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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 June 2017 (08.06.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/096158 A1

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

(21) International Application Number:
PCT/US2016/064611

(22) International Filing Date:
2 December 2016 (02.12.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/263,532 4 December 2015 (04.12.2015) US

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

(72) Inventors: **ZHENG, Xinying**; 7068 Koll Center Parkway, Suite 401, Pleasanton, CA 94566 (US). **SAXONOV, Serge**; 7068 Koll Center Parkway, Suite 401, Pleasanton, CA 94566 (US). **SCHNALL-LEVIN, Michael**; 7068 Koll Center Parkway, Suite 401, Pleasanton, CA 94566 (US). **NESS, Kevin**; 7068 Koll Center Parkway, Suite 401, Pleasanton, CA 94566 (US). **BHARADWAJ, Rajiv**; 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, DJ, 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, KW, 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, 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))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

WO 2017/096158 A1

(54) Title: METHODS AND COMPOSITIONS FOR NUCLEIC ACID ANALYSIS

(57) Abstract: The present invention is directed to methods, compositions and systems for analyzing sequence information while retaining structural and molecular context of that sequence information.

METHODS AND COMPOSITIONS FOR NUCLEIC ACID ANALYSIS

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of United States Provisional Application No. 62/263,532, filed December 4, 2015, which is hereby incorporated by reference in its entirety for all purposes.

BACKGROUND OF THE INVENTION

[0002] Polynucleotide sequencing continues to find increasing use in medical applications such as genetic screening and genotyping of tumors. Many polynucleotide sequencing methods rely on sample processing techniques of the original sample, including random fragmentation of polynucleotides. These processing techniques can provide advantages in terms of throughput and efficiency, but the resultant sequence information obtained from these processed samples can lack important contextual information in terms of the location of particular sequences within the broader linear (two-dimensional) sequence of the original nucleic acid molecule that contained those sequences. Structural context within the three dimensional space of the original sample is also lost with many sample processing and sequencing techniques. There is thus a need for sequencing technologies that retain structural and molecular context of the identified nucleic acid sequences.

SUMMARY OF THE INVENTION

[0003] Accordingly, the present invention provides methods, systems and compositions for providing sequence information that retains both molecular and structural context of the originating nucleic acid molecule.

[0004] In some aspects, the present disclosure provides methods of analyzing nucleic acids while maintaining structural context. Such methods include the steps of: (a) providing a sample containing nucleic acids, where the nucleic acids comprise three dimensional structures; (b) separating portions of the sample into discrete partitions such that portions of the nucleic acid three dimensional structures are also separated into the discrete partitions; (c) obtaining sequence information from the nucleic acids, thereby analyzing nucleic acids while maintaining structural context.

[0005] In some embodiments, the sequence information from obtaining step (c) includes identification of nucleic acids that are in spatial proximity to each other.

[0006] In further embodiments, the sequence information from obtaining step (c) includes identification of nucleic acids that are in spatial proximity to each other.

[0007] In still further embodiments, the obtaining step (c) provides information on intrachromosomal and/or interchromosomal interactions between genomic loci.

[0008] In yet further embodiments, the obtaining step (c) provides information on chromosome conformations.

[0009] In further embodiments, prior to separating step (b), at least some of the three dimensional structures are processed to link different portions of the nucleic acids that are in proximity to each other within the three dimensional structures.

[0010] In any embodiments, the nucleic acids are not isolated from the sample prior to the separating step (b).

[0011] In any embodiments, prior to the obtaining step (c), the nucleic acids within the discrete partitions are barcoded to form a plurality of barcoded fragments, where fragments within a given discrete partition each comprise a common barcode, such that the barcodes identify nucleic acids from a given partition.

[0012] In further embodiments, the obtaining step (c) comprises a sequencing reaction selected from the group consisting of: short read-length sequencing reactions and long read-length sequencing reactions.

[0013] In some aspects, the present disclosure provides methods of analyzing nucleic acids while maintaining structural context that include the steps of (a) forming linked nucleic acids within the sample such that spatially adjacent nucleic acid segments are linked; (b) processing the linked nucleic acids to produce a plurality of ligation products, wherein the ligation products contain portions of the spatially adjacent nucleic acid segments; (c) depositing the plurality of ligation products into discrete partitions; (d) barcoding the ligation products within the discrete partitions to form a plurality of barcoded fragments, wherein fragments within a given discrete partition each comprise a common barcode, thereby associating each fragment with the linked nucleic acid from which it is derived; (e) obtaining sequence information from the plurality of barcoded fragments, thereby analyzing nucleic acids from the sample while maintaining structural context.

[0014] In some aspects, the present disclosure provides methods of analyzing nucleic acids while maintaining structural context that include the steps of: (a) forming

linked nucleic acids within the sample such that spatially adjacent nucleic acid segments are linked; (b) depositing the linked nucleic acids into discrete partitions; (c) processing the linked nucleic acids to produce a plurality of ligation products, wherein the ligation products contain portions of the spatially adjacent nucleic acid segments; (d) barcoding the ligation products within the discrete partitions to form a plurality of barcoded fragments, wherein fragments within a given discrete partition each comprise a common barcode, thereby associating each fragment with the linked nucleic acid from which it is derived; (e) obtaining sequence information from the plurality of barcoded fragments, thereby analyzing nucleic acids from the sample while maintaining structural context.

[0015] In some aspects, the present disclosure provides methods of analyzing nucleic acids while maintaining structural context that include the steps of (a) cross-linking nucleic acids within the sample to form cross-linked nucleic acids, wherein the cross-linking forms covalent links between spatially adjacent nucleic acid segments; (b) depositing the cross-linked nucleic acids into discrete partitions; (c) processing the cross-linked nucleic acids to produce a plurality of ligation products, wherein the ligation products contain portions of the spatially adjacent nucleic acid segments; (d) obtaining sequence information from the plurality of ligation products, thereby analyzing nucleic acids from the sample while maintaining structural context.

[0016] In any embodiments, the sample is a formalin-fixed paraffin sample.

[0017] In any embodiments, the discrete partitions comprise beads. In further embodiments, the beads are gel beads.

[0018] In any embodiments, the sample comprises a tumor sample.

[0019] In any embodiments, the sample comprises a mixture of tumor and normal cells.

[0020] In any embodiments, the sample comprises a nuclear matrix.

[0021] In any embodiments, the nucleic acids comprise RNA.

[0022] In any embodiments, the amount of nucleic acids in the sample is less than 5, 10, 15, 20, 25, 30, 35, 40, 45, or 50 ng/ml.

[0023] In some aspects, the present invention provides a method of analyzing nucleic acids while maintaining structural context, in which the method includes the steps of: (a) providing a sample that contains nucleic acids; (b) applying a library of tags to the sample such that different geographical regions of the sample receive different tags or different concentrations of tags; (c) separating portions of the sample into discrete

partitions such that portions of the library of tags and portions of the nucleic acids are also separated into the discrete partitions; (d) obtaining sequence information from the nucleic acids, and (e) identifying tags or concentrations of tags in the discrete partitions, thereby analyzing nucleic acids while maintaining structural context.

BRIEF DESCRIPTION OF THE DRAWINGS

[0024] FIGURE 1 provides a schematic illustration of molecular context and structural context in accordance with the methods described herein.

[0025] FIGURE 2 provides a schematic illustration of a process described herein.

[0026] FIGURE 3 illustrates a typical workflow for performing an assay to detect sequence information, using the methods and compositions disclosed herein.

[0027] 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.

[0028] FIGURE 5 provides a schematic illustration of a process for barcoding and amplification of chromosomal nucleic acid fragments.

[0029] FIGURE 6 provides a schematic illustration of the use of barcoding of nucleic acid fragments in attributing sequence data to their originating source nucleic acid molecule.

[0030] FIGURE 7 provides a schematic illustration of an exemplary sample preparation method.

DETAILED DESCRIPTION OF THE INVENTION

[0031] 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.

[0032] 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.

[0033] 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.

[0034] 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.

[0035] 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.

[0036] As used herein, the term "comprising" is intended to mean that the compositions and methods include the recited elements, but do not exclude 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).

[0037] 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

[0038] This disclosure provides methods, compositions and systems for characterization of genetic material. In general, the methods, compositions and systems described herein provide methods of analyzing components of a sample while retaining information on the structural as well as molecular context of those components as they were originally in the sample. Although much of the discussion herein is in terms of the analysis of nucleic acids, it will be appreciated that the methods and systems discussed herein can be adapted to apply to other components of a sample, including proteins and other molecules.

[0039] Deoxyribonucleic acid (DNA) is a linear molecule, and as such the genome is often described and assessed in terms of linear dimensions. However, chromosomes are not rigid, and the spatial distance between two genomic loci need not always correspond to their distance along the linear sequence of the genome. Regions separated by many megabases can be immediately adjacent in 3-dimensional space. From the standpoint of regulation, understanding long-range interactions between genomic loci may be useful. For example, gene enhancers, silencers, and insulator elements may possibly function across vast genomic distances. The ability to retain

both structural and molecular context of sequence reads provides the ability to understand such long-range interactions.

[0040] By "retaining structural context" as used herein means that multiple sequence reads or multiple portions of sequence reads are attributable to the original three-dimensional relative location of those sequence reads within the sample. In other words, the sequence reads can be associated with a relative location within the sample with respect to neighboring nucleic acids (and in some situations associated proteins) in that sample. This spatial information is available through the methods discussed herein even if those neighboring nucleic acids are not physically located within the linear sequence of a single originating nucleic acid molecule. Referring to the schematic illustration in Figure 1: in a sample (101), sequences (104) and (105) are located within the linear sequence of two different originating nucleic acid molecules ((102) and (103) respectively), but are located in spatial proximity to each other within the sample. The methods and compositions described herein provide the ability to retain that information on the structural context of sequence reads and thus allow reads from sequences (104) and (105) to be attributed to their relative spatial proximity within the original sample on the original nucleic acid molecules (102) and (103) from which those sequence reads are derived.

[0041] The methods and compositions discussed herein also provide sequence information that retains molecular context. "Retaining molecular context" as used herein means that multiple sequence reads or multiple portions of sequence reads may be attributable to a single originating molecule of a nucleic acid. 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.

[0042] In general, the methods described herein include analyzing nucleic acids while maintaining structural and molecular context. Such analyses include methods in which a sample containing nucleic acids is provided, where the nucleic acids contain three

dimensional structures. Portions of the sample are separated into discrete partitions such that portions of the nucleic acid three dimensional structures are also separated into the discrete partitions -- nucleic acid sequences that are in spatial proximity to each other will tend to be separated into the same partition, thus retaining the three-dimensional information of that spatial proximity even when later-obtained sequence reads are from sequences that were not originally on the same individual originating nucleic acid molecule. Referring again to Figure 1: if sample 101, containing nucleic acid molecules 102 and 103 and 106, is separated into discrete partitions such that subsets of the sample are allocated into different discrete partitions, it is more likely that nucleic acid molecules 102 and 103 will be placed in the same partition with each other than with nucleic acid molecule 106, because of the physical distance between nucleic acid molecule 106 and 102 and 103. As such, nucleic acid molecules within the same discrete partitions are those that were in spatial proximity to each other in the original sample. Sequence information obtained from nucleic acids within the discrete partitions thus provides a way to analyze the nucleic acids, for example through nucleic acid sequencing, and attribute those sequence reads back to the structural context of the originating nucleic acid molecules.

