Title

Single-Cell Multiomics

Permalink

https://escholarship.org/uc/item/23z5c4sc

Journal

Annual Review of Biomedical Data Science, 6(1)

ISSN

2574-3414

Authors

Flynn, Emily

Almonte-Loya, Ana

Fragiadakis, Gabriela K

Publication Date

2023-08-10

DOI

10.1146/annurev-biodatasci-020422-050645

Published in final edited form as:
Annu Rev Biomed Data Sci. 2023 August 10; 6: 313–337. doi:10.1146/annurev-biodatasci-020422-050645.

Single-Cell Multiomics

Emily Flynn[1,*], **Ana Almonte-Loya**[1,2,*], **Gabriela K. Fragiadakis**[1,3]

[1]CoLabs, University of California, San Francisco, California, USA

[2]Biomedical Informatics Program, University of California, San Francisco, California, USA

[3]Division of Rheumatology, Department of Medicine, University of California, San Francisco, California, USA

Abstract

Single-cell RNA sequencing methods have led to improved understanding of the heterogeneity and transcriptomic states present in complex biological systems. Recently, the development of novel single-cell technologies for assaying additional modalities, specifically genomic, epigenomic, proteomic, and spatial data, allows for unprecedented insight into cellular biology. While certain technologies collect multiple measurements from the same cells simultaneously, even when modalities are separately assayed in different cells, we can apply novel computational methods to integrate these data. The application of computational integration methods to multimodal paired and unpaired data results in rich information about the identities of the cells present and the interactions between different levels of biology, such as between genetic variation and transcription. In this review, we discuss both the single-cell technologies for measuring these modalities and describe and characterize a variety of computational integration methods for combining the resulting data to leverage multimodal information toward greater biological insight.

Keywords

multiomics; integration; single-cell; computation; next-generation sequencing; multimodal

1. INTRODUCTION

Recently improvements in molecular biology and microfluidics have aided in the development of single-cell isolation and barcoding technologies. These methods now enable DNA, mRNA, chromatin, and protein profiles to be measured at a single-cell resolution. This technology has significantly advanced our knowledge of biological systems and yielded transformative insights into cellular diversity and development (1). Single-cell measurements paired with the appropriate analytical tools can reveal differences in cell type composition and cell states across conditions, leading to discoveries that increase our biological understanding (2).

gabriela.fragiadakis@ucsf.edu .

[*] These authors contributed equally to this article

Single-cell technologies are particularly powerful in the study of systems with high cellular diversity, such as the immune system (3). The immune system is composed of a diverse array of cell types and states that maintain homeostasis and can detect and respond to threats such as infection and aberrant cell development. Using single-cell methods, we can broaden our understanding of the immune system's complexity, including the heterogeneity, development, differentiation, and microenvironments of cells in health and disease.

A variety of technologies have been developed to examine properties of cells at single-cell resolution, including information on the transcriptome, genome, epigenome, proteome, and spatial organization. While a single modality provides important insights, the combination of single-cell data across modalities produces both richer information about individual cells and insights into the interactions between different elements of the cell state. Multimodal analysis also provides complementary information because each modality has differences in dimensionality, sparsity, and sources of noise. Integrating multiple modalities can improve our ability to identify cell types, and more broadly offer new insights into how different elements of the biological system behave in concert to define cellular behavior.

In this review, we start with a presentation of multiomic technologies available and analytical workflows for omic analyses, and then we discuss tools for data integration of multimodal single-cell data. This is followed by a description of downstream analyses, as well as some of the challenges and next steps in this expanding field.

2. TECHNOLOGIES FOR GENERATING MULTIOMIC DATA

2.1. Transcriptomic Data

Single-cell sequencing methods have been broadly applied in both basic and translational research, and many multiomics technologies such as single-cell RNA sequencing (scRNA-seq) include transcriptomic measurements (Figure 1). In order to measure transcription at a single-cell level, each cell must be isolated from its originating tissue, which can be done using techniques including fluorescence-activated cell sorting (FACS), laser capture microdissection, and microfluidics. Droplet- or microwell-based methods, such as Drop-seq (4) and 10× Genomics Chromium (5), generate pools of full-length complementary DNA (cDNA), enabling unbiased analysis of thousands of cells following Illumina short-read sequencing. Technologies that leverage nanopore long-read sequencing can generate full-length sequences plus information about sequence diversity, splicing, and chimeric transcripts (6). New techniques such as split-pooling offer an alternative to cell isolation by using combinatorial indexing of single cells (7). Most of these methods are also tag based, adding unique molecular identifiers (UMIs) as barcodes at either the 3′ end or 5′ end of the transcript (8, 9), which can help reduce polymerase chain reaction (PCR) bias and artifacts during analysis. Following tagging, reverse transcriptase is used to obtain cDNA from RNA transcripts, which are then amplified and sequenced.

2.2. Genomic Data

Single-cell genome sequencing techniques help elucidate genetic heterogeneity by measuring genetic alterations at single-cell resolution. This permits the analysis of de novo germline mutations and somatic mutations in different cell populations. Pairing transcriptome and genome single-cell sequencing can help uncover mechanisms of gene regulation and genotype–phenotype associations.

