

SYSTEMS AND METHODS FOR IDENTIFYING CELL-ASSOCIATED BARCODES IN MULTI-GENOMIC FEATURE DATA FROM SINGLE-CELL PARTITIONS

INCORPORATION BY REFERENCE

5 **[0001]** The disclosures of any patents, patent applications and publications cited herein are incorporated herein by reference in their entirety.

FIELD

10 **[0002]** The embodiments provided herein are generally related to systems and methods for analysis of genomic nucleic acids and classification of genomic features. Included among embodiments provided herein are systems and methods relating to accurate detection of cell-associated barcodes based on analysis of more than one genomic feature.

BACKGROUND

15 **[0003]** Accurate detection of cell-associated barcodes such as, for example, barcodes from partitions containing one or more cells, is a primary step in the analysis of single-cell molecular datasets from barcoded partitions. Correct cell-calling remains an important challenge for the successful analysis of unbiased genome-wide single-cell molecular datasets. However, this problem becomes even more challenging when addressing the recovery of multiple genomic features from a single cell. Each genomic
20 feature introduces its own unique biochemistry, signal-to-noise profile and measurement inefficiencies. In addition, biology itself can generate diverse combinations of the levels of each genomic feature in a cell.

[0004] As such, there is a need for better classification of cell-associated barcodes in data involving several genomic features from single-cell datasets.

25 SUMMARY

[0005] In accordance with various embodiments, a method is provided for distinguishing cell populations from non-cell populations within a data set, the method comprising receiving a data set at least associated with a plurality of cells, wherein the data set comprises molecule counts of at least two genomic features for each cell; identifying duplicate subsets of data points from the data set; generating deduplicated
30 data by condensing data points from each duplicate subset into a single data point; applying a pre-set threshold to divide the deduplicated data into an initial cell population and a non-cell population, wherein the pre-set threshold is determined using the molecule counts; and generating a refined cell population and a non-cell population by adjusting boundaries of the initial cell population and non-cell population using clustering.

[0006] In accordance with various embodiments, a non-transitory computer-readable medium is provided for storing computer instructions that, when executed by a computer, cause the computer to perform a method for distinguishing cell populations from non-cell populations within a data set, the method comprising: receiving a data set at least associated with a plurality of cells, wherein the data set comprises molecule counts of at least two genomic features for each cell; identifying duplicate subsets of data points from the data set; generating deduplicated data by condensing data points from each duplicate subset into a single data point; applying a pre-set threshold to divide the deduplicated data into an initial cell population and an initial non-cell population, wherein the pre-set threshold is determined using the molecule counts; and generating a refined cell population and a refined non-cell population by adjusting boundaries of the initial cell population and the initial non-cell population using clustering.

[0007] In accordance with various embodiments, a system is provided for distinguishing cell populations from non-cell populations within a data set, comprising a data store configured to store a data set at least associated with a plurality of cells, wherein the data set comprises molecule counts of at least two genomic features for each cell; and a computing device communicatively connected to the data store and configured to receive the data set, the computing device comprising a clustering engine configured to identify duplicate subsets of data points from the data set; generate deduplicated data by condensing data points from each duplicate subset into a single data point; apply a pre-set threshold to divide the deduplicated data into an initial cell population and an initial non-cell population, wherein the pre-set threshold is determined using the molecule counts; and generate a refined cell population and a refined non-cell population by adjusting boundaries of the initial cell population and initial non-cell population using clustering; and a display communicatively connected to the computing device and configured to display a report comprising the refined cell population and refined non-cell population.

[0008] These and other aspects and implementations are discussed in detail herein. The foregoing information and the following detailed description include illustrative examples of various aspects and implementations, and provide an overview or framework for understanding the nature and character of the claimed aspects and implementations. The drawings provide illustration and a further understanding of the various aspects and implementations, and are incorporated in and constitute a part of this specification.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] For a more complete understanding of the principles disclosed herein, and the advantages thereof, reference is now made to the following descriptions taken in conjunction with the accompanying drawings. The accompanying drawings are not intended to be drawn to scale. Like reference numbers and designations in the various drawings indicate like elements. For purposes of clarity, not every component may be labeled in every drawing. In the drawings:

[0010] FIGS. 1A and 1B are schematic illustrations of non-limiting examples of the sequencing workflow for using single cell targeted gene expression sequencing analysis to generate sequencing data for analyzing the expression profile of targeted genes of interest, in accordance with various embodiments.

5 [0011] FIG. 2 is an exemplary flowchart showing a process flow for conducting sequencing data analysis, in accordance with various embodiments.

[0012] FIG. 3 is an exemplary flowchart showing a process flow for generating cell and non-cell populations for joint cell calling, in accordance with various embodiments.

[0013] FIG. 4 is another exemplary flowchart showing a process flow for generating cell and non-cell populations for joint cell calling, in accordance with various embodiments.

10 [0014] FIG. 5 is a schematic diagram of non-limiting examples of a system for generating cell and non-cell populations for joint cell calling, in accordance with various embodiments.

[0015] FIG. 6 are plots depicting an effect of deduplication on barcodes with RNA counts from a particular single-cell sample (plotted on the y-axis) versus associated ATAC counts from the same particular single-cell sample (plotted on the x-axis), in accordance with various embodiments.

15 [0016] FIG. 7 is a plot depicts the initial ordmag classification of deduplicated data, in accordance with various embodiments.

[0017] FIG. 8 is a plot depicting an effect of K-means refinement on barcodes with RNA counts from a particular single-cell sample (plotted on the y-axis) versus associated ATAC counts from the same particular single-cell sample (plotted on the x-axis), in accordance with various embodiments.

20 [0018] FIG. 9 is a plot depicting an effect of the joint cell calling method on sensitivity (plotted on the y-axis) and GEX reads per cell (plotted on the x-axis) with a change of ATAC depth, in accordance with various embodiments.

[0019] FIG. 10 is a block diagram of non-limiting examples illustrating a computer system for use in performing methods provided herein, in accordance with various embodiments.

25 [0020] It is to be understood that the figures are not necessarily drawn to scale, nor are the objects in the figures necessarily drawn to scale in relationship to one another. The figures are depictions that are intended to bring clarity and understanding to various embodiments of apparatuses, systems, and methods disclosed herein. Wherever possible, the same reference numbers will be used throughout the drawings to refer to the same or like parts. Moreover, it should be appreciated that the drawings are not intended to limit
30 the scope of the present teachings in any way.

[0021] The above-identified figures are provided by way of representation and not limitation. The figures may show simplified or partial views, and the dimensions of elements in the figures may be exaggerated or otherwise not in proportion. In addition, as the terms "on," "attached to," "connected to," "coupled to," or similar words are used herein, one element (e.g., a material, a layer, a substrate, etc.) can

be "on," "attached to," "connected to," or "coupled to" another element regardless of whether the one element is directly on, attached to, connected to, or coupled to the other element or there are one or more intervening elements between the one element and the other element. In addition, where reference is made to a list of elements (e.g., elements a, b, c), such reference is intended to include any one of the listed elements by itself, any combination of less than all of the listed elements, and/or a combination of all of the listed elements. Section divisions in the specification are for ease of review only and do not limit any combination of elements discussed.

DETAILED DESCRIPTION

[0022] The following description of various embodiments is exemplary and explanatory only and is not to be construed as limiting or restrictive in any way. Other embodiments, features, objects, and advantages of the present teachings will be apparent from the description and accompanying drawings, and from the claims.

[0023] Provided herein are methods and systems for identifying a subset of cells from a combined genomic sequence dataset. Such systems can be used, for example, for integration of single-cell transcriptomics and epigenomics. It should be appreciated, however, that although the systems and methods disclosed herein can refer to their application in the integration of single-cell transcriptomics and epigenomics workflows, they are equally applicable to other analogous fields.

[0024] The disclosure, however, is not limited to these exemplary embodiments and applications or to the manner in which the exemplary embodiments and applications operate or are described herein. Moreover, the figures may show simplified or partial views, and the dimensions of elements in the figures may be exaggerated or otherwise not in proportion. In addition, as the terms "on," "attached to," "connected to," "coupled to," or similar words are used herein, one element (e.g., a material, a layer, a substrate, etc.) can be "on," "attached to," "connected to," or "coupled to" another element regardless of whether the one element is directly on, attached to, connected to, or coupled to the other element or there are one or more intervening elements between the one element and the other element. In addition, where reference is made to a list of elements (e.g., elements a, b, c), such reference is intended to include any one of the listed elements by itself, any combination of less than all of the listed elements, and/or a combination of all of the listed elements. Section divisions in the specification are for ease of review only and do not limit any combination of elements discussed.

[0025] It should be understood that any use of subheadings herein is for organizational purposes, and should not be read to limit the application of those subheaded features to the various embodiments herein. Each and every feature described herein is applicable and usable in all the various embodiments discussed herein and that all features described herein can be used in any contemplated combination,

regardless of the specific example embodiments that are described herein. It should further be noted that exemplary description of specific features is used, largely for informational purposes, and not in any way to limit the design, subfeature, and functionality of the specifically described feature.

5 [0026] All publications mentioned herein are incorporated herein by reference for the purpose of describing and disclosing devices, compositions, formulations and methodologies which are described in the publication and which might be used in connection with the present disclosure.

[0027] As used herein, the phrase “genomic feature” refers to one or more defined or specified genome elements or regions or functions thereof. In some instances, the genome elements or regions can have some annotated structure and/or function (e.g., a fragment end or a cutsite, a chromosome, a gene, 10 protein coding sequence, mRNA, lncRNA (long non-coding RNA), tRNA, rRNA, repeat sequence, inverted repeat, miRNA, siRNA, etc.) or be a genetic/genomic variant (e.g., single nucleotide polymorphism/variant, insertion/deletion sequence, copy number variation, inversion, etc.) which denotes one or more nucleotides, genome regions, genes or a grouping of genome regions or genes (in DNA or RNA) that have undergone changes as referenced against a particular species or sub-populations within a particular species due to, for 15 example, mutations, recombination/crossover or genetic drift. In some other instances, the genomic feature can be measured by cell surface protein expression, mutational status, or counts of intronic reads.

[0028] As used herein, the phrase “Assay for Transposase-Accessible Chromatin sequencing” or “ATAC sequencing” refers to a sequencing method that probes DNA accessibility with an artificial transposon, which inserts specific sequences into accessible regions of chromatin. Because the transposase 20 can only insert sequences into accessible regions of chromatin not bound by transcription factors and/or nucleosomes, sequencing reads can be used to infer regions of increased chromatin accessibility.

[0029] As used herein, “substantially” means sufficient to work for the intended purpose. The term “substantially” thus allows for minor, insignificant variations from an absolute or perfect state, dimension, measurement, result, or the like such as would be expected by a person of ordinary skill in the 25 field but that do not appreciably affect overall performance. When used with respect to numerical values or parameters or characteristics that can be expressed as numerical values, “substantially” means within one, two, three, four, five, six, seven, nine, or ten percent.

[0030] The term “ones” means more than one.

[0031] As used herein, the term “plurality” can be 2, 3, 4, 5, 6, 7, 8, 9, 10, or more.

30 [0032] As used herein, the terms “comprise”, “comprises”, “comprising”, “contain”, “contains”, “containing”, “have”, “having”, “include”, “includes”, and “including” and their variants are not intended to be limiting, are inclusive or open-ended and do not exclude additional, un-recited additives, components, integers, elements or method steps. For example, a process, method, system, composition, kit, or apparatus

that comprises a list of features is not necessarily limited only to those features but may include other features not expressly listed or inherent to such process, method, system, composition, kit, or apparatus.

5 [0033] Where values are described as ranges, it will be understood that such disclosure includes the disclosure of all possible sub-ranges within such ranges, as well as specific numerical values that fall within such ranges irrespective of whether a specific numerical value or specific sub-range is expressly stated.

10 [0034] Unless otherwise defined, scientific and technical terms used in connection with the present teachings described herein shall have the meanings that are commonly understood by those of ordinary skill in the art. Further, unless otherwise required by context, singular terms shall include pluralities and plural terms shall include the singular. Generally, nomenclatures utilized in connection with, and techniques of, cell and tissue culture, molecular biology, and protein and oligo- or polynucleotide chemistry and hybridization described herein are those well-known and commonly used in the art. Standard techniques are used, for example, for nucleic acid purification and preparation, chemical analysis, recombinant nucleic acid, and oligonucleotide synthesis. Enzymatic reactions and purification techniques are performed according to manufacturer's specifications or as commonly accomplished in the art or as described herein. Standard molecular biological techniques and procedures described herein are generally performed according to conventional methods well known in the art and as described in various general and more specific references that are cited and discussed throughout the instant specification. See, e.g., Sambrook et al., Molecular Cloning: A Laboratory Manual (Third ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. 2000). The nomenclatures utilized in connection with, and the laboratory procedures and standard techniques described herein are those well-known and commonly used in the art.

