(12) STANDARD PATENT (11) Application No. **AU 2016222719 B2**
(19) AUSTRALIAN PATENT OFFICE

(54) Title
Methods for targeted nucleic acid sequence coverage

(51) International Patent Classification(s)
C12Q 1/68 (2006.01)

(21) Application No: **2016222719** (22) Date of Filing: **2016.02.24**

(87) WIPO No: **WO16/138148**

(30) Priority Data

(31) Number	(32) Date	(33) Country
62/119,996	**2015.02.24**	**US**
62/146,834	**2015.04.13**	**US**

(43) Publication Date: **2016.09.01**
(44) Accepted Journal Date: **2022.03.31**

(71) Applicant(s)
10X Genomics, Inc.

(72) Inventor(s)
Schnall-Levin, Michael;Jarosz, Mirna

(74) Agent / Attorney
Houlihan² Pty Ltd, PO Box 611, BALWYN NORTH, VIC, 3104, AU

(56) Related Art
US 2014/0378345 A1

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
1 September 2016 (01.09.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/138148 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/US2016/019382

(22) International Filing Date:
24 February 2016 (24.02.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/119,996 24 February 2015 (24.02.2015) US
62/146,834 13 April 2015 (13.04.2015) US

(71) Applicant: **10X GENOMICS, INC.** [US/US]; 7068 Koll Center Parkway, Suite 401, Pleasanton, CA 94566 (US).

(72) Inventors: **SCHNALL-LEVIN, Michael**; 7068 Koll Center Parkway, Suite 401, Pleasanton, CA 94566 (US). **JAROSZ, Mirna**; 7068 Koll Center Parkway, Suite 401, Pleasanton, CA 94566 (US).

(74) Agents: **TALUKDER, Gargi** et al.; Morgan, Lewis & Bockius Llp, One Market, Spear Tower, San Francisco, CA 94105 (US).

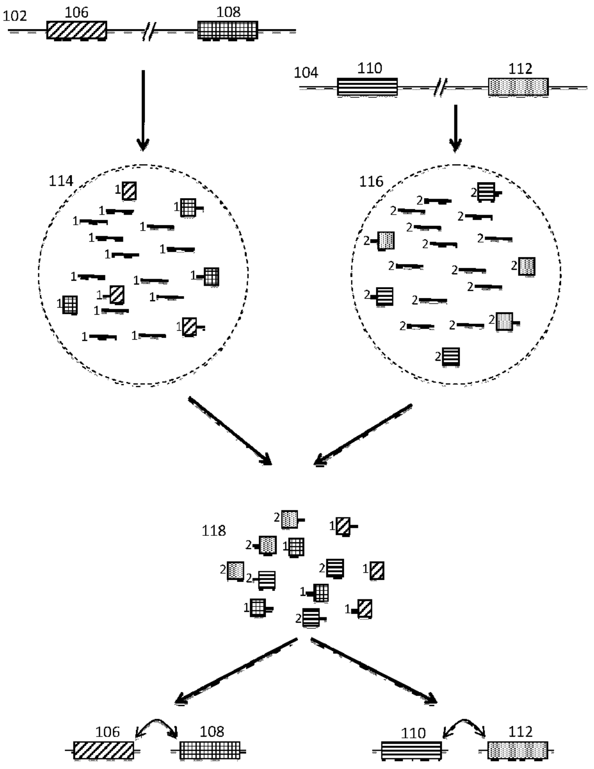
(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC,

[Continued on next page]

(54) Title: METHODS FOR TARGETED NUCLEIC ACID SEQUENCE COVERAGE

(57) **Abstract**: The present invention is directed to methods, compositions and systems for analyzing sequence information from targeted regions of a genome. Such targeted regions may include regions of the genome that are poorly characterized, highly polymorphic, or divergent from reference genome sequences.

Figure 1



SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

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

METHODS FOR TARGETED NUCLEIC ACID SEQUENCE COVERAGE

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of United States Provisional Application No. 62/146,834, filed April 13, 2015, and United States Provisional Application No. 62/119,996, filed February 24, 2015, which are hereby incorporated by reference in their entirety for all purposes.

BACKGROUND OF THE INVENTION

[0002] Despite significant progress in sequencing technologies, about 5-10% of the human genome remains unassembled, unmapped, and poorly characterized. The reference assembly generally annotates these missing regions as multi-megabase heterochromatic gaps. This missing fraction of the genome includes structural features that remain resistant to accurate characterization using generally used sequencing technologies. De novo sequencing of the entire genome is not economically feasible, and thus there remains a need to reduce the costs associated with genome sequencing while retaining the benefits of genomic analysis on a large scale.

SUMMARY OF THE INVENTION

[0003] Accordingly, the present disclosure provides methods, systems and compositions for providing targeted coverage of selected regions of the genome to allow for de novo sequence assembly of those selected regions, and in some aspects, allow for combining that de novo coverage with re-sequencing of remaining regions of the genome with high throughput and high accuracy.

[0004] In some aspects, the present disclosure provides a method for sequencing one or more selected portions of a genome in which the method includes the steps of: (a) providing starting genomic material; (b) distributing individual nucleic acid molecules from the starting genomic material into discrete partitions such that each discrete partition contains an individual nucleic acid molecule; (c) amplifying selected portions of at least some of the individual nucleic acid molecules in the discrete partitions to form a population of amplicons; (d) barcoding the population of amplicons to form a plurality of barcoded fragments of the amplicons, wherein fragments within a given discrete partition each comprise a common barcode, thereby associating each fragment with the individual nucleic acid molecule from which it is derived; (e) obtaining sequence information from the plurality of fragments thereby sequencing one or more selected portions of a genome.

[0005] In further embodiments and in accordance with above, the one or more selected portions of the genome comprise highly polymorphic regions of the genome. In still further embodiments, the sequencing of the one or more selected portions of the genome is a de-novo sequencing.

[0006] In still further embodiments and in accordance with any of the above, the amplifying comprises PCR amplification across a region of at least 3.5 megabasepairs (Mb). In yet further embodiments, the amplifying comprises a PCR amplification utilizing multiple primer pairs staggered across a region of at least 3.0 Mb.

[0007] In some embodiments and in accordance with any of the above, the sequencing reaction is a short read, high accuracy sequencing reaction. In further embodiments, the sequence information generated in the obtaining step retains the molecular context of its originating individual nucleic acid.

[0008] In certain embodiments and in accordance with any of the above, prior to the obtaining step, the plurality of fragments is further enriched for fragments comprising at least a portion of the one or more selected portions of the genome by: (i) hybridizing probes complementary to regions in or near the one or more selected portions of the genome to the fragments to form probe-fragment complexes; (ii) capturing probe-fragment complexes to a surface of a solid support.

[0009] In some embodiments and in accordance with any of the above, the barcoded fragments of the amplicons within the discrete partitions represent about 100X-5000X coverage of the one or more selected portions of the genome. In further embodiments, the barcoded fragments of the amplicons within the discrete partitions represent about 200X-1000X coverage of the one or more selected portions of the genome. In still further embodiments, the barcoded fragments of the amplicons within the discrete partitions represent at least 1000X coverage of the one or more selected portions of the genome. In yet further embodiments, the barcoded fragments of the amplicons within the discrete partitions represent at least 2000X or 5000X coverage of the one or more selected portions of the genome.

[0010] In further aspects, the present disclosure provides a method of obtaining sequence information from one or more poorly characterized portions of a genomic sample, where the method includes the steps of: (a) providing individual first nucleic acid fragment molecules of the genomic sample in discrete partitions; (b) fragmenting the individual first nucleic acid fragment molecules within the discrete partitions to create a plurality of second fragments from each of the individual first nucleic acid fragment molecules; (c) amplifying selected regions of the plurality of second fragments that are poorly characterized to form a population of amplicons; (d) attaching a common barcode sequence to the amplicons within each discrete partition such that each of the amplicons is attributable to the discrete partition in which it is contained; (e) identifying sequences of the amplicons, thereby obtaining sequence information from one or more poorly characterized portions of the genomic sample.

[0011] In certain embodiments, and in accordance with any of the above, the amplifying comprises PCR amplification across a region of at least 3.5 megabasepairs (Mb). In further embodiments, the amplifying comprises a PCR amplification utilizing multiple primer pairs staggered across a region of at least 3.0 Mb. In still further embodiments, the multiple primer pairs contain uracils to prevent amplification of the primer sequences.

[0012] In some embodiments, and in accordance with any of the above, the identifying step preserves the molecular context of the sequences of the amplicons, such that the identifying further comprises identifying amplicons derived from the same individual first nucleic acid fragment molecules. In further embodiments, the method further comprises linking two or more of the individual first fragment molecules in an inferred contig based upon overlapping sequences of the plurality of second fragments, wherein the inferred contig comprises a length N50 of at least 10kb.

[0013] In some embodiments, and in accordance with any of the above, the barcode sequence further comprises additional sequence segments. In further embodiments, additional sequence segments comprise one or more of a member selected from the group consisting of: primers, attachment sequences, random n-mer oligonucleotides, oligonucleotides comprising uracil nucleobases. In yet further embodiments, the barcode is selected from a library of at least 700,000 barcodes.

[0014] In some embodiments, and in accordance with any of the above, the genomic sample within each discrete partition comprises genomic DNA from a single cell. In further embodiments, each discrete partition comprises genomic DNA from a different chromosome.

[0015] In some embodiments, and in accordance with any of the above, the discrete partitions comprise droplets in an emulsion.

[0016] In some embodiments, and in accordance with any of the above, the barcoded amplicons within the discrete partitions represent about 1000X-5000X coverage of the one or more poorly characterized portions of the genome.

[0017] In further aspects, the present application provides a method for obtaining sequence information from one or more portions of a genomic sample while retaining molecular context, the method including the steps of: (a) providing starting genomic material; (b) distributing individual nucleic acid molecules from the starting genomic material into discrete partitions such that each discrete partition contains a first individual nucleic acid molecule; (c) providing a population enriched for fragments comprising at least a portion of the one or more selected portions of the genome; (d) attaching a common barcode sequence to the fragments within each discrete partition such that each of the fragments is attributable to the discrete partition in which it was contained; (e) obtaining sequence information from the fragments, thereby sequencing one or more targeted portions of the genomic sample while retaining molecular context.

[0018] In still further aspects, the present disclosure provides a method for obtaining sequence information from one or more portions of a genomic sample while retaining molecular context, the method including the steps of: (a) providing starting genomic material; (b) distributing individual nucleic acid molecules from the starting genomic material into discrete partitions such that each discrete partition contains a first individual nucleic acid molecule; (c) providing a population within at least some of the discrete partitions that is enriched for sequences of the fragments comprising at least a portion of the one or more selected portions of the genome; (d) attaching a common barcode sequence to the fragments within each discrete partition such that each of the fragments is attributable to the discrete partition in which it was contained; (e) separating discrete partitions containing fragments comprising at least a portion of the one or more selected portions of the genome from discrete partitions containing no fragments comprising the one or more selected portions of the genome; (f) obtaining sequence information from the fragments comprising at least a portion of the one or more selected portions of the genome, thereby sequencing one or more targeted portions of the genomic sample while retaining molecular context.

[0019] In further embodiments and in accordance with any of the above, the providing a population enriched for sequences of the fragments comprising at least a portion of the one or more selected portions of the genome comprises directed PCR amplification of the fragments comprising at least a portion of the one or more selected portions of the genome to produce a population of amplicons comprising at least a portion of the one or more selected portions of the genome. In still further embodiments, this providing step further comprises attaching a detectable label to the amplicons, which in some embodiments may include a fluorescent molecule. In yet further embodiments the step of separating discrete partitions containing fragments comprising at least a portion of the one or more selected portions of the genome from discrete partitions containing no fragments comprising the one or more selected portions of the genome includes sorting the partitions emitting a signal from the detectable labels from the partitions without such a signal.

[0020] In some embodiments and in accordance with any of the above, prior to obtaining sequence information from the fragments, the discrete partitions are combined and the fragments are pooled together. In further embodiments, the step of obtaining sequence information from the fragments is conducted in such a way as to maintain the molecular context of the sequences of the fragments, such that the identifying further comprises identifying fragments derived from the same first individual nucleic acid molecules. In still further embodiments, this obtaining of sequence information includes a sequencing reaction selected from the group consisting of: short read-length sequencing reactions and long read-length sequencing reactions. In yet further embodiments, the sequencing reaction is a short read, high accuracy sequencing reaction.

[0021] In some embodiments and in accordance with any of the above, the discrete partitions comprise droplets in an emulsion. In further embodiments, the barcoded fragments within the discrete partitions represent about 100X-5000X coverage of the one or more selected portions of the genome. In still further embodiments, the barcoded fragments within the discrete partitions represent about 200X-1000X coverage of the one or more selected portions of the genome. In yet further embodiments, the barcoded fragments of the amplicons within the discrete partitions represent at least 1000X coverage of the one or more selected portions of the genome. In still further embodiments, the barcoded fragments within the discrete partitions represent at least 2000X or 5000X coverage of the one or more selected portions of the genome.

[0022] In some aspects and in accordance with any of the above, the present disclosure provides methods for obtaining sequence information from one or more portions of a genomic sample while retaining molecular context, including the steps of (a) providing genomic material; (b) separating individual nucleic acid molecules from the genomic material to form separated individual nucleic acid molecules; (c) providing a population enriched for fragments comprising at least a portion of the one or more selected portions of the genome from the separated individual nucleic acid molecules. In certain embodiments, the separating step is accomplished using any method that allows for one or more nucleic acid molecules to be sorted and processed in relative isolation from other one or more nucleic acid molecules. In some embodiments, the separating is a physical separation into different compartments on a substrate or into distinct partitions. In further embodiments, at least a plurality of the fragments are attributable to the individual nucleic acid molecules from which they are derived. That attribution is obtained using any methods that allow designation of a particular fragment as originating with a particular individual nucleic acid molecule. In certain exemplary embodiments, that attribution is obtained by barcoding fragments. In further aspects, sequence information is obtained from the fragments, thereby sequencing one or more targeted portions of the genomic sample while retaining molecular context.

BRIEF DESCRIPTION OF THE DRAWINGS

[0023] FIGURE 1 provides a schematic illustration of identification and analysis of targeted genomic regions using conventional processes versus the processes and systems described herein.

[0024] FIGURE 2A and B provide schematic illustrations of identification and analysis of targeted genomic regions using processes and systems described herein.

[0025] FIGURE 3 illustrates a typical workflow for performing an assay to detect sequence information, using the methods and compositions disclosed herein.

[0026] FIGURE 4 provides a schematic illustration of a process for combining a nucleic acid sample with beads and partitioning the nucleic acids and beads into discrete droplets.

[0027] FIGURE 5 provides a schematic illustration of a process for barcoding and amplification of chromosomal nucleic acid fragments.

[0028] FIGURE 6A and B provide schematic illustrations of the use of barcoding of nucleic acid fragments in attributing sequence data to their originating source nucleic acid molecule.

[0029] FIGURE 7 provides a schematic illustration of an embodiment of the invention.

[0030] FIGURE 8 provides a schematic illustration of an embodiment of the invention.

[0031] FIGURE 9 shows data from an experiment comparing amplification reactions conducted with template compared with those containing no template (NTC).

[0032] FIGURE 10 shows data from amplification reactions conducted across a range of annealing temperatures.

DETAILED DESCRIPTION OF THE INVENTION

[0033] The practice of the present invention may employ, unless otherwise indicated, conventional techniques and descriptions of organic chemistry, polymer technology, molecular biology (including recombinant techniques), cell biology, biochemistry, and immunology, which are within the skill of the art. Such conventional techniques include polymer array synthesis, hybridization, ligation, phage display, and detection of hybridization using a label. Specific illustrations of suitable techniques can be had by reference to the example herein below. However, other equivalent conventional procedures can, of course, also be used. Such conventional techniques and descriptions can be found in standard laboratory manuals such as *Genome Analysis: A Laboratory Manual Series* (Vols. I-IV), *Using Antibodies: A Laboratory Manual, Cells: A Laboratory Manual, PCR Primer: A Laboratory Manual, and Molecular Cloning: A Laboratory Manual* (all from Cold Spring Harbor Laboratory Press), Stryer, L. (1995) *Biochemistry* (4th Ed.) Freeman, New York, Gait, *“Oligonucleotide Synthesis: A Practical Approach”* 1984, IRL Press, London, Nelson and Cox (2000), *Lehninger, Principles of Biochemistry* 3rd Ed., W. H. Freeman Pub., New York, N.Y. and Berg et al. (2002) *Biochemistry,* 5th Ed., W. H. Freeman Pub., New York, N.Y., all of which are herein incorporated in their entirety by reference for all purposes.

[0034] Note that as used herein and in the appended claims, the singular forms "a," "an," and "the" include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "a polymerase" refers to one agent or mixtures of such agents, and reference to "the method" includes reference to equivalent steps and methods known to those skilled in the art, and so forth.

[0035] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. All publications mentioned herein are incorporated herein by reference for the purpose of describing and disclosing devices, compositions, formulations and methodologies which are described in the publication and which might be used in connection with the presently described invention.

[0036] Where a range of values is provided, it is understood that each intervening value, to the tenth of the unit of the lower limit unless the context clearly dictates otherwise, between the upper and lower limit of that range and any other stated or intervening value in that stated range is encompassed within the invention. The upper and lower limits of these smaller ranges may independently be included in the smaller ranges is also encompassed within the invention, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either both of those included limits are also included in the invention.

[0037] In the following description, numerous specific details are set forth to provide a more thorough understanding of the present invention. However, it will be apparent to one of skill in the art that the present invention may be practiced without one or more of these specific details. In other instances, well-known features and procedures well known to those skilled in the art have not been described in order to avoid obscuring the invention.

[0038] As used herein, the term “comprising” is intended to mean that the compositions and methods include the recited elements, but not excluding others. “Consisting essentially of” when used to define compositions and methods, shall mean excluding other elements of any essential significance to the composition or method. “Consisting of” shall mean excluding more than trace elements of other ingredients for claimed compositions and substantial method steps. Embodiments defined by each of these transition terms are within the scope of this invention. Accordingly, it is intended that the methods and compositions can include additional steps and components (comprising) or alternatively including steps and compositions of no significance (consisting essentially of) or alternatively, intending only the stated method steps or compositions (consisting of).

[0039] All numerical designations, e.g., pH, temperature, time, concentration, and molecular weight, including ranges, are approximations which are varied (+) or (-) by increments of 0.1. It is to be understood, although not always explicitly stated that all numerical designations are preceded by the term “about”. The term “about” also includes the exact value “X” in addition to minor increments of “X” such as “X + 0.1” or “X – 0.1.” It also is to be understood, although not always explicitly stated, that the reagents described herein are merely exemplary and that equivalents of such are known in the art.

I. Overview

[0040] This disclosure provides methods, compositions and systems useful for characterization of genetic material. In particular, the methods, compositions and systems described herein provide increased and redundant coverage of selected portions of the genome such that additional redundant sequence information can be obtained from those selected portion of the genome. In specific instances, that additional sequence information provides enough information to allow for de novo sequencing of those selected portions of the genome.

