

SYSTEMS AND METHODS FOR JOINT INTERACTIVE VISUALIZATION OF GENE EXPRESSION AND DNA CHROMATIN ACCESSIBILITY

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to United States Provisional Patent Application No. 63/061,952, entitled “Systems and Methods for Joint Interactive Visualization of Gene Expression and DNA Chromatin Accessibility,” filed August 6, 2020, and to United States Provisional Patent Application No. 62/976,270, entitled “Methods For Characterizing Cells Using Gene Expression And Chromatin Accessibility,” filed February 13, 2020 each of which is hereby incorporated by reference in its entirety.

TECHNICAL FIELD

[0002] This specification describes technologies relating to visualizing RNA gene expression and DNA chromatin accessibility from the same cells.

BACKGROUND

[0003] The discovery of patterns in a dataset facilitates a number of technical applications such as the discovery of changes in discrete attribute values in genes between different classes (*e.g.*, diseased state, non-diseased state, disease stage, *etc.*). For instance, in the biological arts, advances in RNA-extraction protocols and associated methodologies has led to the ability to perform whole transcriptome shotgun sequencing that quantifies gene expression in biological samples in counts of transcript reads mapped to genes. This has given rise to high throughput transcript identification and the quantification of gene expression for hundreds or even thousands of individual cells in a single dataset. Thus, in the art, datasets containing discrete attribute values (*e.g.*, count of transcript reads mapped to individual genes in a particular cell) for each gene in a plurality of genes for each respective cell in a plurality of cells have been generated. While this is a significant advancement in the art, a number of technical problems need to be addressed to make such data more useful.

[0004] One drawback with such advances in the art the unsatisfactory way in which conventional methods find patterns in such datasets. For instance, such patterns may relate to the discovery of unknown classes among the members of the dataset. For example, the discovery that a dataset of what was thought to be homogenous cells turns out to include cells of two different classes. Such patterns may also relate to the discovery of variables that are statistically associated with known classes. For instance, the discovery that the transcript abundance of a subset of mRNA mapping to a core set of genes discriminates between cells that are in a diseased state versus cells that are not in a diseased state. The discovery of such patterns (e.g., the discovery of genes whose mRNA expression discriminates classes or that define classes) in datasets that are very large, are not amendable to classical statistics because of limited replicate information, and for which such patterns in many instances relate to biological processes that are not well understood remains a technical challenge for which improved tools are needed in the art in order to adequately address such drawbacks.

SUMMARY

[0005] Technical solutions (e.g., computing systems, methods, and non-transitory computer readable storage mediums) for addressing the above identified problems with discovery of patterns in datasets are provided in the present disclosure.

[0006] The following presents a summary of the present disclosure in order to provide a basic understanding of some of the aspects of the present disclosure. This summary is not an extensive overview of the present disclosure. It is not intended to identify key/critical elements of the present disclosure or to delineate the scope of the present disclosure. Its sole purpose is to present some of the concepts of the present disclosure in a simplified form as a prelude to the more detailed description that is presented later.

[0007] One aspect of the present disclosure provides a method for visualizing a pattern in a discrete attribute value dataset. The method is performed at a computer system comprising one or more processing cores and a memory, the memory storing instructions that use the one or more processing cores to perform the method. The method comprises storing the discrete attribute value dataset in the memory. The discrete attribute value dataset comprises a respective discrete attribute value for each corresponding gene in a plurality of genes, for each respective cell in a plurality of cells. The discrete attribute value dataset further comprises a respective

ATAC fragment count for each corresponding ATAC peak in a plurality of ATAC peaks, for each respective cell in the plurality of cells.

[0008] In some embodiments, the plurality of cells comprises 100 cells, the plurality of genes comprises 100 genes, and the plurality of ATAC peaks comprises 50 ATAC peaks.

[0009] Each respective cell in the plurality of cells is assigned to a respective cluster group in a first plurality of cluster groups based on a first clustering of discrete attribute values for the plurality of genes across the plurality of cells. Each respective cell in the plurality of cells is also assigned to a respective cluster group in a second plurality of cluster groups based on a second clustering of ATAC fragment count values for the plurality of ATAC peaks across the plurality of cells.

[0010] The method further comprises displaying, in a first panel, a two-dimensional projection of the plurality of cells based on assignment of the plurality of cells to one of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups.

[0011] The method further comprises indicating within the two-dimensional projection, for each respective cell in the plurality of cells, membership in the other of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups, thereby visualizing the pattern in the discrete attribute value dataset.

[0012] In some embodiments membership of each respective cell in the plurality of cells in the other of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups is indicated by coloring the respective cell a color that is uniquely associated with a cluster group to which the respective cell has been assigned in the other of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups.

[0013] In some embodiments, the method further comprises computing, for each respective gene in the plurality of genes for each respective cluster in the first plurality of clusters or the second plurality of clusters, a difference in the discrete attribute value for the respective gene across the respective subset of cells in the respective cluster relative to the discrete attribute value for the respective gene across the first plurality of clusters or the second plurality of clusters other than the respective cluster, thereby deriving a differential value for each respective gene in the plurality of gene for each respective cluster in the first plurality of clusters or the

second plurality of clusters. In such embodiments, there is displayed in a second panel, concurrently with the first panel, a heat map that comprises a representation of the differential value for each respective gene in the plurality of genes for each respective cluster in the first plurality of clusters or the second plurality of clusters thereby visualizing the pattern in the discrete attribute value dataset. In some such embodiments, the differential value for each respective gene in the plurality of genes for each respective cluster in the first plurality of clusters or the second plurality of clusters is a fold change (*e.g.*, a $\log_2$ fold change or a $\log_{10}$ fold change) in (i) a first measure of central tendency of the discrete attribute value for the respective gene measured in each of the cells in the plurality of cells in the respective cluster and (ii) a second measure of central tendency of the discrete attribute value for the respective gene measured in each of the cells of all clusters in the first plurality of clusters or all clusters in the second plurality of clusters, other than the first respective cluster.

[0014] In some embodiments, the method further comprises normalizing the discrete attribute value prior for each respective gene in the plurality of genes prior to computing the differential value for each respective gene in the plurality of genes for each respective cluster in the first plurality of clusters or the second plurality of clusters. In some such embodiments, the normalizing comprises modeling the discrete attribute value of each gene associated with each cell in the plurality of cells with a negative binomial distribution having a consensus estimate of dispersion.

[0015] In some embodiments, the first clustering of discrete attribute values for the plurality of genes across the plurality of cells is a clustering of a first plurality of dimension reduction values for each respective cell in the plurality of cells, across the plurality of cells, where each respective dimensional reduction value in the first plurality of dimension reduction values for each respective cell in the plurality of cells is derived, using a first dimension reduction algorithm, from the discrete attribute values of each gene in the respective cell. In some such embodiments, the second clustering of ATAC fragment counts for the plurality of ATAC peaks across the plurality of cells is a clustering of a second plurality of dimension reduction values for each respective cell in the plurality of cells, across the plurality of cells, where each respective dimensional reduction value in the second plurality of dimension reduction values for each respective cell in the plurality of cells is derived, using the first dimension reduction algorithm

(*e.g.*, principal component analysis), from the ATAC fragment counts of each ATAC peak in the respective cell.

[0016] In some embodiments, the two-dimensional projection of the plurality of cells is based on the assignment of the plurality of cells to the first plurality of cluster groups, and the two-dimensional projection of the plurality of cells is obtained from t-distributed stochastic neighbor or UMAP embedding of the first plurality of dimension reduction values for each respective cell in the plurality of cells, across the plurality of cells.

[0017] In some embodiments, the two-dimensional projection of the plurality of cells is based on the assignment of the plurality of cells to the second plurality of cluster groups, and the two-dimensional projection of the plurality of cells is obtained from t-distributed stochastic neighbor or UMAP embedding of the second plurality of dimension reduction values for each respective cell in the plurality of cells, across the plurality of cells.

[0018] In some embodiments, the first clustering of discrete attribute values for the plurality of genes across the plurality of cells comprises application of a Louvain modularity algorithm, k-means clustering, a fuzzy k-means clustering algorithm, or Jarvis-Patrick clustering, and the second clustering of ATAC fragment count values for the plurality of ATAC peaks across the plurality of cells comprises application of a Louvain modularity algorithm, k-means clustering, a fuzzy k-means clustering algorithm, or Jarvis-Patrick clustering.

[0019] In some embodiments, the first clustering of discrete attribute values for the plurality of genes across the plurality of cells comprises k-means clustering into a first predetermined number of clusters, or the second clustering of ATAC fragment count values for the plurality of ATAC peaks across the plurality of cells comprises k-means clustering into a second predetermined number of clusters. In some such embodiments, the first or second predetermined number of clusters is an integer value between 2 and 50. In some such embodiments, the method further comprises obtaining the integer value from a user of the computer system.

[0020] In some embodiments, the respective discrete attribute value for each corresponding gene in a plurality of genes, for each respective cell in a plurality of cells represents a whole transcriptome shotgun sequencing experiment that quantifies gene expression from each respective single cell in the plurality of cells in counts of transcript reads mapped to the genes.

[0021] In some embodiments, each gene in a particular cell in the plurality of cells is uniquely represented within the discrete attribute value dataset with a first barcode that is unique to the particular cell.

[0022] In some embodiments, the discrete attribute value of each gene in a particular cell in the plurality of cells is determined after the particular cell has been separated from all the other cells in the plurality of cells into its own microfluidic partition.

[0023] In some embodiments, the plurality of cells comprises 1000 cells, 2000 cells, 5000 cells, 10,000 cells, 25,000 cells, 50,000 cells or 100,000 cells.

[0024] In some embodiments, the plurality of genes comprises 150 genes, 200 genes, 300 genes, 400 genes, 1000 genes, 2000 genes, 3000 genes, 4000 genes, or 5000 genes.

[0025] In some embodiments, the plurality of ATAC peaks comprises 100 ATAC peaks, 200 ATAC peaks, 500 ATAC peaks, 750 ATAC peaks, 1000 ATAC peaks, or 5000 ATAC peaks.

[0026] In some embodiments, the discrete attribute value dataset has a file size of at least 250 megabytes, 500 megabytes, one gigabyte, two gigabytes, or three gigabytes.

[0027] In some embodiments, the discrete attribute value dataset further comprises a feature-linkage matrix. The feature-linkage matrix stores, for each respective gene in the plurality of genes and, for each respective ATAC peak in the plurality of ATAC peaks, a collection of ATAC peaks and genes that are within a threshold distance of the respective gene or respective ATAC peak in a reference genome, and for each respective ATAC peak or respective gene in the collection: a correlation in an ATAC fragment count of the respective ATAC peak or a correlation in the discrete attribute value of the respective gene to the first ATAC peak or first gene across the plurality of cells. In such embodiments, the method further comprises receiving a selection of a first gene in the plurality of genes or a first ATAC peak in the plurality of ATAC peaks. In response to this selection, the feature-linkage matrix is used to obtain and provide a first plot comprising a graphical indicator for each gene in the plurality of genes or each peak in the plurality of peaks linked to the first gene or the first ATAC peak, by order of distance apart from the first gene or the first ATAC peak in the reference genome.

[0028] In some embodiments, the respective graphical indicator for each respective gene in the plurality of genes or each respective peak in the plurality of peaks linked to the first gene or the first ATAC peak is provided for each respective cluster group in the first plurality of cluster groups or each cluster group in the second plurality of cluster groups. The respective graphical indicator is dimensioned within the first plot to represent a proportion of cells in the respective cluster group that have a non-zero value for the discrete attribute value of the respective gene or a non-zero value for the ATAC fragment count of the respective ATAC peak.

[0029] In some embodiments, the feature-linkage matrix further stores, for each respective ATAC peak or respective gene in the collection: a significance in an ATAC fragment count of the respective ATAC peak or a significance in the discrete attribute value of the respective gene to the first ATAC peak or first gene across cells in the plurality of cells. In some such embodiments, method further comprises: limiting the first plot, to each gene in the plurality of genes or each peak in the plurality of peaks that have a threshold correlation or significance in an ATAC fragment count of the respective ATAC peak or discrete attribute value of the respective gene to the first ATAC fragment or first gene.

[0030] In some embodiments, the first plot is limited to each gene in the plurality of genes or each peak in the plurality of peaks that is within a threshold distance (*e.g.*, 1 megabase, 2 megabases, or a value between 0.5 megabases and 10 megabases) of the first gene or first ATAC peak in a reference genome.

[0031] Another aspect of the present disclosure provides a method for characterizing cells, comprising partitioning a plurality of cells or cell nuclei and a plurality of barcode beads into a plurality of partitions, where at least a subset of the plurality of partitions each comprise a cell or cell nucleus of the plurality of cells or cell nuclei and a barcode bead of said plurality of barcode beads, and each bead in the subset of said plurality of partitions comprises a unique barcode sequence. The method further includes generating a plurality of barcoded nucleic acid molecules comprising barcode sequences, where a first subset of the plurality of barcoded nucleic acid molecules comprises a sequence corresponding to a ribonucleic (RNA) molecule and a second subset of said plurality of barcoded nucleic acid molecules comprises a sequence corresponding to a sequence corresponding to a region of accessible chromatin. The method further comprises sequencing the plurality of barcoded nucleic acid molecules, or derivatives generated therefrom,

to generate sequencing information, and using the barcode sequence and the sequencing information to identify cell types in the sequencing information.

[0032] In some embodiments, the methods disclosed herein further comprise using said sequencing information to cluster cells by regions of accessible chromatin. In some embodiments, the methods further comprise using the sequencing information to cluster cells by gene expression. In some embodiments, the methods further comprise using the sequencing information and the cells clustered by gene expression to annotate, identify, or characterize cells clustered by regions of accessible chromatin. In some embodiments, the methods further comprise using the sequencing information and the cells clustered by regions of accessible chromatin to annotate, identify, or characterize cells clustered by gene expression. In some embodiments, the plurality of cells or cell nuclei are derived from a tumor sample or a sample suspected of comprising a tumor. In some embodiments, the methods further comprise using the sequencing information to identify a cell type, a cell state, a tumor-specific gene expression pattern, or a tumor-specific differentially accessible regions of chromatin in the tumor sample or the sample suspected of comprising said tumor. In some embodiments, the methods further comprise using the sequencing information to identify or confirm the presence of a tumor cell in the tumor sample or the sample suspected of comprising said tumor. In some embodiments, the methods further comprise administering a therapeutically effective amount of an agent targeting one or more targets identified in the tumor-specific gene expression pattern or the tumor-specific differentially accessible regions of chromatin. In some embodiments, the tumor is a B cell lymphoma.

[0033] Another aspect of the present disclosure provides a computer system comprising one or more processing cores and a memory, the memory storing instructions that use the one or more processing cores to perform any of the methods disclosed herein.

[0034] Another aspect of the present disclosure provides a non-transitory computer readable storage medium. The non-transitory computer readable storage medium stores instructions, which when executed by a computer system, cause the computer system to perform any of the methods disclosed herein.

[0035] As disclosed herein, any embodiment disclosed herein when applicable can be applied to any aspect.

[0036] Various embodiments of systems, methods and devices within the scope of the appended claims each have several aspects, no single one of which is solely responsible for the desirable attributes described herein. Without limiting the scope of the appended claims, some prominent features are described herein. After considering this discussion, and particularly after reading the section entitled “Detailed Description” one will understand how the features of various embodiments are used.

INCORPORATION BY REFERENCE

[0037] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference in their entireties to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[0038] The implementations disclosed herein are illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings. Like reference numerals refer to corresponding parts throughout the several views of the drawings.

[0039] Figures 1A, 1B, 1C and 1D are an example block diagram illustrating a computing device in accordance with some embodiments of the present disclosure.

[0040] Figures 2A, 2B and 2C collectively illustrate an example method in accordance with an embodiment of the present disclosure, in which optional steps are indicated by broken lines.

[0041] Figure 3 illustrates a user interface for obtaining a dataset in accordance with some embodiments.

[0042] Figure 4A illustrates an example display in which a heat map that comprises a representation of the differential value for each respective gene in a plurality of genes for each cluster in a first plurality of clusters is displayed in a first panel while each respective cell in a plurality of cells is displayed in a second panel based upon a dimension reduced two-dimensional GEX data point for the respective cells in accordance with some embodiments.

[0043] Figure 4B illustrates an example display in which a heat map that comprises a representation of the differential value for each respective ATAC peak in a plurality of ATAC peaks for each cluster in a second plurality of clusters is displayed in a first panel while each respective cell in a plurality of cells is displayed in a second panel based upon a dimension reduced two-dimensional ATAC data point for the respective cells in accordance with some embodiments.

[0044] Figure 5 illustrates the use of linked windows in accordance with some embodiments of the present disclosure.

[0045] Figure 6 illustrates splitting the view of a discrete attribute value dataset into component clusters in accordance with some embodiments of the present disclosure.

[0046] Figure 7 illustrates selection of links to a gene or peak in accordance with some embodiments of the present disclosure.

[0047] Figure 8 illustrates a table of ATAC peaks that are linked to the gene CD69 in accordance with some embodiments of the present disclosure.

[0048] Figure 9 illustrates details of a particular gene represented in a discrete attribute value dataset in accordance with some embodiments of the present disclosure.

[0049] Figure 10 illustrates a linkage map for each cluster in a plurality of clusters for the feature ENSG00000126353 in accordance with some embodiments of the present disclosure.

[0050] Figure 11 illustrates ATAC peak linkages and gene linkages, on a cell cluster by cell cluster basis, to the gene CD69 in accordance with some embodiments of the present disclosure.

[0051] Figure 12 illustrates selection of one arc in a linkage view in accordance with some embodiments of the present disclosure.

[0052] Figure 13A illustrates re-anchoring a linkage view around the feature with the highest positive correlative value to CD69: chr12:9760094-9761921, in accordance with some embodiments of the present disclosure.

[0053] Figure 13B illustrates linkages to the queried gene LEF1 and the average UMIs/cell per cell cluster for LEF1, in accordance with some embodiments of the present disclosure.

[0054] Figure 13C illustrates selection of ATAC Peak chr4:108355456-108356170 which is linked to LEF1 and the average cut sites/cell per cell cluster for ATAC Peak chr4:108355456-108356170 as well as the average UMIs/cell per cell cluster for LEF1, in accordance with some embodiments of the present disclosure.

[0055] Figure 13D illustrates a user query for features linked to the ATAC Peak chr4:108355456, along with the average cut sites/cell per cell cluster 1330 for the searched-for ATAC Peak chr4:108355456, as well as all the average UMIs/cell, per cell cluster, for each of three genes linked to ATAC Peak chr4:108355456, in accordance with some embodiments of the present disclosure.

[0056] Figure 13E illustrates, upon user selection of link chr4:108355456-108356170 - chr4:108433617-108434308 in the arc plot of Figure 13D, the average cut sites/cell, per cell cluster, for the searched-for peak chr4:108355456-108356170 as a first graph and also the average Cut Sites/Cell, per cell cluster, for the linked feature chr4:108433617-108434308, as a second graph, in accordance with some embodiments of the present disclosure.

[0057] Figure 14 illustrates a correlation slider, which filters linkages by absolute r-value, and a significance filter, which filters linkages by significance score (negative log false-discovery-rate) and that collectively provide linkage filters in accordance with some embodiments of the present disclosure.

[0058] Figure 15 illustrates the user selection of ATAC classes for an ATAC category in accordance with some embodiments of the present disclosure.

[0059] Figure 16 illustrates the user selection of GEX classes for a GEX category in accordance with some embodiments of the present disclosure.

[0060] Figure 17 illustrates a locally distinguishing significant feature comparison feature among selected classes (*e.g.*, K-means clusters) of a category in accordance with some embodiments of the present disclosure.

[0061] Figure 18 illustrates an ATAC t-SNE plot, but with the cells clustered (*e.g.*, visualized as colored cells) based on GEX clustering rather than ATAC clustering, in accordance with some embodiments of the present disclosure.

[0062] Figure 19 illustrates an exemplary scheme for tandem ATAC ligation and RNA template switching in accordance with an embodiment of the present disclosure.

[0063] Figure 20 illustrates an example of a user interface where the number of ATAC fragments near the promoter of the gene GZMB are quantified, on a cell by cell basis, in accordance with a scale (*e.g.*, an intensity scale and/or a color scale), in accordance with some embodiments of the present disclosure.

[0064] Figure 21 illustrates an example of the user interface where information on a selected peak is displayed in accordance with some embodiments of the present disclosure.

[0065] Figure 22 illustrates an example of a user interface for selectively presenting information from a discrete attribute dataset 120, including information on fragments, in accordance with some embodiments of the present disclosure.

[0066] Figure 23 illustrates an example of a user interface for selectively presenting information from an ATAC dataset, where the user interface present cluster information and ATAC peak information in accordance with some embodiments of the present disclosure.

[0067] Figures 24A and 24B collectively illustrate cell clustering and annotation of cells using expressed markers, in accordance with an embodiment of the present disclosure. Gene expression (GEX) profiling was performed for 24,000 peripheral blood mononuclear cells (PBMCs) in a joint gene expression and open chromatin profiling assay, and gene expression annotations were obtained using expressed markers for 13 categories of cell types. Gene expression annotations were visualized using a t-distributed stochastic neighbor embedding (t-SNE) plot (2400).

[0068] Figure 25A and 25B collectively illustrate cell clustering and annotation of cells using chromatin accessibility (open chromatin), in accordance with an embodiment of the present disclosure. An assay of transposase-accessible chromatin (ATAC) was performed for 24,000 peripheral blood mononuclear cells (PBMCs) in a joint gene expression and open chromatin profiling assay, and open chromatin annotations were obtained using transcription factor accessibility for 4 categories of cell types (monocytes, NK/CD8 T cells, naïve/memory T cells, and B cells) (2500).

[0069] Figures 26A and 26B collectively illustrate concordance between the two read-outs of Figures 24 (2400) and 25 (2500), in accordance with an embodiment of the present disclosure.

[0070] Figure 27A and 27B collectively illustrate the transfer of gene expression marker-derived annotation into accessible chromatin clustered populations, in accordance with an embodiment of the present disclosure. Top panel: cells were clustered and annotated with ATAC data using transcription factor accessibility (*e.g.*, open chromatin), providing annotated cell clusters with limited specificity to differentiate additional, more precise cell types. Bottom panel: cells clustered using open chromatin (ATAC) clustering were further annotated using gene expression markers.

[0071] Figures 28A, 28B, and 28C collectively illustrate the identification of novel populations of cells that would otherwise be unidentified and/or unannotated when analyzing either gene expression or regions of open chromatin alone, in accordance with an embodiment of the present disclosure. In Figure 28A, populations of cells identified using open chromatin analysis alone would show as large clusters of cells (*e.g.*, B cells (2508)). Such populations can be annotated using gene expression markers, such as those indicated in Figure 28B. Figure 28C illustrates the application of gene expression annotation in order to further stratify sub-populations of cells clustered by open chromatin (*e.g.*, naïve/memory B cells (2722), IgM+IgD+ memory B cells (2724), and plasma B cells (2726)).

[0072] Figures 29A, 29B, 29C, 29D, and 29E collectively illustrate differential gene expression in the annotated cells shown in Figures 28A-C, and the identification and differentiation of prospective naïve vs. memory B cells, in accordance with an embodiment of the present disclosure. A population of cells identified by gene expression (GEX) analysis to be a single cluster of naïve/memory B cells (Figure 29B (2422)), nevertheless appeared as two distinct clusters of naïve/memory B cells in the gene expression-annotated open chromatin clustering (Figure 29A; 2722). Further analysis of the differential gene expression in the gene expression-annotated open chromatin clustering plot revealed two distinct populations of cells (subcluster 1 (2902) and subcluster 2 (2904)) that were visibly separated when using open chromatin clustering (Figure 29C) but were obscured when viewing gene expression cluster analysis alone (Figure 29D). Gene expression analysis of subcluster 1 and subcluster 2 identified subcluster 1 as prospective memory B cells (relatively higher Ig, relatively lower naïve B cell

associated transcripts) and subcluster 2 as prospective naïve B cells (relatively lower Ig, relatively higher naïve B cell associated transcripts), as illustrated in Figure 29E.

[0073] Figures 30A and 30B collectively illustrate cell type annotation of a tumor sample using gene expression (GEX) markers and transcription factor (ATAC) accessibility, in accordance with an embodiment of the present disclosure. Gene expression clustering identified five cell type categories (B cells, replicating B cells, monocytes, T cells, and replicating T cells) (Figure 30A). Open chromatin clustering identified three cell type categories (B cells, monocytes, and T cells) (Figure 30B).

[0074] Figures 31A, 31B, and 31C collectively illustrate the identification of tumor from normal B cells using mutational load (SNVs) (Figure 31A) and the *BANK1* pathway (markers of B cell hyper-activation) (Figure 31B), in gene expression data from the lymph node tumor sample, in accordance with an embodiment of the present disclosure. Two sub-populations of B cells (tumor B cells and normal B cells) in cell populations clustered by gene expression profiling are identified in Figure 31C.

[0075] Figures 32A, 32B, and 32C collectively illustrate gene expression annotation of tumor cells to annotate and identify open chromatin cell populations, in accordance with an embodiment of the present disclosure.

[0076] Figures 33A, 33B, 33C and 33D collectively illustrate, on the basis of covariance of open chromatin and gene expression, the identification of a candidate enhancer region that regulates the expression of the IL4R specifically in the tumor B cells, in accordance with an embodiment of the present disclosure.

[0077] Figures 34A, 34B, and 34C illustrate further comparative analysis of STAT protein activation across multiple cell populations identified by open chromatin clustering (Figure 34A), including epigenetic modifications (Figure 34B) and quantification (Figure 34C), in accordance with an embodiment of the present disclosure.

DETAILED DESCRIPTION

[0078] Reference will now be made in detail to embodiments, examples of which are illustrated in the accompanying drawings. In the following detailed description, numerous specific details are set forth in order to provide a thorough understanding of the present

disclosure. However, it will be apparent to one of ordinary skill in the art that the present disclosure may be practiced without these specific details. In other instances, well-known methods, procedures, components, circuits, and networks have not been described in detail so as not to unnecessarily obscure aspects of the embodiments.

[0079] The implementations described herein provide various technical solutions to detect a pattern in datasets. An example of such datasets are datasets arising from whole transcriptome shotgun sequencing pipelines that quantify gene expression in single cells in counts of transcript reads mapped to genes. Details of implementations are now described in conjunction with the Figures.

[0080] *General terminology.*

[0081] Specific terminology is used throughout this disclosure to explain various aspects of the apparatus, systems, methods, and compositions that are described. This sub-section includes explanations of certain terms that appear in later sections of the disclosure. To the extent that the descriptions in this section are in apparent conflict with usage in other sections of this disclosure, the definitions in this section will control.

[0082] *(i) Subject*

[0083] A “subject” is an animal, such as a mammal (e.g., human or a non-human simian), or avian (e.g., bird), or other organism, such as a plant. Examples of subjects include, but are not limited to, a mammal such as a rodent, mouse, rat, rabbit, guinea pig, ungulate, horse, sheep, pig, goat, cow, cat, dog, primate (e.g. human or non-human primate); a plant such as *Arabidopsis thaliana*, corn, sorghum, oat, wheat, rice, canola, or soybean; an algae such as *Chlamydomonas reinhardtii*; a nematode such as *Caenorhabditis elegans*; an insect such as *Drosophila melanogaster*, mosquito, fruit fly, honey bee or spider; a fish such as zebrafish; a reptile; an amphibian such as a frog or *Xenopus laevis*; a *Dictyostelium discoideum*; a fungi such as *Pneumocystis carinii*, *Takifugu rubripes*, yeast, *Saccharomyces cerevisiae* or *Schizosaccharomyces pombe*; or a *Plasmodium falciparum*.

