



US 20060008824A1

(19) **United States**

(12) **Patent Application Publication**
Ronaghi et al.

(10) **Pub. No.: US 2006/0008824 A1**

(43) **Pub. Date: Jan. 12, 2006**

(54) **METHODS AND COMPOSITIONS FOR
CLONAL AMPLIFICATION OF NUCLEIC
ACID**

(75) Inventors: **Mostafa Ronaghi**, Palo Alto, CA (US);
Foad Mashayekhi, San Francisco, CA
(US)

Correspondence Address:
Ropes & Gray LLP
One International Place
Boston, MA 02110-2624 (US)

(73) Assignee: **Leland Stanford Junior University**,
Palo Alto, CA (US)

(21) Appl. No.: **11/134,683**

(22) Filed: **May 19, 2005**

Related U.S. Application Data

(60) Provisional application No. 60/573,719, filed on May
20, 2004.

Publication Classification

(51) **Int. Cl.**
C12Q 1/68 (2006.01)
C12P 19/34 (2006.01)
(52) **U.S. Cl.** **435/6; 435/91.2**

(57) **ABSTRACT**

Described herein are methods and compositions relating to
amplifying nucleic acid. In certain embodiments, the inven-
tion provides methods for labeling and amplifying a nucleic
acid molecule.

Figure 1.

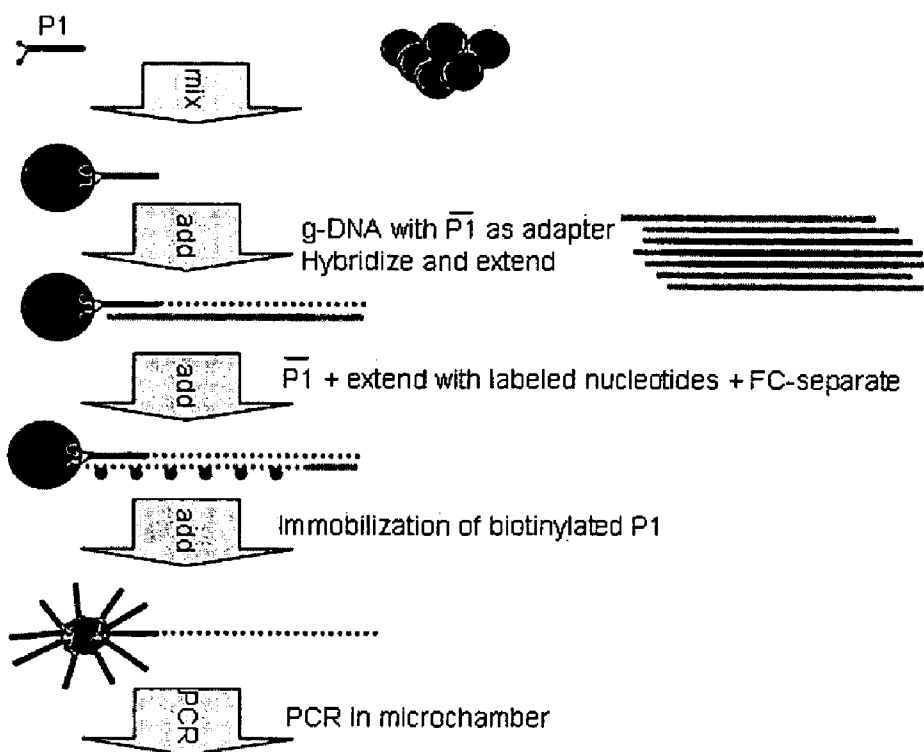
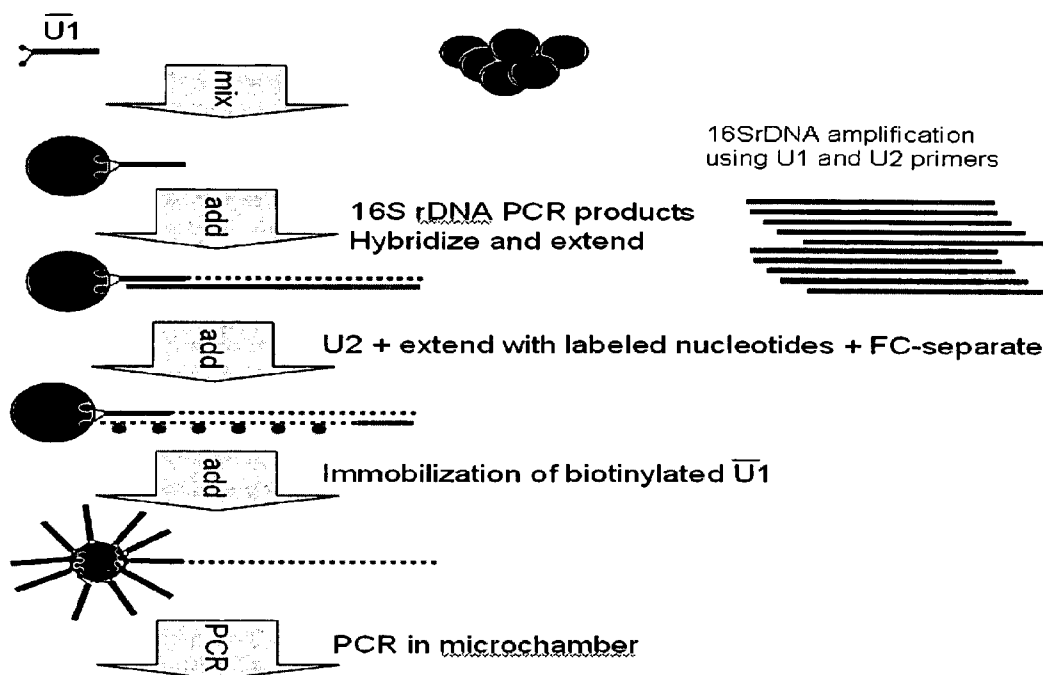


Figure 2.



METHODS AND COMPOSITIONS FOR CLONAL AMPLIFICATION OF NUCLEIC ACID

RELATED APPLICATION

[0001] This application claims the benefit of the filing date of U.S. Provisional Application No. 60/573,719 filed May 20, 2004 and entitled "Methods and Compositions for Clonal Amplification of Nucleic Acid" by Mostafa Ronaghi and Foad Mashayekhi. The entire teachings of the referenced Provisional Application are incorporated herein by reference.