[0043] In further examples, the structural context (also referred to herein as "geographical context") may be maintained by using tags (such as barcode oligonucleotides) to encode the geography of the sample. In some situations, this can include injecting a viral library encoding a collection of barcoded sequences (such as mRNA sequences) to a sample. The barcodes travel through the sample by active processes or by diffusion. When the sample is then further processed in accordance with methods described herein and known in the art, barcodes can be correlated with structural positions to identify nucleic acid sequences from the same geographic location within the sample. In examples in which the barcodes are distributed through the sample through active processes, sequences with the same barcode may be geographically connected and/or connected through the same process. As will be appreciated, this system of using tags to encode structural context can be used alone or in combination with methods described herein utilizing discrete partitions to further retain structural and molecular context. In examples in which tags for encoding spatial locations and barcodes for identifying molecules separated into the same discrete partitions are used, the samples are in essence tagged or "double barcoded" where one set of barcodes is used for identifying spatial locations and one set of barcodes is

partition-specific. In such examples, both sets of barcodes can be used to provide information to retain structural and molecular context of sequence reads generated from the sample.

[0044] In some examples, the sequence information obtained from the nucleic acids provides information on intrachromosomal and/or interchromosomal interactions between genomic loci. In further examples, the sequence information includes information on chromosome conformations.

[0045] In further examples, prior to separation into the discrete partitions, the nucleic acids in the sample may be processed to link different regions of their three dimensional structures such that regions of the sequence that are in proximity to each other within those three dimensional structures are attached to each other. As such, the separation of the sample into discrete partitions will separate those linked regions into the same partition, thereby further ensuring that the structural context of any sequence reads from those nucleic acids is retained.

[0046] In some situations, the linking of nucleic acids may be accomplished using any methods known in the art used to cross-link molecules in spatial proximity. Such cross-linking agents may include without limitation alkylating agents, cisplatin, nitrous oxide, psoralens, aldehydes, acrolein, glyoxal, osmium tetroxide, carbodiimide, mercuric chloride, zinc salts, picric acid, potassium dichromate, ethanol, methanol, acetone, acetic acid, and the like. In specific examples, the nucleic acids are linked using protocols designed for analysis of the three dimensional architecture of genomes, such as the “Hi-C” protocol described for example in Dekker et al., "Capturing chromosome conformation" Science 295:1306-1311 (2002) and Berkum et al., J. Vis. Exp. (39), e1869, doi:10.3791/1869 (2010), which are each hereby incorporated by reference in its entirety for all purposes and in particular for all teachings related to linking nucleic acid molecules. Such protocols generally involve producing a library of molecules by crosslinking the sample so that genomic loci that are in close spatial proximity become linked. In further embodiments, the intervening DNA loop between the crosslink is digested away and then the intrasequence regions are reverse crosslinked for addition to the library. The digesting and reverse crosslinking steps may occur prior to a step of partitioning the sample into discrete partitions, or it may occur within the partitions after the separating step.

[0047] In still further examples, the nucleic acids may undergo a tagging or barcoding step that provides a common barcode for all nucleic acids within a partition. As will be

appreciated, this barcoding may occur with or without the nucleic acid linking/cross-linking steps discussed above. The use of the barcoding technique disclosed herein confers the unique capability of providing individual structural and molecular context for genomic regions – i.e., by attributing certain sequence reads to individual sample nucleic acid molecules, and through variant coordinated assembly, to provide a broader or even longer range inferred context, among multiple sample nucleic acid molecules, and/or to a specific chromosome. The term "genomic region" or "region" as used herein, refers to any defined length of a genome and/or chromosome. For example, a genomic region may refer to the association (i.e., for example, an interaction) between more than one chromosomes. A genomic region can also encompass a complete chromosome or a partial chromosome. In addition, a genomic region can include a specific nucleic acid sequence on a chromosome (i.e., for example, an open reading frame and/or a regulatory gene) or an intergenic noncoding region.

[0048] 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.

[0049] In addition to providing the ability to obtain sequence information from entire or select 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 copy number variations, as described 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 characterization of genomic material.

[0050] 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. As shown in Figure 2, the methods described herein may include an optional step 201 in

which sample nucleic acids are processed to link nucleic acids in spatial proximity to each other. With or without that preliminary processing step (201), the methods described herein will in most examples include a step in which sample nucleic acids containing are partitioned (202). Generally, each partition containing nucleic acids from genomic regions of interest will undergo a process that results in fragments containing barcodes (203). Those fragments may then be pooled (204) prior to sequencing (205). The sequence reads from (205) can be attributed to the originating structural and molecular context (206) generally due to the partition-specific barcodes (203). Each partition may in some examples include more than one nucleic acid, and will in some instances contain several hundred nucleic acid molecules. The barcoded fragments of step 203 can be generated using any methods known in the art – in some examples, oligonucleotides are included with 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 USSNs 14/316,383; 14/316,398; 14/316,416; 14/316,431; 14/316,447; and 14/316,463, each of which is herein incorporated by reference in its entirety for all purposes and in particular for all teachings related to barcoding and amplifying oligonucleotides. 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. An advantage of the methods and systems described herein is that attaching a partition- or sample-specific barcode to the copied fragments preserves the original molecular context of the sequenced fragments, allowing them to be attributed to their original partition and thus their originating sample nucleic acid molecule.

[0051] 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. 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.

[0052] In addition to the above workflow, targeted genomic regions may be 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. The population of targeted genomic regions may further be enriched prior to the above-described pull-down methods by using methods to increase coverage of those targeted regions. Such increased coverage may for example be accomplished using targeted amplification methods, including those described for example in USSN 62/119,996, filed on February 24, 2015, which is hereby incorporated by reference for all purposes and in particular for all teachings related to targeted coverage of nucleic acid molecules.

[0053] 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.

[0054] 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. Further information of targeted coverage of the genome for use

in accordance with methods described herein is provided for example in USSN 62/146,834, filed on April 13, 2015, which is hereby incorporated by reference in its entirety for all purposes, and in particular for all teachings related to targeted coverage of a genome.

[0055] In general, the methods and systems described herein provide nucleic acids for analyses such as sequencing. 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. As described above, the sequencing of nucleic acids is typically carried out in a manner that preserves the structural and 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 the spatial location relative to other nucleic acids in the original sample (structural context) and to the location of that sequence read along the linear sequence of a single originating molecule of a nucleic acid (molecular context). 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.

[0056] 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.

[0057] 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.

[0058] 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".

[0059] 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.

[0060] In one aspect, and in conjunction with any of the 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. This later attribution further allows attribution to the original structural context of those sample nucleic acids in the original sample, because nucleic acids that were close to each other within the three dimensions of the original sample will be more likely to be deposited into the same partition. Thus, attribution of sequence reads to the partitions (and the nucleic acids contained within those partitions) not only provides a molecular context as to the linear location along the original nucleic acid molecule from which that sequence read was derived, but also provides a structural context of identifying sequence reads from nucleic acids that were in close spatial proximity to each other in the three dimensional context of the original sample.

[0061] 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 structural and molecular context described above.

[0062] 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.

[0063] In further examples, the oligonucleotides included with the portions of the sample divided into the discrete partitions 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.

[0064] 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.

[0065] 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 structural and 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.

[0066] 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.

[0067] 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

[0068] In one exemplary aspect, the methods and systems described in the disclosure provide for depositing or partitioning samples into discrete partitions, where each partition maintains separation of its own contents from the contents in other partitions. As discussed in further detail herein, the samples may comprise samples derived from patients, such as cell or tissue samples, which can contain nucleic acids and, in certain situations, associated proteins as well. In specific aspects, the samples used in the methods described herein include formalin fixed paraffin embedded (FFPE) cell and tissue samples and the like, as well as any other sample types where the risk of sample degradation is high.

[0069] 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 Provisional U.S. Patent Application No. 61/977,804, filed April 4, 2014, 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 the case of droplets in an emulsion, partitioning of sample materials 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. Provisional Patent Application No. 62/017,580 (Attorney Docket No. 43487-727.101), filed June 26, 2014, the full disclosure of which is hereby incorporated by reference in its entirety.

[0071] In situations involving samples that are subject to degradation and/or contain low concentrations of components of interest, the samples may be further processed either prior to partitioning or within the partitions to further release the nucleic acids and/or any associated proteins for further analysis. For example, nucleic acids contained in FFPE samples are generally extracted using methods known in the art. To isolate longer nucleic acid molecules, such samples may also be processed by addition of organocatalysts to remove formaldehyde adducts (see for example Karmakar et al., (2015), Nature Chemistry, DOI: 10.1038/NCHEM.2307, which is hereby incorporated by reference in its entirety and in particular for all teachings related to treatment and processing of FFPE samples.)

[0072] Once the samples are introduced into their respective partitions the sample nucleic acids within partitions may be subjected to amplification to increase the amount of nucleic acids for subsequent applications (such as sequencing methods described herein and known in the art). In certain embodiments, this amplification is conducted with a library of primers that are directed to different parts of the genomic sequence, such that the resultant amplification products represent sequences from subsections of the original nucleic acid molecules. In embodiments in which select genomic regions are of interest, this amplification may include one or more rounds of selective amplification such that regions of the genome that are of interest for targeted coverage 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 amplification provides at least 1X, 2X, 5X, 10X, 20X, 30X, 40X or 50X coverage of the whole or select 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.

[0073] Simultaneously with or subsequent to the amplification described above, 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 (e.g., barcode sequences). In particularly preferred aspects, the unique identifiers are provided in the form of oligonucleotides that comprise nucleic acid barcode sequences that may be attached to those 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 exemplary 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.

[0074] 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.

[0075] 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, USSNs 14/175,935; 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, as well as sequencing and other characterizations of genomic material.

[0076] 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.

[0077] 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.

[0078] 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.

[0079] Figure 3 illustrates one particular example method for barcoding and subsequently sequencing a sample nucleic acid. 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.

[0080] 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 coverage of the whole genome or of selected regions of the genome. For embodiments in which targeted coverage is desired, the targeted amplification 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. In embodiments in which whole genome amplification is desired, the amplification may be conducted using primer libraries designed to minimize amplification biases and provide a robust level of coverage across the entire genome.

[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 USSNs 14/175,935; 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 6. 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., and the like. 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. In addition, the sequence reads can be attributed to the structural context of the relative position of the nucleic acid from which those reads are derived in relation to other nucleic acid molecules that were in close spatial proximity within the original sample. 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 and in particular for all teachings related to assembly of genomic sequences.

III. Methods and compositions for retaining structural context

[0092] This disclosure provides methods, compositions and systems for characterization of genetic material. In general, the methods, compositions and systems described herein provide methods of analyzing components of a sample while retaining information on the structural as well as molecular context of those

components as they were originally in the sample. In other words, the description herein relation generally to spatial detection of nucleic acids in a sample, including tissue samples that have been or will be fixed using methods known in the art, such as formalin fixed paraffin embedded samples. As will be appreciated, any of the methods described in this section can be combined with any of the methods described above in the sections entitled "Overview" and "Workflow Overview" as well as with the nucleic acid sequencing methods described in subsequent sections of this specification.

[0093] In general, the methods disclosed herein relate to determining and/or analyzing nucleic acids in a sample, including genomes, particularly the global genome, of a sample. The methods described herein provide the ability to quantitatively or qualitatively analyze the distribution, location or expression of nucleic acid sequences (including genomic sequences) in a sample wherein the spatial context within the sample is retained. The methods disclosed herein provide an advantage over conventional methods of geographic encoding of nucleic acids in a sample, because information on structural context is retained in a high throughput processing method without requiring identification of particular molecular targets (such as specific genes or other nucleic acid sequences) prior to processing the sample for sequence reads. In addition, low amounts of nucleic acid are needed, which is particularly advantageous in samples such as FFPE samples in which the input nucleic acids, particularly DNA, are often fragmented or present in low concentrations.

