

USING FLOWCELL SPATIAL COORDINATES TO LINK READS FOR IMPROVED GENOME ANALYSIS

INCORPORATION BY REFERENCE TO ANY PRIORITY APPLICATIONS

[0001] This application claims priority to U.S. Provisional Application No. 63/511,593 filed June 30, 2023, the content of which is incorporated by reference in its entirety.

BACKGROUND

[0002] Traditional nucleic acid sequencing methods, and several types of next-generation sequencing methods, use a shotgun approach to sequence large genomic DNA fragments, called template genomic sequences. Specifically, template genomic sequences are first fragmented in solution into smaller pieces that are amenable to next-generation sequencing methods on a flowcell. One of the difficulties of this approach is that by the time the smaller sequence fragments from the template genomic sequences have been read, knowledge of their connectivity and proximity to each other in the original template genomic sequence is lost. The process of ordering the sequence fragments to arrive at the sequence of the original template genomic sequence is generally referred to as "assembly." Assembly processes can be computationally intensive and time-consuming. In addition, sequence and assembly errors can become a problem depending upon the sequencing methodology used and the quality of genomic DNA samples under evaluation.

[0003] Moreover, many genomes of interest contain more than one version of each chromosome. For example, the human genome is diploid, having two sets of chromosomes—one set inherited from each parent. Some organisms have polyploid genomes with more than two sets of chromosomes. Examples of polyploid organisms include animals, such as salmon, and many plant species such as wheat, apple, oat and sugar cane. When diploid and polyploid genomes are fragmented and sequenced in typical shotgun methods, phasing information, pertaining to the identity of which fragments came from which set of chromosomes, is lost. This phasing information can be difficult or impossible to reconstruct using typical shotgun methods.

[0004] Somewhat similar yet often more complex difficulties can arise when mixed samples are evaluated. Mixed samples can contain nucleic acid molecules, such as chromosomes, mRNA transcripts, plasmids etc., from two or more organisms. Mixed samples having multiple organisms are often referred to as metagenomic samples. Other examples of mixed samples are different cells or tissues that although being derived from the same organism have different characteristics. Examples include cancerous tissues which may comprise a mixture of healthy cells and cancerous cells, tissues that may comprise pre-cancerous cells and cancerous cells, tissue that may comprise two or more different types of cancerous cells. Indeed, there may be a variety of different types of cancer cells as is the case for cancer samples that have mosaicity. Another example of different cells derived from a single organism are mixtures of maternal and fetal cells obtained from a pregnant female (e.g. from the blood or from tissues). When mixed nucleic acid samples are fragmented and sequenced in typical shotgun methods information pertaining to the identity of which fragments came from which cell, organism or other source is lost. This origin information can be difficult or impossible to reconstruct using typical shotgun methods.

SUMMARY

[0005] The methods disclosed herein each have several aspects, no single one of which is solely responsible for their desirable attributes. Without limiting the scope of the claims, some prominent features will now be discussed briefly. Numerous other embodiments are also contemplated, including embodiments that have fewer, additional, and/or different components, steps, features, objects, benefits, and advantages. The components, aspects, and steps may also be arranged and ordered differently. After considering this discussion, and particularly after reading the section entitled “Detailed Description”, one will understand how the features of the devices and methods disclosed herein provide advantages over other known devices and methods.

[0006] In some embodiments, a method for assigning nucleic acid sequence reads to target polynucleotides is provided, the method including providing transposome complexes, wherein the transposome complexes include a transposase and a first polynucleotide including an end sequence and a first tag; contacting the transposome complexes with target polynucleotides under conditions to fragment the target polynucleotides; amplifying the

fragmented target polynucleotides to form a plurality of nucleic acid clusters on a substrate; obtaining location information for the plurality of nucleic acid clusters on the substrate; determining the nucleic acid sequence reads of the fragmented nucleic acids in each of the nucleic acid clusters; and assigning the nucleic acid sequence reads to the target polynucleotides using the obtained location information.

[0007] In some embodiments, a length of the target polynucleotides is greater than a length of the fragment. In some embodiments, assigning the nucleic acid sequence reads includes determining the distance between each of the clusters and using the determined distance to assign reads to a specific target polynucleotide. In some embodiments, assigning the nucleic acid sequence reads includes determining for a likelihood score that at least a first and a second cluster on the substrate derive from the same target polynucleotide.

[0008] In some embodiments, the method further includes increasing the likelihood score for the first cluster when the spatial distance between at least the first and second clusters are below a threshold value. In some embodiments, the method further includes increasing the likelihood score for the first cluster when a genomic distance between at least the first and second clusters are below a threshold value. In some embodiments, the method further includes increasing the likelihood score for the first cluster when a genomic distance between at least the first and second clusters are below the threshold value. In some embodiments, the method further includes increasing the likelihood score for the first cluster when the spatial distance and a genomic distance between at least the first and second clusters are below a threshold value. In some embodiments, the likelihood score is influenced by a pitch of the substrate, size of the substrate, a pattern of the substrate, temperature, loading density, fragment directionality, or a combination thereof.

[0009] In some embodiments, the method further includes determining whether the target nucleic acid has a variant when the spatial distance between at least the first and second clusters are below a threshold value and when the genomic distance is above the genomic distance threshold. In some embodiments, the spatial distance threshold from a cluster forms a pattern of an ellipse or a circle around the cluster. In some embodiments, the likelihood score of the first cluster is 0. In some embodiments, the likelihood score of the second cluster is above 30. In some embodiments, the likelihood score of one or more other clusters is above 30. In some embodiments, the location information includes a first spatial coordinate and a

second spatial coordinate in a cartesian coordinate system. In some embodiments, the method further includes sorting the plurality of the nucleic acid clusters by their spatial coordinates.

[0010] In some embodiments, the transposome complexes include a second polynucleotide including a region complementary to the transposon end sequence. In some embodiments, the transposome complexes are present on the substrate at a density of at least 103, 104, 105, 106, 107, 108, 109, or 1010 or more complexes per mm². In some embodiments, the transposome complexes include a hyperactive Tn5 transposase.

[0011] In some embodiments, the substrate includes microparticles. In some embodiments, the substrate includes a patterned surface. In some embodiments, the substrate includes wells. In some embodiments, providing the transposome complexes includes providing the transposome complexes in solution. In some embodiments, providing the transposome complexes includes providing the transposome complexes bound to the substrate.

[0012] In some embodiments, a system for assigning sequence reads on a substrate to their original target nucleic acid, is provided the system including a substrate including clusters of amplified fragments of target polynucleotides bound to the substrate in spatial locations; one or more processors having instructions that when executed perform a method including obtaining spatial location information for the clusters of amplified fragments on the substrate; sequencing the target polynucleotides in each of the clusters to determine a sequence read and spatial location from each of the amplified fragments; determining the geographic distance between each cluster on the substrate; and assigning sequence reads from each cluster to a target nucleic acid based on their geographic distance from one another.

[0013] In some embodiments, a length of the target polynucleotides is greater than a length of the fragment. In some embodiments, assigning the nucleic acid sequence reads includes determining the distance between each of the clusters and using the determined distance to assign reads to a specific target polynucleotide. In some embodiments, assigning the nucleic acid sequence reads includes determining for a likelihood score that at least a first and a second cluster on the substrate derive from the same target polynucleotide. In some embodiments, the one or more processors further performs a method including increasing the likelihood score for the first cluster when the spatial distance between at least the first and second clusters is below a threshold value. In some embodiments, the one or more processors further performs a method including increasing the likelihood score for the first cluster when

a genomic distance between at least the first and second clusters is below a threshold value. In some embodiments, the one or more processors further performs a method including increasing the likelihood score for the first cluster when a genomic distance between at least the first and second clusters is below the threshold value. In some embodiments, the one or more processors further performs a method including increasing the likelihood score for the first cluster when the spatial distance and a genomic distance between at least the first and second clusters are below a threshold value. In some embodiments, the likelihood score is influenced by a pitch of the substrate, size of the substrate, a pattern of the substrate, temperature, loading density, fragment directionality, or a combination thereof.

[0014] In some embodiments, the one or more processors further performs a method including determining whether the target nucleic acid has a variant when the spatial distance between the first and second clusters are below a threshold value and when the genomic distance is above the genomic distance threshold. In some embodiments, the spatial distance threshold from a cluster forms a pattern of an ellipse or a circle around the cluster. In some embodiments, the likelihood score of the first cluster is 0. In some embodiments, the likelihood score of the second cluster is above 30. In some embodiments, the likelihood score of one or more other clusters is above 30. In some embodiments, the location information includes a first spatial coordinate and a second spatial coordinate in a 2D coordinate system. In some embodiments, the one or more processors further performs a method including sorting the plurality of the nucleic acid clusters by their spatial coordinates.

[0015] In some embodiments, the transposome complexes include a second polynucleotide including a region complementary to the transposon end sequence. In some embodiments, the transposome complexes are present on the substrate at a density of at least 103, 104, 105, 106, 107, 108, 109, or 1010 or more complexes per mm². In some embodiments, the transposome complexes include a hyperactive Tn5 transposase. In some embodiments, the substrate includes microparticles. In some embodiments, the substrate includes a patterned surface. In some embodiments, the substrate includes wells.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present

invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings (also “figure” and “FIG.” herein).

[0017] FIG. 1 illustrates the difficulty in mapping two identical sequences, A1 and A2.