[0084] *(ii) Nucleic acid and Nucleotide.*

[0085] The terms “nucleic acid” and “nucleotide” are intended to be consistent with their use in the art and to include naturally-occurring species or functional analogs thereof.

Particularly useful functional analogs of nucleic acids are capable of hybridizing to a nucleic acid in a sequence-specific fashion or are capable of being used as a template for replication of a particular nucleotide sequence. Naturally-occurring nucleic acids generally have a backbone containing phosphodiester bonds. An analog structure can have an alternate backbone linkage including any of a variety of those known in the art. Naturally-occurring nucleic acids generally have a deoxyribose sugar (e.g., found in deoxyribonucleic acid (DNA)) or a ribose sugar (e.g., found in ribonucleic acid (RNA)).

[0086] A nucleic acid can contain nucleotides having any of a variety of analogs of these sugar moieties that are known in the art. A nucleic acid can include native or non-native nucleotides. In this regard, a native deoxyribonucleic acid can have one or more bases selected from the group consisting of adenine (A), thymine (T), cytosine (C), or guanine (G), and a ribonucleic acid can have one or more bases selected from the group consisting of uracil (U), adenine (A), cytosine (C), or guanine (G). Useful non-native bases that can be included in a nucleic acid or nucleotide are known in the art.

[0087] (iii) *Barcode*.

[0088] A “barcode” is a label, or identifier, that conveys or is capable of conveying information (e.g., information about an analyte in a sample, a bead, and/or a capture probe). A barcode can be part of an analyte, or independent of an analyte. A barcode can be attached to an analyte. A particular barcode can be unique relative to other barcodes.

[0089] Barcodes can have a variety of different formats. For example, barcodes can include 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 or to another moiety or structure 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 or during sequencing of the sample. Barcodes can allow for identification and/or quantification of individual sequencing-reads (e.g., a barcode can be or can include a unique molecular identifier or “UMI”). In some embodiments, a barcode includes two or more sub-barcodes that together function as a single barcode. For example, a polynucleotide barcode can include two or more polynucleotide sequences (e.g., sub-barcodes) that are separated by one or more non-barcode sequences. More details on barcodes and UMIs is disclosed in United States Provisional

Patent Application No. 62/980,073, entitled “Pipeline for Analysis of Analytes,” filed February 21, 2020, attorney docket number 104371-5033-PR01, which is hereby incorporated by reference.

[0090] (iv) *Biological samples.*

[0091] As used herein, a “biological sample” is obtained from the subject for analysis using any of a variety of techniques including, but not limited to, biopsy, surgery, and laser capture microscopy (LCM), and generally includes tissues or organs and/or other biological material from the subject. Biological samples can include one or more diseased cells. A diseased cell can have altered metabolic properties, gene expression, protein expression, and/or morphologic features. Examples of diseases include inflammatory disorders, metabolic disorders, nervous system disorders, and cancer. Cancer cells can be derived from solid tumors, hematological malignancies, cell lines, or obtained as circulating tumor cells.

[0092] *System.*

[0093] Figure 1A is a block diagram illustrating a visualization system 100 in accordance with some embodiments. The device 100 in some embodiments includes one or more processing units CPU(s) 102 (also referred to as processors), one or more network interfaces 104, a user interface 106 comprising a display 108 and an input module 110, a non-persistent 111, a persistent memory 112, and one or more communication buses 114 for interconnecting these components. The one or more communication buses 114 optionally include circuitry (sometimes called a chipset) that interconnects and controls communications between system components. The non-persistent memory 111 typically includes high-speed random access memory, such as DRAM, SRAM, DDR RAM, ROM, EEPROM, flash memory, whereas the persistent memory 112 typically includes CD-ROM, digital versatile disks (DVD) or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, magnetic disk storage devices, optical disk storage devices, flash memory devices, or other non-volatile solid state storage devices. The persistent memory 112 optionally includes one or more storage devices remotely located from the CPU(s) 102. The persistent memory 112, and the non-volatile memory device(s) within the non-persistent memory 112, comprise non-transitory computer readable storage medium.

[0094] In some implementations, the non-persistent memory 111 or alternatively the non-transitory computer readable storage medium stores the following programs, modules and data structures, or a subset thereof, sometimes in conjunction with the persistent memory 112:

- an optional operating system 116, which includes procedures for handling various basic system services and for performing hardware dependent tasks;
- an optional network communication module (or instructions) 118 for connecting the visualization system 100 with other devices, or a communication network; and
- a browser module 119 for selecting a discrete attribute value dataset 120 from persistent memory and presenting information about the discrete attribute value dataset 120.

[0095] In some implementations, one or more of the above identified elements are stored in one or more of the previously mentioned memory devices, and correspond to a set of instructions for performing a function described above. The above identified modules, data, or programs (*e.g.*, sets of instructions) need not be implemented as separate software programs, procedures, datasets, or modules, and thus various subsets of these modules and data may be combined or otherwise re-arranged in various implementations. In some implementations, the non-persistent memory 111 optionally stores a subset of the modules and data structures identified above. Furthermore, in some embodiments, the memory stores additional modules and data structures not described above. In some embodiments, one or more of the above identified elements is stored in a computer system, other than that of visualization system 100, that is addressable by visualization system 100 so that visualization system 100 may retrieve all or a portion of such data when needed.

[0096] Referring to Figures 1A, 1B, 1C, and 1D, persistent memory 112 stores one or more discrete attribute value datasets 120 (*e.g.*, 120-1, ..., 120-X). In some embodiments, a discrete attribute value dataset 120 comprises a feature-barcode matrix 121, an array of chromosome indices 180, an array of starting positions 183, an array of ending positions 185, and a feature-linkage matrix 187.

[0097] *Feature-barcode matrix.*

[0098] The feature-barcode matrix 121 comprises, for each respective cell 126 in a plurality of cells, (i) a discrete attribute value (gene expression value) 124 for each gene 122 in a plurality

of genes, and (ii) assay for transposase accessible chromatin (ATAC) peaks values for each peak from a plurality of peaks. Typically, a discrete attribute value 124 for a given gene in a given cell is the amount of mRNA that was measured within the given cell for the given gene. Advantageously, the ATAC peak values and the discrete attribute values originate from the same cells.

[0099] As illustrated in Figures 1B and 1C, a feature-barcode matrix 121-1 (shown by way of example) includes information related to cell 1 (126-1), cell 2 (126-2) and other cells up to cell Y (126-Y). As shown for cell 1 (126-1), the cell 1 (126-1) includes discrete attribute value 124-1-1 for gene 1 of cell 1 (122-1-1), discrete attribute value for gene 2 of cell 1 (124-2-1), and other discrete attribute values up to discrete attribute value for gene M of cell 1 (124-M-1). In some embodiments, there are discrete attributes values for 10 or more, 100 or more, 1000 or more, or ten thousand or more genes in the feature-barcode matrix 121 for each cell in the plurality of cells represented in the feature-barcode matrix 121.

[00100] As shown for cell 1 (126-1), the cell 1 (126-1) further comprises ATAC peak fragment count 125-1-1 for ATAC peak 1 of cell 1 (123-1-1), ATAC peak fragment count 125-2-1 for ATAC peak 2 of cell 1 (123-2-1), and other ATAC peak fragment counts up to ATAC peak fragment count for ATAC peak L of cell 1 (125-L-1). In some embodiments, there are ATAC peak fragment counts for 10 or more, 100 or more, 1000 or more, or ten thousand or more ATAC peaks in the feature-barcode matrix 121 for each cell in the plurality of cells represented in the feature-barcode matrix 121.

[00101] For the ATAC data, in typical embodiments, there are no measures of gene expression. The metric for ATAC peak counts 125 is fragments (or UMIs) per called peak 123, where the peak 123 corresponds to a genomic region of accessible chromatin. Thus, for ATAC data, the feature-barcode matrix contains counts of fragments 125 in peaks 123 instead of gene expression.

[00102] In some embodiments, the feature-barcode matrix 121-1 includes ATAC peak data and gene expression discrete attribute values for 500 or more cells, 1000 or more cells, 10,000 or more cells, 15,000 or more cells, 20,000 or more cells, 25,000 or more cells, 30,000 or more cells, or 50,000 or more cells.

[00103] In some embodiments, the dataset further stores a plurality of GEX principal component values 164 and/or a two-dimensional GEX datapoint 166 and/or one or more GEX category 170 assignments for each respective cell 126 in the plurality of cells. Figure 1B illustrates by way of example GEX principal component value 1 (164-1-1) through principal component value N (164-1-N) stored for cell 126-1 of the feature-barcode matrix 121-1 of the discrete attribute value dataset 120-1. The GEX principal components 164 are based upon the principal component analysis of the discrete attribute values 124 across the feature-barcode matrix and not the fragment counts 125.

[00104] In some embodiments, application of principal component analysis can drastically reduce (*e.g.*, reduce by at least 5-fold, at least 10-fold, at least 20-fold, or at least 40-fold) the dimensionality of the GEX data (*e.g.* from approximately 20,000 dimensions to ten dimensions). That is, principal component analysis is used to assign each respective cell a plurality of GEX principal components that collectively describe the variation in the respective cell's mRNA expression levels with respect to expression levels of corresponding mRNA of other cells in the dataset. That is, each respective cell has the same set of GEX principal components, and the principal component analysis assigns different values to these GEX principal components for each respective cell in accordance with the variance in mRNA expression exhibited by the respective cell relative to the other cells.

[00105] Figure 1B also illustrates the GEX cluster assignment for cell 1 (158-1), GEX category assignment 1 for cell 1 (170-1-1), including class 1 for GEX category assignment 1 (172-1-1-1) through class M for GEX category assignment 1 (172-1-1-M), through GEX category assignment Q for cell 1 (170-1-Q), including class 1 for GEX category assignment Q (172-1-Q-1) through class Z for GEX category assignment Q (172-1-Q-Z). These category and class assignments are based upon either (i) the principal component analysis of the discrete attribute values 124 across the plurality of cells represented by the feature-barcode matrix or (ii) the discrete attribute values 124 themselves, and not the ATAC fragment counts 125. In some embodiments each cell represented in the feature-barcode matrix 121-1 has five or more GEX principal component values, ten or more GEX principal component values, or twenty or more GEX principal component values. In some embodiments, the plurality of cells represented by the feature-barcode matrix are clustered into 3 or more, 4 or more, 5 or more, 6 or more, 7 or more, 8 or more, 9 or more, 10 or more, 15 or more or 20 or more GEX clusters.

[00106] In some alternative embodiments, the discrete attribute value dataset 120 stores a GEX two-dimensional datapoint 166 for each respective cell 126 in the plurality of cells (*e.g.*, two-dimensional datapoint for cell 1 166-1 in Figure 1B) based upon the discrete attribute values but not the fragment counts for peaks, but does not store the plurality of principal component values 164. In some embodiments, each discrete attribute value represents a number of mRNA measured in a respective cell that maps to a respective gene in the genome of the cell, and the discrete attribute value dataset further comprises the total RNA counts per cell.

[00107] In some embodiments, the dataset further stores a plurality of ATAC principal component values 165 and/or a two-dimensional ATAC datapoint 167 and/or one or more ATAC category 171 assignments for each respective cell 126 in the plurality of cells.

[00108] In some embodiments, application of principal component analysis can drastically reduce (*e.g.*, reduce by at least 5-fold, at least 10-fold, at least 20-fold, or at least 40-fold) the dimensionality of the ATAC data (*e.g.* from approximately 20,000 dimensions to ten dimensions). That is, principal component analysis is used to assign each respective cell a plurality of ATAC principal components that collectively describe the variation in the respective cell's ATAC fragment counts of ATAC peaks with respect to ATAC fragment counts of corresponding ATAC peaks of other cells in the dataset. That is, each respective cell has the same set of ATAC principal components (*e.g.*, ATAC components 1 through 10), and the principal component analysis assigns different values to these ATAC principal components for each respective cell in accordance with the variance in ATAC fragment count for ATAC peaks exhibited by the respective cell relative to the other cells.

[00109] Figure 1C illustrates by way of example ATAC principal component value 1 (165-1-1) through principal component value P (165-1-P) stored for cell 126-1 of the feature-barcode matrix 121-1 of the discrete attribute value dataset 120-1. The ATAC principal components 165 are based upon the principal component analysis of the ATAC fragment counts 125 across the cells represented by the feature-barcode matrix and not the GEX discrete attribute values 124. Figure 1C also illustrates the ATAC cluster assignment for cell 1 (159-1), ATAC category assignment 1 for cell 1 (171-1-1), including class 1 for category assignment 1 (173-1-1-1) through class M for category assignment 1 (173-1-1-M), through category assignment P for cell 1 (171-1-P), including class 1 for category assignment P (173-1-P-1) though class W for

category assignment P 172-1-P-W. These category and class assignments are based upon either (i) the principal component analysis of the ATAC fragment counts 125 within the feature-barcode matrix or (ii) the ATAC fragment counts 125 themselves, and not the GEX discrete attribute values 125. In some embodiments each cell represented in the feature-barcode matrix 121-1 has five or more ATAC principal component values, ten or more ATAC principal component values, or twenty or more ATAC principal component values. In some embodiments, the plurality of cells represented by the feature-barcode matrix are clustered into 3 or more, 4 or more, 5 or more, 6 or more, 7 or more, 8 or more, 9 or more, 10 or more, 15 or more or 20 or more ATAC clusters. In some embodiments, the number of ATAC clusters represented by the plurality of cells of a given feature-barcode matrix is independent of (e.g. not the same as) the number of GEX clusters represented by the plurality of cells of the given feature-barcode matrix. As a nonlimiting example, in one instance the plurality of cells represented by the feature-barcode matrix are divided into 13 GEX clusters, and this same plurality of cells represented by the feature-barcode matrix are divided into 17 ATAC clusters. Of course, although number of ATAC clusters represented by the plurality of cells of a given feature-barcode matrix is independent of the number of GEX clusters represented by the plurality of cells of the given feature-barcode matrix, it is possible that the number of clusters is coincidentally the same. As a nonlimiting example of this, in another instance the plurality of cells represented by the feature-barcode matrix are divided into 10 GEX clusters, and this same plurality of cells represented by the feature-barcode matrix are divided into 10 ATAC clusters.

[00110] In some embodiments, the feature-barcode matrix 121 also includes inferred counts 194/195 for each gene 193, for each cell, across the plurality of cells by the feature-barcode matrix 121 based on the ATAC data. In such embodiments, as illustrated in Figure 20, the “Nearby Gene” aggregate count 194 sums single ATAC peaks within a given cell that are close to a given gene while the “Promoter Sum” aggregate count 195 sums ATAC peaks that overlap the gene’s promoter region within a given cell. For example, if there are three called ATAC peaks in the dataset that are close to the GZMB gene, and two peaks that are close to the promoter, the GZMB Nearby Gene entry 194 will contain the sum of the fragment counts across those three ATAC peaks for each cell barcodes (e.g., for each cell represented by the feature-barcode matrix 121), and the GZMB Promoter Sum entry 195 will contain the sum of the fragment counts across the two ATAC peaks adjacent to the promoter for each cell barcode (e.g.,

for each cells represented by the feature-barcode matrix 121). As such, for GZMB there would be a plurality of entries 194 and 195, each such entry 194 and 195 representing a different cell in the plurality of cells.

[00111] *Projections.*

[00112] In some embodiments, a 2-D t-Distributed Stochastic Neighboring Entities (t-SNE) projection 196 is computed based on the significant GEX principal component values 164 across the plurality of cells 126 represented by the feature-barcode matrix. t-SNE is a machine learning algorithm that is used for dimensionality reduction. *See* van der Maaten and Hinton, 2008, “Visualizing High-Dimensional Data Using t-SNE,” *Journal of Machine Learning Research* 9, 2579-2605, which is hereby incorporated by reference. The nonlinear dimensionality reduction technique t-SNE is particularly well-suited for embedding high-dimensional data (here, the GEX principal components values 164 computed for each measured cell based upon the measured discrete attribute value 124 (*e.g.*, expression level) of each gene 122 (*e.g.*, expressed mRNA) in a respective cell 126 as determined by principal component analysis into a space of two, which can then be visualized as a two-dimensional visualization. In some embodiments, t-SNE is used to model each high-dimensional object (the GEX principal components 164 of each measured cell) as a two-dimensional point in such a way that similarly expressing cells are modeled as nearby two-dimensional datapoints and dissimilarly expressing cells are modeled as distant two-dimensional datapoints in the t-SNE projection. The t-SNE algorithm comprises two main stages. First, t-SNE constructs a probability distribution over pairs of high-dimensional cell vectors (vector of principal component values) in such a way that similar cell vectors (cells that have similar values for their GEX principal components 164 and thus presumably have similar discrete attribute values 124 across the plurality of genes) have a high probability of being picked, while dissimilarly dissimilar cell vectors (cells that have dissimilar values for their GEX principal components and thus presumably have dissimilar discrete attribute values 124 across the plurality of genes) have a small probability of being picked. Second, t-SNE defines a similar probability distribution over the plurality of cells 126 in the low-dimensional map, and it minimizes the Kullback–Leibler divergence between the two distributions with respect to the locations of the points in the map. In some embodiments the t-SNE algorithm uses the Euclidean distance between objects as the base of its similarity metric. In other embodiments, other

distance metrics are used (e.g., Chebyshev distance, Mahalanobis distance, Manhattan distance, *etc.*).

[00113] In some embodiments, a t-SNE projection 198 is also computed based on the significant ATAC principal component values 165 across the plurality of cells 126 represented by the feature-barcode matrix. In such embodiments, t-SNE embeds the ATAC principal components values 165 computed for each measured cell 126 based upon the measured ATAC fragment counts 125 of each ATAC peak 123 in a respective cell 126 as determined by principal component analysis into a space of two, which can then be visualized as a two-dimensional visualization. In some such embodiments, t-SNE is used to model each high-dimensional object (the ATAC principal components 165 of each measured cell) as a two-dimensional point. As in the case of the GEX t-SNE calculations described above, the ATAC t-SNE algorithm comprises two main stages. First, the ATAC t-SNE constructs a probability distribution over pairs of high-dimensional cell vectors (vector of ATAC principal component values 165) in such a way that similar cell vectors (cells that have similar values for their ATAC principal components 165 and thus presumably have similar ATAC fragment counts 125 across the plurality of ATAC peaks 123) have a high probability of being picked, while dissimilarly dissimilar cell vectors (cells that have dissimilar values for their ATAC principal components and thus presumably have dissimilar ATAC fragment counts 125 across the plurality of ATAC peaks 123) have a small probability of being picked. Second, t-SNE defines a similar probability distribution over the plurality of cells 126 in the low-dimensional map, and it minimizes the Kullback–Leibler divergence between the two distributions with respect to the locations of the points in the map. In some embodiments the ATAC t-SNE algorithm uses the Euclidean distance between objects as the base of its similarity metric. In other embodiments, other distance metrics are used (*e.g.*, Chebyshev distance, Mahalanobis distance, Manhattan distance, *etc.*).

[00114] In some embodiments, a GEX 2-D uniform manifold approximation and projection (UMAP) projection 197 is computed based on the significant GEX principal component values 164 across the plurality of cells 126 represented by the feature-barcode matrix. UMAP is described in McInnes and Healy, 2018, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” ArXiv e-prints 1802.03426, which is hereby incorporated by reference. Correspondingly, in some embodiments, an ATAC 2-D UMAP projection 199 is computed based on the significant ATAC principal component values 165 across the plurality of

cells 126 represented by the feature-barcode matrix. As a result, users are able to explore classifications and scatter plots generated from the two separate analytes: ATAC and GEX.

[00115] *Genomic location of ATAC peaks and GEX genes.*

[00116] In some embodiments, a discrete attribute dataset 120 further includes the genomic location of each ATAC peak and each gene. In some embodiments this information, for each such feature, is stored in each of three location arrays, namely, an array of chromosome indices 180, an array of starting positions 183, and an array of ending positions 185. Thus, the boundaries of a feature i (gene 122 or ATAC peak 123) in the feature-barcode matrix 121 are defined by the i^{th} value of these three arrays. In this way, the location of these features in a genome plot can be determined, and also distances between linked features can be determined. For instance, the chromosome number of a given gene i is located at the i^{th} entry 182 of array of chromosome index 180, the starting position of given gene i is located at the i^{th} entry 184 of the array of starting positions 183, and the ending position of given gene i is located at the i^{th} entry 186 of the array of ending positions 185. In this way, for a feature i , the chromosome, start position and end position of the feature can be found at `location.chromosome[i]`, `location.starts[i]` and `location.ends[i]`, where `location.chromosome[i]`, `location.starts[i]`, `location.ends[i]`, are respectively the array of chromosome indices 180, array of starting positions 183, and array of ending positions 185. In some embodiments, the array of chromosome indices 180, array of starting positions 183, and array of ending positions 185 are integer arrays that can be stored as gzipped blocks.

[00117] *Feature-linkage matrix.*

[00118] Measurement of both gene expression and chromatin accessibility within the same cell provides for the opportunity to link correlated expression and accessibility patterns. Finding strong correlations between open DNA regions and mRNA transcribed from nearby sequences can reveal how the instructions encoded in our DNA trigger the cellular programs executed. For instance, given a feature (gene or peak) of interest, querying the linkage matrix and other structures in the discrete attribute value dataset 120 (e.g., in the form of a .cloupe file) efficiently provides (i) the chromosomal location of the gene or peak of interest, (ii) the identities and locations of the features linked to that gene or peak, and (iii) the correlations of those linkages, and the significance of those linkages. Accordingly, and referring to Figure 1D, some

embodiments of the discrete attribute value dataset 120 further include a feature-linkage matrix 187 between peak features, and between peak to gene features. Because linkages are symmetric (*e.g.*, $\text{linkage}(a, b) = \text{linkage}(b, a)$), in preferred embodiments, such linkages are stored in triangular matrix form.

[00119] In some embodiments, there is a constraint that, in order to be included in the matrix, features (ATAC peaks 123 and genes 122) must be on the same chromosome and within some threshold distance from each other. In some embodiments, this threshold distance is 1 megabase. That is, the two peaks (or peak and gene) must map to within 1 megabase of each other on the same chromosome to be registered as linked within the feature linkage matrix 187. In alternative embodiments, the threshold distance is 1.5 megabases, 2.0 megabases, 2.5 megabases, 3.0 megabases, 3.5 megabases, 4.0 megabases, 4.5 megabases, 5.0 megabases, 5.5 megabases, 6.0 megabases, between 1.5 and 6.0 megabases, or some value greater than 6.0 megabases. Since the feature-barcode matrix 121 includes genes and ATAC peaks from across the gigabases of human and other genomes, this restriction ensures that the feature-linkage matrix 187 is extremely sparse. Thus, in some embodiments, the feature-linkage matrix is stored in persistent memory 112 as a sparse matrix, in compressed-sparse row (CSR) format. In some embodiments, there are four arrays in a feature-linkage matrix 187, a pointer array (P) (188), an index array (X) (190), a correlation array (C) (191), and a significance array (S) 192.

[00120] The correlation array C (191) contains R-value correlations between linked features (from -1, indicating perfect opposites, to 1, indicating perfect alignment). The significance array S (192) contains the statistical significances of the linkages, defined as the negative log of the false discovery rate. A higher significance means the linkage is less likely to be due to random sampling.

[00121] The pointer array (P) 188 is of length $L+M+1$, where $L+M$ is the total number of genes and ATAC peaks represented in the corresponding in the feature barcode matrix 121. For instance, if the feature-barcode matrix provides GEX discrete attribute values for a total of 21 different genes and a total of 30 different ATAC peaks across the plurality of cells 126 represented by the feature-barcode matrix 121, the pointer array has a length of $21+30+1=52$. In some embodiments, the values $P[i]$ and $P[i+1]$ in the pointer array 188 define the minimum and maximum indices in the other three arrays (index array 190, correlation array 191, and

significance array 192) to retrieve information about features linked to the feature i of the feature-barcode matrix. The subset $X[P[i]]: X[P[i+1]]$ contains the row indices of the other features linked to feature i , $C[P[i]]: C[P[i+1]]$ contains the respective correlations between feature i and the other features in X , and likewise $S[P[i]]: S[P[i+1]]$ contains the respective significances between the linkages between feature i and the features over the same range in the index array X (190). For instance, for feature 20, the 20th and the (20+1)th index in the pointer array 180 is consulted to retrieve two values (*e.g.*, 1137 and 1530). These two values, 1137 and 1530, define the minimum and maximum indices that used within the index array 190 to learn the identity of each gene or peak that is linked to feature 20. These two values, 1137 and 1530, further define the minimum and maximum indices that used within the correlation array 191 to learn the correlation of each gene or peak that is linked to feature 20. Finally, these two values, 1137 and 1530, further define the minimum and maximum indices that are used within the significance array 192 to learn the significance of each gene or peak that is linked to feature 20. The sparse format of the index array 190, correlation array 191, and significance array 192 allows efficient storage in persistent memory 112 within a discrete attribute value dataset 120 (*e.g.*, a .cloupe file). Furthermore, in some embodiments, the index array 190, correlation array 191, and significance array 192 are stored in blocked gzip format with indices, such that lookups of the linkages of randomly selected features are efficient, and do not require loading the entire feature-linkage matrix into non-persistent memory 111.

[00122] Although Figures 1A, 1B, 1C, and 1D depict a “visualization system 100,” the figures are intended more as functional description of the various features which may be present in computer systems than as a structural schematic of the implementations described herein. In practice, and as recognized by those of ordinary skill in the art, items shown separately could be combined and some items could be separated. Moreover, although Figure 1A depicts certain data and modules in non-persistent memory 111, some or all of these data and modules may be in persistent memory 112. Further, while discrete attribute value dataset 120 is depicted as resident in persistent memory 112, a portion of discrete attribute value dataset 120 is, in fact, resident in non-persistent memory 111 at various stages of the disclosed methods.

[00123] *Methods.*

[00124] While a system in accordance with the present disclosure has been disclosed with reference to Figures 1A and 1B, a method in accordance with the present disclosure is now detailed with reference to Figures 2A, 2B, and 2C.

[00125] *Block 202.* One aspect of the present disclosure provides a visualization system 100. The visualization system 100 comprises one or more processing cores 102, a non-persistent memory 111 and a persistent memory 111, the persistent memory and the non-persistent memory collectively storing instructions for performing a method. A non-limiting example of a visualization system is collectively illustrated in Figures 1A, 1B, 1C, and 1D. As discussed above, it will be appreciated that the persistent memory and/or the non-persistent memory can be on a single computer, distributed across a network of computers, be represented by one or more virtual machines, or be part of a cloud computing architecture.

[00126] *Block 204 - Storing a discrete attribute value dataset 120 in persistent memory.* The systems and methods of the present disclosure provide for storing a discrete attribute value dataset 120 in persistent memory 112. Referring to Figures 1B, 1C, and 1D, the discrete attribute value dataset 120 comprises a corresponding discrete attribute value 124 for each gene 122 in a plurality of genes for each respective cell 126 in a plurality of cells. The dataset 120 further comprises a corresponding ATAC fragment count 125 for each ATAC peak 123 in a plurality of ATAC peaks for each respective cell 126 in the plurality of cells. Figure 3 illustrates the selection of a particular discrete attribute value dataset 120 using browser module 119. In particular, Figure 3 illustrates how the browser module 119 provides some information regarding a given discrete attribute value dataset 120 such as its type, its name, the number of cells 126 represented by the discrete attribute value dataset 120 (if any), and the last time the discrete attribute value dataset was accessed.