[0094] Although much of the discussion herein is in terms of the analysis of nucleic acids, it will be appreciated that the methods and systems discussed herein can be adapted to apply to other components of a sample, including proteins and other molecules.

[0095] As discussed above, maintaining structural context, also referred to herein as maintaining geographical context and encoding geography, means using methods that allow for obtaining multiple sequence reads or multiple portions of sequence reads that can be attributed to the original three-dimensional relative location of those sequence reads within a sample. In other words, the sequence reads can be associated with a relative location within the sample with respect to neighboring nucleic acids (and in some situations associated proteins) in that sample. This spatial information is available even if those neighboring nucleic acids are not physically located within the linear sequence of a single originating nucleic acid molecule.

[0096] In general, the methods described herein include analyses in which a sample containing nucleic acids is provided, where the nucleic acids contain three dimensional structures. Portions of the sample are separated into discrete partitions such that portions of the nucleic acid three dimensional structures are also separated into the discrete partitions -- nucleic acid sequences that are in spatial proximity to each other will tend to be separated into the same partition, thus retaining the three-dimensional information of that spatial proximity even when later-obtained sequence reads are from sequences that were not originally on the same individual originating nucleic acid molecule. Referring to Figure 1: if sample 101, containing nucleic acid molecules 102 and 103 and 106, is separated into discrete partitions such that subsets of the sample are allocated into different discrete partitions, it is more likely that nucleic acid molecules 102 and 103 will be placed in the same partition with each other than with nucleic acid molecule 106, because of the physical distance between nucleic acid molecule 106 and 102 and 103. As such, nucleic acid molecules within the same discrete partitions are those that were in spatial proximity to each other in the original sample. Sequence information obtained from nucleic acids within the discrete partitions thus provides a way to analyze the nucleic acids, for example through nucleic acid sequencing, and attribute those sequence reads back to the structural context of the originating nucleic acid molecules.

[0097] In some examples, a library of tags is applied to the sample for spatial or geographic encoding of the sample. In certain embodiments, the tags are oligonucleotide tags (which can include “oligonucleotide barcodes” and “DNA barcodes”), but as will be appreciated, any type of tag that is capable of being added into a sample can be used, including without limitation particles, beads, dyes, molecular inversion probes (MIPs), and the like. The library of tags can be applied to the sample through simple diffusion, or through active processes, such as cellular processes within tissue culture or cell culture samples. Cellular transport processes include without limitation osmosis, facilitated diffusion through the involvement of cell transport proteins, passive transport, and active transport through the involvement of cell transport proteins and input of energy from molecules such as ATP. In general, the tags are applied such that different spatial/geographic locations within the sample receive different tags and/or a different concentration of tags. Any further processing of the sample and analysis of the nucleic acids within the sample can be attributed to a particular spatial context through identification of the tags. For example, referring to

Figure 1, addition of a library of tags to sample 101 would result in nucleic acids 102 and 103 having spatial proximity to a different portion or concentration of the library of tags than nucleic acid 106. Any further processing of the sample in accordance to the workflows described herein would then result in nucleic acids 102 and 103 being associated with the same portion/concentration of tags, and thus identification of those tags would indicate that nucleic acids 102 and 103 were in spatial proximity to each other in the original sample 101. Identification of nucleic acid 106 with a different portion/concentration of tags would show that nucleic acid 106 was at a different spatial location than nucleic acids 102 and 103 in the original sample.

[0098] In further examples, partition-specific barcodes are also employed, such that any sequence reads obtained can be attributed back to the partition in which the originating nucleic acid molecules were located. As discussed above, associating sequence reads to a particular partition identifies nucleic acid molecules that were in spatial proximity to each other in the geography of the original sample. Further use of workflows, such as those pictured in Figure 2, also provides information on the molecular context of the sequence reads, such that individual sequence reads can be attributed to the individual nucleic acid molecules from which they originated.

[0099] To enable tagging of samples, the samples may be processed using any methods known in the art to allow application of exogenous molecules such as oligonucleotide tags or other labels. For example, in embodiments in which FFPE samples are used, tags can be applied to the samples by heating the sample to allow embedding of the tags into the sample, and then the sample could be cooled and further processed in accordance with any of the methods described herein, including division into discrete partitions and further analysis to identify sequences of nucleic acids in the sample and the tags that are also in close spatial proximity to those sequence reads, thus retaining structural context of those sequence reads. Other sample processing methods include tissue processing methods that remove extracellular matrix and/or other structural impediments while retaining molecular and protein elements. Such methods include in some non-limiting examples the CLARITY method as well as the use of other tissue clearing and labeling methods, including those described for example in Tomer et al., VOL.9 NO.7 , 2014, Nature Protocols; Kebschull et al., Neuron, Volume 91, Issue 5, 7 September 2016, Pages 975–987; Chung, K. *et al.* Structural and molecular interrogation of intact biological systems. *Nature* 497, 332–337 (2013); Susaki, E.A. *et al.* Whole-brain imaging with single-cell

resolution using chemical cocktails and computational analysis. *Cell* 157, 726–739 (2014); and Lee et al., ACT-PRESTO: Rapid and consistent tissue clearing and labeling method for 3-dimensional (3D) imaging, Scientific Reports, 2016/01/11/online; Vol. 6, p.18631, each of which is hereby incorporated by reference in its entirety for all purposes, and in particular for any teachings related to processing samples for use in structural and molecular interrogation methods.

[0100] In certain embodiments, the methods described herein are used in combination with imaging techniques to identify spatial locations of the tags within the sample, particularly for samples that are immobilized on slides, such as FFPE samples. Such imaging techniques may allow correlation of sequence reads to particular locations on the slides, which allows correlation with other pathological/imaging studies that may have been conducted with those samples. For example, imaging techniques may be used to provide a preliminary identification of a pathology. The sequencing techniques described herein that further provide sequence reads while maintaining structural context could be combined with such imaging analysis to correlate sequence reads with structural context to corroborate or provide further information on that preliminary identification of the pathology. In addition, the imaging techniques may be used in combination with tags with optical properties, such that particular tags are associated with particular regions of the imaged sample. Sequence reads that are correlated with those identified tags could then be further correlated with regions of the imaged sample by virtue of their location with those tags. However, it will be appreciated that the methods described herein are independent of any such imaging techniques, and the ability to retain structural context is not dependent on using an imaging technique for determining spatial information of nucleic acids in the sample.

[0101] In one exemplary aspect, gradients of oligonucleotides are generated in a sample to provide a coordinate system that can be decoded through later processing through sequencing. Such a gradient will allow tagging of cells and/or nucleic acids in the sample with an oligonucleotide or oligonucleotide concentration, which can be mapped to a physical location within the original sample. This coordinate system can be developed by allowing a library of oligonucleotides to diffuse into a sample and/or by injecting oligonucleotides into particular regions of the sample. When using diffusion, standard calculations of diffusion kinetics will provide a correlation between the concentration of the oligonucleotide tags and its spatial location in the original sample.

Thus, any other nucleic acids identified with that concentration of oligonucleotide tags can in turn be correlated to a particular geographic region of the sample.

[0102] In further exemplary embodiments, the methods include processes for analyzing nucleic acids while maintaining structural context in which a library of tags is applied to a sample such that different geographical regions of the sample receive different tags. Portions of the sample, which now contain their original nucleic acids as well as the added tags, are then separated into discrete partitions, such that portions of the library of tags and portions of the nucleic acids that are close to each other in geographic location within the sample end up in the same discrete partition. Sequencing processes, such as those described in detail herein, are used to provide sequence reads of nucleic acids in the discrete partitions. The tags can also be identified before, after or simultaneously with those sequencing processes. The correlation of sequence reads to particular tags (or concentrations of tags in embodiments in which concentration gradients of tags are used) thereby helps to provide the spatial context of the sequence reads. As discussed above, embodiments in which the tags used for spatial encoding are used in conjunction with partition-specific barcoding further provide structural and molecular context for the sequence reads.

IV. Applications of methods and systems to nucleic acid sequencing

[0103] 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.

[0104] In general, the methods and systems described herein accomplish genomic sequencing 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.

[0105] 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%.

[0106] 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.

[0107] In some embodiments, the methods and systems described herein for obtaining sequence information while retaining both structural and molecular context are used for whole genome sequencing. In some embodiments, the methods described herein are used for sequencing of targeted regions of the genome. In further embodiments, the 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 further 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 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 can

contain sequences from both the genomic regions for de novo sequencing and the genomic regions for re-sequencing.

[0108] In specific instances, methods described herein include a step in which the whole or selected regions of the genome are 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 whole or selected regions of the genome. In further embodiments, the amplification provides at least 1X-30X, 2X-25X, 3X-20X, 4X-15X, or 5X-10X coverage of the whole or selected regions of the genome.

[0109] Amplification for coverage of the whole genome and/or select targeted regions of the genome 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 genomic regions of interest – in other words, the library of primer is designed to amplify regions at specific distances along the genome, whether this is across selected regions or across the whole 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.

[0110] 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, a selected region of the genome is selectively amplified in accordance with any of the methods described herein, and that selected region 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 long. 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. 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.

[0111] 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.

[0112] 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 structural and 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 the relative spatial location within the original sample with respect to other nucleic acids (structural context) and/or to the location within the linear sequence of a single originating molecule of a nucleic acid (molecular context).

[0113] 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.

[0114] 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.

[0115] As shown in Figure 2, the methods described herein will in most examples include a step in which samples are partitioned (202). Prior to that partitioning step, there may be an optional step (201) in which nucleic acids in the sample are linked to attach sequence regions that are in close spatial proximity to each other. Generally, each partition containing nucleic acids from genomic regions of interest will undergo some kind of fragmentation process and the original molecular context of the fragments will generally be retained (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. As discussed above, 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. 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.

[0116] Returning to the method exemplified in Figure 2, once the partition-specific barcodes are attached to the copied fragments, the barcoded fragments can optionally then be 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.

[0117] 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 62/072,164, filed on October 29, 2014, 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.

[0118] In some examples, rather than whole genome sequencing, it is desirable to focus on selected regions of the genome. The methods described herein are particularly amenable to such analyses, because the ability to target subsets of the genome, even when those subsets are at large linear distances but potentially in near proximity in the three-dimensional context of the original sample, is an advantageous feature of these methods. 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.

[0119] In general, 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.

[0120] 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. As will be appreciated, step (a) of such a method can include more than one individual nucleic acid molecule.

[0121] 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 structural and molecular context of the sequences of the fragments, such that the identifying further comprises identifying fragments derived from nucleic acids located in close physical proximity within the original sample and/or are located on 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.

[0122] 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.

[0123] 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 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 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.

[0124] In one aspect, and in conjunction with any of the 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.

[0125] 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.

[0126] 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.

[0127] 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. 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.

[0128] 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 for sequences that were located in spatial proximity to each other within the three dimensional space of the original sample and the original nucleic acid molecules from which those sequences were derived.

[0129] The use of the barcoding technique disclosed herein confers the unique capability of providing individual structural and molecular context for sequences and regions of the genome. Such regions of the genome may include 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.

[0130] As noted above, the methods and systems described herein provide individual structural and molecular context for short sequence reads of longer nucleic acids. As used herein, structural context refers to the location of sequences within the three dimensional space of their originating nucleic acid molecules within the original sample.

As discussed above, although the genome is often thought of as linear, chromosomes are not rigid, and the spatial distance between two genomic loci does not necessarily correlate to their distance along the genome – genomic regions separated by several megabases along the linear sequence may be immediately proximal to each other in three- dimensional space. By retaining the information of the original spatial proximity of sequence reads, the methods and compositions described herein provide a way to attribute sequence reads to long-range genomic interactions.