G&T-seq (genome and transcriptome sequencing) (10) uses a biotinylated oligo-dT primer, with the goal of separating mRNA from genomic DNA (gDNA) within the same cell. After cell lysis, the genome and transcriptome are amplified and sequenced in parallel. Using this technology, Macaulay et al. (10) detected single-nucleotide variations in gDNA and mRNA from the same cell. Although a single-cell's genomic copy number can be inferred indirectly from scRNA-seq data (11), only by applying multi-omics approaches can this information be resolved unambiguously. A similar method, DR-seq (gDNA–mRNA sequencing) (12) also amplifies small quantities of gDNA and mRNA from single cells. Isolated single cells are lysed, barcoded, and mRNA is reverse transcribed into cDNA, and both cDNA and gDNA are amplified together then separated for further processing and additional amplification.

2.3. Epigenomic Data

Epigenomic data allow us to study the mechanisms that convert genome content into multiple functional and stable cellular conditions. Chromatin accessibility indicates the physical access to DNA, an essential regulatory mechanism for establishing and maintaining cellular identity (13). Single-nucleus ATAC (assay for transposase-accessible chromatin) sequencing (snATAC-seq) (14) and snRNA-seq (single-nucleus RNA sequencing) have been combined into SHARE-seq (simultaneous high-throughput ATAC and RNA expression with sequencing) to generate paired, cell-type-specific chromatin accessibility over thousands of cells (15). In this process, barcoded Tn5 transposases are first used to label cells in bulk, inserting sequencing adapters into accessible regions of the genome. After this, single nuclei are isolated (frequently using microfluidic devices), and cell-identifying barcodes are introduced. The Chromium Single Cell Multiome ATAC + Gene Expression assay from 10× genomics is one of the most popular protocols for this experiment. While snATAC-seq tells us which chromatin regions are accessible, other methods can identify the locations of specific chromatin proteins bound to DNA. The method scCUT&Tag (single-cell cleavage under targets and tagmentation) (16, 17) uses antibodies targeted to the protein of interest [e.g., RNA Polymerase II, specific transcription factors (TFs), or histone modifications]. The antibodies also tether to a Tn5 transposase–Protein A fusion protein with sequencing adapters, so when the transposase is activated, this results in targeted DNA fragments for sequencing. Paired-Tag (parallel analysis of individual cells for RNA expression and DNA from targeted tagmentation by sequencing) (18) is a multiomic method that adapts scCUT&Tag to assay both histone modifications and gene expression by performing reverse transcription after tagmentation.

Cytosine methylation is a crucial epigenetic layer that indicates the transcriptional potential of genomic DNA (19) and can be detected using bisulfite sequencing. Genomic DNA treated with bisulfite converts unmethylated cytosines into uracils, which, after PCR amplification, are converted to thymidines, allowing the methylation signal to be derived by comparing treated and untreated samples (20). scBS-seq (single-cell bisulfite sequencing) can map methylation locations (21); however, for most loci, the observed sequence reads originate from only one chromosomal copy, making it challenging to identify de novo regulatory elements.

To overcome this limitation, one can combine scBS-seq with scRNA-seq for scM&T-seq (single-cell methylome and transcriptome sequencing) (22). Similar to G&T-seq, single cells are isolated via flow cytometry, and DNA and RNA molecules are separated. RNA transcripts undergo bead capture and amplification while genomic DNA is further processed using bisulfite conversion. This method has been used to analyze the development and epigenetic heterogeneity of mouse embryonic stem cells (22). In addition, scNMT-seq (single-cell nucleosome, methylation, and transcription sequencing) (23) combines scM&T-seq with NOMe-seq (nucleosome occupancy and methylation sequencing) (24) to capture transcriptome, methylome, and chromatin accessibility at the single-cell level.

2.4. Proteomic Data

mRNA and protein levels are not necessarily correlated, in part because of posttranscriptional regulatory mechanisms (5). Many single-cell unimodal proteomic methods, such as high-dimensional flow cytometry (fluorophores) and cytometry by time-of-flight (CyTOF; metal isotopes), use antibodies conjugated to a detectable molecule (25). New multimodal methods have been developed to measure both proteomes and transcriptomes. Indexed FACS followed by scRNA-seq is one of the simpler methods for profiling RNA and a small number of proteins (26). Another method is the proximity extension assay (PEA) for protein measurement in parallel with RNA analysis (27). During PEA, two cDNA sequences are used to tag two different epitopes of the same protein. This facilitates mutual priming and extension into a sequence that can be detected using quantitative PCR.

CITE-seq (cellular indexing of transcriptomes and epitopes) (28) expands the dimensionality of proteins measured by conjugating antibodies to DNA barcodes that bind cell surface proteins [antibody-derived tags (ADT)]. Following incubation with these antibodies, sequencing occurs, producing a library of both mRNA and protein levels. REAP-seq (RNA expression and protein sequencing) (29) also uses DNA-conjugated antibodies, although with a different chemistry. Multiple studies have applied these technologies to examine links between the transcriptome and proteome at the single-cell level. MacParland et al. (30) examined the cellular landscape of the normal human liver using CITE-seq, leveraging both protein and RNA expression to identify 20 hepatic cell populations, including two distinct populations of liver-resident macrophages with inflammatory and noninflammatory immunoregulatory functions. The trimodal assay TEA-seq (simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility) (31) builds on these approaches to produce measures comparable to unimodal assays, although with somewhat lower resolution.