25 [0035] A "polynucleotide", "nucleic acid", or "oligonucleotide" refers to a linear polymer of nucleosides (including deoxyribonucleosides, ribonucleosides, or analogs thereof) joined by internucleosidic linkages. Typically, a polynucleotide comprises at least three nucleosides. Usually oligonucleotides range in size from a few monomeric units, e.g. 3-4, to several hundreds of monomeric units. Whenever a polynucleotide such as an oligonucleotide is represented by a sequence of letters, such as "ATGCCTG," it will be understood that the nucleotides are in 5'->3' order from left to right and that "A" denotes deoxyadenosine, "C" denotes deoxycytidine, "G" denotes deoxyguanosine, and "T" denotes thymidine, unless otherwise noted. The letters A, C, G, and T may be used to refer to the bases themselves, 30 to nucleosides, or to nucleotides comprising the bases, as is standard in the art.

[0036] DNA (deoxyribonucleic acid) is a chain of nucleotides containing 4 types of nucleotides; A (adenine), T (thymine), C (cytosine), and G (guanine), and RNA (ribonucleic acid) is comprised of 4 types of nucleotides; A, U (uracil), G, and C. Certain pairs of nucleotides specifically bind to one another in a complementary fashion (called complementary base pairing). That is, adenine (A) pairs with thymine

(T) (in the case of RNA, however, adenine (A) pairs with uracil (U)), and cytosine (C) pairs with guanine (G). When a first nucleic acid strand binds to a second nucleic acid strand made up of nucleotides that are complementary to those in the first strand, the two strands bind to form a double strand. As used herein, “nucleic acid sequencing data,” “nucleic acid sequencing information,” “nucleic acid sequence,” “genomic sequence,” “genetic sequence,” or “fragment sequence,” or “nucleic acid sequencing read” denotes any information or data that is indicative of the order of the nucleotide bases (e.g., adenine, guanine, cytosine, and thymine/uracil) in a molecule (e.g., whole genome, whole transcriptome, exome, oligonucleotide, polynucleotide, fragment, etc.) of DNA or RNA. It should be understood that the present teachings contemplate sequence information obtained using all available varieties of techniques, platforms or technologies, including, but not limited to: capillary electrophoresis, microarrays, ligation-based systems, polymerase-based systems, hybridization-based systems, direct or indirect nucleotide identification systems, pyrosequencing, ion- or pH-based detection systems, electronic signature-based systems, etc.

[0037] As used herein, the term “cell” is used interchangeably with the term “biological cell.” Non-limiting examples of biological cells include eukaryotic cells, plant cells, animal cells, such as mammalian cells, reptilian cells, avian cells, fish cells or the like, prokaryotic cells, bacterial cells, fungal cells, protozoan cells, or the like, cells dissociated from a tissue, such as muscle, cartilage, fat, skin, liver, lung, neural tissue, and the like, immunological cells, such as T cells, B cells, natural killer cells, macrophages, and the like, embryos (e.g., zygotes), oocytes, ova, sperm cells, hybridomas, cultured cells, cells from a cell line, cancer cells, infected cells, transfected and/or transformed cells, reporter cells and the like. A mammalian cell can be, for example, from a human, mouse, rat, horse, goat, sheep, cow, primate or the like.

[0038] A term “genome”, as used herein, refers to the genetic material of a cell or organism, including animals, such as mammals, e.g., humans and comprises nucleic acids, such as DNA. In humans, total DNA includes, for example, genes, noncoding DNA and mitochondrial DNA. The human genome typically contains 23 pairs of linear chromosomes: 22 pairs of autosomal chromosomes (autosomes) plus the sex-determining X and Y chromosomes. The 23 pairs of chromosomes include one copy from each parent. The DNA that makes up the chromosomes is referred to as chromosomal DNA and is present in the nucleus of human cells (nuclear DNA). Mitochondrial DNA is located in mitochondria as a circular chromosome, is inherited from only the female parent, and is often referred to as the mitochondrial genome as compared to the nuclear genome of DNA located in the nucleus.

[0039] The term “sequencing,” as used herein, generally refers to methods and technologies for determining the sequence of nucleotide bases in one or more polynucleotides. The polynucleotides can be, for example, nucleic acid molecules such as deoxyribonucleic acid (DNA) or ribonucleic acid (RNA), including variants or derivatives thereof (e.g., single stranded DNA). Sequencing can be performed by

various systems currently available, such as, without limitation, a sequencing system by Illumina®, Pacific Biosciences (PacBio®), Oxford Nanopore®, or Life Technologies (Ion Torrent®). Alternatively or in addition, sequencing may be performed using nucleic acid amplification, polymerase chain reaction (PCR) (e.g., digital PCR, quantitative PCR, or real time PCR), or isothermal amplification. Such systems may provide a plurality of raw genetic data corresponding to the genetic information of a subject (e.g., human), as generated by the systems from a sample provided by the subject.

[0040] In some examples, such systems provide “sequencing reads” (also referred to as “fragment sequence reads” or “reads” herein). A read may include a string of nucleic acid bases corresponding to a sequence of a nucleic acid molecule that has been sequenced. In some situations, systems and methods provided herein may be used with proteomic information.

[0041] The phrase “next generation sequencing” (NGS) refers to sequencing technologies having increased throughput as compared to traditional Sanger- and capillary electrophoresis-based approaches, for example with the ability to generate hundreds of thousands of relatively small sequence reads at a time. Some examples of next generation sequencing techniques include, but are not limited to, sequencing by synthesis, sequencing by ligation, and sequencing by hybridization. More specifically, the MISEQ, HISEQ and NEXTSEQ Systems of Illumina and the Personal Genome Machine (PGM), Ion Torrent, and SOLiD Sequencing System of Life Technologies Corp, provide massively parallel sequencing of whole or targeted genomes. The SOLiD System and associated workflows, protocols, chemistries, etc. are described in more detail in PCT Publication No. WO 2006/084132, entitled “Reagents, Methods, and Libraries for Bead-Based Sequencing,” international filing date Feb. 1, 2006, U.S. patent application Ser. No. 12/873,190, entitled “Low-Volume Sequencing System and Method of Use,” filed on Aug. 31, 2010, and U.S. patent application Ser. No. 12/873,132, entitled “Fast-Indexing Filter Wheel and Method of Use,” filed on Aug. 31, 2010, the entirety of each of these applications being incorporated herein by reference thereto.

[0042] The term “read” or “sequencing read” with reference to nucleic acid sequencing refers to the sequence of nucleotides determined for a nucleic acid fragment that has been subjected to sequencing, such as, for example, next generation sequencing (“NGS”). Reads can be any a sequence of any number of nucleotides which defines the read length.

[0043] Methods of processing and sequencing nucleic acids in accordance with the methods and systems described in the present application are also described in further detail in U.S. Ser. Nos. 14/316,383; 14/316,398; 14/316,416; 14/316,431; 14/316,447; and 14/316,463 which are herein incorporated by reference in their entirety for all purposes and in particular for all written description, figures and working examples directed to processing nucleic acids and sequencing and other characterizations of genomic material.

[0044] The term “barcode,” as used herein, generally refers to a label, or identifier, that conveys or is capable of conveying information about an analyte. A barcode can be part of an analyte. A barcode can be independent of an analyte. A barcode can be a tag attached to an analyte (e.g., nucleic acid molecule) or a combination of the tag in addition to an endogenous characteristic of the analyte (e.g., size of the analyte or end sequence(s)). A barcode may be unique. Barcodes can have a variety of different formats. For example, barcodes can include barcode sequences, such as: polynucleotide barcodes; random nucleic acid and/or amino acid sequences; and synthetic nucleic acid and/or amino acid sequences. A barcode can be attached to an analyte in a reversible or irreversible manner. A barcode can be added to, for example, a fragment of a deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) sample before, during, and/or after sequencing of the sample. Barcodes can allow for identification and/or quantification of individual sequencing reads.

[0045] As used herein, the term “cell barcode” refers to any barcodes that have been determined to be associated with a cell, as determined by a “cell calling” step within various embodiments of the disclosure. For example, the cell barcode can be a known nucleotide sequence, which serves as a unique identifier for a single cell partition, such as a single GEM droplet or well. Each cell barcode can contain reads from a single cell.

[0046] As used herein, the term “Gel bead-in-Emulsion” or “GEM” refers to a droplet containing some sample volume and a barcoded gel bead, forming an isolated reaction volume. When referring to the subset of the sample contained in the droplet, the term “partition” may also be used. In various embodiments within the disclosure when used with GEM, the term “barcode” refers to a known nucleotide sequence or a known combination of several nucleotide sequences, which serve as a unique identifier for a single GEM droplet. Each barcode usually contains reads from a single cell. For example, a barcode can comprise one, two, three, four, five, or more known barcode sequences. These barcode sequences attached to the same GEM can be the same or different, but a combination of these barcodes will be unique for the same GEM and be different from another combination of barcode sequences on different GEMs. For example, each GEM has a ATAC DNA barcode oligonucleotide and a gene expression barcode oligonucleotide attached. Although the ATAC DNA barcode oligonucleotide and the gene expression barcode oligonucleotide may be different, they are designed to have a known association, so each genomic feature receives a cell-associated barcode that may comprise a pair of barcode sequences. In alternative embodiments, the ATAC DNA barcode oligonucleotide and the gene expression barcode oligonucleotide may be the same.

[0047] As used herein, the term “GEM well” or “GEM group” refers to a set of partitioned cells (i.e., Gel beads-in-Emulsion or GEMs) from a single 10x Chromium™ Chip channel. One or more sequencing libraries can be derived from a GEM well.

[0048] The terms “adaptor(s)”, “adapter(s)” and “tag(s)” may be used synonymously. An adaptor or tag can be coupled to a polynucleotide sequence to be “tagged” by any approach, including ligation, hybridization, or other approaches. In various embodiments within the disclosure, the term adapter can refer to customized strands of nucleic acid base pairs created to bind with specific nucleic acid sequences, e.g., sequences of DNA.

[0049] The term “bead,” as used herein, generally refers to a particle. The bead may be a solid or semi-solid particle. The bead may be a gel bead. The gel bead may include a polymer matrix (e.g., matrix formed by polymerization or cross-linking). The polymer matrix may include one or more polymers (e.g., polymers having different functional groups or repeat units). Polymers in the polymer matrix may be randomly arranged, such as in random copolymers, and/or have ordered structures, such as in block copolymers. Cross-linking can be via covalent, ionic, or inductive, interactions, or physical entanglement. The bead may be a macromolecule. The bead may be formed of nucleic acid molecules bound together. The bead may be formed via covalent or non-covalent assembly of molecules (e.g., macromolecules), such as monomers or polymers. Such polymers or monomers may be natural or synthetic. Such polymers or monomers may be or include, for example, nucleic acid molecules (e.g., DNA or RNA). The bead may be formed of a polymeric material. The bead may be magnetic or non-magnetic. The bead may be rigid. The bead may be flexible and/or compressible. The bead may be disruptable or dissolvable. The bead may be a solid particle (e.g., a metal-based particle including but not limited to iron oxide, gold or silver) covered with a coating comprising one or more polymers. Such coating may be disruptable or dissolvable.

[0050] The term “macromolecule” or “macromolecular constituent,” as used herein, generally refers to a macromolecule contained within or from a biological particle. The macromolecular constituent may comprise a nucleic acid. In some cases, the biological particle may be a macromolecule. The macromolecular constituent may comprise DNA. The macromolecular constituent may comprise RNA. The RNA may be coding or non-coding. The RNA may be messenger RNA (mRNA), ribosomal RNA (rRNA) or transfer RNA (tRNA), for example. The RNA may be a transcript. The RNA may be small RNA that are less than 200 nucleic acid bases in length, or large RNA that are greater than 200 nucleic acid bases in length. Small RNAs may include 5.8S ribosomal RNA (rRNA), 5S rRNA, transfer RNA (tRNA), microRNA (miRNA), small interfering RNA (siRNA), small nucleolar RNA (snoRNAs), Piwi-interacting RNA (piRNA), tRNA-derived small RNA (tsRNA) and small rDNA-derived RNA (srRNA). The RNA may be double-stranded RNA or single-stranded RNA. The RNA may be circular RNA. The macromolecular constituent may comprise a protein. The macromolecular constituent may comprise a peptide. The macromolecular constituent may comprise a polypeptide.