[00127] In some embodiments, each discrete attribute value 124 is a count of transcript reads within the cell that map to a respective gene in the plurality of genes. In some embodiments, the discrete attribute values 124 of a discrete attribute value dataset 120 represents a whole transcriptome shotgun sequencing experiment that quantifies gene expression from a single cell in counts of transcript reads mapped to the genes. In some such embodiments, microfluidic partitions are used to partition very small numbers of mRNA molecules and to barcode those partitions. In some such embodiments, where discrete attribute values are measured from single

cells, the microfluidic partitions are used to capture individual cells within each microfluidic droplet and then pools of single barcodes within each of those droplets are used to tag all of the contents (*e.g.*, genes 122) of a given cell. For example, in some embodiments, a pool of ~750,000 barcodes is sampled to separately index each cells' transcriptome by partitioning thousands of cells into nanoliter-scale Gel Bead-In-EMulsions (GEMs), where all generated cDNA share a common barcode. Libraries are generated and sequenced from the cDNA and the barcodes are used to associate individual reads back to the individual partitions. In other words, each respective droplet (GEM) is assigned its own barcode and all the contents (*e.g.*, mRNA for genes) in a respective droplet are tagged with the barcode unique to the respective droplet. In some embodiments, such droplets are formed as described in Zheng *et al.*, 2016, Nat Biotechnol. 34(3): 303-311; or in *See the Chromium, Single Cell 3' Reagent Kits v2. User Guide*, 2017, 10X Genomics, Pleasanton, California, Rev. B, page, 2, each of which is hereby incorporated by reference. In some alternative embodiments, equivalent 5' chemistry is used rather than the 3' chemistry disclosed in these references.

[00128] In some embodiments there are tens, hundreds, thousands, tens of thousands, or one hundreds of thousands of such microfluidic droplets. In some such embodiments, at least seventy percent, at least eighty percent, at least ninety percent, at least ninety percent, at least ninety-five percent, at least ninety-eight percent, or at least ninety-nine percent of the respective microfluidic droplets contain either no cell 126 or a single cell 126 while the remainder of the microfluidic droplets contain two or more cells 126. In other words, to achieve single cell resolution, the cells are delivered at a limiting dilution, such that the majority (~90-99%) of generated nanoliter-scale gel bead-in-emulsions (GEMs) contains no cell, while the remainder largely contain a single cell. *See the Chromium, Single Cell 3' Reagent Kits v2. User Guide*, 2017, 10X Genomics, Pleasanton, California, Rev. B, page, 2, which is hereby incorporated by reference. In some alternative embodiments, equivalent 5' chemistry is used rather than the 3' chemistry disclosed in this reference.

[00129] Within an individual droplet, gel bead dissolution releases the amplification primer into the partitioned solution. In some embodiments, upon dissolution of the single cell 3' Gel Bead in a GEM, primers containing (i) an Illumina R1 sequence (read 1 sequencing primer), (ii) a 16 bp 10x Barcode, (iii) a 10 bp Unique Molecular Identifier (UMI) and (iv) a polydT primer sequence are released and mixed with cell lysate and Master Mix. Incubation of the GEMs then

produces barcoded, full-length cDNA from poly-adenylated mRNA. After incubation, the GEMs are broken and the pooled fractions are recovered. *See* the Chromium, Single Cell 3' Reagent Kits v2. User Guide, 2017, 10X Genomics, Pleasanton, California, Rev. B, page, 2, which is hereby incorporated by reference. In some such embodiments, silane magnetic beads are used to remove leftover biochemical reagents and primers from the post GEM reaction mixture. Full-length, barcoded cDNA is then amplified by PCR to generate sufficient mass for library construction.

[00130] In this way, the genes 122 can be mapped to individual genes in the genome of a species and therefore they can be sequenced and, furthermore, the mRNA for genes 122 of a given cell 126 can be distinguished from the mRNA of genes of another cell 126 based on the unique barcode. This contrasts to bulk sequencing techniques in which all the cells are pooled together and the measurement profile is that of the genes of the whole collection of the cells without the ability to distinguish the measurement signal of genes by individual cells. An example of such measurement techniques is disclosed in United States Patent Publication No. 2015/0376609, which is hereby incorporated by reference. As such, in some embodiments, the sequence reads mapping to each gene in a particular cell in the plurality of cells is barcoded with a first barcode that is unique to the particular cell.

[00131] In some embodiments, physical measurement of RNA gene expression and DNA chromatin accessibility is from the same cells. That is, the assay measures both RNA and accessible chromatin from the same cell. Figure 19 shows a non-limiting example schematic of such an assay. In particular, panel 1900 of Figure 19 shows a workflow corresponding to processing of chromatin from a cell, cell bead, or cell nucleus, and panel 1950 of Figure 19 shows a workflow corresponding to processing of an mRNA molecule from the cell, cell bead, or cell nucleus.

[00132] As shown in panel 1900, in bulk solution, chromatin included within a cell, cell bead, or cell nucleus is processed (*e.g.*, as described herein) to provide a template nucleic acid fragment (*e.g.*, tagged fragment) 1904 comprising insert sequence 1908 (*e.g.*, region of open chromatin) and a complement thereof, transposon end sequences 1906 and complements thereof, sequencing primer or portion thereof 1902 (*e.g.*, an R1 sequence), sequencing primer or portion thereof 1910 (*e.g.*, an R2 sequence), and gaps 1907. Template nucleic acid fragment 1904 may

then be partitioned to a droplet or well, as described herein. Within the partition, the cell, cell bead, or cell nucleus comprising template nucleic acid fragment 1904 may be lysed, permeabilized, or otherwise processed to provide access to template nucleic acid fragment 1904 (and one or more RNA molecules) therein. The partition may comprise splint sequence 1912. In some embodiments, the splint sequence 1912 comprises a first sequence 1902' that is complementary to sequencing primer or portion thereof 1902 and a second sequence 1924. In some embodiments sequence 1924 comprises a blocking group (*e.g.*, a 3' blocking group) that prevents extension by reverse transcription. The partition may also include a bead (*e.g.*, gel bead) 1916 coupled to nucleic acid barcode molecules 1918a and 1918b. Nucleic acid barcode molecule 1918a may comprise a flow cell adapter sequence 1920a (*e.g.*, a P5 sequence), a barcode sequence 1922a, and an overhang sequence 1924' that is complementary to sequence 1924 of the splint sequence. Sequence 1924 may hybridize to sequence 1924' to provide a partially double-stranded nucleic acid molecule comprising the sequences of nucleic acid barcode molecule 1918a and the template nucleic acid fragment 1904. In some embodiments, sequence 1924' of nucleic acid barcode molecule 1918a is ligated (*e.g.*, using a ligase) 1926 to sequence 1902 of template nucleic acid fragment 1904. In some instances, 1904 is phosphorylated using a suitable kinase enzyme (*e.g.*, polynucleotide kinase (PNK), such as T4 PNK). In some embodiments, PNK and ATP are added in bulk in the tagmentation reaction (*e.g.*, ATAC) and/or prior to partitioning the cells, cell beads, or cell nuclei. 15U of PNK with 1mM of ATP may be spiked into the tagmentation reaction. For example, less than 95U of PNK may be spiked into the tagmentation reaction. The contents of the partition may then be recovered in bulk solution (*e.g.*, a droplet may be broken) to provide the partially double-stranded nucleic acid molecule comprising nucleic acid barcode molecule 1918a attached to template nucleic acid fragment 1904 in bulk solution. In bulk solution, gaps 1907 may be filled 1928 via a gap filling extension process (*e.g.*, using a DNA polymerase) to provide a double-stranded nucleic acid molecule. This molecule may undergo amplification (*e.g.*, PCR) 1930 to provide a double-stranded amplification product 1932 that includes sequences of the nucleic acid barcode molecule 1918a, the original chromatin molecule, and, optionally, an additional sequence 1934 that may be a flow cell adapter sequence (*e.g.*, a P7 sequence). Gaps may be filled in the partition prior to bulk processing.

[00133] In parallel to the chromatin workflow of panel 1900, an RNA molecule deriving from the same cell, cell bead, or cell nucleus may be processed. As shown in panel 1950, RNA molecule 1958 comprising RNA sequence 1960 and polyA sequence 1962 and bead (*e.g.*, gel bead) 1916 is provided within a partition. In some embodiments, bead (*e.g.*, gel bead 1916, such as the same bead described in panel 1900) is included within the partition and is coupled to nucleic acid barcode molecule 1918b. Nucleic acid barcode molecule 1918b may comprise a flow cell adapter sequence 1968 (*e.g.*, a P5 sequence), a barcode sequence 1922b (*e.g.*, the same barcode sequence as barcode sequence 1922a), a UMI sequence 1966, and a polyT sequence 1964 complementary to polyA sequence 1962. In some instances, nucleic acid barcode molecule 1918b may comprise a sequencing primer sequence 1968 (*e.g.*, an R1 sequence or partial R1 sequence), a barcode sequence 1922b (*e.g.*, the same barcode sequence as barcode sequence 1922a), UMI sequence 1966, and a polyT sequence 1964 complementary to polyA sequence 1962. PolyT sequence 1964 may hybridize to polyA sequence 1962 of RNA molecule 1958. RNA molecule 1958 may be reverse transcribed 1970 off of the polyT sequence 1964 to provide a cDNA molecule comprising cDNA sequence 1972. The reverse transcription process may use a reverse transcriptase with terminal transferase activity that appends sequence 1974 to the resultant cDNA molecule comprising cDNA sequence 1972. In some embodiments sequence 1974 is a polyC sequence. A template switch oligonucleotide 1978 comprising a primer sequence 1980 and a sequence complementary to sequence 1974 (*e.g.*, a polyG sequence) may hybridize to the cDNA molecule. The contents of the partition are recovered in bulk solution (*e.g.*, a droplet may be broken) to provide the cDNA molecule in bulk solution. The cDNA molecule undergoes amplification (*e.g.*, PCR) 1984. Additional amplification (*e.g.*, PCR) 1986 may be performed to provide a double-stranded amplification product 1988 that includes sequences of the nucleic acid barcode molecule 1918b, the original RNA molecule or cDNA corresponding thereto, a flow cell adapter sequence 1998 (*e.g.*, a P7 sequence), and an additional sequence 1990 that may comprise a sequencing primer or portion thereof (*e.g.*, an R2 sequence) 1996, a sample index sequence 1994, and a flow cell adapter sequence (*e.g.*, a P5 sequence) 1992. The barcoded cDNA molecule may also or alternatively undergo fragmentation, end repair, dA tailing, ligation of one or more adapter sequences, and/or nucleic acid amplification. Additional schemes for collecting ATAC and GEX data for the discrete attribute value dataset 120 are described in United States Patent Application No. 16/789,287, entitled “Methods for

Processing Nucleic Acid Molecules,” filed February 12, 2020, which is hereby incorporated by reference.

[00134] Consistent with the workflow of Figure 19, in some embodiments, the discrete attribute value 124 of each gene 122 and the ATAC fragment count 125 of each region of open chromatin (ATAC peak 123) in a particular cell 126 in the plurality of cells is determined after the particular cell 126 has been separated from all the other cells in the plurality of cells into its own microfluidic partition. Such embodiments provide the ability to explore the heterogeneity between cells, which is one form of pattern analysis afforded by the systems and method of the present disclosure. In some such embodiments, because mRNA abundance and regions of open chromatin is being jointly measured, it is possible that the mRNA abundance and regions of open chromatin in each cell sample may vary vastly from cell to cell. As such, the disclosed systems and methods enable the profiling of which genes are being expressed and at what levels in each of the cells, as well as which regions of chromatin are open, and to use these gene profiles and open chromatin profiles (records of GEX discrete attribute values 124 and ATAC fragment counts 125), or principal components derived therefrom, to cluster cells and identify populations of related cells. As discussed above, the cells are clustered two different ways to form two sets of clusters, the GEX clusters 158 and the ATAC clusters 159. That is, the cells are clustered based on their GEX discrete attribute values 124 to form a first plurality of GEX clusters 158. Independent of this clustering, the cells are also clustered based on their ATAC fragment counts 125 to form a second plurality of ATAC clusters 159. This allows for the identification of cells with similar gene profiles and/or regions of open chromatin at different life cycle stages of the cell or different types of cells, different tissue, different organs, or other sources of cell heterogeneity. This further allows for the identification of any linkages between gene expression and open chromatin.

[00135] As such, in some embodiments, each cell 126 corresponds to a single cell, each gene 122 associated with a corresponding cell represents an mRNA (that maps to a gene that is in the genome of the single cell) and the discrete attribute value 124 is a number of copies of the mRNA that have been measured in the single cell. In some such embodiments, the discrete attribute value dataset 120 includes discrete attribute values 124 for 10 or more, 50 or more, 100 or more, 1000 or more, 3000 or more, 5000 or more, 10,000 or more, or 15,000 or more mRNAs in each cell represented by the dataset as well as ATAC fragment counts 125 for 10 or more, 50

or more, 100 or more, 1000 or more, 3000 or more, 5000 or more, 10,000 or more, or 15,000 or more ATAC peaks (open chromatin regions) in each cell represented by the dataset. In some such embodiments, the discrete attribute value dataset 120 includes discrete attribute values 124 for the mRNAs (genes 122) and ATAC fragment counts 125 for the ATAC peaks 123 of 500 or more cells, 5000 or more cells, 100,000 or more cells, 250,000 or more cells, 500,000 or more cells, 1,000,000 or more cells, 10 million or more cells or 50 million or more cells. In some embodiments, each single cell is a human cell. In some embodiments, each cell 126 represents a different human cell. In some embodiments, the discrete attribute value dataset 120 includes data for human cells of several different classes (*e.g.*, representing different deceased states and/or wild type states). In such embodiments, the discrete attribute value 124 for a respective mRNA of a gene 122 in a given cell 126 is the number of mRNAs for the respective gene that were measured in the given cell. This will either be zero or some positive integer. Moreover, the ATAC fragment count 125 for a respective ATAC peak 123 (open chromatin region) in a given cell 126 is the count of unique UMI for the respective ATAC peak 123 that were measured in the given cell. This will either be zero or some positive integer.

[00136] In some embodiments, the discrete attribute value 124 for a given gene 122 for a given cell 126 is a number in the set $\{0, 1, \dots, 100\}$. In some embodiments, the discrete attribute value 124 for a given gene 122 for a given cell 126 is a number in the set $\{0, 1, \dots, 50\}$. In some embodiments, the discrete attribute value 124 for a given gene 122 for a given cells 126 is a number in the set $\{0, 1, \dots, 30\}$. In some embodiments, the discrete attribute value 124 for a given gene 122 for a given cell 126 is a number in the set $\{0, 1, \dots, N\}$, where N is a positive integer, *e.g.*, a number in the range of 30 to 10,000, *etc.*

[00137] In some embodiments, the ATAC fragment count 125 for a given ATAC peak 123 for a given cell 126 is a number in the set $\{0, 1, \dots, 100\}$. In some embodiments, the ATAC fragment count 125 for a given ATAC peak 123 for a given cell 126 is a number in the set $\{0, 1, \dots, 50\}$. In some embodiments, the ATAC fragment count 125 for a given ATAC peak 123 for a given cell 126 is a number in the set $\{0, 1, \dots, 30\}$. In some embodiments, the ATAC fragment count 125 for a given ATAC peak 123 for a given cell 126 is a number in the set $\{0, 1, \dots, M\}$, where M is a positive integer, *e.g.*, a number in the range of 30 to 10,000, *etc.*

[00138] In some embodiments, the discrete attribute value dataset 120 includes discrete attribute values for 1000 or more, 3000 or more, 5000 or more, 10,000 or more, or 15,000 or more genes 122 in each cell 126 represented by the dataset. In some embodiments, the discrete attribute value dataset 120 includes ATAC fragment counts 125 for 1000 or more, 3000 or more, 5000 or more, 10,000 or more, or 15,000 or more ATAC peaks (*e.g.*, open chromatin regions) in each cell 126 represented by the dataset. In some such embodiments, the discrete attribute value dataset 120 includes respective discrete attribute values 124 for the genes and ATAC fragment counts 125 for ATAC peaks 123 of 500 or more cells, 5000 or more cells, 100,000 or more cells, 250,000 or more cells, 500,000 or more cells, 1,000,000 or more cells, 10 million or more cells, or 50 million or more cells.

[00139] As the above ranges indicate, the systems and methods of the present disclosure support very large discrete attribute value datasets 120 that may have difficulty being stored in the persistent memory 112 of conventional devices due to persistent memory 112 size limitations in conventional devices. Moreover, the systems and methods of the present disclosure are designed for data in which the sparsity is significantly more than twenty percent. The number of zero-valued elements divided by the total number of elements (*e.g.*, $m \times n$ for an $m \times n$ matrix) is called the sparsity of the matrix (which is equal to 1 minus the density of the matrix). While there are approximately twenty thousand genes in the human genome, most genes are not being expressed in a cell at any given time. Thus, it is expected that such data will have a sparsity approaching two percent in many instances. Thus, advantageously, to address the size constraints of the persistent memory (*e.g.*, magnetic drives or solid state drives) 112 limitations of conventional computers, in some embodiments, the discrete attribute value dataset 120 is represented in a compressed sparse matrix representation that may be searched both on a feature basis (gene 122 basis or ATAC peak 123) and a cell 126 basis. To accomplish this, the discrete attribute value dataset 120 redundantly represents the corresponding discrete attribute value 124 for each gene 122 in a plurality of genes as well as the corresponding ATAC fragment count 125 for each ATAC peak 123 for each respective cell 126 in a plurality of cells in both a compressed sparse row format and a compressed sparse column format in which genes or ATAC peaks for a respective cell that have a null discrete attribute data value are optionally discarded.

[00140] In some embodiments, the average density of the feature-barcode matrices 121 that are used in the systems and methods of the present disclosure are on the order of two percent.

Thus, if the features (genes and ATAC peaks) were viewed as a dense matrix, then only two percent of them would have data that is not zero. With a sparse matrix, all the zeroes are discarded. And so the sparse matrix allows for the dataset to fit in persistent memory 112. But with typical discrete attribute value datasets 120 of the present disclosure the memory footprint is still too high once the data for half a million cells 126 or more is used. For this reason, both the row-oriented and column-oriented sparse-matrix representations of the data are stored in persistent memory 112 in some embodiments in compressed blocks (*e.g.*, bgzf blocks) to support quick differential-expression analysis, which requires examination of the data (*e.g.* the discrete attribute values of genes) for individual cells. In the case of the gene “gene 3,” access to the discrete attribute data for gene 3 works by looking at the address in the dataset for gene 3, which thereby identifies the block in which the data for gene 3 resides. As such, when doing differential expression for a subset of the cells in the discrete attribute value dataset 120, the address of the individual cell is first needed.

[00141] Accordingly, in some embodiments, the discrete attribute value dataset 120 is stored in compressed sparse row (CSR) format. Here the term “compressed sparse row” is used interchangeably with the term “compressed sparse column” (CSC) format. The CSR format stores a sparse $m \times n$ matrix M in row form using three (one-dimensional) arrays (A , IA , JA). Here, NNZ denotes the number of nonzero entries in M (note that zero-based indices shall be used here) and the array A is of length NNZ and holds all the nonzero entries of M in left-to-right top-to-bottom (“row-major”) order. The array IA is of length $m + 1$. It is defined by this recursive definition:

[00142] $IA[0] = 0$;

[00143] $IA[i] = IA[i - 1] + (\text{number of nonzero elements on the } (i - 1)^{\text{th}} \text{ row in the original matrix})$.

[00144] Thus, the first m elements of IA store the index into A of the first nonzero element in each row of M , and the last element $IA[m]$ stores NNZ , the number of elements in A , which can be also thought of as the index in A of first element of a phantom row just beyond the end of the matrix M . The values of the i^{th} row of the original matrix is read from the elements $A[IA[i]]$ to $A[IA[i + 1] - 1]$ (inclusive on both ends), *e.g.* from the start of one row to the last index just before the start of the next.

[00145] The third array, JA, contains the column index in M of each element of A and hence is of length NNZ as well.

[00146] For example, the matrix M

$$[00147] \begin{pmatrix} 0 & 0 & 0 & 0 \\ 5 & 8 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 6 & 0 & 0 \end{pmatrix}$$

[00148] is a 4×4 matrix with 4 nonzero elements, hence

$$[00149] A = [5 \ 8 \ 3 \ 6]$$

$$[00150] IA = [0 \ 0 \ 2 \ 3 \ 4]$$

$$[00151] JA = [0 \ 1 \ 2 \ 1]$$

[00152] In one implementation of the matrix M above, each row represents a different cell 126 and each element of a given row represents a different feature (gene 122 or ATAC peak 123) associated with the different cell. Further, the value at a given matrix element represents the discrete attribute value 124 if the feature is a gene 122 or the ATAC fragment count 125 if the feature is an ATAC peak 123. In some embodiments, the ATAC data is stored in CSR format in a first matrix and the GEX data is stored in CSR format in a second matrix. In some embodiments, the ATAC data and the GEX data are stored in CSR format in the same matrix.

[00153] In some embodiments, the discrete attribute value dataset 120 is also stored in compressed sparse column (CSC or CCS) format. A CSC is similar to CSR except that values are read first by column, a row index is stored for each value, and column pointers are stored. For instance, CSC is (val, row_ind, col_ptr), where val is an array of the (top-to-bottom, then left-to-right) non-zero values of the matrix; row_ind is the row indices corresponding to the values; and, col_ptr is the list of val indexes where each column starts. In some embodiments, the ATAC data is stored in compressed sparse column format in a first matrix and the GEX data is stored in compressed sparse column format in a second matrix. In some embodiments, the ATAC data and the GEX data are stored in compressed sparse column format in the same matrix.

[00154] In addition to redundantly representing the corresponding discrete attribute value 124 for each gene 122 in a plurality of genes and the ATAC fragment count 125 for each ATAC peak

123 in a plurality of ATAC peaks for each respective cell 126 in a plurality of cells in both a compressed sparse row format and a compressed sparse column format, in some embodiments, the discrete attribute value dataset 120 is compressed in accordance with a blocked compression algorithm. In some such embodiments, this involves compressing the A and JA data structures but not the IA data structures using a block compression algorithm such as bgzf and storing this in persistent memory 112. Moreover, an index for compressed A and an index for compressed JA enable random seeks of the compressed data. In this way, although the discrete attribute value dataset 120 is compressed, it can be efficiently obtained and restored. All that needs to be done to obtain specific discrete attribute values 124 is seek to the correct block in persistent memory 112 and un-compress the block that contains the values and read them from within that block. Thus, certain operations, for example, like computing a differential heat map described below with reference to Figure 4A, is advantageously fast with the systems and method of present disclosure because it is known ahead of time which block of compressed data the desired attribute values 124 are in. That is, the systems and methods of the present disclosure know which row that a particular sought after cell is from looking at the row address value of the sparse matrix, which is stored outside of the compressed values. So, all that is needed is to figure out which block has the sought after gene data and what their discrete attribute values are, the algorithm jumps to the spot in the correct block (*e.g.*, bgzf block) that contains the data.

[00155] In some embodiments, in addition to the ATAC data, the discrete attribute value dataset 120 represents a whole transcriptome shotgun sequencing (RNA-seq) experiment that quantifies gene expression from a single cell in counts of transcript reads mapped to the genes.

[00156] *Block 206 – clustering the dataset.* In some embodiments, once a discrete attribute value dataset 120 is selected, *e.g.*, using the interface illustrated in Figure 3, the discrete attribute values 124 and/or the ATAC fragments counts 125 in the discrete attribute value dataset 120 are used by the browser module 119 to perform cluster visualization, as illustrated in Figure 4A, Figure 24 (*e.g.*, discrete attribute values using gene expression), Figure 25 (*e.g.*, ATAC), and Figures 30A and 30B (*e.g.*, gene expression and/or ATAC). In typical embodiments, either GEX principal component values 164 or ATAC principal component values 165 stored in the discrete attribute value dataset 120 that have been computed by the method of principal component analysis using the discrete attribute values 124 of the genes 122 or the ATAC fragment counts 125 of the ATAC peaks 123 across the plurality of cells 126 of the discrete attribute value

dataset 120 are used to perform cluster visualization, as illustrated in Figure 4A for the case of GEX principal component values 164.

[00157] Principal component analysis (PCA) is a mathematical procedure that reduces a number of correlated variables into a fewer uncorrelated variables called “principal components.” The first principal component is selected such that it accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The purpose of PCA is to discover or to reduce the dimensionality of the dataset, and to identify new meaningful underlying variables. PCA is accomplished by establishing actual data in a covariance matrix or a correlation matrix. The mathematical technique used in PCA is called Eigen analysis: one solves for the eigenvalues and eigenvectors of a square symmetric matrix with sums of squares and cross products. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component. The eigenvector associated with the second largest eigenvalue determines the direction of the second principal component. The sum of the eigenvalues equals the trace of the square matrix and the maximum number of eigenvectors equals the number of rows (or columns) of this matrix. *See, for example, Duda, Hart, and Stork, Pattern Classification, Second Edition, John Wiley & Sons, Inc., NY, 2000, pp. 115-116, which is hereby incorporated by reference.*

[00158] In some embodiments, principal component analysis, or other forms of data reduction, such as subset selection (*e.g.*, as disclosed in Hastie, 2001, *The Elements of Statistical Learning*, Springer, New York, pp. 55-57), discrete methods (*e.g.*, as disclosed in Furnival & Wilson, 1974, “Regression by Leaps and Bounds,” *Technometrics* 16(4), 499-511), forward/backward stepwise selection (*e.g.*, as disclosed in Berk, 1978, “Comparing Subset Regression Procedures,” *Technometrics* 20:1, 1-6), shrinkage methods (*e.g.*, as disclosed in Hastie, 2001, *The Elements of Statistical Learning*, Springer, New York, pp. 59-66), ridge regression (*e.g.*, as disclosed in Hastie, 2001, *The Elements of Statistical Learning*, Springer, New York, pp. 59-64), lasso techniques (*e.g.*, as disclosed in Hastie, 2001, *The Elements of Statistical Learning*, Springer, New York, pp. 64-65, 69-72, 330-331), derived input direction methods (*e.g.*, principal component regression (PCR), partial least squares (PLS), *etc.* as disclosed, for example, in Viyayakurma and Schaal, 2000, “Locally Weighted Projection Regression: An O(n) Algorithm for Incremental Real Time Learning in High Dimensional Space, Proc. of Seventeenth International Conference on Machine Learning (ICML2000), pp.

1079-1086), or combinations thereof, are used to reduce the dimensionality of the GEX or ATAC data down to a certain number of dimensions (*e.g.*, ten dimensions, or another number of dimensions) termed principal components or features (*e.g.*, principal components 164 for GEX and principal component 165 for ATAC).

[00159] Referring to block 208, in some embodiments, such clustering is performed at a prior time on a remote computer system. That is, in some embodiments, the cluster assignment of each cell 126 was already performed prior to accessing the discrete attribute value dataset 120. In such embodiments, the discrete attribute value dataset 120 includes the GEX cluster assignment 158 and of each cell, as illustrated in Figure 1B, as well as the ATAC cluster assignment 159 of each cell.

[00160] In some embodiments, the cluster assignment of each cell 126 is not performed prior to accessing the discrete attribute value dataset 120 but rather all the principal component analysis computation of the GEX principal component values 164 and the ATAC principal component values 165 is performed after accessing the discrete attribute value dataset 120.