[0131] Similarly, the retention of individual molecular context possible with the methods described herein provides 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.

[0132] The methods, compositions, and systems described herein allow for characterization of long-range interactions across the genome as well as characterization of associated proteins and other molecules within a sample. Like the higher-level organization of proteins, the bending and folding of DNA and chromatin create functionally significant structures at a wide variety of scales. At small scales, it is well known that DNA is often wound around proteins such as histones to create a structure known as the nucleosome. These nucleosomes pack into larger `chromatin fibers`, and the packing pattern has been implicated as being affected by cellular processes such as transcription. Functional structures also exist at larger scales: regions separated by many megabases long the linear sequence of the genome can be immediately adjacent in 3-dimensional space. Such long-range interactions between genomic loci may play a role in functional characteristics: for example, gene enhancer, silencer and insulator elements may all function across vast genomic distances and their primary mode of action could involve a direct physical association with target genes, noncoding RNAs and/or regulatory elements. Long-range interactions are not limited to elements located in cis, i.e. along the same chromosome, but can also occur between genomic loci located in trans, i.e. on different chromosomes. The existence of long-range interactions can complicate efforts to understand the pathways that regulate cellular processes, because the interacting regulatory elements could lie at a great genomic distance from a target gene, even on another chromosome. In the case of oncogenes and other disease-associated genes, identification of long-range genetic regulators can be of great use in identifying the genomic variants responsible for the disease state and the process by which the disease state is brought about. Thus, the ability to retain structural and molecular context in accordance with the methods described herein provides a way to identify long-range genomic interactions and characterize any associated proteins as well.

[0133] The methods described herein are particularly useful for characterization of nucleic acids from an FFPE tissue sample, including a historic FFPE tissue sample. FFPE samples generally present challenges to nucleic acid characterization, because the nucleic acids are often fragmented or otherwise degraded, which can limit the amount of information that can be obtained using conventional methods. The structural and molecular context information that is retained in the methods described herein provides a unique opportunity with such samples, because that contextual information can provide characterizations of long range genomic interactions even for degraded

samples, because that long-range information is accessible through short read sequencing technologies. Applications of FFPE nucleic acid characterizations include comparisons of sequences from one or more historic samples to sequences from a sample from a subject, e.g., a cancer patient to provide diagnostic or prognostic information. For example, the status of one or more molecular markers in a historic sample can be correlated with one or more treatment outcomes, and the correlation of a treatment outcome with molecular marker status in one or more historic samples can be used to predict treatment outcomes for the subject, e.g., a cancer patient. These predictions can be the basis for determining whether or not to recommend a drug treatment option to the subject.

V. Samples

[0134] As will be appreciated, the methods and systems discussed herein can be used to obtain 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.

[0135] In particular aspects, samples of use in the present invention include formalin fixed paraffin embedded (FFPE) cell and tissue samples and the like, including any other sample types where the risk of sample degradation is high. Other types of fixed samples include without limitation samples that were fixed using: acrolein, glyoxal, osmium tetroxide, carbodiimide, mercuric chloride, zinc salts, picric acid, potassium dichromate, ethanol, methanol, acetone, and/or acetic acid.

[0136] In further embodiments, the samples of use in the methods and systems described herein comprise nuclear matrix. "Nuclear matrix" refers to any composition comprising nucleic acids and protein. The nucleic acids may be organized into

chromosomes, wherein the proteins (i.e., for example, histones) may become associated with the chromosomes having a regulatory function.

[0137] The methods and systems provided herein are particularly useful for nucleic acid sequencing applications in which the starting nucleic acids (e.g., DNA, mRNA, etc.) – or starting target nucleic acids – are present in small quantities, or where nucleic acids that are targeted for analysis, are present at a relatively low proportion of the total nucleic acids within a sample. In one aspect, the present disclosure provides a method of analyzing nucleic acids where the input nucleic acid molecules are present at an amount of less than 50 nanograms (ng). In further embodiments, the nucleic acid molecules are at an input amount of less than less than 40 ng. In some embodiments, the amount is less than 20 ng. In some embodiments, the amount is less than 10 ng. In some embodiments, the amount is less than 5 ng. In some embodiments, the amount is less than 1 ng. In some embodiments, the amount is less than 0.1 ng. Methods for isolating and analyzing nucleic acids where the starting input amount is a small quantity are further described for example in USSN 14/752,602, filed on June 26, 2015, which is hereby incorporated by reference in its entirety for all purposes and in particular for all teachings related to isolation and characterization of nucleic acids derived from samples in which the nucleic acids are present in small quantities.

[0138] As will be appreciated, samples can be processed using methods known in the art at any point during the methods described herein. For example, samples can be processed prior to partitioning or after the sample has been partitioned into discrete partitions.

[0139] In certain embodiments, the samples are processed to ensure that longer nucleic acid strands are retained. In embodiments in which FFPE samples are used, such samples may be subjected to processing to remove formaldehyde adducts to improve nucleic acid yields. Such processing methods may include in one non-limiting example the use of water-soluble organocatalysts to speed the reversal of formaldehyde adducts from RNA and DNA bases, as described in Karmakar et al., (2015), Nature Chemistry, DOI: 10.1038/NCHEM.2307, which is hereby incorporated by reference in its entirety and in particular for all teachings related to treatment and processing of FFPE samples.

[0140] 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., USSN 14/752,641, filed June 26, 2015, the full disclosure of which is hereby incorporated by reference in its entirety.

[0141] 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.

[0142] 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.

VI. Embodiments

[0143] In some aspects, the present disclosure provides methods of analyzing nucleic acids while maintaining structural context. Such methods include the steps of: (a) providing a sample containing nucleic acids, where the nucleic acids comprise three dimensional structures; (b) separating portions of the sample into discrete partitions such that portions of the nucleic acid three dimensional structures are also separated into the discrete partitions; (c) obtaining sequence information from the nucleic acids, thereby analyzing nucleic acids while maintaining structural context.

[0144] In some embodiments, the sequence information from obtaining step (c) includes identification of nucleic acids that are in spatial proximity to each other.

[0145] In any embodiments, the obtaining step (c) provides information on intrachromosomal and/or interchromosomal interactions between genomic loci.

[0146] In any embodiments, the obtaining step (c) provides information on chromosome conformations.

[0147] In any embodiments, prior to separating step (b), at least some of the three dimensional structures are processed to link different portions of the nucleic acids that are in proximity to each other within the three dimensional structures.

[0148] In any embodiments, the sample is a formalin-fixed paraffin sample.

[0149] In any embodiments, the nucleic acids are not isolated from the sample prior to the separating step (b).

[0150] In any embodiments, the discrete partitions comprise beads.

[0151] In any embodiments, the beads are gel beads.

[0152] In any embodiments, prior to the obtaining step (c), the nucleic acids within the discrete partitions are barcoded to form a plurality of barcoded fragments, where fragments within a given discrete partition each comprise a common barcode, such that the barcodes identify nucleic acids from a given partition.

[0153] In any embodiments, the obtaining step (c) comprises a sequencing reaction selected from the group consisting of: short read-length sequencing reactions and long read-length sequencing reactions.

[0154] In any embodiments, the sample comprises a tumor sample.

[0155] In any embodiments, the sample comprises a mixture of tumor and normal cells.

[0156] In any embodiments, the sample comprises a nuclear matrix.

[0157] In any embodiments, the nucleic acids comprise RNA.

[0158] In any embodiments, the amount of nucleic acids in the sample is less than 5, 10, 15, 20, 25, 30, 35, 40, 45, or 50 ng/ml.

[0159] In some aspects, the present disclosure provides methods of analyzing nucleic acids while maintaining structural context that include the steps of (a) forming linked nucleic acids within the sample such that spatially adjacent nucleic acid segments are linked; (b) processing the linked nucleic acids to produce a plurality of ligation products, wherein the ligation products contain portions of the spatially adjacent nucleic acid segments; (c) depositing the plurality of ligation products into discrete partitions; (d) barcoding the ligation products within the discrete partitions to form a plurality of barcoded fragments, wherein fragments within a given discrete partition each comprise a common barcode, thereby associating each fragment with the linked nucleic acid from which it is derived; (e) obtaining sequence information from the plurality of barcoded fragments, thereby analyzing nucleic acids from the sample while maintaining structural context.

[0160] In further embodiments, the processing step (b) includes blunt-end ligation under conditions favoring intramolecular ligation, such that the spatially adjacent nucleic acid segments are ligated within the same molecule.

[0161] In any embodiments, the conditions favoring intramolecular ligation comprise diluting the sample to reduce concentration of the nucleic acids under 10 ng/μL.

[0162] In any embodiments, the nucleic acids are not isolated from the sample prior to the step (a).

[0163] In any embodiments, prior to step forming (a), the nucleic acids are immunoprecipitated such that associated DNA binding proteins remain bound to the nucleic acids.

[0164] In any embodiments, the partitions comprise beads.

[0165] In any embodiments, the beads are gel beads.

[0166] In any embodiments, the sample comprises a tumor sample.

[0167] In any embodiments, the sample comprises a mixture of tumor and normal cells.

[0168] In any embodiments, the processing step includes reversal of the linking subsequent to forming the ligation products.

[0169] In any embodiments, the obtaining step (e) provides information on intrachromosomal and/or interchromosomal interactions between genomic loci.

[0170] In any embodiments, the obtaining step (e) provides information on chromosome conformations.

[0171] In any embodiments, the chromosome conformations are associated with disease states.

[0172] In any embodiments, the processing step results in ligation products comprising nucleic acids that were originally in close spatial proximity in the sample.

[0173] In any embodiments, 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.

[0174] In any embodiments, the sequencing reaction is a short read, high accuracy sequencing reaction.

[0175] In any embodiments, the forming step (a) includes cross-linking nucleic acids in the sample.

[0176] In any embodiments, the forming step (a) results in covalent links between spatially adjacent nucleic acid segments.

[0177] In some aspects, the present disclosure provides methods of analyzing nucleic acids while maintaining structural context that include the steps of: (a) forming linked nucleic acids within the sample such that spatially adjacent nucleic acid

segments are linked; (b) depositing the linked nucleic acids into discrete partitions; (c) processing the linked nucleic acids to produce a plurality of ligation products, wherein the ligation products contain portions of the spatially adjacent nucleic acid segments; (d) barcoding the ligation products within the discrete partitions to form a plurality of barcoded fragments, wherein fragments within a given discrete partition each comprise a common barcode, thereby associating each fragment with the linked nucleic acid from which it is derived; (e) obtaining sequence information from the plurality of barcoded fragments, thereby analyzing nucleic acids from the sample while maintaining structural context.

[0178] In further embodiments, the processing step (c) includes blunt-end ligation under conditions favoring intramolecular ligation, such that the spatially adjacent nucleic acid segments are ligated within the same molecule.

[0179] In any embodiments, the sample is a formalin-fixed paraffin sample.

[0180] In any embodiments, the sample comprises a nuclear matrix.

[0181] In any embodiments, the nucleic acids comprise RNA.

[0182] In any embodiments, the nucleic acids are not isolated from the sample prior to step (a).

[0183] In any embodiments, prior to the forming step (a), the nucleic acids are immunoprecipitated such that associated DNA binding proteins remain bound to the nucleic acids.

[0184] In any embodiments, the partitions comprise beads.

[0185] In any embodiments, the beads are gel beads.

[0186] In any embodiments, the sample comprises a tumor sample.

[0187] In any embodiments, the sample comprises a mixture of tumor and normal cells.