[0018] FIG. 2 is a flow diagram illustrating a method according to some embodiments.

[0019] FIGS. 3A-3B are schematical diagrams illustrating the disambiguation of multi-mapped reads according to some embodiments.

[0020] FIGS. 4A-4B illustrate reads based on spatial location and genomic distance according to some embodiments. FIG. 4A illustrates an update to the MAPQ value and FIG. 4B illustrates when the MAPQ value is not updated.

[0021] FIG. 5 shows a non-limiting example of a solid support according to some embodiments.

[0022] FIG. 6 illustrates clustering on a solid support according to some embodiments.

[0023] FIG. 7 illustrates clustering on a solid support according to some embodiments.

[0024] FIG. 8 illustrates non-limiting examples of metrics related to spatial distance information from the clusters according to some embodiments.

[0025] FIG. 9 illustrates non-limiting examples of spatial information obtained from a nucleic acid.

[0026] FIG. 10 is a flow diagram illustrating a non-limiting example of updating a MAPQ score of a MAP0 read.

[0027] FIG. 11 is a flow diagram illustrating a non-limiting example of updating a MAPQ score of a MAP0 read.

DETAILED DESCRIPTION

[0028] While various embodiments of the 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 may 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.

[0029] Embodiments of the invention relate to systems and methods for sequencing target nucleic acids by fragmenting the target nucleic acid and distributing the fragments onto a flow cell. As the fragments are distributed along the flow cell, they bind capture primers and are then used to create clusters by well-known technologies, such as those provided by Illumina Inc. (San Diego, CA). It has been discovered that fragments which were derived from the same template genomic sequence are more likely to bind to the flow cell in spatially nearby positions as compared to fragments that are from different template genomic sequences, particularly when the fragmentation is performed directly on the flow cell using immobilized transposome complexes on the surface of the flow cell. This spatial information can be used to help guide assembly and variant calling of the original template genomic sequence, as will be described in more detail below.

[0030] For example, one embodiment is a method for assigning nucleic acid sequence reads to target polynucleotides, which includes providing transposome complexes. In some embodiments, the transposome complexes include a transposase and a first polynucleotide having end sequences which can be used to fragment the target polynucleotides and insert into each fragment an end sequence or tag which can be used to bind to capture probes located on the substrate. The method can include contacting the transposome complexes with the target polynucleotides under conditions to fragment the target polynucleotides and add capture sequences to the ends of each fragment. In some embodiments, the capture sequences include P5 or P7 sequences as provided by Illumina, Inc. In some embodiments, the complexed strand and transposome is in solution, and is then brought towards a substrate and immobilized thereon. In some embodiments, prior to immobilization of the transposome complexes on the substrate, one or more of the transposome complexes bind the target polynucleotides in solution. In this embodiment, the transposome complexes in solution become immobilized to the substrate.

[0031] Once the fragments have been bound to substrate, the bound fragments can be amplified to form a plurality of nucleic acid clusters on the substrate. The location of each cluster on the flow cell can then be determined before, during or after performing sequencing by synthesis reactions (SBS) to obtain the nucleotide sequence of each fragment located in

each cluster. Once the nucleotide sequence of each cluster has been determined, the method can start to map those reads to determine the original target polynucleotide from which the read originated. In some embodiments, the mapping process takes into account the spatial location of each cluster, such that clusters which are closer to each other on the flow cell are more likely to have originated from the same target polynucleotide. In some embodiments, the library preparation steps are performed on the flow cell, which may reduce the complexity and the amount of equipment required for the systems. Furthermore, by mapping the sequenced fragments to target polynucleotides using the spatial information accompanying each cluster, the method performs more accurate mapping operations as compared to methods that do not take the spatial location of each cluster into account during the mapping process. Therefore, spatial information that includes relative distances between various clusters on a flow cell is leveraged to adjust mapping information, thereby increasing the read quality of previously identified multi-mapped reads. In the past, identified multi-mapped reads may have been discarded. Increasing the read quality of these previously discarded reads may improve the alignment information and quality of information used in certain genomic analysis applications including, but not limited to, variant calling.

[0032] In order to manage the spatial location of each cluster during the mapping process, each flow cell is divided into swaths and tiles. Each swath is a longitudinal stripe of the flow cell, and each tile is an area within each stripe. More information on this schema can be found with reference to Figs. 5, 6 and 7. In some embodiments, each tile on the flow cell is given a unique tile number, which includes the corresponding swath where the tile is located. When the nucleotide sequence of a cluster is determined, it is stored using a filename or readname which includes not only the sequence information but also the identification of the swath and tile. A non-limiting example of storing such information includes storing the information in a specific file format, such as a fastq file format. Then, during the assembly process, if the process determines that a particular fragment maps to more than one possible target polynucleotide, the process reads the spatial location of the cluster and uses that spatial location to determine if one of the mapping assignments is more likely than the other based on the spatial location of each read on the flow cell. Reads which are within a relatively short distance with one another are more likely to have derived from the same target polynucleotide, so the process assigns a read which is mapped to multiple target polynucleotides to a particular

one based on their relative spatial locations on the flow cell. Therefore, spatial distances between clusters (e.g., on a flow cell) may be used to improve the speed and quality of mapping reads to their target polynucleotides.

[0033] In some embodiments, the sequencing system first determines the sequence and location of reads in each cluster on the flow cell. The system creates a BAM file for each such read which includes the read sequence and identification of one or more possible target polynucleotides. For any BAM file which includes more than one possible target polynucleotide, the system attempts to disambiguate the reads by referring to the spatial location information contained with each read. As mentioned above, if some reads are very near one another on the flow cell, then the system determines that one read is more likely to have originated from a particular target polynucleotide than another read. If a particular read can be disambiguated, then the BAM file entry for that read is updated to indicate only the correct target polynucleotide. This will be explained more fully in the sections and examples below.

Definitions

[0034] The section headings used herein are for organizational purposes only and are not to be construed as limiting the subject matter described.

[0035] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as is commonly understood by one of ordinary skill in the art. The use of the term “including” as well as other forms, such as “include”, “includes,” and “included,” is not limiting. The use of the term “having” as well as other forms, such as “have”, “has,” and “had,” is not limiting. As used in this specification, whether in a transitional phrase or in the body of the claim, the terms “comprise(s)” and “comprising” are to be interpreted as having an open-ended meaning. That is, the above terms are to be interpreted synonymously with the phrases “having at least” or “including at least.” For example, when used in the context of a process, the term “comprising” means that the process includes at least the recited steps, but may include additional steps. When used in the context of a compound, composition, or device, the term “comprising” means that the compound, composition, or device includes at least the recited features or components, but may also include additional features or components.

[0036] The terms “polynucleotide,” “oligonucleotide,” “nucleic acid” and “nucleic acid molecules” are used interchangeably herein and refer to a covalently linked sequence of nucleotides of any length (i.e., ribonucleotides for RNA, deoxyribonucleotides for DNA, analogs thereof, or mixtures thereof) in which the 3' position of the pentose of one nucleotide is joined by a phosphodiester group to the 5' position of the pentose of the next. The terms should be understood to include, as equivalents, analogs of either DNA, RNA, cDNA, or antibody-oligo conjugates made from nucleotide analogs and to be applicable to single stranded (such as sense or antisense) and double stranded polynucleotides. The term as used herein also encompasses cDNA, that is complementary or copy DNA produced from an RNA template, for example by the action of reverse transcriptase. This term refers only to the primary structure of the molecule. Thus, the term includes, without limitation, triple-, double- and single-stranded deoxyribonucleic acid (“DNA”), as well as triple-, double- and single-stranded ribonucleic acid (“RNA”). The nucleotides include sequences of any form of nucleic acid. As apparent from the examples below and elsewhere herein, a nucleic acid can have a naturally occurring nucleic acid structure or a non-naturally occurring nucleic acid analog structure. A nucleic acid can contain phosphodiester bonds; however, in some embodiments, nucleic acids may have other types of backbones, comprising, for example, phosphoramidate, phosphorothioate, phosphorodithioate, O-methylphosphoroamidite and peptide nucleic acid backbones and linkages. Nucleic acids can have positive backbones; non-ionic backbones, and non-ribose based backbones. Nucleic acids may also contain one or more carbocyclic sugars. The nucleic acids used in methods or compositions herein may be single stranded or, alternatively double stranded, as specified. In some embodiments a nucleic acid can contain portions of both double stranded and single stranded sequence, for example, as demonstrated by forked adapters. A nucleic acid can contain any combination of deoxyribo- and ribonucleotides, and any combination of bases, including uracil, adenine, thymine, cytosine, guanine, inosine, xanthine, hypoxanthine, isocytosine, isoguanine, and base analogs such as nitropyrrole (including 3-nitropyrrole) and nitroindole (including 5-nitroindole), etc. In some embodiments, a nucleic acid can include at least one promiscuous base. A promiscuous base can base-pair with more than one different type of base and can be useful, for example, when included in oligonucleotide primers or inserts that are used for random hybridization in complex nucleic acid samples such as genomic DNA samples. An example of a promiscuous

base includes inosine that may pair with adenine, thymine, or cytosine. Other examples include hypoxanthine, 5-nitroindole, acyclic 5-nitroindole, 4-nitropyrrole, 4-nitroimidazole and 3-nitropyrrole. Promiscuous bases that can base-pair with at least two, three, four or more types of bases can be used.