2.5. Spatial Organization Data

In order to understand a system and a cell’s immediate environment it is often necessary to profile its spatial organization. A major challenge of single-cell sequencing involves matching the transcriptomic or other cellular properties with their position within a tissue. Several approaches including imaging-based techniques and sequencing-based techniques have been developed.

Imaging-based technologies, such as those leveraging multiplexed fluorescence in situ hybridization (FISH), offer high-spatial-resolution detection of mRNAs at the single-cell level (32). Single-molecule RNA imaging approaches, such as single-molecule FISH (33), involve multiple short DNA probes conjugated to the same fluorescence dye. Methods such as MERFISH (multiplexed error-robust FISH) (34) expand the number of genes measured from tens to thousands by using multiple rounds of hybridization. However, the higher number of hybridization rounds results in long imaging times, large amounts of data, and increased error (35).

Sequencing-based technologies capitalize on barcoding positional information in an array to capture and sequence mRNA using untargeted probes. This includes the Visium spatial transcriptomic platform (10× Genomics) (36) whose capture area is approaching single-cell resolution. Another technology, ZipSeq, enables the researcher to identify tissue regions of interest through real-time imaging that then are labeled with printed barcodes (zipcodes) to be subsequently sequenced using a standard droplet-based scRNA-seq workflow (37). Additional methods such as XYZeq use a spatially barcoded array to label cells rather than capture mRNA, such that the tissue can be dissociated and labeled cells sequenced using scRNA-seq (38).

3. COMPUTATIONAL INTEGRATION OF MULTIOMIC DATA

Analysis of multiomic data requires processing the different data types; this can be done for each modality entirely separately, or by integrating the data types at one of several steps in the analysis. Below, we describe the workflow for single-cell mono-omic analysis as a scaffold for contextualizing multiomic integration (Figure 2).

3.1. Single-Cell Mono-Omic Analysis

Standard computational processing for scRNA-seq can be divided into different steps, including data alignment, quality control (QC), normalization, integration, and visualization (Figure 2). There are several tools that enable this analysis, including some developed by companies that sell the reagents for the single-cell experiments, such as Loupe Browser by 10× Genomics (https://www.10xgenomics.com/products/loupe-browser), along with web-based interfaces that provide user-friendly tools (40).

The first step involves aligning raw data to the genome, producing a count matrix of features (e.g., genes, transcripts) per cell. Alignment methods include STAR (spliced transcripts alignment to a reference) (41) and Kallisto (42), which are also used for bulk RNA-seq data, as well as Cell Ranger (10× Genomics) (5), which is specific to single-cell sequencing.

QC filtering is an essential procedure, since barcodes that represent unwanted droplets, such as dying cells, empty droplets, or contamination, are often present. The most general filtering criteria include removing cells with high numbers of mitochondrial-encoded genes and with low numbers of mRNA transcripts and library sizes (43). Identifying low-quality cells can be difficult and is highly dependent on the biological properties of the sample. Another data quality challenge is the identification of doublets, or droplets containing more than one cell.

The next step is to normalize gene expression data to account for technical and biological variation. Commonly, expression is scaled to a fixed number of counts per cell (often 10,000) and log-normalized. This is followed by the selection of highly variable genes, with the goal of reducing data size to facilitate further processing, while keeping relevant biological features for downstream clustering and visualization. Features are then scaled or centered; at this step, desired covariates may also be regressed out.

Application of dimensionality reduction methods, usually principal component analysis (PCA), is the next step for summarizing the data. After this, a shared neighborhood graph is built based on the distances between cells in the reduced space, and graph-based community detection methods, such as Louvain (44) or Leiden (45) clustering, are used to group similar cells into clusters. These clusters are then examined for differential expression (DE) of key markers and annotated with cell types. Clusters are visualized with t-distributed stochastic neighbor embedding (t-SNE) (46) or uniform manifold approximation and projection (UMAP) (47), which are graph-based dimensionality reduction methods designed to help preserve structure for visualization.

Different modalities require different bioinformatics approaches. For snATAC-seq data, after mapping raw transcripts to a reference genome (also performed in scRNA-seq), peaks are called. Alternate QC metrics, including TSS enrichment and fragment size distribution, are used for filtering; after this, because of the sparsity of snATAC-seq data, latent semantic indexing is used for dimensionality reduction instead of PCA. This is followed by clustering and visualization, similar to scRNA-seq. Software such as ChromVar (48), Signac (49), and ArchR (50) provides a framework for the analysis of single-cell chromatin data by performing several analysis tasks, including dimensionality reduction, integration, and the discovery of enriched DNA sequence motifs.

3.2. Batch Correction and Mono-Omic Integration

Batch effect correction across libraries is often necessary. Methods such as those provided by the limma package (51) and ComBat (52) fit a linear model containing a blocking term for the batch structure, but these existing methods were designed for bulk RNA-seq and make assumptions that might not be suitable for scRNA-seq. Other methods were developed for single-cell data (53, 54), for example, identifying mutual nearest neighbors (MNNs) between batches and calculating the difference in expression values between cells, which are then used for correction (55). Scanorama (56) and Seurat integration (57, 58) also leverage a variation of the MNN method. Another method, Harmony (59), uses a low-dimensional space to group the data into distinct clusters, favoring diverse clusters from different datasets. Seurat and Harmony, as well as multiple additional methods, such as LIGER (linked inference of genomic experimental relationships) (60), CyCombine (61), and Cobolt (62), can be used for both batch correction and multimodal integration (see Section 4 below for more about these methods).