[0051] The term “molecular tag,” as used herein, generally refers to a molecule capable of binding to a macromolecular constituent. The molecular tag may bind to the macromolecular constituent with high

affinity. The molecular tag may bind to the macromolecular constituent with high specificity. The molecular tag may comprise a nucleotide sequence. The molecular tag may comprise a nucleic acid sequence. The nucleic acid sequence may be at least a portion or an entirety of the molecular tag. The molecular tag may be a nucleic acid molecule or may be part of a nucleic acid molecule. The molecular tag may be an oligonucleotide or a polypeptide. The molecular tag may comprise a DNA aptamer. The molecular tag may be or comprise a primer. The molecular tag may be, or comprise, a protein. The molecular tag may comprise a polypeptide. The molecular tag may be a barcode.

[0052] The term “partition,” as used herein, generally, refers to a space or volume that may be suitable to contain one or more species or conduct one or more reactions. A partition may be a physical compartment, such as a droplet or well. The partition may isolate space or volume from another space or volume. The droplet may be a first phase (e.g., aqueous phase) in a second phase (e.g., oil) immiscible with the first phase. The droplet may be a first phase in a second phase that does not phase separate from the first phase, such as, for example, a capsule or liposome in an aqueous phase. A partition may comprise one or more other (inner) partitions. In some cases, a partition may be a virtual compartment that can be defined and identified by an index (e.g., indexed libraries) across multiple and/or remote physical compartments. For example, a physical compartment may comprise a plurality of virtual compartments.

[0053] The term “subject,” as used herein, generally refers to an animal, such as a mammal (e.g., human) or avian (e.g., bird), or other organism, such as a plant. For example, the subject can be a vertebrate, a mammal, a rodent (e.g., a mouse), a primate, a simian or a human. Animals may include, but are not limited to, farm animals, sport animals, and pets. A subject can be a healthy or asymptomatic individual, an individual that has or is suspected of having a disease (e.g., cancer) or a pre-disposition to the disease, and/or an individual that is in need of therapy or suspected of needing therapy. A subject can be a patient. A subject can be a microorganism or microbe (e.g., bacteria, fungi, archaea, viruses).

[0054] The term “sample,” as used herein, generally refers to a “biological sample” of a subject. The sample may be obtained from a tissue of a subject. The sample may be a cell sample. A cell may be a live cell. The sample may be a cell line or cell culture sample. The sample can include one or more cells. The sample can include one or more microbes. The biological sample may be a nucleic acid sample or protein sample. The biological sample may also be a carbohydrate sample or a lipid sample. The biological sample may be derived from another sample. The sample may be a tissue sample, such as a biopsy, core biopsy, needle aspirate, or fine needle aspirate. The sample may be a fluid sample, such as a blood sample, urine sample, or saliva sample. The sample may be a skin sample. The sample may be a cheek swab. The sample may be a plasma or serum sample. The sample may be a cell-free or cell free sample. A cell-free sample may include extracellular polynucleotides. Extracellular polynucleotides may be isolated from a bodily sample that may be selected from the group consisting of blood, plasma, serum, urine, saliva,

mucosal excretions, sputum, stool and tears. In some embodiments, the term “sample” can refer to a cell or nuclei suspension extracted from a single biological source (blood, tissue, etc.).

[0055] The sample may comprise any number of macromolecules, for example, cellular macromolecules. The sample may or may include one or more constituents of a cell, but may not include other constituents of the cell. An example of such cellular constituents is a nucleus or an organelle. The sample may be or may include DNA, RNA, organelles, proteins, or any combination thereof. The sample may be or include a chromosome or other portion of a genome. The sample may be or may include a bead (e.g., a gel bead) comprising a cell or one or more constituents from a cell, such as DNA, RNA, a cell nucleus, organelles, proteins, or any combination thereof, from the cell. The sample may be or may include a matrix (e.g., a gel or polymer matrix) comprising a cell or one or more constituents from a cell, such as DNA, RNA, nucleus, organelles, proteins, or any combination thereof, from the cell.

[0056] As used herein, the term “PCR duplicates” refers to duplicates created during PCR amplification. During PCR amplification of the fragments, each unique fragment that is created may result in multiple read-pairs sequenced with near identical barcodes and sequence data. These duplicate reads are identified computationally, and collapsed into a single fragment record for downstream analysis.

SINGLE CELL SEQUENCING AND DATA ANALYSIS WORKFLOWS

Single Cell Sequencing Workflow

[0057] In accordance with various embodiments, sequencing data may be obtained by single-cell sequencing methods such as droplet-based single cell sequencing as discussed below, sci-CAR (Single-cell Combinatorial Indexing Chromatin Accessibility and mRNA; Cao J et al., *Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science* 361: 1380-1385 (2018), incorporated by reference in their entirety), SNARE-seq (Single-Nucleus Chromatin Accessibility and mRNA Expression sequencing; Chen et al., *High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat Biotechnol* 37, 1452–1457 (2019), incorporated by reference in their entirety), or a combination of.

[0058] Any known single cell sequencing methods can be used to provide single cell sequencing data for feature linkage methods and systems in various embodiments. In various embodiments, single cells can be separated into partitions such as droplets or wells, wherein each partition comprises a single cell with a known identifier like a barcode. The barcode can be attached to a support, for example, a bead, such as a solid bead or a gel bead.

[0059] In accordance with various embodiments, a general schematic workflow is provided in FIGS. 1A and 1B to illustrate a non-limiting example process for using single cell sequencing technology to generate sequencing data. Such sequencing data can be used for characterizing cells and cell features in

accordance with various embodiments. The workflow can include various combinations of features, whether it has more or less features than that illustrated in FIGS 1A and 1B. As such, FIGS. 1A and 1B simply illustrates one example of a possible workflow.

Gel beads-in-EMulsion (GEMs) Generation

5 **[0060]** The workflow 100 provided in FIG. 1A begins with Gel beads-in-EMulsion (GEMs) generation. The bulk cell suspension containing the cells is mixed with a gel beads solution 140 or 144 containing a plurality of individually barcoded gel beads 142 or 146. In various embodiments, this step results in partitioning the cells into a plurality of individual GEMs 150, each including a single cell, and a
10 barcoded gel bead 142 or 146. This step also results in a plurality of GEMs 152, each containing a barcoded gel bead 142 or 146 but no nuclei. Detail related to GEM generation, in accordance with various embodiments disclosed herein, is provided below. Further details can be found in US Patent Nos. 10343166 and 10583440, US Published Application Nos. US20180179590A1, US20190367969A1, US20200002763A1, and US20200002764A1, and Published International PCT Application No. WO 2019/040637, each of which is incorporated herein by reference in its entirety.

15 **[0061]** In various embodiments, GEMs can be generated by combining barcoded gel beads, individual cells, and other reagents or a combination of biochemical reagents that may be necessary for the GEM generation process. Such reagents may include, but are not limited to, a combination of biochemical reagents (e.g., a master mix) suitable for GEM generation and partitioning oil. The barcoded gel beads 142 or 146 of the various embodiments herein may include a gel bead attached to oligonucleotides containing
20 (i) an Illumina® P5 sequence (adapter sequence), (ii) a 16 nucleotide (nt) 10x Barcode, and (iii) a Read 1 (Read 1N) sequencing primer sequence. It is understood that other adapter, barcode, and sequencing primer sequences can be contemplated within the various embodiments herein.

[0062] In various embodiments, GEMS are generated by partitioning the cells using a microfluidic chip. To achieve single cell resolution per GEM, the cells can be delivered at a limiting dilution, such that
25 the majority (e.g., ~90-99%) of the generated GEMs do not contain any cells, while the remainder of the generated GEMs largely contain a single cell.

Barcoding Nucleotide Fragments

[0063] The workflow 100 provided in FIG. 1A further includes lysing the cells and barcoding the RNA molecules or fragments for producing a plurality of uniquely barcoded single-stranded nucleic acid
30 molecules or fragments. Upon generation of the GEMs 150, the gel beads 142 or 146 can be dissolved releasing the various oligonucleotides of the embodiments described above, which are then mixed with the RNA molecules or fragments resulting in a plurality of uniquely barcoded single-stranded nucleic acid molecules or fragments 160 following a nucleic acid extension reaction, e.g., reverse transcription of mRNA to cDNA, within the GEMs 150. Detail related to generation of the plurality of uniquely barcoded

single-stranded nucleic acid molecules or fragments 160, in accordance with various embodiments disclosed herein, is provided below.

[0064] In various embodiments, upon generation of the GEMs 150, the gel beads 142 or 146 can be dissolved, and oligonucleotides of the various embodiments disclosed herein, containing a capture sequence, e.g., a poly(dT) sequence or a template switch oligonucleotide (TSO) sequence, a unique molecular identifier (UMI), a unique 10x Barcode, and a Read 1 sequencing primer sequence can be released and mixed with the RNA molecules or fragments and other reagents or a combination of biochemical reagents (e.g., a master mix necessary for the nucleic acid extension process). Denaturation and a nucleic acid extension reaction, e.g., reverse transcription, within the GEMs can then be performed to produce a plurality of uniquely barcoded single-stranded nucleic acid molecules or fragments 160. In various embodiments herein, the plurality of uniquely barcoded single-stranded nucleic acid molecules or fragments 160 can be 10x barcoded single-stranded nucleic acid molecules or fragments. In one non-limiting example of the various embodiments herein, a pool of ~750,000, 10x barcodes are utilized to uniquely index and barcode nucleic acid molecules derived from the RNA molecules or fragments of each individual cell.

[0065] Accordingly, the in-GEM barcoded nucleic acid products of the various embodiments herein can include a plurality of 10x barcoded single-stranded nucleic acid molecules or fragments that can be subsequently removed from the GEM environment and amplified for library construction, including the addition of adaptor sequences for downstream sequencing. In one non-limiting example of the various embodiments herein, each such in-GEM 10x barcoded single-stranded nucleic acid molecule or fragment can include a unique molecular identifier (UMI), a unique 10x barcode, a Read 1 sequencing primer sequence, and a fragment or insert derived from an RNA fragment of the cell, e.g., cDNA from an mRNA via reverse transcription. Additional adaptor sequence may be subsequently added to the in-GEM barcoded nucleic acid molecules after the GEMs are broken.

In various embodiments, after the in-GEM barcoding process, the GEMs 150 are broken and pooled barcoded nucleic acid molecules or fragments are recovered. The 10x barcoded nucleic acid molecules or fragments are released from the droplets, i.e., the GEMs 150, and processed in bulk to complete library preparation for sequencing, as described in detail below. In various embodiments, following the amplification process, leftover biochemical reagents can be removed from the post-GEM reaction mixture. In one embodiment of the disclosure, silane magnetic beads can be used to remove leftover biochemical reagents. Additionally, in accordance with embodiments herein, the unused barcodes from the sample can be eliminated, for example, by Solid Phase Reversible Immobilization (SPRI) beads.

Library Construction

[0066] The workflow 100 provided in FIG. 1A further includes a library construction step. In the library construction step of workflow 100, a library 170 containing a plurality of double-stranded DNA molecules or fragments are generated. These double-stranded DNA molecules or fragments can be utilized for completing the subsequent sequencing step. Detail related to the library construction, in accordance with various embodiments disclosed herein, is provided below.

[0067] In accordance with various embodiments disclosed herein, an Illumina® P7 sequence and P5 sequence (adapter sequences), a Read 2 (Read 2N) sequencing primer sequence, and a sample index (SI) sequence(s) (e.g., i7 and/or i5) can be added during the library construction step via PCR to generate the library 170, which contains a plurality of double stranded DNA fragments. In accordance with various embodiments herein, the sample index sequences can each comprise of one or more oligonucleotides. In one embodiment, the sample index sequences can each comprise of four to eight or more oligonucleotides. In various embodiments, when analyzing the single cell sequencing data for a given sample, the reads associated with all four of the oligonucleotides in the sample index can be combined for identification of a sample. Accordingly, in one non-limiting example, the final single cell gene expression analysis sequencing libraries contain sequencer compatible double-stranded DNA fragments containing the P5 and P7 sequences used in Illumina® bridge amplification, sample index (SI) sequence(s) (e.g., i7 and/or i5), a unique 10x barcode sequence, and Read 1 and Read 2 sequencing primer sequences.

[0068] Various embodiments of single cell sequencing technology within the disclosure can at least include platforms such as One Sample, One GEM Well, One Flowcell; One Sample, One GEM well, Multiple Flowcells; One Sample, Multiple GEM Wells, One Flowcell; Multiple Samples, Multiple GEM Wells, One Flowcell; and Multiple Samples, Multiple GEM Wells, Multiple Flowcells platform. Accordingly, various embodiments within the disclosure can include sequence dataset from one or more samples, samples from one or more donors, and multiple libraries from one or more donors.