FUNDING

[0002] Work described herein was funded, in whole or in part, by National Institutes of Health Grant Number 2PO1HG00205. The United States government has certain rights in the invention.

BACKGROUND

[0003] The essence of biology is an understanding of all of the species and their biological mechanisms. Speciation and biological function are primarily determined by the organism's DNA sequence. The development of vastly improved DNA sequence determination for personalized medicine and ecological studies could complete the revolution initiated by the Human Genome Project.

[0004] The Human Genome Project was markedly accelerated during the last several years and was completed with blistering speed. The complete nucleotide sequence of the human genome was completed in approximately 13 years (Collins et al., 2003) and the genome was published in 2003 (www.genome.gov/11006929). There are currently over 3×10^{10} nucleotides in public databases (www.ncbi.nlm.nih.gov/genbank/genbankstats), and genome sequences of over 185 organisms have been fully sequenced, as well as parts of the genomes of over 100,000 taxonomic species (www.integratedgenomics.com/gold).

[0005] The Human Genome Project was essentially accomplished by a reduction in the cost of DNA sequencing by three orders of magnitude. To reduce the cost by another two or three orders of magnitude, a highly integrated platform will be needed. Although the Human Genome Project took gel-based Sanger sequencing and achieved a decrease in cost and increase in throughput by over three orders of magnitude, the project was unable to develop any competitive alternative technology for genome sequencing.

[0006] Large bacterial genomes and genomes from complex organisms are often fragmented in two steps in order to simplify the assembly process. After constructing large-insert libraries, each clone is further fragmented and smaller libraries are prepared for bi-directional shotgun sequencing. Although the material cost of creating libraries is minimal, the process has proven to be laborious for large-scale genome sequencing. Therefore, alternative strategies need to be developed.

SUMMARY

[0007] The present invention provides a method of labeling and amplifying a nucleic acid molecule. In certain embodiments, the subject invention provides a method of labeling and amplifying a nucleic acid molecule, comprising

attaching a unique primer to the nucleic acid molecule to be amplified and amplifying said labeled nucleic acid molecule. In certain embodiments, the unique primer is an adapter comprising a unique barcode. In additional embodiments, the unique primer is attached to the nucleic acid molecule by ligating the adapter comprising a unique barcode to the nucleic acid molecule. In further embodiments, the unique barcode comprises a unique sequence of five nucleotides. In yet other embodiments, the unique primer identifies a source of nucleic acid. In additional embodiments, the unique barcode identifies a source of nucleic acid. In certain embodiments, the source of nucleic acid is selected from the group consisting of: a BAC clone, a bacterium, and a tissue sample. In other embodiments, the labeled nucleic acid molecule is amplified by a sequencing-by-synthesis method. In certain embodiments, the labeled nucleic acid molecule is amplified by a method selected from the group consisting of: massively parallel signature sequencing, BEAM, and polony technology.

[0008] The invention additionally relates to a method of amplifying a nucleic acid molecule, comprising (a) labeling the nucleic acid molecule with a unique primer; (b) hybridizing said nucleic acid molecule to one or more anti-primers immobilized on a solid surface; (c) incubating the solid surface of (b) with polymerase and nucleotides under conditions permitting extension of said one or more anti-primers; (d) incubating the solid surface of (c) under conditions permitting denaturing of the nucleic acid molecules from the extended anti-primers; removing the supernatant; incubating the solid surface with polymerase, labeled nucleotides, and one or more oligonucleotides that hybridize to the immobilized extended anti-primers of (c), under conditions permitting extension of said hybridized oligonucleotides; and identifying said solid surfaces comprising the extended hybridized oligonucleotides; (e) immobilizing additional anti-primers on the solid surfaces of (d); and (f) incubating the solid surfaces of (e) with polymerase, nucleotides, and one or more oligonucleotides that hybridize to the immobilized anti-primers of (e), under conditions permitting extension of said hybridized oligonucleotides, wherein the nucleic acid molecule is thereby amplified. An "anti-primer" is an oligonucleotide that hybridizes to a unique primer of the subject invention. In certain embodiments, an anti-primer will be complementary, in whole or in part, to a unique primer of the subject invention. In certain embodiments, a unique primer is an adapter comprising a unique barcode. In further embodiments, the unique barcode comprises a unique sequence of five nucleotides. In additional embodiments, invention further relates to (g) sequencing the immobilized amplified nucleic acid molecules of (f). In certain embodiments of the invention, the nucleic acid molecules are sequenced by pyrosequencing. In further embodiments, the unique primer identifies a source of nucleic acid. In other embodiments, the unique barcode identifies a source of nucleic acid. In certain embodiments, the source of nucleic acid is selected from the group consisting of: a BAC clone, a bacterium, and a tissue sample. In certain embodiments, the solid surface is a bead. In certain further embodiments, the bead is selected from the group consisting of: a magnetic bead and a Sepharose bead.

[0009] The present invention additionally relates to a nucleic acid molecule attached to a solid surface and comprising a unique primer. In certain embodiments, the unique primer is an adapter that comprises a unique barcode. In

further embodiments, the unique barcode comprises a unique sequence of five nucleotides. In additional embodiments, the solid surface is a bead.

[0010] In additional embodiments, the present invention relates to a solid surface comprising one or more nucleic acid molecules prepared by a method comprising: (a) labeling the nucleic acid molecule with a unique primer; (b) hybridizing said nucleic acid molecule to one or more anti-primers immobilized on a solid surface; (c) incubating the solid surface of (b) with polymerase and nucleotides under conditions permitting extension of said one or more anti-primers; (d) incubating the solid surface of (c) under conditions permitting denaturing of the nucleic acid molecules from the immobilized extended anti-primers; removing the supernatant; incubating the solid surface with polymerase, labeled nucleotides, and one or more oligonucleotides that hybridize to the extended anti-primers of (c), under conditions permitting extension of said hybridized oligonucleotides; and identifying said solid surfaces comprising the extended hybridized oligonucleotides; (e) immobilizing additional anti-primers on the solid surfaces of (d); and (f) incubating the solid surfaces of (e) with polymerase, nucleotides, and one or more oligonucleotides that hybridize to the immobilized anti-primers of (e), under conditions permitting extension of said hybridized oligonucleotides, wherein the nucleic acid molecule is thereby amplified.