[0037] As used herein, the term "fragment," when used in reference to a first nucleic acid, is intended to mean a second nucleic acid having a part or portion of the sequence of the first nucleic acid. Generally, the fragment and the first nucleic acid are separate molecules. The fragment can be derived, for example, by physical removal from the larger nucleic acid, by replication or amplification of a region of the larger nucleic acid, by degradation of other portions of the larger nucleic acid, a combination thereof or the like. The term can be used analogously to describe sequence data or other representations of nucleic acids. As used herein, the term "haplotype" refers to a set of alleles at more than one locus inherited by an individual from one of its parents. A haplotype can include two or more loci from all or part of a chromosome. Alleles include, for example, single nucleotide polymorphisms (SNPs), short tandem repeats (STRs), gene sequences, chromosomal insertions, chromosomal deletions etc. The term "phased alleles" refers to the distribution of the particular alleles from a particular chromosome, or portion thereof. Accordingly, the "phase" of two alleles can refer to a characterization or representation of the relative location of two or more alleles on one or more chromosomes.

[0038] As used herein, the term "nucleotide sequence" is intended to refer to the order and type of nucleotide monomers in a nucleic acid polymer. A nucleotide sequence is a characteristic of a nucleic acid molecule and can be represented in any of a variety of formats including, for example, a depiction, image, electronic medium, series of symbols, series of numbers, series of letters, series of colors, etc. The information can be represented, for example, at single nucleotide resolution, at higher resolution (e.g. indicating molecular structure for nucleotide subunits) or at lower resolution (e.g. indicating chromosomal regions, such as haplotype blocks). A series of "A," "T," "G," and "C" letters is a well-known sequence representation for DNA that can be correlated, at single nucleotide resolution, with the actual sequence of a DNA molecule. A similar representation is used for RNA except that "T" is replaced with "U" in the series.

[0039] As used herein, the term "solid support" refers to a rigid substrate that is insoluble in aqueous liquid. The substrate can be non-porous or porous. The substrate can optionally be capable of taking up a liquid (e.g. due to porosity) but will typically be sufficiently rigid that the substrate does not swell substantially when taking up the liquid and does not contract substantially when the liquid is removed by drying. A nonporous solid support is generally impermeable to liquids or gases. Exemplary solid supports include, but are not limited to, glass and modified or functionalized glass, plastics (including acrylics, polystyrene and copolymers of styrene and other materials, polypropylene, polyethylene, polybutylene, polyurethanes, Teflon™, cyclic olefins, polyimides etc.), nylon, ceramics, resins, Zeonor, silica or silica-based materials including silicon and modified silicon, carbon, metals, inorganic glasses, optical fiber bundles, and polymers. Particularly useful solid supports for some embodiments are located within a flow cell apparatus. Exemplary flow cells are set forth in further detail below.

[0040] As used herein, the term "flow cell" is intended to mean a chamber having a surface across which one or more fluid reagents can be flowed. Generally, a flow cell will have an ingress opening and an egress opening to facilitate flow of fluid. A flow cell can have multiple surfaces. Examples of flow cells and related fluidic systems and detection platforms that can be readily used in the methods of the present disclosure are described, for example, in Bentley et al, Nature 456:53-59 (2008), WO 04/018497; US 7,057,026; WO 91/06678; WO 07/123744; US 7,329,492; US 7,211,414; US 7,315,019; US 7,405,281, and US 2008/0108082, each of which is incorporated herein by reference.

[0041] In many embodiments, a solid support to which nucleic acids are attached in a method set forth herein will have a continuous or monolithic surface. Thus, fragments can attach at spatially random locations wherein the distance between nearest neighbor fragments (or nearest neighbor clusters derived from the fragments) will be variable. The resulting arrays will have a variable or random spatial pattern of features. Alternatively, a solid support used in a method set forth herein can include an array of features that are present in a repeating pattern. In such embodiments, the features provide the locations to which modified nucleic acid polymers, or fragments thereof, can attach. Particularly useful repeating patterns are hexagonal patterns, rectilinear patterns, grid patterns, patterns having reflective symmetry, patterns having rotational symmetry, or the like. The features to which a modified nucleic acid

polymer, or fragment thereof, attach can each have an area that is smaller than about 1mm^2 , $500\text{ }\mu\text{m}^2$, $100\text{ }\mu\text{m}^2$, $25\text{ }\mu\text{m}^2$, $10\text{ }\mu\text{m}^2$, $5\text{ }\mu\text{m}^2$, $1\text{ }\mu\text{m}^2$, 500 nm^2 , or 100 nm^2 . Alternatively, or additionally, each feature can have an area that is larger than about 100 nm^2 , 250 nm^2 , 500 nm^2 , $1\text{ }\mu\text{m}^2$, $2.5\text{ }\mu\text{m}^2$, $5\text{ }\mu\text{m}^2$, $10\text{ }\mu\text{m}^2$, $100\text{ }\mu\text{m}^2$, or $500\text{ }\mu\text{m}^2$. A cluster or colony of nucleic acids that result from amplification of fragments on an array (whether patterned or spatially random) can similarly have an area that is in a range above or between an upper and lower limit selected from those exemplified above.

[0042] For embodiments that include an array of features on a surface, the features can be discrete, being separated by interstitial regions. Alternatively, some or all of the features on a surface can be abutting (i.e. not separated by interstitial regions). Whether the features are discrete or abutting, the average size of the features and/or average distance between the features can vary such that arrays can be high density, medium density or lower density. High density arrays are characterized as having features with average pitch of less than about $15\text{ }\mu\text{m}$. Medium density arrays have average feature pitch of about 15 to $30\text{ }\mu\text{m}$, while low density arrays have average feature pitch of greater than $30\text{ }\mu\text{m}$. An array useful in the invention can have feature pitch of, for example, less than $100\text{ }\mu\text{m}$, $50\text{ }\mu\text{m}$, $10\text{ }\mu\text{m}$, $5\text{ }\mu\text{m}$, $1\text{ }\mu\text{m}$ or $0.5\text{ }\mu\text{m}$. Alternatively or additionally, the feature pitch can be, for example, greater than $0.1\text{ }\mu\text{m}$, $0.5\text{ }\mu\text{m}$, $1\text{ }\mu\text{m}$, $5\text{ }\mu\text{m}$, $10\text{ }\mu\text{m}$, $50\text{ }\mu\text{m}$, or $100\text{ }\mu\text{m}$.

[0043] As used herein, the term "source" is intended to include an origin for a nucleic acid molecule, such as a tissue, cell, organelle, compartment, or organism. The term can be used to identify or distinguish an origin for a particular nucleic acid in a mixture that includes origins for several other nucleic acids. A source can be a particular organism in a metagenomic sample having several different species of organisms. In some embodiments the source will be identified as an individual origin (e.g. an individual cell or organism). Alternatively, the source can be identified as a species that encompasses several individuals of the same type in a sample (e.g. a species of bacteria or other organism in a metagenomic sample having several individual members of the species along with members of other species as well).

[0044] As used herein, the term "surface," when used in reference to a material, is intended to mean an external part or external layer of the material. The surface can be in contact with another material such as a gas, liquid, gel, polymer, organic polymer, second surface of a similar or different material, metal, or coat. The surface, or regions thereof, can be substantially

flat. The surface can have surface features such as wells, pits, channels, ridges, raised regions, pegs, posts or the like. The material can be, for example, a solid support, gel, or the like.

[0045] As an example, in some embodiments, fragments derived from a long nucleic acid molecule captured at the surface of a flow cell occur in a line across the surface of the flow cell (e.g. if the nucleic acid was stretched out prior to fragmentation or amplification) or in a cloud on the surface. Further, a physical map of the immobilized nucleic acid can then be generated. The physical map thus correlates the physical relationship of clusters after immobilized nucleic acid is amplified. Specifically, the physical map is used to calculate the probability that sequence data obtained from any two clusters are linked, as described in the incorporated materials of WO 2012/025250. Alternatively or additionally, the physical map can be indicative of the genome of a particular organism in a metagenomic sample. In this latter case the physical map can indicate the order of sequence fragments in the organism's genome; however, the order need not be specified and instead the mere presence of two or more fragments in a common organism (or other source or origin) can be sufficient basis for a physical map that characterizes a mixed sample and one or more organisms therein.

[0046] In some embodiments, the physical map is generated by imaging the solid support to establish the location of the immobilized nucleic acid molecules across the surface. In some embodiments, the immobilized nucleic acid is imaged by adding an imaging agent to the solid support and detecting a signal from the imaging agent. In some embodiments, the imaging agent is a detectable label. Suitable detectable labels, include, but are not limited to, protons, haptens, radionuclides, enzymes, fluorescent labels, chemiluminescent labels, and/or chromogenic agents. For example, in some embodiments, the imaging agent is an intercalating dye or non-intercalating DNA binding agent. Any suitable intercalating dye or non-intercalating DNA binding agent as are known in the art can be used, including, but not limited to those set forth in U.S. 2012/0282617, which is incorporated herein by reference.

[0047] In certain embodiments, a plurality of modified nucleic acid molecules is flowed onto a flow cell comprising a plurality of nano-channels. As used herein, the term nano-channel refers to a narrow channel into which a long linear nucleic acid molecule is stretched. In some embodiments, no more than 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900 or no more than 1000 individual long strands of nucleic acid are stretched across each nano-channel. In some

embodiments the individual nano-channels are separated by a physical barrier that prevents individual long strands of target nucleic acid from interacting with multiple nano-channels. In some embodiments, the solid support comprises at least 10, 50, 100, 200, 500, 1000, 3000, 5000, 10000, 30000, 50000, 80000 or at least 100000 nano-channels.