[0188] In any embodiments, the processing step (c) results in ligation products comprising nucleic acids that were originally in close spatial proximity in the sample.

[0189] In any embodiments, the obtaining step (e) provides information on intrachromosomal and/or interchromosomal interactions between genomic loci.

[0190] In any embodiments, 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.

[0191] In any embodiments, the sequencing reaction is a short read, high accuracy sequencing reaction.

[0192] In some aspects, the present disclosure provides methods of analyzing nucleic acids while maintaining structural context that include the steps of (a) cross-linking nucleic acids within the sample to form cross-linked nucleic acids, wherein the cross-linking forms covalent links between spatially adjacent nucleic acid segments; (b) depositing the cross-linked nucleic acids into discrete partitions; (c) processing the cross-linked nucleic acids to produce a plurality of ligation products, wherein the ligation products contain portions of the spatially adjacent nucleic acid segments; (d) obtaining sequence information from the plurality of ligation products, thereby analyzing nucleic acids from the sample while maintaining structural context.

[0193] In further embodiments the processing step (b) includes blunt-end ligation under conditions favoring intramolecular ligation, such that the spatially adjacent nucleic acid segments are ligated within the same molecule.

[0194] In any embodiments, the sample is a formalin-fixed paraffin sample.

[0195] In any embodiments, the sample comprises a nuclear matrix.

[0196] In any embodiments, the nucleic acids comprise RNA.

[0197] In any embodiments, the nucleic acids are not isolated from the sample prior to the cross-linking step (a).

[0198] In any embodiments, the amount of nucleic acids in the sample is less than 5, 10, 15, 20, 25, 30, 35, 40, 45, or 50 ng/ml.

[0199] In any embodiments, prior to the cross-linking step (a), the nucleic acids are immunoprecipitated such that associated DNA binding proteins remain bound to the nucleic acids.

[0200] In any embodiments, prior to the obtaining step (d), the ligation products are associated with a barcode.

[0201] In any embodiments, ligation products within the same partition receive common barcodes, such that the barcodes identify ligation products from a given partition.

[0202] In any embodiments, the obtaining step (d) comprises a sequencing reaction selected from the group consisting of: short read-length sequencing reactions and long read-length sequencing reactions.

[0203] 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 1: Sample preparation

[0204] Sample preparation methods were modified to provide long DNA molecules from FFPE samples. Figure 7 illustrates an exemplary workflow, with modifications indicated for preparing FFPE samples for both whole genome sequencing (WGS) and whole exome sequencing (WES). For example, after DNA extraction, a standard thermalcycling protocol was modified at 701 to move the 98 degree denaturation step from the end of each cycle to the beginning. In addition, a 70 degree hold was added for 2 minutes at the end of each cycle.

[0205] During the post cycling cleanup 702 and the WES library preparation and target enrichments steps 704 and 705, 1.8X Solid Phase Reversible Immobilisation (SPRI) beads over normal protocols were used.

[0206] Another modification included changing conditions during the shearing step 703, in which an ultrasonicator with a peak incident power of about 450 was used, as opposed to a standard sonicator with a peak incident power of 50.

[0207] An additional modification that may be used in certain situations is to first process the FFPE sample with organocatalysts in order to remove formaldehyde adducts, as for example described in Karmakar et al., (2015), Nature Chemistry, DOI: 10.1038/NCHEM.2307. Such protocols include adding 5 mM organocatalysts in 30 mM pH 7 Tris buffer to the samples to effect adduct reversal. Effective organocatalysts include without limitation water-soluble bifunctional catalysts, such as the anthranilate and phosphanilate catalysts described in Karmakar et al. Reversal of the adducts has the effect of improving the yield of nucleic acid yields from the sample.

Example 2: Barcoding of FFPE Samples

[0208] FFPE samples (which can include FFPE samples on a slide) can be tagged with DNA barcodes applied in spatially well-defined pattern, such as those used in DNA microarray printing. The DNA barcode (henceforth called barcode-1) is either long so that it will not diffuse out in subsequent steps or is covalently applied to the FFPE

sample. To enable barcoding DNA to get embedded into FFPE slide, the sample is heated, and then the barcodes are added. The barcodes are generally a library of barcodes such that different barcodes are provided in different parts of the slide. The barcodes may also be added in different concentrations in different parts of the slide to assist in the geographic encoding – in that situation, the library of barcodes may comprise identical or different barcodes. After the barcodes are added, the slide is then cooled and then separated into portions generally through cutting in ways such as using laser microdissection, mechanical/acoustic means, and the like. Fluorophores or Qdots may also be used instead of barcodes, however, barcoding enables massively parallel random encapsulation of sample portions while retaining local spatial information (e.g., tumor vs normal cells).

[0209] The portions of samples containing the barcodes can then be put in a sequencing system, including a droplet based system such as the 10X Genomics Chromium™ system, such that a single barcoded portion is encapsulated per droplet.

[0210] Deparaffinization of the sample can be carried out in the droplet by heating. Paraffin is immiscible in water but soluble in certain oils and thus the paraffin can be easily removed from the droplet upon heating the droplets on-chip. Xylene could also be used in a liquid-liquid extraction process to de-paraffinize the sample portions and ready their nucleic acid contents for further processing.

[0211] Further steps include de-cross-linking methylene bridges of the deparaffinized sample. For this step, specialized chemical means can be used to remove the crosslinks and thereby enable access to the contained nucleic acids for any subsequent processing, including the nucleic acid barcoding, amplification, and library preparation steps discussed herein (see for example Figure 2). Note that the spatial barcoding DNA is also encapsulated in the droplet. The second barcoding step of the individual nucleic acids will serve to barcode the nucleic acids and the barcode used to spatially encode the sample. Sequence reads can then be stitched together to provide information that can then be compared to the original spatial location in the sample and hence related to pathological data.

[0212] In alternative versions of this spatial encoding workflow, the de-cross-linking step is first performed within the droplet and then the nucleic acids in the sample, including genomic DNA as well as the spatial encoding barcodes, are attached to particles or are otherwise isolated from the sample. The nucleic acids are then re-

encapsulated and subjected to the workflow of barcoding and sequencing in methods described herein, including that pictured in Figure 2.

[0213] 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.

CLAIMS

What is claimed:

1. A method of analyzing nucleic acids while maintaining structural context, the method comprising:
   (a) providing a sample comprising nucleic acids, wherein the nucleic acids comprise three dimensional structures;
   (b) separating portions of the sample into discrete partitions such that portions of the nucleic acid three dimensional structures are also separated into the discrete partitions;
   (c) obtaining sequence information from the nucleic acids, thereby analyzing nucleic acids while maintaining structural context.

2. The method of claim 1, wherein the sequence information from obtaining step (c) includes identification of nucleic acids that were in spatial proximity to each other in the sample.

3. The method of claims 1-2, wherein the obtaining step (c) provides information on intrachromosomal and/or interchromosomal interactions between genomic loci.

4. The method of claims 1-2, wherein the obtaining step (c) provides information on chromosome conformations.

5. The method of claims 1-4, wherein prior to separating step (b), at least some of the three dimensional structures are processed to link different portions of the nucleic acids that are in proximity to each other within the three dimensional structures.

6. The method of claims 1-5, wherein the sample is a formalin-fixed paraffin sample.

7. The method of claims 1-6, wherein the nucleic acids are not isolated from the sample prior to the separating step (b).

8. The method of claims 1-7, wherein the discrete partitions comprise beads.

9. The method of claim 8, wherein the beads are gel beads.

10. The method of claims 1-9, wherein prior to the obtaining step (c), the nucleic acids within the discrete partitions are barcoded to form a plurality of barcoded fragments, wherein fragments within a given discrete partition each comprise a common barcode, such that the barcodes identify nucleic acids from a given partition.

11. The method of claims 1-10, wherein the obtaining step (c) comprises a sequencing reaction selected from the group consisting of: short read-length sequencing reactions and long read-length sequencing reactions.

12. The method of claims 1-11, wherein the sample comprises a tumor sample.

13. The method of claims 1-11, wherein the sample comprises a mixture of tumor and normal cells.

14. The method of claims 1-13, wherein the sample comprises a nuclear matrix.

15. The method of claims 1-14, wherein the nucleic acids comprise RNA.

16. A method of analyzing nucleic acids from a sample while maintaining structural context of nucleic acids within the sample, the method comprising:
    (a) forming linked nucleic acids within the sample such that spatially adjacent nucleic acid segments are linked;
    (b) processing the linked nucleic acids to produce a plurality of ligation products, wherein the ligation products contain portions of the spatially adjacent nucleic acid segments;
    (c) depositing the plurality of ligation products into discrete partitions;
    (d) barcoding the ligation products within the discrete partitions to form a plurality of barcoded fragments, wherein fragments within a given discrete partition each comprise a common barcode, thereby associating each fragment with the linked nucleic acid from which it is derived;
    (e) obtaining sequence information from the plurality of barcoded fragments, thereby analyzing nucleic acids from the sample while maintaining structural context.

17. The method of claim 16, wherein the processing step (b) includes blunt-end ligation under conditions favoring intramolecular ligation, such that the spatially adjacent nucleic acid segments are ligated within the same molecule.

18. The method of claim 17, wherein the conditions favoring intramolecular ligation comprise diluting the sample to reduce concentration of the nucleic acids under 10 ng/μL.

19. The method of claims 16-18, wherein the sample is a formalin-fixed paraffin sample.

20. The method of claims 16-18, wherein the sample comprises a nuclear matrix.

21. The method of claims 16-20, wherein the nucleic acids comprise RNA.

22. The method of claims 16-21, wherein the nucleic acids are not isolated from the sample prior to the step (a).

23. The method of claims 16-21, wherein prior to step (a), the nucleic acids are immunoprecipitated such that associated DNA binding proteins remain bound to the nucleic acids.

24. The method of claims 16-23, wherein the amount of nucleic acids in the sample is less than 50 ng/ml.

25. The method of claims 16-24, wherein the partitions comprise beads.

26. The method of claim 25, wherein the beads are gel beads.

27. The method of claims 16-26, wherein the sample comprises a tumor sample.

28. The method of claims 16-27, wherein the sample comprises a mixture of tumor and normal cells.

29. The method of claims 16-28, wherein the processing step includes reversal of the linking subsequent to forming the ligation products.

30. The method of claims 16-29, wherein the obtaining step (e) provides information on intrachromosomal and/or interchromosomal interactions between genomic loci.

31. The method of claims 16-29, wherein the obtaining step (e) provides information on chromosome conformations.

32. The method of claim 31, wherein the chromosome conformations are associated with disease states.

33. The method of claims 16-32, wherein the processing step results in ligation products comprising nucleic acids that were originally in close spatial proximity in the sample.

34. The method of claims 16-33, 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.

35. The method of claim 34, wherein the sequencing reaction is a short read, high accuracy sequencing reaction.

36. The method of claims 16-35, wherein the forming step (a) comprises cross-linking nucleic acids in the sample.

37. The method of claims 16-36, wherein the forming step (a) results in covalent links between spatially adjacent nucleic acid segments.

38. A method of analyzing nucleic acids from a sample while maintaining structural context of nucleic acids within the sample, the method comprising:
    (a) forming linked nucleic acids within the sample such that spatially adjacent nucleic acid segments are linked;
    (b) depositing the linked nucleic acids into discrete partitions;

(c) processing the linked nucleic acids to produce a plurality of ligation products, wherein the ligation products contain portions of the spatially adjacent nucleic acid segments;

(d) barcoding the ligation products within the discrete partitions to form a plurality of barcoded fragments, wherein fragments within a given discrete partition each comprise a common barcode, thereby associating each fragment with the linked nucleic acid from which it is derived;

(e) obtaining sequence information from the plurality of barcoded fragments, thereby analyzing nucleic acids from the sample while maintaining structural context.