[0041] In general, the methods, compositions, and systems described herein provide genetic characterization of selected regions of a genome. This genetic characterization is of sufficient depth to allow de novo sequencing of the selected regions of the genome. This de novo sequencing is of particular use for regions of the genome that are poorly characterized, are highly polymorphic, and/or diverge from reference sequences. As will be appreciated, a significant percentage (at least 5-10% according to, for example Altemose et al., *PLOS* Computational Biology, May 15, 2014, Vol. 10, Issue 5) of the human genome remains unassembled, unmapped, and poorly characterized. The reference assembly generally annotates these missing regions as multi-megabase heterochromatic gaps, found primarily near centromeres and on the short arms of the acrocentric chromosomes. This missing fraction of the genome includes structural features that remain resistant to accurate characterization using generally used sequencing technologies. Exemplary regions that are resistant to accurate characterization include areas that have close homologous pseudogenes (for example SMN1/2 CYP2D6), areas that have substantial repeated sequences throughout the genome, including without limitation transposons (such as SINEs, LINEs), and particularly areas that have tremendous variation for which reference sequences serve as a poor guide (such as the regions encoding the genes for the human leukocyte antigen (HLA) complex). The methods, compositions, and systems described herein combine selective amplification of the regions of interest with the ability to maintain molecular context, thereby allowing for de novo sequencing of genomic regions that are generally poorly characterized, as well as optionally providing long range molecular context of these regions in the larger genome.

[0042] In specific instances, methods described herein include a step in which selected regions of the genome are selectively amplified prior to sequencing. This amplification, which is generally conducted using methods known in the art (including without limitation PCR amplification) provides at least 1X, 10X, 20X, 50X, 100X, 200X, 500X, 1000X, 1500X, 2000X, 5000X, or 10000X coverage of the selected regions of the genome, thereby providing a quantity of nucleic acids to allow de novo sequencing of those selected regions. In further embodiments, the amplification provides at least 1X-20X, 50X-100X, 200X-1000X, 1500X-5000X, 5000X-10,000X, 1000X-10000X, 1500X-9000X, 2000X-8000X, 2500X-7000X, 3000X-6500X, 3500X-6000X, 4000X-5500X coverage of the selected regions of the genome.

[0043] The amplification is generally conducted through extension of primers complementary to sequences within or near the selected regions of the genome. In some cases, a library of primers is used that is designed to tile across the regions of interest – in other words, the library of primers is designed to amplify regions at specific distances along the selected regions of the genome. In some instances, the selective amplification utilizes primers that are complementary to every 10, 15, 20, 25, 50, 100, 200, 250, 500, 750, 1000, or 10000 bases along the selected regions of the genome. In still further examples, the tiled library of primers is designed to capture a mixture of distances – that mixture can be a random mixture of distances or intelligently designed such that specific portions or percentages of the selected regions are amplified by different primer pairs.

[0044] In general, the methods and systems described herein accomplish targeted genomic sequencing by providing for the determination of the sequence of selected regions of the genome, and this sequencing information is obtained using methods that have the advantages of the extremely low sequencing error rates and high throughput of short read sequencing technologies.

[0045] Sequencing of nucleic acids is typically carried out in a manner that preserves the molecular context of sequence reads or portions of sequence reads. By that is meant that multiple sequence reads or multiple portions of sequence reads may be attributable to a single originating molecule of a nucleic acid. By ‘attributable to’ is meant that the sequence reads can be identified as falling within the linear sequence of bases of their particular originating molecules of a nucleic acid – in other words, if fragments 1 and 2 are generated from originating nucleic acid molecule A, then the sequencing is carried out in a manner such that sequence reads from fragments 1, 2, 3 and 4 retain their molecular context and it is readily ascertained that fragments 1 and 2 are derived from originating molecule A.

[0046] While this single molecule of a nucleic acid may be of any of a variety of lengths, in preferred aspects, it will be a relatively long molecule, allowing for preservation of long range molecular context. In particular, the single originating molecule is preferably substantially longer than the typical short read sequence length, e.g., longer than 200 bases, and is often at least 1000 bases or longer, 5000 bases or longer, 10,000 bases or longer, 20,000 bases or longer, 30,000 bases or longer, 40,000 bases or longer, 50,000 bases or longer, 60,000 bases or longer, 70,000 bases or longer, 80,000 bases or longer, 90,000 bases or longer, or 100,000 bases or longer, and in some cases up to 1 megabase or longer.

[0047] In general, as shown in Figure 1, the methods and systems described herein may be used to characterize nucleic acids, particularly nucleic acids from selected regions of the genome, while retaining molecular context. As shown, two discrete individual nucleic acids 102 and 104 are illustrated, each having a number of regions of interest, e.g., region 106 and 108 in nucleic acid 102, and regions 110 and 112 in nucleic acid 104. The regions of interest in each nucleic acid are linked within (e.g., originate from) the same nucleic acid molecule, but in some cases these regions may be relatively separated from each other, e.g., more than 1kb apart, more than 5 kb apart, more than 10kb apart, more than 20kb apart, more than 30 kb apart, more than 40kb apart, more than 50 kb apart, and in some cases, as much as 100 kb apart or more. The regions of interest are generally discrete and separate parts of the genome – in some cases, such regions are poorly characterized regions. The regions of interest may also denote individual genes, gene groups, exons. As shown, each nucleic acid 102 and 104 is separated. As illustrated in Figure 1, each nucleic acid is separated into its own partition 114 and 116, respectively; however, as will be appreciated, the methods described herein are not limited to the use of such partitions and any method of separating of nucleic acid molecules can be used and then those separated nucleic acid molecules can be further processed in accordance with any of the methods disclosed herein. As noted elsewhere herein, partitions such as 114 and 116 in Figure 1 are, in many cases, aqueous droplets in a water in oil emulsion. Within each droplet, portions of each fragment are copied in a manner that preserves the original molecular context of those fragments, *e.g.*, as having originated from the same molecule. Such molecular context can be preserved using any method that allows for attribution of the fragment to the original nucleic acid molecule from which it was derived. As shown in Figure 1, one method by which this is achieved is through the inclusion in each copied fragment of a barcode sequence, *e.g.*, barcode sequence “1” or “2” as illustrated, that is representative of the droplet into which the originating fragment was partitioned. For whole genome sequence analysis applications, one could simply pool all of the copied fragments and their associated barcodes, in order to sequence and reassemble the full range sequence information from each of the originating nucleic acids 102 and 104. However, in many cases, it is more desirable to only analyze specific targeted portions of the overall genome, in order to provide greater focus on scientifically relevant portions of the genome, and to minimize the time and expense of performing sequencing on less relevant or irrelevant portions of the genome. Other sequencing methods that assist in preserving molecular context include single molecule sequencing processes, such as SMRT sequencing available from Pacific Biosciences, and nanopore sequencing described by, e.g., Oxford Nanopore, and Truseq SLR processes available from Illumina, Inc.

[0048] In accordance with the above, in addition to the barcoding step, there may be one or more steps of selective amplification, such that if nucleic acids 102 or 104 contain selected genomic regions of interest, amplicons from those regions will form a larger percentage of the fragments in each of the partitions 114 and 116. This amplification step will generally take place prior to or simultaneously with the attachment of the barcodes in accordance with the methods described herein, although in some embodiments the amplification step may also occur subsequent to attachment of the barcodes.

[0049] Because the pooled fragments within library 118 retain their original molecular context, e.g., through the retention of the barcode information, they may be reassembled into their original molecular contexts with embedded (at times, long range) linkage information, e.g., with inferred linkage as between each of the assembled regions of interest 106:108 and 110:112. By way of example, one may identify direct molecular linkage between two disparate targeted portions of the genome, e.g., two or more exons, and that direct molecular linkage may be used to identify structural variations and other genomic characteristics. For situations in which selective amplification is utilized to increase the amount of nucleic acid fragments containing portions of selected regions of the genome, then the ability to identify the molecular context also provides a way to sequence those selected regions of the genome, often at a depth of coverage that allows for de novo assembly of those regions.

[0050] In certain situations, sequencing methods described herein include a combination of deep coverage of the selected regions with lower level linked reads across longer ranges of the genome. As will be appreciated, this combination of de novo and re-sequencing provides an efficient way to sequence an entire genome and/or large portions of a genome. Targeted coverage of poorly characterized and/or highly polymorphic regions through the selective amplification methods described herein provides the amount of nucleic acid material at a coverage level necessary for de novo sequence assembly of those regions, whereas linked genomic sequencing over other regions of the genome allows for high throughput analysis of the remainder of the genome by providing sequence information as to discrete regions which are linked together through preservation of their molecular context. The methods and compositions described herein are uniquely amenable to allowing for a combination of de novo and linked read sequencing, because the same sequencing platform and sequencing library can be used for both types of coverage. The population of nucleic acids and/or nucleic acid fragments that are sequenced in accordance with the methods described herein contain sequences from both the genomic regions for de novo sequencing and the genomic regions for re-sequencing – the proportion of nucleic acids covering the regions of interest for de novo sequencing is higher than the nucleic acids covering the other regions of the genome due to the targeted amplification methods described in further detail herein. Such methods are further amenable for de novo assembly of haplotypes, because the methods described herein allow phase information to be retained during assembly.

[0051] In addition to providing the ability to obtain sequence information from selected regions of the genome, the methods and systems described herein can also provide other characterizations of genomic material, including without limitation haplotype phasing, identification of structural variations, and identifying copy number variations, as described in U.S. Patent Application Nos. 14/752,589 and 14/752,602, which are herein incorporated by reference in their entirety for all purposes and in particular for all written description, figures and working examples directed to characterization of genomic material.

[0052] Methods of processing and sequencing nucleic acids in accordance with the methods and systems described in the present application are also described in further detail in U.S. Patent Application Nos. 14/316,383; 14/316,398; 14/316,416; 14/316,431; 14/316,447; and 14/316,463 which are herein incorporated by reference in their entirety for all purposes and in particular for all written description, figures and working examples directed to processing nucleic acids and sequencing and other characterizations of genomic material.

[0053] Generally, methods of the invention include steps as illustrated in Figure 2, which provides a schematic overview of methods of the invention discussed in further detail herein. As will be appreciated, the method outlined in Figure 2 is an exemplary embodiment that may be altered or modified as needed and as described herein.

[0054] As shown in Figure 2, the methods described herein will in most examples include a step in which sample nucleic acids containing the targeted regions of interest are separated, for example into partitions (201). Generally, each partition containing nucleic acids from genomic regions of interest will undergo a targeted enrichment to produce a population of fragments in which a large proportion will contain sequences from a selected genomic region (202). Those fragments are then further fragmented or copied in such a way as to preserve the original molecular context of the fragments (203), usually by barcoding the fragments that are specific to the partition in which they are contained, although any other methods of attributing the original molecular context of the fragments can be used. Each partition may in some examples include more than one nucleic acid, and will in some instances contain several hundred nucleic acid molecules – in situations in which multiple nucleic acids are within a partition, any particular locus of the genome will generally be represented by a single individual nucleic acid prior to barcoding. The barcoded fragments of step 203 can be generated using any methods known in the art – in some examples, oligonucleotides are the samples within the distinct partitions. Such oligonucleotides may comprise random sequences intended to randomly prime numerous different regions of the samples, or they may comprise a specific primer sequence targeted to prime upstream of a targeted region of the sample. In further examples, these oligonucleotides also contain a barcode sequence, such that the replication process also barcodes the resultant replicated fragment of the original sample nucleic acid. Such barcodes can be added using any method known in the art, including addition of barcode sequences during amplification methods that amplify segments of the individual nucleic acid molecules as well as insertion of barcodes into the original individual nucleic acid molecules using transposons, including methods such as those described in Amini et al., Nature Genetics 46: 1343-1349 (2014) (advance online publication on October 29, 2014). A particularly elegant process for use of these barcode oligonucleotides in amplifying and barcoding samples is described in detail in U.S. Patent Application Nos. USSNs 14/316,383; 14/316,398; 14/316,416; 14/316,431; 14/316,447; and 14/316,463 which are herein incorporated by reference in their entirety for all purposes and in particular for all written description, figures and working examples directed to processing nucleic acids and sequencing and other characterizations of genomic material. Extension reaction reagents, e.g., DNA polymerase, nucleoside triphosphates, co-factors (e.g., Mg^{2+} or Mn^{2+} etc.), that are also contained in the partitions, then extend the primer sequence using the sample as a template, to produce a complementary fragment to the strand of the template to which the primer annealed, and the complementary fragment includes the oligonucleotide and its associated barcode sequence. Annealing and extension of multiple primers to different portions of the sample can result in a large pool of overlapping complementary fragments of the sample, each possessing its own barcode sequence indicative of the partition in which it was created. In some cases, these complementary fragments may themselves be used as a template primed by the oligonucleotides present in the partition to produce a complement of the complement that again, includes the barcode sequence. In further examples, this replication process is configured such that when the first complement is duplicated, it produces two complementary sequences at or near its termini to allow the formation of a hairpin structure or partial hairpin structure, which reduces the ability of the molecule to be the basis for producing further iterative copies.

[0055] Returning to the method exemplified in Figure 2, once the partition-specific barcodes are attached to the copied fragments, the barcoded fragments are then pooled (204). The pooled fragments are then sequenced (205) and the sequences of the fragments are attributed to their originating molecular context (206), such that the targeted regions of interest are both identified and also linked with that originating molecular context. An advantage of the methods and systems described herein is that attaching a partition- or sample-specific barcode to the copied fragments prior to enriching the fragments for targeted genomic regions preserves the original molecular context of those targeted regions, allowing them to be attributed to their original partition and thus their originating sample nucleic acid molecule.

[0056] In addition to the above workflow, targeted genomic regions may be further enriched, isolated or separated, *i.e.,* “pulled down,” for further analysis, particularly sequencing, using methods that include both chip-based and solution-based capture methods. Such methods utilize probes that are complementary to the genomic regions of interest or to regions near or adjacent to the genomic regions of interest. For example, in hybrid (or chip-based) capture, microarrays containing capture probes (usually single-stranded oligonucleotides) with sequences that taken together cover the region of interest are fixed to a surface. Genomic DNA is fragmented and may further undergo processing such as end-repair to produce blunt ends and/or addition of additional features such as universal priming sequences. These fragments are hybridized to the probes on the microarray. Unhybridized fragments are washed away and the desired fragments are eluted or otherwise processed on the surface for sequencing or other analysis, and thus the population of fragments remaining on the surface is enriched for fragments containing the targeted regions of interest (e.g., the regions comprising the sequences complementary to those contained in the capture probes). The enriched population of fragments may further be amplified using any amplification technologies known in the art. Exemplary methods for such targeted pull down enrichment methods are described in US Patent Application No. 14/927,297, filed on October 29, 2015, which is hereby incorporated by reference in its entirety for all purposes and in particular for all teachings related to targeted pull down enrichment methods and sequencing methods, including all written description, figures and examples.

[0057] As noted above, the methods and systems described herein provide individual molecular context for short sequence reads of longer nucleic acids. Such individual molecular context can be provided by any method or composition that allows attribution of the shorter sequence reads to the originating individual nucleic acid. As used herein, individual molecular context refers to sequence context beyond the specific sequence read, e.g., relation to adjacent or proximal sequences, that are not included within the sequence read itself, and as such, will typically be such that they would not be included in whole or in part in a short sequence read, e.g., a read of about 150 bases, or about 300 bases for paired reads. In particularly preferred aspects, the methods and systems provide long range sequence context for short sequence reads. Such long range context includes relationship or linkage of a given sequence read to sequence reads that are within a distance of each other of longer than 1 kb, longer than 5 kb, longer than 10 kb, longer than 15 kb, longer than 20 kb, longer than 30 kb, longer than 40 kb, longer than 50 kb, longer than 60 kb, longer than 70 kb, longer than 80 kb, longer than 90 kb or even longer than 100 kb, or longer. As will be appreciated, by providing long range individual molecular context, one can also derive the phasing information of variants within that individual molecular context, e.g., variants on a particular long molecule will be, by definition commonly phased.

[0058] By providing longer range individual molecular context, the methods and systems of the invention also provide much longer inferred molecular context (also referred to herein as a “long virtual single molecule read”). Sequence context, as described herein, can include mapping or providing linkage of fragments across different (generally on the kilobase scale) ranges of full genomic sequence. These methods include mapping the short sequence reads to the individual longer molecules or contigs of linked molecules, as well as long range sequencing of large portions of the longer individual molecules, e.g., having contiguous determined sequences of individual molecules where such determined sequences are longer than 1 kb, longer than 5 kb, longer than 10 kb, longer than 15 kb, longer than 20 kb, longer than 30 kb, longer than 40 kb, longer than 50 kb, longer than 60 kb, longer than 70 kb, longer than 80 kb, longer than 90 kb or even longer than 100 kb. As with sequence context, the attribution of short sequences to longer nucleic acids, e.g., both individual long nucleic acid molecules or collections of linked nucleic acid molecules or contigs, may include both mapping of short sequences against longer nucleic acid stretches to provide high level sequence context, as well as providing assembled sequences from the short sequences through these longer nucleic acids.

[0059] Furthermore, while one may utilize the long range sequence context associated with long individual molecules, having such long range sequence context also allows one to infer even longer range sequence context. By way of one example, by providing the long range molecular context described above, one can identify overlapping variant portions, e.g., phased variants, translocated sequences, etc., among long sequences from different originating molecules, allowing the inferred linkage between those molecules. Such inferred linkages or molecular contexts are referred to herein as “inferred contigs”. In some cases when discussed in the context of phased sequences, the inferred contigs may represent commonly phased sequences, e.g., where by virtue of overlapping phased variants, one can infer a phased contig of substantially greater length than the individual originating molecules. These phased contigs are referred to herein as “phase blocks”.

[0060] By starting with longer single molecule reads (e.g., the “long virtual single molecule reads” discussed above), one can derive longer inferred contigs or phase blocks than would otherwise be attainable using short read sequencing technologies or other approaches to phased sequencing. See, e.g., published U.S. Patent Application No. 2013-0157870. In particular, using the methods and systems described herein, one can obtain inferred contig or phase block lengths having an N50 (where the sum of the block lengths that are greater than the stated N50 number is 50% of the sum of all block lengths) of at least about 10kb, at least about 20kb, at least about 50kb. In more preferred aspects, inferred contig or phase block lengths having an N50 of at least about 100kb, at least about 150kb, at least about 200kb, and in many cases, at least about 250kb, at least about 300 kb, at least about 350 kb, at least about 400 kb, and in some cases, at least about 500 kb or more, are attained. In still other cases, maximum phase block lengths in excess of 200 kb, in excess of 300 kb, in excess of 400 kb, in excess of 500 kb, in excess of 1 Mb, or even in excess of 2 Mb may be obtained.