3.3. Bulk Multiomic Integration

Computational methods for single-cell multiomic integration have been greatly influenced by the existing methodology in the bulk space. Integrative multiomic methods for bulk measurements (i.e., per-sample rather than per-cell measurements) have been developed and reviewed extensively (63, 64). These methods can be broadly categorized into three classes of approaches. Joint dimensionality reduction methods seek to produce a shared lower-dimensional space from multiple omic datasets that maximize covariance. These include multiple co-inertia analysis (65), joint and individual variance explained (66), canonical correlation analysis (CCA) (67), multiomic factor analysis (MOFA) (68), generalizations of non-negative matrix factorization (NMF) (such as intNMF, jointNMF, and multiNMF) (69), and partial least squares (DIABLO) (70). Several of these have been adapted for use on single-cell multiomic data, which are discussed in Section 4.2.2 below. Similarity- or network-based methods, such as similarity network fusion (71), first compute similarities per modality, followed by integration. Finally, statistical methods model the underlying distribution of the data, using either a Bayesian prior (e.g., Bayesian consensus clustering) (72) or different distributional assumptions (e.g., PARADIGM, iCluster) (73, 74). These integrative multiomic methods have been applied to examine underlying structure in the data such as disease heterogeneity (68) and to relate biological measurements to categories and outcomes of interest such as cancer type (75) and survival (64).

4. SINGLE-CELL MULTIOMIC INTEGRATION

Multimodal single-cell experiments can yield either paired (from the same cell) or unpaired (from different cells) measurements (Figure 3). Paired data require specific experimental protocols, yielding high-resolution insights into individual cells. Unpaired measurements may result from the same sample separated into portions for each modality, or from two datasets measuring similar populations of cells (e.g., same species and tissues), and could be gathered together by different laboratories for different experiments.

The type of measurement (paired versus unpaired) greatly affects downstream computational integration, as one is a problem of vertical integration (different measurements from the same cells) and the other is diagonal integration (different measurements from different cells)—horizontal integration involves the same type of measurement across different cells and is discussed above in Section 3.2 (76). Because of these key differences, the following sections separately cover methods designed for paired (Section 4.1) and unpaired (Section 4.2) data. We also address the case where the data are partially paired (Section 4.3) (Table 1).

Multimodal integration often requires transformation of the input data to a shared feature space in order to analyze the modalities together. For example, snATAC-seq data include locations of accessibility peaks while scRNA-seq data involve counts per gene. For integration, the accessibility peaks are converted to a gene matrix; the simplest method for doing this involves summing the counts over the promoter and gene body. Other methods, such as Cicero (77), model the relationship between accessibility and expression and convert snATAC-seq data to the gene level. Certain multimodal data integration methods leverage matrices generated from paired data to perform this conversion (78), while others are able to take the raw peak data as input to their models (62, 79).

4.1. Methods for Paired Data

Paired associations can be examined via (*a*) transferring labels from one modality to another, (*b*) creating a shared low-dimensional space, or (*c*) late integration (63, 80), which we define as data integration after dimensionality reduction.

4.1.1. Label transfer.—The simplest way of examining relationships between cells with paired multimodal data involves clustering in one modality and then overlaying measurements from other modalities to further understand the behavior of the cells in each cluster. For paired spatial transcriptomic assays, such as Visium, XYZeq, and ZipSeq, this is often a key component of the analysis: Transcriptomic data are processed following the mono-omic workflow, and then cell labels are transferred to the spatial locations. Initial efforts for paired assays, such as for CITE-seq (28), involved defining a low-dimensional space and clustering based on gene expression, followed by examination of the cell surface markers within each cluster. While this allows for examination of multiple data types, it misses information added by the cell surface markers that could separate cell populations in clustering. For T and natural killer (NK) cell subpopulations, which are hard to distinguish based on transcriptomic data (28), using clusters based on transcription alone may reduce our ability to identify these populations.

4.1.2. Joint dimensionality reduction.—In the case where multiple modalities can aid in distinguishing between cell types, computational methods for paired multimodal integration can be used to create a shared low-dimensional space. The low-dimensional space can also be used for clustering and data visualization. There are several ways to generate this shared low-dimensional space, including by matrix decomposition and by neural network–based methods.

4.1.2.1. Matrix decomposition.: Single-cell experiments generate hundreds to thousands of measurements per cell, leading to large cell–measured feature matrices. As one of the first steps in single-cell mono-omic analysis, dimensionality reduction is performed to identify underlying axes of variation and reduce the number of dimensions for downstream analysis. In most cases, PCA is used. Matrix decomposition divides a matrix into the product of two matrices, a matrix of latent factors-by-samples and a matrix of factors-by-features, such that each data point in the original matrix is made up of a linear combination of the underlying factors multiplied by their loadings. For dimensionality reduction, the number of factors selected is less than the original number of features, so each sample (in this case, cell) can be represented by a smaller number of values. PCA works by identifying independent factors that explain the maximum amount of variation in a dataset. Multiple additional matrix decomposition techniques exist, including independent component analysis and NMF; these can also be leveraged to combine multiple data types into a shared space (81).