[00161] For clustering in accordance with one embodiment of the systems and methods of the present disclosure, regardless at what stage it is performed, consider the case in which each cell 126 is associated with ten genes 122. In such instances, each cell 126 can be expressed as a vector:

$$\vec{X}_{10} = \{ x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10} \}$$

[00162] where X_i is the discrete attribute value 124 for the gene i 124 associated with the cell 126. Thus, if there are one thousand cells 126, 1000 GEX vectors are defined. Those cells 126 that exhibit similar discrete attribute values across the set of genes 122 of the dataset 120 will tend to cluster together. For instance, in the case where each cell 126 is an individual cell, the genes 122 correspond to mRNA mapped to individual genes within such individual cells, and the discrete attribute values 124 are mRNA counts for such mRNA, it is the case in some embodiments that the discrete attribute value dataset 120 includes mRNA data from one or more cell types (*e.g.*, diseased state and non-diseased state), two or more cell types, three or more cell types. In such instances, it is expected that cells of like type will tend to have like values for mRNA across the set of genes (mRNA) and therefor cluster together. For instance, if the discrete attribute value dataset 120 includes class a: cells from subjects that have a disease, and

class b: cells from subjects that do not have a disease, an ideal clustering classifier will cluster the discrete attribute value dataset 120 into two groups, with one cluster group uniquely representing class a and the other cluster group uniquely representing class b. The same is true for ATAC data. Consider the case in which each cell 126 is associated with ten ATAC peaks 123. In such instances, each cell 126 can be expressed as a vector:

$$\vec{Y}_{10} = \{ y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10} \}$$

where Y_i is the ATAC fragment count 125 for the ATAC peak i 123 associated with the cell 126. Thus, if there are one thousand cells 126, 1000 ATAC vectors are defined, one for each cell. Those cells 126 that exhibit similar ATAC fragment counts across the set of ATAC peaks (open chromatin regions) of the dataset 120 will tend to cluster together.

[00163] For GEX clustering in accordance with another embodiment of the systems and methods of the present disclosure, regardless at what stage it is performed, consider the case in which each cell 126 is associated with ten GEX principal component values 164 that collectively represent the variation in the discrete attribute values of a large number of genes 122 across the cells in the dataset. In such instances, each cell 126 can be expressed as a vector:

$$\vec{W}_{10} = \{ w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9, w_{10} \}$$

where W_i is the GEX principal component value 164 i associated with the cell 126. Thus, if there are one thousand cells 126, one those vectors are defined. Those cells 126 that exhibit similar GEX principal component values 164 across the set of principal component values 164 will tend to cluster together. For instance, in the case where each cell 126 is an individual cell, the genes 122 correspond to mRNA mapped to individual genes within such individual cells, and the discrete attribute values 124 are mRNA counts for such mRNA, it is the case in some embodiments that the discrete attribute value dataset 120 includes mRNA data from one or more cell types (*e.g.*, diseased state and non-diseased state), two or more cell types, three or more cell types, or more than four cell types. In such instances, it is expected that cells of like type will tend to have like principal component values 164 across the set of principal components and therefor cluster together. For instance, if the discrete attribute value dataset 120 includes class a: cells from subjects that have a disease, and class b: cells from subjects that have a disease, an ideal clustering classifier will cluster the discrete attribute value dataset 120 into two groups,

with one cluster group uniquely representing class a and the other cluster group uniquely representing class b. In some embodiments, the ATAC principal component values 165 are likewise used to cluster the cells in the dataset into a plurality of ATAC clusters.

[00164] Clustering is described on pages 211-256 of Duda and Hart, *Pattern Classification and Scene Analysis*, 1973, John Wiley & Sons, Inc., New York, (hereinafter “Duda 1973”) which is hereby incorporated by reference in its entirety. As described in Section 6.7 of Duda 1973, the clustering problem is described as one of finding natural groupings in a dataset. To identify natural groupings, two issues are addressed. First, a way to measure similarity (or dissimilarity) between two samples is determined. This metric (similarity measure) is used to ensure that the samples in one cluster are more like one another than they are to samples in other clusters. Second, a mechanism for partitioning the data into clusters using the similarity measure is determined.

[00165] Similarity measures are discussed in Section 6.7 of Duda 1973, where it is stated that one way to begin a clustering investigation is to define a distance function and to compute the matrix of distances between all pairs of samples in a dataset. If distance is a good measure of similarity, then the distance between samples in the same cluster will be significantly less than the distance between samples in different clusters. However, as stated on page 215 of Duda 1973, clustering does not require the use of a distance metric. For example, a nonmetric similarity function $s(x, x')$ can be used to compare two vectors x and x' . Conventionally, $s(x, x')$ is a symmetric function whose value is large when x and x' are somehow “similar.” An example of a nonmetric similarity function $s(x, x')$ is provided on page 216 of Duda 1973.

[00166] Once a method for measuring “similarity” or “dissimilarity” between points in a dataset has been selected, clustering requires a criterion function that measures the clustering quality of any partition of the data. Partitions of the dataset that extremize the criterion function are used to cluster the data. See page 217 of Duda 1973. Criterion functions are discussed in Section 6.8 of Duda 1973.

[00167] More recently, Duda *et al.*, *Pattern Classification*, Second edition, John Wiley & Sons, Inc. New York, which is hereby incorporated by reference, has been published. Pages 537-563 describe clustering in detail. More information on clustering techniques can be found in Kaufman and Rousseeuw, 1990, *Finding Groups in Data: An Introduction to Cluster Analysis*,

Wiley, New York, N.Y.; Everitt, 1993, Cluster analysis (Third Edition), Wiley, New York, N.Y.; and Backer, 1995, *Computer-Assisted Reasoning in Cluster Analysis*, Prentice Hall, Upper Saddle River, N.J. Referring to blocks 210-212, particular exemplary clustering techniques that can be used in the systems and methods of the present disclosure to cluster a plurality of vectors, (where each respective vector in the plurality of vectors comprises: (i) the discrete attribute values 124 of the genes 122 of a corresponding cell 126, (ii) the GEX principal components 164 of a corresponding cell 126, (iii) the ATAC fragment counts 125 of the ATAC peaks 123 of a corresponding cell 126, or (iv) the ATAC principal components 165 of a corresponding cell 126) includes, but is not limited to, hierarchical clustering (agglomerative clustering using nearest-neighbor algorithm, farthest-neighbor algorithm, the average linkage algorithm, the centroid algorithm, or the sum-of-squares algorithm), k-means clustering, fuzzy k-means clustering algorithm, and Jarvis-Patrick clustering.

[00168] Thus, in some embodiments, the discrete attribute values 124 for each gene 122 in the plurality of genes for each respective cell 126 in the plurality of cells, or principal component values 164 derived from the discrete attribute values 124, is clustered thereby assigning each respective cell 126 in the plurality of cell to a corresponding GEX cluster 158 in a first plurality of clusters and thereby assigning a cluster attribute value to each respective cell in the plurality of cells. Further, the ATAC fragment counts 125 for each ATAC peak 123 in the plurality of ATAC peaks for each respective cell 126 in the plurality of cells, or principal component values 165 derived from the ATAC fragment counts 125, is clustered thereby assigning each respective cell 126 in the plurality of cell to a ATAC corresponding cluster 159 in a second plurality of clusters and thereby assigning a cluster attribute value to each respective cell in the plurality of cells.

[00169] Referring to block 214, in one embodiment of the present disclosure k-means clustering is used for both the ATAC and GEX clustering. The goal of k-means clustering is to cluster the discrete attribute value dataset 120 based upon the principal components (or the discrete attribute values or ATAC fragment counts) of individual cells into K partitions. Referring to block 214, in some embodiments, K is a number between 2 and 50 inclusive for both the ATAC and the GEX clustering. In some embodiments, the number K is set to a predetermined number such as 10. In some embodiments, the number K is optimized for a

particular discrete attribute value dataset 120. Referring to block 216, in some embodiments, a user sets the number K using browser module 119.

[00170] Figure 4A illustrates an instance in which the atac_gex_intron dataset 120, constituting mRNA GEX data and ATAC data from 9,461 different cells, has been clustered, based on GEX principal components 164 into thirteen clusters 158. In some embodiments, for k-means clustering, the user selects in advance how many clusters the clustering algorithm will compute prior to clustering. In some embodiments, no predetermined number of clusters is selected. Instead, clustering is performed until predetermined convergence criteria are achieved. In embodiments where a predetermined number of clusters is determined, k-means clustering of the present disclosure is then initialized with K cluster centers $\mu_1, \dots, \mu_K$ randomly initialized in two dimensional space. As discussed above, for each respective cell 126 i in the dataset, a vector X_i is constructed of each principal component value 164 associated with the respective cell 126. In the case where K is equal to 10, ten such vectors $\vec{X}$ are selected to be the centers of the ten clusters. Then, each remaining vector $\vec{X}_i$, corresponding to the cells 126 which were not selected to be cluster centers, is assigned to its closest cluster center:

$$\mathbf{[00171]} \quad \mathcal{C}_k = \{n: k = \arg \min_k \|\vec{X}_i - \mu_k\|^2\}$$

[00172] where $\mathcal{C}_k$ is the set of examples closest to μ_k using the objective function:

$$\mathbf{[00173]} \quad J(\mu, r) = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|\vec{X}_i - \mu_k\|^2$$

[00174] where $\mu_1, \dots, \mu_K$ are the K cluster centers and $r_{nk} \in \{0, 1\}$ is an indicator denoting whether a cell 126 $\vec{X}_i$ belongs to a cluster k . Then, new cluster centers μ_k are recomputed (mean / centroid of the set $\mathcal{C}_k$):

$$\mathbf{[00175]} \quad \mu_k = \frac{1}{|\mathcal{C}_k|} \sum_{n \in \mathcal{C}_k} \vec{X}_i$$

[00176] Then, all vectors $\vec{X}_i$, corresponding to the cells 126 are assigned to the closest updated cluster centers as before. This is repeated while not converged. Any one of a number of convergence criteria can be used. One possible convergence criteria is that the cluster centers do not change when recomputed. The k-means clustering computes a score for each respective cell 126 that takes into account the distance between the respective cell and the centroid of the cluster

158 that the respective cell has been assigned. In some embodiments this score is stored as a cluster attribute value for the cell 126.

[00177] Once the clusters are identified, as illustrated in Figure 4A, individual clusters can be selected to display. For instance, referring to Figure 4A, clusters 158 can be individually selected or deselected to display or remove them from panel 420. In Figure 4A, each cluster 158 is assigned a different color. Each point in panel 420 represents one of the 9,461 cells in the `atac_gex_intron` dataset 120. Moreover, each cell in panel 420 is colored by its GEX cluster assignment 158.

[00178] As discussed above, the cells of the discrete attribute value dataset are clustered in two different ways, GEX and ATAC. Affordance 450 can be used to recolor the cells in panel 420 in accordance with the ATAC cluster assignments 159. Figure 4B illustrates the same view as Figure 4A, only now the cells are colored using the 11 ATAC cluster assignment 159 instead of the 13 GEX cluster assignments. Moreover, in Figure 4B, each cell in panel 420 is colored by its ATAC cluster assignment 159. As such, Figure 4B illustrates displaying, in a first panel 420, a two-dimensional projection (GEX t-SNE) of the plurality of cells based on assignment of the plurality of cells to one of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups (in the case of Figure 4B, the GEX cluster groups 158). Figure 4B further illustrates indicating within the two-dimensional projection, for each respective cell in the plurality of cells, membership in the other of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups (in the case of Figure 4B, the ATAC cluster groups 159) thereby visualizing the pattern in the discrete attribute value dataset. Figure 4B further illustrates membership of each respective cell in the plurality of cells in the other of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups (in the case of Figure 4B, the ATAC cluster groups 159) by coloring the respective cell a color that is uniquely associated with a cluster group to which the respective cell has been assigned in the other of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups (in the case of Figure 4B, the ATAC cluster groups 159). For example, membership of cells in another plurality of cluster groups is illustrated in Figures 27, 28A-C, 29A-D, 32A-B, and 33A-D. Furthermore, in some embodiments, concordance between the first plurality of cluster groups and the second plurality of cluster groups is determined, as illustrated in Figures 26A-B.

[00179] Returning to Figure 4A, in accordance with the systems and methods of the present disclosure, in typical embodiments each respective cluster 158 in the plurality of clusters consists of a unique different subset of the plurality of cells 126. Moreover, because in typical embodiments the discrete attribute value dataset 120 is too large to load into the non-persistent memory 111, in typical embodiments this clustering loads less than the entirety of the discrete attribute value dataset 120 into the non-persistent memory 111 at any given time during the clustering. For instance, in embodiments where the discrete attribute value dataset 120 has been compressed using bgzf, only a subset of the blocks of the discrete attribute value dataset 120 are loaded into non-persistent memory during the clustering of the discrete attribute value dataset 120. Once one subset of the blocks of the discrete attribute value dataset 120 have been loaded from persistent memory 112 into non-persistent memory 111 and processed in accordance with the clustering algorithm (e.g., k-means clustering), the subset of blocks of data is discarded from non-persistent memory 111 and a different subset of blocks of the discrete attribute value dataset 120 are loaded from persistent memory 112 into non-persistent memory 111 and processed in accordance with the clustering algorithm of the clustering module 152.

[00180] In some embodiments k-means clustering is used to assign cells 126 to clusters 158. In some such embodiments the k-means clustering uses as input the GEX principal component values 164 for each cell 126 as the basis for clustering the cells into cluster. Thus, the k-means algorithm computes like clusters of cells from the higher dimensional data (the set of GEX principal component values 164) and then after some resolution, the k-means clustering tries to minimize error. In this way, the k-means clustering provides cluster assignments 158, which are recorded in the discrete attribute value dataset 120. In some embodiments, with k-means clustering, the user decides in advance how many clusters 158 there will be. In some embodiments, the feature of k-means cluster is exploited by running a series of k-means clustering runs, with each different run having a different number of clusters (a different value for K). Thus, in some embodiments, a separate k-means clustering is performed on the GEX principal component data values 164 of each cell 122, ranging from two clusters to ten clusters, with each k-means clustering identifying a separability score (quality score) and all the results of each clustering embedded in the discrete attribute value dataset 120 from K=2 through K=10. In some such embodiments, such clustering is performed for K=2 through K=25. In some such embodiments, such clustering is performed for K=2 through K=100. The clustering that is

displayed by default in such embodiments is the k-means clustering (1, ... N) that has the highest separability score. In Figure 4A, the cells of each cluster assignment 158 are displayed in panel 420 in a different color (*e.g.*, the cells assigned to cluster 1 are displayed in red, the cells assigned to cluster 6 are displayed in green, etc.). In other embodiments, the cells of each cluster assignment 158 is displayed in panel 420 with a different dot pattern or hash pattern.

[00181] In some embodiments k-means clustering is also used to assign cells 126 to clusters 159. In some such embodiments the k-means clustering uses as input the ATAC principal component values 165 for each cell 126 as the basis for clustering the cells into cluster. Thus, the k-means algorithm computes like clusters of cells from the higher dimensional data (the set of ATAC principal component values 165) and then after some resolution, the k-means clustering tries to minimize error. In this way, the k-means clustering provides cluster assignments 159, which are recorded in the discrete attribute value dataset 120. In some embodiments, with k-means clustering, the user decides in advance how many clusters 159 there will be. In some embodiments, the feature of k-means cluster is exploited by running a series of k-means clustering runs, with each different run having a different number of clusters (a different value for K). Thus, in some embodiments, a separate k-means clustering is performed on the ATAC principal component data values 165 of each cell 122, ranging from two clusters to ten clusters, with each k-means clustering identifying a separability score (quality score) and all the results of each clustering embedded in the discrete attribute value dataset 120 from K=2 through K=10. In some such embodiments, such clustering is performed for K=2 through K=25. In some such embodiments, such clustering is performed for K=2 through K=100. The clustering that is displayed by default in such embodiments is the k-means clustering (1, ... N) that has the highest separability score. In Figure 4B, the cells assigned to each cluster assignment 159 is displayed in a different color. In other embodiments, the cells of each cluster assignment 159 is displayed in panel 420 with a different dot pattern or hash pattern.

[00182] The k-means clustering algorithm is an attempt to elucidate like clusters 158 (or clusters 159) within the data. There is no guarantee that the clusters 158 or clusters 159 represent physiologically significant events. In other words, *a priori*, it is not known what the clusters 158 or what cluster 159 mean. What is known is that the algorithm has determined that there are differences between the cells 126 that are being represented by different colors or different hash patterns or symbols. The systems and methods of the present disclosure provide

tools for determining whether there is any meaning behind the differences between the clusters such as the heat map of panel 404.

[00183] Referring to block 214, in some embodiments of the present disclosure, rather than using k-means clustering for clustering ATAC or GEX data, a Louvain modularity algorithm is used. *See*, Blondel *et al.*, July 25, 2008, “Fast unfolding of communities in large networks,” arXiv:0803.0476v2 [physical.coc-ph], which is hereby incorporated by reference. In some embodiments, the user can choose a clustering algorithm. In some embodiments, the user can choose between at least K-means clustering and a Louvain modularity algorithm. In some embodiments, the clustering of the dataset comprises application of a Louvain modularity algorithm to a map, the map comprising a plurality of nodes and a plurality of edges. Each node in the plurality of nodes represents a cell in the plurality of cells. The coordinates in N-dimensional space of a respective node in the plurality of nodes are a set of principal components of the corresponding cell in the plurality of cells. The set of principal components is either derived from the corresponding discrete attribute values 124 of the plurality of genes for the corresponding cell or from the corresponding ATAC fragment counts 125 of the plurality of ATAC peaks 123 for the corresponding cell, where N is the number of principal components in each set of principal components. An edge exists in the plurality of edges between a first node and a second node in the plurality of nodes when the first node is among the k nearest neighboring nodes of the second node in the first plurality of node, where the k nearest neighboring nodes to the second node is determined by computing a distance in the N-dimensional space between each node in the plurality of nodes, other than the second node, and the second node. In some embodiments, the distance is a Euclidean distance. In other embodiments, other distance metrics are used (*e.g.*, Chebyshev distance, Mahalanobis distance, Manhattan distance, *etc.*). In typical embodiments, the nodes and the edges are not weighted for the Louvain modularity algorithm. In other words, each node and each edge receives the same weight in such embodiments

[00184] *Block 218 - Computing differential attribute values for each cluster.* Once each cell 126 has been assigned to a respective cluster 158 based on GEX data, the systems and methods of the present disclosure are able to compute, for each respective gene 122 in the plurality of genes for each respective cluster 158 in a first plurality of clusters, a difference in the discrete attribute value 124 for the respective gene 122 across the respective subset of cells 126 in the

respective cluster 158 relative to the discrete attribute value 124 for the respective gene 122 across the first plurality of clusters 158 other than the respective cluster, thereby deriving a differential value for each respective gene 122 in the plurality of genes for each cluster 158 in the first plurality of clusters. For instance, in some such embodiments, a differential expression algorithm is invoked to find the top expressing genes that are different between cell classes or other forms of cell labels. This is a form of the general differential expressional problem in which there is one set of expression data and another set of expression data and the question to be addressed is determining which genes are differentially expressed between the datasets.

[00185] In some embodiments differential expression is computed as the $\log_2$ fold change in (i) the average number of transcripts (discrete attribute value 124 for gene 122) measured in each of the cells of the subject cluster 158 that map to a particular gene and (ii) the average number of transcripts measured in each of the cells of all clusters other than the subject cluster that map to the particular gene. Thus, consider the case in which the subject cluster contains 50 cells and on average each of the 50 cells contain 100 transcripts for gene A. The remaining clusters in the first plurality of clusters (the GEX clusters) collectively contain 250 cells and on average each of the 250 cells contain 50 transcripts for gene A. Here, the fold change in expression for gene A is $100/50$ and the $\log_2$ fold change is $\log_2(100/50) = 1$. In Figure 4A, lower panel, the $\log_2$ fold change is computed in this manner for each gene in represented in the discrete attribute set 120 and such values are presented in a color coded heat map.

[00186] Referring to block 220 of Figure 2B, in some embodiments, the differential value for each respective gene 122 in the plurality of genes for each respective cluster 158 in the first plurality of clusters (the GEX clusters) is a fold change in (i) a first measure of central tendency of the discrete attribute value 124 for the gene measured in each of the cells 126 in the plurality of cells in the respective cluster 158 and (ii) a second measure of central tendency of the discrete attribute value 124 for the respective gene 122 measured in each of the cells 126 of all clusters 158 in the first plurality of clusters other than the respective cluster. In some embodiments, the first measure of central tendency is an arithmetic mean, weighted mean, midrange, midhinge, trimean, Winsorized mean, median, or mode of all the discrete attribute value 124 for the gene measured in each of the cells 126 in the plurality of cells in the respective cluster 158. In some embodiments, the second measure of central tendency is an arithmetic mean, weighted mean, midrange, midhinge, trimean, Winsorized mean, median, or mode of all the discrete attribute

values 124 for the gene 122 measured in each of the cells 126 in the plurality of cells 126 in all clusters of the first plurality of clusters other than the respective cluster. Referring to block 222, in some embodiments the fold change is a $\log_2$ fold change. Referring to block 224, in some embodiments the fold change is a $\log_{10}$ fold change.

[00187] Moreover, once each cell 126 has also been assigned to a respective cluster 159 based on ATAC data, the systems and methods of the present disclosure are able to compute, for each respective ATAC peak 123 in the plurality of ATAC peaks for each respective cluster 159 in a second plurality of clusters (the ATAC clusters), a difference in the ATAC fragment count 125 for the respective ATAC peak 123 across the respective subset of cells 126 in the respective cluster 159 relative to the ATAC fragment count 125 for the ATAC peak 123 across the second plurality of clusters 159 other than the respective cluster, thereby deriving a differential value for each respective ATAC peak 123 in the plurality of ATAC peaks for each cluster 159 in the second plurality of clusters (the ATAC clusters). For instance, in some such embodiments, a differential abundance algorithm is invoked to find the top ATAC peaks that are different between cell classes or other forms of cell labels. This is a form of the general differential abundance problem in which there is one set of abundance data and another set of abundance data and the question to be addressed is determining which peaks are differentially abundant between the datasets.

[00188] In some embodiments differential abundance is computed as the $\log_2$ fold change in (i) the average fragment count (ATAC fragment count for peak 123) measured in each of the cells of the subject cluster 159 that map to a particular ATAC peak 123 and (ii) the average number of fragments measured in each of the cells of all clusters other than the subject cluster that map to the particular ATAC peak 123. In Figure 4B, lower panel, the $\log_2$ fold change is computed in this manner for each ATAC peak (123-1 through 123-L) represented in the discrete attribute set 120 and such values are presented in a color coded heat map.

[00189] In some embodiments, the differential value for each respective ATAC peak 123 in the plurality of ATAC peaks 123 for each respective cluster 159 in the second plurality of clusters is a fold change in (i) a first measure of central tendency of the ATAC fragment count 125 for the ATAC peak 123 measured in each of the cells 126 in the plurality of cells in the respective cluster 159 and (ii) a second measure of central tendency of the ATAC fragment count

125 for the respective ATAC peak 123 measured in each of the cells 126 of all clusters 159 in the second plurality of clusters (the ATAC clusters) other than the respective cluster. In some embodiments, the first measure of central tendency is an arithmetic mean, weighted mean, midrange, midhinge, trimean, Winsorized mean, median, or mode of all the ATAC fragment counts 125 for the ATAC peak 123 measured in each of the cells 126 in the plurality of cells in the respective cluster 159. In some embodiments, the second measure of central tendency is an arithmetic mean, weighted mean, midrange, midhinge, trimean, Winsorized mean, median, or mode of all the ATAC fragment count for the ATAC peak 123 measured in each of the cells 126 in the plurality of cells 126 in all clusters other than the respective cluster.

[00190] Given that measurement of discrete attribute values 124 for genes 122 (*e.g.*, count of mRNA that maps to a given gene in a given cell) is typically noisy, the variance of the discrete attribute values 124 for genes 122 in each cell 126 (*e.g.*, count of mRNA that maps to given gene in a given cell) in a given cluster 158 of such cells 126 is taken into account in some embodiments. This is analogous to the t-test which is a statistical way to measure the difference between two samples. Here, in some embodiments, statistical methods that take into account that a discrete number of genes 122 are being measured (as the discrete attribute values 124 for a given gene 122) for each cell 126 and that model the variance that is inherent in the system from which the measurements are made are implemented.

[00191] Thus, referring to block 226 of Figure 2B, in some embodiments, each discrete attribute value 124 is normalized prior to computing the differential value 162 for each respective gene 122 in the plurality of first genes for each respective cluster 158 in the plurality of clusters. Referring to block 228 of Figure 2B, in some embodiments, the normalizing comprises modeling the discrete attribute value 124 of each gene associated with each cell in the plurality of cells with a negative binomial distribution having a consensus estimate of dispersion without loading the entire dataset into non-persistent memory 111. Such embodiments are useful, for example, for RNA-seq experiments that produce discrete attribute values 124 for genes 122 (*e.g.*, digital counts of mRNA reads that are affected by both biological and technical variation). To distinguish the systematic changes in expression between conditions from noise, the counts are frequently modeled by the Negative Binomial distribution. *See* Yu, 2013, “Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size,” *Bioinformatics* 29, pp. 1275-1282, which is hereby incorporated by reference.

[00192] The negative binomial distribution for a discrete attribute value 124 for a given gene 122 includes a dispersion parameter for the discrete attribute value 124 which tracks the extent to which the variance in the discrete attribute value 124 exceeds an expected value. *See* Yu, 2013, “Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size,” *Bioinformatics* 29, pp. 1275-1282, and Cameron and Trivedi, 1998, “Regression Analysis of Count Data,” *Econometric Society Monograph* 30, Cambridge University Press, Cambridge, UK, each of which is hereby incorporated by reference. Rather than relying upon an independent dispersion parameter for the discrete attribute value 124 of each gene 122, some embodiments of the disclosed systems and methods advantageously use a consensus estimate across the discrete attribute values 124 of all the genes 122. This is termed herein the “consensus estimate of dispersion.” The consensus estimate of dispersion is advantageous for RNA-seq experiments in which whole transcriptome shotgun sequencing (RNA-seq) technology quantifies gene expression in biological samples in counts of transcript reads mapped to the genes, which is one form of experiment used to acquire the disclosed discrete attribute values 124 in some embodiments, thereby concurrently quantifying the expression of many genes. The genes share aspects of biological and technical variation, and therefore a combination of the gene-specific estimates and of consensus estimates can yield better estimates of variation. *See* Yu, 2013, “Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size,” *Bioinformatics* 29, pp. 1275-1282 and Anders and Huber, 2010, “Differential expression analysis for sequence count data,” *Genome Biol* 11, R106, each of which are hereby incorporated by reference. For instance, in some such embodiments, sSeq is applied to the discrete attribute value 124 of each gene 122. sSeq is disclosed in Yu, 2013, “Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size,” *Bioinformatics* 29, pp. 1275-1282, which is hereby incorporated by reference. sSeq scales very well with the number of genes that are being compared. In typical experiments in accordance with the present disclosure, each cluster 158 may include hundreds, thousands, tens of thousands, hundreds of thousands, or more cells 126, and each respective cell 126 may contain mRNA expression data for hundreds, or thousands of different genes. As such, sSeq is particularly advantageous when testing for differential expression in such large discrete attribute value datasets 120. Of all the RNA-seq methods, sSeq is advantageously faster. Other single-cell differential expression methods exist and can be used

in some embodiments, but they are designed for smaller-scale experiments. As such sSeq, and more generally techniques that normalize discrete attribute values by modeling the discrete attribute value 124 of each gene 122 associated with each cell 126 in the cells with a negative binomial distribution having a consensus estimate of dispersion without loading the entire discrete attribute value dataset 120 into non-persistent memory 111, are practiced in some embodiments of the present disclosure. In some embodiments, in the case where parameters for the sSeq calculations are calculated, the discrete attribute values for each of the genes is examined in order to get a dispersion value for all the genes. Here, although all the discrete attribute values for the genes are accessed to make the calculation, the discrete attribute values are not all read from persistent memory 112 at the same time. In some embodiments, discrete attribute values are obtained by traversing through blocks of compressed data, a few blocks at a time. That is, a set of blocks, consisting of the few compressed blocks, in the dataset are loaded into non-persistent memory from persistent memory and are analyzed to determine which genes the set of blocks represent. An array of discrete attribute values across the plurality of cells, for each of the genes encoded in the set of blocks, is determined and used calculate the variance, or other needed parameters, for these genes across the plurality of cells. This process is repeated in which new set of blocks is loaded into non-persistent memory from persistent memory, analyzed to determine which genes are encoded in the new set of blocks, and then used to compute the variance, or other needed parameters, for these genes across the plurality of cells for each of the genes encoded in the new set of blocks, before discarding the set of blocks from non-persistent memory. In this way, only a limited amount of the discrete attribute value dataset 120 is stored in non-persistent memory 111 at any given time (*e.g.*, the data for a particular block that contain the discrete attribute values for a particular gene). Further, the systems and methods of the present disclosure are able to compute variance in discrete attribute values for a given gene because it has got all the discrete attribute values for that particular gene across the entire discrete attribute value dataset 120 stored in a single bgzf block, in some embodiments. Once the variance, or other needed parameter is computed for the genes (or discrete attribute values of the genes), the accessed set of bgzf blocks (which is a subset of the total number of bgzf blocks in the dataset), which had been loaded into non-persistent memory 111 to perform the computation, is dropped from non-persistent memory and another set of bgzf blocks for which such computations is to be performed is loaded into the non-persistent memory 111 from the

persistent memory 112. In some embodiments, such processes run in parallel (*e.g.*, one process for each gene) when there are multiple processing cores 102. That is, each processing core concurrently analyzes a different respective set of blocks in the dataset and computes statistics for those genes represented in the respective set of blocks.

