and preferably from about 4 to about 20 nucleotides. Since the barcode sequence is sequenced along with the template nucleic acid or may be sequenced in a separate read, the oligonucleotide length should be of minimal length so as to permit the longest read from the template nucleic acid attached. Generally, the barcode sequences are spaced from the template nucleic acid molecule by at least one base.

Methods of the invention involve attaching the barcode sequences to the template nucleic acids. Template nucleic acids are able to be fragmented or sheared to desired length, e.g. generally from 100 to 500 bases or longer, using a variety of mechanical, chemical and/or enzymatic methods. DNA may be randomly sheared via sonication, exposed to a DNase or one or more restriction enzymes, a transposase, or nicking enzyme. RNA may be fragmented by brief exposure to an RNase, heat plus magnesium, or by shearing. The RNA may be converted to cDNA before or after fragmentation.

Barcode sequence is integrated with template using, for example, a ligase or a polymerase. The ligase may be any enzyme capable of ligating an oligonucleotide (RNA or DNA) to the template nucleic acid molecule. Suitable ligases include T4 DNA ligase and T4 RNA ligase (such ligases are available commercially, from New England Biolabs). Methods for using ligases are well known in the art. The polymerase may be any enzyme capable of adding nucleotides to the 3' and the 5' terminus of template nucleic acid molecules. Barcode sequence can be incorporated via a PCR reaction as part of the PCR primer.

The ligation may be blunt ended or via use of complementary over hanging ends. In certain embodiments, following fragmentation, the ends of the fragments may be repaired, trimmed (e.g. using an exonuclease), or filled (e.g., using a polymerase and dNTPs), to form blunt ends. Upon generating blunt ends, the ends may be treated with a polymerase and dATP to form a template independent addition to the 3'-end and the 5-end of the fragments, thus producing a single A overhanging. This single A is used to guide ligation of fragments with a single T overhanging from the 5'-end in a method referred to as T-A cloning.

Alternatively, because the possible combination of overhangs left by the restriction enzymes are known after a restriction digestion, the ends may be left as is, i.e., ragged ends. In certain embodiments double stranded oligonucleotides with complementary over hanging ends are used.

Surface Attachment

Methods of the invention may involve attaching or immobilizing barcoded nucleic acid templates to solid supports. Such methods are described for example in Sabot et al. (U.S. patent application number 2009/0226975), Adessi et al. (U.S. patent number 7,115,400), and Kawashima et al. (U.S. patent application number 2005/0100900), the content of each of which is incorporated by reference herein in its entirety.

The term immobilized as used herein is intended to encompass direct or indirect attachment to a solid support via covalent or non-covalent bond(s). In certain embodiments of the invention, covalent attachment may be used, but generally all that is required is that template remain immobilized on the support. Typically oligonucleotides are immobilized such that a 3' end is available for enzymatic extension and at least a portion of the sequence is capable of hybridizing to a complementary sequence. Immobilization can occur via hybridization to a surface attached oligonucleotide. Alternatively, immobilization can occur by means other than base-pairing hybridization, such as the covalent attachment set forth above.

Substrates or supports for use in the invention include, but are not limited to, latex beads, dextran beads, polystyrene surfaces, polypropylene surfaces, polyacrylamide gel, gold surfaces, glass surfaces and silicon wafers. In certain embodiments, the solid support may include an inert substrate or matrix that has been functionalized, for example by the application of a layer or coating of an intermediate material including reactive groups that permit covalent attachment to

molecules such as polynucleotides.

Amplification

In certain embodiments, methods of the invention involve amplifying barcoded nucleic acid templates prior to sequencing the templates. Such methods are described for example in Sabot et al. (U.S. patent application number 2009/0226975), Adessi et al. (U.S. patent number 7,115,400), and Kawashima et al. (U.S. patent application number 2005/0100900), the content of each of which is incorporated by reference herein in its entirety.

Primer oligonucleotides or amplification sequences are polynucleotide sequences that are capable of annealing specifically to a single stranded polynucleotide sequence to be amplified under conditions encountered in a primer annealing step of an amplification reaction. Generally, the terms nucleic acid, polynucleotide and oligonucleotide are used interchangeably herein. The different terms are not intended to denote any particular difference in size, sequence, or other property unless specifically indicated otherwise. For clarity of description the terms may be used to distinguish one species of molecule from another when describing a particular method or composition that includes several molecular species.

Primers may additionally include non-nucleotide chemical modifications, for example to facilitate covalent attachment of the primer to a solid support. Certain chemical modifications may themselves improve the function of the molecule as a primer or may provide some other useful functionality, such as providing a cleavage site that enables the primer (or an extended polynucleotide strand derived therefrom) to be cleaved from a solid support. Useful chemical modifications can also provide reversible modifications that prevent hybridization or extension of the primer until the modification is removed or reversed. Similarly, other molecules attached to a surface in accordance with the invention can include cleavable linker moieties and or reversible modifications that alter a particular chemical activity of function of the molecule.

A plurality of oligonucleotides used in the methods set forth herein can include species that function as capture oligonucleotides. The capture oligonucleotides may include a template specific portion, namely a sequence of nucleotides capable of annealing to a primer binding sequence in a single stranded polynucleotide molecule of interest such as one that is to be amplified. The primer binding sequences will generally be of known sequence and will therefore be complementary to a region of known sequence of the single stranded polynucleotide

molecule. The capture oligonucleotides may include a capture sequence and an amplification sequence. For example, a capture oligonucleotide may be of greater length than amplification primers that are attached to the same substrate, in which case the 5' end of the capture sequences may comprise a region with the same sequence as one of the amplification primers. A portion of a template, such as the 3' end of the template, may be complementary to the 3' of the capture sequences. The 5' end of the template may contain a region that comprises a sequence identical to one of the amplification primers such that upon copying the template, the copy can hybridize to the immobilized amplification primer. Thus, an oligonucleotide species that is useful in the methods set forth herein can have a capture sequence, an amplification sequence or both. Conversely, an oligonucleotide species can lack a capture sequence, an amplification sequence or both. In this way the hybridization specificity of an oligonucleotide species can be tailored for a particular application of the methods.