Factor analysis is a widely used technique for identifying a small number of latent factors that drive observed variation in a dataset. MOFA (68) is a method for identifying shared and modality-specific factors within a dataset. The authors of MOFA+ (82) extended the original bulk-focused method to accommodate the large sizes of single-cell datasets and to consider group relationships between cells. Consideration of group structure is particularly important in single-cell analysis because each sample contains a set of cells, rather than a single measurement.

Another method, scAI (single-cell aggregation and integration) (83), iteratively performs matrix decomposition, learning both a low-dimensional representation and a cell–cell similarity matrix across paired transcriptomic and epigenomic measurements. The cell–cell similarity matrix can be used for linking between modalities, while the shared dimensionality reduction is used for clustering and downstream cell type identification.

Argelaguet et al. (82) applied MOFA+ to a mouse scNMT-seq dataset of 1,800 cells at three time points in mouse development, and used the resulting latent factors to identify different lineages in development. Jin et al.'s (83) application of scAI to an sci-CAR dataset from 8,800 kidney cells led to better identification of subpopulations than with scRNA-seq alone. The authors also used the joint clustering and cell similarity matrix to perform motif discovery and examine cell-type-specific TF patterns.

4.1.2.2. Neural networks.: Matrix decomposition methods such as PCA and factor analysis assume linearity of the feature space, which is not the case for most biological data. For example, bulk gene expression measurements follow a negative binomial distribution; for single-cell data, this is often zero-inflated because of dropout (84). Neural network methods allow for the modeling of nonlinear relationships. In particular, several multimodal integration methods leverage variational autoencoders (VAEs), which, similar to PCA and factor analysis, perform dimensionality reduction, identifying a reduced latent space to describe the data. VAEs consist of an encoder, a decoder, and a loss function. The encoder iteratively learns the latent space (or encoding) from the data and the decoder regenerates the data from the reduced-dimensionality latent space. A loss function minimizes the error in data reconstruction, keeping the maximum amount of information in this latent space, and regularization is used to reduce the number of latent variables, avoid overfitting, and increase model interpretability.

The method totalVI (total variational inference) (85), which is part of the Python library scvi-tools (86), uses VAEs to integrate scRNA-seq and protein data from CITE-seq, while considering each modality's different background and noise distributions. Leveraging an RNA noise model from scVI (single-cell variational inference) (87), totalVI also includes a protein model that separates protein signal from background, where background includes both the antibodies in empty droplets and nonspecific binding. Following this correction, totalVI learns a joint probability distribution for integration and joint dimensionality reduction. This results in a 20-dimensional latent space for integration, clustering, and visualization. A new addition to scvi-tools (86), multiVI (88), focuses on integrating scRNA-seq and snATAC-seq using the scVI (87) RNA and the peakVI (89) accessibility models for each modality, respectively.

Other VAEs for multimodal integration include scMVAE (single-cell multimodal VAE) (90), which leverages a VAE with a probabilistic Gaussian mixture model because it has been shown to work better with sparse inputs, which is important for integrating epigenomic data. scMM (79) extends this, using a mixture of experts multimodal VAE (MVAE), which models the raw count data in each data type separately (the experts), so that the posterior distributions are estimated separately and then mixed evenly in the joint representation. For chromatin accessibility data, scMM uses a zero-inflated negative binomial model to model peak counts. The authors also provide a procedure for generating so-called pseudocells from different latent variables to aid in model interpretability. Another method, Cobolt (62), works for both unimodal and multimodal data. Both scMM and Cobolt work with peak summaries directly and do not summarize them to the gene level.

Application of totalVI (85) to a CITE-seq dataset of over 32,000 mouse spleen and lymph node cells allowed for better characterization of variation in B cell populations, identifying a unique subpopulation of mature B cells. Zuo & Chen (90) applied the scMVAE method to paired snATAC-seq and scRNA-seq data from a mixture of human cell lines to predict TF–target gene pairs, two-thirds of which were in an existing regulatory network database.

4.1.3. Late integration using shared neighborhood graphs.—In a single-cell analysis workflow, following dimensionality reduction, a neighborhood graph is generated and then clustered. A UMAP or t-SNE is often derived from the neighborhood graph for two-dimensional visualization. While matrix decomposition and VAEs perform dimensionality reduction, shared neighborhood graphs work downstream of standard dimensionality reductions (e.g., PCA) to perform joint clustering.

The method CiteFuse (91) uses network fusion to combine low-dimensional representations of ADT and RNA data, allowing for downstream identification of ligand–receptor interactions. Weighted nearest neighbors (WNN) (92) (available in Seurat v4) learns cell-specific weights for each modality and uses these to generate a shared nearest-neighbor graph across modalities. The cell-specific modality weights are based on how well a cell's neighbors in each modality predict the cell's profiles for both modalities. These weights make WNN robust to cross-modality variation in data quality and number of features, as well as the amount of information provided for each cell (e.g., antibody data are generally more helpful for identifying immune cells).

Hao et al. (92) demonstrated that applying WNN to a cord blood mononuclear cell CITE-seq dataset improved separation between CD8$^+$ and CD4$^+$ T cells and identified a subpopulation of NK cells. Application of WNN to bone marrow mononuclear cell and peripheral blood mononuclear cell (PBMC) CITE-seq datasets with larger antibody panels, as well as to Multiome from 10× Genomics and SHARE-seq datasets, also demonstrated improved separation of cell types when including multiple modalities to define clusters.