39. The method of claim 38, wherein the processing step (c) includes blunt-end ligation under conditions favoring intramolecular ligation, such that the spatially adjacent nucleic acid segments are ligated within the same molecule.

40. The method of claims 38-39, wherein the sample is a formalin-fixed paraffin sample.

41. The method of claims 38-39, wherein the sample comprises a nuclear matrix.

42. The method of claims 38-41, wherein the nucleic acids comprise RNA.

43. The method of claims 38-42, wherein the nucleic acids are not isolated from the sample prior to step (a).

44. The method of claims 38-42, wherein prior to step (a), the nucleic acids are immunoprecipitated such that associated DNA binding proteins remain bound to the nucleic acids.

45. The method of claims 38-44, wherein the partitions comprise beads.

46. The method of claim 45, wherein the beads are gel beads.

47. The method of claims 38-46, wherein the sample comprises a tumor sample.

48. The method of claims 38-46, wherein the sample comprises a mixture of tumor and normal cells.

49. The method of claims 38-48, wherein the processing step (c) results in ligation products comprising nucleic acids that were originally in close spatial proximity in the sample.

50. The method of claims 38-49, wherein the obtaining step (e) provides information on intrachromosomal and/or interchromosomal interactions between genomic loci.

51. The method of claims 38-50, 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.

52. The method of claim 51, wherein the sequencing reaction is a short read, high accuracy sequencing reaction.

53. A method of analyzing nucleic acids from a sample while maintaining structural context of nucleic acids within the sample, the method comprising:

(a) cross-linking nucleic acids within the sample to form cross-linked nucleic acids, wherein the cross-linking forms covalent links between spatially adjacent nucleic acid segments;

(b) depositing the cross-linked nucleic acids into discrete partitions;

(c) processing the cross-linked nucleic acids to produce a plurality of ligation products, wherein the ligation products contain portions of the spatially adjacent nucleic acid segments;

(d) obtaining sequence information from the plurality of ligation products, thereby analyzing nucleic acids from the sample while maintaining structural context.

54. The method of claim 53, wherein the processing step (b) includes blunt-end ligation under conditions favoring intramolecular ligation, such that the spatially adjacent nucleic acid segments are ligated within the same molecule.

55. The method of claims 53-54, wherein the sample is a formalin-fixed paraffin sample.

56. The method of claims 53-54, wherein the sample comprises a nuclear matrix.

57. The method of claims 53-56, wherein the nucleic acids comprise RNA.

58. The method of claims 53-57, wherein the nucleic acids are not isolated from the sample prior to the cross-linking step (a).

59. The method of claims 53-57, wherein prior to the cross-linking step (a), the nucleic acids are immunoprecipitated such that associated DNA binding proteins remain bound to the nucleic acids.

60. The method of claims 53-59, wherein prior to the obtaining step (d), the ligation products are associated with a barcode.

61. The method of claim 60, wherein ligation products within the same partition receive common barcodes, such that the barcodes identify ligation products from a given partition.

62. The method of claims 53-61, wherein the obtaining step (d) comprises a sequencing reaction selected from the group consisting of: short read-length sequencing reactions and long read-length sequencing reactions.

63. A method of analyzing nucleic acids while maintaining structural context, the method comprising:

(a) providing a sample comprising nucleic acids;

(b) applying a library of tags to the sample such that different geographical regions of the sample receive different tags;

(c) separating portions of the sample into discrete partitions such that portions of the library of tags and portions of the nucleic acids are also separated into the discrete partitions;

(d) obtaining sequence information from the nucleic acids, and

(e) identifying tags in the discrete partitions,

thereby analyzing nucleic acids while maintaining structural context.

64. The method of claim 63, wherein the library of tags comprises oligonucleotide tags.

65. The method of claims 63-64, wherein the library of tags comprises mRNA tags.

66. The method of claims 63-65, wherein the applying step (b) is accomplished through diffusion.

67. The method of claim 66, wherein the applying step (b) results in different concentrations of tags in different spatial locations in the sample.

68. The method of claim 67, wherein the identifying step (e) comprises determining concentration of tags in the discrete partitions.

69. The method of claims 63-65, wherein the applying step (b) is accomplished through one or more active processes within the sample.

70. The method of claim 69, wherein presence of identical tags at different spatial locations in the sample indicates that elements of the sample in those different spatial locations are connected through one or more active processes.

71. The method of claims 69-70, wherein the one or more active processes include cellular transport processes.

72. The method of claims 63-71, wherein the applying step (b) comprises randomly distributing different portions of the library of tags to different spatial locations of the sample.

73. The method of claims 63-71, wherein the applying step (b) comprises adding a predetermined portion or concentration of the library of tags to identified spatial locations of the sample.

74. The method of claims 63-73, wherein the obtaining step (d) and the identifying step (e) are performed simultaneously.

75. The method of claims 63-74, the obtaining step (d) comprises a sequencing reaction selected from the group consisting of: short read-length sequencing reactions and long read-length sequencing reactions.

76. The method of claim 75, wherein the sequencing reaction is a short read, high accuracy sequencing reaction.

77. The method of claims 63-76, wherein the sample comprises a tumor sample.

78. The method of claims 63-76, wherein the sample comprises a mixture of tumor and normal cells.

79. The method of claims 63-76, wherein the sample is a formalin-fixed paraffin sample.

80. The method of claims 63-76, wherein the sample comprises a nuclear matrix.

81. The method of claims 63-76, wherein the nucleic acids comprise RNA.

82. The method of claims 63-82, wherein the partitions comprise beads.

83. The method of claim 82, wherein the beads are gel beads.

84. The method of claims 63-83, wherein the amount of nucleic acids in the discrete partitions is less than 10 ng/ml.

85. The method of claims 63-83, wherein the amount of nucleic acids in the discrete partitions is less than 1 ng/ml.

Figure 1

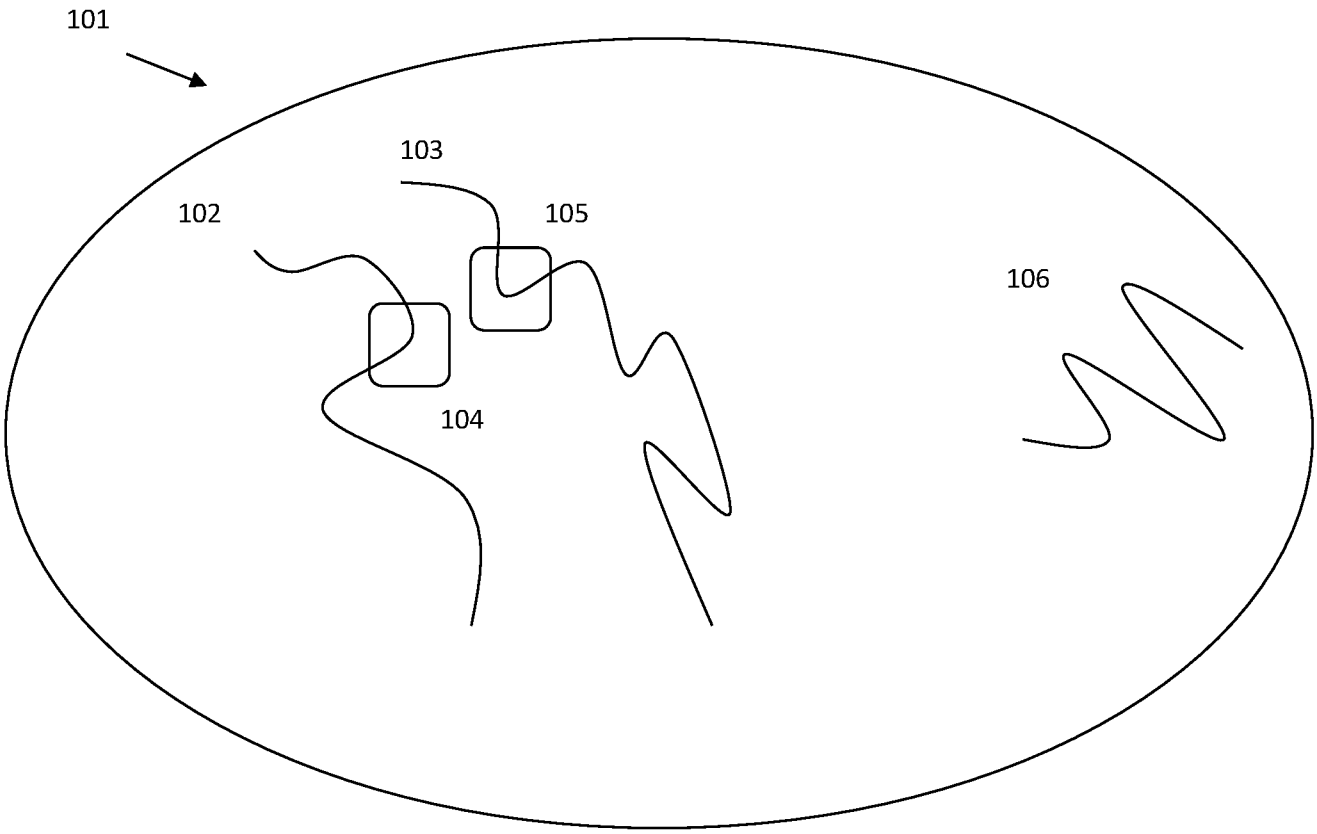
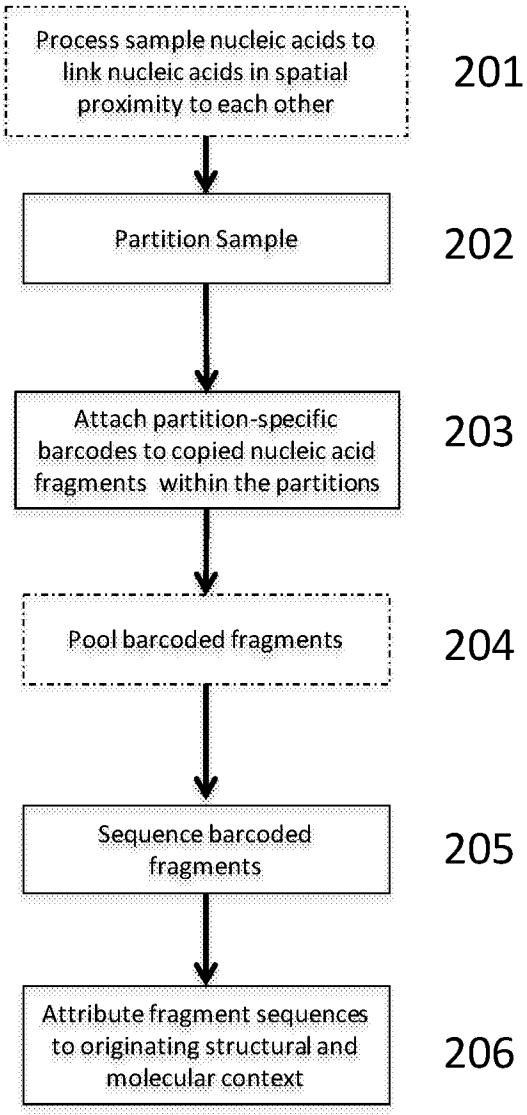


Figure 2


Process sample nucleic acids to link nucleic acids in spatial proximity to each other
201
Partition Sample
202
Attach partition-specific barcodes to copied nucleic acid fragments within the partitions
203
Pool barcoded fragments
204
Sequence barcoded fragments
205
Attribute fragment sequences to originating structural and molecular context
206