Targeted Gene Enrichment by Hybridization Capture

[0069] FIG. 1B depicts an example of a workflow for generating a targeted sequencing library using a hybridization capture approach. As shown, step 153 starts with obtaining a library of double stranded barcoded nucleic acid molecules from single cells (e.g., by partitioning single cells into droplets or wells with barcoding reagents including beads having nucleic acid barcode molecules) is denatured to provide single stranded molecules in step 154. To generate a targeted library using the single stranded molecules, a plurality of oligonucleotide probes designed to cover a panel of selected genes is provided. Each gene in the panel is represented by a plurality of labeled (e.g., biotinylated) oligonucleotide probes, which is allowed to hybridize to the single stranded molecules in step 155 to enrich for genes of interest (e.g. Target 1 and Target 2). To allow for capture, step 155 further includes the addition of supports (e.g.,

beads) that comprise a molecule having affinity for the labels on each labeled oligonucleotide probe. In one embodiment, the oligonucleotide label comprises biotin and the supports comprise streptavidin beads. Following hybridization capture, cleanup steps 156 and 157 are performed (e.g., one or more washing steps to remove unhybridized or off-target library fragments). Captured library fragments are then subjected to nucleic acid extension/amplification to generate a final targeted library for sequencing in step 158. This workflow allows the generation of targeted libraries from gene expression assays. In general, this workflow may be used to enrich any library of fragments having inserts or targets (light gray bar regions) that represent genes, e.g., cDNA transcribed from mRNA of single cells. It should be appreciated, however, that although the description above describes targeted gene enrichment through the use of hybridization capture probes, the methods disclosed herein can also work with other targeted gene enrichment techniques.

[0070] The workflow 100 provided in FIG. 1 further includes a sequencing step. In this step, the library 170 can be sequenced to generate a plurality of sequencing data 180. The fully constructed library 170 can be sequenced according to a suitable sequencing technology, such as a next-generation sequencing protocol, to generate the sequencing data 180. In various embodiments, the next-generation sequencing protocol utilizes the Illumina® sequencer for generating the sequencing data. It is understood that other next-generation sequencing protocols, platforms, and sequencers such as, e.g., MiSeq™, NextSeq™ 500/550 (High Output), HiSeq 2500™ (Rapid Run), HiSeq™ 3000/4000, and NovaSeq™, can be also used with various embodiments herein.

Sequencing Data Input and Data Analysis Workflow

[0071] The workflow 100 provided in FIG. 1 further includes a sequencing data analysis workflow 190. With the sequencing data 180 in hand, the data can then be output, as desired, and used as an input data 185 for the downstream sequencing data analysis workflow 190 for targeted gene expression analysis, in accordance with various embodiments herein. Sequencing the single cell libraries produces standard output sequences (also referred to as the “sequencing data”, “sequence data”, or the “sequence output data”) that can then be used as the input data 185, in accordance with various embodiments herein. The sequence data contains sequenced fragments (also interchangeably referred to as “fragment sequence reads”, “sequencing reads” or “reads”), which in various embodiments include RNA sequences of the targeted RNA fragments containing the associated 10x barcode sequences, adapter sequences, and primer oligo sequences.

[0072] The various embodiments, systems and methods within the disclosure further include processing and inputting the sequence data. A compatible format of the sequencing data of the various embodiments herein can be a FASTQ file. Other file formats for inputting the sequence data is also contemplated within the disclosure herein. Various software tools within the embodiments herein can be employed for processing and inputting the sequencing output data into input files for the downstream data

analysis workflow. One example of a software tool that can process and input the sequencing data for downstream data analysis workflow is the cellranger-atac mkfastq tool within the Cell Ranger™ Targeted Gene Expression analysis pipeline (or the scRNA equivalent Cell Ranger™ analysis tool). It is understood that, various systems and methods with the embodiments herein are contemplated that can be employed to independently analyze the inputted single cell targeted gene expression analysis sequencing data for studying cellular gene expression, in accordance with various embodiments.

Single Cell Sequencing Data Analysis Workflow

[0073] In accordance with various embodiments, a general schematic workflow is provided in FIG. 2 to illustrate a non-limiting example process of a sequencing data analysis workflow for analyzing the single cell sequencing data for gene expression analysis and the single cell ATAC sequencing data for identifying genome-wide differential accessibility of gene regulatory elements. The workflow can include various combinations of features, whether it be more or less features than that illustrated in FIG. 2. As such, FIG. 2 simply illustrates one example of a possible sequencing data analysis workflow.

[0074] FIG. 2 provides an example schematic workflow 200, which is an expansion of the sequencing data analysis workflow 190 of FIG. 1, in accordance with various embodiments. It should be appreciated that the methodologies described in the workflow 200 of FIG. 2 and accompanying disclosure can be implemented independently of the methodologies for generating single cell sequencing data described in FIG. 1. Therefore, FIG. 2 can be implemented independently of the sequencing data generating workflow as long as it is capable of sufficiently analyzing single cell sequencing data for gene expression analysis and identifying genome-wide differential accessibility of gene regulatory elements in accordance with various embodiments.

[0075] In various embodiments, the example data analysis workflow 200 can include one or more of the following analysis steps, a gene expression data processing step 210, an ATAC data processing step 220, a joint cell calling step 230, a gene expression analysis step 240, an ATAC analysis step 250, and an ATAC and gene expression analysis step 260.

[0076] Not all the steps within the disclosure of FIG. 2 need to be utilized as a group. Therefore, some of the steps within FIG. 2 are capable of independently performing the necessary data analysis as part of the various embodiments disclosed herein. Accordingly, it is understood that, certain steps within the disclosure can be used either independently or in combination with other steps within the disclosure, while certain other steps within the disclosure can only be used in combination with certain other steps within the disclosure. Further, one or more of the steps or filters described below, presumably defaulted to be utilized as part of the computational pipeline for analyzing both the gene expression sequencing data and the single cell ATAC sequencing data, can also not be utilized per user input. It is understood that the reverse is also contemplated. It is further understood that additional steps for analyzing the sequencing data generated by

the single cell sequencing workflow are also contemplated as part of the computational pipeline within the disclosure.

Gene Expression Data Processing

5 [0077] The gene expression data processing step 210 can comprise processing the barcodes in the single cell sequencing data set for fixing the occasional sequencing error in the barcodes so that the sequenced fragments can be associated with the original barcodes, thus improving the data quality.

10 [0078] The barcode processing step 210 can include checking each barcode sequence against a “whitelist” of correct barcode sequences. The barcode processing step can further include counting the frequency of each whitelist barcode. The barcode processing step can also include various barcode correction steps as part of the various embodiments disclosed herein. For example, one may attempt to correct the barcodes that are not included on the whitelist by finding all the whitelisted barcodes that are within 2 differences (Hamming distance ≤ 2) of the observed sequence, and then scoring them based on the abundance of that barcode in the read data and quality value of the incorrect bases. As another example, an observed barcode that is not present in the whitelist can be corrected to a whitelist barcode if it has > 90% probability of being the real barcode based.

15 [0079] The gene expression data processing step 210 can further comprise aligning the read sequences (also referred to as the “reads”) to a reference sequence. In the alignment step of the various embodiments herein, a reference-based analysis is performed by aligning the read sequences (also referred to as the “reads”) to a reference sequence. The reference sequence for the various embodiments herein can include a reference transcriptome sequence (including genes and introns) and its associated genome annotations, which include gene and transcript coordinates. The reference transcriptome and annotations of various embodiments herein can be obtained from reputable, well-established consortia, including but not limited to NCBI, GENCODE, Ensembl, and ENCODE. In various embodiments, the reference sequence can include single species and/or multi-species reference sequences. In various embodiments, systems and methods within the disclosure can also provide pre-built single and multi-species reference sequences. In various embodiments, the pre-built reference sequences can include information and files related to regulatory regions including, but not limited to, annotation of promoter, enhancer, CTCF binding sites, and DNase hypersensitivity sites. In various embodiments, systems and methods within the disclosure can also provide building custom reference sequences that are not pre-built.

25 [0080] Various embodiments herein can be configured to correct for sequencing errors in the UMI sequences, before UMI counting. Reads that were confidently mapped to the transcriptome can be placed into groups that share the same barcode, UMI, and gene annotation. If two groups of reads have the same barcode and gene, but their UMIs differ by a single base (i.e., are Hamming distance 1 apart), then one of

the UMIs was likely introduced by a substitution error in sequencing. In this case, the UMI of the less-supported read group is corrected to the UMI with higher support.

5 [0081] After grouping the reads by barcode, UMI (possibly corrected), and gene annotation, if two or more groups of reads have the same barcode and UMI, but different gene annotations, the gene annotation with the most supporting reads is kept for UMI counting, and the other read groups can be discarded. In case of a tie for maximal read support, all read groups can be discarded, as the gene cannot be confidently assigned.

10 [0082] After these two filtering steps, each observed barcode, UMI, gene combination is recorded as a UMI count in an unfiltered feature-barcode matrix, which contains every barcode from fixed list of known-good barcode sequences. This includes background and cell associated barcodes. The number of reads supporting each counted UMI is also recorded in the molecule info file.

15 [0083] The gene expression data processing step 210 can further comprise annotating the individual cDNA fragment reads as exonic, intronic, intergenic, and by whether they align to the reference genome with high confidence. In various embodiments, a fragment read is annotated as exonic if at least a portion of the fragment intersects an exon. In various embodiments, a fragment read is annotated as intronic if it is non-exonic and intersects an intron. The annotation process can be determined by the alignment method and its parameters/settings as performed, for example, using the STAR aligner.

20 [0084] The gene expression data processing step 210 can further comprise unique molecule processing to better identify certain subpopulations such as for example, low RNA content cells, a unique molecule processing step can be performed prior to cell calling. For low RNA content cells, such a step is important, particularly when low RNA content cells are mixed into a population of high RNA content cells. The unique molecule processing can include a high content (e.g., RNA content) capture step and a low content capture step.

ATAC Data Processing

25 [0085] The ATAC data processing step 220 can comprise processing the barcodes in the single cell ATAC sequencing data for fixing the occasional sequencing error in the barcodes so that the sequenced fragments can be associated with the original barcodes, thus improving the data quality.

30 [0086] The barcode processing step can include checking each barcode sequence against a “whitelist” of correct barcode sequences. The barcode processing step can further include counting the frequency of each whitelist barcode. The barcode processing step can also include various barcode correction steps as part of the various embodiments disclosed herein. For example, one may attempt to correct the barcodes that are not included on the whitelist by finding all the whitelisted barcodes that are within 2 differences (Hamming distance ≤ 2) of the observed sequence, and then scoring them based on the abundance of that barcode in the read data and quality value of the incorrect bases. As another example,

an observed barcode that is not present in the whitelist can be corrected to a whitelist barcode if it has > 90% probability of being the real barcode based.

[0087] The ATAC data processing step 220 can further comprise aligning the read sequences (also referred to as the “reads”) to a reference sequence. One of more sub-steps can be utilized for trimming off adapter sequences, primer oligo sequences, or both in the read sequence before the read sequence is aligned to the reference genome.

[0088] The ATAC data processing step 220 can further comprise marking sequencing and PCR duplicates and outputting high quality de-duplicated fragments. One or more sub-steps can be employed for identifying duplicate reads, such as sorting aligned reads by 5' position to account for transposition event and identifying groups of read-pairs and original read-pair. The process may further include filters that, when activated in various embodiments herein, can determine whether a fragment is mapped with MAPQ > 30 on both reads (i.e., includes a barcode overlap for reads with mapping quality below 30), not mitochondrial, and not chimerically mapped.

[0089] The ATAC data processing step 220 can comprise a peak calling analysis that includes counting cut sites in a window around each base-pair of the genome and thresholding it to find regions enriched for open chromatin. Peaks are regions in the genome enriched for accessibility to transposase enzymes. Only open chromatin regions that are not bound by nucleosomes and regulatory DNA-binding proteins (e.g., transcription factors) are accessible by transposase enzymes for ATAC sequencing. Therefore, the ends of each sequenced fragment of the various embodiments herein can be considered to be indicative of a region of open chromatin. Accordingly, the combined signal from these fragments can be analyzed in accordance with various embodiments herein to determine regions of the genome enriched for open chromatin, and thereby, to understand the regulatory and functional significance of such regions. Therefore, using the sites as determined by the ends of the fragments in the position-sorted fragment file (e.g., the fragments.tsv.gz file) described above, the number of transposition events at each base-pair along the genome can be counted. In one embodiment within the disclosure, the cut sites in a window around each base-pair of the genome is counted.

Joint Cell Calling

[0090] The joint cell calling step 230 can comprise a cell calling analysis that includes associating a subset of barcodes observed in both the single cell gene expression library and the single cell ATAC library to the cells loaded from the sample. Identification of these cell barcodes can allow one to then analyze the variation and quantification in data at a single cell resolution. More details of an example workflow for generating cell populations and non-cell populations for joint cell calling will be provided in FIG. 3.