[0011] In further embodiments, the present invention relates to a kit comprising: (a) a unique primer; (b) a solid surface comprising immobilized anti-primers thereon; (c) oligonucleotides that hybridize to the immobilized anti-primers of (b); and (d) instructions for amplifying a nucleic acid molecule. In certain embodiments, the unique primer of (a) is an adapter comprising a unique barcode. In additional embodiments, the unique barcode comprises a unique sequence of five nucleotides. In further embodiments, the instructions of (d) include instructions for amplifying a nucleic acid molecule according to a method comprising: (a) labeling the nucleic acid molecule with a unique primer; (b) hybridizing said nucleic acid molecule to one or more anti-primers immobilized on a solid surface; (c) incubating the solid surface of (b) with polymerase and nucleotides under conditions permitting extension of said one or more anti-primers; (d) incubating the solid surface of (c) under conditions permitting denaturing of the nucleic acid molecules from the extended anti-primers; removing the supernatant; incubating the solid surface with polymerase, labeled nucleotides, and one or more oligonucleotides that hybridize to the immobilized extended anti-primers of (c), under conditions permitting extension of said hybridized oligonucleotides; and identifying said solid surfaces comprising the extended hybridized oligonucleotides; (e) immobilizing additional anti-primers on the solid surfaces of (d); and (f) incubating the solid surfaces of (e) with polymerase, nucleotides, and one or more oligonucleotides that hybridize to the immobilized anti-primers of (e), under conditions permitting extension of said hybridized oligonucleotides, wherein the nucleic acid molecule is thereby amplified.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] FIG. 1 is a schematic relating to barcoded clonal amplification.

[0013] FIG. 2 is a schematic relating to clonal amplification.

DETAILED DESCRIPTION OF THE INVENTION

[0014] Described herein are methods for labelling nucleic acid molecules for clonal amplification and identification. The present invention relates to labelling nucleic acid molecules with a nucleic acid barcode, which enables easy identification of the nucleic acid after amplification and sequencing of the barcoded nucleic acid molecule. The present invention also relates to barcodes, which are nucleic acid tags that may be ligated to a nucleic acid molecule that is to be amplified and, in certain embodiments, sequenced. In certain preferred embodiments, the barcodes of the present invention comprise five nucleotides, preferably in a uniquely identifiable sequence.

[0015] In certain embodiments, the subject invention relates to a nucleic acid molecule that includes an adapter. An adapter of the invention comprises nucleic acid that hybridizes to a primer, such as an immobilized primer described herein. In certain embodiments, the present invention relates to a nucleic acid molecule that includes a barcode sequence. For example, in certain embodiments, the present invention relates to a nucleic acid molecule that includes an adapter that comprises a barcode sequence. In certain further embodiments, the invention provides a distinct barcode sequence such that each member of, e.g., a Bacterial Artificial Chromosome (BAC) library, can later be identified if isolated individually or as part of an enriched population of nucleic acid molecules. For example, a BAC library may be fragmented into 1000 base pair-long DNA fragments. The fragmented genomic DNA can then be saturated with a unique barcode for ligation. Genomic DNA from one BAC will then carry a unique barcode (e.g., five nucleotides in a unique sequence) that can later be identified easily, for example, after amplification and sequencing of the genomic DNA fragment. The barcoded clonal amplification described herein can be used to facilitate the sequencing of a genome. For example, the present invention provides for easier fragment assembly of genomic DNA fragments, such as barcoded nucleic acid fragments as described herein that have been amplified and sequenced.

[0016] Genome sequencing can be divided into three steps: upstream processing including sample preparation, DNA sequencing, and fragment assembly.

[0017] Prior to sequencing, a genome is usually fragmented into smaller pieces. Viral genomes are relatively small and can often be sequenced in one step of a restriction digest. However, large bacterial genomes and genomes from complex organisms are often fragmented in two steps in order to simplify the assembly process. Genomes are currently being partially restricted with enzymes into large fragments of 100-250 kb in size and cloned into Bacterial Artificial Chromosomes (BACs) and P1-derived Artificial Chromosomes (PACs). BAC vectors are preferable for the construction of genomic libraries mostly due to their smaller vector size (about 8 kb versus about 15 kb for PACs) which has a favorable effect on the economies of clone-by-clone shotgun sequencing. After constructing these large-insert libraries, each clone is further fragmented and smaller libraries are prepared for bi-directional shotgun sequencing. Although the material cost of creating libraries is minimal, the process has proven to be laborious for large-scale genome sequencing.

[0018] An alternative strategy is subcloning the large-insert and smaller genomes onto beads, a process known as clonal amplification. This process involves chopping the DNA into small pieces and cloning the chopped DNA onto beads prior to amplification, producing a unique set of DNA on each bead. Initially, each bead will carry a single DNA molecule. After amplification, each bead will contain thousands to millions of identical copies of the DNA molecule. This amplification is important in order to achieve sufficient signal for detection.

[0019] Several techniques have been introduced and tested, most notably polony technology (Mitra and Church, 1999), BEAM by Vogelstein's group (Dressman et al., 2003), and a cloning strategy developed for massively parallel signature sequencing (MPSS) (Brenner et al., 2000b). 454 Life Sciences, which employs Pyrosequencing as a technique for genome sequencing, has used a micro-emulsion technique to clone and amplify (Leamon and al., 2003) unique DNA fragments onto the beads, a strategy similar to BEAMing. This technique has been used to sequence six genomes, one of which is the Adenovirus genome and has been put in public databases (Sarkis and al., 2003). All the above-mentioned methods for clonal amplification may be appropriate for sequencing-by-synthesis.

[0020] Overview of Ultra-Fast DNA Sequencing (UDS):

[0021] With the availability of the human genome sequence, increasing emphasis has been focused on technologies for genotyping or resequencing specific areas of the genome for SNP discovery. Currently, chip-based techniques have dominated this field (Elahi et al., 2004). However, it is believed that de novo sequencing will be needed to develop a deeper understanding beyond polygenic factors that contribute to human diseases. When evaluating a technology, the following parameters are usually considered: the throughput, cost per instrument, cost per base, accuracy, and read-length of individual fragments. Here, we summarize a set of UDS technologies that have the potential to offer de novo sequencing fulfilling the above-mentioned criteria although they are still in relatively early stages of development. These technologies include electrophoretic methods, sequencing-by-hybridization (SBH), sequencing-by-synthesis (SBS), massively parallel signature sequencing (MPSS), and non-enzymatic single-molecule sequencing methods.