[0061] In one aspect, and in conjunction with any of the capture methods described above and later herein, the methods and systems described herein provide for the separation of sample nucleic acids for further processing in accordance with any of the methods described herein. Such separation can be of any form that allows the nucleic acids to undergo further processing and reactions in relative isolation from other nucleic acids from which they are separated. The separating can be in terms of single nucleic acids each separated from all other nucleic acids, or into groups of two or more nucleic acids, which are then separated from other groups of nucleic acids. In some exemplary embodiments, such separating includes compartmentalization, depositing or partitioning of sample nucleic acids, or fragments thereof, into discrete compartments or partitions (referred to interchangeably herein as partitions), where each partition maintains separation of its own contents from the contents of other partitions. Unique identifiers or other means of attribution (in some examples, barcodes), may be previously, subsequently or concurrently delivered to the separated nucleic acids in order to allow for the later attribution of the characteristics, e.g., nucleic acid sequence information, to the sample nucleic acids from which that information is derived. In certain exemplary embodiments in which the nucleic acids are separated into compartments or partitions, the identifier can be included within or introduced to a particular compartment, and particularly to relatively long stretches of contiguous sample nucleic acids that may be originally deposited into the partitions.

[0062] The sample nucleic acids utilized in the methods described herein typically represent a number of overlapping portions of the overall sample to be analyzed, e.g., an entire chromosome, exome, or other large genomic portion. These sample nucleic acids may include whole genomes, individual chromosomes, exomes, amplicons, or any of a variety of different nucleic acids of interest. The sample nucleic acids are typically partitioned such that the nucleic acids are present in the partitions in relatively long fragments or stretches of contiguous nucleic acid molecules. Typically, these fragments of the sample nucleic acids may be longer than 1 kb, longer than 5 kb, longer than 10 kb, longer than 15 kb, longer than 20 kb, longer than 30 kb, longer than 40 kb, longer than 50 kb, longer than 60 kb, longer than 70 kb, longer than 80 kb, longer than 90 kb or even longer than 100 kb, which permits the longer range molecular context described above.

[0063] The sample nucleic acids are also typically partitioned at a level whereby a given partition has a very low probability of including two overlapping fragments of a genomic locus. This is typically accomplished by providing the sample nucleic acid at a low input amount and/or concentration during the partitioning process. As a result, in preferred cases, a given partition may include a number of long, but non-overlapping fragments of the starting sample nucleic acids. The sample nucleic acids in the different partitions are then associated with unique identifiers, where for any given partition, nucleic acids contained therein possess the same unique identifier, but where different partitions may include different unique identifiers. Moreover, because the partitioning step allocates the sample components into very small volume partitions or droplets, it will be appreciated that in order to achieve the desired allocation as set forth above, one need not conduct substantial dilution of the sample, as would be required in higher volume processes, e.g., in tubes, or wells of a multiwell plate. Further, because the systems described herein employ such high levels of barcode diversity, one can allocate diverse barcodes among higher numbers of genomic equivalents, as provided above. In particular, previously described, multiwell plate approaches (see, e.g., U.S. Published Application No. 2013-0079231 and 2013-0157870) typically only operate with a hundred to a few hundred different barcode sequences, and employ a limiting dilution process of their sample in order to be able to attribute barcodes to different cells/nucleic acids. As such, they will generally operate with far fewer than 100 cells, which would typically provide a ratio of genomes:(barcode type) on the order of 1:10, and certainly well above 1:100. The systems described herein, on the other hand, because of the high level of barcode diversity, e.g., in excess of 10,000, 100,000, 500,000, etc. diverse barcode types, can operate at genome:(barcode type) ratios that are on the order of 1:50 or less, 1:100 or less, 1:1000 or less, or even smaller ratios, while also allowing for loading higher numbers of genomes (e.g., on the order of greater than 100 genomes per assay, greater than 500 genomes per assay, 1000 genomes per assay, or even more) while still providing for far improved barcode diversity per genome.

[0064] Often, the sample is combined with a set of oligonucleotide tags that are releasably-attached to beads prior to the partitioning step. Methods for barcoding nucleic acids are known in the art and described herein. In some examples, methods are utilized as described in Amini et al, 2014, *Nature* Genetics, Advance Online Publication), which is herein incorporated by reference in its entirety for all purposes and in particular for all teachings related to attaching barcodes or other oligonucleotide tags to nucleic acids. In further examples, the oligonucleotides may comprise at least a first and second region. The first region may be a barcode region that, as between oligonucleotides within a given partition, may be substantially the same barcode sequence, but as between different partitions, may and, in most cases is a different barcode sequence. The second region may be an N-mer (either a random N-mer or an N-mer designed to target a particular sequence) that can be used to prime the nucleic acids within the sample within the partitions. In some cases, where the N-mer is designed to target a particular sequence, it may be designed to target a particular chromosome (e.g., chromosome 1, 13, 18, or 21), or region of a chromosome, e.g., an exome or other targeted region. In some cases, the N-mer may be designed to target a particular gene or genetic region, such as a gene or region associated with a disease or disorder (e.g., cancer). Within the partitions, an amplification reaction may be conducted using the second N-mer to prime the nucleic acid sample at different places along the length of the nucleic acid. As a result of the amplification, each partition may contain amplified products of the nucleic acid that are attached to an identical or near-identical barcode, and that may represent overlapping, smaller fragments of the nucleic acids in each partition. The bar-code can serve as a marker that signifies that a set of nucleic acids originated from the same partition, and thus potentially also originated from the same strand of nucleic acid. Following amplification, the nucleic acids may be pooled, sequenced, and aligned using a sequencing algorithm. Because shorter sequence reads may, by virtue of their associated barcode sequences, be aligned and attributed to a single, long fragment of the sample nucleic acid, all of the identified variants on that sequence can be attributed to a single originating fragment and single originating chromosome. Further, by aligning multiple co-located variants across multiple long fragments, one can further characterize that chromosomal contribution. Accordingly, conclusions regarding the phasing of particular genetic variants may then be drawn, as can analyses across long ranges of genomic sequence – for example, identification of sequence information across stretches of poorly characterized regions of the genome. Such information may also be useful for identifying haplotypes, which are generally a specified set of genetic variants that reside on the same nucleic acid strand or on different nucleic acid strands. Copy number variations may also be identified in this manner.

[0065] The described methods and systems provide significant advantages over current nucleic acid sequencing technologies and their associated sample preparation methods. Ensemble sample preparation and sequencing methods are predisposed towards primarily identifying and characterizing the majority constituents in the sample, and are not designed to identify and characterize minority constituents, e.g., genetic material contributed by one chromosome, from a poorly characterized or highly polymorphic region of the genome, or material from one or a few cells, or fragmented tumor cell DNA molecule circulating in the bloodstream, that constitute a small percentage of the total DNA in the extracted sample. The methods described herein include selective amplification methods that increase the genetic material from these minority constituents, and the ability to retain the molecular context of this genetic material further provides genetic characterization of these constituents. The described methods and systems also provide a significant advantage for detecting populations that are present within a larger sample. As such, they are particularly useful for assessing haplotype and copy number variations – the methods disclosed herein are also useful for providing sequence information over regions of the genome that are poorly characterized or are poorly represented in a population of nucleic acid targets due to biases introduced during sample preparation.

[0066] The use of the barcoding technique disclosed herein confers the unique capability of providing individual molecular context for a given set of genetic markers, i.e., attributing a given set of genetic markers (as opposed to a single marker) to individual sample nucleic acid molecules, and through variant coordinated assembly, to provide a broader or even longer range inferred individual molecular context, among multiple sample nucleic acid molecules, and/or to a specific chromosome. These genetic markers may include specific genetic loci, e.g., variants, such as SNPs, or they may include short sequences. Furthermore, the use of barcoding confers the additional advantages of facilitating the ability to discriminate between minority constituents and majority constituents of the total nucleic acid population extracted from the sample, e.g. for detection and characterization of circulating tumor DNA in the bloodstream, and also reduces or eliminates amplification bias during optional amplification steps. In addition, implementation in a microfluidics format confers the ability to work with extremely small sample volumes and low input quantities of DNA, as well as the ability to rapidly process large numbers of sample partitions (droplets) to facilitate genome-wide tagging.

[0067] As described previously, an advantage of the methods and systems described herein is that they can achieve the desired results through the use of ubiquitously available, short read sequencing technologies. Such technologies have the advantages of being readily available and widely dispersed within the research community, with protocols and reagent systems that are well characterized and highly effective. These short read sequencing technologies include those available from, e.g., Illumina, inc. (GAIIx, NextSeq, MiSeq, HiSeq, X10), Ion Torrent division of Thermo-Fisher (Ion Proton and Ion PGM), pyrosequencing methods, as well as others.

[0068] Of particular advantage is that the methods and systems described herein utilize these short read sequencing technologies and do so with their associated low error rates and high throughputs. In particular, the methods and systems described herein achieve the desired individual molecular readlengths or context, as described above, but with individual sequencing reads, excluding mate pair extensions, that are shorter than 1000 bp, shorter than 500 bp, shorter than 300 bp, shorter than 200 bp, shorter than 150 bp or even shorter; and with sequencing error rates for such individual molecular readlengths that are less than 5%, less than 1%, less than 0.5%, less than 0.1%, less than 0.05%, less than 0.01%, less than 0.005%, or even less than 0.001%.

II. Work flow overview

[0069] The methods and systems described in the disclosure provide for separating nucleic acids into different groups or into different regions such that the separated nucleic acids can undergo further processing and/or reactions in relative isolation from one or more other nucleic acids. Such separating can in certain exemplary instances include depositing or partitioning individual samples (e.g., nucleic acids) into discrete partitions, where each partition maintains separation of its own contents from the contents in other partitions. As used herein, the partitions refer to containers or vessels that may include a variety of different forms, e.g., wells, tubes, micro or nanowells, through holes, or the like. In preferred aspects, however, the partitions are flowable within fluid streams. These vessels may be comprised of, e.g., microcapsules or micro-vesicles that have an outer barrier surrounding an inner fluid center or core, or they may be a porous matrix that is capable of entraining and/or retaining materials within its matrix. In preferred aspect, however, these partitions may comprise droplets of aqueous fluid within a non-aqueous continuous phase, e.g., an oil phase. A variety of different vessels are described in, for example, U.S. Patent Application No. 13/966,150, filed August 13, 2013. Likewise, emulsion systems for creating stable droplets in non-aqueous or oil continuous phases are described in detail in, e.g., Published U.S. Patent Application No. 2010-0105112. In certain cases, microfluidic channel networks are particularly suited for generating partitions as described herein. Examples of such microfluidic devices include those described in detail in U.S. Patent Application No. 14/682,952, filed April 9, 2015, the full disclosure of which is incorporated herein by reference in its entirety for all purposes. Alternative mechanisms may also be employed in the partitioning of individual cells, including porous membranes through which aqueous mixtures of cells are extruded into non-aqueous fluids. Such systems are generally available from, e.g., Nanomi, Inc.

[0070] In methods utilizing droplets in an emulsion, partitioning of sample materials, e.g., nucleic acids, into discrete partitions may generally be accomplished by flowing an aqueous, sample containing stream, into a junction into which is also flowing a non-aqueous stream of partitioning fluid, e.g., a fluorinated oil, such that aqueous droplets are created within the flowing stream partitioning fluid, where such droplets include the sample materials. As described below, the partitions, e.g., droplets, also typically include co-partitioned barcode oligonucleotides. The relative amount of sample materials within any particular partition may be adjusted by controlling a variety of different parameters of the system, including, for example, the concentration of sample in the aqueous stream, the flow rate of the aqueous stream and/or the non-aqueous stream, and the like. The partitions described herein are often characterized by having extremely small volumes. For example, in the case of droplet based partitions, the droplets may have overall volumes that are less than 1000 pL, less than 900 pL, less than 800 pL, less than 700 pL, less than 600 pL, less than 500 pL, less than 400pL, less than 300 pL, less than 200 pL, less than 100pL, less than 50 pL, less than 20 pL, less than 10 pL, or even less than 1 pL. Where co-partitioned with beads, it will be appreciated that the sample fluid volume within the partitions may be less than 90% of the above described volumes, less than 80%, less than 70%, less than 60%, less than 50%, less than 40%, less than 30%, less than 20%, or even less than 10% the above described volumes. In some cases, the use of low reaction volume partitions is particularly advantageous in performing reactions with very small amounts of starting reagents, e.g., input nucleic acids. Methods and systems for analyzing samples with low input nucleic acids are presented in U.S. Patent Application Nos. 14/752,589 and 14/752,602, the full disclosure of which are hereby incorporated by reference in their entirety.

[0071] Once the samples are introduced into their respective partitions, in accordance with the methods and systems described herein, the sample nucleic acids within partitions are generally subjected to selective amplification, such that regions of the genome that are of interest for targeted coverage to allow for de novo sequencing are present in higher proportion in comparison to other regions of the genome (although, as will be appreciated, those other regions of the genome may also be amplified, but to a lesser extent, as they are not of interest for de novo coverage). In certain embodiments, the genomic regions of interest are amplified to provide at least 1X, 2X, 5X, 10X, 20X, 30X, 40X or 50X coverage of those selected regions of the genome. In further embodiments, all of the nucleic acids within a partition are amplified, but selected genomic regions are amplified in a targeted way such that at least 1-5, 2-10, 3-15, 4-20, 5-25, 6-30, 7-35, 8-40, 9-45, or 10-50 times more amplicons are produced from those selected genomic regions than from other parts of the genome.

[0072] Simultaneously with or subsequent to the selective amplification of selected regions of the genome, the nucleic acids (or fragments thereof) within the partitions are provided with unique identifiers such that, upon characterization of those nucleic acids they may be attributed as having been derived from their respective origins. Accordingly, the sample nucleic acids are typically co-partitioned with the unique identifiers. In some exemplary embodiments, such unique identifiers are barcode sequences. For the sake of clarity, much of the discussion herein is directed to identifiers comprising barcode sequences, but, as will be appreciated, any unique identifiers that can be used to retain molecular context for sequence reads can be used in accordance with the methods described herein. In some preferred aspects, the unique identifiers are provided in the form of oligonucleotides that comprise nucleic acid barcode sequences that may be attached to the nucleic acid samples. The oligonucleotides are partitioned such that as between oligonucleotides in a given partition, the nucleic acid barcode sequences contained therein are the same, but as between different partitions, the oligonucleotides can, and preferably have differing barcode sequences. In preferred aspects, only one nucleic acid barcode sequence will be associated with a given partition, although in some cases, two or more different barcode sequences may be present.

[0073] The nucleic acid barcode sequences will typically include from 6 to about 20 or more nucleotides within the sequence of the oligonucleotides. These nucleotides may be completely contiguous, i.e., in a single stretch of adjacent nucleotides, or they may be separated into two or more separate subsequences that are separated by one or more nucleotides. Typically, separated subsequences may typically be from about 4 to about 16 nucleotides in length.

[0074] The co-partitioned oligonucleotides also typically comprise other functional sequences useful in the processing of the partitioned nucleic acids. These sequences include, e.g., targeted or random/universal amplification primer sequences for amplifying the genomic DNA from the individual nucleic acids within the partitions while attaching the associated barcode sequences, sequencing primers, hybridization or probing sequences, e.g., for identification of presence of the sequences, or for pulling down barcoded nucleic acids, or any of a number of other potential functional sequences. Again, co-partitioning of oligonucleotides and associated barcodes and other functional sequences, along with sample materials is described in, for example, U.S. Patent Application Nos. 14/316,383; 14/316,398; 14/316,416; 14/316,431; 14/316,447; and 14/316,463, the full disclosures of which is hereby incorporated by reference in their entireties.

[0075] Briefly, in one exemplary process, beads are provided that each may include large numbers of the above described oligonucleotides releasably attached to the beads, where all of the oligonucleotides attached to a particular bead may include the same nucleic acid barcode sequence, but where a large number of diverse barcode sequences may be represented across the population of beads used. Typically, the population of beads may provide a diverse barcode sequence library that may include at least 1000 different barcode sequences, at least 10,000 different barcode sequences, at least 100,000 different barcode sequences, or in some cases, at least 1,000,000 different barcode sequences. Additionally, each bead may typically be provided with large numbers of oligonucleotide molecules attached. In particular, the number of molecules of oligonucleotides including the barcode sequence on an individual bead may be at least bout 10,000 oligonucleotides, at least 100,000 oligonucleotide molecules, at least 1,000,000 oligonucleotide molecules, at least 100,000,000 oligonucleotide molecules, and in some cases at least 1 billion oligonucleotide molecules.

[0076] The oligonucleotides may be releasable from the beads upon the application of a particular stimulus to the beads. In some cases, the stimulus may be a photo-stimulus, e.g., through cleavage of a photo-labile linkage that may release the oligonucleotides. In some cases, a thermal stimulus may be used, where elevation of the temperature of the beads environment may result in cleavage of a linkage or other release of the oligonucleotides form the beads. In some cases, a chemical stimulus may be used that cleaves a linkage of the oligonucleotides to the beads, or otherwise may result in release of the oligonucleotides from the beads.

[0077] In accordance with the methods and systems described herein, the beads including the attached oligonucleotides may be co-partitioned with the individual samples, such that a single bead and a single sample are contained within an individual partition. In some cases, where single bead partitions are desired, it may be desirable to control the relative flow rates of the fluids such that, on average, the partitions contain less than one bead per partition, in order to ensure that those partitions that are occupied, are primarily singly occupied. Likewise, one may wish to control the flow rate to provide that a higher percentage of partitions are occupied, e.g., allowing for only a small percentage of unoccupied partitions. In preferred aspects, the flows and channel architectures are controlled as to ensure a desired number of singly occupied partitions, less than a certain level of unoccupied partitions and less than a certain level of multiply occupied partitions.

[0078] Figure 3 illustrates one particular example method for barcoding and subsequently sequencing a sample nucleic acid, particularly for use for a copy number variation or haplotype assay. First, a sample comprising nucleic acid may be obtained from a source, 300, and a set of barcoded beads may also be obtained, 310. The beads are preferably linked to oligonucleotides containing one or more barcode sequences, as well as a primer, such as a random N-mer or other primer. Preferably, the barcode sequences are releasable from the barcoded beads, e.g., through cleavage of a linkage between the barcode and the bead or through degradation of the underlying bead to release the barcode, or a combination of the two. For example, in certain preferred aspects, the barcoded beads can be degraded or dissolved by an agent, such as a reducing agent to release the barcode sequences. In this example, a low quantity of the sample comprising nucleic acid, 305, barcoded beads, 315, and optionally other reagents, e.g., a reducing agent, 320, are combined and subject to partitioning. By way of example, such partitioning may involve introducing the components to a droplet generation system, such as a microfluidic device, 325. With the aid of the microfluidic device 325, a water-in-oil emulsion 330 may be formed, wherein the emulsion contains aqueous droplets that contain sample nucleic acid, 305, reducing agent, 320, and barcoded beads, 315. The reducing agent may dissolve or degrade the barcoded beads, thereby releasing the oligonucleotides with the barcodes and random N-mers from the beads within the droplets, 335. The random N-mers may then prime different regions of the sample nucleic acid, resulting in amplified copies of the sample after amplification, wherein each copy is tagged with a barcode sequence, 340. Preferably, each droplet contains a set of oligonucleotides that contain identical barcode sequences and different random N-mer sequences. Subsequently, the emulsion is broken, 345 and additional sequences (e.g., sequences that aid in particular sequencing methods, additional barcodes, etc.) may be added, via, for example, amplification methods, 350 (e.g., PCR). Sequencing may then be performed, 355, and an algorithm applied to interpret the sequencing data, 360. Sequencing algorithms are generally capable, for example, of performing analysis of barcodes to align sequencing reads and/or identify the sample from which a particular sequence read belongs. In addition, and as is described herein, these algorithms may also further be used to attribute the sequences of the copies to their originating molecular context.