[0048] As used herein, the term "target," when used in reference to a nucleic acid polymer, is intended to linguistically distinguish the nucleic acid, for example, from other nucleic acids, modified forms of the nucleic acid, fragments of the nucleic acid, and the like. Any of a variety of nucleic acids set forth herein can be identified as target nucleic acids, examples of which include genomic DNA (gDNA), messenger RNA (mRNA), copy or complimentary DNA (cDNA), and derivatives or analogs of these nucleic acids.

[0049] As used herein, the term "transposase" is intended to mean an enzyme that is capable of forming a functional complex with a transposon element-containing composition (e.g., transposons, transposon ends, transposon end compositions) and catalyzing insertion or transposition of the transposon element-containing composition into a target DNA with which it is incubated, for example, in an in vitro transposition reaction. The term can also include integrases from retrotransposons and retroviruses. Transposases, transposomes and transposome complexes are generally known to those of skill in the art, as exemplified by the disclosure of US Pat. App. Pub. No. 2010/0120098, which is incorporated herein by reference. Although many embodiments described herein refer to Tn5 transposase and/or hyperactive Tn5 transposase, it will be appreciated that any transposition system that is capable of inserting a transposon element with sufficient efficiency to tag a target nucleic acid can be used. In particular embodiments, a preferred transposition system is capable of inserting the transposon element in a random or in an almost random manner to tag the target nucleic acid. As used herein, the term "transposome" is intended to mean a transposase enzyme bound to a nucleic acid. Typically the nucleic acid is double stranded. For example, the complex can be the product of incubating a transposase enzyme with double-stranded transposon DNA under conditions that support non-covalent complex formation. Transposon DNA can include, without limitation, Tn5 DNA, a portion of Tn5 DNA, a transposon element composition, a mixture of transposon element compositions or other nucleic acids capable of interacting with a transposase such as the hyperactive Tn5 transposase.

[0050] As used herein, the term "transposon element" is intended to mean a nucleic acid molecule, or portion thereof, that includes the nucleotide sequences that form a transposome with a transposase or integrase enzyme. Typically, the nucleic acid molecule is a double stranded DNA molecule. In some embodiments, a transposon element is capable of forming a functional complex with the transposase in a transposition reaction. As non-limiting examples, transposon elements can include the 19-bp outer end ("OE") transposon end, inner end ("IE") transposon end, or "mosaic end" ("ME") transposon end recognized by a wild-type or mutant Tn5 transposase, or the R1 and R2 transposon end as set forth in the disclosure of US Pat. App. Pub. No. 2010/0120098, which is incorporated herein by reference. Transposon elements can comprise any nucleic acid or nucleic acid analogue suitable for forming a functional complex with the transposase or integrase enzyme in an in vitro transposition reaction. For example, the transposon end can comprise DNA, RNA, modified bases, non-natural bases, modified backbone, and can comprise nicks in one or both strands.

[0051] A standard NGS sequencing run yields millions of short sequences that are eventually mapped on a reference genome. A percentage of good-quality reads (1-5%) are discarded because of ambiguous genomic location. Increasing read length (2x500 or long-read sequencing), designing a specialized algorithm to map reads on specific regions of the genome (targeted callers), using expensive and time-consuming library preparation (Illumina CLR), or a combination thereof may be implemented to address the need for disambiguating such reads that would normally be discarded. However, such approaches are costly, laborious, and time intensive. Spatial information (X and Y coordinates) obtained from a solid support surface) can be leveraged to identify fragments that are generated from a single long input fragment and subsequently be used to improve mapping reads in ambiguous positions.

[0052] FIG. 1 is a diagram 100 showing the difficulty resolving sequence reads from identical sequences A1 and A2 on a reference genome. Sequence reads that are found within the illustrated A1 and A2 sequences are poor candidates for mapping due to each sequence read resolving to either one of the A1 or A2 sequences. This can lead to such sequencing reads being considered as "multi-mapped" reads because they map to more than a single position on the target reference genome. Because such reads cannot be mapped to any particular location on the reference genome, these poorly mapped reads cannot easily be used for variant calling, and thus are often discarded even if their read quality is high.

[0053] To overcome such a challenge, FIG. 2 provides a method 200 for assigning nucleic acid sequence reads to target polynucleotides. The method 200 includes providing a substrate having transposome complexes immobilized thereon in step 202. The transposome complexes include a transposase and a first polynucleotide including an end sequence and a first tag in some embodiments. The method 200 then includes contacting the transposome complexes in step 204 with target polynucleotides under conditions to fragment the target polynucleotides. The method 200 then includes amplifying the fragmented target polynucleotides in step 206 to form a plurality of nucleic acid clusters on the substrate. The method 200 then includes obtaining location information in step 208 for the plurality of nucleic acid clusters on the substrate. After the location information from step 208 has been obtained, the method 200 then includes determining the nucleic acid sequence reads of the fragmented nucleic acids in each of the nucleic acid clusters in step 210. The method 200 then includes assigning the nucleic acid sequence reads to the target polynucleotides using the obtained location information in step 212. In some embodiments, a length of the target polynucleotides is greater than a length of the fragment.

[0054] In some embodiments, assigning the nucleic acid sequence reads in step 212 includes determining the distance between each of the clusters and using the determined distance to assign reads to a specific target polynucleotide. In some embodiments, assigning the nucleic acid sequence reads includes determining for a likelihood score that sequence reads from at least a first and a second cluster on the substrate derive from the same target polynucleotide. In some embodiments, sequence reads from more than two clusters are determined to be derived from the same target polynucleotide based on the likelihood score. In some embodiments, the method further includes increasing the likelihood score for sequence reads from at least the first cluster when the spatial distance and a genomic distance between the reads derived from first and second clusters are below a threshold value. In some embodiments, assigning the nucleic acid sequence reads to a particular target polynucleotide includes determining for a likelihood score that a plurality of sequence reads from clusters on the substrate derive from the same target polynucleotide. In some embodiments, the method further includes increasing the likelihood score for sequence reads from one cluster when the spatial distance and a genomic distance between one cluster and one or more other clusters are below a threshold value.

[0055] In some embodiments, the method further includes determining whether the target nucleic acid has a variant when the spatial distance between at least the first and second clusters are below a threshold value and when the genomic distance is above the genomic distance threshold. In some embodiments, the method further includes determining whether the target nucleic acid has a variant when the spatial distances between the first cluster and the second cluster, and the first cluster and one or more other clusters are below a threshold value and when the genomic distance between the genomic locations of reads derived from the first cluster and the second cluster is above the genomic distance threshold.

[0056] In some embodiments, the spatial distance threshold from a cluster forms a pattern of an ellipse or a circle around the cluster. Other patterns are contemplated including, but not limited to, symmetrical patterns, asymmetrical patterns, rectilinear patterns, hexagonal patterns, or the like. In some embodiments, the likelihood score includes a score representing the likelihood that a read from a particular cluster maps to a target polynucleotide. This likelihood score may refer to a mapping quality (MAPQ) score, e.g., likelihood that the mapping location of a read to a particular target polynucleotide is correct. In some embodiments, the likelihood score of a read from a first cluster is 0. In some embodiments, the likelihood score of a read from second cluster is above 30. In some embodiments, the likelihood score of reads from clusters other than the first and second clusters is above 30. In some embodiments, the location information includes a first spatial coordinate and a second spatial coordinate in a 2D coordinate system. In some embodiments, the method further includes sorting the plurality reads determined from the nucleic acid clusters by their spatial coordinates. In some embodiments, the method further includes sorting each read from each cluster by the read's MAPQ score.

[0057] In some embodiments, the transposome complexes include a second polynucleotide including a region complementary to the transposon end sequence. In some embodiments, the transposome complexes are present on the substrate at a density of at least 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 , 10^9 , or 10^{10} or more complexes per mm^2 . In some embodiments, the transposome complexes include a hyperactive Tn5 transposase.

[0058] In some embodiments, the likelihood score is influenced by a pitch of the substrate, size of the substrate, a pattern of the substrate, temperature, input nucleic acid size distribution, or a combination thereof. In some embodiments, the likelihood score is influenced

by a pitch of the substrate, size of the substrate, a pattern of the substrate, temperature, loading density, fragment directionality, or a combination thereof. In some embodiments, the substrate includes microparticles. In some embodiments, the substrate includes a patterned surface. In some embodiments, the substrate includes wells. In some embodiments, the substrate includes a flow cell.

[0059] FIG. 3A shows a method 300 of updating a standard BAM file using spatial information to disambiguate reads which are mapping to more than one target polynucleotide. As illustrated, the standard BAM file includes both uniquely mapped reads and multi-mapped reads, where the read could be assigned to more than one target polynucleotide. The uniquely mapped reads may be used for variant calling without any further processing. Such variant calls are stored in a variant call file ("vcf"). However, multi-mapped reads where a read cannot be assigned to a single target polynucleotide are not suitable for variant calling due to the ambiguity of whether such reads are mapped to a particular target (reference) sequence. In one example, multi-mapped reads are reads that are mapped to more than one target polynucleotide because the reads have not been resolved adequately. As shown, unique sites within the target polynucleotide or genome can be leveraged to resolve a particular read.