[0091] The process may further include correction of gel bead artifacts, such as gel bead multiples (where a cell shares more than one barcoded gel bead) and barcode multiplets (which occurs when a cell associated gel bead has more than one barcode). In some embodiments, the steps associated with cell calling and correction of gel bead artifacts are utilized together for performing the necessary analysis as part of the various embodiments herein.

[0092] In accordance with various embodiments, the record of mapped high-quality fragments that passed all the filters of the various embodiments disclosed in the steps above and were indicated as a fragment in the fragment file (e.g., the fragments.tsv file), are recorded. With the peaks determined in the peak calling step disclosed herein, the number of fragments that overlap any peak regions, for each barcode, can be utilized to separate the signal from noise, i.e., to separate barcodes associated with cells from non-cell barcodes. It is to be understood that such method of separation of signal from noise works better in practice as compared to naively using the number of fragments per barcode.

[0093] Various methods, in accordance with various embodiments herein, can be employed to for cell calling. In various embodiments, the cell calling can be performed in two steps. In the first step of cell calling of the various embodiments herein, the barcodes that have fraction of fragments overlapping called peaks lower than the fraction of genome in peaks are identified. When this first step is employed in the cell calling process of the various embodiments herein, the peaks are padded by 2000 bp on both sides so as to account for the fragment length for this calculation.

Gene Expression Analysis

[0094] The gene expression analysis step 240 can comprise generating a feature-barcode matrix that summarizes that gene expression counts per each cell. The feature-barcode matrix can include only detected cellular barcodes. The generation of the feature-barcode matrix can involve compiling the valid non-filtered UMI counts per gene (e.g., output from the 'Unique Molecule Processing' step discussed herein) from each cell-associated barcode (e.g., output from the 'Cell Calling step discussed above) together into the final output count matrix, which can then be used for downstream analysis steps.

[0095] The gene expression analysis step 240 can comprise various dimensionality reduction, clustering and t-SNE projection tools. Dimensionality reduction tools of the various embodiments herein are utilized to reduce the number of random variables under consideration by obtaining a set of principal variables. In accordance with various embodiments herein, clustering tools can be utilized to assign objects of the various embodiments herein to homogeneous groups (called clusters) while ensuring that objects in different groups are not similar. T-SNE projection tools of the various embodiments herein can include an algorithm for visualization of the data of the various embodiments herein. In accordance with various embodiments, systems and methods within the disclosure can further include dimensionality reduction, clustering and t-SNE projection tools. In some embodiments, the analysis associated with dimensionality

reduction, clustering, and t-SNE projection for visualization are utilized together for performing the necessary analysis as part of the various embodiments herein. Various analysis tools for dimensionality reduction include Principal Component Analysis (PCA), Latent Semantic Analysis (LSA), and Probabilistic Latent Semantic Analysis (PLSA), clustering, and t-SNE projection for visualization that allow one to group
5 and compare a population of cells with another.

[0096] In some embodiments, the systems and methods within the disclosure are directed to identifying differential gene expression. As the data is sparse at single cell resolution, dimensionality reduction in accordance with various embodiments herein can be performed to cast the data into a lower dimensional space.

10 [0097] In accordance with various embodiments, the gene expression analysis step 240 can comprise a differential expression analysis that performs differential analysis to identify genes whose expression is specific to each cluster, Cell Ranger tests, for each gene and each cluster, whether the in-cluster mean differs from the out-of-cluster mean.

ATAC Analysis

15 [0098] The ATAC analysis step 250 can comprise determining the peak-barcode matrix. In accordance with various embodiments, at ATAC analysis step 250, a raw peak-barcode matrix can be generated first, which is a count matrix consisting of the counts of fragment ends (or cut sites) within each peak region for each barcode. This raw peak-barcode matrix captures the enrichment of open chromatin per barcode. The raw matrix can then be filtered to consist only of cell barcodes by filtering out the non-
20 cell barcodes from the raw peak-barcode matrix, which can then be used in the various dimensionality reduction, clustering and visualization steps of the various embodiments herein.

[0099] The ATAC analysis step 250 can comprise various dimensionality reduction, clustering and t-SNE projection tools, similar to as described above in the gene expression analysis step 240.

25 [0100] The ATAC analysis step 250 can comprise annotating the peaks by performing gene annotations and discovering transcription factor-motif matches on each peak. It is contemplated that peak annotation can be employed with subsequent differential analysis steps within various embodiments of the disclosure. Various peak annotation procedures and parameters are contemplated and are discussed in detail below.

30 [0101] Peaks are regions enriched for open chromatin, and thus have potential for regulatory function. It is therefore understood that observing the location of peaks with respect to genes can be insightful. Various embodiments herein, e.g., bedtools closest -D=b, can be used to associate each peak with genes based on closest transcription start sites (TSS) that are packaged within the reference. In accordance with some embodiments within the disclosure, a peak is associated with a gene if the peak is within 1000 bases upstream or 100 bases downstream of the TSS. Additionally, in accordance with some

embodiments within the disclosure, genes can be associated to putative distal peaks that are much further from the TSS and are less than 100kb upstream or downstream from the ends of the transcript. This association can be adopted by companion visualization software of the various embodiments herein, e.g., Loupe Cell Browser. In another embodiment, this association can be used to construct and visualize derived features such as promoter-sums that can pool together counts from peaks associated with a gene.

[0102] The ATAC analysis step 250 can further comprise a transcription factor (TF) motif enrichment analysis. TF motif enrichment analysis includes generating a TF-barcode matrix consisting of the peak-barcode matrix (i.e., pooled cut-site counts for peaks) having a TF motif match, for each motif and each barcode. It is contemplated that the TF motif enrichment can then be utilized for subsequent analysis steps, such as differential accessibility analysis, within various embodiments of the disclosure. Detail related to TF motif enrichment analysis is provided below.

[0103] The ATAC analysis step 250 can further comprise a differential accessibility analysis that performs differential analysis of TF binding motifs and peaks for identifying differential gene expression between different cells or groups of cells. Various algorithms and statistical models within the disclosure, such as a Negative Binomial (NB2) generalized linear model (GLM), can be employed for the differential accessibility analysis.

ATAC and Gene Expression Analysis

[0104] The ATAC and gene expression analysis step 260 can comprise a feature linkage analysis for detecting correlations between pairs of genomic features, for example, between peaks and genes from single cell datasets. Such correlations can be denoted as feature linkages and can be used for inferring enhancer-gene targeting relationships and constructing transcriptional networks.

[0105] In various embodiments, data from the joint cell calling 230 can be further processed by the ATAC and gene expression analysis step 260 to identify correlations between the single cell gene expression library and the single cell ATAC library.

[0106] The features with strong linkages or correlations are considered to be “co-expressed” and enrich for a shared regulatory mechanism. For example, the accessibility of an enhancer and the expression of its target gene can display a very synchronized differential pattern across a heterogeneous population of cells. A highly accessible enhancer leads to an elevated level of transcription factor (TF) binding, which in turn leads to elevated (or repressed) gene expression. On the other hand, when the enhancer is inaccessible, no TF can bind to the enhancer, and thus transcription activation is at minimum, which leads to reduced target gene expression.

GENERATING CELL AND NON-CELL POPULATIONS FOR JOINT CELL CALLING

[0107] FIG. 3 provides a workflow 300 for generating cell populations and non-cell populations for joint cell calling. In step 310, multi-omic data matrix comprising molecular counts of at least two genomic features for each cell may be received.

5 [0108] In various embodiments, multi-omic data matrix comprising data for two, three, four, five, six or more genomic features for each single cell can be generated, received, or processed. For example, two genomic features per single cell can be measured, wherein the first genomic feature is a gene, and the second genomic feature is an open genomic region. In additional and alternative embodiments, at least three genomic features can be measured, wherein the first genomic feature is a gene, and the second
10 genomic feature is an open genomic region, and the third genomic feature is measured by protein abundance.

[0109] The data matrix can be output by gene expression data processing step 210 and ATAC data processing steps 220 and can be input into joint cell calling step 230, which includes associating a subset of barcodes observed in the library to the cells loaded from the sample based on multi-dimensional data,
15 for example, gene expression data and ATAC data.

[0110] In additional and alternative embodiments, the data matrix can be generated by sci-CAR (Single-cell Combinatorial Indexing Chromatin Accessibility and mRNA; Cao J et al., *Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science* 361: 1380–1385 (2018)) or SNARE-seq (Single-Nucleus Chromatin Accessibility and mRNA Expression sequencing; Chen et al.,
20 *High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat Biotechnol* 37, 1452–1457 (2019)).

[0111] In duplicate identification and deduplication step 320, duplicate data, for example, data associated with duplicate barcodes, can be identified and de-duplicated from the multi-omic data matrix. Non-cell barcodes, especially those with really low molecule counts (e.g., less than 1, 2, 3, 4, 5, 6, 7, 8, 9,
25 10 or any intermediate ranges or values) in either assay, are generally derived from whitelist contamination (e.g., free barcodes present in solution along with the gel beads) and are highly over-represented in the data before de-duplication. In some embodiments, de-duplication collapses these non-cell barcodes into a much smaller number of unique counts, thereby de-emphasizing the data from whitelist contamination.

[0112] Real cells would have orders of magnitude counts more than contamination. Two cell
30 barcodes would rarely have exactly the same or substantially the same counts between at least two genomic features, for example, RNAs and cut sites. In other words, real cell barcodes would not have the same RNA UMI counts and ATAC cutsite counts (e.g., same (x, y) counts on a two-dimensional scatter plot). If some barcodes have exactly the same or substantially the same two genomic features, for example, exactly the

same or substantially the same RNA UMI counts and ATAC cutsite counts, these barcodes can be identified because they are presumed to be contamination, e.g., whitelist contamination.

5 [0113] In various embodiments, the data associated with these identified barcodes can be processed differently from the data not associated with these barcodes in de-duplication. For example, the data associated with these barcodes can be collapsed to a single count of barcode. In additional and alternative aspects, the data associated with these barcodes can be removed from downstream analysis, such as initialization and refining boundaries to generate cell populations and non-cell populations.

10 [0114] Instead of counting data per barcode, after deduplication, one just counts unique pairs of counts. The density distribution of the unique pairs of counts becomes much more representative of the observed (x,y) scatterplot (e.g., x refers to ATAC cutsites in peaks, y refers to UMI counts for RNA) and thereby is more accurately able to represent the separation between cells and non-cells. For example, there may be 12 points that have the same locus (x, y) and deduplication masks 11 of these from the downstream initialization and boundary refinement steps of joint cell calling.

15 [0115] In various embodiments, deduplication allows for suppression of noise without using thresholds or making assumptions about the count distribution profile because noise can vary widely from sample to sample. In additional and alternative embodiments, deduplication functions to de-emphasize whitelist contamination.

20 [0116] In Initialization step 330, deduplicated data may be used for initialization in clustering. In some embodiments, the initialization may involve non-random selections of data points to be clustered. For example, the initialization step 330 may involve determining a threshold that separates cell data from non-cell data, for example, separates cell barcodes from non-cell barcodes. After a threshold is determined, the threshold may be applied to the deduplicated data to generate initial cell populations and non-cell populations.

25 [0117] Initialization step 330 can happen after duplicate identification and deduplication step 320 because duplicate identification and deduplication 320 helps to get a reasonable estimate of a threshold that separates cell and non-cell barcodes using the deduplicated data. After independently defining a threshold for each dimension (i.e. x and y axis) barcodes above both thresholds are classified as cells and the remaining as classified as non-cells. As such, one threshold is set on x, one threshold on y, and each of these use a thresholding step.

30 [0118] In the initialization step 330, for each axis or dimension (e.g., data from each assay), a thresholding method based on an order of magnitude (“ordmag”) may be used for the deduplicated data for initial separation of a cell population and a non-cell population. For example, a threshold value of molecular counts is determined as a threshold to separate the barcodes into initial cell barcodes and initial non-cell barcodes. To determine a threshold, the thresholding can include using a cutoff based on total molecular

counts of each barcode to identify cells. This thresholding can include sorting barcodes by molecular counts (e.g., UMI counts or ATAC cutsites in peaks). With m representing the 99th percentile of the top N sorted barcodes by total UMI counts, the count value of $m/10$ determines a threshold. The percentile is a parameter that can be varied. For example, the percentile can range from about 50% to about 99.9% or any ranges or values derived therefrom. Because of the possible interrelated nature of percentile and fold change parameter (e.g., the $m/10$ value for the denominator - by default equal to 10), an adjustment to the accompanying fold-change parameter can take place. For example, if the top 99st percentile of the top N sorted barcodes by total molecular counts would have a total molecular count of 500, the count value of $m/10$ would be a count of 50, with any barcodes with a molecule count exceeding that value (e.g., a count of 50) being considered meeting the threshold.