The length of primer binding sequences need not be the same as those of known sequences of polynucleotide template molecules and may be shorter, being particularly 16-50 nucleotides, more particularly 16-40 nucleotides and yet more particularly 20-30 nucleotides in length. The desired length of the primer oligonucleotides will depend upon a number of factors. However, the primers are typically long (complex) enough so that the likelihood of annealing to sequences other than the primer binding sequence is very low. Accordingly, known sequences that flank a template sequence can include a primer binding portion and other portions such as a capture sequence, barcode sequence or combination thereof.

In certain embodiments of the invention, amplification primers for solid phase amplification are immobilized by covalent attachment to the solid support at or near the 5' end of the primer, such that a portion of the primer is free to anneal to its cognate template and the 3' hydroxyl group is free to function in primer extension.

The chosen attachment chemistry will typically depend on the nature of the solid support and any functionalization or derivatization applied to it. In the case of nucleic acid embodiments, the primer itself may include a moiety which may be a non-nucleotide chemical modification to facilitate attachment. For example, the primer may include a sulfur containing nucleophile such as a phosphorothioate or thiophosphate at the 5' end. In the case of solid supported polyacrylamide hydrogels, this nucleophile may bind to a bromoacetamide group present in the hydrogel. In one embodiment, the means of attaching primers to the solid support is via St

phosphorothioate attachment to a hydrogel comprised of polymerized acrylamide and N-(5-bromoacetamidylpentyl) acrylamide (BRAPA).

A uniform, homogeneously distributed lawn of immobilized oligonucleotides may be formed by coupling (grafting) a solution of oligonucleotide species onto the solid support. The solution can contain a homogenous population of oligonucleotides but will typically contain a mixture of different oligonucleotide species. The mixture can include, for example, at least two, three or more different species of oligonucleotide. Each surface that is exposed to the solution therefore reacts with the solution to create a uniform density of immobilized sequences over the whole of the exposed solid support. As such, a portion of the surface having a mixture of different immobilized sequences can be surrounded by an area of the surface having a mixture of the same immobilized sequences. A suitable density of amplification oligonucleotides is at least 1 fmol/mm^2 (6 x 10^{10} per cm^2), or more optimally at least 10 fmol/mm^2 (6 x 10^{11} per cm^2). The density of the capture oligonucleotides can be controlled to give an optimum cluster density of 10^6 - 10^9 clusters per cm^2 and optimum cluster brightness. The ratio of capture oligonucleotide species to the amplification oligonucleotide species can be any desired value including, but not limited to at least 1:100, 1:1000 or 1:100000 depending on the desired cluster density and brightness. Similar densities or ratios of other molecular species can be used in embodiments where molecules other than nucleic acids are attached to a surface.

In a particular embodiment, for each cluster of template molecules, a complementary copy of a single stranded polynucleotide template molecule is attached to the solid support by hybridization. Methods of hybridization for formation of stable duplexes between complementary sequences by way of Watson-Crick base-pairing are known in the art. The immobilized capture oligonucleotides can include a region of sequence that is complementary to a region or template specific portion of the single stranded template polynucleotide molecule. An extension reaction may then be carried out in which the capture sequence is extended by sequential addition of nucleotides to generate a complementary copy of the single stranded polynucleotide sequence attached to the solid support via the capture oligonucleotide. The single stranded polynucleotide sequence not immobilized to the support may be separated from the complementary sequence under denaturing conditions and removed, for example by washing.

The terms separate and separating, when used in reference to strands of a nucleic acid, refer to the physical dissociation of the DNA bases that interact within for example, a Watson-

Crick DNA-duplex of the single stranded polynucleotide sequence and its complement. The terms also refer to the physical separation of these strands. Thus, the term can refer to the process of creating a situation wherein annealing of another primer oligonucleotide or polynucleotide sequence to one of the strands of a duplex becomes possible. After the first extension reaction, the duplex is immobilized through a single 5' attachment, and hence strand separation can result in loss of one of the strands from the surface. In cases where both strands of the duplex are immobilized, separation of the strands means that the duplex is converted into two immobilized single strands.

In one aspect of the invention, one or more of the amplification primers can be modified to prevent hybridization of a region or template specific portion of the single stranded polynucleotide molecule. Alternatively or additionally, one or more of the amplification primers may be modified to prevent extension of the primer during one or more extension reactions, thus preventing copying of the hybridized templates. These modifications can be temporary or permanent.

Generally, the capture sequences will include a region of the same sequence as the plurality of amplification oligonucleotides. Once the 3' end of the extended immobilized template copy has hybridized to one of the amplification primers and been extended, the resulting duplex will be immobilized at both ends and all of the bases in the capture oligonucleotide sequence will have been copied. Thus the capture oligonucleotide may include both the amplification primer sequence, plus a further sequence that is complementary to the end of the template. Typically the sequence complementary to the end of the template will not be present in any of the amplification primers. Alternatively, the amplification primers can contain the sequences complementary to the ends of the single stranded templates, but the amplification primers can be reversibly blocked to prevent hybridization and/or extension during one or more extension step, such as a first extension step in a particular amplification process.

According to one aspect of the invention, one or more of the amplification primers may include a modification that acts as a reversible block to either template hybridization or extension or both. By way of non-limiting example, such modifications can be presence of an additional sequence of nucleotides that is complementary to the amplification primer. This additional sequence can be present in a portion of the amplification primer and thus acts as an intramolecular hairpin duplex, or a 3' blocking group that prevents extension of the primer.