[0022] Electrophoretic Sequencing:

[0023] The well-proven and elegant Sanger DNA sequencing method (Sanger et al., 1977), which is based on the electrophoretic separation of DNA fragments with single-base resolution, is the dominant DNA sequencing technology and has served the community for 27 years. The current cost per base is about 0.1 cents for advanced Genome Centers, however, it ranges from 5 to 20 cents per base for other laboratories. Typically, 99.99% accuracy can be achieved with as few as three raw reads covering a given nucleotide. This technology is currently being miniaturized and integrated by several groups, including the Mathies group (Enrich et al., 2002) and researchers at the Whitehead BioMEMS laboratory in collaboration with Network BioSciences (Koutny and Al., 2000; Paegel et al., 2003).

[0024] Sequencing-By-Hybridization (SBH):

[0025] SBH was proposed by several groups more than 15 years ago (Bains and Smith, 1988; Drmanac, 1998; Fodor et

al., 1995; Southern, 1989). The basic principle of SBH is that differential hybridization of oligonucleotide probes can be used to decode a target DNA sequence. The most common approach involves immobilizing the DNA to be sequenced on a solid support and then performing serial hybridizations with short probe oligonucleotides (5 to 8 nucleotide long). The extent to which specific probes bind can be used to infer the unknown sequence. The strategy has been applied to both genome resequencing and de novo sequencing (Drmanac and al., 2001; Drmanac, 1998). Affymetrix and its spin-off Perlegen have used a different approach to SBH by hybridizing sample DNA to high density oligonucleotide arrays (Lipshutz and al., 1995). The current maximum density of Affymetrix arrays is about one oligonucleotide feature per 5 micron square, where each feature consists of roughly 100,000 copies of a defined 25 base pair oligonucleotide. For each base pair of a reference genome to be resequenced, there are four features on the chip. By hybridizing labeled sample DNA to the chip and determining which of the four features yields the strongest signal for each base pair in the reference sequence, a DNA sample can be rapidly resequenced. Miniaturization, bioinformatics, and the availability of a reference human genome sequence has permitted Perlegen to greatly extend this approach and develop an oligonucleotide array for resequencing human chromosome 21 (Patil et al., 2001).

[0026] SBH technology possesses a unique set of advantages and challenges. It can be used to obtain an impressive amount of sequence (>109 bases) from many distinct chromosomes. The primary challenge of SBH is the cross-hybridization of probes to incorrect targets due to repetitive elements or chance similarities. These factors render a substantial fraction of Chromosome 21 (more than 50%) inaccessible (Patil et al., 2001), and might also contribute to the 3% false-positive SNP detection rate observed in that study. It is also worth noting that SBH still requires sample preparation steps, as the relevant fraction of the genome must be PCR-amplified prior to hybridization.

[0027] Sequencing-By-Synthesis (SBS):

[0028] SBS methods generally involve enzymatic extension by polymerase through the iterative addition of nucleotides. Each nucleotide addition, called a cycle, only queries one or a few bases, but thousands or millions of features can be processed in parallel when the assay is performed in array format. These array features may be ordered or randomly dispersed. The progression of the primer strand can be monitored either by fluorescence using labeled nucleotides or enzymatically using natural nucleotides. Several groups reported on SBS using fluorescent nucleotides more than a decade ago (Metzker et al., 1994). Other groups have made advances in microfabrication and integration of SBS using fluorescent nucleotides. Most notable is the work performed by Quake's (Braslavsky et al., 2003) and Church's (Mitra et al., 2003; Mitra and Church, 1999) groups. Pyrosequencing is another SBS-based technique which was developed recently (Ronaghi, 2001; Ronaghi et al., 1996; Ronaghi et al., 1998). All of the above-mentioned SBS methods require an amplification step prior to sequencing in order to achieve the required signal fidelity.

[0029] Several groups have worked on developing single-molecule DNA sequencing techniques based on SBS, most notably Solexa (www.solexa.com), Genovox (www.genovox.com).

voxx.de), Nanofluidics (Levene and al., 2003) (in collaboration with Craighead's group), and Helicos (Braslavsky et al., 2003) (in collaboration with Quake's group). With respect to the ease and reliability of detecting extension events, SBS methods that sequence amplified molecules have an obvious advantage over single-molecule methods.

[0030] Massively Parallel Signature Sequencing (MPSS):

[0031] This method relies on cycles of restriction digestion and ligation (Brenner et al., 2000a). In MPSS, array features are sequenced at each cycle by employing a type II restriction enzyme to cleave within a target sequence, leaving a four base-pair overhang. Sequence-specific ligation of a fluorescent linker is then used to query the identity of the overhang. The achievable 8 to 16 bp read-lengths (which involves 3 to 4 cycles) are adequate for many purposes (Brenner et al., 2000a), but not for whole-genome sequencing.

[0032] Non-Enzymatic Single-Molecule DNA Sequencing:

[0033] Another UDS single-molecule approach is nanopore sequencing (Deamer and Akeson, 2000; Deamer and Branton, 2002; Li et al., 2003). As DNA passes through a nanopore, different base-pairs obstruct the pore to varying degrees, resulting in fluctuations in the electrical conductance of the pore. The pore conductance can therefore be measured and used to infer the DNA sequence. The accuracy of base-calling ranges from 60% for single events to 99.9% for 15 events (Winters-Hilt et al., 2003). However, the method has thus far been limited to the terminal base-pairs of a specific type of hairpin. This method has a great deal of long-term potential for extraordinarily rapid sequencing with little to no sample preparation. However, it is likely that significant pore engineering will be necessary before we can achieve a single-base resolution. Rather than engineering a pore to probe single nucleotides, Visigen (<http://www.visigenbio.com/tech.html>) and Li-Cor (Williams, 2001) are attempting to engineer DNA polymerases or fluorescent nucleotides to provide real-time, base-specific signals while synthesizing DNA at its natural pace.

[0034] Pyrosequencing:

[0035] Pyrosequencing relies on the real-time detection of inorganic pyrophosphate (PPi) released on successful incorporation of nucleotides during DNA synthesis. PPi is immediately converted to adenosine triphosphate (ATP) by ATP sulfurylase, and the level of ATP generated is detected via luciferase-producing photons. Unused ATP and deoxynucleotides are washed in the three-enzyme system of Pyrosequencing (Ronaghi et al., 1996) and degraded in the four-enzyme system of Pyrosequencing (Ronaghi et al., 1998) by the nucleotide-degrading enzyme apyrase. The presence or absence of PPi, and therefore the incorporation or nonincorporation of each nucleotide added, is ultimately assessed on the basis of whether or not photons are detected.