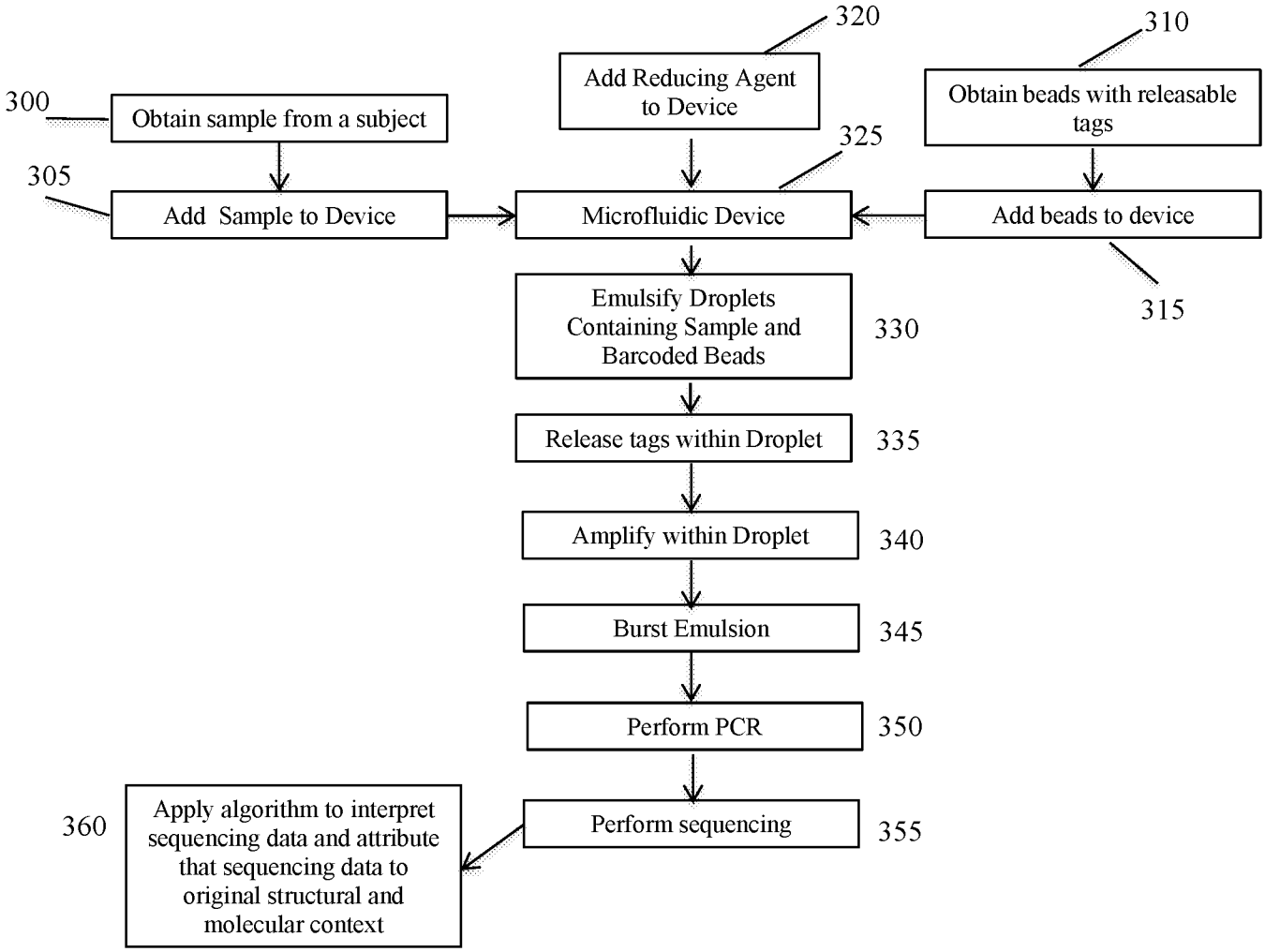

320
310
Add Reducing Agent to Device
Obtain beads with releasable tags
300
Obtain sample from a subject
325
305
Add Sample to Device
Microfluidic Device
Add beads to device
315
Emulsify Droplets Containing Sample and Barcoded Beads
330
Release tags within Droplet
335
Amplify within Droplet
340
Burst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

Figure 3

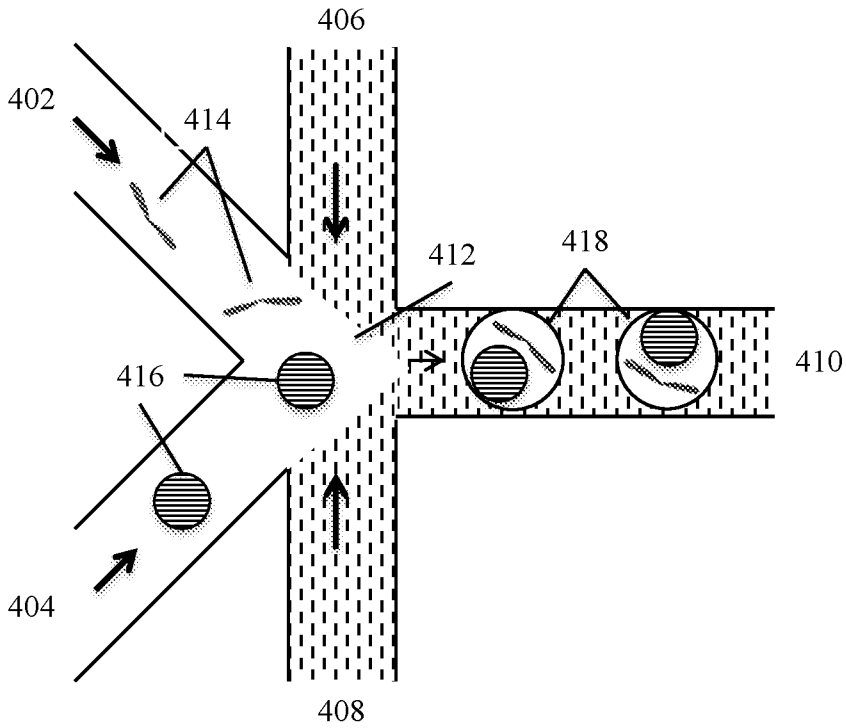


Figure 4

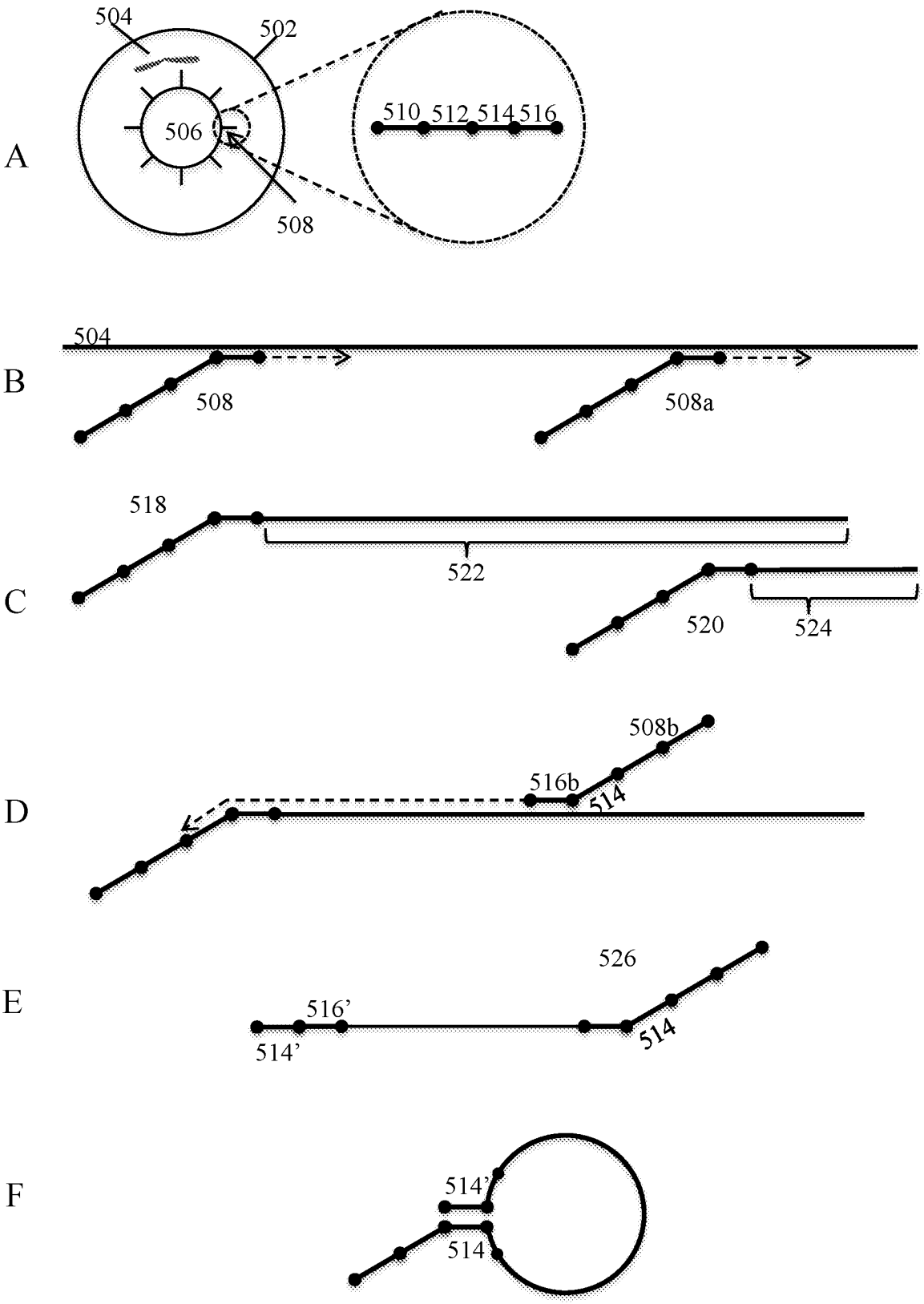

A
504
502
506
508
510 512 514 516
B
504
508
508a
C
518
522
520
524
D
508b
516b
514
E
526
516'
514'
514
F
514'
514