Alternatively, the additional sequence can be found on a separate oligonucleotide that hybridizes to the amplification primer. A particular feature of such a modification is that it can be removed, altered or reversed such that the functionality of the modified primer oligonucleotide is restored and the primer is able to undergo hybridization and extension during later steps of the methods. Among other examples, the blocking group may be a small chemical species such as a 3' phosphate moiety that can be removed enzymatically, may be an a basic nucleotide such that the 3' end of the primer is not capable of hybridization (and thereby extension), or may be a sequence of nucleotides that can be selectively excised from the immobilized strands, for example, using restriction endonucleases that selectively cleave particular sequences or deglycosylases that selectively cleave oligonucleotides having exogenous bases such as uracil deoxyribonucleotides or 8-oxoguanine.

In one embodiment a plurality of three types of oligonucleotides (for example comprising capture sequences, forward and reverse primers) are immobilized to a solid support. Alternatively the three oligonucleotides may be forward amplification, blocked forward amplification and reverse amplification, where the unblocked forward primer acts as the capture sequence.

The single stranded polynucleotide molecules may have originated in single-stranded form, as DNA or RNA or may have originated in double-stranded DNA (dsDNA) form (e.g. genomic DNA fragments, PCR and amplification products and the like). Thus a single stranded polynucleotide may be the sense or antisense strand of a polynucleotide duplex. Methods of preparation of single stranded polynucleotide molecules suitable for use in the method of the invention using standard techniques are well known in the art. The precise sequence of the primary polynucleotide molecules may be known or unknown during different steps of the methods set forth herein. It will be understood that a double stranded polynucleotide molecule can be hybridized to an immobilized capture oligonucleotide as exemplified herein for single stranded polynucleotide molecules, so long as a single stranded region of the double stranded polynucleotide is available and at least partially complementary to the capture oligonucleotide sequence.

An exemplary method for the isolation of one strand of a double stranded molecular construct is described herein. A sample of unknown sequence may be fragmented and have barcode sequences attached at both ends of the fragment. Adapters are then attached to the ends

of each fragment. One strand of the adapters may contain a moiety for surface immobilization, for example a biotin that can be captured onto a streptavidin surface. The adapters may be mismatch adapters, for example as described in co-pending application US 2007/0128624, the contents of which are incorporated herein by reference in their entirety. Amplification of the mismatch or forked adapters using a pair of amplification primers, one of which carries a biotin modification means that one strand of each duplex carries a biotin modification. Immobilization of the strands onto a streptavidin surface means that the non-biotinylated strand can be eluted simply by denaturation/strand separation. The eluted constructs will be in single stranded form and upon exposure to hybridization conditions can be used to hybridize against the immobilized capture sequences which can be extended.

In a particular embodiment, the single stranded polynucleotide molecules are DNA molecules. More particularly, the single stranded polynucleotide molecules represent genomic DNA molecules, or amplicons thereof, which include both intron and exon sequence (coding sequence), as well as non-coding regulatory sequences such as promoter and enhancer sequences. Still yet more particularly, the single stranded polynucleotide molecules are human genomic DNA molecules, or amplicons thereof.

In a particular embodiment, a single stranded target polynucleotide molecule has two regions of known sequence. Yet more particularly, the regions of known sequence will be at the 5' and 3' termini of the single stranded polynucleotide molecule such that the single stranded polynucleotide molecule will be of the structure:

5'[known sequence I]-[barcode sequence]-[target polynucleotide sequence]-[barcode sequence]-[known sequence II]-3'

Typically "known sequence I" and "known sequence II" will consist of more than 20, or more than 40, or more than 50, or more than 100, or more than 300 consecutive nucleotides. The precise length of the two sequences may or may not be identical. The primer binding sequences generally will be of known sequence and will therefore particularly be complementary to a sequence within “known sequence I” and “known sequence II” of the single stranded polynucleotide molecule. The length of the primer binding sequences need not be the same as those of known sequence I or II, and may be shorter, being particularly 16-50 nucleotides, more

particularly 16-40 nucleotides and yet more particularly 20-30 nucleotides in length. "Known sequence I" can be the same as "known sequence II" or the two can be different.

Methods of hybridization for formation of stable duplexes between complementary sequences by way of Watson-Crick base pairing are known in the art. A region or part of the single stranded polynucleotide template molecules can be complementary to at least a part of the immobilized capture sequence oligonucleotides. Since the amplification oligonucleotides are either modified to prevent hybridization and/or extension, or are non-complementary to the known ends of the template strands, only the capture sequences will be capable of hybridization and extension. An extension reaction may then be carried out wherein the capture sequence primer is extended by sequential addition of nucleotides to generate a complementary copy of the single stranded template polynucleotide attached to the solid support via the capture sequence oligonucleotide. The single stranded template polynucleotide sequence not immobilized to the support may be separated from the complementary sequence under denaturing conditions and removed, for example by washing. The distance between the individual capture sequence oligonucleotides on the surface therefore controls the density of the single stranded template polynucleotides and hence the density of clusters formed later on the surface is also controlled.

In certain embodiments in which the modified forward primer oligonucleotides are blocked and are unable to be extended, generally all of the amplification primer oligonucleotides will hybridize to the single stranded template polynucleotides. When the extension reaction is carried out only the unmodified forward capture primer oligonucleotides are extended by sequential addition of nucleotides to generate a complementary copy of the single stranded template polynucleotide attached to the solid support via the unmodified forward primer oligonucleotide. The single stranded template polynucleotide sequences not hybridized to the support may be separated from the un-extended blocked forward primer oligonucleotides under denaturing conditions and removed, for example by washing with a chemical denaturant such as formamide. The distance between the individual unmodified forward primer oligonucleotides on the surface therefore controls the density of the single stranded template polynucleotides and hence the density of clusters formed later on the surface is also controlled.