[0060] Similarly, some multi-mapped reads may be linked to other multi-mapped reads. In this circumstance, the system may use differentiating sites to correct this issue. The system may also look to disambiguate some multi-mapped reads by using spatial information of the particular read and the possible multi-mapped target polynucleotides to determine if one of the target polynucleotides is more likely than others to be the correctly mapped sequence based on the spatial location on the flow cell of each read that is attributed to the target polynucleotide. If the multi-mapped read is disambiguated, then the system may perform variant calling and update the BAM file to include the correct mapping location of the particular multi-mapped read. In some embodiments, the system updates the mapping quality (MAPQ) score for the particular multi-mapped read to indicate that the mapping quality has increased if the system has determined the correct target polynucleotide based on the spatial location information. The updated MAPQ score indicates the disambiguation of a multi-mapped read. This enlarges the pool of available reads (updated MAPQ reads from one or more of the previously identified multi-mapped reads and the uniquely mapped reads) for variant calling, the output of which is stored in a vcf file.

[0061] FIG. 3B shows a method 350 of updating a standard BAM file using spatial information to disambiguate reads which are mapped to more than one target polynucleotide, similarly to the method 300 shown in FIG. 3A. However, in a step 355 of FIG. 3B, the method 350 may update a MAPQ value and assign a unique position to MAPQ0 reads, even when the read may be mapped to a segmental duplication. For a multi-mapped read which is mapped to a segmental duplication, there are at least two homologous regions where the read may be matched. In a non-limiting example, proximal reads with High MAPQ (>30) are fetched from these homologous regions, e.g., high MAPQ reads from flanking regions, 75kb or more in length, of the multi-mapped read and then all alternative alignment locations are saved. For each MAPQ0 read, the proximal high MAPQ reads are searched on the flow cell (max flow cell distance = 50kb). If proximal high MAPQ reads are found, then the distance between them and all alternative alignments are measured. The mean value of the distances when more than one proximal link read is found is kept. The MAPQ score of the multi-reads is updated only when one of the alternative alignments has an average smaller than 50kb, otherwise the multi-mapped read is left as MAPQ0. When the distance of all of the alternative alignments is greater than 50kb, or more than one of the alternative alignments is less than 50kb, then the MAPQ score of the multi-mapped read remains zero. Distances greater or shorter than 50 kb are also contemplated.

[0062] This may be understood more completely with reference to FIG. 4A which shows the position of several clusters on a flow cell and the potential location of those reads on the target polynucleotide or genome. As shown, the unknown multi-mapped read is indicated as having a MAPQ of 0 (MAPQ0) due to the uncertainty that the indicated alignment is correct. In the upper alignment on the genome, the MAPQ0 read is shown as having a distance on the genome of less than an average of 50kb from the other reads having MAPQ scores of 30, the latter of which are found to also be geographically nearby to the MAPQ0 read. Because the distance $d1$ is less than 50kb, the system determines that there is high likelihood that this assignment of the MAPQ0 read to this position on the genome is accurate, and so the MAPQ0 score is updated to a MAPQ score of 20. In this non-limiting example, the system can update the MAPQ score based on various factors to indicate that the assignment to this position on the genome is likely to be correct.

[0063] As shown on the bottom genomic diagram, the MAPQ0 read is possibly located within 50kb of a set of three other MAPQ30 reads and also possibly a distance d2 of more than 50kb from the same reads. Given this disparity in the read distances, the system determines that the proper assignment of this read is the position within 50kb of the other MAPQ30 reads and the MAPQ score can be updated to be a MAPQ of 20 to indicate that the assignment to this position is likely to be correct.

[0064] As shown in FIG 4B, it may not always be possible to update the MAPQ score of each multi-mapped read. For example, as shown in the clusters on the flow cell in FIG. 4B, the MAPQ0 read may be located spatially very near to other reads on the flow cell (distances greater or shorter than 50 flow cell units are contemplated). In a non-limiting example, the threshold for nearness is 100 flow cell units. However, as shown in the upper genome diagram of FIG. 4B, the average distance for each of those reads on the genome is more than 50kb, where $d1 > 50\text{kb}$. In this circumstance, it's not clear that the MAPQ0 read is correctly assigned to this portion of the genome and so the MAPQ score would stay as zero.

[0065] Similarly, for the lower genome diagram in FIG. 4B, the MAPQ0 read may be positioned near other reads on the flow cell as shown in the clusters thereon, but the multiple possible locations of the MAPQ0 read on the lower genome diagram of the Figure make it difficult to assign one location on the genome. For example, multiple MAPQ0 reads may be located less than 50kb from the other MAPQ30 reads in either of their possible locations (e.g., $d1 < 50\text{kb}$ and $d2 < 50\text{kb}$). Thus, assigning the MAPQ0 reads to one location or the other is a challenge. In another possible scenario, both of the possible MAPQ0 reads are located more than 50kb away from the other MAPQ0 reads on the genome, which makes it a challenge to assign the MAPQ0 read to one position or the other. A dynamic distance threshold is contemplated based multiple factors, including (but not limited to) number of links, length, directions, etc.

[0066] In some embodiments, a flow cell 500 includes a plurality of lanes 510 as shown in FIG. 5. Each lane 510 includes a plurality of surfaces. As shown, in some embodiments of the flow cell 500, a lane includes a top surface 512 and a bottom surface 514. In some embodiments, each surface is subdivided into a plurality of tiles 520. As shown, a cluster 530 is located on a tile 520 that is designated as 1201. This designation serves as an illustrative example only and is not limited to the alphanumeric characters shown in the figure.

In some embodiments, the tile 520 includes 2D X-Y coordinates as shown to provide the spatial information between clusters. The X-Y coordinates are derived from FQU (fastq units). In some embodiments, the subdivision of the surface into tiles 520 is an artificial separation so that the surface of the flow cell is not separated into physical tiles, but instead the images captured by a camera can be segmented into tiles. As shown, the tiles 520 are subdivided by swath, which is a width of a camera. In some embodiments, the tile 520 denotes the size of an image that can be captured by the camera. In some embodiments, the X-Y coordinates are pixel values. In some embodiments, 1 unit of a tile 520 can be approximated to be $1/10^{\text{th}}$ of a pixel. A physical separation is contemplated in some embodiments where the tile can have physical barriers, wells, other structures which separate one portion of the flow cell from another portion of the flow cell.

[0067] In a non-limiting example shown in FIG. 6, a plurality of clusters 610, 620, 630 are shown in tile 600. As shown, spatial information, including X-Y coordinates, for clusters 610, 620, 630 are obtained by the camera that processes the pixel value of the digital image, the processing of which is shown in the readnames 612, 622, 632 for clusters 610, 620, 630 respectively. By way of example only, such readnames 612, 622, 632 for clusters 610, 620, 630 respectively, shown in FIG. 6, follow the format of: Instrument:Run:Flowcell ID:Lane #:Surface (1 or 2, top or bottom):Swath:Tile-X:Tile-Y. Other formats of the readnames are contemplated so long as the X-Y spatial information is obtained. For readname 612, which includes information for cluster 610, the instrument value is A01298; the run number is 280; the flow cell ID is HYC2MDRX2; number of lanes is 2; for 1201, the identified Surface is 1, the identified Swath is 2, and the identified Tile is 1; the Tile X-coordinate is 26928; and the Tile Y-coordinate is 18349. FIG. 7 shows how the Tile X-Y coordinates can be converted to flow cell X-Y coordinates, where the flow cell X-coordinate = Tile X-coordinate x Tile number; and the flow cell Y-coordinate = Tile Y-coordinate x Swath number.

[0068] In some embodiments, spatial information is used to link reads together. In some embodiments, spatial information is used to link reads together, where the link between the reads can be physical or non-physical. In some embodiments, spatial information is used to link reads together to form a longer linked read with one or more read subpairs. The linked reads have expected properties such as, but not limited to, expected length distribution, distance between pairs, and number of pairs. These properties can be leveraged in genome

analysis. In one non-limiting example, Fragment 1 ----- Fragment 2 are linked using spatial information that confirm Fragments 1 and 2 are from the same polynucleotide. The length of the linked read construct is the length of Fragment 1 and Fragment 2 plus 5 units, where each unit is indicated by “-”. Therefore, the distance between Fragment 1 and 2 is 5 units, such as 5 flow cell units. In another non-limiting example, Fragment 3 ----- Fragment 4 ----- Fragment 5 are linked using spatial information to confirm that Fragments 3, 4, and 5 are from the same polynucleotide. The distance between Fragments 3 and 4 is 5 units, where each unit is indicated by “-”; and the distance between Fragments 4 and 5 is 6 units.

[0069] Linking reads can be performed in a genomic reference dependent or independent manner. For genomic independent linking, links are formed between reads using spatial information only. In a non-limiting example, reads are linked when they are 50 flow cell distance units apart. In this example, reference-based information (including, but not limited to, genomic information and the like) is not considered in linking these reads. For a genomic dependent linking, links are formed between reads using spatial and genomic information. In a non-limiting example of a genomic dependent linking, a pair of reads or multiple reads are linked when they are 50 flowcell distance units apart and within 10 kbp distance when the reads are aligned to a reference genome.