[00193] Following such normalization, in some embodiments, for each respective gene 122, an average (or some other measure of central tendency) discrete attribute value 124 (*e.g.*, count of the gene 122) for each gene 122 is calculated for each cluster 158 of cells 126. Thus, in the case where there is a first cluster 158-1 and second cluster 158-2 of cells 126, the average (or some other measure of central tendency) discrete attribute value 124 of the gene “A” across all the cells 126 of the first cluster 158-1, and the average (or some other measure of central tendency) discrete attribute value 124 of gene “A” across all the cells 126 of the second cluster 158-2 is calculated and, from this, the differential value for each gene with respect to the first cluster is calculated. This is repeated for each of the genes 122 in a given cluster. It is further repeated for each cluster 158 in the plurality of clusters. In some embodiments, there are other factors that are considered, like adjusting the initial estimate of the variance in the discrete attribute value 124 when the data proves to be noisy. In the case where there are more than two clusters, the average (or some other measure of central tendency) discrete attribute value 124 of gene A across all the cells 126 of the first cluster 158-1 and the average (or some other measure of central tendency) discrete attribute value 124 of gene A across all the cells 126 of the remaining cluster 158-2, is calculated and used to compute the differential value.

[00194] Moreover, in some embodiments, for each respective ATAC peak 123, an average (or some other measure of central tendency) ATAC fragment count 125 (*e.g.*, count of the ATAC peak 123) for each ATAC peak 123 is calculated for each cluster 159 of cells 126. Thus, in the case where there is a first cluster 159-1 and a second cluster 159-2 of cells 126, the average (or some other measure of central tendency) ATAC fragment count 125 of the ATAC peak “A” across all the cells 126 of the first cluster 159-1, and the average (or some other measure of central tendency) ATAC fragment count 125 of ATAC peak “A” across all the cells 126 of the second cluster 159-2 is calculated and, from this, the differential value for each ATAC peak with respect to the first cluster 159-1 is calculated. This is repeated for each of the ATAC peaks 123 in a given cluster. It is further repeated for each cluster 159 in the second plurality of clusters. In some embodiments, there are other factors that are considered, like adjusting the initial

estimate of the variance in the ATAC fragment count 125 when the data proves to be noisy. In the case where there are more than two clusters, the average (or some other measure of central tendency) ATAC fragment count 125 of ATAC peak A across all the cells 126 of the first cluster 159-1 and the average (or some other measure of central tendency) ATAC fragment count 125 of ATAC peak A across all the cells 126 of the remaining cluster 159-2, is calculated and used to compute the differential value.

[00195] *Block 230 - Display a heat map.* With reference to Figure 4A, once the differential value for each respective gene 122 in the plurality of genes for each respective cluster 158 in the first plurality of clusters has been computed or otherwise obtained, a heat map 402 of these differential values is displayed in a first panel 404 of an interface 400. In Figure 4A, the heat map 402 comprises a representation of the differential value for each respective gene 122 in the plurality of genes for each cluster 158 in the first plurality of clusters. As illustrated in Figure 4A, the differential value for each gene 122 in the plurality of genes (*e.g.*, genes 122-1 to 122-M) for each cluster 158 (*e.g.*, clusters 158-1, 158-4, 158-7, and 158-10) is illustrated in a color coded way to represent the $\log_2$ fold change in accordance with color key 408. In accordance with color key 408, those genes 122 that are upregulated in the cells 126 of a particular cluster 158 relative to all other clusters in the first plurality of clusters are assigned more positive values, whereas those genes 122 that are down-regulated in the cells 126 of a particular cluster 158 relative to all other clusters in the first plurality of clusters are assigned more negative values. In some embodiments, the heat map can be exported to persistent storage (*e.g.*, as a PNG graphic, JPG graphic, or other file formats). Heat maps obtained using differential values for genes are further illustrated in Figures 29E and 32C.

[00196] With reference to Figure 4B, once the differential value for each respective ATAC peak 123 in the plurality of ATAC peaks for each respective cluster 159 in the second plurality of clusters has been computed or otherwise obtained, a heat map 402 of these differential values is displayed in a first panel 404 of an interface 400. In Figure 4B, the heat map 402 comprises a representation of the differential value for each ATAC peak 123 in the plurality of ATAC peaks 123 for each cluster 159 in the second plurality of clusters. As illustrated in Figure 4B, the differential value for each ATAC peak 123 in the plurality of ATAC peaks (*e.g.*, ATAC peaks 123-1 to 122-L) for each cluster 159 (*e.g.*, clusters 158-1, 158-4, 158-7, and 158-9) is illustrated in a color coded way to represent the $\log_2$ fold change in accordance with color key 408. In

accordance with color key 408, those ATAC peaks 123 that are upregulated in the cells 126 of a particular cluster 159 relative to all other clusters in the second plurality of clusters are assigned more positive values, whereas those ATAC peaks 123 that are down-regulated in the cells 126 of a particular cluster 159 relative to all other clusters in the second plurality of clusters are assigned more negative values. In some embodiments, the heat map is exported to persistent storage (*e.g.*, as a PNG graphic, JPG graphic, or other file formats).

[00197] *Block – 232 plot a two dimensional plot of the cells in the dataset.* With reference to Figures 4A and 4B, in some embodiments, a two-dimensional visualization (*e.g.*, as illustrated in Figures 24, 25, and/or 30) of the discrete attribute value dataset 120 is also provided in a second panel 420. In some embodiments, the two-dimensional visualization in the second panel 420 is computed by a back end pipeline that is remote from visualization system 100 and is stored as two-dimensional datapoints 166 (Figure 1B) and/or as projections 196 or 197 (Figure 1C) in the discrete attribute value dataset 120. In some embodiments, the two-dimensional visualization 420 is computed by the visualization system 100 (Figure 1A).

[00198] Because the initial data is sparse, in some embodiments, in the case of GEX data, the two-dimensional visualization 420 of Figure 4A is prepared by computing a plurality of GEX principal component values 164 for each respective cell 126 in the plurality of cells based upon respective discrete attribute values 124 for each gene 122 in the respective cell 126. In some embodiments, the plurality of GEX principal component values is ten. In some embodiments, the plurality of GEX principal component values is between 5 and 100. In some embodiments, the plurality of GEX principal component values is between 5 and 50. In some embodiments, the plurality of GEX principal component values is between 8 and 35. Then, a dimension reduction technique is applied to the plurality of GEX principal components values for each respective cell 126 in the plurality of cells thereby determining a two-dimensional data point 166 for each cell 126 in the plurality of cells. Each respective cell 126 in the plurality of cells is then plotted in the second panel of Figure 4A based upon the two-dimensional data point for the respective cell.

[00199] Likewise, in some embodiments, in the case of ATAC data, the two-dimensional visualization 420 of Figure 4B is prepared by computing a plurality of ATAC principal component values 165 for each respective cell 126 in the plurality of cells based upon respective

ATAC fragment counts 125 for each ATAC peak 123 in the respective cell 126. In some embodiments, the plurality of ATAC principal component values is ten. In some embodiments, the plurality of principal component values is between 5 and 100. In some embodiments, the plurality of ATAC principal component values is between 5 and 50. In some embodiments, the plurality of ATAC principal component values is between 8 and 35. Then, a dimension reduction technique is applied to the plurality of ATAC principal components values for each respective cell 126 in the plurality of cells thereby determining a two-dimensional data point 167 for each cell 126 in the plurality of cells. Each respective cell 126 in the plurality of cells is then plotted in the second panel of Figure 4B based upon the two-dimensional data point for the respective cell.

[00200] *t-SNE*.

[00201] In some embodiments, the dimension reduction technique that is applied to the GEX principal components or the ATAC principal components is t-Distributed Stochastic Neighboring Entities (t-SNE) (*e.g.*, as illustrated in Figures 24, 25, and/or 30). In the case of GEX data, the collection of two-dimensional datapoints 166 upon application of t-SNE is stored as a GEX t-SNE projection 196. In the case of ATAC data, the collection of two-dimensional datapoints 167 upon application of t-SNE is stored as an ATAC t-SNE projection 198 (Figure 1D). One embodiment of the present disclosure provides a back end pipeline for computing such t-SNE projections that is performed on a computer system other than the visualization system 100. In other embodiments, the visualization system 100 computes the t-SNE projections. t-SNE is particularly well-suited for embedding high-dimensional data (*e.g.*, the GEX principal components values 164 or the ATAC principal components values 165) computed for each measured cell based upon the measured discrete attribute value (*e.g.*, expression level) of each gene 122 (*e.g.*, expressed mRNA) or the ATAC fragments counts for each ATAC peak in a respective cell as determined by principal component analysis into a space of two, which can then be visualized as a two-dimensional visualization (*e.g.* the scatter plot of second panel 420) of Figure 4A (GEX t-SNE projection 196) or Figure 4B (ATAC t-SNE projection 198). In some embodiments, t-SNE is used to model each high-dimensional object (the principal components of each measured cell) as a two-dimensional point in such a way that cells with similar principal component values are modeled as nearby two-dimensional datapoints 166 / 167 and cells with

dissimilar principal component values are modeled as distant two-dimensional datapoints 166 / 167 in the two-dimensional plot.

[00202] *UMAP.*

[00203] In some embodiments, the dimension reduction technique that is applied to the GEX principal components or the ATAC principal components is a uniform manifold approximation and projection (UMAP). In the case of GEX data, the collection of two-dimensional datapoints 166 upon application of UMAP is stored as a GEX UMAP projection 197. In the case of ATAC data, the collection of two-dimensional datapoints 167 upon application of UMAP is stored as an ATAC UMAP projection 199 (Figure 1D). One embodiment of the present disclosure provides a back end pipeline for computing such UMAP projections that is performed on a computer system other than the visualization system 100. In other embodiments, the visualization system 100 computes the UMAP projections. UMAP is particularly well-suited for embedding high-dimensional data (*e.g.*, the GEX principal components values 164 or the ATAC principal components values 165) computed for each measured cell based upon the measured discrete attribute value (*e.g.*, expression level) of each gene 122 (*e.g.*, expressed mRNA) or the ATAC fragments counts for each ATAC peak in a respective cell as determined by principal component analysis into a space of two, which can then be visualized as a two-dimensional visualization. In some embodiments, UMAP is used to model each high-dimensional object (the principal components of each measured cell) as a two-dimensional point in such a way that cells with similar principal component values are modeled as nearby two-dimensional datapoints 166 / 167 and cells with dissimilar principal component values are modeled as distant two-dimensional datapoints 166 / 167 in the two-dimensional plot.

[00204] Affordance 460 of Figures 4A and 4B can be used to select among the GEX t-SNE projection 196, GEX UMAP projection 197, ATAC t-SNE projection 198, and ATAC UMAP projection 199 for the discrete attribute value dataset 120.

[00205] *Other dimension reduction methods.*

[00206] In some embodiments, referring to block 238 of Figure 2C, rather than using t-SNE or UMAP, the dimension reduction technique used to reduce the principal component values 164 or 165 to a corresponding two-dimensional datapoint 166 or 167 is Sammon mapping, curvilinear components analysis, stochastic neighbor embedding, Isomap, maximum variance

unfolding, locally linear embedding, or Laplacian Eigenmaps. These techniques are described in van der Maaten and Hinton, 2008, “Visualizing High-Dimensional Data Using t-SNE,” Journal of Machine Learning Research 9, 2579-2605, which is hereby incorporated by reference. In some embodiments, the user has the option to select the dimension reduction technique. In some embodiments, the user has the option to select t-SNE, Sammon mapping, curvilinear components analysis, stochastic neighbor embedding, Isomap, maximum variance unfolding, locally linear embedding, or Laplacian Eigenmaps.

[00207] Referring to block 234 of Figure 2C, and as illustrated in Figures 4A and 4B, in some embodiments each cluster 158 in the first plurality of clusters (GEX clusters) or the second plurality of clusters (ATAC clusters) is assigned a different graphic or color code. Further, each respective cell 126 in the plurality of entities is coded in the second panel 420 with the different graphic or color code for the cluster 158 the respective cell has been assigned.

[00208] Referring to block 240, in some embodiments, each of the respective plurality of principal component values is derived from the discrete attribute values of each gene or ATAC peak in a corresponding cell in the plurality of cells by principal component analysis. In some embodiments, such analysis is performed on a computer system remote from the visualization system 100 prior to storing the discrete attribute value dataset 120 in persistent memory. In such embodiments the dataset includes each respective plurality of principal component values.

[00209] Now that the overall functionality of the systems and methods of the present disclosure has been introduced, attention turns to additional features afforded by the present disclosure. As illustrated in Figure 4A, for each cluster 158, there is a row in the lower panel 404 that illustrates the fold change (*e.g.* $\log_2$ fold change) of the average discrete attribute value 124 for each respective gene 122 across the cells 126 of the cluster 158 represented by the row compared to the average discrete attribute value 124 of the respective gene 122 in the remainder of the population of cells represented by the discrete attribute value dataset 120.

[00210] Correspondingly, as illustrated in Figure 4B, for each cluster 159, there is a row in the lower panel 404 that illustrates the fold change (*e.g.* $\log_2$ fold change) of the average ATAC peak fragment count 125 for each ATAC peak 123 across the cells 126 of the cluster 159 represented by the row compared to the average ATAC peak fragment count 125 of the

respective ATAC peak 123 in the remainder of the population of cells represented by the discrete attribute value dataset 120.

[00211] Referring back to Figure 4A, legend 408 on the right of the lower panel 404 indicates the $\log_2$ fold change compared to the average expression in the population. For instance, in one color coding scheme, red means higher abundance (higher discrete attribute values 124), blue means lower abundance (lower discrete attribute values 124), in a given cluster 158 as compared to the average abundance in the population. In Figure 4A, $\log_2$ fold change in expression refers to the $\log_2$ fold value of (i) the average number of transcripts (discrete attribute value) measured in each of the cells of the subject cluster that map to a particular gene 122 and (ii) the average number of transcripts measured in each of the cells of all clusters other than the subject cluster that map to the particular gene.

[00212] Referring to Figure 4B, here legend 408 on the right of the lower panel 404 indicates the $\log_2$ fold change compared to the average fragment count in the population. For instance, in one color coding scheme, red means higher fragment counts (higher ATAC fragment counts 125), blue means lower fragment counts (lower ATAC fragment counts 125), in a given cluster 159 as compared to the average fragment count in the population. In Figure 4B, $\log_2$ fold change in fragment counts refers to the $\log_2$ fold value of (i) the average number of ATAC fragment counts measured in each of the cells of the subject cluster that map to a particular ATAC fragment 123 and (ii) the average number of ATAC fragment counts measured in each of the cells of all clusters other than the subject cluster that map to the particular ATAC peak.

[00213] *Linked windows.* To allow users to see common characteristics or compare different images at once, one aspect of the present disclosure makes use of novel linked windows. Referring to Figure 4B, clicking on the “Add Window” affordance 465 brings up a list of projections 470 for the discrete attribute value dataset 120 to open in a linked window. Thus, referring to Figure 4B, the GEX t-SNE projection 196 is visible in panel 420 and the user has the option of adding a window for the ATAC t-SNE projection 198, ATAC UMAP projection 199, GEX UMAP projection 197, or another instance of the GEX t-SNE projection 196. In fact, the user can add multiple instances of any of these projections by successive use of the “Add Window” affordance. Clicking on one of the projections listed in panel 470 opens that projection in a smaller window within the operating system 116. Actions taken in the main

window, such as changing the active category, or showing expression or accessibility for a particular feature, will propagate to the linked windows.

[00214] Figure 5 shows three concurrent views of the same discrete attribute value dataset 120: the ATAC UMAP projection 199, the GEX t-SNE projection 196 and the GEX UMAP projection 197, with the same feature, ATAC peak chr2:2311671099-231673368 highlighted. In other words, for each projection, each respective cell in the plurality of cells represented by the discrete attribute dataset 120 are color coded in accordance with scale 502 by the ATAC fragment count 125 at ATAC peak chr2:2311671099-231673368. Zoom, pan and split operations are separate per window. In this manner, it is also easy to see whether there is coherence between gene expression-derived clusters 158 and ATAC-derived clusters 159.

[00215] With both ATAC and GEX projections displayed, a user can toggle between the ATAC graph-based clusters 159 and the GEX graph-based clusters 158 using a combination of affordances 502 and 502. Specifically, affordance 502 can be used to toggle between “Feature Accessibility Expression” as illustrated in Figure 5 and “Categories” as illustrated in Figures 4A and 6. When affordance 502 is in “Categories” mode, the user can then select between GEX clusters 158 and ATAC clusters 159 using affordance 450 as illustrated in Figures 4B and 6. In particular, Figure 4B shows that the user can use panel 450 to select between ATAC graph-based clusters, ATAC K-means clusters, GEX graph-based clusters, and GEX K-means clusters. Such toggling between ATAC graph-based clusters and GEX graph-based clusters provides one basis for determining whether certain clusterings appear concordant between the GEX and ATAC data. Moreover, using the split view affordance 602 in the main window 420 allows for cross-projection overlap analysis in further detail. For instance, in Figure 6 affordance 502 is toggled to “Categories,” affordance 450 is used to select GEX graph-based clustering, and affordance 602 is used to select ATAC graph-based clusters. Thus, in Figure 6, the cells in the plurality of cells in the discrete attribute dataset 120 for each ATAC graph based cluster 159 are shown in separate subpanels in the main panel 420 of Figure 6. Moreover, each such cell is mapped into the subpanel based on its two-dimensional UMAP coordinates. Finally, each cell is uniquely colored not by which graph based ATAC cluster 159 it has been assigned, but rather, by which GEX based cluster 158 it has been assigned. Thus, from Figure 6, it is apparent that there is a strong correlation between the GEX Graph-Based Cluster 7 and the ATAC Graph-

Based Cluster 8 because almost all of the cells in ATAC Graph-Based Cluster 8 in the main panel 402 are colored with the color assigned to GEX Graph-Based Cluster 7.

[00216] Referring to Figure 5, in some embodiments, linked windows (e.g., windows 510 and 520 open initially in miniaturized view as illustrated in Figure 5, where only the projection and a button 530 to expand the window to a full panel is shown. However, when using a mouse cursor to hover over a linked window (e.g., window 510), more options are revealed that provide a subset of common actions, such as the ability to pan and zoom a linked window. However, the linked windows are still predominantly controlled by manipulating the original, or anchor window 500.

[00217] Referring to Figure 5, changes to the anchor window 500 will propagate automatically to the other linked windows (e.g., windows 510 and 520), such as selection of active clusters (which clusters are displayed across all the linked windows), selecting an individual cluster, creating a new cluster or modifying a cluster, selecting one or more genes or ATAC peaks to show feature expression (gene, peak), changing cluster membership, changing individual cluster colors or the active expression color scale, and in selecting transcription factor motifs. However, features such as panning, zooming, and window sizes remain independent in the anchor and linked windows.

[00218] Referring to Figure 5, it is possible to expand a linked window from mini-mode to access the full range of visualization options by clicking on the expand affordance 530. Clicking affordance 530 within the linked window again will shrink the linked window (e.g., 510, 520, *etc.*) back to mini-mode. In some embodiments, changes to the discrete attribute value dataset 120 in any window are saved through the anchor window. Thus, referring to Figure 5, any change to the discrete attribute value dataset 120 in windows 510 or 520 must be saved through window 500 in such embodiments. As such, using linked windows avoids having to jump back and forth, making the investigation fluid and intuitive.

[00219] It is also possible to have other linked windows open to concurrently view additional discrete attribute value datasets 120. To avoid confusion, when multiple attribute value datasets are displayed, the color of the button 540 (Figure 5) on the main window 500 and the affordances 530 of the windows linked to the main window of each discrete attribute value dataset will adopt its own common unique color that is different than that of any other discrete attribute value

datasets 120 that are displayed. For instance, if an additional discrete attribute dataset 120 was displayed along with the one illustrated in Figure 5, the logo 540 of main window 500, and the affordance 530 of windows 510 and 520 of the originally displayed discrete attribute value dataset would have a common first color while the logo 540 of the main window 500 of the additional discrete attribute dataset 120 along with the displayed affordances 530 of windows linked to it would have a common second color different from the first color.

[00220] Moreover, linked windows are not limited to spatial discrete attribute value datasets 120. Most gene expression datasets have both t-SNE and UMAP projections 121 (see United States Patent Application No. 16/442,800 entitled “Systems and Methods for Visualizing a Pattern in a Dataset,” filed June 17, 2019) that can be linked and viewed at the same time in a similar fashion.

[00221] While linked windows have been illustrated in conjunction with showing mRNA-based UMI abundance as well as ATAC peak count, they can also be used to illustrate the quantification of other analytes, arranged in two-dimensional space using dimension reduction algorithms such as t-SNE or UMAP, including any combination of intracellular proteins (*e.g.*, transcription factors), cell methylation status, other forms of accessible chromatin (*e.g.*, DNase-seq, and/or MNase-seq), metabolites, barcoded labelling agents (*e.g.*, the oligonucleotide tagged antibodies) and V(D)J sequences of an immune cell receptor (*e.g.*, T-cell receptor), perturbation agent (*e.g.*, a CRISPR crRNA/sgRNA, TALEN, zinc finger nuclease, and/or antisense oligonucleotides). Any of these can be compared to each other or to mRNA-based UMI abundance and/or ATAC peak count.

[00222] *Linkage matrix visualization and analysis.*

[00223] *Linkage table.* Advantageously, some embodiments of the present disclosure provide the ability to review linkages between features, such as ATAC peaks and genes that are in the above described feature-linkage matrix 187, in a tabular fashion or in an interactive graphical fashion. In some embodiments, the tabular view is accessed by clicking on the list-with-arcs affordance (*e.g.*, icon) 702 of Figure 7. This brings up a Feature Links table 704 which remained empty as illustrated in Figure 7 until a search query is entered into search query box 706. A user may type a gene or peak of interest into search query box 706. The search query box will generate autocomplete options for both genes and ATAC peaks based on gene and

ATAC peaks represented in the active discrete attribute value dataset 120. When a feature is autocompleted or otherwise entered into the search query box 706, a query will generate a table of linked features and display them in the Feature Links Table 704, as shown for gene CD69 in Figure 8. This table contains the name of both the original feature of interest (CD69 in Figure 8) and the linked features 802, the significance 192 of such linkages, and the correlation 191 and the distance between features. Clicking on any of column headers in table 704 will sort the table of linked features by the values of that column. For instance, clicking on the column header of significance 192 column a first time will sort the linked features of table 704 from highest to lowest significance to CD69. Then, clicking on the column header of significance 192 column a second time will sort the linked features from lowest to highest significance to CD69. In this manner, a user can rapidly get a picture of other features strongly linked to the source feature (CD69 in Figure 8).

[00224] Linkage table 704 can also be accessed from other locations. For instance, referring to Figure 9, clicking on a gene in the significant genes table 900 brings up the menu of options 902. Option 904 of menu 902 is the option to view features linked to the selected gene, TRABD2A in Figure 9, as the source feature in a linkage matrix table. In addition, referring back to Figure 8, clicking on any one of the features in the ‘Feature Link’ column 802 brings up a menu similar to that of menu 902, through which the user can recenter the linkage matrix table around the selected new feature. A user may also use the menu to view expression or accessibility levels graphically over the 2-D projection, or add the feature to a list for later retrieval.

[00225] *Linkage viewer.*

[00226] Referring to Figure 10, the systems and methods of the present disclosure also provide a linkage mode (linkage viewer 1110) within an ATAC peak viewer 1002. The ATAC peak viewer 1002 is accessed by clicking on affordance 1004. From the Peak Viewer panel 1002, selecting “Linkage” 1006 from the mode selector 1008 loads the linkage viewer 1010. In a similar manner to the tabular view of Figure 8, typing in the name of a feature at search query prompt 1012 will trigger an autocomplete menu.

[00227] Entering CD69 into prompt 1012 of Figure 10 will bring up a view like the one shown in Figure 11, where the linkage viewer 1010 is now anchored to the selected feature of

interest (in this case, gene CD69). In the linkage viewer, there are three plots. First, there is an arc plot 1102, which illustrates the locations and correlations of the features linked to the subject feature, here CD69. These locations and correlations are drawn from the feature-linkage matrix 187, as discussed above in conjunction with Figure 1D. Second, there is a peak graph 1104, showing the positions and frequencies of peaks, on a cell cluster by cell cluster basis, that are linked to the subject feature. Third, there is a gene expression panel 1108, showing abundance, on a cell cluster by cell cluster basis, of either (i) all genes linked to the feature (gene or ATAC peak) anchoring the view or (ii) the gene anchoring the current view. In Figure 10, gene expression panel 1108 shows, on a cell cluster by cell cluster basis, the abundance of the anchoring gene CD69. Figure 11 thus illustrates how a feature-linkage matrix 187 is used to service the selection of a first gene 122 in the plurality of genes or a first ATAC peak 123 in the plurality of ATAC peaks at affordance 1105. In response to this selection at affordance 1105, the feature-linkage matrix 187 is used to obtain and provide a first plot 1010 comprising a graphical indicator for each gene in the plurality of genes and/or each peak in the plurality of peaks linked to the first gene or the first ATAC peak, by order of distance apart from the first gene or the first ATAC peak in the reference genome. In some embodiments, as illustrated in Figure 11, the respective graphical indicator for each respective gene in the plurality of genes or each respective peak in the plurality of peaks linked to the first gene or the first ATAC peak is provided for each respective cluster group. Figure 11, for instance, shows how a respective graphical indicator 1110 of a particular gene linked to CD69 is provided for each respective cluster group in the displayed plurality of cluster groups one through thirteen. In some embodiments, as illustrated for the graphical indicators 1110-1 through 1110-13 of Figure 11, each respective graphical indicator is dimensioned within the first plot 1110 to represent a proportion of cells in the respective cluster group that have a non-zero value for the discrete attribute value of the respective linked gene or a non-zero value for the ATAC fragment count of the respective linked ATAC peak. Thus, in Figure 11, each graphical indicator 1110-1 through 1110-13 is sized to reflect the proportion of cells in the respective cluster group that have a non-zero value for the discrete attribute value of the respective gene represented by the corresponding graphical indicator 1110.

[00228] By default, the genomic window represented by the linkage viewer 1010 will span all the features linked to the anchor feature. The arc plot 1102 and peak graph 1104 share the

same genomic x-axis, which are physical locations on the genome. The features linked to CD69 roughly span chr12:9600000-10200000. In some embodiments, this window is 2 megabases wide as discussed above in conjunction with the feature-linkage matrix 187 illustrated in Figure 1D. A user may zoom into a narrower region of the linkage view 1010 by dragging their mouse over a window in the arc plot 1102 or peak graph 1104.