Following the attachment of the complementary single stranded template polynucleotides, the modified/blocked primers can be treated to reverse, remove or alter the modification such that they become functionally equivalent to the unmodified forward primer

oligonucleotides. For example, the double stranded structure may be removed either by denaturation, for example by heating or treatment with an alkaline solution when it is formed by a separate hybridized polynucleotide. Alternatively, where the hybridized polynucleotide is covalently linked, enzymatic digestion could be used to sequence-selectively cleave the strand, followed by denaturation. Such methods for removing the double stranded structure are known in the art and would be apparent to the skilled person (Sambrook and Russell, Molecular Cloning, A Laboratory Manual, third edition, Cold Spring Harbor Laboratory Press (2001)).

In one embodiment of the invention, the single stranded template polynucleotide molecule can be attached to the solid support by ligation to double stranded primers immobilized to the solid support using ligation methods known in the art (Sambrook and Russell, supra). Such methods utilize ligase enzymes such as DNA ligase to effect or catalyze the joining of the ends of the two polynucleotide strands, in this case, the single stranded template polynucleotide molecule and the primer oligonucleotide ligate such that covalent linkages are formed. In this context "joining" means covalent linkage of two polynucleotide strands that were not previously covalently linked. Thus, an aim of the invention can also be achieved by modifying the 3' end of a subset of primer oligonucleotides such that they are unable to ligate to the single stranded template polynucleotides. By way of non-limiting example, the addition of 2'3'dideoxy AMP (dideoxyAMP) by the enzyme terminal deoxynucleotidyl transferase (TdT) effectively prevents T4 DNA ligase from ligating treated molecules together.

An alternative method would be to have the capture sequences as duplex strands and the amplification sequences as single strands. Upon ligation of the single strands to the capture duplexes (which would be the only immobilized species carrying a free 5' phosphate) the 3' end of the immobilized strand can be extended as described above. Upon denaturation of the hybridized template sequence, amplification of the immobilized strand can proceed as described. Other such methods for attaching single strands will be apparent to others skilled in the art.

In a next step according to particular embodiments of the present invention, suitable conditions are applied to the immobilized single stranded polynucleotide molecule and the plurality of amplification primer oligonucleotides such that the single stranded polynucleotide molecule hybridizes to an amplification primer oligonucleotide to form a complex in the form of a bridge structure. Suitable conditions such as neutralizing and/or hybridizing buffers are well known in the art (See Sambrook et al., supra; Ausubel et al., Current Protocols in Molecular

Biology, John Wiley and Sons, Baltimore, Md. (1998)). The neutralising and/or hybridising buffer may then be removed.

Next by applying suitable conditions for extension an extension reaction is performed. The primer oligonucleotide of the complex is extended by sequential addition of nucleotides to generate an extension product complimentary to the single stranded polynucleotide molecule. The resulting duplex is immobilized at both 5' ends such that each strand is immobilized.

Suitable conditions such as extension buffers/solutions comprising an enzyme with polymerase activity are well known in the art (See Sambrook et al., supra; Ausubel et al. supra). In a particular embodiment dNTP's may be included in the extension buffer. In a further embodiment dNTP's could be added prior to the extension buffer. This bridge amplification technique can be carried out as described, for example, in Adessi et al. (U.S. patent number 7,115,400), and Kawashima et al. (U.S. patent application number 2005/0100900), the contents of which are incorporated herein by reference.

Examples of enzymes with polymerase activity which can be used in the present invention are DNA polymerase (Klenow fragment, T4 DNA polymerase), heat-stable DNA polymerases from a variety of thermostable bacteria (such as Taq, VENT, Pfu, or Tfl DNA polymerases) as well as their genetically modified derivatives (TaqGold, VENTexo, or Pfu exo). A combination of RNA polymerase and reverse transcriptase can also be used to generate the extension products. Particularly the enzyme has strand displacement activity, more particularly the enzyme will be active at a pH of about 7 to about 9, particularly pH 7.9 to pH 8+, yet more particularly the enzymes are Est or Klenow.

The nucleoside triphosphate molecules used are typically deoxyribonucleotide triphosphates, for example dATP, dTTP, dCTP, dGTP, or are ribonucleoside triphosphates for example ATP, UTP, CTP, GTP, The nucleoside triphosphate molecules may be naturally or non-naturally occurring.

After the hybridization and extension steps, the support and attached nucleic acids can be subjected to denaturation conditions. A flow cell can be used such that, the extension buffer is generally removed by the influx of the denaturing buffer. Suitable denaturing buffers are well known in the art (See Sambrook et al., supra; Ausubel et al. supra). By way of example it is known that alterations in pH and low ionic strength solutions can denature nucleic acids at substantially isothermal temperatures. Formamide and urea form new hydrogen bonds with the

bases of nucleic acids disrupting hydrogen bonds that lead to Watson-Crick base pairing. In a particular embodiment the concentration of formamide is 50% or more. These result in single stranded nucleic acid molecules. If desired, the strands may be separated by treatment with a solution of very low salt (for example less than 0.01 M cationic conditions) and high pH (>12) or by using a chaotropic salt (e.g. guanidinium hydrochloride). In a particular embodiment a strong base is used. A strong base is a basic chemical compound that is able to deprotonate very weak acids in an acid base reaction. The strength of a base is indicated by its pK.sub.b value, compounds with a pK_b value of less than about 1 are called strong bases and are well known to one skilled in the art. In a particular embodiment the strong base is Sodium Hydroxide (NaOH) solution used at a concentration of from 0.05 M to 0.25 M, particularly 0.1 M.