[0036] There is a minimal time lapse between these events, and the conditions of the reaction are such that iterative addition of nucleotides and PPi detection are possible. Prior to the start of the Pyrosequencing reactions, an amplicon is generated by PCR in which one of the primers is biotinylated at its 5' terminus. The biotinylated double-stranded DNA PCR products are then linked to a solid surface coated with streptavidin and denatured. The two

strands are separated, and the strand bound to the solid surface is usually used as the template. After hybridizing a sequencing primer to this strand, DNA synthesis under Pyrosequencing conditions can commence.

[0037] In Pyrosequencing, 1 pmol of DNA template molecules can generate the same number of ATP molecules per nucleotide incorporated, which, in turn, can generate more than 6×10^9 photons at a wavelength of 560 nm. This amount of light is easily detected by a photodiode, photomultiplier tube, or charge-coupled device (CCD) camera.

Experimental Design And Methods:

[0038] The procedure for the standard Pyrosequencing reaction comprises in vitro amplification, template preparation, and Pyrosequencing. For highly parallel genome sequencing the amplification step will be rate-limiting. To circumvent this rate limiting step, clonal amplification is proposed, in a step where large numbers of DNA are clonally amplified on a single bead.

[0039] The subject invention relates to a labelling strategy, which will allow multiplexing the sequencing of a large number of clones within a single chamber. This labelling step can also be applied to the strategies described above and by others (Brenner et al., 2000b; Dressman et al., 2003; Leamon and al., 2003; Mitra and Church, 1999). Described herein is a strategy to clonally amplify a unique DNA on a solid surface such as magnetic beads.

[0040] Preparation of Beads Carrying Single P1 Molecule:

[0041] P1 is a 30-base long general oligonucleotide, which can be any unique sequence such as, e.g., 5'-TTTTTTTG-CAAATGTTATCGAGGTCCGGC, that has two biotin molecules at the 5' end. Two biotin molecules permit efficient immobilization and also allow use of a higher temperature without being detached from the streptavidin molecule (Dressman et al., 2003). The ratio of magnetic beads to P1 primers is designed to be large so that most beads that capture a primer molecule will capture exactly one. The solution will be saturated with fluorescently labeled anti-P1 (seen as P1 bar in FIG. 1), which will allow isolation of all beads carrying P1 and recycling of the remaining beads.

[0042] Preparation of Unique Adapters:

[0043] The adapters are double stranded oligonucleotides. For example, adapters will be synthesized on the order of 100 unique adapters. The coding strand of the adapter carries the same sequence as P1 followed by a unique sequence of five nucleotides (red part of the oligonucleotide). For the complementary strand of the adapter, 100 different coding strands will be synthesized. Synthesis of non-coding strands starts with five unique nucleotides (complementary to the one chosen for the coding strands) followed by the complementary sequence of P1 (shown as P1 bar in FIG. 1). The 5' end of the latter strand will be phosphorylated. Both strands are mixed together in 100 different tubes to make 100 different adapters.

[0044] Preparation of Genomic DNA for Clonal Amplification:

[0045] The BAC library will be nebulized to roughly 1000 bp-long DNA fragments. The fragmented DNA from each BAC will then be end repaired and gel purification will be used to select DNA fragment sizes between 800-1200 nucle-

otides long. Next, the fragmented genomic DNA will be saturated with one of the unique adapters for ligation. At this step, the adapters can simply be ligated from one end (the end carrying phosphorylated 5'-end). Genomic DNA carrying adapters will be gel purified from non-ligated adapters. Now, all genomic DNA from one BAC will carry a unique barcode (a unique sequence of 5 nucleotides) that can later be identified easily in from the read.

[0046] Single Molecule Capture of Genomic DNA on Bead and Extension:

[0047] Since the beads selected previously have a single attached P1 primer, they can only capture a single DNA fragment. Therefore, we saturate the beads with denatured genomic DNA for hybridization overnight. Though there are two capture sites on each DNA fragment, at least 10 cycles of extension are performed (each cycle comprises denaturation, hybridization, and extension) with polymerase and nucleotides present; therefore, all captured DNA fragments should be extended. After these extension cycles, all the magnetic beads are isolated and washed.

[0048] Labeling of Captured DNA:

[0049] Isolated beads from the previous step will be treated with alkali to denature the DNA and then the supernatant is removed. Anti-P1 is added together with polymerase and nucleotides. The nucleotides contain a small proportion of labeled nucleotides (labeled dUTP), which allows us to label the captured DNA and later subject it to flow cytometry for isolation.

[0050] Saturation of Beads with Biotinylated P1:

[0051] Biotinylated P1 is added to all isolated beads to saturate the beads with immobilized P1. All beads are isolated, and the leftovers are removed.

[0052] Clonal Amplification on Microfluidic Chambers:

[0053] The microfluidic chambers will be filled with isolated beads. Anti-P1 primer is added together with the PCR solution and cycles of amplification are performed to extend all P1 primers. After amplification, all the left-over substrates are removed and beads are treated with alkali to remove the non-biotinylated strand. Anti-P1 may then be added to perform, e.g., Pyrosequencing.

[0054] Sepharose beads can also be used instead of magnetic beads. In certain embodiments of the invention, Streptavidin-coated Sepharose beads may be used. For example, Sepharose beads that can be used include version 1.1 of the integrated chip on a PCR machine from MJ Research. The well geometry of this platform is similar to the one from 454 Corporation (Leamon and al., 2003). For clonal amplification using magnetic beads, version 1.2 of the integrated chip may be used, which has almost 6 times smaller wells. The same PCR conditions can be applied.

[0055] Clonal Amplification Using Plasmid DNA:

[0056] In certain embodiments, the methods of the invention may additionally be applied to a plasmid that is chopped with a shearing machine to 100-nucleotide long DNA. The shearing machine provides very homogenous fragment size down to 50 bases. The chopped plasmid DNA will be ligated to adapters and gel purified. Subsequently fragments will be cloned onto 30 μ m Streptavidin-coated Sepharose beads and the above-mentioned process will be followed. The beads

will then be loaded into the microfluidic chip with 35 μ m well sizes bundled with the 128-feature CMOS chip. Clonally amplified DNA on beads will be subjected to complete extension having all four nucleotides present with all Pyrosequencing reagents. The efficiency will be measured by quantifying the light signal generated from Pyrosequencing extension reaction. The clonal amplification strategy of the present invention may be applied to different lengths of DNA fragments up to 2000 nucleotides long. After clonal amplification, cyclic Pyrosequencing can be performed to deduce the sequence and assemble plasmid DNA. This step will indicate the predicted coverage in sequencing that is enough for accurate sequence determination. In addition, the efficiency of clonal amplification can be monitored and PCR performed in designed format.