[0070] In some embodiments, a system for assigning sequence reads on a substrate to their original target nucleic acid, the system including a substrate including clusters of amplified fragments of target polynucleotides bound to the substrate in spatial locations; and one or more processors. The one or more processors include instructions that when executed perform a method that includes obtaining spatial location information for the clusters of amplified fragments on the substrate; sequencing the target polynucleotides in each of the clusters to determine a sequence read and spatial location from each of the amplified fragments; determining the geographic distance between each cluster on the substrate; and assigning sequence reads from each cluster to a target nucleic acid based on their geographic distance from one another. In some embodiments, the obtaining of the spatial location information is performed during the sequencing of the target nucleic acids. In some embodiments, an alignment of the target nucleic acids to the target polynucleotides is determined as the sequencing of the target nucleic acids is being performed. In some embodiments the one of

more processors further include instructions that when executed perform a method that includes defining a definitive position of each cluster on the substrate.

[0071] In some embodiments, a length of the target polynucleotides is greater than a length of the fragment. In some embodiments, the assigning of the nucleic acid sequence reads includes determining the distance between each of the clusters and using the determined distance to assign reads to a specific target polynucleotide. In some embodiments, the assigning of the nucleic acid sequence reads includes determining for a likelihood score that at least a first and a second cluster on the substrate derive from the same target polynucleotide.

[0072] In some embodiments, the one or more processors further performs a method including increasing the likelihood score for at least the first cluster when the spatial distance between at least the first and second clusters are below a threshold value. In some embodiments, the one or more processors further performs a method including increasing the likelihood score for at least the first cluster when a genomic distance between at least the first and second clusters are below the threshold value. In some embodiments, the one or more processors further performs a method including increasing the likelihood score for at least the first cluster when the spatial distance and a genomic distance between at least the first and second clusters are below a threshold value. In some embodiments, the likelihood score is influenced by a pitch of the substrate, size of the substrate, a pattern of the substrate, temperature, input nucleic acid size distribution, or a combination thereof. In some embodiments, the substrate includes microparticles. In some embodiments, the substrate includes a patterned surface. In some embodiments, the substrate includes wells. In some embodiments, the substrate includes nanowells. In one non-limiting example, a pitch of the nanowell affects the likelihood score.

[0073] In some embodiments, the one or more processors further perform a method including determining whether the target nucleic acid has a variant when the spatial distance between at least the first and second clusters are below a threshold value and when the genomic distance is above the genomic distance threshold. In some embodiments, the spatial distance threshold from a cluster forms a pattern of an ellipse or a circle around the cluster. In some embodiments, the likelihood score of the first cluster is 0. In some embodiments, the likelihood score of the second cluster is above 30. In some embodiments, the location information includes a first spatial coordinate and a second spatial coordinate in a 2D coordinate system.

In some embodiments, the one or more processors further performs a method further including sorting the plurality of the nucleic acid clusters by their spatial coordinates.

[0074] In some embodiments, transposome complexes include a second polynucleotide including a region complementary to the transposon end sequence. In some embodiments, the transposome complexes are present on the substrate at a density of at least 103, 104, 105, 106 complexes per mm². In some embodiments, the transposome complexes include a hyperactive Tn5 transposase.

EXAMPLE 1

[0075] FIG. 8 provides a non-limiting example 800 of flow cell grid geometry 805 and spatial metrics 810. As shown at grid 805, the nanowells of the flow cell are organized into a grid. The nanowell grid 805 is overlayed with a pixel grid (not shown) which is derived from an image of the flow cell grid 805 by a CCD camera. In this example, 1 pixel in the captured image represents 10 FQU's and each pixel size is 345 nm. Various cameras at different magnification levels will have different pixel sizes and the pixel size disclosed in FIG. 8 is by way of example only. With the pixel size and the pixel to FQU ratio known, the physical FQU size is determined to be 34.5 nm, in this example. Other metrics are obtained with this information includes the nanowell pitch, diameter, and interstitial distance. In this example, the nanowell pitch is shown to be 624 nm, 1.81 pixels, and 18.1 FQU's; the nanowell diameter is 360 nm, 1.04 pixels, 10.4 FQU's; and the nanowell interstitial distance is 282 nm, 0.82 pixel, and 8.2 FQU's. Also shown in this example, the nanowell grid is organized in a hexagonal pattern.

EXAMPLE 2

[0076] FIG. 9 provides a non-limiting example 900 of a relationship between spatial and genomic metrics, the latter with respect to an exemplary linear strand of DNA 905. As shown, based on the information obtained by a camera (e.g., pixel size), the pitch between two nanowells is 642 nm and the FQU is 18. Because the distance between the bases 910 of B-DNA is 0.34 nm, the number of base pairs of linear DNA representing the distance of the pitch is approximately 1,888 bp. Accordingly, 10 nanowells have a pitch of 6,420 nm, which is also approximately 18,880 bp. Likewise, a length from a linear DNA fragment can be

leveraged to determine a nanowell count, wherein 10kb of DNA represents 95 FQU and 3,400 nm, which is 5.2 nanowells.

EXAMPLE 3

[0077] FIG. 10 provides a non-limiting example of a workflow for updating a MAPQ score of a MAPQ0 read. At block 1010, alignment data is obtained, which includes two homologous regions (A and B). The alignment data is obtained from BAM, BED, SEG, or any other file formats. After the alignment data has been obtained, the workflow proceeds to blocks 1020 and 1030. Block 1020 includes fetching high MAPQ (>30) reads from regions A and B (+75kb flanks). Block 1030 includes searching for MAPQ0 reads from regions A and B (+75kb flanks) and saving all possible alternative alignment (AA) locations. Once the search for MAPQ reads has been completed in block 1030, the workflow moves to a decision state 1032 to determine whether a MAPQ0 read has been found. If a MAPQ0 read has been found, then the workflow moves to block 1040 to search for high MAPQ reads fetched from block 1020 that are within a predetermined distance threshold from each MAPQ0 read. In this example, for each MAPQ0 read, the search in block 1040 searches for proximal reads with high MAPQ in regions A and B. For example, in one embodiment, the maximum flow cell distance is equal to 50. If a MAPQ0 read is not found at the decision block 1032, then the workflow moves to the end and the process 1000 terminates. It should be realized that following the fetch of the high MAPQ reads from regions A and B at block 1020, the workflow 1000 moves to the block 1040 to search for high MAPQ reads within a distance threshold from each MAPQ0 read.

[0078] Following the search performed in block 1040, the workflow 1000 moves to a decision state 1042 to determine whether high MAPQ reads that are within the distance threshold from each MAPQ0 read has been found. If such high MAPQ reads are found, then the workflow moves to block 1050 to obtain a genomic distance (GD) between high MAPQ reads and all AAs. If a high MAPQ read is not found, then the workflow moves to the end. In this example, measuring the distance between the high MAPQ reads and all alternative alignments (keep the mean value of the distances when more than one proximal link read is found). Following the obtaining of the GD, the workflow moves to a decision state 1052 to determine whether if one of the AA has an average GD within a GD threshold. In one example,

if one of the alternative alignments has an average GD smaller than 50kb, then the workflow proceeds to block 1060 to update the MAPQ score of the MAPQ0 read. If all of the AA > 50kb or if more than one AA < 50kb and the rest are ≥ 50 kb, then the workflow moves to block 1061 and the MAPQ score of MAPQ0 is not updated.

[0079] In some file formats such as BAM, these files can be rewritten with the updated MAPQ reads, tagged as having been updated, and set as primary alignment.

EXAMPLE 4

[0080] FIG. 11 provides a non-limiting example of a workflow of updating a MAPQ score of a MAPQ0 read. At block 1110, alignment data is obtained, which includes two homologous regions (A and B). The alignment data is obtained from BAM, BED, SEG, or any other file formats. After the alignment data has been obtained, the workflow proceeds to blocks 1120 and 1130. Block 1120 includes fetching high MAPQ (>30) reads from A and B (+75kb flanks). Block 1130 includes searching for MAPQ0 reads from A and B (+75kb flanks) and saving all possible alternative alignment (AA) locations. Once the search in block 1130 is finished, the workflow moves to a decision state 1132 to determine whether a MAPQ0 read has been found. If a MAPQ0 read has been found, then the workflow moves to block 1140 to search for high MAPQ reads fetched from block 1120 that are within a distance threshold from each MAPQ0 read. If a MAPQ0 read is not found, then the workflow moves to block 1135 where the regional window to obtain alignment data is expanded to be greater than the starting +75kb flanks. In this example, for each MAPQ0 read, the search in block 1140 searches for proximal reads with high MAPQ in A and B (max flow cell distance = 50).

[0081] Following the search performed in block 1140, the workflow moves to a decision state 1142 to determine whether high MAPQ reads that are within the distance threshold from each MAPQ0 read has been found. If such high MAPQ reads are found, then the workflow moves to block 1150 to obtain a genomic distance (GD) between high MAPQ reads and all AAs. If a high MAPQ read is not found, then the workflow moves to block 1135 where the regional window to obtain alignment data is expand to be greater than the starting +75kb flanks. In this example, measuring the distance between the high MAPQ reads and all alternative alignments (keep the mean value of the distances when more than one proximal link read is found). Following the obtaining of the GD, the workflow moves to a decision state

1152 to determine whether if one of the AA has an average GD within a GD threshold. In this example, if one of the alternative alignments has an average GD smaller than 50kb, then the workflow proceeds to block 1160 to update the MAPQ score of the MAPQ0 read. If all of the AA > 50kb or if more than one AA < 50kb and the rest are ≥ 50 kb, then the workflow moves to block 1161 and the MAPQ score of MAPQ0 is not updated.