Following the hybridization, extension and denaturation steps exemplified above, two immobilized nucleic acids will be present, the first being the first template single stranded polynucleotide molecule (that was initially immobilized) and the second being a nucleic acid complementary thereto, extending from one of the immobilized primer oligonucleotides. Both the original immobilized single stranded polynucleotide molecule and the immobilized extended primer oligonucleotide formed are then able to initiate further rounds of amplification by subjecting the support to further cycles of hybridization, extension and denaturation.

It may be advantageous to perform optional washing steps in between each step of the amplification method. For example an extension buffer without polymerase enzyme with or without dNTP's could be applied to the solid support before being removed and replaced with the full extension buffer.

Such further rounds of amplification can be used to produce a nucleic acid colony or cluster comprising multiple immobilized copies of the single stranded polynucleotide sequence and its complementary sequence.

The initial immobilization of the single stranded polynucleotide molecule means that the single stranded polynucleotide molecule can hybridize with primer oligonucleotides located at a distance within the total length of the single stranded polynucleotide molecule. Other surface bound primers that are out of reach will not hybridize to the polynucleotide. Thus the boundary of the nucleic acid colony or cluster formed is limited to a relatively local area surrounding the location in which the initial single stranded polynucleotide molecule was immobilized.

Once more copies of the single stranded polynucleotide molecule and its complement

have been synthesized by carrying out further rounds of amplification, i.e. further rounds of hybridization, extension and denaturation, then the boundary of the nucleic acid colony or cluster being generated will be able to be extended further, although the boundary of the colony formed is still limited to a relatively local area around the location in which the initial single stranded polynucleotide molecule was immobilized. For example the size of each amplified cluster may be 0.5-5 microns.

It can thus be seen that the method of the present invention allows the generation of a plurality of nucleic acid colonies from multiple single immobilized single stranded polynucleotide molecules and that the density of these colonies can be controlled by altering the proportions of modified capture/amplification oligonucleotides used to graft the surface of the solid support.

In one embodiment, the hybridization, extension and denaturation steps are all carried out at the same, substantially isothermal temperature. For example the temperature is from 37°C to about 75°C, particularly from 50°C to 70°C, yet more particularly from 60°C to 65°C. In a particular embodiment the substantially isothermal temperature may be the optimal temperature for the desired polymerase.

In a particular aspect, the method according to the first aspect of the invention is used to prepare clustered arrays of nucleic acid colonies, analogous to those described in U.S. Pat. No. 7,115,400, US 2005/0100900 A1, WO 00/18957 and WO 98/44151 (the contents of which are herein incorporated by reference), by solid-phase amplification.

In yet another aspect more than one capture sequences and more than two amplification sequences, for example, at least three or four or more, different amplification primer sequences may be grafted to the solid support. In this manner more than one library, with common sequences which differ between the libraries, could be utilized to prepare clusters, such as, for example libraries prepared from two different patients. Whilst the cluster may overlap in space, they would be able to be sequenced one after the other due to the differences between the ends of the templates. For example, two different samples can be captured using two different capture sequences. These can be amplified from the same two amplification primers. The samples can be differentiated due to the two different capture sequences, which can be used as the sites for hybridization of two different sequencing primers. The use of different capture sequences thereby gives rise to a method of sample indexing using different sequencing primers.

Clustered arrays formed by the methods of the invention are suitable for use in applications usually carried out on ordered arrays such as micro-arrays. Such applications by way of non-limiting example include hybridization analysis, gene expression analysis, protein binding analysis, sequencing, genotyping, nucleic acid methylation analysis and the like. The clustered array may be sequenced before being used for downstream applications such as, for example, hybridization with fluorescent RNA or binding studies using fluorescent labelled proteins.

Sequencing Methods

The invention also encompasses methods of sequencing amplified nucleic acids generated by solid-phase amplification. Thus, the invention provides a method of nucleic acid sequencing comprising amplifying a pool of nucleic acid templates using solid-phase amplification as described above and carrying out a nucleic acid sequencing reaction to determine the sequence of the whole or a part of at least one amplified nucleic acid strand produced in the solid-phase amplification reaction.

Sequencing can be carried out using any suitable sequencing technique. A particularly useful method is one wherein nucleotides are added successively to a free 3' hydroxyl group, resulting in synthesis of a polynucleotide chain in the 5' to 3' direction. The nature of the nucleotide added may be determined after each nucleotide addition or at the end of the sequencing process. Sequencing techniques using sequencing by ligation, wherein not every contiguous base is sequenced, and techniques such as massively parallel signature sequencing (MPSS) where bases are removed from, rather than added to the strands on the surface are also within the scope of the invention.

The initiation point for the sequencing reaction may be provided by annealing of a sequencing primer to a product of the solid-phase amplification reaction. In this connection, one or both of the adaptors added during formation of the template library may include a nucleotide sequence which permits annealing of a sequencing primer to amplified products derived by whole genome or solid-phase amplification of the template library.

The products of solid-phase amplification reactions wherein both forward and reverse amplification primers are covalently immobilized on the solid surface are so-called bridged structures formed by annealing of pairs of immobilized polynucleotide strands and immobilized

complementary strands, both strands being attached to the solid support at the 5' end. Arrays comprised of such bridged structures provide inefficient templates for typical nucleic acid sequencing techniques, since hybridization of a conventional sequencing primer to one of the immobilized strands is not favored compared to annealing of this strand to its immobilized complementary strand under standard conditions for hybridization.

In order to provide more suitable templates for nucleic acid sequencing, it may be advantageous to remove or displace substantially all or at least a portion of one of the immobilized strands in the bridged structure in order to generate a template which is at least partially single-stranded. The portion of the template which is single-stranded will thus be available for hybridization to a sequencing primer. The process of removing all or a portion of one immobilized strand in a `bridged` double-stranded nucleic acid structure may be referred to herein as linearization, and is described in further detail in WO07010251, the contents of which are incorporated herein by reference in their entirety.