[0119] The denominator is a parameter that can be varied. For example, the denominator can range from about 2 to about 50, about 3 to about 20, any iterative ranges in between, about 10, about 20, about 30 and so on.

[0120] In additional and alternative embodiments, top 1% of the barcodes based on molecular counts may be masked as outliers, and other 99% of the barcodes based on molecular counts may be used to determine a threshold as described above; of the barcodes with molecular counts greater than this threshold as selected barcodes, again the top 1% of the selected barcodes may be masked as outliers, and the other 99% of the selected barcodes may be used to determine a new threshold. This process may be repeated to determine a new threshold based on 99% of the subsequently selected barcodes to define a final threshold.

[0121] In additional and alternative embodiments, the thresholding method may involve sorting barcodes based on molecular counts and pick barcodes that meet a selected number of molecular counts as cells or meeting a threshold.

[0122] In the initialization step 330, after using a threshold to separate the barcodes into an initial cell population and an initial non-cell population, e.g., initial cell barcodes and initial non-cell barcodes, initial centroids of the initial cell population and initial non-cell population may be determined using the (x, y) counts in each population. The centroid or geometric center of a plane figure is the arithmetic mean position of all the points in the plane figure, for example, all the data points in the initial cell populations or all the data points in the initial non-cell population. For example, the centroid is the point at which a cutout of the shape (e.g., the initial cell cloud or the initial non-cell cloud) could be perfectly balanced on the tip of a pin. Centroids can extend to any object in n-dimensional space: a centroid can be the mean position of all the points in all of the coordinate directions of the initial cell population or the initial non-cell population.

[0123] Additionally, or alternatively, other methods that make reasonable guesses as to the centroids or center of mass of the initial cell and initial non-cell population can also be used to initialize a clustering algorithm, such as a K-means algorithm. For example, the kmeans++ or any variations thereof may be used for initialization: the initial centroid may be selected randomly, but the next centroid to be selected may be chosen as far apart from the initial centroid as possible. For another example, the k-cluster centers are initialized by choosing k-random points from the data to be clustered as initial points. For even another example, each point in the data can be randomly assigned to a random cluster ID. Then, the points may be grouped by their cluster ID and averaged (per cluster ID) to yield the initial points. In further embodiments, the effect of the initialization step on the outcome of the clustering algorithm can be evaluated and serve as a basis for selecting a desired initialization method for clustering.

[0124] In boundary refinement step 340, after initialization, a clustering method such as k-means clustering, may be used to reclassify the initial centroids for boundary refinement. After initialization to separate data into two groups of an initial cell population and an initial non-cell population, the centroids can be calculated for these initial groups and can then be used to initialize a standard K-means clustering algorithm. After initialization, the k-means algorithm proceeds in two principal phases that are repeated until convergence: (a) For each data point being clustered, compare its distance to each of the k cluster centers and assign it to (make it a member of) the cluster to which it is the closest; (b) For each cluster center, change its position to the centroid of all of the points in its cluster (from the memberships just computed). The centroid is computed as the average of all the data points in the cluster. This process is iterated until membership (and hence cluster centers) ceases to change between iterations. At this point the clustering terminates, and the output is a set of final cluster centers and a mapping of each point to the cluster to which it belongs. The clusters for k-means can be generated quickly with k from 1 to 10. In some embodiments, k is 2. In other embodiments, k is more than 2.

[0125] The initialization step 330 determines an initial estimate where initial clusters (e.g., cells and non-cells) are, and boundary refinement step 340 using clustering (e.g., K-means cluster) to refine this initial classification to get more accurate joint cell calling by utilizing the multi-dimensional information (e.g., x and y) together. For example, two groups are defined based on the ordmag thresholding. Centroids are calculated for these two groups and used to initialize the K-means algorithms set to k=2. Basically, this is a way to get more accurate assignment of barcodes than in the initialization/thresholding step.

[0126] At step 350, an optional filtering step may be applied to exclude noise barcodes using ATAC counts or RNA counts. For example, filtering may be based on RNA counts to filter dead cells using elevated monoconidial UMI counts.

[0127] Filtering step 350 and duplicate identification and deduplication step 320 can be reversed or run concurrently and are independent steps. For example, filtering 350 may follow deduplication 320

but precede initialization 330. In other embodiments, filtering 350 may precede deduplication 320 or occur concurrently as deduplication 320.

[0128] The filtering step 350 may include the correction of gel bead artifacts (whitelist contamination, barcode multiplets, gel bead multiplets) to addressed signal splitting. In various
5 embodiments, transposing bulk nuclei suspensions can include incubating the nuclei suspension with a transposition mix that includes a Transposase enzyme, e.g., a Tn5 transposase. The transposase enters the nuclei and preferentially fragments the DNA in open regions of the chromatin by a process called transposing. More specifically, in various embodiments herein, the process results in transposing the nuclei in a bulk solution. Simultaneously during this process, adapter sequences can be added to the ends of the
10 DNA fragments by the transposase. This process results in adapter-tagged DNA fragments inside an individually transposed nucleus. The transposase fragments the free DNA, producing many small fragments with adapters on either end. Fragments can be sequenced if two transposase enzymes cut at close (<1kb) locations in the same orientation, so the fragment between them has the correct set of adapters. Therefore, if three transposase enzymes cut at nearby locations in the same orientation, two fragments are
15 produced that share a cut site between them. Each of those fragments are then barcoded independently by whatever barcodes are available in the GEM. A transposase can tagment the free DNA at sites that produce directly adjacent fragments, which generally would and should possess identical barcodes. However, in certain circumstances, largely due to barcoding-related errors during the sequencing process, those adjacent fragments can be tagged with different barcodes. Therefore, rather than producing a signal that would
20 normally be from one barcode common to both fragments, the signal is incorrectly split between two different barcodes, causing inaccuracy in the data and inaccuracy in downstream computational analysis. These can generally be referred to as multiplets. In the case of a gel-bead based sequencing platform, those barcoding-related errors can be referred to as gel bead artifacts, though artifacts could potentially occur in various sequencing platforms, and thus the general designation as multiplets.

[0129] Gel bead artifacts such as multiplets can be identified in various ways, one of which is by
25 use of an adjacency matrix such as adjacency matrix. With adjacency matrices, a grid is set up that lists all the nodes on both the x-axis (horizontal) and the y-axis (vertical). Then, values are filled in to the matrix to indicate if there is or is not an “edge” between every pair of nodes. Typically, a 0 indicates no edge and a 1 indicates an edge. To apply these matrices to analyze for multiplets generally, one would count of pairs
30 of adjacent fragments identified through the adjacency matrix, and then keep track of how often the pair of barcodes is seen.

[0130] In various embodiments, methods are provided for generating cell populations and non-cell populations for joint cell calling. The methods can be implemented via computer software or hardware. The methods can also be implemented on a computing device/system that can include a combination of

engines for generating cell populations and non-cell populations for joint cell calling. In various embodiments, the computing device/system can be communicatively connected to one or more of a data source, sample analyzer (e.g., a genomic sequence analyzer), and display device via a direct connection or through an internet connection.

5 **[0131]** Referring now to FIG. 4 a flowchart illustrating a non-limiting example method 400 for generating cell populations and non-cell populations for joint cell calling is disclosed, in accordance with various embodiments. The method can comprise, at step 402, receiving a data set at least associated with a plurality of cells, wherein the data set comprises molecule counts of at least two genomic features for each cell. For example, at least two genomic features can be gene expression features assay for transposase-accessible chromatin (ATAC) features.

10 **[0132]** The method can comprise, at step 404, identifying duplicate subsets of data points from the data set. If some barcodes have exactly the same or substantially the same two genomic features, for example, exactly the same or substantially the same RNA UMI counts and ATAC cutsite counts, these barcodes can be identified because they are presumed to be contamination, e.g., whitelist contamination.

15 **[0133]** The method can comprise, at step 406, generating deduplicated data by condensing data points from each duplicate subset into a single data point. Deduplication would allow to only count unique pairs of counts and remove other counts from future analysis. The density distribution of the unique pairs of counts becomes much more representative of the observed (x,y) scatterplot (e.g., x refers to ATAC cutsites in peaks, y refers to UMI counts for RNA) and thereby is more accurately able to represent the separation between cells and non-cells.

20 **[0134]** The method can comprise, at step 408, applying a pre-set threshold to divide the deduplicated data into an initial cell population and an initial non-cell population, wherein the pre-set threshold is determined using the molecule counts of at least two genomic features for each single cell. This initial separation serves as initialization for optimal clustering. For example, the pre-set threshold can be determined by ordmag. After a threshold is determined, the threshold may be applied to the deduplicated data to generate an initial cell population and an initial non-cell population for initialization.

25 **[0135]** The method can comprise, at step 410, generating a refined cell population and a refined non-cell population by adjusting boundaries of the initial cell population and initial non-cell population using clustering. In accordance with various embodiments, the clustering includes K-means clustering, for example, K equals to 2 or K is more than 2.

30 **[0136]** In accordance with various embodiments, FIG. 5 illustrates a non-limiting example system for joint cell calling, in accordance with various embodiments. The system 500 includes a genomic sequence analyzer 502, a data storage unit 504, a computing device/analytics server 506, and a display 514.

[0137] The genomic sequence analyzer 502 can be communicatively connected to the data storage unit 504 by way of a serial bus (if both form an integrated instrument platform) or by way of a network connection (if both are distributed/separate devices). The genomic sequence analyzer 502 can be configured to process, analyze and generate two or more genomic sequence datasets from a sample, such as the single cell gene expression libraries and the single cell ATAC libraries of the various embodiments herein. Each fragment in the single cell gene expression libraries includes an associated barcode and unique identifier sequence (i.e., UMI). In various embodiments, the genomic sequence analyzer 502 can be a next-generation sequencing platform or sequencer such as the Illumina® sequencer, MiSeq™, NextSeq™ 500/550 (High Output), HiSeq 2500™ (Rapid Run), HiSeq™ 3000/4000, and NovaSeq.

[0138] In various embodiments, the generated genomic sequence datasets can then be stored in the data storage unit 504 for subsequent processing. In various embodiments, one or more raw genomic sequence datasets can also be stored in the data storage unit 504 prior to processing and analyzing. Accordingly, in various embodiments, the data storage unit 504 can be configured to store one or more genomic sequence datasets, e.g., the genomic sequence datasets of the various embodiments herein that includes a plurality of fragment sequence reads with their associated barcodes and unique identifier sequences from the single cell gene expression libraries and the single cell ATAC libraries. In various embodiments, the processed and analyzed genomic sequence datasets can be fed to the computing device/analytics server 506 in real-time for further downstream analysis.

[0139] In various embodiments, the data storage unit 504 is communicatively connected to the computing device/analytics server 506. In various embodiments, the data storage unit 504 and the computing device/analytics server 506 can be part of an integrated apparatus. In various embodiments, the data storage unit 504 can be hosted by a different device than the computing device/analytics server 506. In various embodiments, the data storage unit 504 and the computing device/analytics server 506 can be part of a distributed network system. In various embodiments, the computing device/analytics server 506 can be communicatively connected to the data storage unit 504 via a network connection that can be either a “hardwired” physical network connection (e.g., Internet, LAN, WAN, VPN, etc.) or a wireless network connection (e.g., Wi-Fi, WLAN, etc.). In various embodiments, the computing device/analytics server 506 can be a workstation, mainframe computer, distributed computing node (part of a “cloud computing” or distributed networking system), personal computer, mobile device, etc.

[0140] In various embodiments, the computing device/analytics sever 506 is configured to host one or more upstream data processing engines 508, a clustering engine 510, and one or more downstream data processing engines 512.

[0141] Examples of upstream data processing engines 508 can include, but are not limited to: alignment engine, cell barcode processing engine (for correcting sequencing barcode sequencing errors),

alignment engine (for aligning the fragment sequence reads to a reference genome), duplicate marking engine (identifying duplicate reads by sorting reads), peak calling engine (for counting cut sites in a window around each base-pair of the genome and thresholding it to find regions enriched for open chromatin), annotation engine (for annotating each of the aligned fragment sequence reads with relevant information),
5 etc.

[0142] The clustering engine 510 can be configured to receive one or more genomic sequence datasets that are stored in the data storage unit 504. The genomic sequence datasets comprise a plurality of fragment sequence reads (generated from the sequencing of a single cell gene expression library and a single cell ATAC library), each with an associated barcode sequence. In various embodiments, the
10 clustering engine 510 can be configured to receive processed and analyze genomic sequence datasets from the genomic sequence analyzer 502 in real-time.