[0082] In some file formats such as BAM, these files can be rewritten with the updated MAPQ reads, tagged as having been updated, and set as primary alignment.

Additional Notes

[0083] Various embodiments of the present disclosure may be a system, a method, and/or a computer program product at any possible technical detail level of integration. The computer program product may include a computer readable storage medium (or mediums) having computer readable program instructions thereon for causing a processor to carry out aspects of the present disclosure.

[0084] For example, the functionality described herein may be performed as software instructions are executed by, and/or in response to software instructions being executed by, one or more hardware processors and/or any other suitable computing devices. The software instructions and/or other executable code may be read from a computer readable storage medium (or mediums). Computer readable storage mediums may also be referred to herein as computer readable storage or computer readable storage devices.

[0085] The computer readable storage medium can be a tangible device that can retain and store data and/or instructions for use by an instruction execution device. The computer readable storage medium may be, for example, but is not limited to, an electronic storage device (including any volatile and/or non-volatile electronic storage devices), a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing. A non-exhaustive list of more specific examples of the computer readable storage medium includes the following: a portable computer diskette, a hard disk, a solid state drive, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), a static random access memory (SRAM), a portable compact disc read-only memory (CD-ROM), a digital versatile disk (DVD), a memory stick, a floppy disk,

a mechanically encoded device such as punch-cards or raised structures in a groove having instructions recorded thereon, and any suitable combination of the foregoing. A computer readable storage medium, as used herein, is not to be construed as being transitory signals per se, such as radio waves or other freely propagating electromagnetic waves, electromagnetic waves propagating through a waveguide or other transmission media (e.g., light pulses passing through a fiber-optic cable), or electrical signals transmitted through a wire.

[0086] Computer readable program instructions described herein can be downloaded to respective computing/processing devices from a computer readable storage medium or to an external computer or external storage device via a network, for example, the Internet, a local area network, a wide area network and/or a wireless network. The network may comprise copper transmission cables, optical transmission fibers, wireless transmission, routers, firewalls, switches, gateway computers, and/or edge servers. A network adapter card or network interface in each computing/processing device receives computer readable program instructions from the network and forwards the computer readable program instructions for storage in a computer readable storage medium within the respective computing/processing device.

[0087] Computer readable program instructions (as also referred to herein as, for example, "code," "instructions," "module," "application," "software application," and/or the like) for carrying out operations of the present disclosure may be assembler instructions, instruction-set-architecture (ISA) instructions, machine instructions, machine dependent instructions, microcode, firmware instructions, state-setting data, configuration data for integrated circuitry, or either source code or object code written in any combination of one or more programming languages, including an object oriented programming language such as Smalltalk, C++, or the like, and procedural programming languages, such as the "C" programming language or similar programming languages. Computer readable program instructions may be callable from other instructions or from itself, and/or may be invoked in response to detected events or interrupts. Computer readable program instructions configured for execution on computing devices may be provided on a computer readable storage medium, and/or as a digital download (and may be originally stored in a compressed or installable format that requires installation, decompression or decryption prior to execution) that may then be stored on a computer readable storage medium. Such computer readable program instructions

may be stored, partially or fully, on a memory device (e.g., a computer readable storage medium) of the executing computing device, for execution by the computing device. The computer readable program instructions may execute entirely on a user's computer (e.g., the executing computing device), partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider). In some embodiments, electronic circuitry including, for example, programmable logic circuitry, field-programmable gate arrays (FPGA), or programmable logic arrays (PLA) may execute the computer readable program instructions by utilizing state information of the computer readable program instructions to personalize the electronic circuitry, in order to perform aspects of the present disclosure.

[0088] Aspects of the present disclosure are described herein with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems), and computer program products according to embodiments of the disclosure. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer readable program instructions.

[0089] These computer readable program instructions may be provided to a processor of a general-purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks. These computer readable program instructions may also be stored in a computer readable storage medium that can direct a computer, a programmable data processing apparatus, and/or other devices to function in a particular manner, such that the computer readable storage medium having instructions stored therein comprises an article of manufacture including instructions which implement aspects of the function/act specified in the flowchart(s) and/or block diagram(s) block or blocks.

[0090] The computer readable program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other device to cause a series of operational steps to be performed on the computer, other programmable apparatus or other device to produce a computer implemented process, such that the instructions which execute on the computer, other programmable apparatus, or other device implement the functions/acts specified in the flowchart and/or block diagram block or blocks. For example, the instructions may initially be carried on a magnetic disk or solid-state drive of a remote computer. The remote computer may load the instructions and/or modules into its dynamic memory and send the instructions over a telephone, cable, or optical line using a modem. A modem local to a server computing system may receive the data on the telephone/cable/optical line and use a converter device including the appropriate circuitry to place the data on a bus. The bus may carry the data to a memory, from which a processor may retrieve and execute the instructions. The instructions received by the memory may optionally be stored on a storage device (e.g., a solid-state drive) either before or after execution by the computer processor.

[0091] The flowchart and block diagrams in the Figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods, and computer program products according to various embodiments of the present disclosure. In this regard, each block in the flowchart or block diagrams may represent a service, module, segment, or portion of instructions, which comprises one or more executable instructions for implementing the specified logical function(s). In some alternative implementations, the functions noted in the blocks may occur out of the order noted in the Figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. In addition, certain blocks may be omitted in some implementations. The methods and processes described herein are also not limited to any particular sequence, and the blocks or states relating thereto can be performed in other sequences that are appropriate.

[0092] It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts or carry out combinations of special purpose hardware and computer instructions. For example, any of the processes, methods, algorithms, elements, blocks,

applications, or other functionality (or portions of functionality) described in the preceding sections may be embodied in, and/or fully or partially automated via, electronic hardware such as application-specific processors (e.g., application-specific integrated circuits (ASICs)), programmable processors (e.g., field programmable gate arrays (FPGAs)), application-specific circuitry, and/or the like (any of which may also combine custom hard-wired logic, logic circuits, ASICs, FPGAs, etc. with custom programming/execution of software instructions to accomplish the techniques).

[0093] Any of the above-mentioned processors, and/or devices incorporating any of the above-mentioned processors, may be referred to herein as, for example, “computers,” “computer devices,” “computing devices,” “hardware computing devices,” “hardware processors,” “processing units,” and/or the like. Computing devices of the above-embodiments may generally (but not necessarily) be controlled and/or coordinated by operating system software, such as Mac OS, iOS, Android, Chrome OS, Windows OS (e.g., Windows XP, Windows Vista, Windows 7, Windows 8, Windows 10, Windows 11, Windows Server, etc.), Windows CE, Unix, Linux, SunOS, Solaris, Blackberry OS, VxWorks, or other suitable operating systems. In other embodiments, the computing devices may be controlled by a proprietary operating system. Conventional operating systems control and schedule computer processes for execution, perform memory management, provide file system, networking, I/O services, and provide a user interface functionality, such as a graphical user interface (“GUI”), among other things.

[0094] Reference throughout the specification to “one example”, “another example”, “an example”, and so forth, means that a particular element (e.g., feature, structure, and/or characteristic) described in connection with the example is included in at least one example described herein, and may or may not be present in other examples. In addition, it is to be understood that the described elements for any example may be combined in any suitable manner in the various examples unless the context clearly dictates otherwise.

[0095] It is to be understood that the ranges provided herein include the stated range and any value or sub-range within the stated range, as if such value or sub-range were explicitly recited. For example, a range from about 2 kbp to about 20 kbp should be interpreted to include not only the explicitly recited limits of from about 2 kbp to about 20 kbp, but also to include individual values, such as about 3.5 kbp, about 8 kbp, about 18.2 kbp, etc., and sub-ranges,

such as from about 5 kbp to about 10 kbp, etc. Furthermore, when “about” and/or “substantially” are/is utilized to describe a value, this is meant to encompass minor variations (up to +/- 10%) from the stated value.

[0096] While several examples have been described in detail, it is to be understood that the disclosed examples may be modified. Therefore, the foregoing description is to be considered non-limiting.

[0097] While certain examples have been described, these examples have been presented by way of example only, and are not intended to limit the scope of the disclosure. Indeed, the novel methods described herein may be embodied in a variety of other forms. Furthermore, various omissions, substitutions and changes in the methods described herein may be made without departing from the spirit of the disclosure. The accompanying claims and their equivalents are intended to cover such forms or modifications as would fall within the scope and spirit of the disclosure.

[0098] Features, materials, characteristics, or groups described in conjunction with a particular aspect, or example are to be understood to be applicable to any other aspect or example described in this section or elsewhere in this specification unless incompatible therewith. All of the features disclosed in this specification (including any accompanying claims, abstract and drawings), and/or all of the steps of any method or process so disclosed, may be combined in any combination, except combinations where at least some of such features and/or steps are mutually exclusive. The protection is not restricted to the details of any foregoing examples. The protection extends to any novel one, or any novel combination, of the features disclosed in this specification (including any accompanying claims, abstract and drawings), or to any novel one, or any novel combination, of the steps of any method or process so disclosed.

[0099] Furthermore, certain features that are described in this disclosure in the context of separate implementations can also be implemented in combination in a single implementation. Conversely, various features that are described in the context of a single implementation can also be implemented in multiple implementations separately or in any suitable sub-combination. Moreover, although features may be described above as acting in certain combinations, one or more features from a claimed combination can, in some cases, be

excised from the combination, and the combination may be claimed as a sub-combination or variation of a sub-combination.