Bridged template structures may be linearized by cleavage of one or both strands with a restriction endonuclease or by cleavage of one strand with a nicking endonuclease. Other methods of cleavage can be used as an alternative to restriction enzymes or nicking enzymes, including inter alia chemical cleavage (e.g. cleavage of a diol linkage with periodate), cleavage of abasic sites by cleavage with endonuclease (for example `USER`, as supplied by NEB, part number M5505S), or by exposure to heat or alkali, cleavage of ribonucleotides incorporated into amplification products otherwise comprised of deoxyribonucleotides, photochemical cleavage or cleavage of a peptide linker.

Following the cleavage step, regardless of the method used for cleavage, the product of the cleavage reaction may be subjected to denaturing conditions in order to remove the portion(s) of the cleaved strand(s) that are not attached to the solid support. Suitable denaturing conditions, for example sodium hydroxide solution, formamide solution or heat, will be apparent to the skilled reader with reference to standard molecular biology protocols (Sambrook et al., supra; Ausubel et al. supra). Denaturation results in the production of a sequencing template which is partially or substantially single-stranded. A sequencing reaction may then be initiated by hybridization of a sequencing primer to the single-stranded portion of the template.

Thus, the invention encompasses methods wherein the nucleic acid sequencing reaction comprises hybridizing a sequencing primer to a single-stranded region of a linearized

amplification product, sequentially incorporating one or more nucleotides into a polynucleotide strand complementary to the region of amplified template strand to be sequenced, identifying the base present in one or more of the incorporated nucleotide(s) and thereby determining the sequence of a region of the template strand.

One sequencing method which can be used in accordance with the invention relies on the use of modified nucleotides having removable 3' blocks, for example as described in WO04018497, US 2007/0166705A1 and U.S. Pat. No. 7,057,026, the contents of which are incorporated herein by reference in their entirety. Once the modified nucleotide has been incorporated into the growing polynucleotide chain complementary to the region of the template being sequenced there is no free 3'-OH group available to direct further sequence extension and therefore the polymerase can not add further nucleotides. Once the nature of the base incorporated into the growing chain has been determined, the 3' block may be removed to allow addition of the next successive nucleotide. By ordering the products derived using these modified nucleotides, it is possible to deduce the DNA sequence of the DNA template. Such reactions can be done in a single experiment if each of the modified nucleotides has a different label attached thereto, known to correspond to the particular base, to facilitate discrimination between the bases added during each incorporation step. Alternatively, a separate reaction may be carried out containing each of the modified nucleotides separately.

The modified nucleotides may carry a label to facilitate their detection. A fluorescent label, for example, may be used for detection of modified nucleotides. Each nucleotide type may thus carry a different fluorescent label, for example, as described in WO07135368, the contents of which are incorporated herein by reference in their entirety. The detectable label need not, however, be a fluorescent label. Any label can be used which allows the detection of an incorporated nucleotide.

One method for detecting fluorescently labeled nucleotides comprises using laser light of a wavelength specific for the labelled nucleotides, or the use of other suitable sources of illumination. The fluorescence from the label on the nucleotide may be detected by a CCD camera or other suitable detection means. Suitable instrumentation for recording images of clustered arrays is described in WO07123744, the contents of which are incorporated herein by reference in their entirety.

The invention is not intended to be limited to use of the sequencing method outlined

above, as essentially any sequencing methodology which relies on successive incorporation of nucleotides into a polynucleotide chain can be used. Suitable alternative techniques include, for example, the Genome Sequencers from Roche/454 Life Sciences (Margulies et al. (2005) Nature, 437:376-380; U.S. Pat. Nos. 6,274,320; 6,258,568; 6,210,891), and the SOLiD system from Applied Biosystems (solid.appliedbiosystems.com), and the sequencer from Ion Torrent (www.iontorrent.com).

Incorporation by Reference

References and citations to other documents, such as patents, patent applications, patent publications, journals, books, papers, web contents, have been made throughout this disclosure. All such documents are hereby incorporated herein by reference in their entirety for all purposes.

Equivalents

The invention may be embodied in other specific forms without departing from the spirit or essential characteristics thereof. The foregoing embodiments are therefore to be considered in all respects illustrative rather than limiting on the invention described herein.

What is claimed is:

1. A method for validating the integrity of a nucleic acid template in a multiplex sequencing reaction, the method comprising:

obtaining a template nucleic acid;

incorporating at least two sequence identifiers into the template; and

sequencing the template.

2. The method according to claim 1, wherein a first identifier is incorporated into a 5' end of the template and a second identifier is incorporated into a 3' end of the template.

3. The method according to claim 2, wherein the first and second identifiers are the same.

4. The method according to claim 2, wherein the first and second identifiers are different.

5. The method according to claim 1, wherein prior to sequencing, the template is attached to a substrate.

6. The method according to claim 2, wherein the template is directly attached to the substrate.

7. The method according to claim 2, wherein the template is indirectly attached to the substrate.

8. The method according to claim 5, further comprising, amplifying the template.

9. The method according to claim 1, wherein sequencing is sequencing by synthesis.

10. The method according to claim 9, wherein the sequencing by synthesis is single molecule sequencing by synthesis.

11. The method according to claim 1, wherein sequencing comprises:

hybridizing a primer to the template to form a template/primer duplex;

contacting the duplex with a polymerase enzyme in the presence of a detectably labeled nucleotides under conditions that permit the polymerase to add nucleotides to the primer in a template-dependent manner;

detecting a signal from the incorporated labeled nucleotide; and

sequentially repeating the contacting and detecting steps at least once, wherein sequential detection of incorporated labeled nucleotide determines the sequence of the nucleic acid.