[0143] In various embodiments, the clustering engine 510 can be configured identifying duplicate reads of data points from the data set that comprises molecule counts of at least two genomic features for each cell. If some barcodes have exactly the same or substantially the same two genomic features, for
15 example, exactly the same or substantially the same RNA UMI counts and ATAC cutsite counts, these barcodes can be identified because they are presumed to be contamination, e.g., whitelist contamination.

[0144] In various embodiments, the clustering engine 510 can be configured to generate deduplicated data by condensing data points from each duplicate subset into a single data point. Deduplication would allow to only count unique pairs of counts and remove other counts from future
20 analysis. The density distribution of the unique pairs of counts becomes much more representative of the observed (x,y) scatterplot (e.g., x refers to ATAC cutsites in peaks, y refers to UMI counts for RNA) and thereby is more accurately able to represent the separation between cells and non-cells.

[0145] In various embodiments, the clustering engine 510 can be configured to apply a pre-set threshold to divide the deduplicated data into an initial cell population and an initial non-cell population,
25 wherein the pre-set threshold is determined using the molecule counts of at least two genomic features for each single cell. This initial separation serves as initialization for optimal clustering. For example, the pre-set threshold may be determined by calculating a lower bound for the region of interest using a percentile of the sorted barcodes, wherein all barcodes having a molecule count above the lower bound are included for further analysis. In accordance with various embodiments, the percentile can be the 1st percentile
30 barcode of ranked barcodes. For example, the lower bound can be 10 percent of the count of the 1st percentile barcode. After a threshold is determined, the threshold may be applied to the deduplicated data to generate an initial cell population and an initial non-cell population for initialization.

[0146] In various embodiments, the clustering engine 510 can be configured to generate a refined cell population and a refined non-cell population by adjusting boundaries of the initial cell population and

non-cell population using clustering. For example, the clustering can be a K-means clustering method, for example, K equals to 2 or K is more than 2.

[0147] The identified subset of cell then be further processed by one or more downstream data processing engines 512. Examples of downstream data processing engines 512 can include, but are not limited to: a joint cell calling engine (for grouping fragment sequence reads and gene expression sequence reads as being from a unique cell), feature barcode matrix engine (for creating a feature barcode matrix), differential analysis engine (for identifying genes whose expression is specific to each cell cluster), peak barcode matrix engine (for creating a peak barcode matrix), differential accessibility analysis engine (for identifying differential gene expression between different cells or groups of cells), secondary analysis engine (including dimensionality reduction, clustering, t-SNE projection), peak annotation engine (for mapping peaks to a gene), etc.

[0148] After the downstream processing has been performed, an output of the results can be displayed as a result or summary on a display or client terminal 514 that is communicatively connected to the computing device/analytics server 506. In various embodiments, the display or client terminal 514 can be a client computing device. In various embodiments, the display or client terminal 514 can be a personal computing device having a web browser (e.g., INTERNET EXPLORER™, FIREFOX™, SAFARI™, etc.) that can be used to control the operation of the genomic sequence analyzer 502, data storage unit 504, upstream data processing engines 508, clustering engine 510, and the downstream data processing engines 512.

[0149] It should be appreciated that the various engines can be combined or collapsed into a single engine, component or module, depending on the requirements of the particular application or system architecture. In various embodiments engines 508/510/512 can comprise additional engines or components as needed by the particular application or system architecture.

25 EXAMPLES

[0150] FIG. 6 are plots depicting an effect of deduplication on barcodes with RNA counts from a particular single-cell sample (plotted on the y-axis) versus associated ATAC counts from the same particular single-cell sample (plotted on the x-axis), in accordance with various embodiments. Initial centroids refer to the centroids calculated on the ordmag-derived cell/non-cell populations. Updated centroids are calculated based on the final classification after boundary refinement.

[0151] In the scatterplots the grayscale of the points represents the density of data in that region. A region is defined by splitting the plot into a 100 x 100 grid. For the panel on the left, where density is calculated over all barcodes, the lower left corner of the plot between (10,10) has the highest density. In fact, the maximum density is $2^{17.5}$ i.e. that region represents $2^{17.5}$ data points.

[0152] In contrast, for the panel on the right, the density is calculated after de-duplication and now the lower-left corner is no longer the region with the highest amount of data. And overall the range of the density of the points is much more compressed i.e. region with highest density is only representing 2^3 points.

5 [0153] FIG. 7 is a plot depicting the initial ordmag classification of deduplicated data, with RNA counts from a particular single-cell sample (plotted on the y-axis) versus associated ATAC counts from the same particular single-cell sample (plotted on the x-axis), in accordance with various embodiments. After independently defining a threshold for each dimension (i.e. x and y axis), barcodes above both thresholds are classified as cells and the remaining as classified as non-cells. The barcodes above both thresholds and
10 classified as cells are illustrated by the cluster of points shown on the upper right quadrant in FIG. 7

[0154] As such, one threshold is set on x, one threshold on y, and each of these use ordmag. This classification is where the initial centroids are derived from and is used to initialize the K means algorithm in the boundary refinement step. Therefore, this plot only has the initial centroids but not the updated centroids. Ordmag gets an initial estimate where cells are, and boundary refinement using K-means
15 clustering refines this classification to get more accurate joint cell calling by utilizing the x,y information together.

[0155] FIG. 8 is a plot depicting an effect of K-means refinement on barcodes with RNA counts from a particular single-cell sample (plotted on the y-axis) versus associated ATAC counts from the same particular single-cell sample (plotted on the x-axis), in accordance with various embodiments. This graph
20 shows the final outcome of the classification after K-means refinement. Compared with the previous plot in FIG. 7, the classification into cell barcodes (upper right) and non-cell barcodes (lower left) is much more consistent with the observed separation between the two populations.

[0156] FIG. 9 is a plot depicting an effect of joint cell calling method on sensitivity (plotted on the y-axis) and GEX reads per cell (plotted on the x-axis) with different ATAC depths represented by the
25 different lines and numbers (full, 50000, 25000, 10000, 2000, 1000, and 100) . Cell calls at full sequencing depth of ATAC and GEX presumably represent the real situation (or truth). On the x-axis is plotted the GEX reads/cell and different lines denote the ATAC reads/cell. The y-axis corresponds to the sensitivity i.e. recall or true positive rate for that combination of ATAC and GEX sequencing depth with reference to the truth.

30 [0157] This is an evaluation of the joint cell calling method. It establishes that, even at low sequencing depth, one can drop down to 2k GEX reads/cell and 2k ATAC reads/cell and can still keep sensitivity to 90%.

Computer Implemented System

[0158] In various embodiments, the methods for generating cell populations and non-cell populations from a multi genomic feature sequence dataset for joint cell calling can be implemented via computer software or hardware. That is, as depicted in FIG. 5, the methods disclosed herein can be implemented on a computing device 506 that includes upstream processing engines 508, a clustering engine 510 and downstream processing engines 512. In various embodiments, the computing device 506 can be communicatively connected to a data storage unit 504 and a display device 514 via a direct connection or through an internet connection.

[0159] It should be appreciated that the various engines depicted in FIG. 5 can be combined or collapsed into a single engine, component or module, depending on the requirements of the particular application or system architecture. Moreover, in various embodiments, the upstream processing engines 508, clustering engine 510 and downstream processing engines 512 can comprise additional engines or components as needed by the particular application or system architecture.

[0160] FIG. 10 is a block diagram illustrating a computer system 1000 upon which embodiments of the present teachings may be implemented. In various embodiments of the present teachings, computer system 1000 can include a bus 1002 or other communication mechanism for communicating information and a processor 1004 coupled with bus 1002 for processing information. In various embodiments, computer system 1000 can also include a memory, which can be a random-access memory (RAM) 1006 or other dynamic storage device, coupled to bus 1002 for determining instructions to be executed by processor 1004. Memory can also be used for storing temporary variables or other intermediate information during execution of instructions to be executed by processor 1004. In various embodiments, computer system 1000 can further include a read only memory (ROM) 1010 or other static storage device coupled to bus 1002 for storing static information and instructions for processor 1004. A storage device 1012, such as a magnetic disk or optical disk, can be provided and coupled to bus 1002 for storing information and instructions.

[0161] In various embodiments, computer system 1000 can be coupled via bus 1002 to a display 1014, such as a cathode ray tube (CRT) or liquid crystal display (LCD), for displaying information to a computer user. An input device 1016, including alphanumeric and other keys, can be coupled to bus 1002 for communication of information and command selections to processor 1004. Another type of user input device is a cursor control 1018, such as a mouse, a trackball or cursor direction keys for communicating direction information and command selections to processor 1004 and for controlling cursor movement on display 1014. This input device 1016 typically has two degrees of freedom in two axes, a first axis (i.e., x) and a second axis (i.e., y), that allows the device to specify positions in a plane. However, it should be

understood that input devices 1016 allowing for 3-dimensional (x, y and z) cursor movement are also contemplated herein.

[0162] Consistent with certain implementations of the present teachings, results can be provided by computer system 1000 in response to processor 1004 executing one or more sequences of one or more instructions contained in memory 1006. Such instructions can be read into memory 1006 from another computer-readable medium or computer-readable storage medium, such as storage device 1012. Execution of the sequences of instructions contained in memory 1006 can cause processor 1004 to perform the processes described herein. Alternatively, hard-wired circuitry can be used in place of or in combination with software instructions to implement the present teachings. Thus, implementations of the present teachings are not limited to any specific combination of hardware circuitry and software.

[0163] The term “computer-readable medium” (e.g., data store, data storage, etc.) or “computer-readable storage medium” as used herein refers to any media that participates in providing instructions to processor 1004 for execution. Such a medium can take many forms, including but not limited to, non-volatile media, volatile media, and transmission media. Examples of non-volatile media can include, but are not limited to, dynamic memory, such as memory 1006. Examples of transmission media can include, but are not limited to, coaxial cables, copper wire, and fiber optics, including the wires that comprise bus 1002.

[0164] Common forms of computer-readable media include, for example, a floppy disk, a flexible disk, hard disk, magnetic tape, or any other magnetic medium, a CD-ROM, any other optical medium, punch cards, paper tape, any other physical medium with patterns of holes, a RAM, PROM, and EPROM, a FLASH-EPROM, another memory chip or cartridge, or any other tangible medium from which a computer can read.

[0165] In addition to computer-readable medium, instructions or data can be provided as signals on transmission media included in a communications apparatus or system to provide sequences of one or more instructions to processor 1004 of computer system 1000 for execution. For example, a communication apparatus may include a transceiver having signals indicative of instructions and data. The instructions and data are configured to cause one or more processors to implement the functions outlined in the disclosure herein. Representative examples of data communications transmission connections can include, but are not limited to, telephone modem connections, wide area networks (WAN), local area networks (LAN), infrared data connections, NFC connections, etc.

[0166] It should be appreciated that the methodologies described herein, flow charts, diagrams and accompanying disclosure can be implemented using computer system 1000 as a standalone device or on a distributed network or shared computer processing resources such as a cloud computing network.

[0167] The methodologies described herein may be implemented by various means depending upon the application. For example, these methodologies may be implemented in hardware, firmware, software, or any combination thereof. For a hardware implementation, the processing unit may be implemented within one or more application specific integrated circuits (ASICs), digital signal processors (DSPs), digital signal processing devices (DSPDs), programmable logic devices (PLDs), field programmable gate arrays (FPGAs), processors, controllers, micro-controllers, microprocessors, electronic devices, other electronic units designed to perform the functions described herein, or a combination thereof.

[0168] In various embodiments, the methods of the present teachings may be implemented as firmware and/or a software program and applications written in conventional programming languages such as C, C++, Python, etc. If implemented as firmware and/or software, the embodiments described herein can be implemented on a non-transitory computer-readable medium in which a program is stored for causing a computer to perform the methods described above. It should be understood that the various engines described herein can be provided on a computer system, such as computer system 1000, whereby processor 1004 would execute the analyses and determinations provided by these engines, subject to instructions provided by any one of, or a combination of, memory components 1006/1010/1012 and user input provided via input device 1016.

[0169] While the present teachings are described in conjunction with various embodiments, it is not intended that the present teachings be limited to such embodiments. On the contrary, the present teachings encompass various alternatives, modifications, and equivalents, as will be appreciated by those of skill in the art.

[0170] In describing the various embodiments, the specification may have presented a method and/or process as a particular sequence of steps. However, to the extent that the method or process does not rely on the particular order of steps set forth herein, the method or process should not be limited to the particular sequence of steps described, and one skilled in the art can readily appreciate that the sequences may be varied and still remain within the spirit and scope of the various embodiments.