[00229] The arc plot 1102 illustrates the direction and magnitude of the linkages between the anchor feature and linked features. In the embodiment illustrated in Figure 11, blue arcs (left of the anchoring feature in plot 1102) represent positive correlations, implying that cells with higher levels of CD69 expression have open chromatin at those linked peak regions, and regions from cells not expressing CD69 may be inaccessible. Red arcs (right of the anchoring feature in arc plot 1102) are negative correlations, implying the opposite. The height of the peaks are proportional to the correlative R-value 191; the closer to -1 or 1, the higher the R value, and the stronger the correlation. Other embodiments of the present disclosure alter the color, width or transparency of the peak curves of Figure 1102 in relation to the statistical significance of the linkage. For instance, more arcs representing more statically significant links are thicker than arc representing less statically significant arcs in some embodiments. The y-axis of the arc plot 1102 shows the absolute correlation of all the peaks, which is dynamic relative to the linked feature with the highest r-value.

[00230] The peak graph 1104 below the arc plot shows the location, size, and frequency of all peaks linked to the anchor feature within the active clustering. In Figure 11, the active clustering is the precomputed Louvain graph-based method over the gene expression-derived clusters. Changing the active clustering will change the clusters within the peak graph. For instance, the clustering can be changed to k-means GEX clustering, Louvain graph-based ATAC clustering, or k-means ATAC clustering, to name some non-limiting examples of alternative clusterings. The heights of the peaks in each cluster track are proportional to the percentage of cells within that cluster where the chromatin is open within that peak region.

[00231] Clicking on an arc in the arc plot 1102 will highlight the arc in the arc plot, and the target peak within the peak graph 1104. Figure 12 shows this. Arc 1204 has been selected in arc plot 1102 and so the linked peak 1202 corresponding to arc 1204 has been highlighted in the

peak graph 1104. In this manner, the differences between clusters of open chromatin frequencies for the linked peak can be visualized.

[00232] In some embodiments, gene expression panel 1108 can show the differences in gene expression between clusters for all genes linked to the anchor feature. To show this more precisely, Figure 13A re-anchors the linkage view around the feature with the highest positive correlative value to CD69: chr12:9760094-9761921. The gene expression panel 1108 now shows the distribution of expression among all genes 1302 linked to that peak, as determined by the feature-linkage matrix 187, within the currently selected cell clusters. For instance, in the case of chr12:9760094-9761921, there are a total of 10 linked genes 1302 (1302-1 through 1302-10), and the expression of each of these genes, on a cluster by cluster basis is now displayed in the gene expression panel 1108.

[00233] The order of display of genes 1302 in the gene expression panel 1108, from top to bottom, is in genomic order in Figure 13A. In other embodiments, the genes 1302 are presented in order of highest to lowest significance 192. In other embodiments, the genes 1302 are presented in order of lowest to highest significance 192. In other embodiments, the genes 1302 are presented in order of highest to lowest correlative value 191. In other embodiments, the genes 1302 are presented in order of lowest to highest correlative value 192. In still other embodiments, the genes 1302 are presented in order of closest to furthest genetic distance to the anchoring feature. In still other embodiments, the genes 1302 are presented in order of furthest to closest genetic distance to the anchoring feature. In some embodiments, only the N genes closest to the anchoring feature are shown in panel 1108, where N is a positive integer between 1 and 20 (*e.g.*, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or 20).

[00234] Hovering over a row 1302 in the gene expression panel 1108 shows the identity of the linked gene, as well as the average expression per cluster. In some embodiments, this is absolute expression (UMI/cell). In other embodiments this is relative (average log-normalized count per cell). The gene expression panel 1108 shows a clear correlative pattern and anti-pattern. Genes positively correlated to the specified peak (rows 1-4, including CD69 on row 1302-4) have higher levels of gene expression in clusters 1, 2, 6, and 7, whereas genes negatively correlated to the anchoring peak have higher levels of gene expression in clusters 3, 4, 8 and 9. Cluster 5, on the other hand, shows high level of expressions regardless throughout all the cell

clusters. The y-axis of each gene expression plot is independent due to differences in overall abundance; the purpose of the panel is to show common patterns of relative expression.

[00235] In some embodiments (not shown in Figure 13A), the accessibility or expression of the anchor feature is illustrated, on a cell cluster by cell cluster basis, above the gene expression panel 1108.

[00236] In some embodiments (not shown in Figure 13A), a user clicks on a gene annotation in genes display 1304, a peak in peak graph 1104, or an arc in arc plot 1102. In response, a distribution (ATAC peak count or GEX data depending on the nature of the selected feature), on a cell cluster by cell cluster basis, of the selected feature is illustrated in gene expression panel 1108. In some such embodiments, the distribution, on a cell cluster by cell cluster basis, of the anchoring feature is also shown in gene expression panel 1108. In this way, a direct comparison of the distribution of the selected linked feature with the anchor feature, on a cell cluster by cluster basis, can be made. Referring to Figure 13B, when a user is searching for a gene (*e.g.*, LEF1 at affordance 1105 as illustrated in Figure 13B), they will see the average UMIs/cell per cell cluster 158 or 159 for that gene in plot 1108. When a user clicks on a particular linked ATAC peak 123 (*e.g.*, selection of ATAC Peak chr4:108355456-108356170) in the arc plot 1102, they will also see the average cut sites/cell per cell cluster 158 or 159 for the linked ATAC peak 123 (*e.g.* chr4:108355456-108356170) as illustrated in plot 1320 of Figure 13C.

Conversely, referring to Figure 13D, when a user is searching for a peak (*e.g.*, entry of ATAC Peak chr4:108355456-108356170 at affordance 1105), they will see the average cut sites/cell, per cell cluster 158 or 159 in plot 1330 for the searched-for ATAC peak 123 (ATAC Peak chr4:108355456-108356170) as well as all the average UMIs/cell per cell cluster 158 or 159 for all linked genes. In some embodiments, hovering over a column in panel 1330 of Figure 13D shows the numerical average cut sites/cell of the ATAC peak 123 (chr4:108355456-108356170) across the cells in the cell cluster represented by the selected column. In the case of ATAC Peak chr4:108355456-108356170, there three linked genes: LEF1 represented by plot 1340-1, RPL34 represented by plot 1340-2, and OSTC represented by plot 1340-3. Referring to Figure 13D, when a user clicks on a link in arc plot 1102 (*e.g.*, the chr4:108355456-108356170 - chr4:108433617-108434308 link), the average cut sites/cell, per cell cluster 158 or 159, for the searched-for peak (chr4:108355456-108356170 in Figure 13E) is displayed as graph 1360 and also the average UMIs/Cell or average Cut Sites/Cell, per cell cluster 158 or 159, for the linked

feature, whether it be linked to a gene or another peak, is displayed as graph 1370 (as illustrated for the linked feature chr4:108433617-108434308 in Figure 13E).

[00237] In some embodiments, as illustrated in Figure 13A, when the linkage viewer 1010 is anchored on a peak instead of a gene, dashed arcs in the arc plot 1102 are peak-to-peak links, while solid arcs are peak-to-gene links. In alternative embodiments, not shown, when the linkage viewer 1010 is anchored on a peak instead of a gene, arcs in the arc plot 1102 representing peak-to-peak links have a first linewidth and/or color and arcs in the arc plot 1102 representing peak-to-gene links have a second linewidth and/or color different from the first linewidth and/or color.

[00238] In addition, in some embodiments, the gene annotations 1304 between the arc plot 1102 and the peak graph 1104 highlight linked genes in black, while greying out genes in the genomic window that did not have strong linkage to the anchor peak.

[00239] *Linkage filter options.*

[00240] In some embodiments, one or more tools for filtering linkages are provided. These one or more tools are useful, for example, when displaying a wide genomic region. These one or more tools are also useful, for example, when displaying an anchor feature linked to a large number of other features. Referring to Figure 14, clicking on the options icon 1402 an options menu 1404. The options menu includes a correlation slider (filter) 1406 that filters linkages by absolute r-value 191 and, optionally, a significance slider (filter) 1408, which filters linkages by significance score (negative log false-discovery-rate) 192. Linkages that do not meet the filtering criterion or criteria will be hidden from the arc plot 1102, and linked genes that do not meet the filtering criterion or criteria are greyed out in the gene expression panel 1304.

[00241] In some embodiments, not shown in Figure 14, options menu 1404 includes an additional filter that allows a user to limit the view within linkage viewer 1010 to gene-to-peak linkages. That is, only gene-to-peak linkages are displayed in the arc plot 1102 of linkage viewer 1010 when this filter is enforced, and when the filter is not enforced, both peak-to-peak and gene-to-peak linkages are displayed in the linkage viewer.

[00242] *Additional Features.*

[00243] Referring to Figure 5, selection of a particular ATAC peak 123 in the lower panel 404 causes each respective cell 126 in the upper panel 420 to be colored on a color scale that

represents the ATAC fragment count 125 of that respective ATAC peak 123 in each respective cell 126. So, for example, referring to Figure 5, when a user clicks on the ATAC peak 550 entitled chr2:231671099-231673368 in the lower panel 404, each respective cell 126 in the upper panel 420 is colored to reflect the ATAC fragment count 125 for chr2:231671099-231673368 in the respective cell 126. From Figure 5, upper panel, it is seen that high expression of chr2:231671099-231673368 is exhibited in portion 560 of the ATAC UMAP. In instances where the lower panel is showing genes rather than ATAC peaks, selection of such genes will cause the cells in the upper panel 420 to be colored to reflect the discrete attribute value for the selected gene. This is seen for example, in Figure 9, where the cells in the upper panel 420 have been colored to reflect the CCR7 discrete attribute value of CCR7 in each cell.

[00244] In some embodiments, the user can select more than one gene 122 or ATAC peak 123 in the lower panel 404 and thereby cause the upper panel 420 to concurrently illustrate the discrete attribute value 124 or ATAC fragment count 125 of each of the more than one gene 122 or more than one ATAC fragment count 125 in each respective cell 126 in the discrete attribute value dataset 120 at the same time.

[00245] Referring to Figure 4B, an alternate view to the bottom panel 404 of Figure 4B is shown by clicking on affordance 470. Upon selection of affordance 470, a tabular representation of the log₂ ATAC fragment counts 125 of the heat map of Figure 4B is illustrated in column format in Figure 5, whereas the heat map of Figure 4B shows the log₂ ATAC fragment counts 125 for ATAC peaks 123-1 through 123-L in a row for each cluster. For instance, in Figure 4B, the average ATAC fragment count 125 in cluster 159-1 for each ATAC peaks 123-1 through 123-L is shown in the first row in panel 404. Turning to Figure 5, the user can select any respective cluster 159 by selecting the column label for the respective cluster (column labels for clusters 3, 4 and 5 are marked as 570-3, 570-4, and 570-5, respectively in Figure 5). This will re-rank all the ATAC peaks 123 such that those ATAC peaks 123 that are associated with the most significant ATAC fragment counts 125 in the selected cluster 159 are ranked first. Selecting the same cluster 159 by again selecting the column label for the respective cluster will re-rank all the ATAC peaks 123 such that those ATAC peaks 123 that are associated with the least significant ATAC fragment counts 125 in the selected cluster 159 are ranked first. Moreover, a p-value 580 is provided for the ATAC fragment count 125 of each ATAC peak 123 in the selected cluster to provide the statistical significance of the ATAC fragment count 125 in the selected cluster 159

relative to the ATAC fragment count 125 of the same ATAC peak 123 in all the other clusters 159. For instance, in Figure 5, ATAC cluster 159-5 has been selected through column header 570-5 and thus the p-values 580 listed are the ATAC fragment count 125 of each ATAC peak 123 in the cluster 159-5 to provide the statistical significance of the ATAC fragment count 125 in cluster 159-5 relative to the ATAC fragment count 125 of the same ATAC peak 123 in all the other clusters 159. In some embodiments, these p-values are calculated based upon the ATAC fragment count 125, not the $\log_2$ values used for visualization in the heat map 402 of Figure 5B. In some embodiments, p-values are annotated with a star system, in which four stars means there is a significant difference between the selected cluster (cluster 158-5 in Figure 5) and the rest of the clusters for a given ATAC peak, whereas fewer stars means that there is a less significant difference in the ATAC fragment count 125 between the ATAC peak in the selected cluster relative to all the other clusters. While the above-discussion has been in relation to ATAC peaks 123, the same tools are provided when genes 122 and their discrete attribute values are displayed instead. Thus, the systems and methods of the present disclosure provide quantitative inspection of the difference in discrete attribute value 158 in any one cluster 158 or the ATAC fragment count 159 in any one cluster 158 (GEX clusters) or cluster 159 (ATAC clusters) relative to the rest of the clusters. The ranked table (e.g., 404 of Figure 5) can be exported using affordance 590.

[00246] Moreover, a user can identify a set of cells of interest using the lasso selection tool 592 and the draw selection tool 594 provided in the upper panel of Figure 5. Once identified, such cells can be assigned a custom classification and/or exported.

[00247] Referring again to Figure 4A, the heatmap 402 provides a $\log_2$ differential that is optimal in instances where the cell 126 is a cell and the discrete attribute value 124 represents the number of transcripts that map to a given gene in the cell in order to provide a sufficient dynamic range over the number of transcripts seen per gene in the given cell. In some embodiments, $\log_{10}$ differential expression is used instead. However, it is expected that $\log_{10}$ does not provide sufficient dynamic range for appropriate visualization of the relative expression of gene data in the k-means clusters in some instances. This is because the distinction between zero and one count in the raw data is also fairly important. Because of this, it is not desirable to drown the difference between zero and one with the difference between nine and ten. The difference between zero and one in the discrete attribute value 124 differential (between one cluster and the

other clusters) is a significant jump and so a log scale that is able to at least have that floor where “zero” is one color in the heat map 402 and “one” is something that is visually different from “zero.” Hence the $\log_2$ scale is used in the heat map 402 illustrated in Figure 4A.

[00248] Figures 4 through 14 illustrate the analysis of data that comes out of a cell analysis (e.g., single cell sequencing) pipeline. Another aspect of the present disclosure handles situations in which the pipeline consists of multiple GEX classes 172 and/or ATAC classes 173 of cells 126. That is, situations in which each such sample comprises first discrete attribute values 124 for each respective gene 122 (e.g., mRNA that map to a particular gene in a plurality of genes) and/or first ATAC fragment counts 125 for each respective ATAC peak 123 in each cell 126 in a first plurality of cells under a first condition (therefore representing first class 172 and/or first class 173), second discrete attribute values 124 for each respective gene 122 and/or second ATAC fragment counts 125 for each respective ATAC peak 123 in each cell in a second plurality of different cells under a second condition (therefore representing a second class 172 and/or second class 173), and so forth. In other situations, each such sample consists of first discrete attribute values 124 for each respective gene 122 (e.g., mRNA that map to a particular gene in a plurality of genes) and/or first ATAC fragment counts 125 for each respective ATAC peak 123 in each cell 126 in a first plurality of cells of a first type (first class 172 and/or first class 173), second discrete attribute values 124 for each respective gene 122 and/or second ATAC fragment counts 125 for each respective ATAC peak 123 in each cell in a second plurality of cells of a second type (second class 172 and/or second class 173), and so forth, where each such class 172 and/or 173 refers to a different cell type, a different disease state, a different tissue type, a different organ type, a different species, or different assay conditions or any of the foregoing. In some embodiments, the discrete attribute value dataset 120 contains data for cells from two or more such classes, three or more such classes, four or more such classes, five or more such classes, ten or more such classes 172, or 100 or more such classes 172 and/or classes 173.

[00249] Referring to Figure 15, in some embodiments, there are a plurality of ATAC categories 171 and each cell 126 is in each such category 171. In some such embodiments, each category 171 has one or more sub-categories, termed classes 173, which can be individually selected. In some embodiments, all such data is preloaded into a single discrete attribute value dataset 120. One example of a category 171 is ATAC K-means clustering (where the ATAC K-

means clustering is the “category” 170 and each K-means cluster 159 is an example of a class 173).

[00250] Other non-limiting examples of categories not shown in Figure 15 include “LibraryID” (where LibraryID is the category 171 and which library a cell originated from is the class 173), and “Cohort” (where Cohort is the category 171, and which patient sub-population a cell originated within the cohort is the class 173). Selection of such a category leads to a view in panel 420 in which each cell is color coded by a phenotypic status (*e.g.*, acute myeloid leukemia status associated with the selected category). The spatial representation of the cells in the panel 420 does not change by selection of such categories, only the labeling for the cells changes.

[00251] Turning to Figure 15, by selecting affordance 502, a dropdown menu is provided. The dropdown menu shows all the different categories 170 and/or categories 171 that are associated with each cell in the discrete attribute value dataset 120 as illustrated, for example, in Figure 4B). Returning to Figure 15, in the dataset illustrated in Figure 15, the category is “ATAC K-means” 171-1, the selection of which will provide the view of Figure 15 panel 420, in which each cell 126 is color coded by its K-means cluster 159 identity 173. As such, the K-means clustering itself is deemed a category 171 and the clusters 159 are each deemed a different class 173 of the category 171.

[00252] In some embodiments, where there is a category 171 in a discrete attribute value dataset 120 having ATAC classes 173, each respective cell in the discrete attribute value dataset 120 is a member of each respective ATAC category 171 and one of the ATAC classes 173 of each respective ATAC category 171. In some such embodiments, where the dataset 120 comprises a plurality of categories 171, each respective cell in the discrete attribute value dataset 120 is a member of each respective ATAC category 171, and a single class 173 of each respective ATAC category 171.

[00253] In some embodiments where there is a category 171 in a discrete attribute value dataset 120 that has no underlying classes 173, a subset of the cells in the dataset 120 are a member of the category 171. In some embodiments where there is a category 171 in a discrete attribute value dataset 120 having subclasses 173, only a portion of the respective cells in the dataset 120 are a member of the category 171. Moreover, each cell in the portion of the respective cells is independently in any one of the respective classes 173 of the category 171.

[00254] A user can select or deselect any category 171. Moreover, a user can select or deselect any combination of classes 173 in a selected category 171. Referring to Figure 15, in some embodiments, the user is able to click on a single category 173 (the categories are labeled as 173-1-2, 173-1-3, 173-1-4, and 173-1-5, respectively, in Figure 15) to highlight the cells belonging to that category (cluster) in the plot 420. In this way, by successively clicking on categories 173, the user is able to turn on or off the coloring of the cells corresponding to these categories in panel 410.

[00255] Because panel 420 is a GEX t-SNE projection, in two-dimensional space there is an appearance that the clusters 173 overlap each other. However, in the multi-dimensional space in which the clustering was performed, the clusters 173 do not overlap each other.

[00256] Referring to Figure 16, in some embodiments, there are a plurality of GEX categories 170 and each cell 126 is in each such category 170. In such embodiments, each GEX category 170 has one or more sub-categories, termed GEX classes 172, which can be individually selected. In some embodiments, all such data is preloaded into a single discrete attribute value dataset 120. Examples of categories 170 are illustrated in Figure 4B and include ATAC graph-based, ATAC K-means clustering, GEX graph-based clustering, and GEX K-means clustering. In Figure 16 GEX K-means 170-1 is selected and GEX K-means is the category 170 and each GEX K-means cluster 158 is an example of a GEX class 172.

[00257] In the dataset illustrated in Figure 16, selection of the category “GEX K-means” 170-1 provides the view of Figure 16 panel 420, in which each cell 126 is color coded by its GEX K-means cluster 158 identity. As such, the GEX K-means clustering itself is deemed a category 170 and the GEX clusters 158 are each deemed a different GEX class 172 of the GEX category 170.

[00258] In some embodiments, where there is a category 170 in a discrete attribute value dataset 120 having classes 172, each respective cell in the discrete attribute value dataset 120 is a member of each respective category 170 and one of the classes 172 of each respective category 170. In some such embodiments, where the dataset comprises a plurality of categories 170, each respective cell in the discrete attribute value dataset 120 is a member of each respective category 170, and a single class of each respective category 170.

[00259] In some embodiments where there is a category 170 in a discrete attribute value dataset 120 that has no underlying classes 172, a subset of the cells in the dataset 120 are a member of the category 170.

[00260] In some embodiments where there is a category 170 in a discrete attribute value dataset 120 having subclasses 172, only a portion of the respective cells in the dataset 120 are a member of the category 170. Moreover, each cell in the portion of the respective cells is independently in any one of the respective classes 172 of the category 170.

[00261] A user can select or deselect any category 170. A user can also select or deselect any combination of classes 172 in a selected category 170. Thus, referring to Figure 16, in some embodiments, the user is able to click on a single category 172 (the categories are labeled as 172-1-2, 172-1-3, 172-1-4, 172-1-5, and 172-1-6, respectively, in Figure 16) to highlight the cells belong to that category (cluster) in the plot 420. Selection of a highlighted category causes the cells in panel 420 that belong to that category to no longer be colored by that category. Instead, they are greyed out. Because panel 420 is a t-SNE projection, in two-dimensional space there is an appearance that the clusters 172 overlap each other. However, in the multi-dimensional space in which the clustering was performed, the clusters 172 do not overlap each other.

[00262] The presentation of the data in the manner depicted for example in Figures 4-16 advantageously provides the ability to determine the genes 122 whose discrete attribute values 124 separates (discriminates) classes 172 within a selected category based upon their discrete attribute values. The presentation of the data in the manner depicted for example in Figures 4-16 advantageously also provides the ability to determine the ATAC peaks 123 whose ATAC fragment count 125 separates (discriminates) classes 173 within a selected category 172 based upon their ATAC fragment count 125. To further assist with this, the significant genes (*e.g.*, Sig. genes) affordance 1602 is selected thereby providing two options “globally distinguishing” as illustrated in Figure 16 and locally distinguishing (not shown in Figure 16).

[00263] Referring to Figure 16, the significant feature comparison 1602 identifies the feature (genes 122 or ATAC peaks 123) whose values (discrete attribute values 124 or ATAC fragment counts 125) within the selected classes statistically discriminate with respect to the entire discrete attribute value dataset 120 (*e.g.*, finds genes expressed highly within the selected categories 170, relative to all the categories in the dataset 120 or finds ATAC peaks that high counts within the

selected categories, relative to all the categories in the dataset 120). Referring to Figure 17, the locally distinguishing option identifies the features (genes or ATAC peaks) whose values (GEX discrete attribute values 124 or ATAC counts 125) discriminate the selected classes without considering the values (discrete attribute values 124 or ATAC counts) in classes of cells that have not been selected.

[00264] In some embodiments, visualization system 100 comprises a plurality of processing cores 102 and the identification of features (genes or ATAC peaks whose values (discrete attribute values 124 or ATAC fragment counts 125) discriminate classes under either the globally distinguishing or locally distinguishing algorithms makes use of the processing cores 102 to independently build up needed statistics (*e.g.*, a measure of central tendency of the discrete attribute value) of individual features across a class and/or one or more categories of a class of cells (or the entire dataset).

[00265] To further illustrate, turning to Figure 17 in which the “GEX K-Means” category 170 option has been selected and the data for the cells in the GEX K-Means classes 172-2-1, 172-2-3 and 172-2-4 have been deselected, the locally distinguishing affordance 1202 of Figure 16 identifies the cells 126 whose discrete attribute values (*e.g.*, mRNA counts) uniquely discriminate the selected cluster 158-2 (class 172-1-2) from cluster classes 172-1-5 and 172-1-6 in the discrete attribute value dataset 120 without considering the values (discrete attribute values 124 or ATAC counts) in classes 172-1-1, 172-1-3, and 1-172-4. Such discriminating genes are listed out in the lower panel 404 of Figure 17. By contrast, as illustrated in Figure 16, the globally distinguishing option identifies the genes 122 in lower panel 404 whose discrete attribute values 124 discriminate the difference between the class 172-1-1 (158-1) and all the remaining classes 172 of the GEX K-Means category 170.

[00266] Advantageously, the systems and method of the present disclosure allow for the creation of new categories using the upper panel 420 and any number of classes within such categories using lasso 592 or draw selection tool 594 of Figure 17. So, for example, if a user would like to identify cell subtypes, this can be done by selecting a number of cells displayed in the upper panel 420 with the lasso tools. Moreover, they can also be selected from the lower panel 404 (*e.g.*, the user can select a number of cells by their discrete attribute values or ATAC fragment counts). In this way, a user can drag and create a class within a category. The user is

prompted to name the new category and the new class within the category. The user can create multiple classes of cells within a category. Once the classes of a category have been defined in this way, the user can compute the genes whose discrete attribute values discriminate between the identified user defined classes. In some such embodiments, such operations proceed faster than with categories that make use of all the cells in the discrete attribute value dataset 120 because fewer numbers of cells are involved in the computation. In some embodiments, the speed of the algorithm to identified genes that discriminate classes is proportional to the number of classes in the category times the number of cells that are in the analysis. For instance, in some embodiments identification of discriminating genes in the case where there are two classes and twenty-five cells takes about four to five seconds on a standard client device 100.

[00267] In some embodiments, a discrete attribute value dataset 120 can have data for up to 750,000 cells and still identify genes or ATAC peaks that discriminate between classes of a category in real time (*e.g.*, less than 30 minutes, less than 10 minutes, less than 5 minutes, or less than 1 minute).

[00268] *Discrete attribute value pipeline.*

[00269] As discussed above, methods for collecting ATAC and GEX data for discrete attribute value datasets 120 are described in United States Patent Application No. 16/789,287, entitled “Methods for Processing Nucleic Acid Molecules,” filed February 12, 2020, which is hereby incorporated by reference. In some embodiments, the Cell Ranger™ analysis pipelines perform secondary analysis and visualization of such data. In addition to performing standard analysis steps such as demultiplexing, alignment, and gene counting, Cell Ranger™ leverages the bar codes described in United States Patent Application No. 16/789,287, entitled “Methods for Processing Nucleic Acid Molecules,” filed February 12, 2020, which is hereby incorporated by reference to generate GEX expression data and ATAC fragment count data with single-cell resolution in the form of the discrete attribute value dataset 120. This data type enables applications including cell clustering, cell type classification, differential gene expression, and differential ATAC fragment count at a scale of hundreds to millions of cells. Moreover, as discussed above, because the pipeline delivers this information by indexing discrete attribute value 124 from cells on an individual cell basis using barcodes, the data from such single cells can be combined with the data from other pipelines that make use of barcodes to track data from

single cells, such as the V(D)J Pipeline described in United States Patent Application No. 15/984,324, entitled “Methods for Clonotype Screening,” filed May 19, 2019, which is hereby incorporated by reference, to provide unique biological insight into underlying molecular mechanisms associated with cell samples.

[00270] In the embodiments of the present disclosure, techniques for analyzing biological samples are provided. The techniques involve acquiring a sample (*e.g.*, a tumor biopsy, a sample of any tissue, body fluid, *etc.*) and processing the sample to acquire data from each cell in the sample for computational analysis. Each cell in the sample is barcoded, at a minimum, as discussed above. The acquired data is stored in a certain manner, for example, in specific data structure(s), for consumption by one or more processors (or processing cores) that are configured to access the data structures and to perform computational analysis such that biologically meaningful patterns within the sample are detected. The computational analysis and associated computer-generated visualization of results of the computational analysis on a graphical user interface allow for the observation of properties of the sample that would not otherwise be detectable. In particular, in some embodiments, each cell of the sample is subjected to analysis and characteristics of each cell within the sample are obtained such that it becomes possible to characterize the sample based on differentiation among different types of cells in the sample. For example, the clustering analysis, as well as other techniques of data analysis described above, reveal distributions of cell populations and sub-populations within a sample that would not be otherwise discernable. This leads to the discovery of novel cell types, or to the novel discovery of relationships between (A) aspects of the cellular phenotypes, such as genome (*e.g.*, genomic rearrangements, structural variants, copy number variants, single nucleotide polymorphisms, loss of heterozygosity, rare variants), epigenome (*e.g.*, DNA methylation, histone modification, chromatin assembly, protein binding), transcriptome (*e.g.*, gene expression, alternative splicing, non-coding RNAs, small RNAs), proteome (*e.g.*, protein abundance, protein-protein interactions, cytokine screening), metabolome (*e.g.*, absence, presence, or amount of small molecules, drugs, metabolites, and lipids), and/or phenome (*e.g.*, functional genomics, genetics screens, morphology), and (B) particular phenotypic states, such as absence or presence of a marker, participation in a biological pathway, disease state, absence or presence of a disease state, to name a few non-limiting examples. The identification of different classes of cells within the sample allows for taking an action with respect to the sample or with respect to a source of

the sample. For example, depending on a distribution of cell types within a biological sample that is a tumor biopsy obtained from a subject, a specific treatment can be selected and administered to the subject.