[0079] As will be appreciated, prior to or simultaneously with tagging with the barcode sequence 340, the samples can be amplified in accordance with any of the methods described herein to provide targeted coverage of selected regions of the genome. This targeted coverage generally results in a larger population of amplicons representing sequences of the nucleic acids (or portions of thereof) in a partition containing those selected regions of the genome as compared to amplicons from other regions of the genome. As a result, there will be a larger number of the amplified copies containing barcode sequence 340 within a partition from the selected regions of the genome than from other regions of the genome.

[0080] In some embodiments and in accordance with any of the above, different amplification protocols are used to favor amplification of fragments containing portions of selected regions of the genome than the protocols used to attach barcode sequences to the fragments. In one non-limiting example, the selective amplification using targeted PCR primers are conducted under standard PCR amplification thermal cycling conditions, whereas the amplification for attachment of the barcodes is conducted with a sharp drop in temperature followed by a slow ramp of increasing temperature to allow for the priming and extension of the random N-mers.

[0081] As noted above, while single occupancy may be the most desired state, it will be appreciated that multiply occupied partitions or unoccupied partitions may often be present. An example of a microfluidic channel structure for co-partitioning samples and beads comprising barcode oligonucleotides is schematically illustrated in Figure 4. As shown, channel segments 402, 404, 406, 408 and 410 are provided in fluid communication at channel junction 412. An aqueous stream comprising the individual samples 414 is flowed through channel segment 402 toward channel junction 412. As described elsewhere herein, these samples may be suspended within an aqueous fluid prior to the partitioning process.

[0082] Concurrently, an aqueous stream comprising the barcode carrying beads 416 is flowed through channel segment 404 toward channel junction 412. A non-aqueous partitioning fluid is introduced into channel junction 412 from each of side channels 406 and 408, and the combined streams are flowed into outlet channel 410. Within channel junction 412, the two combined aqueous streams from channel segments 402 and 404 are combined, and partitioned into droplets 418, that include co-partitioned samples 414 and beads 416. As noted previously, by controlling the flow characteristics of each of the fluids combining at channel junction 412, as well as controlling the geometry of the channel junction, one can optimize the combination and partitioning to achieve a desired occupancy level of beads, samples or both, within the partitions 418 that are generated.

[0083] As will be appreciated, a number of other reagents may be co-partitioned along with the samples and beads, including, for example, chemical stimuli, nucleic acid extension, transcription, and/or amplification reagents such as polymerases, reverse transcriptases, nucleoside triphosphates or NTP analogues, primer sequences and additional cofactors such as divalent metal ions used in such reactions, ligation reaction reagents, such as ligase enzymes and ligation sequences, dyes, labels, or other tagging reagents. The primer sequences may include random primer sequences or targeted PCR primers directed to amplifying selected regions of the genome or a combination thereof.

[0084] Once co-partitioned, the oligonucleotides disposed upon the bead may be used to barcode and amplify the partitioned samples. A particularly elegant process for use of these barcode oligonucleotides in amplifying and barcoding samples is described in detail in U.S. Patent Application Nos. 14/316,383; 14/316,398; 14/316,416; 14/316,431; 14/316,447; and 14/316,463, the full disclosures of which are hereby incorporated by reference in their entireties. Briefly, in one aspect, the oligonucleotides present on the beads that are co-partitioned with the samples and released from their beads into the partition with the samples. The oligonucleotides typically include, along with the barcode sequence, a primer sequence at its 5’ end. The primer sequence may be random or structured. Random primer sequences are generally intended to randomly prime numerous different regions of the samples. Structured primer sequences can include a range of different structures including defined sequences targeted to prime upstream of a specific targeted region of the sample as well as primers that have some sort of partially defined structure, including without limitation primers containing a percentage of specific bases (such as a percentage of GC N-mers), primers containing partially or wholly degenerate sequences, and/or primers containing sequences that are partially random and partially structured in accordance with any of the description herein. As will be appreciated, any one or more of the above types of random and structured primers may be included in oligonucleotides in any combination.

[0085] Once released, the primer portion of the oligonucleotide can anneal to a complementary region of the sample. Extension reaction reagents, e.g., DNA polymerase, nucleoside triphosphates, co-factors (e.g., Mg2+ or Mn2+ etc.), that are also co-partitioned with the samples and beads, then extend the primer sequence using the sample as a template, to produce a complementary fragment to the strand of the template to which the primer annealed, with complementary fragment includes the oligonucleotide and its associated barcode sequence.

Annealing and extension of multiple primers to different portions of the sample may result in a large pool of overlapping complementary fragments of the sample, each possessing its own barcode sequence indicative of the partition in which it was created. In some cases, these complementary fragments may themselves be used as a template primed by the oligonucleotides present in the partition to produce a complement of the complement that again, includes the barcode sequence. In some cases, this replication process is configured such that when the first complement is duplicated, it produces two complementary sequences at or near its termini, to allow the formation of a hairpin structure or partial hairpin structure, which reduces the ability of the molecule to be the basis for producing further iterative copies. A schematic illustration of one example of this is shown in Figure 5.

[0086] As the figure shows, oligonucleotides that include a barcode sequence are co-partitioned in, e.g., a droplet 502 in an emulsion, along with a sample nucleic acid 504. As noted elsewhere herein, the oligonucleotides 508 may be provided on a bead 506 that is co-partitioned with the sample nucleic acid 504, which oligonucleotides are preferably releasable from the bead 506, as shown in panel A. The oligonucleotides 508 include a barcode sequence 512, in addition to one or more functional sequences, e.g., sequences 510, 514 and 516. For example, oligonucleotide 508 is shown as comprising barcode sequence 512, as well as sequence 510 that may function as an attachment or immobilization sequence for a given sequencing system, e.g., a P5 sequence used for attachment in flow cells of an Illumina Hiseq or Miseq system. As shown, the oligonucleotides also include a primer sequence 516, which may include a random or targeted N-mer for priming replication of portions of the sample nucleic acid 504. Also included within oligonucleotide 508 is a sequence 514 which may provide a sequencing priming region, such as a “read1” or R1 priming region, that is used to prime polymerase mediated, template directed sequencing by synthesis reactions in sequencing systems. In many cases, the barcode sequence 512, immobilization sequence 510 and R1 sequence 514 may be common to all of the oligonucleotides attached to a given bead. The primer sequence 516 may vary for random N-mer primers, or may be common to the oligonucleotides on a given bead for certain targeted applications.

[0087] Based upon the presence of primer sequence 516, the oligonucleotides are able to prime the sample nucleic acid as shown in panel B, which allows for extension of the oligonucleotides 508 and 508a using polymerase enzymes and other extension reagents also co-portioned with the bead 506 and sample nucleic acid 504. As shown in panel C, following extension of the oligonucleotides that, for random N-mer primers, would anneal to multiple different regions of the sample nucleic acid 504; multiple overlapping complements or fragments of the nucleic acid are created, e.g., fragments 518 and 520. Although including sequence portions that are complementary to portions of sample nucleic acid, e.g., sequences 522 and 524, these constructs are generally referred to herein as comprising fragments of the sample nucleic acid 504, having the attached barcode sequences. As will be appreciated, the replicated portions of the template sequences as described above are often referred to herein as “fragments” of that template sequence. Notwithstanding the foregoing, however, the term “fragment” encompasses any representation of a portion of the originating nucleic acid sequence, e.g., a template or sample nucleic acid, including those created by other mechanisms of providing portions of the template sequence, such as actual fragmentation of a given molecule of sequence, e.g., through enzymatic, chemical or mechanical fragmentation. In preferred aspects, however, fragments of a template or sample nucleic acid sequence will denote replicated portions of the underlying sequence or complements thereof.

[0088] The barcoded nucleic acid fragments may then be subjected to characterization, e.g., through sequence analysis, or they may be further amplified in the process, as shown in panel D. For example, additional oligonucleotides, e.g., oligonucleotide 508b, also released from bead 506, may prime the fragments 518 and 520. In particular, again, based upon the presence of the random N-mer primer 516b in oligonucleotide 508b (which in many cases will be different from other random N-mers in a given partition, e.g., primer sequence 516), the oligonucleotide anneals with the fragment 518, and is extended to create a complement 526 to at least a portion of fragment 518 which includes sequence 528, that comprises a duplicate of a portion of the sample nucleic acid sequence. Extension of the oligonucleotide 508b continues until it has replicated through the oligonucleotide portion 508 of fragment 518. As noted elsewhere herein, and as illustrated in panel D, the oligonucleotides may be configured to prompt a stop in the replication by the polymerase at a desired point, e.g., after replicating through sequences 516 and 514 of oligonucleotide 508 that is included within fragment 518. As described herein, this may be accomplished by different methods, including, for example, the incorporation of different nucleotides and/or nucleotide analogues that are not capable of being processed by the polymerase enzyme used. For example, this may include the inclusion of uracil containing nucleotides within the sequence region 512 to prevent a non-uracil tolerant polymerase to cease replication of that region. As a result a fragment 526 is created that includes the full-length oligonucleotide 508b at one end, including the barcode sequence 512, the attachment sequence 510, the R1 primer region 514, and the random N-mer sequence 516b. At the other end of the sequence will be included the complement 516’ to the random N-mer of the first oligonucleotide 508, as well as a complement to all or a portion of the R1 sequence, shown as sequence 514’. The R1 sequence 514 and its complement 514’ are then able to hybridize together to form a partial hairpin structure 528. As will be appreciated because the random N-mers differ among different oligonucleotides, these sequences and their complements would not be expected to participate in hairpin formation, e.g., sequence 516', which is the complement to random N-mer 516, would not be expected to be complementary to random N-mer sequence 516b. This would not be the case for other applications, e.g., targeted primers, where the N-mers would be common among oligonucleotides within a given partition. By forming these partial hairpin structures, it allows for the removal of first level duplicates of the sample sequence from further replication, e.g., preventing iterative copying of copies. The partial hairpin structure also provides a useful structure for subsequent processing of the created fragments, e.g., fragment 526.

[0089] All of the fragments from multiple different partitions may then be pooled for sequencing on high throughput sequencers as described herein. Because each fragment is coded as to its partition of origin, the sequence of that fragment may be attributed back to its origin based upon the presence of the barcode. This is schematically illustrated in Figure 6A. As shown in one example, a nucleic acid 604 originated from a first source 600 (e.g., individual chromosome, strand of nucleic acid, etc.) and a nucleic acid 606 derived from a different chromosome 602 or strand of nucleic acid are each partitioned along with their own sets of barcode oligonucleotides as described above.

[0090] Within each partition, each nucleic acid 604 and 606 is then processed to separately provide overlapping set of second fragments of the first fragment(s), e.g., second fragment sets 608 and 610. This processing also provides the second fragments with a barcode sequence that is the same for each of the second fragments derived from a particular first fragment. As shown, the barcode sequence for second fragment set 608 is denoted by “1” while the barcode sequence for fragment set 610 is denoted by “2”. A diverse library of barcodes may be used to differentially barcode large numbers of different fragment sets. However, it is not necessary for every second fragment set from a different first fragment to be barcoded with different barcode sequences. In fact, in many cases, multiple different first fragments may be processed concurrently to include the same barcode sequence. Diverse barcode libraries are described in detail elsewhere herein.

[0091] The barcoded fragments, e.g., from fragment sets 608 and 610, may then be pooled for sequencing using, for example, sequence by synthesis technologies available from Illumina or Ion Torrent division of Thermo Fisher, Inc. Once sequenced, the sequence reads from the pooled fragments 612 can be attributed to their respective fragment set, e.g., as shown in aggregated reads 614 and 616, at least in part based upon the included barcodes, and optionally, and preferably, in part based upon the sequence of the fragment itself. The attributed sequence reads for each fragment set are then assembled to provide the assembled sequence for each sample fragment, e.g., sequences 618 and 620, which in turn, may be further attributed back to their respective original chromosomes or source nucleic acid molecules (600 and 602). Methods and systems for assembling genomic sequences are described in, for example, U.S. Patent Application

No. 14/752,773, filed June 26, 2015, the full disclosure of which is hereby incorporated by reference in its entirety.

[0092] In some embodiments and as illustrated in Figure 6B, included with the partitions containing fragment sets 608 or 610 are primer sets 613. The primer sets 613 are in further embodiments directed to selected regions of the genome, such that prior to, simultaneously with or subsequent to providing the barcode sequences (barcode “1” for 608 and “2” for 610), the fragment sets 608 and 610 are amplified such that the selected regions of the genome are covered to an additional extent over other regions of the genome. In the exemplary embodiment pictured in Figure 6B, fragment set 608 contain sequences from the selected regions of the genome to which primer sets 613 are directed, but fragment set 610 does not contain sequences from those selected regions of the genome. As such, there will be increased coverage (e.g., more copies) of fragments from set 608 than from set 610. Thus, the pooled fragments 612 contains barcoded fragments contain fragments that have been amplified in a targeted way, allowing for a larger proportion of sequence reads from fragment set 608 (the “1” barcoded fragments) than from fragment set 610 (the “2” barcoded fragments). In addition, due to the barcodes, that larger proportion of sequence reads from set 608 can, like the remainder of the fragments in pooled set 612, be attributed back to their respective original source nucleic acid molecules 600 and 602 (shown in Figure 6A).

III. Application of methods and systems to nucleic acid sequencing

[0093] The methods, compositions, and systems described herein are particularly amenable for use in nucleic acid sequencing technologies. Such sequencing technologies can include any technologies known in the art, including short-read and long-read sequencing technologies. In certain aspects, the methods, compositions and systems described herein are used in short read, high accuracy sequencing technologies.

[0094] The methods, compositions, and systems described herein allow for genetic characterization of regions of the genome that are poorly characterized, are highly polymorphic, and/or diverge from reference sequences. In particular, the methods, compositions and systems described herein provide increased and redundant coverage of selected portions of the genome such that additional redundant sequence information can be obtained from those selected portion of the genome. In specific instances, that additional sequence information (e.g., increased coverage of targeted regions of the genome) provides enough information to allow for de novo sequencing of those selected portions of the genome. This de novo sequencing is of particular use for regions of the genome that are poorly characterized, are highly polymorphic, and/or diverge from reference sequences. As will be appreciated, a significant percentage (at least 5-10%

according to, for example Altemose et al., *PLOS* Computational Biology, May 15, 2014, Vol. 10, Issue 5) of the human genome remains unassembled, unmapped, and poorly characterized. The reference assembly generally annotates these missing regions as multi-megabase heterochromatic gaps, found primarily near centromeres and on the short arms of the acrocentric chromosomes. This missing fraction of the genome includes structural features that remain resistant to accurate characterization using generally used sequencing technologies. Additional exemplary regions that are resistant to accurate characterization include without limitation areas that have close homologous pseudogenes (for example SMN1/2 Cyp2d6), areas that have substantial repeated sequences throughout the genome, including without limitation transposons (such as SINEs, LINEs), as well as areas that have tremendous variation for which reference sequences serve as a poor guide (such as the regions encoding the genes for the human leukocyte antigen (HLA) complex). The methods, compositions, and systems described herein combine selective amplification of the regions of interest with the ability to maintain molecular context, thereby allowing for de novo sequencing of genomic regions that are generally poorly characterized.

[0095] In specific instances, methods described herein include a step in which selected regions of the genome are selectively amplified prior to sequencing. This amplification, which is generally conducted using methods known in the art (including without limitation PCR amplification) provides at least 1X, 2X, 3X, 4X, 5X, 6X, 7X, 8X, 9X, 10X, 11X, 12X, 13X, 14X, 15X, 16X, 17X, 18X, 19X, or 20X coverage of the selected regions of the genome, thereby providing a quantity of nucleic acids to allow de novo sequencing of those selected regions. In further embodiments, the amplification provides at least 1X-30X, 2X-25X, 3X-20X, 4X-15X, or 5X-10X coverage of the selected regions of the genome.

[0096] The amplification is generally conducted through extension of primers complementary to sequences within or near the selected regions of the genome. In some cases, a library of primers is used that is designed to tile across the regions of interest – in other words, the library of primer is designed to amplify regions at specific distances along the selected regions of the genome. In some instances, the selective amplification utilizes primers that are complementary to every 10, 15, 20, 25, 50, 100, 200, 250, 500, 750, 1000, or 10000 bases along the selected regions of the genome. In still further examples, the tiled library of primers is designed to capture a mixture of distances – that mixture can be a random mixture of distances or intelligently designed such that specific portions or percentages of the selected regions are amplified by different primer pairs. In further embodiments, the primer pairs are designed such that each pair amplifies about 1-5%, 2-10%, 3-15%, 4-20%, 5-25%, 6-30%, 7-35%, 8-40%, 9-45%, or 10-50% of any contiguous region of a selected portion of the genome.

[0097] In certain embodiments and in accordance with any of the description above, the amplification occurs across a region of the genome that is at least 3 megabasepairs long (Mb). In further embodiments, the selected region of the genome that is selectively amplified in accordance with any of the methods described herein is at least 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, or 10 Mb. In yet further embodiments, the selected region of the genome is about 2-20, 3-18, 4-16, 5-14, 6-12, or 7-10 Mb in length. As discussed above, amplification may occur across these regions using a single primer pair complementary to sequences at the ends or near the ends of these regions. In other embodiments, amplification is conducted with a library of primer pairs that are tiled across the length of the region, such that regular segments, random segments, or some combination of different segment distances along the region are amplified, with the extent of coverage in accordance with the description above.

[0098] In some embodiments, the primers used in selective amplification of selected regions of the genome contain uracils so that the primers themselves are not amplified.

[0099] In general, the methods and systems described herein accomplish targeted genomic sequencing by providing for the determination of the sequence of selected regions of the genome, and this sequencing information is generally obtained using methods that have the advantages of the extremely low sequencing error rates and high throughput of short read sequencing technologies. As described previously, an advantage of the methods and systems described herein is that they can achieve the desired results through the use of ubiquitously available, short read sequencing technologies. Such technologies have the advantages of being readily available and widely dispersed within the research community, with protocols and reagent systems that are well characterized and highly effective. These short read sequencing technologies include those available from, e.g., Illumina, inc. (GAIIx, NextSeq, MiSeq, HiSeq, X10), Ion Torrent division of Thermo-Fisher (Ion Proton and Ion PGM), pyrosequencing methods, as well as others.