[0057] Development of Assembly Methodology and Software Assembly:

[0058] The Pyrosequencing array will produce a large number of reads, each of which is prefixed by a pentamer that uniquely identifies the BAC clone source of the read within the array. With this information the reads can be separated into separate "clones". A whole-genome assembly methodology may comprise the following:

[0059] (1): Assemble the reads from each clone separately to obtain a set of contigs per clone. (2): Identify all pairwise overlaps between clone assemblies; filter out ones likely to be repeat-induced and obtain a set of unique overlaps. (3): Construct the overlap graph of contigs in clone assemblies, and perform sequencing-error correction. (4): Merge and extend contigs using the overlap graph, and use a merge operation between adjacent clones for further repeat resolution.

[0060] The principle of the subject methodology is that within short windows, the majority of the genome contains few or no high fidelity repeats that are longer than one read length. Therefore, sufficient shotgun sequencing from within the window will lead to a correct and gap-free assembly. To cluster reads within short windows, the following strategy may be followed: (A) in (1) above, reads are clustered within windows of one clone length; (1) will lead to reasonably contiguous assemblies of the majority of a mammalian genome. (B) In (4), reads are clustered into even shorter windows defined by successive BAC clone ends. The quality of assembly produced by the above procedure will depend on the following key parameters: the depth of clone coverage of the genome, the depth of shotgun sequence coverage of each clone, the read length, and the sequencing error rate. To assess assembly quality, computational simulations of sequencing may be performed as well as assembling genomes of varying size and repeat content and under different parameter settings. To develop such an assembly technique, Arachne (Batzoglou et al., 2002; Batzoglou et al., 2000; Jaffe et al., 2003) may be used as a module. Arachne is publicly available in source code and has been used successfully to assemble several genomes including the mouse (Mouse Genome Sequencing Consortium 2002).

[0061] The subject invention can also be used for expression or RNA profiling. For example, total RNA molecules can be extracted from a sample, such as a bacterial sample. Biotinylated oligo dT primers or biotinylated random hexamers are used to amplify the RNA molecules. In certain embodiments, biotinylated random pentamers, heptamers,

or octamers may be used. After amplification, the cDNA molecules are immobilized on a solid support. Enzymatic restriction can be performed to generate overhang ends onto which an adapter can be ligated. For multiplex analysis, different samples may be ligated with adapters containing a unique tag, such as a nucleic acid barcode as described herein. The fragments are then detached and the same process for clonal amplification can be performed as for genome sequencing as described above. In certain embodiments, the subject invention relates to a method of clonal amplification for nucleic acid isolated from a bacterial sample. For example, described herein is a strategy to clonally amplify a unique bacterial nucleic acid on a solid surface such as magnetic beads.

[0062] Preparation of Beads Carrying Single U1-Bar Molecule:

[0063] U1-BAR is a 30-base long general oligonucleotide having two biotin molecules at the 5' end. Two biotin molecules permit efficient immobilization and also allow use of a higher temperature without being detached from the streptavidin molecule (Dressman et al., 2003). The ratio of magnetic beads to U1-BAR primers is designed to be large so that most beads that capture a primer molecule will capture exactly one. The solution will be saturated with fluorescent labeled U1 (seen as U1-BAR bar in FIG. 2), which will allow isolation of all beads carrying U1-BAR and recycling of the remaining beads.

[0064] Preparation of 16S rDNA for Clonal Amplification:

[0065] Sample containing isolated bacterial DNA from flora will be used for amplification using U1 and U2 universal primers U1 5'-AGAGTTTGATITGGCTCAG-3' (*E. coli* positions 8 through 27, International Union of Biochemistry nomenclature) and U2 5'-CGGITACCTTGT-TACGACTT-3' (*Escherichia coli* positions 1493 through 1513), respectively. These primers amplify 97% of all bacteria available in the databases. Most of the remaining 3% of bacteria are ones living in environments such as high salt, hot spring water, etc.

[0066] Single Molecule Capture of 16S rDNA on Bead and Extension:

[0067] Since the beads selected previously have a single attached U1-BAR primer, they can only capture a single DNA fragment. Therefore, the beads are saturated with denatured 16S rDNA amplicons for hybridization overnight. After ten extension cycles using U2-bar, all the magnetic beads are isolated and washed.

[0068] Labeling of Captured DNA:

[0069] Isolated beads from the previous step will be treated with alkali to denature the DNA and then the supernatant is removed. Anti-U2-BAR is added together with polymerase and nucleotides. The nucleotides contain a small proportion of labeled nucleotides (labeled dUTP), which allow labeling the captured DNA and later subjecting it to flow cytometry for isolation.

[0070] Saturation of Beads with Biotinylated U1-BAR:

[0071] Biotinylated U1-BAR is added to all isolated beads to saturate the beads with immobilized U1-BAR. All beads are isolated and the leftovers are removed.

[0072] Clonal Amplification on Microfluidic Chambers:

[0073] The microfluidic chambers will be filled with isolated beads. U2-bar primer is added together with the PCR solution and cycles of amplification are performed to extend all U1-BAR primers. After amplification, all the left-over substrates are removed and beads are treated with alkali to remove the non-biotinylated strand. U2-bar is added to perform Pyrosequencing to generate signature sequences of different bacteria.

REFERENCES

- [0074]** Bains, W. and Smith, G. C. (1988) A novel method for nucleic acid sequence determination. *Journal of Theoretical Biology* 135(3), 303-7.
- [0075]** Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P. and Lander, E. S. (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Res* 12(1), 177-89.
- [0076]** Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. and Lander, E. S. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res* 10(7), 950-8.
- [0077]** Braslavsky, I., Hebert, B., Kartalov, E. and Quake, S. R. (2003) Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci USA* 100(7), 3960-4.
- [0078]** Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S., Moon, K., Burcham, T., Pallas, M., DuBridge, R., Kirchner, J., Fearon, K., Mao, J. and Corcoran, K. (2000a) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 18(6), 630-634.
- [0079]** Brenner, S., Williams, S. R., Vermaas, E. H., Storck, T., Moon, K., McCollum, C., Mao, J. I., Luo, S., Kirchner, J. J., Eletr, S., DuBridge, R. B., Burcham, T. and Albrecht, G. (2000b) In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. USA* 97(4), 1665-70.
- [0080]** Collins, F. S., Morgan, M. and Patrinos, A. (2003) The human genome project: lessons from large-scale biology. *Science* 300, 286-290.
- [0081]** Deamer, D. W. and Akeson, M. (2000) Nanopores and nucleic acids: prospects for ultrarapid sequencing. *Trends Biotechnol* 18, 147-151.
- [0082]** Deamer, D. W. and Branton, D. (2002) Characterization of nucleic acids by nanopore analysis. *Acc Chem Res* 35, 817-825.
- [0083]** Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. and Vogelstein, B. (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci USA* 100, 8817-8822.