[00271] In the embodiments of the present disclosure, a sample can include multiple second cells, and the described techniques allow for the determination of feature (*e.g.*, mRNA sequences that map to a particular gene in a plurality of genes and/or ATAC peaks) having values (*e.g.*, certain discrete attribute values or ATAC fragment counts) within each cell in the sample. For example, in some embodiments, in the case of genes, each discrete attribute value is a count of transcript reads within a single cell that map to a respective gene in the plurality of genes. As another example, in some embodiments, in the case of ATAC peaks, each ATAC fragment count is a count of ATAC fragments within a single cell that map to a respective ATAC peak in the plurality of ATAC peaks. In the described embodiments, as discussed above, such counts can be in the form of a count of UMIs. In this way, gene expression or ATAC fragment count per cell is determined and a discrete attribute value dataset is generated that includes discrete attribute values for each gene in each cell as well as ATAC fragment counts for each ATAC peak in each cell.

[00272] Furthermore, the techniques in accordance with the described embodiments allow clustering and otherwise analyzing the discrete attribute value dataset so as to identify patterns within the dataset and thereby assign each cell to a type or class. As used in this context here, a class refers to a different cell type, a different disease state, a different tissue type, a different organ type, a different species, or different assay conditions or any other feature or factor that allows for the differentiation of cells (or groups of cells) from one another. The discrete attribute value dataset includes any suitable number of cell classes of any suitable type. Moreover, as discussed above, the described techniques (including barcoding and computational analysis and visualization) provide the basis for identifying relationships between cellular phenotype and overall phenotypic state of an organism that is the source of the biological sample from which the sample was obtained that would not otherwise be discernable.

[00273] *Transposase accessible chromatin (ATAC) pipeline.*

[00274] An assay for transposase accessible chromatin (ATAC) has been described above and further details for such assays is provided in United States Patent Application No.

16/789,287, entitled “Methods for Processing Nucleic Acid Molecules,” filed February 12, 2020, which is hereby incorporated by reference. In such embodiments, a count 125 of each ATAC peak 123 in a plurality of ATAC peaks is acquired. In some embodiments, ATAC with high-throughput sequencing (ATAC-seq) maps chromatin accessibility by probing for DNA accessibility using hyperactive Tn5 transposase. Tn5 transposase inserts sequencing adapters into accessible regions of chromatin. See Buenrostro *et al.*, 2015, “ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide,” *Curr Protoc Mol Biol.* 109:21.29.1-9. In the systems and methods of the present disclosure, however, single cell ATAC techniques are provided that allow for the acquisition of ATAC data from each cell in a sample. The chromatin profiling of numerous cells (*e.g.*, tens of thousands of single cells) in parallel allows workers to see how chromatin compaction and DNA-binding proteins regulate gene expression at high resolution.

[00275] For the ATAC portion of the discrete attribute value dataset 120, in typical embodiments, there are no measures of gene expression. The primary entity of rows in the matrix for ATAC data within the discrete attribute value dataset 120 is fragments (or UMIs) 125 per called ATAC peak 123, where the peak 123 corresponds to a genomic region of accessible chromatin. Referring to Figure 1B, the ATAC matrix for discrete attribute value dataset 120-1 comprises ATAC peak elements 123-1-1 through 123-L-1, together with their measured ATAC fragment counts 125-1-1 through 125-L-1. For ATAC data, fragment counts 125 in peaks 123 instead of gene expression is recorded. In some embodiments, ATAC matrices will also contain aggregate rows as described below. As such, the ATAC pipeline generates a feature-barcode matrix, identifies clusters 159 based on the feature-barcode matrix, identifies significant features 165 (using, *e.g.*, a principal component analysis or other forms of data reduction such as those disclosed herein), and computes a 2-D t-SNE projection 198 or 2-D UMAP 199 projection based on the significant elements of the matrix. Because the ATAC pipeline performs a type of analysis that is different from the GEX pipeline discussed above, different types of graphical user interface components are needed to display the results of the ATAC pipeline, as discussed below.

[00276] In some embodiments, an ATAC pipeline identifies distinct peak regions, which can be regions encompassing from hundreds to thousands of nucleotides. In these regions, fragments derived from open chromatin sites can be detected. A feature metadata module (a label class)

can be used to associate ATAC peaks 123 with nearby promoter genes, to encode user-specified tags, such as, *e.g.*, a gene target of CRISPR guide RNA, whether an antibody is a positive or negative control, a source reference genome of a gene feature in the feature data module (matrix), *etc.*

[00277] In some embodiments, a visualization system is provided that is configured to present a graphical user interface on a display of a computing device. In some embodiments, a graphical user interface, also referred to herein as browser module 119, is configured to receive a discrete attribute value dataset 120 that includes GEX data as well as ATAC data. In some embodiments, prior to being accessed by the browser module 119, the data for the discrete attribute value dataset 120 is transformed into a format suitable for representation of the data on the graphical user interface. The data is presented on the graphical user interface in the format that allows manipulating the data based on user input or in other manner. One or more patterns within the data can be detected by the computing hardware executing the browser module 119. The graphical user interface is configured to display the discrete attribute value dataset 120 and various patterns in the manner that allows for revealing previously unknown groups or patterns within the data (*e.g.*, within cells from a biological sample). Moreover, the graphical user interface is configured to receive instructions (*e.g.*, user input) in response to which new patterns can be defined within the data. In some embodiments, user-specific tags are not encoded into the discrete attribute value dataset 120. In alternative embodiments, user-specific tags are encoded into the discrete attribute value dataset 120. In some embodiments, rather than making use of user-specific tags, nearby promoters, nearby genes, peak region function, transcription factor motifs, and unique transcript IDs are encoded in the discrete attribute value dataset 120. As one nonlimiting example, Figure 1C, discussed above, describes how ATAC nearby gene inferred count 194 and promoter sum inferred count 195 is stored for each gene 193 in the plurality of genes represented by the discrete attribute value dataset 120. In some embodiments, user-specific tags are encoded in the discrete attribute value dataset 120.

[00278] As discussed above, in some embodiments, a browser module 119 is executed in conjunction with display 108 of a visualization system 100. Figure 4A illustrates an example of browser module 119. In some embodiments, for representation and manipulation of ATAC data, the user interface includes certain features specific to the ATAC peak data. In some embodiments, the discrete attribute value dataset 120 includes an identifier indicating whether

the dataset 120 includes gene expression data (*e.g.* in the form of discrete attribute value dataset 120), ATAC peaks, or both. For example, in Figure 3, icon 320 to the left of the “atac_gex_intron” discrete attribute value dataset 120 indicates that this dataset includes both GEX and ATAC data. Various other icons in Figure 3 illustrate other forms of discrete attribute value datasets 120. It should be appreciated that the representations in Figure 3 are shown by way of example only, as any other indicators can be used to differentiate between discrete attribute value datasets 120, including gene expression data and data sets including both gene expression and ATAC data.

[00279] In some embodiments, the ATAC pipeline and the GEX pipeline are configured to generate clusters and/or t-SNE projections and/or UMAP projections based on subsets of features in the discrete attribute dataset 120. For example, in one example embodiment, when the discrete attribute value dataset 120 comprises antibody data, discrete attribute values 124 related to genes 122 in the discrete attribute value dataset 120 are used to determine clusters 158 and create a GEX t-SNE projection 196 or GEX UMAP projection 197, and ATAC fragment counts 125 related to ATAC peaks are used to construct an ATAC t-SNE projection 198 or ATAC UMAP projection 199. As shown in Figure 4B, in some embodiments, a projection selector affordance 460 receives input indicating selection of a GEX t-SNE projection 196, GEX UMAP projection 197, ATAC t-SNE projection 198, or ATAC UMAP projection 199 within the discrete attribute value dataset 120. For instance, to generate the ATAC t-SNE projection 198, the plurality of ATAC peaks 123 is dimension reduced using, for example, principal component analysis, across the cells of the discrete attribute dataset 120 to form reduction components (*e.g.*, principal components) 165 and these reduction components are further reduced by t-SNE to generate the two dimensional plot of Figure 18 in which cells with similar ATAC reduction components (*e.g.*, principal components) 165 are clustered together into ATAC clusters 159 and cells with dissimilar reduction components (*e.g.*, principal components) 165 lie further apart from each other in panel 420. Moreover, the plotted cells are color coded by which GEX cluster 159 they are in. In this way, Figure 18 illustrates an ATAC t-SNE plot, but with the cells clustered (*e.g.*, colored) based on GEX clustering rather than ATAC clustering.

[00280] As discussed above, in some embodiments the GEX t-SNE projection 196, GEX UMAP projection 197, ATAC t-SNE projection 198, and/or ATAC UMAP projection 199 for the discrete attribute value dataset 120 are computed in upstream processing rather than by

browser module 119. In some embodiments, to generate additional projections, a configuration file is specified that identifies the barcode sequence corresponding to a feature barcode, its type, as well as additional metadata. In some embodiments, such feature barcodes are separate and apart from the barcodes for the GEX data. An example of a line from an antibody configuration file, which represents one form of feature, is as follows:

[00281] id, name, read, pattern, sequence, feature_type

[00282] CD3, CD3_UCHT1_TotalC, R2, ^NNNNNNNNNN(BC)NNNNNNNNNN,
CTCATTGTAACCT, Antibody Capture

[00283] The first two columns are the feature ID (“id”) and display name “name” that will propagate to the browser module 119, the read and pattern columns indicate to the pipeline how to extract the barcode sequence from the raw sequencing read, the sequence column specifies the identifying sequence per feature, and the “feature_type” column specifies the feature type. For such a configuration file, all features of type “Antibody Capture” are combined to create an additional projection. If the configuration file had multiple feature types, the pipeline can generate a projection for each distinct feature type. The additional projections can be stored in the same format as the t-SNE projection from versions of browser module 119 discussed above.

[00284] Each ATAC peak 123 corresponds to an open area of the genome at a particular location. There can be multiple distinct ATAC peak 123 locations near a particular gene, or a particular gene transcription start site. Both ATAC peak types can be marked as being “nearby” that gene. Thus, in some embodiments the “nearby gene inferred count 194” is a representation of open sites of the genome per cell near that particular gene. Figure 20 illustrates a representation (a log₂ fold change) of ATAC peaks around the GZMB promoter that were combined, on an individual cell by cell basis, to form an aggregate feature (“GZMB Promoter Sum”) for each cell in the plurality of cells of the discrete attribute value dataset 120. In this way, the combined number of ATAC fragments near each gene per cell is superimposed, in panel 420, on any supported projection of the plurality of cells of the discrete attribute value dataset 120. In Figure 20, the projection that is displayed is an ATAC t-SNE projection. Thus, in Figure 20, the individual cells are arranged in panel 420 by an ATAC t-SNE projection but are colored based on their individual GZMB Promoter Sum.

[00285] A discrete attribute value dataset can include features of multiple types, and differential expression/chromatin accessibility can be segmented by feature type. Accordingly, in some embodiments, the browser module 119 is configured to receive input indicating selection of a type of a feature with respect to which a differential expression/accessibility is desired to be performed.

[00286] In some embodiments, default options for a discrete attribute value dataset 120 are “peak,” “motif,” and either a “promoter” or “nearby gene” options.

[00287] Since multiple feature types are supported by the browser module 119 in the described embodiments, in some in embodiments, in any user interface element where a user is entering a feature name to be autocompleted, the feature type is displayed in the autocompletion entry. This allows a user to distinguish, *e.g.*, between a CD4 gene and a CD4 protein.

[00288] In some embodiments, as discussed above, differential discrete attribute values for genes in each cluster 158 can be computed. Moreover, in some embodiments, as discussed above, differential ATAC fragment counts for ATAC fragments in each cluster 159 can be computed. For example, once each cell has been assigned to a respective cluster, the systems and methods of the present disclosure are able to compute, for each respective gene or ATAC peak, for each respective cluster, a difference in the discrete attribute value for the respective gene or the respective ATAC peak across the respective subset of cells in the respective cluster relative to the discrete attribute value or ATAC fragment count for the respective gene or ATAC peak across the plurality of clusters other than the respective cluster, thereby deriving a differential value for each respective gene or ATAC peak in the genes or ATAC peaks for each cluster. In this way, for example, the top most abundant genes or ATAC peaks that are different between cell classes or other forms of cell labels can be determined. This allows determining which genes or ATAC peaks discriminate between GEX or ATAC classes, GEX or ATAC categories, and/or GEX or ATAC clusters.

[00289] *ATAC peak viewer.*

[00290] In some embodiments of the present disclosure, browser module 119 is configured to display ATAC data in an intuitive manner. In some embodiments browser module 119 includes a set of tools that support representation and analysis of ATAC data and such tools are referred to herein as an ATAC Peak Viewer.

[00291] An ATAC peak 123 in accordance with the present disclosure is a representation of a region of open, accessible chromatin determined by analyzing fragments read between transposase cut sites. Unlike gene expression, where a gene is expected at the same location in a genome as per a reference genome, the location of ATAC peaks 123 can vary from run to run as different cells can have different regions of open chromatin accessible to regulatory elements. Accordingly, in the embodiments described herein, each ATAC peak 123 has a name associated therewith, with the name corresponding to the peak's genomic location (e.g., "chr2:237835-238281," in one example). The name of each ATAC peak 123 is stored in the feature-barcode matrix 121 of a discrete attribute dataset 120 similar to the manner in which gene names 122 are stored.

[00292] The described techniques provide computational visualization of ATAC peaks 123 in a manner that allows identifying each peak's genomic context, including nearby genes, exons, and untranslated regions (UTRs). In some embodiments, gene, exon and UTR information is stored in the discrete attribute value dataset in a manner similar to gene annotations. The gene, exon, and UTR transcripts can be derived from genomic reference information. In some embodiments, the discrete attribute value dataset 120 includes contig indexes, start positions, end positions, gene names, gene IDs, and strandedness. The discrete attribute value dataset also supports annotation of functional regions in a gene and includes a transcript annotation module. In one embodiment, the transcript annotation module includes the following data structures:

[00293] • array (string) of transcript IDs of length T (distinct transcript IDs for all transcript annotations);

[00294] • array (uint) of transcript object types of length $U+E=A$, the total number of annotations, which is made of U UTR annotations and E exon annotations (this indicates whether the information in all subsequent arrays pertains to an exon or a UTR);

[00295] • array (int) of indices into the gene ID array, of length A (the gene ID of an annotation);

[00296] • array (int) of indices into the transcript ID array, of length A (the transcript ID of an annotation);

[00297] • contig index array (int) of length A;

[00298] • start position array (int) of length A;

[00299] • end position array (int) of length A;

[00300] • strand array (int) of length A.

[00301] Following these notations, a single transcript annotation is formed from entries at position i for all arrays of length A . In this manner, a transcript annotation is formed from a gene ID, transcript ID, contig index, start position, end position, and strand. Moreover, the locations of multiple exons and UTRs per transcript and multiple transcripts per gene are encoded. Information on one, two, or more than two transcripts per gene can be stored in this way.

[00302] In some embodiments, the gene and transcript annotation information is sorted by respective genomic positions. In some embodiments, in response to a query, such as, *e.g.*, a range query indicating an interval of interest, a server executing the browser module 119 can perform a binary search by position to find elements within the interval with performance $O(\log A)$. The performance is improved due to the fact that the annotations are grouped by chromosome/contig, such that the search for gene annotations can start within a smaller subset of the data.

[00303] ATAC peak locations are also sorted and stored in memory in a similar manner. The ATAC peaks are representations of open chromatin regions that can be determined by identifying an overlap between fragments exposed by two transposase cut sites. Furthermore, analyzing the fragments themselves can reveal additional information and structure. Accordingly, in some embodiments, the ATAC pipeline is configured to generate a fragments module (stored separately from the discrete attribute dataset) that includes the following columns: a cell barcode column storing a barcode of a cell on which the fragment was found, and a contig, start and end position columns storing the contig, start, and end positions of the fragment, respectively. The fragments module can be in the form of a tab delimited data structure, which can be optionally compressed (*e.g.* gzipped). In some embodiments, the ATAC pipeline is configured to compute a tabix index over the compressed (*e.g.*, gzipped) tab-delimited information stored in the fragments module. A tabix index module stores the computed tabix index that allows for faster lookup of the on-disk chunks of fragments within a particular interval. The tabix index can be computed as described, for example, in Li, 2011, "Tabix: fast retrieval of sequence features from generic TAB-delimited files," *Bioinformatics*, 27(5): 718-

719. Other techniques can be used additionally or alternatively. In some embodiments, the tabix file module is embedded into the attribute value dataset 120, whereas the fragments module, which may be hundreds of megabytes to gigabytes in size, is stored separately from the discrete attribute value dataset.

[00304] Referring to Figure 10, visual representation of ATAC peaks and associated information is provided in the a user interface rendered by the browser 119 and in some embodiments includes an affordance 1004 that is configured to receive input indicating a selection of an option to display a peak viewer panel 1002. The peak viewer panel can be configured to display a relative distribution of ATAC peaks and occurrence of fragments between different clusters in a currently-selected category. As shown in Figure 10, query box 102 is responsive to an input a gene or a genomic region (peak). For example, as shown in Figure 10, when the user input includes “CD69,” ATAC peaks per each GEX cluster 159 (13 clusters, in this example) near the CD69A gene are displayed. As another example (Figure 13A), when the user input includes, *e.g.*, “chr:12:9760094-9761921,” all ATAC peaks and genes, and their respective quantities in cells on a cluster by cluster basis, within a base pair region (*e.g.*, 1 megabase) of the selected (anchoring) peak are displayed.

[00305] As shown in Figure 11 illustrating an example of a visual representation of ATAC peaks 123 identified near the CD69 gene in cells from a certain sample, each peak 123 can be represented as a corresponding rectangle that spans a certain genomic region within which the peak was detected. Each “active” (selected) cluster has a track (row) in the visualization. In this example, the height of the peak rectangle within each track is proportional to the percentage of cells within that cluster that have a fragment site within a particular peak for which the peak count for that cell at that peak in the feature dataset (*e.g.*, a peak-barcode matrix) is nonzero. In this example, peak heights are normalized to the highest frequency among all clusters. However, it should be appreciated that the normalization is optional. If an input device (*e.g.*, a computer mouse or another device, including user’s finger if the display is a touch screen) is detected over a representation of a peak, information related to the peak, such as, *e.g.*, the peak’s position, frequency within the cluster, and whether the peak region either encloses or is near a functional region of the gene, can be displayed. Such annotation is illustrated at element 2104 in Figure 21.

[00306] If a pointer of an input device (*e.g.*, a computer mouse) is disposed over a peak (or otherwise positioned with respect to the peak, depending on the specific implementation) in Figure 11, information on the peak's position, frequency within a cluster, and whether the peak region either encloses or is near a functional region of the gene can be displayed. In the example of Figure 11, referring again to the example of the CD69 gene, when the mouse is positioned over a rectangle corresponding to a peak, the processor instructs the user interface to display the peak's location, the potential function of the peak's region if any (*e.g.*, CD69 Promoter) and the relative frequency within that cluster (*e.g.*, a number between 0 percent and 100 percent).

[00307] As discussed above, embodiments of the present disclosure store and employ gene and transcript annotations. In some implementations of the browser module 119, gene annotations for a genomic region are displayed as lines 2102, *e.g.*, above the cluster peak tracks, as illustrated in Figure 21. In some embodiments, the gene can be represented by a solid line. However, it should be appreciated that various other representations of a gene can be generated additionally or alternatively. For example, in some implementations, a gene boundary may be dashed, and a recorded transcript may be in the form of a solid line. Any other types of representations can be used, including, *e.g.*, vertical lines that indicate gene boundaries. In some embodiments, the exons of the gene have thicker line widths, or are otherwise marked, relative to the introns of the gene. In some embodiments, exonic regions and untranslated regions (UTRs) can be represented in the user interface in any suitable manner. For example, in one implementation, exonic regions of the gene annotation are shown as solid rectangles (*e.g.*, black or of any other color) adjacent to (*e.g.*, atop) the gene annotation, and UTRs are in the form of unfilled rectangles. Visual representation(s) of one, two, or more than two transcripts per gene in a reference genome can be generated. In some implementations, a strand of a gene is shown by a backward (<) or forward (>) arrow next to the gene annotation. In response to detecting that an input device is positioned such that it hovers over the gene name, gene position, gene ID, and transcript ID (as found in the reference genome) may be displayed in some implementations.

[00308] The representation of ATAC peaks and associated information in accordance with embodiments of the present disclosure can be modified in various ways in response to user input. For example, user input can be received indicating an instruction to zoom out, zoom in, pan left and pan right using collective affordances 2106.

[00309] In some embodiments, the default representation of ATAC peaks displays the peaks, *e.g.*, as shown in Figure 21. Furthermore, the browser module 119 is configured to display representation of the overlap of open fragments. This information can be displayed in response to input specifying a location of a fragments module (or file) storing the fragment information associated with the ATAC dataset. For example, the fragment information can be acquired (*e.g.*, from a fragments dataset in response to a user input (*e.g.*, with response to a “folder” icon) including an instruction to load a “.tsv.gz” file. In some embodiments, a user input can indicate a URL indicating a location of the fragments module file. In some embodiments, a back-end server performs a “sanity check” to verify that the fragments module is correctly associated with the stored fragments tabix index. If the proper association is verified, the fragment information, together with peak information, can be employed for every range and gene query. On the back-end, the server interrogates the tabix index stored in the discrete attribute dataset 120, which will return the byte offset of the chunks within the block-gzipped fragments file containing fragments within the visible region. Since the beginning and ending chunks may also contain fragment locations outside the visible window, the server may apply overlap filtering to ensure that each fragment is within the visible window.

[00310] In some embodiments, when fragments information is loaded into memory for use by processing core(s) executing the browser module 119, a user interface is instructed to present respective fragments information along with a representation of the peaks. For example, as shown in Figure 23, the user interface can present fragments information in the form of grey (in this example) rectangles “behind” the peaks. This information is generated as rolling window sums of fragment cut sites within a certain distance (*e.g.*, default may be 200 bp or a window of another size) from that spot in the genome. The y-value of each track is the total number of fragments found in cells among that cluster with a start or end position within n base pairs of the underlying genomic location on the x-axis, where n can be user-specified. In the example of the user interface in Figure 23, heights of the grey rectangles representing fragments are normalized to the largest number of cut sites found at any point in any one cluster. In some embodiments, these may be further normalized by cell count per cluster. If a pointer of an input device is detected over the trace, the number of fragment cut sites (either the start or end of the fragment) within a certain distance of that point is displayed. This indicates accessibility of the chromatin

of that point. The representation of the fragment peaks in relation to functional elements in the gene allows assessing biological context of those peaks.

[00311] The ATAC peaks and associated information can be displayed in an interactive manner, such that one or more elements in the user interface can be modified automatically or in response to user input. For example, in some embodiments, user input can be received instructing the server to modify the representation of the peaks. For example, a selection of a “disable auto height” option will result in a change in the allocation for each cluster from a percentage of the peak viewer height to a fixed number of pixels. In this view, the gene annotation track drops from above the clusters to an overlay atop it. As shown in Figure 23, in response to user input (which can be, for example, moving the scroll bar up and down), cluster representation can be displayed underneath the gene annotation. A user can instruct the user interface to return to the auto height representation by selecting the “enable auto height” option. Furthermore, when peak distribution and fragment cut site location data is loaded, the user interface provides an option to “hide” fragment traces or peaks with the “show/hide peak distribution” and “show/hide cut sites” options displayed on the user interface. Similarly, gene annotations can be displayed or removed from the display.

[00312] Furthermore, as mentioned above, values of the fragment traces are functions of the windows around each position in the genome. The described techniques allow a user to vary the “openness” metric, based on the assumption that chromatin may be less accessible downstream from a particular cut site. Thus, the user interface can allow selecting the width of the window used to mark locations as being accessible by choosing, e.g., a value of a “smoothing window size” option. Smaller windows may reveal further nuances in the distribution of cut sites.

[00313] In some embodiments, underlying data and/or the graphical representation of the peak information can be stored in a suitable format. For example, in response to an indication that an “export” option is selected, the server hardware can instruct the browser module 119 to store the information. In this way, the frequency of each currently visible peak within each cluster can be stored, e.g., as a CSV file that is exported as a .bedgraph file representing the windowed cut site sums for each cluster. The graphical representation of the peaks can be stored in PNG, SVG, or any other suitable format. It should be noted that the embodiments of the

present disclosure are not limited with respect to a format in which the peak data and peak's graphical representation can be stored.

[00314] Another aspect of the present disclosure provides a method for characterizing cells, comprising partitioning a plurality of cells or cell nuclei and a plurality of barcode beads into a plurality of partitions, where at least a subset of the plurality of partitions each comprise a cell or cell nucleus of the plurality of cells or cell nuclei and a barcode bead of said plurality of barcode beads, and each bead in the subset of said plurality of partitions comprises a unique barcode sequence. The method further includes generating a plurality of barcoded nucleic acid molecules comprising barcode sequences, where a first subset of the plurality of barcoded nucleic acid molecules comprises a sequence corresponding to a ribonucleic (RNA) molecule and a second subset of said plurality of barcoded nucleic acid molecules comprises a sequence corresponding to a sequence corresponding to a region of accessible chromatin. The method further comprises sequencing the plurality of barcoded nucleic acid molecules, or derivatives generated therefrom, to generate sequencing information, and using the barcode sequence and the sequencing information to identify cell types in the sequencing information.

[00315] In some embodiments, the methods disclosed herein further comprise using said sequencing information to cluster cells by regions of accessible chromatin. In some embodiments, the methods further comprise using the sequencing information to cluster cells by gene expression. In some embodiments, the methods further comprise using the sequencing information and the cells clustered by gene expression to annotate, identify, or characterize cells clustered by regions of accessible chromatin. In some embodiments, the methods further comprise using the sequencing information and the cells clustered by regions of accessible chromatin to annotate, identify, or characterize cells clustered by gene expression. In some embodiments, the plurality of cells or cell nuclei are derived from a tumor sample or a sample suspected of comprising a tumor. In some embodiments, the methods further comprise using the sequencing information to identify a cell type, a cell state, a tumor-specific gene expression pattern, or a tumor-specific differentially accessible regions of chromatin in the tumor sample or the sample suspected of comprising said tumor. In some embodiments, the methods further comprise using the sequencing information to identify or confirm the presence of a tumor cell in the tumor sample or the sample suspected of comprising said tumor. In some embodiments, the methods further comprise administering a therapeutically effective amount of an agent targeting

one or more targets identified in the tumor-specific gene expression pattern or the tumor-specific differentially accessible regions of chromatin. In some embodiments, the tumor is a B cell lymphoma.