[00100] Of particular advantage is that the methods and systems described herein utilize these short read sequencing technologies and do so with their associated low error rates. In particular, the methods and systems described herein achieve the desired individual molecular readlengths or context, as described above, but with individual sequencing reads, excluding mate pair extensions, that are shorter than 1000 bp, shorter than 500 bp, shorter than 300 bp, shorter than 200 bp, shorter than 150 bp or even shorter; and with sequencing error rates for such individual molecular readlengths that are less than 5%, less than 1%, less than 0.5%, less than 0.1%, less than 0.05%, less than 0.01%, less than 0.005%, or even less than 0.001%.

[0100] Methods of processing and sequencing nucleic acids in accordance with the methods and systems described in the present application are also described in further detail in USSNs 14/316,383; 14/316,398; 14/316,416; 14/316,431; 14/316,447; and 14/316,463 which are herein incorporated by reference in their entirety for all purposes and in particular for all written description, figures and working examples directed to processing nucleic acids and sequencing and other characterizations of genomic material.

[0101] Regardless of the sequencing platform used, in general and in accordance with any of the methods described herein, sequencing of nucleic acids is typically carried out in a manner that preserves the molecular context of sequence reads or portions of sequence reads. By that is meant that multiple sequence reads or multiple portions of sequence reads may be attributable to a single originating molecule of a nucleic acid. By ‘attributable to’ is meant that the sequence reads can be identified as falling within the linear sequence of bases of their particular originating molecules of a nucleic acid – in other words, and in reference to Figure 7, if fragments 703, 704, 705 and 706 are generated from originating nucleic acid molecules 701 and 702, then the sequencing is carried out in a manner such that sequence reads from fragments 703, 704, 705 and 706 retain their molecular context and it is readily ascertained that fragments 703 and 704 are derived from originating molecule 701 whereas 705 and 706 are derived from originating molecule 702, even if all the fragments are pooled together for the sequencing reaction. In addition, the sequencing is generally conducted such that not only is the originating molecule ascertained, but so also is the relative position of each fragment along that linear molecule – e.g., it can be determined that fragment 703 is “upstream” from fragment 704 along the linear sequence of originating nucleic acid 701. In general, molecular context is retained through the use of any identifier or any other method of distinguishing one or more fragments from other fragments. In general, such identifiers are used on fragments that have been separated into groups or into individual entities. In some examples, such separation is separation into discrete partitions, although, as will be appreciated, any other methods of separating molecules can be used. In still further examples, the identifiers used are barcodes, and the linear position is determined both through barcoding as well as algorithmic assembly of sequence reads from overlapping fragments. Although for the sake of clarity much of the discussion herein is in terms of separation into partitions and/or barcoding, it will be appreciated that any methods of separating originating nucleic acid molecules and any methods of identifying or otherwise attributing fragments are of use in the methods and systems described herein.

[0102] As will be appreciated, while the single originating molecule of a nucleic acid may be of any of a variety of lengths, in preferred aspects, it will be a relatively long molecule, allowing for preservation of long range molecular context. In particular, the single originating molecule is preferably substantially longer than the typical short read sequence length, e.g., longer than 200 bases, and is often at least 1000 bases or longer, 5000 bases or longer, 10,000 bases or longer, 20,000 bases or longer, 30,000 bases or longer, 40,000 bases or longer, 50,000 bases or longer, 60,000 bases or longer, 70,000 bases or longer, 80,000 bases or longer, 90,000 bases or longer, or 100,000 bases or longer, and in some cases 1 megabase or longer.

[0103] In certain situations, sequencing methods described herein include a combination of deep coverage of the selected regions with lower level linked reads across longer ranges of the genome. As will be appreciated, this combination of de novo and re-sequencing provides an efficient way to sequence an entire genome and/or large portions of a genome. Targeted coverage of poorly characterized and/or highly polymorphic regions through the selective amplification methods described herein provides the amount of nucleic acid material necessary for de novo sequence assembly, whereas linked genomic sequencing over other regions of the genome maintains high throughput sequencing of the remainder of the genome. The methods and compositions described herein are uniquely amenable to allowing for this combination of de novo and linked read sequencing, because the same sequencing platform can be used for both types of coverage. The population of nucleic acids and/or nucleic acid fragments that are sequenced in accordance with the methods described herein contain sequences from both the genomic regions for de novo sequencing and the genomic regions for re-sequencing – the proportion of nucleic acids covering the regions of interest for de novo sequencing is higher than the nucleic acids covering the other regions of the genome due to the targeted amplification methods described in further detail herein.

[0104] In general, as shown in Figure 1, the methods and systems described herein may be used to characterize nucleic acids, particularly nucleic acids from selected regions of the genome, while retaining molecular context. As shown, two discrete individual nucleic acids 102 and 104 are illustrated, each having a number of regions of interest, e.g., region 106 and 108 in nucleic acid 102, and regions 110 and 112 in nucleic acid 104. The regions of interest in each nucleic acid are linked within (e.g., originate from) the same nucleic acid molecule, but in some cases these regions may be relatively separated from each other, e.g., more than 1 kb apart, more than 5 kb apart, more than 10kb apart, more than 20kb apart, more than 30 kb apart, more than 40kb apart, more than 50 kb apart, and in some cases, as much as 100 kb apart. The regions of interest are generally discrete and separate parts of the genome – in some cases, such regions are poorly characterized regions. The regions of interest may also denote individual genes, gene groups, exons. As shown, each nucleic acid 102 and 104 is separated into its own partition 114 and 116, respectively. As noted elsewhere herein, these partitions are, in many cases, aqueous droplets in a water in oil emulsion. Within each droplet, portions of each fragment are copied in a manner that preserves the original molecular context of those fragments, *e.g.*, as having originated from the same molecule. As shown, this is achieved through the inclusion in each copied fragment of a barcode sequence, *e.g.*, barcode sequence “1” or “2” as illustrated, that is representative of the droplet into which the originating fragment was partitioned. For whole genome sequence analysis applications, one could simply pool all of the copied fragments and their associated barcodes, in order to sequence and reassemble the full range sequence information from each of the originating nucleic acids 102 and 104. However, in many cases, it is more desirable to only analyze specific targeted portions of the overall genome, in order to provide greater focus on scientifically relevant portions of the genome, and to minimize the time and expense of performing sequencing on less relevant or irrelevant portions of the genome.

[0105] In accordance with the above, in addition to the barcoding step, there may be one or more steps of selective amplification, such that if nucleic acids 102 or 104 contain selected genomic regions of interest, amplicons from those regions will form a larger percentage of the fragments in each of the partitions 114 and 116. This amplification step will generally take place prior to or simultaneously with the attachment of the barcodes in accordance with the methods described herein, although in some embodiments the amplification step may also occur subsequent to attachment of the barcodes.

[0106] Because the pooled fragments within library 118 retain their original molecular context, e.g., through the retention of the barcode information, they may be reassembled into their original molecular contexts with embedded (at times, long range) linkage information, e.g., with inferred linkage as between each of the assembled regions of interest 106:108 and 110:112. By way of example, one may identify direct molecular linkage between two disparate targeted portions of the genome, e.g., two or more exons, and that direct molecular linkage may be used to identify structural variations and other genomic characteristics. For situations in which selective amplification is utilized to increase the amount of nucleic acid fragments containing portions of selected regions of the genome, then the ability to identify the molecular context also provides a way to sequence those selected regions of the genome, often at a depth that allows for de novo assembly of those regions.

[0107] Generally, methods of the invention include steps as illustrated in Figure 2, which provides a schematic overview of methods of the invention discussed in further detail herein. As will be appreciated, the method outlined in Figure 2 is an exemplary embodiment that may be altered or modified as needed and as described herein.

[0108] As shown in Figure 2, the methods described herein will in most examples include a step in which sample nucleic acids containing the targeted regions of interest are partitioned (201). Generally, each partition containing nucleic acids from genomic regions of interest will undergo a targeted enrichment to produce a population of fragments in which a large proportion will contain sequences from a selected genomic region (202). Those fragments are then further fragmented or copied in such a way as to preserve the original molecular context of the fragments (203), usually by barcoding the fragments that are specific to the partition in which they are contained. Each partition may in some examples include more than one nucleic acid, and will in some instances contain several hundred nucleic acid molecules – in situations in which multiple nucleic acids are within a partition, any particular locus of the genome will generally be represented by a single individual nucleic acid prior to barcoding. The barcoded fragments of step 203 can be generated using any methods known in the art – in some examples, oligonucleotides are the samples within the distinct partitions. Such oligonucleotides may comprise random sequences intended to randomly prime numerous different regions of the samples, or they may comprise a specific primer sequence targeted to prime upstream of a targeted region of the sample. In further examples, these oligonucleotides also contain a barcode sequence, such that the replication process also barcodes the resultant replicated fragment of the original sample nucleic acid. A particularly elegant process for use of these barcode oligonucleotides in amplifying and barcoding samples is described in detail in U.S. Patent Application Nos. 14/316,383; 14/316,398; 14/316,416; 14/316,431; 14/316,447; and 14/316,463 which are herein incorporated by reference in their entirety for all purposes and in particular for all written description, figures and working examples directed to processing nucleic acids and sequencing and other characterizations of genomic material. Extension reaction reagents, e.g., DNA polymerase, nucleoside triphosphates, co-factors (e.g., Mg^{2+} or Mn^{2+} etc.), that are also contained in the partitions, then extend the primer sequence using the sample as a template, to produce a complementary fragment to the strand of the template to which the primer annealed, and the complementary fragment includes the oligonucleotide and its associated barcode sequence. Annealing and extension of multiple primers to different portions of the sample can result in a large pool of overlapping complementary fragments of the sample, each possessing its own barcode sequence indicative of the partition in which it was created. In some cases, these complementary fragments may themselves be used as a template primed by the oligonucleotides present in the partition to produce a complement of the complement that again, includes the barcode sequence. In further examples, this replication process is configured such that when the first complement is duplicated, it produces two complementary sequences at or near its termini to allow the formation of a hairpin structure or partial hairpin structure, which reduces the ability of the molecule to be the basis for producing further iterative copies.

[0109] Returning to the method exemplified in Figure 2, once the partition-specific barcodes are attached to the copied fragments, the barcoded fragments are then pooled (204). The pooled fragments are then sequenced (205) and the sequences of the fragments are attributed to their originating molecular context (206), such that the targeted regions of interest are both identified and also linked with that originating molecular context. An advantage of the methods and systems described herein is that attaching a partition- or sample-specific barcode to the copied fragments prior to enriching the fragments for targeted genomic regions preserves the original molecular context of those targeted regions, allowing them to be attributed to their original partition and thus their originating sample nucleic acid.

[0110] In addition to the above workflow, targeted genomic regions may be further enriched, isolated or separated, *i.e.,* “pulled down,” for further analysis, particularly sequencing, using methods that include both chip-based and solution-based capture methods. Such methods utilize probes that are complementary to the genomic regions of interest or to regions near or adjacent to the genomic regions of interest. For example, in hybrid (or chip-based) capture, microarrays containing capture probes (usually single-stranded oligonucleotides) with sequences that taken together cover the region of interest are fixed to a surface. Genomic DNA is fragmented and may further undergo processing such as end-repair to produce blunt ends and/or addition of additional features such as universal priming sequences. These fragments are hybridized to the probes on the microarray. Unhybridized fragments are washed away and the desired fragments are eluted or otherwise processed on the surface for sequencing or other analysis, and thus the population of fragments remaining on the surface is enriched for fragments containing the targeted regions of interest (e.g., the regions comprising the sequences complementary to those contained in the capture probes). The enriched population of fragments may further be amplified using any amplification technologies known in the art. Exemplary methods for such targeted pull down enrichment methods are described in USSN 14/927,297, filed on October 29, 2015, which is hereby incorporated by reference in its entirety for all purposes and in particular for all teachings related to targeted pull down enrichment methods and sequencing methods, including all written description, figures and examples.

[0111] In some aspects, methods for coverage of selected regions of the genome include methods in which the discrete partitions containing nucleic acid molecules and/or fragments thereof from those selected regions are themselves sorted for further processing. As will be appreciated, this sorting of the discrete partitions may take place in any combination with other methods of selective amplification and/or targeted pull-down of genomic regions of interest described herein, in particular in any combination with the steps of the work flow described above.

[0112] In general, such methods of sorting of the discrete partitions includes steps in which partitions containing at least a portion of the one or more selected portions of the genome are separated from partitions that do not contain any sequences from those portions of the genome. These methods include the steps of providing a population enriched for sequences of the fragments comprising at least a portion of the one or more selected portions of the genome within the discrete partitions containing sequences from those portions of the genome. Such enrichment is generally accomplished through the use of directed PCR amplification of the fragments within the discrete partitions that include at least a portion of the one or more selected portions of the genome to produce a population. This directed PCR amplification thus produces amplicons comprising at least a portion of the one or more selected portions of the genome. In certain embodiments, these amplicons are attached to a detectable label, which in some non-limiting embodiments may include a fluorescent molecule. In general, such attachment occurs such that only those amplicons generated from the fragments containing the one or more selected portions of the genome are attached to the detectable label. In some embodiments, the attachment of the detectable labels occurs during the selective amplification of the one or more selected portions of the genome. Such detectable labels may in further embodiments include without limitation fluorescent labels, electrochemical labels, magnetic beads, and nanoparticles. This attachment of the detectable label can be accomplished using methods known in the art. In yet further embodiments, discrete partitions containing fragments comprising at least a portion of the one or more selected portions of the genome are sorted based on signals emitted from the detectable labels attached to the amplicons within those partitions.

[0113] In further embodiments, the steps of sorting discrete partitions containing selected portions of the genome from those that do not contain such sequences include the steps of (a) providing starting genomic material; (b) distributing individual nucleic acid molecules from the starting genomic material into discrete partitions such that each discrete partition contains a first individual nucleic acid molecule; (c) providing a population within at least some of the discrete partitions that is enriched for sequences of the fragments comprising at least a portion of the one or more selected portions of the genome; (d) attaching a common barcode sequence to the fragments within each discrete partition such that each of the fragments is attributable to the discrete partition in which it was contained; (e) separating discrete partitions containing fragments comprising at least a portion of the one or more selected portions of the genome from discrete partitions containing no fragments comprising the one or more selected portions of the genome; (f) obtaining sequence information from the fragments comprising at least a portion of the one or more selected portions of the genome, thereby sequencing one or more targeted portions of the genomic sample while retaining molecular context.

[0114] In further embodiments and in accordance with any of the above, prior to obtaining sequence information from the fragments, the discrete partitions are combined and the fragments are pooled together. In further embodiments, the step of obtaining sequence information from the fragments is conducted in such a way as to maintain the molecular context of the sequences of the fragments, such that the identifying further comprises identifying fragments derived from the same first individual nucleic acid molecules. In still further embodiments, this obtaining of sequence information includes a sequencing reaction selected from the group consisting of: short read-length sequencing reactions and long read-length sequencing reactions. In yet further embodiments, the sequencing reaction is a short read, high accuracy sequencing reaction.

[0115] In still further embodiments and in accordance with any of the above, the discrete partitions comprise droplets in an emulsion. In further embodiments, the barcoded fragments within the discrete partitions represent about 1X-10X coverage of the one or more selected portions of the genome. In still further embodiments, the barcoded fragments within the discrete partitions represent about 2X-5X coverage of the one or more selected portions of the genome. In yet further embodiments, the barcoded fragments of the amplicons within the discrete partitions represent at least 1X coverage of the one or more selected portions of the genome. In still further embodiments, the barcoded fragments within the discrete partitions represent at least 2X or 5X coverage of the one or more selected portions of the genome.

[0116] In addition to providing the ability to obtain sequence information from selected regions of the genome, the methods and systems described herein can also provide other characterizations of genomic material, including without limitation haplotype phasing, identification of structural variations, and identifying copy number variations, as described in US Patent Application Nos. 14/752,589 and 14/752,602, which are herein incorporated by reference in their entirety for all purposes and in particular for all written description, figures and working examples directed to characterization of genomic material.

[0117] As noted above, the methods and systems described herein provide individual molecular context for short sequence reads of longer nucleic acids. As used herein, individual molecular context refers to sequence context beyond the specific sequence read, e.g., relation to adjacent or proximal sequences, that are not included within the sequence read itself, and as such, will typically be such that they would not be included in whole or in part in a short sequence read, e.g., a read of about 150 bases, or about 300 bases for paired reads. In particularly preferred aspects, the methods and systems provide long range sequence context for short sequence reads. Such long range context includes relationship or linkage of a given sequence read to sequence reads that are within a distance of each other of longer than 1 kb, longer than 5 kb, longer than 10 kb, longer than 15 kb, longer than 20 kb, longer than 30 kb, longer than 40 kb, longer than 50 kb, longer than 60 kb, longer than 70 kb, longer than 80 kb, longer than 90 kb or even longer than 100 kb, or longer. As will be appreciated, by providing long range individual molecular context, one can also derive the phasing information of variants within that individual molecular context, e.g., variants on a particular long molecule will be, by definition commonly phased.

[0118] By providing longer range individual molecular context, the methods and systems of the invention also provide much longer inferred molecular context (also referred to herein as a "long virtual single molecule read"). Sequence context, as described herein can include mapping or providing linkage of fragments across different (generally on the kilobase scale) ranges of full genomic sequence. These methods include mapping the short sequence reads to the individual longer molecules or contigs of linked molecules, as well as long range sequencing of large portions of the longer individual molecules, e.g., having contiguous determined sequences of individual molecules where such determined sequences are longer than 1 kb, longer than 5 kb, longer than 10 kb, longer than 15 kb, longer than 20 kb, longer than 30 kb, longer than 40 kb, longer than 50 kb, longer than 60 kb, longer than 70 kb, longer than 80 kb, longer than 90 kb or even longer than 100 kb. As with sequence context, the attribution of short sequences to longer nucleic acids, e.g., both individual long nucleic acid molecules or collections of linked nucleic acid molecules or contigs, may include both mapping of short sequences against longer nucleic acid stretches to provide high level sequence context, as well as providing assembled sequences from the short sequences through these longer nucleic acids.

[0119] Furthermore, while one may utilize the long range sequence context associated with long individual molecules, having such long range sequence context also allows one to infer even longer range sequence context. By way of one example, by providing the long range molecular context described above, one can identify overlapping variant portions, e.g., phased variants, translocated sequences, etc., among long sequences from different originating molecules, allowing the inferred linkage between those molecules. Such inferred linkages or molecular contexts are referred to herein as "inferred contigs". In some cases when discussed in the context of phased sequences, the inferred contigs may represent commonly phased sequences, e.g., where by virtue of overlapping phased variants, one can infer a phased contig of substantially greater length than the individual originating molecules. These phased contigs are referred to herein as "phase blocks".

[0120] By starting with longer single molecule reads (e.g., the "long virtual single molecule reads" discussed above), one can derive longer inferred contigs or phase blocks than would otherwise be attainable using short read sequencing technologies or other approaches to phased sequencing. See, e.g., published U.S. Patent Application No. 2013-0157870. In particular, using the methods and systems described herein, one can obtain inferred contig or phase block lengths having an N50 (where the sum of the block lengths that are greater than the stated N50 number is 50% of the sum of all block lengths) of at least about 10kb, at least about 20kb, at least about 50kb. In more preferred aspects, inferred contig or phase block lengths having an N50 of at least about 100kb, at least about 150kb, at least about 200kb, and in many cases, at least about 250kb, at least about 300 kb, at least about 350 kb, at least about 400 kb, and in some cases, at least about 500 kb or more, are attained. In still other cases, maximum phase block lengths in excess of

200 kb, in excess of 300 kb, in excess of 400 kb, in excess of 500 kb, in excess of 1 Mb, or even in excess of 2 Mb may be obtained.