4.2. Methods for Unpaired Data

While paired data provide important insights about within-cell relationships, they can only be produced by experimental technologies designed for multiple measurements (e.g., G&T-seq, scM&T-seq) and it is not always possible to use these methods. This could be in part due to cost, timing, or limitations in available technology and information provided. Examination of unpaired data also allows for the combination of data across batches and the reanalysis of existing datasets. As a result, methods for integrating unpaired single-cell measurements are required. Similar to methods for paired data, we organize these methods into three categories: label transfer, joint dimensionality reduction, and late integration.

4.2.1. Label transfer.—

Label transfer is generally used when one modality provides less information than the other modality, such as with sparse epigenomic measurements or spatial data. One example, scJoint (93), performs a neural network–based dimensionality reduction to generate a shared space and then uses the *k*-nearest neighbors to identify the nearest labeled neighbors of an unlabeled cell and transfer labels. Mixing in the low, shared-dimensional space is thereby improved using the transferred labels.

4.2.2. Joint dimensionality reduction.—

While label transfer can be helpful in many contexts, it biases the results to one modality and may miss the signal provided by the combination of modalities. Generating a shared low-dimensional representation, for downstream clustering and annotation, provides a powerful technique for combining datasets. While the goal of dimensionality reduction is similar to the case of paired data, different methodologies are required to address the challenge of diagonal integration (e.g., different cells, different measurements).

4.2.2.1. Matrix decomposition.:

Above we described how methods for matrix decomposition are regularly used in unimodal analysis (e.g., PCA) and can help with identifying latent factors in paired data (e.g., with factor analysis). For unpaired data, matrix decomposition methods can also be used to identify shared factors. Below, we highlight two NMF methods, as well as a method that uses CCA to perform integration.

NMF is a dimensionality reduction technique that results in factors with only positive entries. These positive, sparse loadings make the relationships between samples and factors (as well as features and loadings) more interpretable; the loadings in the sample-by-factor matrix can be interpreted as soft-clustering assignments of samples to latent factors (81, 94). One method, coupled NMF (78), along with the use of NMF to identify shared factors, also uses a conversion (or coupling) matrix generated from publicly available paired data to translate between snATAC-seq and scRNA-seq measurements.

LIGER (95) is another NMF-based method that leverages integrative NMF (iNMF) to combine datasets. iNMF(94) expands upon NMF to combine datasets into sets of shared and dataset-specific latent features. Following iNMF, LIGER creates a shared factor graph by identifying cells across datasets with similar factor neighborhoods. Clustering is performed on this graph, and the factor loadings of the smaller dataset are quantile normalized to match the larger one. UINMF (96) is an extension of LIGER that learns low-dimensional representation from both shared and unshared features. When combining other modalities with scRNA-seq, there are often many unshared features; for spatial data this is often genes, and in accessibility data this can be intergenic regions. UINMF improves integration with little additional computational cost by including this information. Another expansion of LIGER to include online learning (97) was designed for integrating large datasets with fixed amounts of memory. Online iNMF breaks datasets into mini-batches, which are used to iteratively update the NMF factors and loadings.

CCCA identifies pairs of maximally correlated canonical variables, which are linear combinations of the variables in each dataset. Stuart et al. (58) used CCA and anchor finding (Seurat v3) to combine unpaired multimodal datasets. Briefly, the datasets are reduced into a joint low-dimensional space with CCA, and then pairs of cells (anchors) between datasets are identified in this space, filtered, and scored based on their shared nearest neighbor overlap. One dataset (the query) is then transformed based on the other (reference) by the weighted average of the integration anchor vectors, resulting in a corrected expression matrix.

Duren et al. (78) applied coupled NMF to examine the effects of retinoic acid treatment on mouse embryonic stem cells. Using both expression of TFs from scRNA-seq and TF motif enrichment in chromatin-accessible regions from snATAC-seq, the authors constructed cluster-specific gene regulatory networks (GRNs) based on proximity (see Section 5.5).

Both Welch et al. (95) and Stuart et al. (58) each applied their respective methods to mouse cortex samples to integrate scRNA-seq data with epigenomic and spatial measurements. The authors demonstrated that integration across data types leads to identification of cell types not found within the sparser modality alone (snATAC-seq, methylation, or spatial). Integration with spatial data also allowed for the examination of the expression localization of genes not labeled by spatial methods.

4.2.2.2. Neural networks.: While most neural network–based integration methods exist for paired data, a small number can be used with unpaired data. scDART (single-cell deep learning model for ATAC-seq and RNA-seq trajectory integration) (98) is a model designed for integrating snATAC-seq and scRNA-seq data. scDART learns the gene–chromatin region relationships from the data instead of using a predefined gene activity matrix, which allows for nonlinear relationships between regions and genes and does not assume colocalization of regulation and expression. scMoGNN (99) is a graph neural network–based method. The authors of this model used it because it aggregates information from the neighborhood graph, and it has had success in single-cell unimodal analysis (100). scMoGNN starts with cell–feature bipartite graphs for each modality and learns separate low-dimensional spaces for each modality that are then aggregated in a joint network. The method also allows for the addition of known information on between-modality interactions to the graph.