- [0084] Drmanac, R. and al., e. (2001) DNA sequencing by hybridization with arrays of samples or probes. *Methods Mol Biol* 170, 173-179.
- [0085] Drmanac, S. (1998) Accurate sequencing by hybridization for DNA diagnostics and individual genomics. *Nature Biotechnology* 16, 54-58.
- [0086] Elahi, E., Kumm, J. and Ronaghi, M. (2004) Global genetic analysis. *J Biochem Mol Biol* 37(1), 11-27.
- [0087] Emrich, C. A., Tian, H., Medintz, I. L. and Mathies, R. A. (2002) Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis. *Anal. Chem.* 74, 5076-5083.
- [0088] Fodor, S. P. A., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. and Solas, D. (1995) Light-directed, spatially addressable parallel chemical synthesis. *Science* 251(4995), 767-773.
- [0089] Jaffe, D. B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J. P., Zody, M. and Lander, E. S. (2003) Whole-Genome Sequence Assembly for Mammalian Genomes: Arachne 2. *Genome Res* 13, 91-96.
- [0090] Koutny, L. and Al., E. (2000) Eight hundred-base sequencing in a microfabricated electrophoretic device. *Anal. Chem.* 72, 2288-3391.
- [0091] Leamon, J. H. and al., e. (2003) A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* 24, 3769-3777.
- [0092] Levene, M. J. and al., e. (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 299, 682-686.
- [0093] Li, J., Gershow, M., Stein, D., Brandin, E. and Golovchenko, J. A. (2003) DNA molecules and configurations in a solid-state nanopore microscope. *Nat Mater* 2, 611-615.
- [0094] Lipshutz, R. J. and al., e. (1995) Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* 19, 442-447.
- [0095] Metzker, M. L., Raghavachari, R., Richards, S., Jacutin, S. E., Civitello, A., Burgess, K. and Gibbs, R. A. (1994) Termination of DNA synthesis by novel 3'-modified-deoxyribonucleoside 5'-triphosphates. *Nucleic Acids Research* 22(20), 4259-67.
- [0096] Mitra, R. D., Butty, V. L., Shendure, J., Williams, B. R., Housman, D. E. and Church, G. M. (2003) Digital genotyping and haplotyping with polymerase colonies. *Proc Natl Acad Sci USA* 100(10), 5926-31.
- [0097] Mitra, R. D. and Church, G. M. (1999) In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res* 27, 1-6.
- [0098] Paegel, B. M., Blazej, R. G. and Mathies, R. A. (2003) Microfluidic devices for DNA sequencing: sample preparation and electrophoretic analysis. *Curr Opin Biotechnol* 14, 42-50.
- [0099] Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stem, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P. and Cox, D. R. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294(5547), 1719-23.
- [0100] Ronaghi, M. (1998) Pyrosequencing: a tool for sequence-based DNA analysis. Doctoral thesis ISBN 91-7170-297-0.
- [0101] Ronaghi, M. (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Res* 11(1), 3-11.
- [0102] Ronaghi, M. (2003) Pyrosequencing for SNP genotyping. *Methods Mol Biol* 212, 189-195.
- [0103] Ronaghi, M. and Elahi, E. (2002) Pyrosequencing for microbial typing. *J Chromatogr B Analyt Technol Biomed Life Sci.* 782, 67-72.
- [0104] Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry* 242(1), 84-9.
- [0105] Ronaghi, M., Nygren, M., Lundeberg, J. and Nyren, P. (1999) Analyses of secondary structures in DNA by pyrosequencing. *Analytical Biochemistry* 267(1), 65-71.
- [0106] Ronaghi, M., Uhlen, M. and Nyren, P. (1998) A sequencing method based on real-time pyrophosphate. *Science* 281(5375), 363.
- [0107] Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA.* 74(12), 5463-7.
- [0108] Sarkis, G. and al., e. (2003) Sequence analysis of the pAdEasy-1 recombinant adenoviral construct using the 454 Life Sciences sequencing-by-synthesis method. NCBI AY370911, gi:34014919.
- [0109] Southern, E. M. (1989) Analysing polynucleotide sequences. Patent WO/10977.
- [0110] Williams, J. (2001) Heterogenous assay for pyrophosphate, U.S. Pat. No. 6,306,607, USA.
- [0111] Winters-Hilt, S. and al., e. (2003) Accurate classification of basepairs on termini of single DNA molecules. *Biochys J* 84, 967-976.

INCORPORATION BY REFERENCE

[0112] All publications and patents mentioned herein are hereby incorporated by reference in their entirety as if each individual publication or patent was specifically and individually indicated to be incorporated by reference. In case of conflict, the present application, including any definitions herein, will control.

[0113] Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims.

 SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 3

<210> SEQ ID NO 1
 <211> LENGTH: 30
 <212> TYPE: DNA
 <213> ORGANISM: Artificial Sequence
 <220> FEATURE:
 <223> OTHER INFORMATION: P1 primer

<400> SEQUENCE: 1

ttttttttgc aaatgttatc gaggtccggc 30

<210> SEQ ID NO 2
 <211> LENGTH: 20
 <212> TYPE: DNA
 <213> ORGANISM: Artificial Sequence
 <220> FEATURE:
 <223> OTHER INFORMATION: U1 primer
 <220> FEATURE:
 <221> NAME/KEY: misc_feature
 <222> LOCATION: 11, 12
 <223> OTHER INFORMATION: n = inosine

<400> SEQUENCE: 2

agagtttgat nntggctcag 20

<210> SEQ ID NO 3
 <211> LENGTH: 20
 <212> TYPE: DNA
 <213> ORGANISM: Artificial Sequence
 <220> FEATURE:
 <223> OTHER INFORMATION: U2 primer
 <220> FEATURE:
 <221> NAME/KEY: misc_feature
 <222> LOCATION: 4
 <223> OTHER INFORMATION: n = inosine

<400> SEQUENCE: 3

cggtacacctt gttacgactt 20

What is claimed is:

1. A method of labeling and amplifying a nucleic acid molecule, comprising attaching a unique primer to the nucleic acid molecule to be amplified and amplifying said labeled nucleic acid molecule.