[0121] In one aspect, and in conjunction with any of the capture methods described above and later herein, the methods and systems described herein provide for the compartmentalization, depositing or partitioning of sample nucleic acids, or fragments thereof, into discrete compartments or partitions (referred to interchangeably herein as partitions), where each partition maintains separation of its own contents from the contents of other partitions. Unique identifiers, e.g., barcodes, may be previously, subsequently or concurrently delivered to the partitions that hold the compartmentalized or partitioned sample nucleic acids, in order to allow for the later attribution of the characteristics, e.g., nucleic acid sequence information, to the sample nucleic acids included within a particular compartment, and particularly to relatively long stretches of contiguous sample nucleic acids that may be originally deposited into the partitions.

[0122] The sample nucleic acids utilized in the methods described herein typically represent a number of overlapping portions of the overall sample to be analyzed, e.g., an entire chromosome, exome, or other large genomic portion. These sample nucleic acids may include whole genomes, individual chromosomes, exomes, amplicons, or any of a variety of different nucleic acids of interest. The sample nucleic acids are typically partitioned such that the nucleic acids are present in the partitions in relatively long fragments or stretches of contiguous nucleic acid molecules. Typically, these fragments of the sample nucleic acids may be longer than 1 kb, longer than 5 kb, longer than 10 kb, longer than 15 kb, longer than 20 kb, longer than 30 kb, longer than 40 kb, longer than 50 kb, longer than 60 kb, longer than 70 kb, longer than 80 kb, longer than 90 kb or even longer than 100 kb, which permits the longer range molecular context described above.

[0123] The sample nucleic acids are also typically partitioned at a level whereby a given partition has a very low probability of including two overlapping fragments of the starting sample nucleic acid. This is typically accomplished by providing the sample nucleic acid at a low input amount and/or concentration during the partitioning process. As a result, in preferred cases, a given partition may include a number of long, but non-overlapping fragments of the starting sample nucleic acids. The sample nucleic acids in the different partitions are then associated with unique identifiers, where for any given partition, nucleic acids contained therein possess the same unique identifier, but where different partitions may include different unique identifiers. Moreover, because the partitioning step allocates the sample components into very small volume partitions or droplets, it will be appreciated that in order to achieve the desired allocation as set forth above, one need not conduct substantial dilution of the sample, as would be required in higher volume processes, e.g., in tubes, or wells of a multiwell plate. Further, because the systems described herein employ such high levels of barcode diversity, one can allocate diverse barcodes among higher numbers of genomic equivalents, as provided above. In particular, previously described, multiwell plate approaches (see, e.g., U.S. Published Application No. 2013-0079231 and 2013-0157870) typically only operate with a hundred to a few hundred different barcode sequences, and employ a limiting dilution process of their sample in order to be able to attribute barcodes to different cells/nucleic acids. As such, they will generally operate with far fewer than 100 cells, which would typically provide a ratio of genomes:(barcode type) on the order of 1:10, and certainly well above 1:100. The systems described herein, on the other hand, because of the high level of barcode diversity, e.g., in excess of 10,000, 100,000, 500,000, 600,000, 700,000 etc. diverse barcode types, can operate at genome:(barcode type) ratios that are on the order of 1:50 or less, 1:100 or less, 1:1000 or less, or even smaller ratios, while also allowing for loading higher numbers of genomes (e.g., on the order of greater than 100 genomes per assay, greater than 500 genomes per assay, 1000 genomes per assay, or even more) while still providing for far improved barcode diversity per genome.

[0124] Often, the sample is combined with a set of oligonucleotide tags that are releasably-attached to beads prior to the partitioning step. In some examples, amplification methods are used to add barcodes to the resultant amplification products, which in some examples contain smaller segments (fragments) of the full originating nucleic acid molecule from which they are derived. In some examples, methods using transposons are utilized as described in Amini et al, Nature Genetics 46: 1343-1349 (2014) (advance online publication on October 29, 2014), which is herein incorporated by reference in its entirety for all purposes and in particular for all teachings related to attaching barcodes or other oligonucleotide tags to nucleic acids. In further examples, methods of attaching barcodes can include the use of nicking enzymes or polymerases and/or invasive probes such as recA to produce gaps along double stranded sample nucleic acids - barcodes can then be inserted into those gaps.

[0125] In examples in which amplification is used to tag nucleic acid fragments, the oligonucleotide tags may comprise at least a first and second region. The first region may be a barcode region that, as between oligonucleotides within a given partition, may be substantially the same barcode sequence, but as between different partitions, may and, in most cases is a different barcode sequence. The second region may be an N-mer (either a random N-mer or an N-mer designed to target a particular sequence) that can be used to prime the nucleic acids within the sample within the partitions. In some cases, where the N-mer is designed to target a particular sequence, it may be designed to target a particular chromosome (e.g., chromosome 1, 13, 18, or 21), or region of a chromosome, e.g., an exome or other targeted region. As discussed herein, the N-mer may also be designed to selected regions of the genome that tend to be poorly characterized or are highly polymorphic or divergent from the reference sequence. In some cases, the N-mer may be designed to target a particular gene or genetic region, such as a gene or region associated with a disease or disorder (e.g., cancer). Within the partitions, an amplification reaction may be conducted using the second N-mer to prime the nucleic acid sample at different places along the length of the nucleic acid. As a result of the amplification, each partition may contain amplified products of the nucleic acid that are attached to an identical or near-identical barcode, and that may represent overlapping, smaller fragments of the nucleic acids in each partition. The bar-code can serve as a marker that signifies that a set of nucleic acids originated from the same partition, and thus potentially also originated from the same strand of nucleic acid. Following amplification, the nucleic acids may be pooled, sequenced, and aligned using a sequencing algorithm. Because shorter sequence reads may, by virtue of their associated barcode sequences, be aligned and attributed to a single, long fragment of the sample nucleic acid, all of the identified variants on that sequence can be attributed to a single originating fragment and single originating chromosome. Further, by aligning multiple co-located variants across multiple long fragments, one can further characterize that chromosomal contribution. Accordingly, conclusions regarding the phasing of particular genetic variants may then be drawn, as can analyses across long ranges of genomic sequence – for example, identification of sequence information across stretches of poorly characterized regions of the genome. Such information may also be useful for identifying haplotypes, which are generally a specified set of genetic variants that reside on the same nucleic acid strand or on different nucleic acid strands. Copy number variations may also be identified in this manner.

[0126] The described methods and systems provide significant advantages over current nucleic acid sequencing technologies and their associated sample preparation methods. Ensemble sample preparation and sequencing methods are predisposed towards primarily identifying and characterizing the majority constituents in the sample, and are not designed to identify and characterize minority constituents, e.g., genetic material contributed by one chromosome, from a poorly characterized or highly polymorphic region of the genome, or material from one or a few cells, or fragmented tumor cell DNA molecule circulating in the bloodstream, that constitute a small percentage of the total DNA in the extracted sample. The methods described herein include selective amplification methods that increase the genetic material from these minority constituents, and the ability to retain the molecular context of this genetic material further provides genetic characterization of these constituents. The described methods and systems also provide a significant advantage for detecting populations that are present within a larger sample. As such, they are particularly useful for assessing haplotype and copy number variations – the methods disclosed herein are also useful for providing sequence information over regions of the genome that are poorly characterized or are poorly represented in a population of nucleic acid targets due to biases introduced during sample preparation.

[0127] The use of the barcoding technique disclosed herein confers the unique capability of providing individual molecular context for a given set of genetic markers, i.e., attributing a given set of genetic markers (as opposed to a single marker) to individual sample nucleic acid molecules, and through variant coordinated assembly, to provide a broader or even longer range inferred individual molecular context, among multiple sample nucleic acid molecules, and/or to a specific chromosome. These genetic markers may include specific genetic loci, e.g., variants, such as SNPs, or they may include short sequences. Furthermore, the use of barcoding confers the additional advantages of facilitating the ability to discriminate between minority constituents and majority constituents of the total nucleic acid population extracted from the sample, e.g. for detection and characterization of circulating tumor DNA in the bloodstream, and also reduces or eliminates amplification bias during optional amplification steps. In addition, implementation in a microfluidics format confers the ability to work with extremely small sample volumes and low input quantities of DNA, as well as the ability to rapidly process large numbers of sample partitions (droplets) to facilitate genome-wide tagging.

[0128] As noted above, the methods and systems described herein provide individual molecular context for short sequence reads of longer nucleic acids. As used herein, individual molecular context refers to sequence context beyond the specific sequence read, e.g., relation to adjacent or proximal sequences, that are not included within the sequence read itself, and as such, will typically be such that they would not be included in whole or in part in a short sequence read, e.g., a read of about 150 bases, or about 300 bases for paired reads. In particularly preferred aspects, the methods and systems provide long range sequence context for short sequence reads. Such long range context includes relationship or linkage of a given sequence read to sequence reads that are within a distance of each other of longer than 1 kb, longer than 5 kb, longer than 10 kb, longer than 15 kb, longer than 20 kb, longer than 30 kb, longer than 40 kb, longer than 50 kb, longer than 60 kb, longer than 70 kb, longer than 80 kb, longer than 90 kb or even longer than 100 kb, or longer. By providing longer range individual molecular context, the methods and systems of the invention also provide much longer inferred molecular context. Sequence context, as described herein can include lower resolution context, e.g., from mapping the short sequence reads to the individual longer molecules or contigs of linked molecules, as well as the higher resolution sequence context, e.g., from long range sequencing of large portions of the longer individual molecules, e.g., having contiguous determined sequences of individual molecules where such determined sequences are longer than 1 kb, longer than 5 kb, longer than 10 kb, longer than 15 kb, longer than 20 kb, longer than 30 kb, longer than 40 kb, longer than 50 kb, longer than 60 kb, longer than 70 kb, longer than 80 kb, longer than 90 kb or even longer than 100 kb. As with sequence context, the attribution of short sequences to longer nucleic acids, e.g., both individual long nucleic acid molecules or collections of linked nucleic acid molecules or contigs, may include both mapping of short sequences against longer nucleic acid stretches to provide high level sequence context, as well as providing assembled sequences from the short sequences through these longer nucleic acids.

IV. Samples

[0129] As will be appreciated, the methods and systems discussed herein can be used to obtain targeted sequence information from any type of genomic material. Such genomic material may be obtained from a sample taken from a patient. Exemplary samples and types of genomic material of use in the methods and systems discussed herein include without limitation polynucleotides, nucleic acids, oligonucleotides, circulating cell-free nucleic acid, circulating tumor cell (CTC), nucleic acid fragments, nucleotides, DNA, RNA, peptide polynucleotides, complementary DNA (cDNA), double stranded DNA (dsDNA), single stranded DNA (ssDNA), plasmid DNA, cosmid DNA, chromosomal DNA, genomic DNA (gDNA), viral DNA, bacterial DNA, mtDNA (mitochondrial DNA), ribosomal RNA, cell-free DNA, cell free fetal DNA (cffDNA), mRNA, rRNA, tRNA, nRNA, siRNA, snRNA, snoRNA, scaRNA, microRNA, dsRNA, viral RNA, and the like. In summary, the samples that are used may vary depending on the particular processing needs.

[0130] Any substance that comprises nucleic acid may be the source of a sample. The substance may be a fluid, e.g., a biological fluid. A fluidic substance may include, but not limited to, blood, cord blood, saliva, urine, sweat, serum, semen, vaginal fluid, gastric and digestive fluid, spinal fluid, placental fluid, cavity fluid, ocular fluid, serum, breast milk, lymphatic fluid, or combinations thereof. The substance may be solid, for example, a biological tissue. The substance may comprise normal healthy tissues, diseased tissues, or a mix of healthy and diseased tissues. In some cases, the substance may comprise tumors. Tumors may be benign (non-cancer) or malignant (cancer). Non-limiting examples of tumors may include : fibrosarcoma, myxosarcoma, liposarcoma, chondrosarcoma, osteogenic sarcoma, chordoma, angiosarcoma, endotheliosarcoma, lymphangiosarcoma, lymphangioendotheliosarcoma, synovioma, mesothelioma, Ewing's sarcoma, leiomyosarcoma, rhabdomyosarcoma, gastrointestinal system carcinomas, colon carcinoma, pancreatic cancer, breast cancer, genitourinary system carcinomas, ovarian cancer, prostate cancer, squamous cell carcinoma, basal cell carcinoma, adenocarcinoma, sweat gland carcinoma, sebaceous gland carcinoma, papillary carcinoma, papillary adenocarcinomas, cystadenocarcinoma, medullary carcinoma, bronchogenic carcinoma, renal cell carcinoma, hepatoma, bile duct carcinoma, choriocarcinoma, seminoma, embryonal carcinoma, Wilms' tumor, cervical cancer, endocrine system carcinomas, testicular tumor, lung carcinoma, small cell lung carcinoma, non-small cell lung carcinoma, bladder carcinoma, epithelial carcinoma, glioma, astrocytoma, medulloblastoma, craniopharyngioma, ependymoma, pinealoma, hemangioblastoma, acoustic neuroma, oligodendroglioma, meningioma, melanoma, neuroblastoma, retinoblastoma, or combinations thereof. The substance may be associated with various types of organs. Non-limiting examples of organs may include brain, liver, lung, kidney, prostate, ovary, spleen, lymph node (including tonsil), thyroid, pancreas, heart, skeletal muscle, intestine, larynx, esophagus, stomach, or combinations thereof. In some cases, the substance may comprise a variety of cells, including but not limited to: eukaryotic cells, prokaryotic cells, fungi cells, heart cells, lung cells, kidney cells, liver cells, pancreas cells, reproductive cells, stem cells, induced pluripotent stem cells, gastrointestinal cells, blood cells, cancer cells, bacterial cells, bacterial cells isolated from a human microbiome sample, etc. In some cases, the substance may comprise contents of a cell, such as, for example, the contents of a single cell or the contents of multiple cells. Methods and systems for analyzing individual cells are provided in, e.g., U.S. Patent Application No. 14/752,641, filed June 26, 2015, the full disclosure of which is hereby incorporated by reference in its entirety.

[0131] Samples may be obtained from various subjects. A subject may be a living subject or a dead subject. Examples of subjects may include, but not limited to, humans, mammals, non-human mammals, rodents, amphibians, reptiles, canines, felines, bovines, equines, goats, ovines, hens, avines, mice, rabbits, insects, slugs, microbes, bacteria, parasites, or fish. In some cases, the subject may be a patient who is having, suspected of having, or at a risk of developing a disease or disorder. In some cases, the subject may be a pregnant woman. In some case, the subject may be a normal healthy pregnant woman. In some cases, the subject may be a pregnant woman who is at a risking of carrying a baby with certain birth defect.

[0132] A sample may be obtained from a subject by any means known in the art. For example, a sample may be obtained from a subject through accessing the circulatory system (e.g., intravenously or intra-arterially via a syringe or other apparatus), collecting a secreted biological sample (e.g., saliva, sputum urine, feces, etc.), surgically (e.g., biopsy) acquiring a biological sample (e.g., intra-operative samples, post-surgical samples, etc.), swabbing (e.g., buccal swab, oropharyngeal swab), or pipetting.

[0133] While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

EXAMPLES

Example: Targeted coverage of TP53 gene

[0134] An amplification reaction targeting the TP53 gene was conducted. Tumor protein p53, also known as p53, cellular tumor antigen p53 (UniProt name), phosphoprotein p53, tumor suppressor p53, antigen NY-CO-13, or transformation-related protein 53 (TRP53), is a protein that is encoded by the TP53 gene in humans. The p53 protein is crucial in multicellular organisms, where it regulates the cell cycle and, thus, functions as a tumor suppressor, preventing cancer. As such, p53 has been described as "the guardian of the genome" because of its role in conserving stability by preventing genome mutation. Hence TP53 is classified as a tumor suppressor gene.

[0135] Targeted amplification of the region of the genome containing the TP53 gene (which is about 19149 bp in length) was conducted using a total of 96 primers spanning the entire gene in a multiplex reaction. The primers were designed to tile across this region of the genome about 400 bp apart. The amplification reaction was conducted with a temperature gradient for the annealing step, 14 cycles, and an input amount of about 3 ng of DNA. The thermocycling protocol used for this example was as follows:

Initial Denaturation	98°C	30 seconds
18 Cycles	98°C	10 seconds
	30-55°C	15 seconds
	72°C	15 seconds
Final Extension	72°C	2 minutes
Hold	4C	

[0136] An exemplary workflow for this type of reaction is pictured in Figure 8. As will be appreciated, this is an exemplary embodiment of a method in accordance with the invention described herein and can be altered or expanded using known methods. As shown in Figure 8, the selected region of the genome (in this case, the TP53 gene) is amplified using target specific primers, such as those pictured as 802 and 803. In addition, a primer with barcode 801 was also incorporated into the amplicons, which can in certain embodiments as described herein provide molecular context for the subsequent sequence reads (808).

[0137] The primers 802 and 803 had in this experiment “tails” R1 and R2, which rendered the resultant amplicons amenable to sequencing on specific platforms, such as the Illumina platform. The amplification with the SI primer (806) further provided a sample index that is also used with the

Illumina platform. As will be appreciated, sequences that are useful for other sequencing platforms can be used in place of the R1 and R2 and S1 primers.

[0138] Figure 9 shows that the amplification reaction was specific, as the no template controls (NTC) showed no product. Figure 10 provides the fold-enrichment seen as a result of the above-described protocol across a range of temperatures.

[0139] The present specification provides a complete description of the methodologies, systems and/or structures and uses thereof in example aspects of the presently-described technology. Although various aspects of this technology have been described above with a certain degree of particularity, or with reference to one or more individual aspects, those skilled in the art could make numerous alterations to the disclosed aspects without departing from the spirit or scope of the technology hereof. Since many aspects can be made without departing from the spirit and scope of the presently described technology, the appropriate scope resides in the claims hereinafter appended. Other aspects are therefore contemplated. Furthermore, it should be understood that any operations may be performed in any order, unless explicitly claimed otherwise or a specific order is inherently necessitated by the claim language. It is intended that all matter contained in the above description and shown in the accompanying drawings shall be interpreted as illustrative only of particular aspects and are not limiting to the embodiments shown. Unless otherwise clear from the context or expressly stated, any concentration values provided herein are generally given in terms of admixture values or percentages without regard to any conversion that occurs upon or following addition of the particular component of the mixture. To the extent not already expressly incorporated herein, all published references and patent documents referred to in this disclosure are incorporated herein by reference in their entirety for all purposes. Changes in detail or structure may be made without departing from the basic elements of the present technology as defined in the following claims.