4.2.2.3. Manifold alignment.: Several methods for unpaired integration leverage the technique of manifold alignment, which has been used in other domains to join multiple data types resulting from observations of the same events. Manifold alignment works by generating a low-dimensional space or manifold for each modality, and then aligning the manifolds to create a shared space.

MATCHER (manifold alignment to characterize experimental relationships) (101) adapted manifold alignment to combine single-cell datasets from different modalities, specifically focusing on applications related to development—assuming the single-cell data exist on a 1D manifold that only moves in one direction. Another manifold alignment method, MAGAN (manifold aligning generative adversarial network) (102), does not require input data to exist on a 1D trajectory. MAGAN also leverages generative adversarial networks (GANs) (103), which are a pair of neural networks that compete to fool each other. The authors use GANs in part because they scale better for massive datasets than most other graph-based manifold alignment methods.

Welch et al. (101) applied MATCHER to two paired single-cell transcriptome and methylome datasets, demonstrating its ability to accurately align these data types and examine the epigenome/transcriptome relationship during stem cell development. Amodio et al. (102) applied MAGAN to combine scRNA-seq and flow cytometry data, as well as CyTOF data, demonstrating that this method can align similar cell populations.

4.2.3. Late integration.—Late integration of unpaired multiomic data involves alignment of modalities that have already been processed and projected into a low-dimensional space. Aligning these data often occurs by jointly clustering across modalities, or training models of the relationship between data types.

4.2.3.1. Joint clustering.: Some late integration methods combine modalities during clustering. Harmony (59) is a soft-clustering-based method that is generally used for integrating across mono-omic batches, but it can also be applied to multimodal datasets. Harmony starts with principal component space representations of the datasets, and then iteratively assigns cells to soft clusters while maximizing the diversity of the datasets within each cluster in order to obtain clusters that are organized by cell type rather than dataset. CyCombine (61) is a method for combining cytometry measurements, such as those from CITE-seq or CyTOF, across batches, panels, and modalities. CyCombine clusters cells using a self-organizing map (104) for dimensionality reduction and clustering and then applies ComBat (105) to each cluster of cells to perform batch correction.

Korunsky et al. (59) applied Harmony to a spatial (MERFISH) and scRNA-seq dataset to generate a shared low-dimensional space. This improved the cell type labels for the spatial dataset, identified cell population locations, and predicted the localization TFs not measured in the spatial data. Pedersen et al. (61) demonstrated that CyCombine allowed for the examination of similar populations across technologies, integrating over 6,700 PBMCs from spectral flow, CyTOF, and CITE-seq.

4.2.3.2. Bayesian methods.: Satija et al. (106) developed an integration method for combining spatial and scRNA-seq data that leverages Bayesian methods and implemented this in their software Seurat (v3.2+). Their integration method first trains a LASSO (least absolute shrinkage and selection operator) model (107) on the scRNA-seq data to impute the expression of each of the spatial or landmark genes. The spatial data are divided into bins, and for each bin, a multivariate normal model is trained to predict the joint expression of landmark genes from scRNA-seq. Finally, for each cell in the scRNA-seq data, the posterior probability of each bin is estimated and summarized to identify the cell’s location. clonealign (108) is a method for combining unpaired single-cell genomic and transcriptomic data to assign expression to clones in cancer data. For each clone, clonealign models the relationship between copy number variation and expression, adding a term for noise and covariates of interest, and then estimates the assignment of each cell across clones using the variational Bayes method.

Satija et al. (106) combined FISH and scRNA-seq data from zebrafish embryos to generate high-resolution maps of gene expression localization and identify rare cell populations and their locations. Campbell et al. (108) applied clonealign to identify clone-specific differentially regulated genes in an ovarian cell line not identified by the scRNA-seq data alone.

4.3. Methods for Partially Paired Data

Many of the methods for paired data, including MOFA+ (82), totalVI (85), multiVI (88), scMM (79), and Cobolt (62), also work in the case that the data are only partially paired (i.e., only a portion of cells have measurements from multiple modalities). Other methods were developed specifically for this case; for example, Hao et al. (109) introduced a method for multimodal bridge integration (Seurat) to allow for mapping across multiple modalities to an scRNA-seq reference. The method uses dictionary learning to identify underlying elements of each modality in a paired dataset, and then uses this as a bridge to impute measurements in the unpaired data.

For data from one modality (i.e., if the second modality is fully missing), it is also possible to computationally generate profiles for another modality if a model exists to predict these measurements. BABEL (110) is a deep learning method for translating between expression and chromatin accessibility. The method includes scRNA-seq and snATAC-seq encoder/decoder pairs to translate in either direction. The model is trained on data from PBMCs, as well as colon, colorectal and lymphoblastoid cell lines—as such, it shows better performance in data on cell types more similar to those in the training dataset, but it can also be applied in other contexts.

5. DOWNSTREAM ANALYSIS

Following multimodal integration, clustering, and cell type identification, a variety of methods can be applied to examine associations within and between modalities. Many of these methods are also applied to unimodal datasets following clustering, with the key difference being that the clusters are defined based on a single data type rather than multiple data types.