[00316] *EXAMPLES.*

[00317] Example 1. Joint ATAC and Gene Expression Profiling of Peripheral Blood Mononuclear Cells

[00318] As described herein, in some embodiments, complex biological systems can be better understood by profiling a plurality of different modalities within single cells. For example, in some instances, each cell in a plurality of cells can be assayed using gene expression analysis to obtain patterns in discrete attribute values for genes and/or differential values thereof (*e.g.*, fold change and/or relative gene expression), as well as using open chromatin analysis (*e.g.*, ATAC fragment count values) to obtain information on regions of chromatin accessibility and/or associated cell populations. Subsequently, analysis of these individual modalities (*e.g.*, clustering based on discrete attribute values and/or clustering based on ATAC fragment count values) for the plurality of cells can be combined to obtain indications of membership in each of the respective modalities for each respective cell in the plurality of cells (*e.g.*, membership in one or more of gene expression clusters and/or open chromatin clusters). Linkages obtained using assignment of single cells to multiple modality classes (*e.g.*, a single population of cells, where each respective cell is analyzed for both gene expression and open chromatin profiling) are likely to be more robust than, for example, single modality analysis on each of separate sub-populations of cells (*e.g.*, a first subset of cells analyzed for gene expression and a second subset of cells analyzed for open chromatin). For example, in some instances, comparisons drawn from a plurality of separate analyses of a respective plurality of sub-populations provide inferred linkages, in contrast to true linkages that can be obtained from a plurality of analyses performed on a single cell. Accordingly, a joint open chromatin (ATAC) and gene expression (GEX) profiling assay was performed for 24,000 peripheral blood mononuclear cells (PBMCs), in accordance with some embodiments of the present disclosure.

[00319] Each cell in the plurality of PBMCs was analyzed and clustered by gene expression (GEX) profiling using expressed markers (Figure 24) and open chromatin analysis using an assay of transposase-accessible chromatin (ATAC) (Figure 25). For each modality, the plurality

of cells was visualized as a dimension reduced projection using t-distributed stochastic neighbor embedding (t-SNE) (plots 2400 and 2500). For the gene expression analysis, each cell in the plurality of cells was assigned to a respective cluster group (*e.g.*, annotated) from 13 gene expression clusters (*e.g.*, 2402, 2404, 2406, 2408, 2410, 2412, 2414, 2416, 2418, 2420, 2422, 2424, and/or 2426). For the open chromatin analysis, each cell in the plurality of cells was assigned to a respective cluster group (*e.g.*, annotated) from 4 chromatin accessibility clusters (*e.g.*, 2502, 2504, 2506, and/or 2508). Concordance observed between the two modalities (*e.g.*, gene expression and open chromatin) is illustrated in Figures 26A and 26B. Figure 26A shows concordance between gene expression annotations (RNA (2400)) and open chromatin annotations (ATAC (2500)) for each respective cell in the plurality of PBMCs used in the joint ATAC and GEX profiling assay. Figure 26B maps the relationships between gene expression clusters and open chromatin clusters as indicated by the overlap of cells assigned to each cluster type (*e.g.*, co-annotations of PBMCs shared between cell type annotations; indicated by gray). Clusters identified by gene expression profiling (RNA) are indicated to the left and cell type annotations identified by open chromatin profiling (ATAC) are indicated to the right of Figure 26B.

[00320] Gene expression marker-derived annotations were transferred onto cell populations clustered by accessible chromatin analysis, as illustrated in Figure 27. The top panel of Figure 27 illustrates cells that were clustered and annotated with ATAC data using transcription factor accessibility (*e.g.*, a two-dimensional projection of the plurality of cells based on assignment of the plurality of cells to open chromatin cluster groups), while the bottom panel shows these same clusters differentially annotated with the assignment of each cell to a respective gene expression cluster, obtained from gene expression clustering analysis (*e.g.*, an indication of membership, for each cell in the plurality of cells, in a respective gene expression cluster group). Whereas single modality annotation of cell clusters (*e.g.*, ATAC clustering only) results with limited specificity, Figure 27 shows that additional annotation of such cell clusters with a second modality (*e.g.*, gene expression markers and/or gene expression clustering annotation) can be used to differentiate additional, more precise cell types and provide further context around particular cell types.

[00321] Additional gene expression analysis was performed to identify novel or stratified populations of cells otherwise undetected when using analysis of either gene expression or

regions of open chromatin alone. For example, as illustrated in Figure 28A, one or more populations of cells identified using open chromatin analysis alone may be visualized as large clusters or groups of clusters of cells. In particular, as illustrated in Figure 28A, B cells are visualized as a plurality of large cluster groups comprising multiple smaller populations of cells. (2508). Similarly, populations of cells identified using gene expression markers alone may include mixed populations of cell subtypes, as illustrated in Figure 28B. In particular, naïve and memory B cells are grouped together in a single cluster (2422). Whereas Figure 28C shows that the combination of different modalities (*e.g.*, open chromatin clustering and annotation using gene expression analysis) allows the stratification of the large B cell cluster (2508) into three separate sub-populations (*e.g.*, naïve/memory B cells (2722), IgM+IgD⁺ memory B cells (2724), and plasma B cells (2726)), it further indicates the presence of two possible sub-populations of cells in the naïve/memory B cell ATAC/GEX-annotated cluster.

[00322] Figures 29A and 29B further show a comparison of the naïve/memory B cell cluster using two different methods of analysis. Whereas GEX-clustering and annotation alone indicate only a single population of naïve/memory B cells (Figure 29B; 2422), ATAC-clustering coupled with GEX-annotation indicated two sub-populations (Figure 29A; 2722). Further analysis of the differential gene expression in the gene expression-annotated open chromatin clustering plot revealed that the two distinct sub-populations of cells (subcluster 1 (2902) and subcluster 2 (2904)) that were visible when viewing open chromatin clustering (Figure 29C) represented prospective memory B cells and naïve B cells, respectively. These sub-populations were indistinguishable when viewing gene expression clustering alone, with memory B cells and naïve B cells clustering together (Figure 29D). Gene expression analysis of the memory B cells in subcluster 1 and the naïve B cells in subcluster 2 identified differential gene expression of representative markers, where prospective memory B cells expressed relatively higher Ig, relatively lower naïve B cell associated transcripts and prospective naïve B cells expressed relatively lower Ig, relatively higher naïve B cell associated transcripts, as illustrated in Figure 29E. Together, the data indicate that additional annotation of open chromatin (ATAC) clusters with gene expression markers can provide additional context around particular cell types and subtypes that would be lacking not only when using only ATAC clustering and annotation, but also when using only gene expression clustering and annotation. Thus, in some embodiments, both gene expression analysis and open chromatin analysis are improved by the additional

annotation of a second modality, for example, by facilitating the identification of novel populations of cells that would otherwise be unidentified and/or unannotated when analyzing either gene expression or regions of open chromatin alone.

[00323] Example 2. Functional Characterization of a Small B Cell Lymphoma

[00324] A sample of an intra-abdominal lymph node tumor was obtained from a subject in order to perform a functional characterization of a small B cell lymphoma and the signaling pathways thereof. A pathology report for the subject indicated a diagnosis of diffuse small lymphocytic lymphoma of the lymph node (*e.g.*, malignant lymphoma, small B cell, diffuse type, IHC: CD20(+), CD3(-)). Approximately 9000 nuclei (*e.g.*, cells) from the tumor sample were profiled using gene expression analysis and single cell open chromatin (ATAC-seq). For each modality (*e.g.*, GEX analysis and ATAC-seq), the plurality of cells was visualized as a dimension reduced projection using t-distributed stochastic neighbor embedding (t-SNE). Gene expression clustering assigned tumor cells to one of five cell type categories (B cells, replicating B cells, monocytes, T cells, and replicating T cells), as illustrated in Figure 30A. Open chromatin clustering assigned tumor cells to one of three cell type categories (B cells, monocytes, and T cells), as illustrated in Figure 30B.

[00325] Using mutational load (SNVs) and gene expression markers for the *BANK1* pathway (markers of B cell hyper-activation), tumor B cells were distinguished from normal B cells. Cells expressing differential levels of SNVs consistent with tumor cells are indicated in black in Figure 31A, while cells expressing differential levels of *BANK1* pathway genes consistent with tumor cells are indicated in black in Figure 31B. Application of differential tumor and normal gene expression data allowed the identification of two sub-populations of B cells (tumor B cells and normal B cells) in cell populations clustered by gene expression profiling, as illustrated in Figure 31C.

[00326] The stratification of clusters is visualized in dimension reduced projections of gene expression clustering and ATAC-seq clustering of lymph node tumor cells in Figures 32A and 32B. In both instances, application of differential tumor and normal gene expression data allowed the identification of two sub-populations of B cells (tumor B cells and normal B cells) in cell populations clustered by gene expression profiling (Figure 32A) and open chromatin analysis (Figure 32B). Figure 32C illustrates differential gene expression for representative gene

markers compared between normal B cells (left panel) and tumor (right panel) B cells. *FCRL5* / *FCRL3* encode members of the immunoglobulin receptor superfamily and the Fc-receptor like family. These genes are implicated in B cell development and lymphomagenesis. *MIR155HG* represents a microRNA host gene. The long RNA transcribed from this gene is expressed at high levels in lymphoma and may function as an oncogene. *RASGRF1* is a guanine nucleotide exchange factor (GEF) and is involved in MAP-Erk pathway. *IL4R* is a receptor to a key inflammatory signaling factor, pro-growth and pro-metastatic. *XAF1* encodes a protein which binds to and counteracts the inhibitory effect of a member of the IAP (inhibitor of apoptosis) protein family. *BANK1* is a tumor suppressor in B cell lymphoma.

[00327] The covariance of open chromatin and gene expression was further used to identify a candidate enhancer region that regulates the expression of the *IL4R* specifically in the tumor B cells, as illustrated in Figures 33A-D. Signal transducer and activator of transcription (STAT) proteins are critical mediators of cytokine signaling. Among the seven STAT proteins, STAT6 is activated by IL-4 and IL-13 and plays a predominant role in the immune system. Gene expression characterization (Figure 33A; top panel) and accessible chromatin characterization (Figure 33A; bottom panel) of *Stat3* and *Stat6* was performed, using a dimension reduced visualization of gene expression clustering analysis (compare, for example, with Figure 33B). The data indicate that the *IL4R*-mediated STAT6 signaling pathway is activated in the lymph node tumor sample. Figure 33C further illustrates the activation of STAT proteins only in clusters of tumor B cells identified using a dimension reduced visualization of open chromatin analysis (ATAC) (compare, for example, with Figure 33D), indicating that STAT proteins accumulate in the cytoplasm of cells but are only active upon nuclear translocations.

[00328] Further comparative analysis of STAT protein activation across multiple cell populations is illustrated in Figures 34A-C. For example, enhancer regions identified by open chromatin clustering (Figure 34A) were assessed for the presence of epigenetic modifications, as illustrated in Figure 34B (gray shaded column). Quantification of *IL4R* expression, promoter peaks, and linked peaks revealed increased levels of each consistent in the tumor B cell population but not in the T cell, monocyte, or normal B cell populations (Figure 34C). Thus, the data indicate that the combination of gene expression annotations and open chromatin clustering can uncover tumor-specific differentially accessible chromatin regions that are not revealed

when analyzing gene expression data alone and can be used to discover novel open chromatin regions associated with specific cell states.

[00329] CONCLUSION

[00330] All references cited herein are incorporated herein by reference in their entirety and for all purposes to the same extent as if each individual publication or patent or patent application was specifically and individually indicated to be incorporated by reference in its entirety for all purposes.

[00331] Plural instances may be provided for components, operations or structures described herein as a single instance. Finally, boundaries between various components, operations, and data stores are somewhat arbitrary, and particular operations are illustrated in the context of specific illustrative configurations. Other allocations of functionality are envisioned and may fall within the scope of the implementation(s). In general, structures and functionality presented as separate components in the example configurations may be implemented as a combined structure or component. Similarly, structures and functionality presented as a single component may be implemented as separate components. These and other variations, modifications, additions, and improvements fall within the scope of the implementation(s).

[00332] It will also be understood that, although the terms first, second, etc. may be used herein to describe various elements, these elements should not be limited by these terms. These terms are only used to distinguish one element from another. For example, a first subject could be termed a second subject, and, similarly, a second subject could be termed a first subject, without departing from the scope of the present disclosure. The first subject and the second subject are both subjects, but they are not the same subject.

[00333] The terminology used in the present disclosure is for the purpose of describing particular embodiments only and is not intended to be limiting of the present disclosure. As used in the description of the present disclosure and the appended claims, the singular forms “a”, “an” and “the” are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will also be understood that the term “and/or” as used herein refers to and encompasses any and all possible combinations of one or more of the associated listed items. It will be further understood that the terms “comprises” and/or “comprising,” when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or

components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

[00334] As used herein, the term “if” may be construed to mean “when” or “upon” or “in response to determining” or “in response to detecting,” depending on the context. Similarly, the phrase “if it is determined” or “if [a stated condition or event] is detected” may be construed to mean “upon determining” or “in response to determining” or “upon detecting (the stated condition or event)” or “in response to detecting (the stated condition or event),” depending on the context.

[00335] The foregoing description included example systems, methods, techniques, instruction sequences, and computing machine program products that embody illustrative implementations. For purposes of explanation, numerous specific details were set forth in order to provide an understanding of various implementations of the inventive subject matter. It will be evident, however, to those skilled in the art that implementations of the inventive subject matter may be practiced without these specific details. In general, well-known instruction instances, protocols, structures and techniques have not been shown in detail.

[00336] The foregoing description, for purpose of explanation, has been described with reference to specific implementations. However, the illustrative discussions above are not intended to be exhaustive or to limit the implementations to the precise forms disclosed. Many modifications and variations are possible in view of the above teachings. The implementations were chosen and described in order to best explain the principles and their practical applications, to thereby enable others skilled in the art to best utilize the implementations and various implementations with various modifications as are suited to the particular use contemplated.

What is claimed:

1. A method for visualizing a pattern in a discrete attribute value dataset, the method comprising:
 - at a computer system comprising one or more processing cores and a memory:
 - storing the discrete attribute value dataset in the memory, wherein the discrete attribute value dataset comprises:
 - a respective discrete attribute value for each corresponding gene in a plurality of genes, for each respective cell in a plurality of cells, and
 - a respective ATAC fragment count for each corresponding ATAC peak in a plurality of ATAC peaks, for each respective cell in the plurality of cells, wherein
 - the plurality of cells comprises 100 cells,
 - the plurality of genes comprises 100 genes,
 - the plurality of ATAC peaks comprises 50 ATAC peaks,
 - each respective cell in the plurality of cells is assigned to a respective cluster group in a first plurality of cluster groups based on a first clustering of discrete attribute values for the plurality of genes across the plurality of cells, and
 - each respective cell in the plurality of cells is also assigned to a respective cluster group in a second plurality of cluster groups based on a second clustering of ATAC fragment count values for the plurality of ATAC peaks across the plurality of cells;
 - displaying, in a first panel, a two-dimensional projection of the plurality of cells based on assignment of the plurality of cells to one of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups; and
 - indicating within the two-dimensional projection, for each respective cell in the plurality of cells, membership in the other of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups, thereby visualizing the pattern in the discrete attribute value dataset.
2. The method of claim 1, wherein membership of each respective cell in the plurality of cells in the other of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups is indicated by coloring the respective cell a color that is uniquely associated with a cluster group to

which the respective cell has been assigned in the other of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups.

3. The method of claim 1, the method further comprising:

computing, for each respective gene in the plurality of genes for each respective cluster in the first plurality of clusters or the second plurality of clusters, a difference in the discrete attribute value for the respective gene across the respective subset of cells in the respective cluster relative to the discrete attribute value for the respective gene across the first plurality of clusters or the second plurality of clusters other than the respective cluster, thereby deriving a differential value for each respective gene in the plurality of gene for each respective cluster in the first plurality of clusters or the second plurality of clusters, and

displaying in a second panel, concurrently with the first panel, a heat map that comprises a representation of the differential value for each respective gene in the plurality of genes for each respective cluster in the first plurality of clusters or the second plurality of clusters thereby visualizing the pattern in the discrete attribute value dataset.

4. The method of claim 3, wherein the differential value for each respective gene in the plurality of genes for each respective cluster in the first plurality of clusters or the second plurality of clusters is a fold change in (i) a first measure of central tendency of the discrete attribute value for the respective gene measured in each of the cells in the plurality of cells in the respective cluster and (ii) a second measure of central tendency of the discrete attribute value for the respective gene measured in each of the cells of all clusters in the first plurality of clusters or all clusters in the second plurality of clusters, other than the first respective cluster.

5. The method of claim 4, wherein the fold change is a $\log_2$ fold change or a $\log_{10}$ fold change.

6. The method of any one of claims 3-5, wherein the method further comprises normalizing the discrete attribute value prior for each respective gene in the plurality of genes prior to computing the differential value for each respective gene in the plurality of genes for each respective cluster in the first plurality of clusters or the second plurality of clusters.

7. The method of claim 6, wherein the normalizing comprises modeling the discrete attribute value of each gene associated with each cell in the plurality of cells with a negative binomial distribution having a consensus estimate of dispersion.

8. The method of claim 1, wherein:

the first clustering of discrete attribute values for the plurality of genes across the plurality of cells is a clustering of a first plurality of dimension reduction values for each respective cell in the plurality of cells, across the plurality of cells, wherein each respective dimensional reduction value in the first plurality of dimension reduction values for each respective cell in the plurality of cells is derived, using a first dimension reduction algorithm, from the discrete attribute values of each gene in the respective cell, and

the second clustering of ATAC fragment counts for the plurality of ATAC peaks across the plurality of cells is a clustering of a second plurality of dimension reduction values for each respective cell in the plurality of cells, across the plurality of cells, wherein each respective dimensional reduction value in the second plurality of dimension reduction values for each respective cell in the plurality of cells is derived, using the first dimension reduction algorithm, from the ATAC fragment counts of each ATAC peak in the respective cell.

9. The method of claim 8, wherein the first dimension reduction algorithm is principal component analysis.

10. The method of claim 8 or 9, wherein

the two-dimensional projection of the plurality of cells is based on the assignment of the plurality of cells to the first plurality of cluster groups, and

the two-dimensional projection of the plurality of cells is obtained from t-distributed stochastic neighbor or UMAP embedding of the first plurality of dimension reduction values for each respective cell in the plurality of cells, across the plurality of cells.

11. The method of claim 8 or 9, wherein

the two-dimensional projection of the plurality of cells is based on the assignment of the plurality of cells to the second plurality of cluster groups, and

the two-dimensional projection of the plurality of cells is obtained from t-distributed stochastic neighbor or UMAP embedding of the second plurality of dimension reduction values for each respective cell in the plurality of cells, across the plurality of cells.

12. The method of any one of claims 1-11, wherein

the first clustering of discrete attribute values for the plurality of genes across the plurality of cells comprises application of a Louvain modularity algorithm, k-means clustering, a fuzzy k-means clustering algorithm, or Jarvis-Patrick clustering, and

the second clustering of ATAC fragment count values for the plurality of ATAC peaks across the plurality of cells comprises application of a Louvain modularity algorithm, k-means clustering, a fuzzy k-means clustering algorithm, or Jarvis-Patrick clustering.

13. The method of any one of claim 1-12, wherein

the first clustering of discrete attribute values for the plurality of genes across the plurality of cells comprises k-means clustering into a first predetermined number of clusters, or

the second clustering of ATAC fragment count values for the plurality of ATAC peaks across the plurality of cells comprises k-means clustering into a second predetermined number of clusters.

14. The method of claim 13, wherein the first or second predetermined number of clusters is an integer value between 2 and 50.

15. The method of claim 14, the method further comprising obtaining the integer value from a user of the computer system.

16. The method of any one of claims 1-15, wherein the respective discrete attribute value for each corresponding gene in a plurality of genes, for each respective cell in a plurality of cells represents a whole transcriptome shotgun sequencing experiment that quantifies gene expression from each respective single cell in the plurality of cells in counts of transcript reads mapped to the genes.

17. The method of claim 1, wherein each gene in a particular cell in the plurality of cells is uniquely represented within the discrete attribute value dataset with a first barcode that is unique to the particular cell.

18. The method of claim 1, wherein the discrete attribute value of each gene in a particular cell in the plurality of cells is determined after the particular cell has been separated from all the other cells in the plurality of cells into its own microfluidic partition.

19. The method of any one of claims 1-18, wherein the plurality of cells comprises 1000 cells, 2000 cells, 5000 cells, 10,000 cells, 25,000 cells, 50,000 cells or 100,000 cells.

20. The method of any one of claims 1-19, wherein the plurality of genes comprises 150 genes, 200 genes, 300 genes, 400 genes, 1000 genes, 2000 genes, 3000 genes, 4000 genes, or 5000 genes.

21. The method of any one of claims 1-20, wherein the plurality of ATAC peaks comprises 100 ATAC peaks, 200 ATAC peaks, 500 ATAC peaks, 750 ATAC peaks, 1000 ATAC peaks, or 5000 ATAC peaks.

22. The method of any one of claims 1-21, wherein the discrete attribute value dataset has a file size of at least 250 megabytes, 500 megabytes, one gigabyte, two gigabytes, or three gigabytes.

23. The method of any one of claims 1-22, wherein

the discrete attribute value dataset further comprises a feature-linkage matrix that stores, for each respective gene in the plurality of genes and for each respective ATAC peak in the plurality of ATAC peaks:

a collection of ATAC peaks and genes that are within a threshold distance of the respective gene or respective ATAC peak in a reference genome, and for each respective ATAC peak or respective gene in the collection:

a correlation in an ATAC fragment count of the respective ATAC peak or a correlation in the discrete attribute value of the respective gene to the first ATAC peak or first gene across the plurality of cells, and

the method further comprises:

receiving a selection of a first gene in the plurality of genes or a first ATAC peak in the plurality of ATAC peaks, and

using the feature-linkage matrix to obtain and provide a first plot comprising an indicator for each gene in the plurality of genes or each peak in the plurality of peaks linked to the first gene or the first ATAC peak, by order of distance apart from the first gene or the first ATAC peak in the reference genome.

24. The method of claim 23, wherein the respective indicator for each respective gene in the plurality of genes or each respective peak in the plurality of peaks linked to the first gene or the first ATAC peak:

is provided for each respective cluster group in the first plurality of cluster groups or each cluster group in the second plurality of cluster groups, and

is dimensioned to represent a proportion of cells in the respective cluster group that have a non-zero value for the discrete attribute value of the respective gene or a non-zero value for the ATAC fragment count of the respective ATAC peak.

25. The method of claim 23 or 24 wherein the feature-linkage matrix further stores, for each respective ATAC peak or respective gene in the collection:

a significance in an ATAC fragment count of the respective ATAC peak or a significance in the discrete attribute value of the respective gene to the first ATAC peak or first gene across cells in the plurality of cells, and the method further comprises:

limiting the first plot, to each gene in the plurality of genes or each peak in the plurality of peaks that have a threshold correlation or significance in an ATAC fragment count of the respective ATAC peak or discrete attribute value of the respective gene to the first ATAC fragment or first gene.

26. The method of any one of claims 23-25, wherein the first plot is limited to each gene in the plurality of genes or each peak in the plurality of peaks that is within a threshold distance of the first gene or first ATAC peak in a reference genome.

27. The method of claim 26, wherein the threshold distance is 1 megabase, 2 megabases, or a value between 0.5 megabases and 10 megabases.

28. A computer system comprising one or more processing cores and a memory, the memory storing instructions that use the one or more processing cores to perform a method comprising:
storing a discrete attribute value dataset in the memory, wherein the discrete attribute value dataset comprises:

- a respective discrete attribute value for each corresponding gene in a plurality of genes, for each respective cell in a plurality of cells, and
- a respective ATAC fragment count for each corresponding ATAC peak in a plurality of ATAC peaks, for each respective cell in the plurality of cells, wherein
 - the plurality of cells comprises 100 cells,
 - the plurality of genes comprises 100 genes,
 - the plurality of ATAC peaks comprises 50 ATAC peaks,
- each respective cell in the plurality of cells is assigned to a respective cluster group in a first plurality of cluster groups based on a first clustering of discrete attribute values for the plurality of genes across the plurality of cells, and
- each respective cell in the plurality of cells is also assigned to a respective cluster group in a second plurality of cluster groups based on a second clustering of ATAC fragment count values for the plurality of ATAC peaks across the plurality of cells;
- displaying, in a first panel, a two-dimensional projection of the plurality of cells based on assignment of the plurality of cells to one of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups; and
- indicating within the two-dimensional projection, for each respective cell in the plurality of cells, membership in the other of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups, thereby visualizing the pattern in the discrete attribute value

29. A non-transitory computer readable storage medium, wherein the non-transitory computer readable storage medium stores instructions, which when executed by a computer system, cause the computer system to perform a method for visualizing a pattern in a discrete attribute value dataset, the method comprising:

obtaining the discrete attribute value dataset, wherein the discrete attribute value dataset comprises:

a respective discrete attribute value for each corresponding gene in a plurality of genes, for each respective cell in a plurality of cells, and

a respective ATAC fragment count for each corresponding ATAC peak in a plurality of ATAC peaks, for each respective cell in the plurality of cells, wherein

the plurality of cells comprises 100 cells,

the plurality of genes comprises 100 genes,

the plurality of ATAC peaks comprises 50 ATAC peaks,

each respective cell in the plurality of cells is assigned to a respective cluster group in a first plurality of cluster groups based on a first clustering of discrete attribute values for the plurality of genes across the plurality of cells, and

each respective cell in the plurality of cells is also assigned to a respective cluster group in a second plurality of cluster groups based on a second clustering of ATAC fragment count values for the plurality of ATAC peaks across the plurality of cells;

displaying, in a first panel, a two-dimensional projection of the plurality of cells based on assignment of the plurality of cells to one of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups; and

indicating within the two-dimensional projection, for each respective cell in the plurality of cells, membership in the other of (i) the first plurality of cluster groups or (ii) the second plurality of cluster groups, thereby visualizing the pattern in the discrete attribute value.

30. A method for characterizing cells, comprising

(a) partitioning a plurality of cells or cell nuclei and a plurality of barcode beads into a plurality of partitions, wherein at least a subset of said plurality of partitions each comprise a cell or cell nucleus of said plurality of cells or cell nuclei and a barcode bead of said plurality of

barcode beads, wherein each bead in said subset of said plurality of partitions comprise a unique barcode sequence;

(b) generating a plurality of barcoded nucleic acid molecules comprising barcode sequences, wherein a first subset of said plurality of barcoded nucleic acid molecules comprises a sequence corresponding to a ribonucleic (RNA) molecule; and a second subset of said plurality of barcoded nucleic acid molecules comprises a sequence corresponding to a sequence corresponding to a region of accessible chromatin;

(c) sequencing said plurality of barcoded nucleic acid molecules or derivatives generated therefrom to generate sequencing information; and

(d) using said barcode sequence and said sequencing information to identify cell types in said sequencing information.

31. The method of claim 30, further comprising using said sequencing information to cluster cells by regions of accessible chromatin.

32. The method of claim 30 or claim 31, further comprising using said sequencing information to cluster cells by gene expression.

33. The method of claim 32, further comprising using said sequencing information and said cells clustered by gene expression to annotate, identify, or characterize cells clustered by regions of accessible chromatin.

34. The method of any one of claims 31-33, further comprising using said sequencing information and said cells clustered by regions of accessible chromatin to annotate, identify, or characterize cells clustered by gene expression.

35. The method of any one of claims 30-34, where said plurality of cells or cell nuclei are derived from a tumor sample or a sample suspected of comprising a tumor.

36. The method of any one of claims 30-35, further comprising using said sequencing information to identify a cell type, a cell state, a tumor-specific gene expression pattern, or a

tumor-specific differentially accessible regions of chromatin in said tumor sample or said sample suspected of comprising said tumor.

37. The method of any one of claims 35-36, further comprising using said sequencing information to identify or confirm the presence of a tumor cell in said tumor sample or said sample suspected of comprising said tumor.

38. The method of any one of claims 36-37, further comprising administering a therapeutically effective amount of an agent targeting one or more targets identified in said tumor-specific gene expression pattern or said tumor-specific differentially accessible regions of chromatin.

39. The method of any one of claims 35-38, wherein said tumor is a B cell lymphoma.