2016222719 09 Mar 2022

THE CLAIMS DEFINING THE INVENTION ARE AS FOLLOWS:

1. A method for obtaining sequence information from one or more portions of a genomic sample while retaining molecular context, the method comprising:
(a) providing starting genomic material;
(b) distributing individual nucleic acid molecules from the starting genomic material into discrete partitions such that each discrete partition contains an individual nucleic acid molecule, wherein the genomic material within each discrete partition comprises genomic DNA from a single cell;
(c) amplifying one or more targeted portions of at least some of the individual nucleic acid molecules in the discrete partitions to form a plurality of amplicons, wherein the amplifying is conducted with a library of primer pairs and wherein the library of primer pairs is designed to amplify at specific distances along the targeted portions;
(d) simultaneously with or following step (c), attaching a common barcode sequence to the amplicons within each discrete partition such that each of the amplicons is attributable to the discrete partition in which it was contained;
(e) obtaining sequence information from the amplicons, thereby sequencing one or more targeted portions of the genomic sample while retaining molecular context, wherein the library of primer pairs is selected to amplify genomic regions that are poorly characterized, highly polymorphic, or divergent from reference genome sequences.

2. The method of claim 1, wherein the one or more targeted portions comprise a contiguous region of the genome of at least 3.0 Mb in length.

3. The method of claim 1 or claim 2, wherein the obtaining step (e) comprises a sequencing reaction selected from the group consisting of: short read-length sequencing reactions and long read-length sequencing reactions.

4. The method of any one of claims 1 to 3, wherein the obtaining step (e) preserves the molecular context of the sequences of the amplicons, wherein the method further comprises identifying amplicons derived from the same first individual nucleic acid molecules.

5. The method of any one of claims 1 to 4, wherein the method further comprises linking two or more of the individual first fragment molecules in an inferred contig

based upon overlapping sequences of the plurality of amplicons, wherein the inferred contig comprises a length N50 of at least 10kb.

6. The method of any one of claims 1 to 5, wherein the barcode sequence further comprises additional sequence segments.

7. The method of claim 6, wherein the additional sequence segments comprise one or more of a member selected from the group consisting of: primers, attachment sequences, random n-mer oligonucleotides, oligonucleotides comprising uracil nucleobases.

8. The method of any one of claims 1 to 7, wherein the attaching step (d) comprises attaching a barcode selected from a library of at least 700,000 barcodes.

9. The method of any one of claims 1 to 8, wherein the discrete partitions comprise droplets in an emulsion.

10. The method of any one of claims 1 to 9, wherein prior to the obtaining step (e), the amplicons are further amplified such that the resultant amplification products are capable of forming partial or complete hairpin structures.

11. The method of any one of claims 1 to 10, wherein the barcoded amplicons within the discrete partitions represent about 100X-5,000X coverage of the one or more targeted portions.

12. The method of any one of claims 1 to 10, wherein the barcoded amplicons within the discrete partitions represent at least 1X coverage of the one or more targeted portions.

13. The method of any one of claims 1 to 10, wherein the barcoded amplicons within the discrete partitions represent at least 2,000X coverage of the one or more targeted portions.

14. The method of any one of claims 1 to 10, wherein the barcoded amplicons within the discrete partitions represent at least 5,000X coverage of the one or more targeted portions.

Figure 1

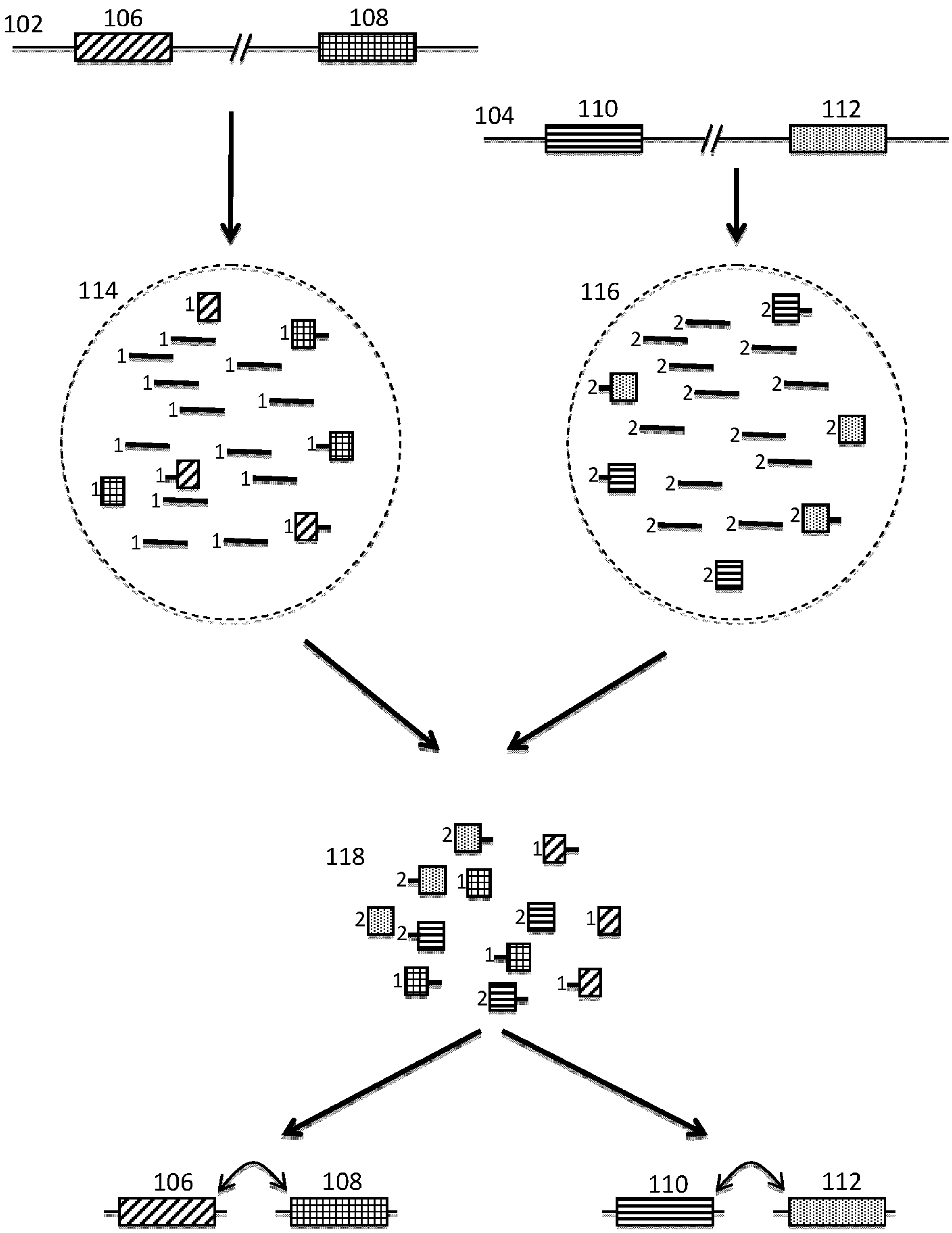


Figure 2

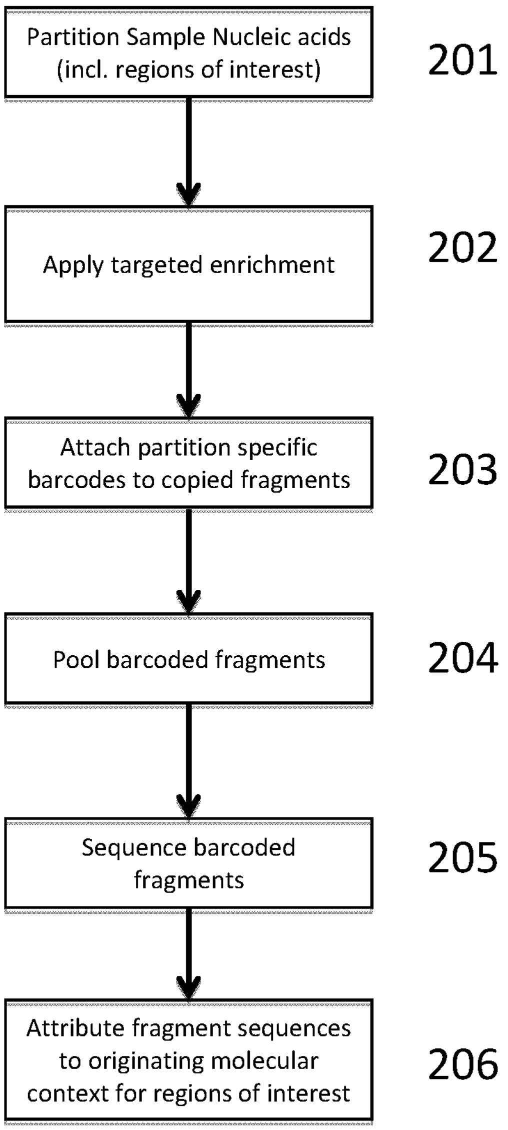
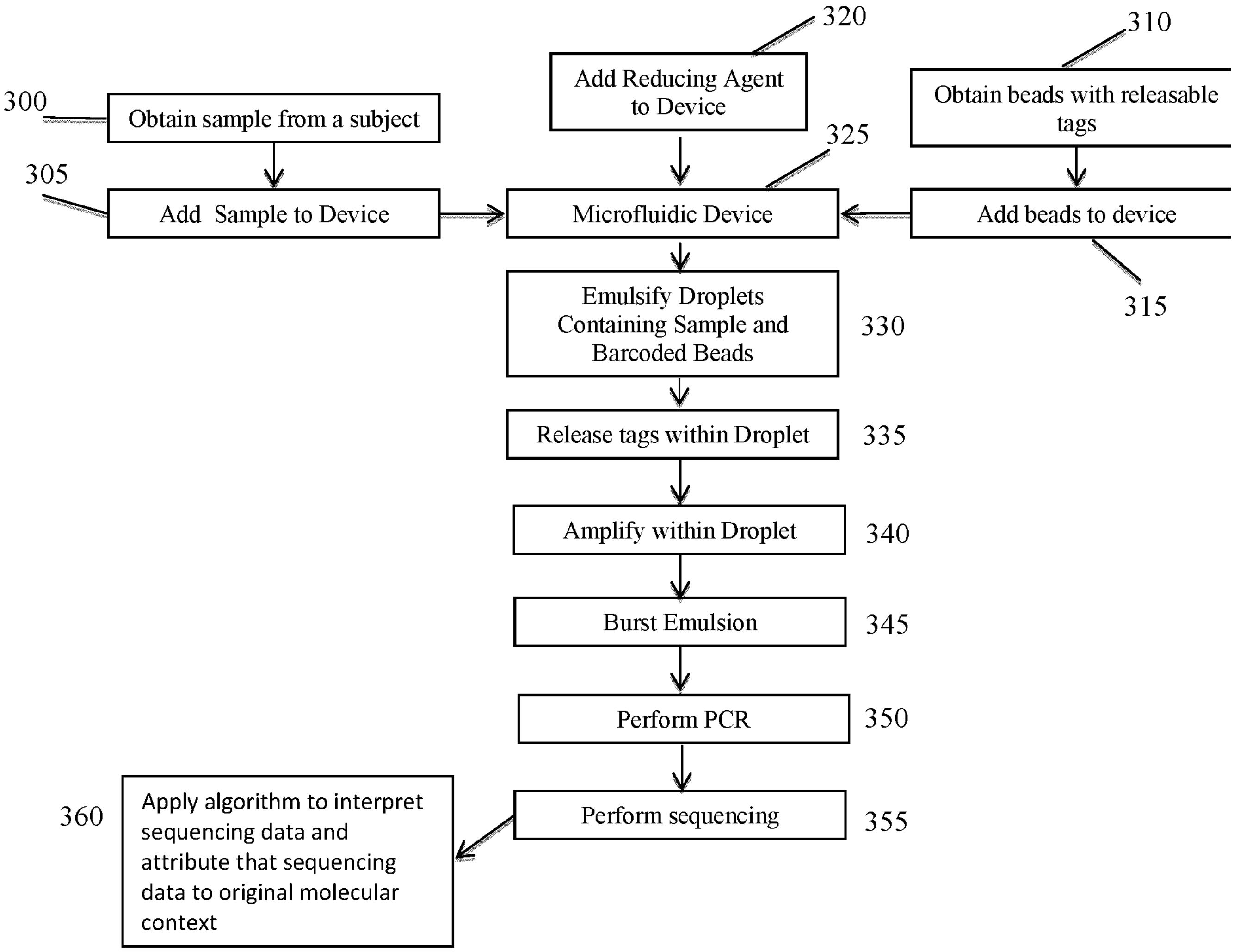
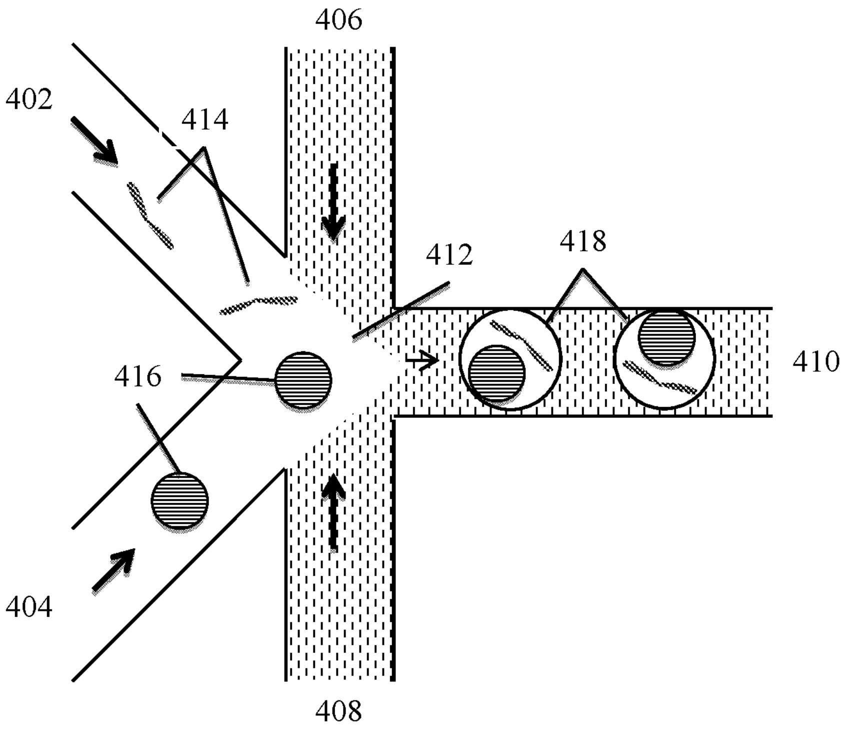

Partition Sample Nucleic acids (incl. regions of interest)
201
Apply targeted enrichment
202
Attach partition specific barcodes to copied fragments
203
Pool barcoded fragments
204
Sequence barcoded fragments
205
Attribute fragment sequences to originating molecular context for regions of interest
206

Figure 3



320
310
Add Reducing Agent to Device
Obtain beads with releasable tags
300
Obtain sample from a subject
325
305
Add Sample to Device
Microfluidic Device
Add beads to device
Emulsify Droplets Containing Sample and Barcoded Beads
330
315
Release tags within Droplet
335
Amplify within Droplet
340
Burst Emulsion
345
Perform PCR
350
360
Apply algorithm to interpret sequencing data and attribute that sequencing data to original molecular context
Perform sequencing
355

Figure 4



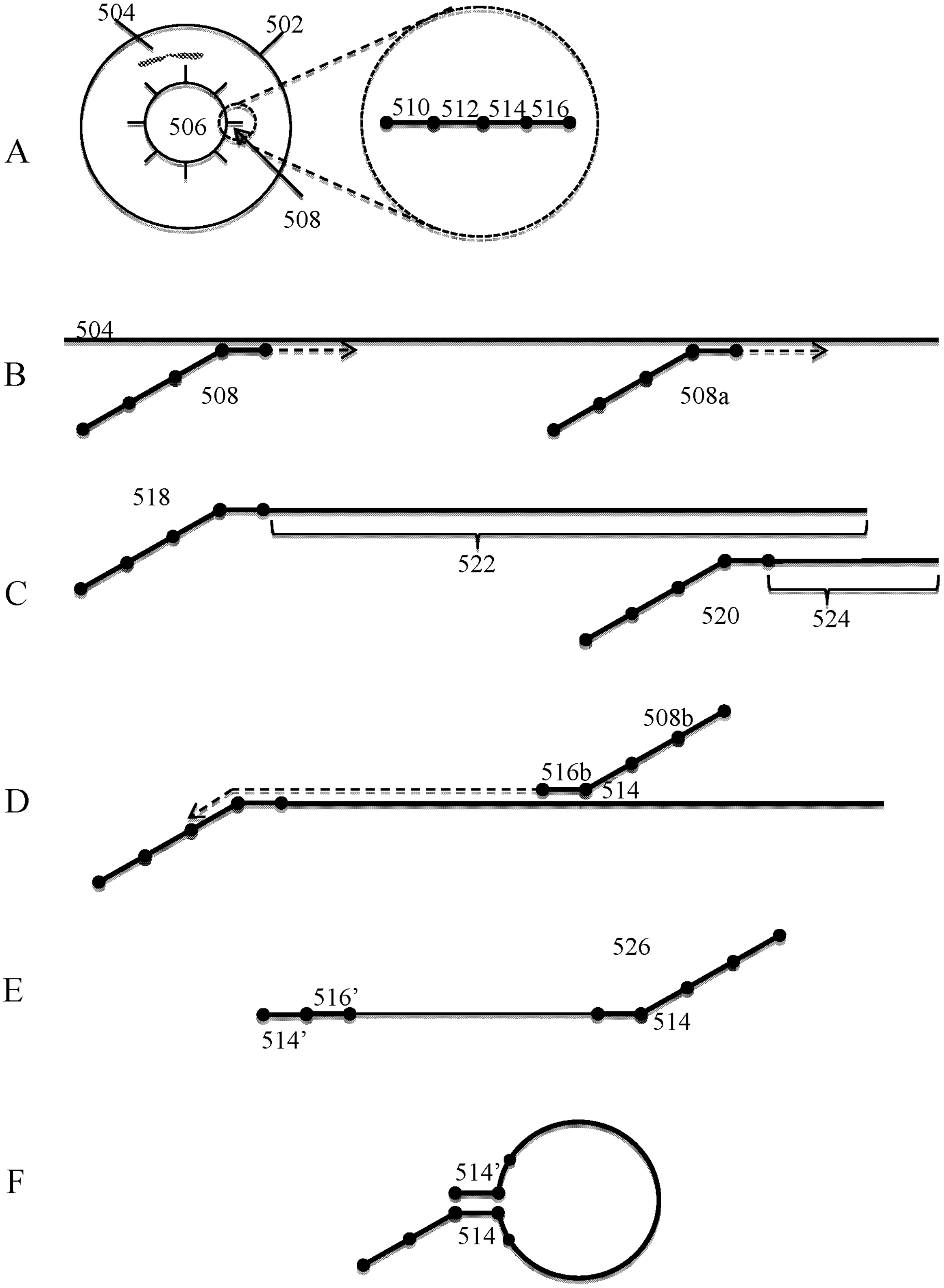

A
504
502
506
508
510
512
514
516
B
504
508
508a
C
518
522
520
524
D
508b
516b
514
E
526
516'
514'
514
F
514'
514