[0100] Moreover, while operations may be depicted in the drawings or described in the specification in a particular order, such operations need not be performed in the particular order shown or in sequential order, or that all operations be performed, to achieve desirable results. Other operations that are not depicted or described can be incorporated in the example methods and processes. For example, one or more additional operations can be performed before, after, simultaneously, or between any of the described operations. Further, the operations may be rearranged or reordered in other implementations. Those skilled in the art will appreciate that in some examples, the actual steps taken in the processes illustrated and/or disclosed may differ from those shown in the figures. Depending on the example, certain of the steps described above may be removed or others may be added. Furthermore, the features and attributes of the specific examples disclosed above may be combined in different ways to form additional examples, all of which fall within the scope of the present disclosure.

[0101] For purposes of this disclosure, certain aspects, advantages, and novel features are described herein. Not necessarily all such advantages may be achieved in accordance with any particular example. Thus, for example, those skilled in the art will recognize that the disclosure may be embodied or carried out in a manner that achieves one advantage or a group of advantages as taught herein without necessarily achieving other advantages as may be taught or suggested herein.

[0102] Conditional language, such as “can,” “could,” “might,” or “may,” unless specifically stated otherwise, or otherwise understood within the context as used, is generally intended to convey that certain examples include, while other examples do not include, certain features, elements, and/or steps. Thus, such conditional language is not generally intended to imply that features, elements, and/or steps are in any way required for one or more examples or that one or more examples necessarily include logic for deciding, with or without user input or prompting, whether these features, elements, and/or steps are included or are to be performed in any particular example.

[0103] Conjunctive language such as the phrase “at least one of X, Y, and Z,” unless specifically stated otherwise, is otherwise understood with the context as used in general to convey that an item, term, etc. may be either X, Y, or Z. Thus, such conjunctive language

is not generally intended to imply that certain examples require the presence of at least one of X, at least one of Y, and at least one of Z.

[0104] Language of degree used herein, such as the terms “approximately,” “about,” “generally,” and “substantially” represent a value, amount, or characteristic close to the stated value, amount, or characteristic that still performs a desired function or achieves a desired result.

[0105] The scope of the present disclosure is not intended to be limited by the specific disclosures of preferred examples in this section or elsewhere in this specification, and may be defined by claims as presented in this section or elsewhere in this specification or as presented in the future. The language of the claims is to be interpreted broadly based on the language employed in the claims and not limited to the examples described in the present specification or during the prosecution of the application, which examples are to be construed as non-exclusive.

WHAT IS CLAIMED IS:

1. A method for assigning nucleic acid sequence reads to target polynucleotides comprising:

providing transposome complexes, wherein the transposome complexes comprise a transposase and a first polynucleotide comprising an end sequence and a first tag;

contacting the transposome complexes with target polynucleotides under conditions to fragment the target polynucleotides;

amplifying the fragmented target polynucleotides to form a plurality of nucleic acid clusters on a substrate;

obtaining location information for the plurality of nucleic acid clusters on the substrate;

determining the nucleic acid sequence reads of the fragmented nucleic acids in each of the nucleic acid clusters; and

assigning the nucleic acid sequence reads to the target polynucleotides using the obtained location information.

2. The method of claim 1, wherein a length of the target polynucleotides is greater than a length of the fragment.

3. The method of claim 1, wherein assigning the nucleic acid sequence reads comprises determining the distance between each of the clusters and using the determined distance to assign reads to a specific target polynucleotide.

4. The method of claim 3, wherein assigning the nucleic acid sequence reads comprises determining for a likelihood score that at least a first and a second cluster on the substrate derive from the same target polynucleotide.

5. The method of claim 4, further comprising increasing the likelihood score for the first cluster when the spatial distance between at least the first and second clusters are below a threshold value.

6. The method of claim 4, further comprising increasing the likelihood score for the first cluster when a genomic distance between at least the first and second clusters are below a threshold value.

7. The method of claim 5, further comprising increasing the likelihood score for the first cluster when a genomic distance between at least the first and second clusters are below the threshold value.

8. The method of claim 4, further comprising increasing the likelihood score for the first cluster when the spatial distance and a genomic distance between at least the first and second clusters are below a threshold value.

9. The method of any one of claims 5-8, wherein the likelihood score is influenced by a pitch of the substrate, size of the substrate, a pattern of the substrate, temperature, loading density, fragment directionality, or a combination thereof.

10. The method of claim 9, further comprising determining whether the target nucleic acid has a variant when the spatial distance between at least the first and second clusters are below a threshold value and when the genomic distance is above the genomic distance threshold.

11. The method of claim 10, wherein the spatial distance threshold from a cluster forms a pattern of an ellipse or a circle around the cluster.

12. The method of claim 4, wherein the likelihood score of the first cluster is 0.

13. The method of claim 4, wherein the likelihood score of the second cluster is above 30.

14. The method of claim 4, wherein the likelihood score of one or more other clusters is above 30.

15. The method of any one of claims 1-14, wherein the location information comprises a first spatial coordinate and a second spatial coordinate in a cartesian coordinate system.

16. The method of claim 12, further comprising sorting the plurality of the nucleic acid clusters by their spatial coordinates.

17. The method of claim 1, wherein said transposome complexes comprise a second polynucleotide comprising a region complementary to the transposon end sequence.

18. The method of claim 1, wherein the transposome complexes are present on the substrate at a density of at least 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 , 10^9 , or 10^{10} or more complexes per mm^2 .

19. The method of claim 1, wherein said transposome complexes comprise a hyperactive Tn5 transposase.
20. The method of claim 1, wherein the substrate comprises microparticles.
21. The method of claim 1, wherein the substrate comprises a patterned surface.
22. The method of claim 1, wherein the substrate comprises wells.
23. The method of any of claims 1-22, wherein providing the transposome complexes comprises providing the transposome complexes in solution.
24. The method of any of claims 1-23, wherein providing the transposome complexes comprises providing the transposome complexes bound to the substrate.
25. A system for assigning sequence reads on a substrate to their original target nucleic acid, comprising:
 - a substrate comprising clusters of amplified fragments of target polynucleotides bound to the substrate in spatial locations;
 - one or more processors having instructions that when executed perform a method comprising:
 - obtaining spatial location information for the clusters of amplified fragments on the substrate;
 - sequencing the target polynucleotides in each of the clusters to determine a sequence read and spatial location from each of the amplified fragments;
 - determining the geographic distance between each cluster on the substrate; and
 - assigning sequence reads from each cluster to a target nucleic acid based on their geographic distance from one another.
26. The system of claim 25, wherein a length of the target polynucleotides is greater than a length of the fragment.
27. The system of claim 25, wherein assigning the nucleic acid sequence reads comprises determining the distance between each of the clusters and using the determined distance to assign reads to a specific target polynucleotide.

28. The system of claim 27, wherein assigning the nucleic acid sequence reads comprises determining for a likelihood score that at least a first and a second cluster on the substrate derive from the same target polynucleotide.

29. The system of claim 28, wherein the one or more processors further performs a method comprising increasing the likelihood score for the first cluster when the spatial distance between at least the first and second clusters is below a threshold value.

30. The system of claim 28, wherein the one or more processors further performs a method comprising increasing the likelihood score for the first cluster when a genomic distance between at least the first and second clusters is below a threshold value.

31. The system of claim 29, wherein the one or more processors further performs a method comprising increasing the likelihood score for the first cluster when a genomic distance between at least the first and second clusters is below the threshold value.

32. The system of claim 28, wherein the one or more processors further performs a method comprising increasing the likelihood score for the first cluster when the spatial distance and a genomic distance between at least the first and second clusters are below a threshold value.

33. The system of any one of claims 29-32, wherein the likelihood score is influenced by a pitch of the substrate, size of the substrate, a pattern of the substrate, temperature, loading density, fragment directionality, or a combination thereof.

34. The system of claim 28, wherein the one or more processors further performs a method comprising determining whether the target nucleic acid has a variant when the spatial distance between the first and second clusters are below a threshold value and when the genomic distance is above the genomic distance threshold.

35. The system of claim 34, wherein the spatial distance threshold from a cluster forms a pattern of an ellipse or a circle around the cluster.

36. The system of claim 28, wherein the likelihood score of the first cluster is 0.

37. The system of claim 28, wherein the likelihood score of the second cluster is above 30.

38. The system of claim 28, wherein the likelihood score of one or more other clusters is above 30.

39. The system of any one of claims 25-38, wherein the location information comprises a first spatial coordinate and a second spatial coordinate in a 2D coordinate system.

40. The system of any one of claims 28-32, wherein the one or more processors further performs a method comprising sorting the plurality of the nucleic acid clusters by their spatial coordinates.

41. The system of claim 25, wherein said transposome complexes comprise a second polynucleotide comprising a region complementary to the transposon end sequence.

42. The system of claim 25, wherein the transposome complexes are present on the substrate at a density of at least 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 , 10^9 , or 10^{10} or more complexes per mm^2 .

43. The system of claim 25, wherein said transposome complexes comprise a hyperactive Tn5 transposase.

44. The system of claim 25, wherein the substrate comprises microparticles.

45. The system of claim 25, wherein the substrate comprises a patterned surface.

46. The system of claim 25, wherein the substrate comprises wells.