12. The method according to claim 11, wherein the detectably labeled nucleotides are optically labeled nucleotides.

13. The method according to claim 12, wherein the optically labeled nucleotides are fluorescently labeled nucleotides.

14. The method according to claim 1, wherein the pair of identifiers are first and second barcode sequences that are attached to the template.

15. The method according to claim 14, wherein a first barcode sequence is attached to a 5' end of the template and a second barcode sequence is attached to a 3' end of the template.

16. The method according to claim 15, wherein the first and second barcode sequences are the same.

17. The method according to claim 15, wherein the first and second barcode sequences are different.

18. A method for analyzing a nucleic acid, the method comprising:

obtaining a nucleic acid template;

attaching a pair of barcode sequences to the template; and

sequencing the template.

19. The method according to claim 18, wherein a first barcode sequence is attached to a 5' end of the template and a second barcode sequence is attached to a 3' end of the template.

20. The method according to claim 18, wherein prior to sequencing, the template is attached to a substrate.

21. The method according to claim 20, wherein the template is directly attached to the substrate.

22. The method according to claim 20, wherein the template is indirectly attached to the substrate.

23. The method according to claim 20, further comprising, amplifying the template.

24. The method according to claim 18, wherein sequencing is sequencing by synthesis.

25. The method according to claim 24, wherein the sequencing by synthesis is single molecule sequencing by synthesis.

26. The method according to claim 18, wherein sequencing comprises:

hybridizing a primer to the template to form a template/primer duplex;

contacting the duplex with a polymerase enzyme in the presence of a detectably labeled nucleotides under conditions that permit the polymerase to add nucleotides to the primer in a template-dependent manner;

detecting a signal from the incorporated labeled nucleotide; and

sequentially repeating the contacting and detecting steps at least once, wherein sequential detection of incorporated labeled nucleotide determines the sequence of the nucleic acid.

27. The method according to claim 26, wherein the detectably labeled nucleotides are optically labeled nucleotides.

28. The method according to claim 27, wherein the optically labeled nucleotides are fluorescently labeled nucleotides.

29. The method according to claim 19, wherein the first and second barcode sequences are the same.

30. The method according to claim 19, wherein the first and second barcode sequences are different.

31. A method for analyzing a nucleic acid, the method comprising:

obtaining a nucleic acid template;

attaching a first barcode sequence to a 5' end of the template and a second barcode sequence to a 3' end of the template;

amplifying the template; and

sequencing the template.

32. The method according to claim 31, wherein prior to amplifying, the template is immobilized to a solid support.

33. The method according to claim 31, wherein the first and second barcode sequences are the same.

34. The method according to claim 31, wherein the first and second barcode sequences are different.

35. The method according to claim 31, wherein sequencing comprises:

hybridizing a primer to the template to form a template/primer duplex;

contacting the duplex with a polymerase enzyme in the presence of detectably labeled nucleotides under conditions that permit the polymerase to add nucleotides to the primer in a template-dependent manner;

detecting a signal from the incorporated labeled nucleotide; and

sequentially repeating the contacting and detecting steps at least once, wherein sequential detection of incorporated labeled nucleotide determines the sequence of the nucleic acid.

36. A method for validating detection of an analyte, the method comprising detecting the presence of two or more identifiers that are uniquely associated with said analyte.

37. The method of claim 36, wherein said identifiers are nucleic acid sequence tags.

38. The method of claim 37, wherein said sequence tags are associated with a moiety that binds said analyte.

39. The method of claim 36, wherein said analyte is a protein.

40. The method of claim 39, wherein said identifiers are peptides.

Abstract

The invention generally relates to methods for maintaining the integrity and identification of a nucleic acid template in a multiplex sequencing reaction. In certain embodiments, methods of the invention involve obtaining a template nucleic acid, incorporating a pair of sequence identifiers into the template, and sequencing the template.

1795192

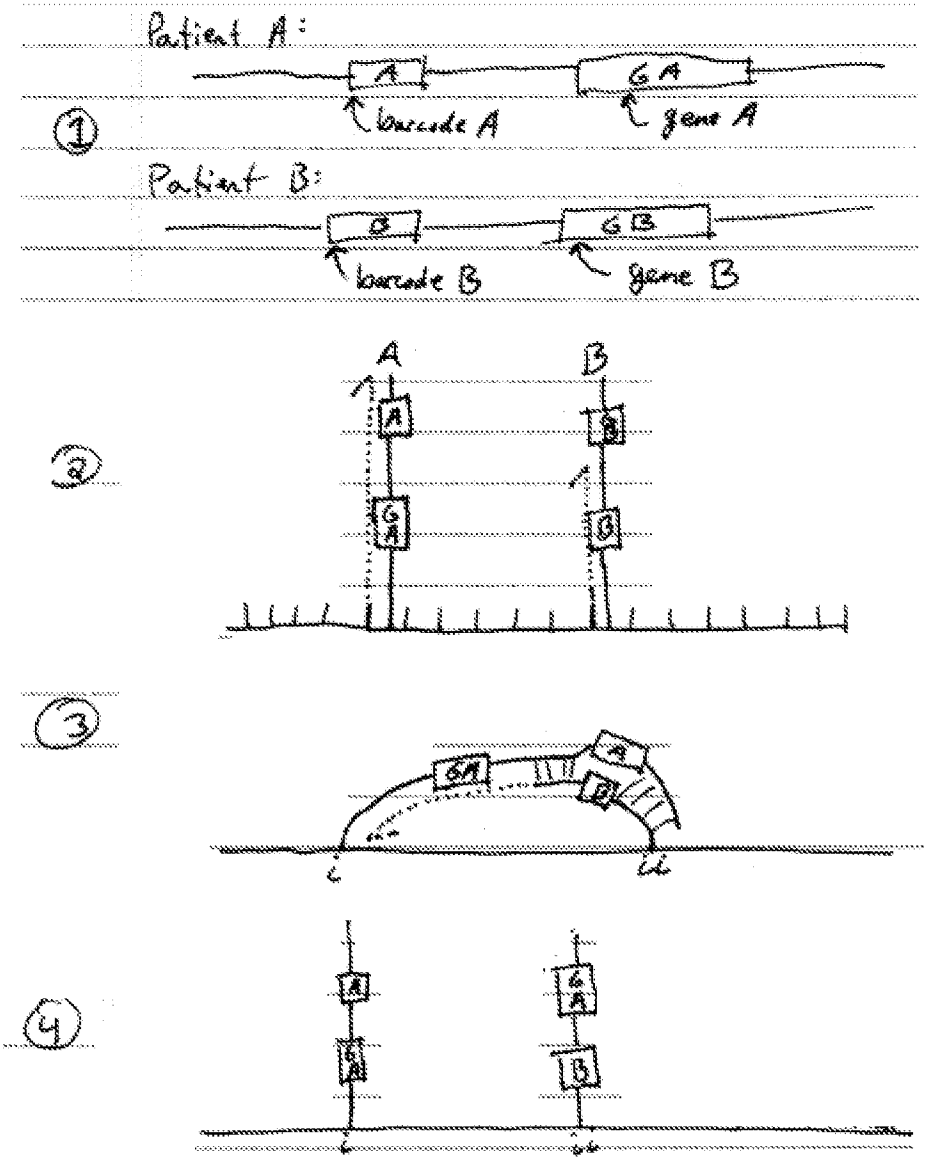
Patient A:
A
GA
barcode A
gene A
1
Patient B:
B
GB
barcode B
gene B
A
B
2
3
4