Figure 5

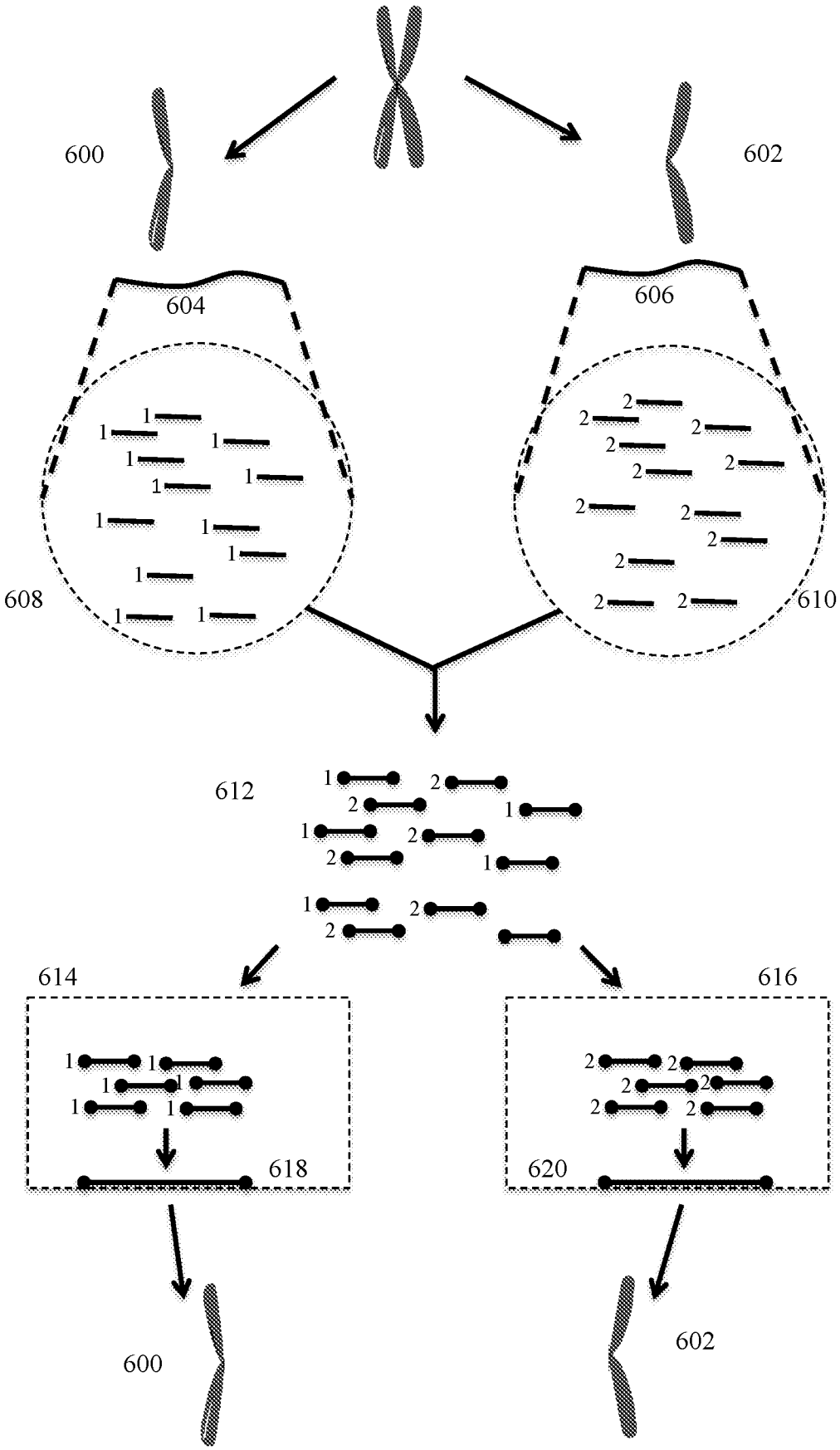


Figure 6

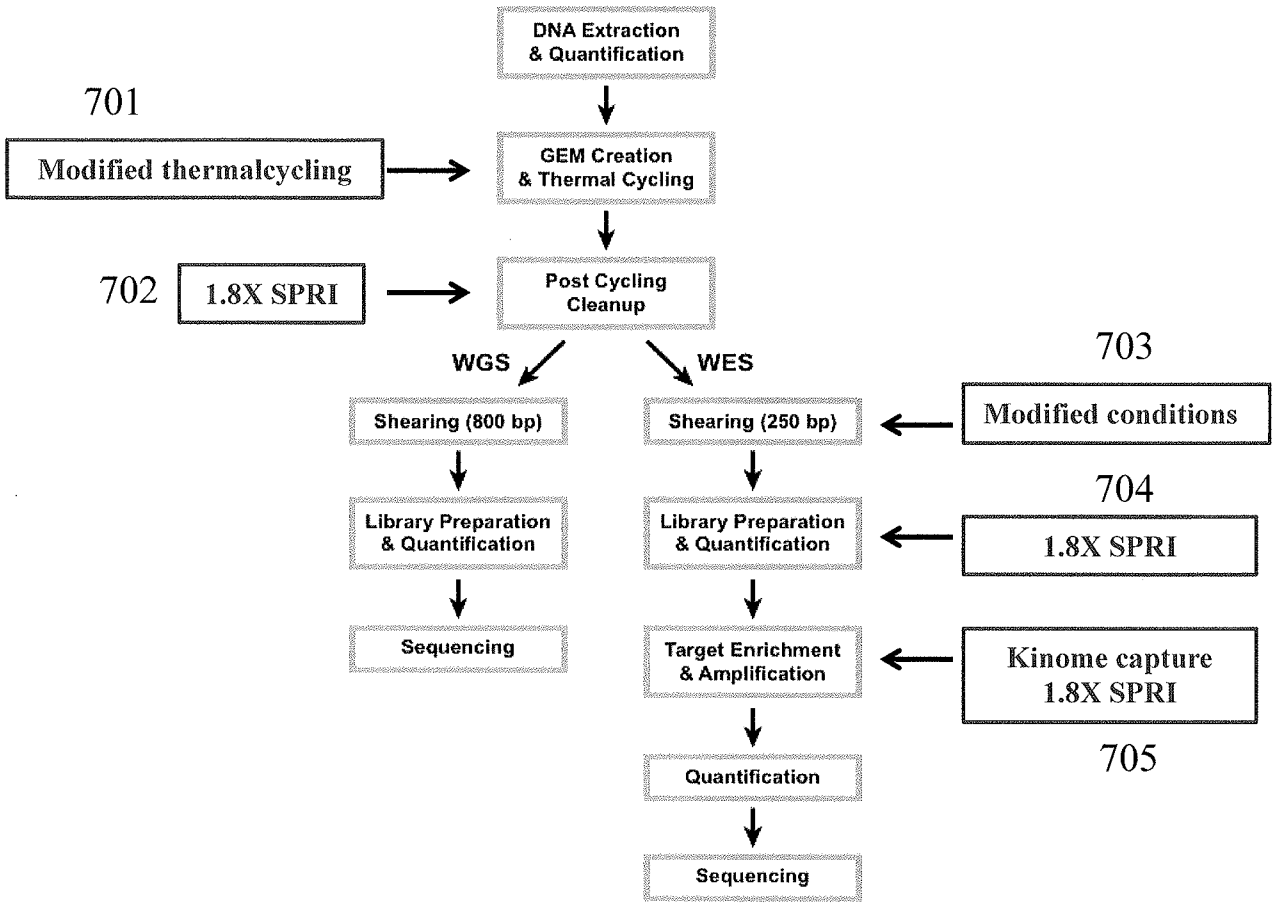

DNA Extraction & Quantification
701
Modified thermalcycling
GEM Creation & Thermal Cycling
702
1.8X SPRI
Post Cycling Cleanup
WGS
WES
703
Modified conditions
Shearing (800 bp)
Shearing (250 bp)
Library Preparation & Quantification
704
1.8X SPRI
Library Preparation & Quantification
Sequencing
Target Enrichment & Amplification
Kinome capture 1.8X SPRI
705
Quantification
Sequencing

Figure 7

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2016/064611

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68
ADD.

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

B. FIELDS SEARCHED

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

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2013/134261 A1 (HARVARD COLLEGE [US]; GEN HOSPITAL CORP [US]) 12 September 2013 (2013-09-12) figures 1, 2 -----	1-15
X	ASSAF ROTEM ET AL: "Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state", NATURE BIOTECHNOLOGY, vol. 33, no. 11, 12 October 2015 (2015-10-12), pages 1165-1172, XP055235766, US ISSN: 1087-0156, DOI: 10.1038/nbt.3383 figure 1 ----- -/--	1-15

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
1 February 2017	04/04/2017

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Aslund, Fredrik

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2016/064611

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	H CHRISTINA FAN ET AL: "Whole-genome molecular haplotyping of single cells", NATURE BIOTECHNOLOGY, vol. 29, no. 1, 1 January 2011 (2011-01-01), pages 51-57, XP055026438, ISSN: 1087-0156, DOI: 10.1038/nbt.1739 figure 1	1-15
X	WO 2013/150083 A1 (MAX PLANCK GESELLSCHAFT [DE]) 10 October 2013 (2013-10-10) figures 1, 3	1-15
A	WO 2013/126741 A1 (RAINDANCE TECHNOLOGIES INC [US]) 29 August 2013 (2013-08-29) claim 1; figure 7	1-15
X	TAKASHI NAGANO ET AL: "Single-cell Hi-C reveals cell-to-cell variability in chromosome structure", NATURE, vol. 502, no. 7469, 25 September 2013 (2013-09-25), pages 59-64, XP055341040, United Kingdom ISSN: 0028-0836, DOI: 10.1038/nature12593 figure 1	1-15

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2016/064611

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

1-15

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-15

A method of analyzing nucleic acids while maintaining structural context, the method comprising:(a) providing a sample comprising nucleic acids, wherein the nucleic acids comprise three dimensional structures;(b) separating portions of the sample into discrete partitions such that portions of the nucleic acid three dimensional structures are also separated into the discrete partitions;(c) obtaining sequence information from the nucleic acids, thereby analyzing nucleic acids while maintaining structural context.

---

2. claims: 16-52

A method of analyzing nucleic acids from a sample while maintaining structural context of nucleic acids within the sample, the method comprising:
(a) forming linked nucleic acids within the sample such that spatially adjacent nucleic acid segments are linked;
(b) optionally (claim 16) processing the linked nucleic acids to produce a plurality of ligation products, wherein the ligation products contain portions of the spatially adjacent nucleic acid segments;
(c) depositing the plurality of ligation products into discrete partitions and optionally carrying out step b) (claim 38);
(d) barcoding the ligation products within the discrete partitions to form a plurality of barcoded fragments, wherein fragments within a given discrete partition each comprise a common barcode, thereby associating each fragment with the linked nucleic acid from which it is derived;
(e) obtaining sequence information from the plurality of barcoded fragments, thereby analyzing nucleic acids from the sample while maintaining structural context.

---

3. claims: 53-62

A method of analyzing nucleic acids from a sample while maintaining structural context of nucleic acids within the sample, the method comprising:(a) cross-linking nucleic acids within the sample to form cross-linked nucleic acids, wherein the cross-linking forms covalent links between spatially adjacent nucleic acid segments;(b) depositing the cross-linked nucleic acids into discrete partitions;(c) processing the cross-linked nucleic acids to produce a plurality of ligation products, wherein the ligation products contain portions of the spatially adjacent nucleic acid segments;(d) obtaining sequence information from the plurality of ligation products, thereby analyzing nucleic

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

acids from the sample while maintaininq structural context.

---

4. claims: 63-85

A method of analyzing nucleic acids while maintaining structural context, the method comprising:(a) providing a sample comprising nucleic acids;(b) applying a library of tags to the sample such that different geographical regions of the sample receive different tags;(c) separating portions of the sample into discrete partitions such that portions of the library of tags and portions of the nucleic acids are also separated into the discrete partitions;(d) obtaining sequence information from the nucleic acids, and(e) identifying tags in the discrete partitions,thereby analyzing nucleic acids while maintaining structural context.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2016/064611

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2013134261 A1	12-09-2013	EP 2823064 A1	14-01-2015
		US 2015057163 A1	26-02-2015
		WO 2013134261 A1	12-09-2013
WO 2013150083 A1	10-10-2013	CA 2868689 A1	10-10-2013
		CA 2868691 A1	10-10-2013
		EP 2647426 A1	09-10-2013
		EP 2833997 A1	11-02-2015
		EP 2833998 A1	11-02-2015
		US 2015072867 A1	12-03-2015
		US 2015141269 A1	21-05-2015
		WO 2013150082 A1	10-10-2013
		WO 2013150083 A1	10-10-2013
WO 2013126741 A1	29-08-2013	EP 2817418 A1	31-12-2014
		US 2013225418 A1	29-08-2013
		WO 2013126741 A1	29-08-2013