2. The method of claim 1, wherein the unique primer is an adapter comprising a unique barcode.

3. The method of claim 2, wherein the unique primer is attached to the nucleic acid molecule by ligating the adapter comprising a unique barcode to the nucleic acid molecule.

4. The method of claim 2, wherein the unique barcode comprises a unique sequence of five nucleotides.

5. The method of claim 1, wherein the unique primer identifies a source of nucleic acid.

6. The method of claim 2, wherein the unique barcode identifies a source of nucleic acid.

7. The method of claim 5 or claim 6, wherein the source of nucleic acid is selected from the group consisting of: a BAC clone, a bacterium, and a tissue sample.

8. The method of claim 1, wherein the labeled nucleic acid molecule is amplified by a sequencing-by-synthesis method.

9. The method of claim 8, wherein the labeled nucleic acid molecule is amplified by a method selected from the group consisting of: massively parallel signature sequencing, BEAM, and polony technology.

10. A method of amplifying a nucleic acid molecule, comprising

- (a) labeling the nucleic acid molecule with a unique primer;
- (b) hybridizing said nucleic acid molecule to one or more anti-primers immobilized on a solid surface;
- (c) incubating the solid surface of (b) with polymerase and nucleotides under conditions permitting extension of said one or more anti-primers;
- (d) incubating the solid surface of (c) under conditions permitting denaturing of the nucleic acid molecules from the extended anti-primers; removing the supernatant; incubating the solid surface with polymerase, labeled nucleotides, and one or more oligonucleotides that hybridize to the immobilized extended anti-prim-

ers of (c), under conditions permitting extension of said hybridized oligonucleotides; and identifying said solid surfaces comprising the extended hybridized oligonucleotides;

- (e) immobilizing additional anti-primers on the solid surfaces of (d); and
- (f) incubating the solid surfaces of (e) with polymerase, nucleotides, and one or more oligonucleotides that hybridize to the immobilized anti-primers of (e), under conditions permitting extension of said hybridized oligonucleotides, wherein the nucleic acid molecule is thereby amplified.
- 11. The method of claim 10, wherein the unique primer of (a) is an adapter comprising a unique barcode.
- 12. The method of claim 11, wherein the unique barcode comprises a unique sequence of five nucleotides.
- 13. The method of claim 10, further comprising:
 - (g) sequencing the immobilized amplified nucleic acid molecules of (f).
- 14. The method of claim 13, wherein the nucleic acid molecules are sequenced by pyrosequencing.
- 15. The method of claim 10, wherein the unique primer identifies a source of nucleic acid.
- 16. The method of claim 11, wherein the unique barcode identifies a source of nucleic acid.
- 17. The method of claim 15 or claim 16, wherein the source of nucleic acid is selected from the group consisting of: a BAC clone, a bacterium, and a tissue sample.
- 18. The method of claim 10, wherein the solid surface is a bead.
- 19. The method of claim 18, wherein the bead is selected from the group consisting of: a magnetic bead and a Sepharose bead.
- 20. A nucleic acid molecule attached to a solid surface and comprising a unique primer.
- 21. The method of claim 20, wherein the unique primer is an adapter that comprises a unique barcode.
- 22. The method of claim 21, wherein the unique barcode comprises a unique sequence of five nucleotides.
- 23. The nucleic acid molecule of claim 20, wherein the solid surface is a bead.
- 24. A solid surface comprising one or more nucleic acid molecules prepared by a method comprising:
 - (a) labeling the nucleic acid molecule with a unique primer;
 - (b) hybridizing said nucleic acid molecule to one or more anti-primers immobilized on a solid surface;
 - (c) incubating the solid surface of (b) with polymerase and nucleotides under conditions permitting extension of said one or more anti-primers;
 - (d) incubating the solid surface of (c) under conditions permitting denaturing of the nucleic acid molecules from the extended anti-primers; removing the supernatant; incubating the solid surface with polymerase, labeled nucleotides, and one or more oligonucleotides that hybridize to the immobilized extended anti-prim-

ers of (c), under conditions permitting extension of said hybridized oligonucleotides; and identifying said solid surfaces comprising the extended hybridized oligonucleotides;

- (e) immobilizing additional anti-primers on the solid surfaces of (d); and
- (f) incubating the solid surfaces of (e) with polymerase, nucleotides, and one or more oligonucleotides that hybridize to the immobilized anti-primers of (e), under conditions permitting extension of said hybridized oligonucleotides, wherein the nucleic acid molecule is thereby amplified.
- 25. A kit comprising:
 - (a) a unique primer;
 - (b) a solid surface comprising immobilized anti-primers thereon;
 - (c) oligonucleotides that hybridize to the immobilized anti-primers of (b); and
 - (d) instructions for amplifying a nucleic acid molecule.
- 26. The kit of claim 25, wherein the unique primer of (a) is an adapter comprising a unique barcode.
- 27. The kit of claim 26, wherein the unique barcode comprises a unique sequence of five nucleotides.
- 28. The kit of claim 25, wherein the instructions of (d) include instructions for amplifying a nucleic acid molecule according to a method comprising:
 - (a) labeling the nucleic acid molecule with a unique primer;
 - (b) hybridizing said nucleic acid molecule to one or more anti-primers immobilized on a solid surface;
 - (c) incubating the solid surface of (b) with polymerase and nucleotides under conditions permitting extension of said one or more anti-primers;
 - (d) incubating the solid surface of (c) under conditions permitting denaturing of the nucleic acid molecules from the extended anti-primers; removing the supernatant; incubating the solid surface with polymerase, labeled nucleotides, and one or more oligonucleotides that hybridize to the immobilized extended anti-primers of (c), under conditions permitting extension of said hybridized oligonucleotides; and identifying said solid surfaces comprising the extended hybridized oligonucleotides;
 - (e) immobilizing additional anti-primers on the solid surfaces of (d); and
 - (f) incubating the solid surfaces of (e) with polymerase, nucleotides, and one or more oligonucleotides that hybridize to the immobilized anti-primers of (e), under conditions permitting extension of said hybridized oligonucleotides, wherein the nucleic acid molecule is thereby amplified.

* * * * *