FIGURE 1

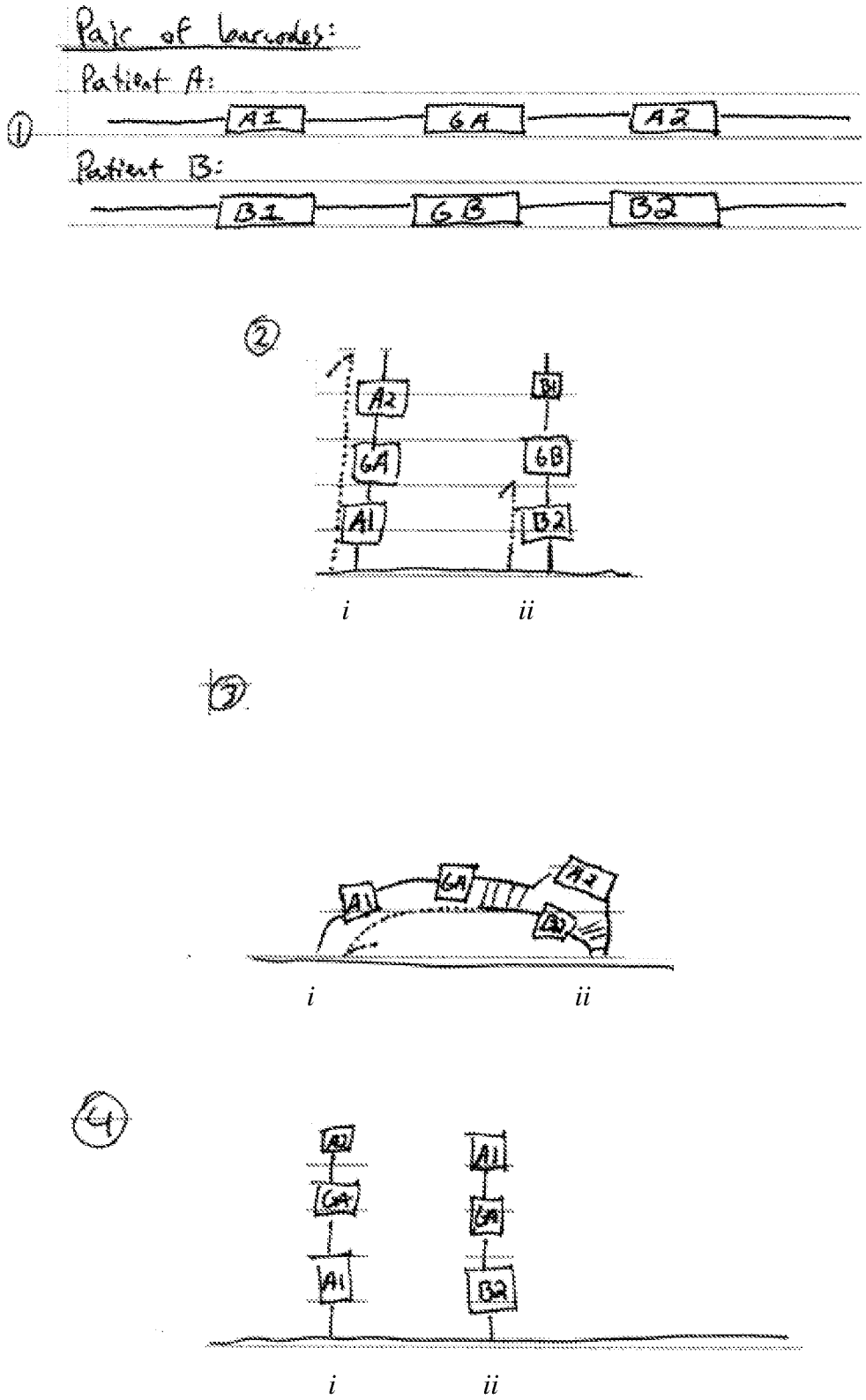
Pair of barcodes:
Patient A:
A1
GA
A2
Patient B:
B1
GB
B2
i
ii
i
ii
i
ii

FIGURE 2