RECITATION OF EMBODIMENTS

[0171] Embodiment 1: A method for distinguishing cell populations from non-cell populations within a data set, the method comprising receiving a data set at least associated with a plurality of cells, wherein the data set comprises molecule counts of at least two genomic features for each cell; identifying duplicate subsets of data points from the data set; generating deduplicated data by condensing data points from each duplicate subset into a single data point; applying a pre-set threshold to divide the deduplicated data into an initial cell population and a non-cell population, wherein the pre-set threshold is determined

using the molecule counts; and generating a refined cell population and a non-cell population by adjusting boundaries of the initial cell population and non-cell population using clustering.

[0172] Embodiment 2: The method of Embodiment 1, wherein at least one of the at least two genomic features comprise a gene.

5 [0173] Embodiment 3: The method of any one of Embodiments 1 to 2, wherein at least one of the at least two genomic features comprise open genomic regions.

[0174] Embodiment 4: The method of any one of Embodiments 1 to 3, further comprising filtering the data set to remove gel bead artifacts.

10 [0175] Embodiment 5: The method of any one of Embodiments 1 to 4, wherein the data set comprises barcodes, each barcode corresponding to each single cell of the plurality of cells.

[0176] Embodiment 6: The method of any one of Embodiments 1 to 5, wherein the pre-set threshold is determined by ranking barcodes in the deduplicated data based on molecular counts of each barcode and determining the pre-set threshold for selecting barcodes using a pre-set percentile of ranked barcodes, wherein any barcodes having a molecular count above the pre-set threshold are classified as being
15 in the initial cell population.

[0177] Embodiment 7: The method of any one of Embodiments 1 to 6, wherein adjusting boundaries comprises obtaining centroids of the initial cell population and the initial non-cell population.

[0178] Embodiment 8: The method of Embodiment 7, wherein adjusting boundaries comprises initializing a K-means clustering with the centroids.

20 [0179] Embodiment 9: The method of any one of Embodiments 1 to 8, wherein adjusting boundaries comprises using K-means clustering with K=2.

[0180] Embodiment 10: The method of any one of Embodiments 1 to 8, wherein adjusting boundaries comprises using K-means clustering with K more than 2.

[0181] Embodiment 11: A non-transitory computer-readable medium storing computer
25 instructions that, when executed by a computer, cause the computer to perform a method for distinguishing cell populations from non-cell populations within a data set, the method comprising: receiving a data set at least associated with a plurality of cells, wherein the data set comprises molecule counts of at least two genomic features for each cell; identifying duplicate subsets of data points from the data set; generating deduplicated data by condensing data points from each duplicate subset into a single data point; applying a
30 pre-set threshold to divide the deduplicated data into an initial cell population and an initial non-cell population, wherein the pre-set threshold is determined using the molecule counts; and generating a refined cell population and a refined non-cell population by adjusting boundaries of the initial cell population and the initial non-cell population using clustering.

[0182] Embodiment 12: The non-transitory computer-readable medium of Embodiment 11, wherein at least one of the at least two genomic features comprise a gene.

5 [0183] Embodiment 13: The non-transitory computer-readable medium of any one of Embodiments 11 to 12, wherein at least one of the at least two genomic features comprise open genomic regions.

[0184] Embodiment 14: The non-transitory computer-readable medium of any one of Embodiments 11 to 13, further comprising filtering the data set to remove gel bead artifacts.

10 [0185] Embodiment 15: The non-transitory computer-readable medium of any one of Embodiments 11 to 14, wherein the data set comprises barcodes, each barcode corresponding to each single cell of the plurality of cells.

[0186] Embodiment 16: The non-transitory computer-readable medium of any one of Embodiments 11 to 15, wherein the pre-set threshold is determined by ranking barcodes in the deduplicated data based on molecular counts of each barcode and determining the pre-set threshold for selecting barcodes using a pre-set percentile of ranked barcodes, wherein any barcodes having a molecular count above the pre-set threshold are classified as being in the initial cell populations.

[0187] Embodiment 17: The non-transitory computer-readable medium of any one of Embodiments 11 to 16, wherein adjusting boundaries comprises obtaining centroids of the initial cell population and initial non-cell population.

20 [0188] Embodiment 18: The non-transitory computer-readable medium of Embodiment 17, wherein adjusting boundaries comprises initializing a K-means clustering with the centroids.

[0189] Embodiment 19: The non-transitory computer-readable medium of any one of Embodiments 11 to 18, wherein adjusting boundaries comprises using K-means clustering with K=2.

25 [0190] Embodiment 20: The non-transitory computer-readable medium of any one of Embodiments 11 to 18, wherein adjusting boundaries comprises using K-means clustering with K more than 2.

30 [0191] Embodiment 21: A system for distinguishing cell populations from non-cell populations within a data set, comprising: a data store configured to store a data set at least associated with a plurality of cells, wherein the data set comprises molecule counts of at least two genomic features for each cell; and a computing device communicatively connected to the data store and configured to receive the data set, the computing device comprising a clustering engine configured to identify duplicate subsets of data points from the data set; generate deduplicated data by condensing data points from each duplicate subset into a single data point; apply a pre-set threshold to divide the deduplicated data into an initial cell population and an initial non-cell population, wherein the pre-set threshold is determined using the molecule counts; and generate a refined cell population and a refined non-cell population by adjusting boundaries of the initial

cell population and initial non-cell population using clustering; and a display communicatively connected to the computing device and configured to display a report comprising the refined cell population and refined non-cell population.

5 **[0192]** Embodiment 22: The system of Embodiment 21, wherein at least one of the at least two genomic features comprise a genes.

[0193] Embodiment 23: The system of any one of Embodiments 21 to 22, wherein at least one of the at least two genomic features comprise open genomic regions.

[0194] Embodiment 24: The system of any one of Embodiments 21 to 23, further comprising filtering the data set to remove gel bead artifacts.

10 **[0195]** Embodiment 25: The system of any one of Embodiments 21 to 24, wherein the data set comprises barcodes, each barcode corresponding to each single cell of the plurality of cells.

[0196] Embodiment 26: The system of any one of Embodiments 21 to 25, wherein the pre-set threshold is determined by ranking barcodes in the deduplicated data based on molecular counts of each barcode and determining the pre-set threshold for selecting barcodes using a pre-set percentile of ranked
15 barcodes, wherein any barcodes having a molecular count above the pre-set threshold are classified as being in the initial cell populations.

[0197] Embodiment 27: The system of any one of Embodiments 21 to 26, wherein adjusting boundaries comprises obtaining centroids of the initial cell population and initial non-cell population.

20 **[0198]** Embodiment 28: The system of Embodiment 27, wherein adjusting boundaries comprises initializing a K-means clustering with the centroids.

[0199] Embodiment 29: The system of any one of Embodiments 21 to 28, wherein adjusting boundaries comprises using K-means clustering with K=2.

[0200] Embodiment 30: The system of any one of Embodiments 21 to 28, wherein adjusting boundaries comprises using K-means clustering with K more than 2.

25

What is claimed:

1. A method for distinguishing cell populations from non-cell populations within a data set, the method comprising:
 - receiving a data set at least associated with a plurality of cells, wherein the data set comprises molecule counts of at least two genomic features for each cell;
 - identifying duplicate subsets of data points from the data set;
 - generating deduplicated data by condensing data points from each duplicate subset into a single data point;
 - applying a pre-set threshold to divide the deduplicated data into an initial cell population and an initial non-cell population, wherein the pre-set threshold is determined using the molecule counts; and
 - generating a refined cell population and a refined non-cell population by adjusting boundaries of the initial cell population and initial non-cell population using clustering.
2. The method of claim 1, wherein at least one of the at least two genomic features comprise a gene.
3. The method according to any one of claims 1-2, wherein at least one of the at least two genomic features comprise open genomic regions.
4. The method according to any one of claims 1-3, further comprising filtering the data set to remove gel bead artifacts.
5. The method according to any one of claims 1-4, wherein the data set comprises barcodes, each barcode corresponding to each single cell of the plurality of cells.
6. The method of claim 5, wherein the pre-set threshold is determined by ranking barcodes in the deduplicated data based on molecular counts of each barcode and determining the pre-set threshold for selecting barcodes using a pre-set percentile of ranked barcodes, wherein any barcodes having a molecular count above the pre-set threshold are classified as being in the initial cell population.
7. The method according to any one of claims 1-6, wherein adjusting boundaries comprises obtaining centroids of the initial cell population and the initial non-cell population.

8. The method of claim 7, wherein adjusting boundaries comprises initializing a K-means clustering with the centroids.

9. The method according to any one of claims 1-8, wherein adjusting boundaries comprises
5 using K-means clustering with K=2.

10. The method according to any one of claims 1-8, wherein adjusting boundaries comprises using K-means clustering with K more than 2.

10 11. A non-transitory computer-readable medium storing computer instructions that, when executed by a computer, cause the computer to perform a method for distinguishing cell populations from non-cell populations within a data set, the method comprising:

receiving a data set at least associated with a plurality of cells, wherein the data set comprises molecule counts of at least two genomic features for each cell;

15 identifying duplicate subsets of data points from the data set;

generating deduplicated data by condensing data points from each duplicate subset into a single data point;

applying a pre-set threshold to divide the deduplicated data into an initial cell population and an initial non-cell population, wherein the pre-set threshold is determined using the molecule counts; and

20 generating a refined cell population and a refined non-cell population by adjusting boundaries of the initial cell population and the initial non-cell population using clustering.

12. The non-transitory computer-readable medium of claim 11, wherein at least one of the at least two genomic features comprise a gene.

25

13. The non-transitory computer-readable medium according to any one of claims 11-12, wherein at least one of the at least two genomic features comprise open genomic regions.

14. The non-transitory computer-readable medium according to any one of claims 11-13,
30 further comprising filtering the data set to remove gel bead artifacts.

15. The non-transitory computer-readable medium according to any one of claims 11-14, wherein the data set comprises barcodes, each barcode corresponding to each single cell of the plurality of cells.

16. The non-transitory computer-readable medium according to any one of claims 11-15, wherein the pre-set threshold is determined by ranking barcodes in the deduplicated data based on molecular counts of each barcode and determining the pre-set threshold for selecting barcodes using a pre-set percentile of ranked barcodes, wherein any barcodes having a molecular count above the pre-set threshold are classified as being in the initial cell population.

17. The non-transitory computer-readable medium according to any one of claims 11-16, wherein adjusting boundaries comprises obtaining centroids of the initial cell population and initial non-cell population.

10

18. The non-transitory computer-readable medium of claim 17, wherein adjusting boundaries comprises initializing a K-means clustering with the centroids.

19. The non-transitory computer-readable medium according to any one of claims 11-18, wherein adjusting boundaries comprises using K-means clustering with $K=2$.

15

20. The non-transitory computer-readable medium according to any one of claims 11-18, wherein adjusting boundaries comprises using K-means clustering with K more than 2.

20

21. A system for distinguishing cell populations from non-cell populations within a data set, comprising:

a data store configured to store a data set at least associated with a plurality of cells, wherein the data set comprises molecule counts of at least two genomic features for each cell; and

5 a computing device communicatively connected to the data store and configured to receive the data set, the computing device comprising a clustering engine configured to

identify duplicate subsets of data points from the data set;

generate deduplicated data by condensing data points from each duplicate subset into a single data point;

10 apply a pre-set threshold to divide the deduplicated data into an initial cell population and an initial non-cell population, wherein the pre-set threshold is determined using the molecule counts; and

generate a refined cell population and a refined non-cell population by adjusting boundaries of the initial cell population and initial non-cell population using clustering;

15 and

a display communicatively connected to the computing device and configured to display a report comprising the refined cell population and refined non-cell population.

22. The system of claim 21, wherein at least one of the at least two genomic features
20 comprise a genes.

23. The system according to any one of claims 21-22, wherein at least one of the at least two genomic features comprise open genomic regions.

25 24. The system according to any one of claims 21-23, further comprising filtering the data set to remove gel bead artifacts.

25. The system according to any one of claims 21-24, wherein the data set comprises barcodes, each barcode corresponding to each single cell of the plurality of cells.

30

26. The system according to any one of claims 21-25, wherein the pre-set threshold is determined by ranking barcodes in the deduplicated data based on molecular counts of each barcode and determining the pre-set threshold for selecting barcodes using a pre-set percentile of ranked barcodes, wherein any barcodes having a molecular count above the pre-set threshold are classified as being in the
5 initial cell population.

27. The system according to any one of claims 21-26, wherein adjusting boundaries comprises obtaining centroids of the initial cell population and initial non-cell population.

10 28. The system of claim 27, wherein adjusting boundaries comprises initializing a K-means clustering with the centroids.

29. The system according to any one of claims 21-28, wherein adjusting boundaries comprises using K-means clustering with $K=2$.
15

30. The system according to any one of claims 21-28, wherein adjusting boundaries comprises using K-means clustering with K more than 2.

20