Figure 5

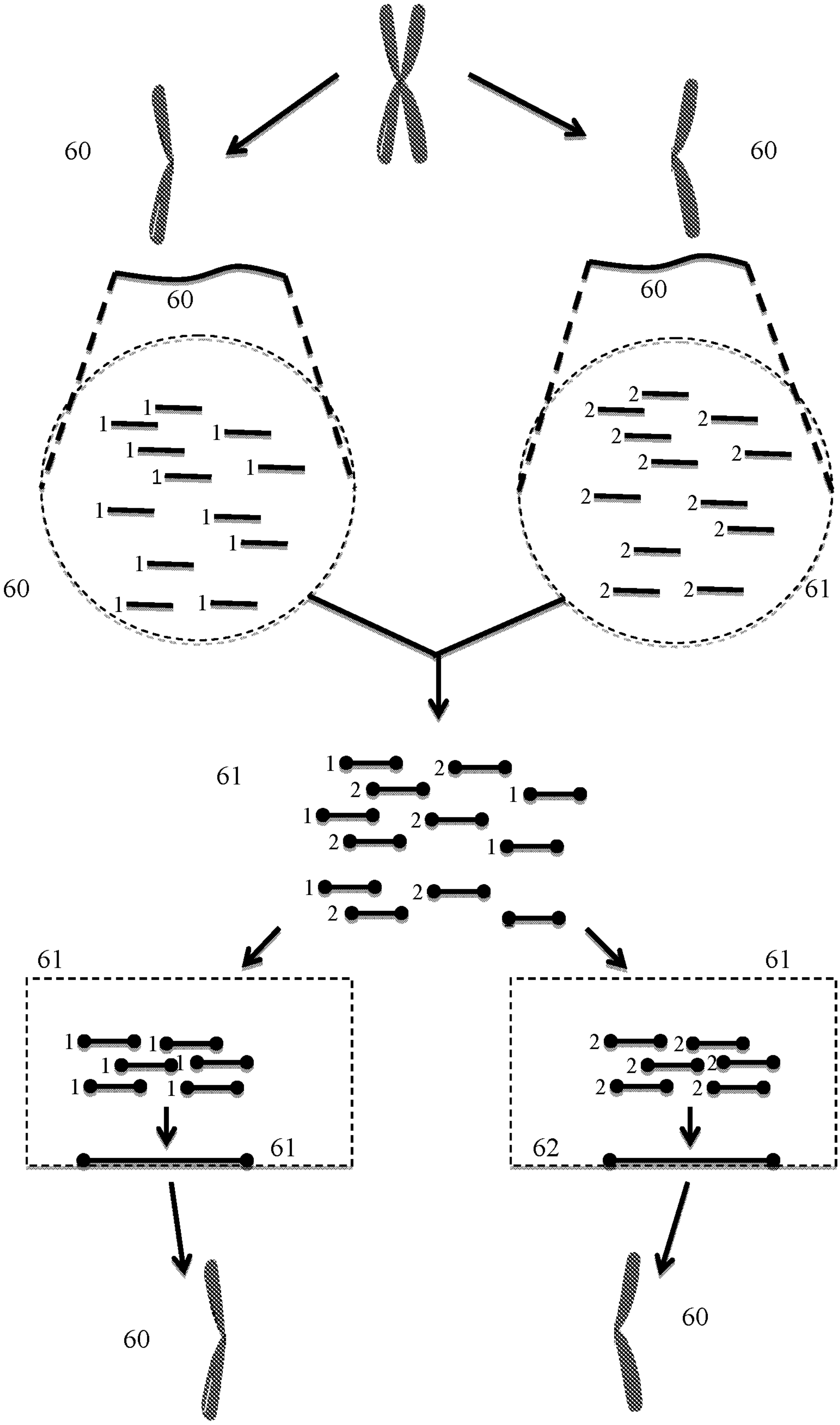


Figure 6A

Figure 6B

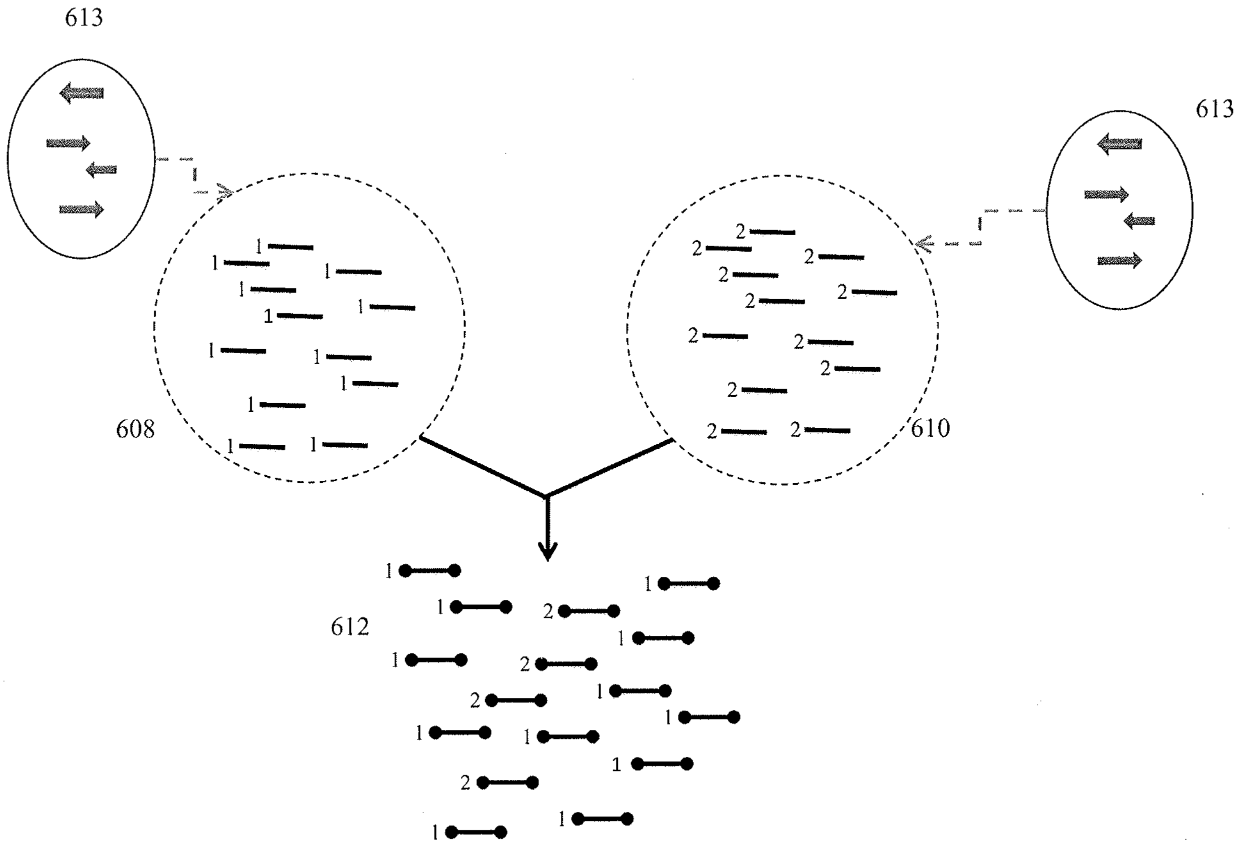


Figure 7

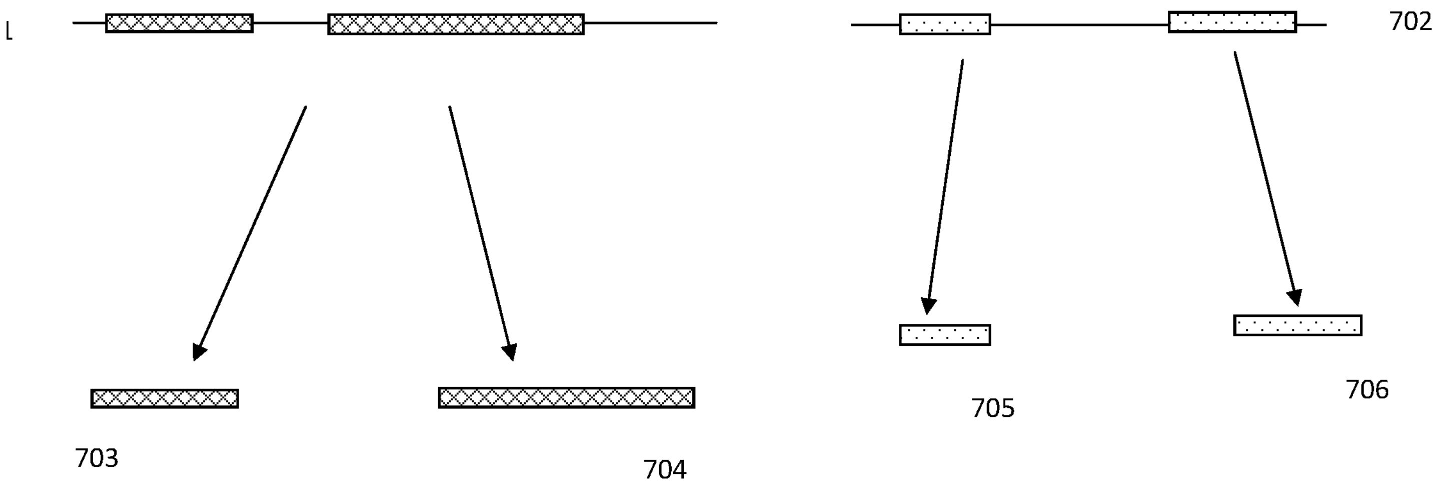
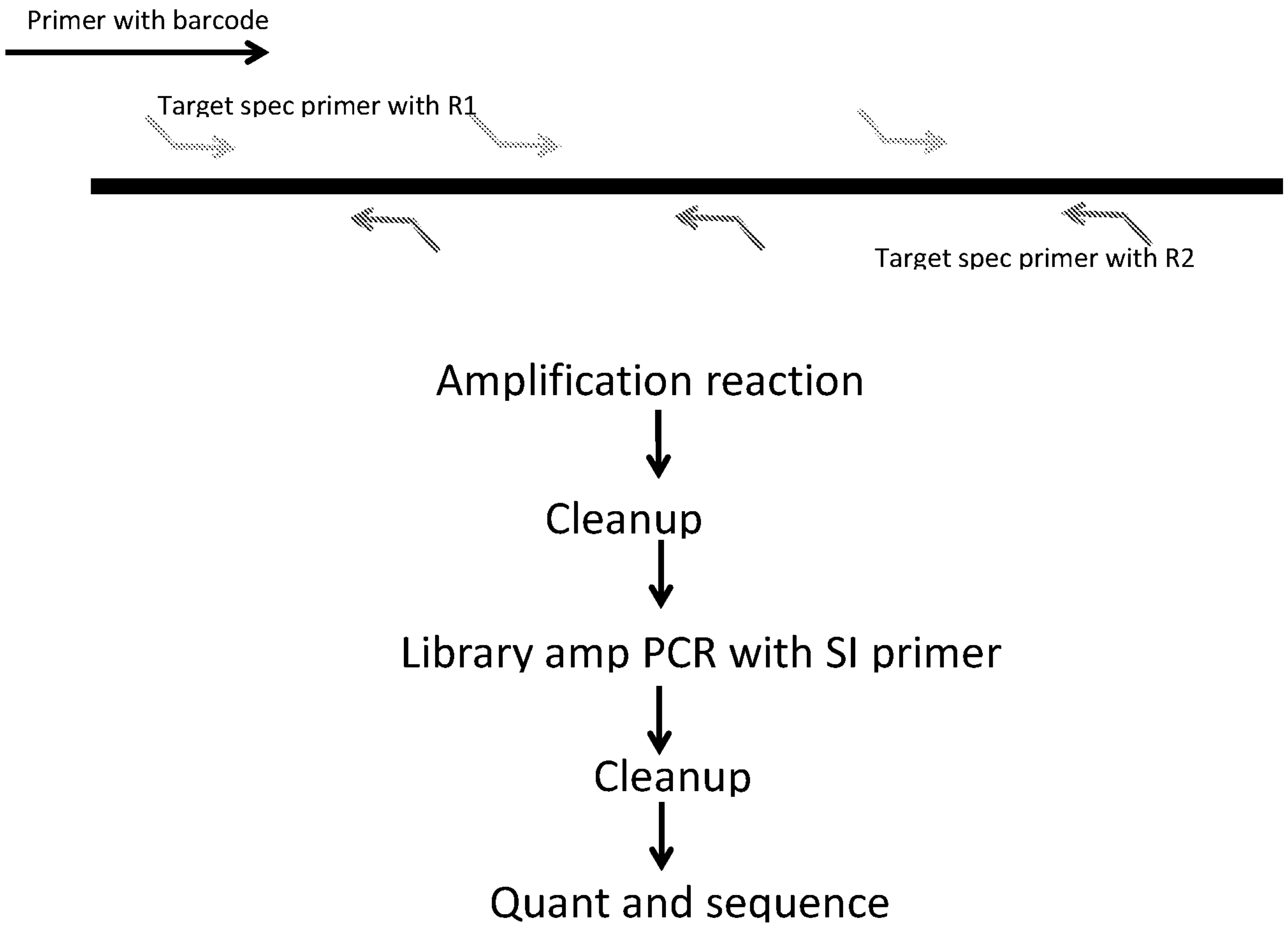


Figure 8



Amplification reaction

Cleanup

Library amp PCR with SI primer

Cleanup

Quant and sequence

Figure 9

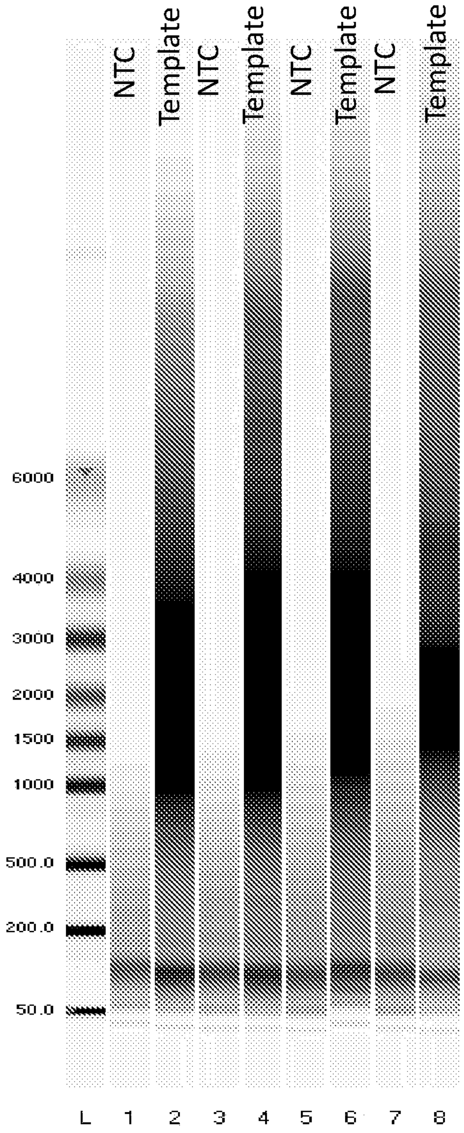
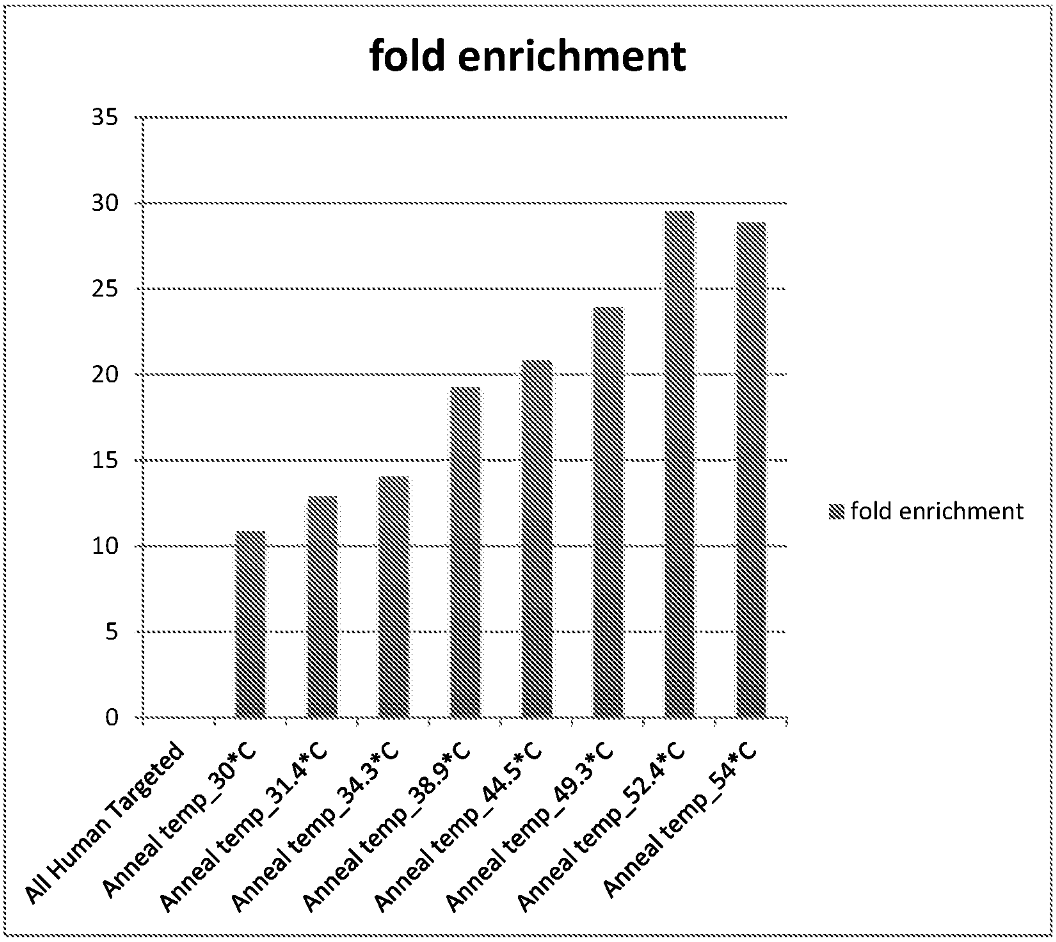

NTC
Template
NTC
Template
NTC
Template
NTC
Template
6000
4000
3000
2000
1500
1000
500.0
200.0
50.0
L
1
2
3
4
5
6
7
8

Figure 10


fold enrichment
35
30
25
20
15
10
5
0
▨ fold enrichment
All Human Targeted
Anneal temp_30*C
Anneal temp_31.4*C
Anneal temp_34.3*C
Anneal temp_38.9*C
Anneal temp_44.5*C
Anneal temp_49.3*C
Anneal temp_52.4*C
Anneal temp_54*C