Downstream methods are applied at the single-cell or pseudobulk level. For pseudobulk analysis, the measurements of cells within a single cluster or group of clusters are summarized together, making them appear similar to the bulk assay of that modality. Pseudobulk summarization can occur across the whole cluster (or group of clusters)—or can be separated by biological samples within that cluster (e.g., for n samples in a cluster, we could calculate n pseudobulk measurements).

5.1. Differential Expression Analysis

DE analysis helps with the identification of cell types and with the examination of differences between groups of samples. Tools originally developed for bulk RNA-seq data [e.g., DESeq2 (111), edgeR (112)] are regularly applied to single-cell data, after summarization to pseudobulk measurements, to identify differentially expressed genes (DEG). In addition, an increasing number of DEG analysis methods have been developed for single-cell data, with varying strengths and limitations (113, 114). For example MAST (84), which uses a Hurdle model and a normalization procedure that transforms the UMI counts into a dense matrix, is especially useful for handling zero-inflation and zero-deflation present in single-cell data. Monocle (115) uses a generalized additive model and has proven to be successful in handling the response variables of both categorical and continuous data (116). Methods such as SCDE (single-cell differential expression) (117) that use a mixture model are commonly used to capture the different abundance of specific transcripts in each cell.

Differential analysis tests are specific to the feature types collected by each modality. For example, for snATAC-seq data, differences in peak locations or accessible regions are generally examined.

5.2. Gene Set Enrichment Analysis

Aggregating genes to the gene set or pathway level both reduces the number of tests and aids in the interpretability of DEG results. Most gene set enrichment methods test for significant associations with curated gene sets from databases such as MSigDB (Molecular Signatures Database) (118, 119). Gene set enrichment analysis (GSEA) (118, 120) is a method developed for bulk gene expression data that ranks genes by expression (usually average log-fold change across conditions) and uses a Kolmogorov–Smirnov test to determine whether a set of genes is overenriched at the top or bottom of this ranked list. Single-sample GSEA (121) ranks the gene expression of each sample separately; a variation of this is often applied to single-cell analysis, with rankings calculated at the cell rather than sample level. GSVA (gene set variation analysis) (122) is another sample-based nonparametric gene set enrichment method created for bulk data, which can be applied to individual cells rather than samples. Several single-cell-specific methods have been developed for gene enrichment [e.g., Pagoda (123), AUCell (124)]. These are reviewed elsewhere (125, 126), along with their relative performance compared to bulk methods.

WGCNA (weighted gene co-expression network analysis) (127) does not rely on existing sets of genes; rather, it constructs a network from the data, with each edge weight corresponding to the co-expression of the genes it connects. The method is helpful for grouping highly correlated genes into clusters, which can then be analyzed for DE.

5.3. Quantitative Trait Loci

In quantitative trait locus (QTL) analyses, tests for associations between genetic variants and changes in another modality (e.g., expression, methylation, splicing) are performed to identify QTLs [e.g., expression QTL (eQTL), methylation QTL, or splicing QTL]. The majority of these analyses have been performed to examine associations with eQTLs, so we focus on this modality here, but similar principles apply for other modalities.

Previous bulk RNA-seq studies have identified both widespread and tissue-specific eQTLs (128). While single-cell eQTL analyses have lower statistical power because of increased sparsity and smaller dataset sizes than bulk analysis, they are better at detecting cell type and context-dependent associations (129). For single-cell eQTL (sc-eQTL) analysis, annotated clusters are summarized to pseudobulk measurements and tested for associations between genetic variants and expression. Genetic data are often gathered separately, with either a genotyping array or whole-genome sequencing; however, it is possible to call variants from scRNA-seq alone (130).

Most eQTL analyses focus on *cis*-eQTLs, which are proximal (with a defined region, often 1–2 kb) to the gene they are associated with, because subsetting to a local region reduces the number of tests performed. CRISPR/Cas9 genome editing combined with paired genomic and transcriptome measurements (e.g., G&T-seq) can be used to generate experimental eQTLs, which are helpful for identifying *trans*-eQTLs (nonproximal) (129).

Multiple sc-eQTL studies (131–133) have identified both known (i.e., matching those from bulk) and novel cell- and context-specific associations. For example, Yazar et al. (133) identified 26,597 *cis*- and 990 *trans*-eQTLs in a dataset of 1.3 million PBMCs from 982 donors across 14 immune cell types, many of which were associated with known autoimmune disease risk loci.

5.4. Trajectory Inference

Single-cell omics provides a powerful tool for understanding dynamic processes such as cell development and activation. By assuming that transcriptomic similarity is the driving force of differentiation, computational trajectory inference or pseudotime methods can track a progression along a differentiation pathway (134). The structure of the dynamic process can either be linear, bifurcating (or tree shaped), or nonlinear (e.g., cyclic). The majority of these algorithms require previous knowledge to determine the process's direction and are applied following a standard workflow, after which these cells are ordered and abstracted into a graph (135). The most popular trajectory inference methods include Monocle (115, 116), slingshot (136), and PAGA (partition-based graph abstraction) (137).

Incorporating pseudotime with different biological measurements (e.g., epigenetic, protein level, and spatial information) can provide a better representation by which to understand processes in differentiation and development (138). MATCHER (101), which was discussed above in the context of paired data, uses a manifold alignment approach; it has been used to reconstruct the correlation between transcriptomic and epigenetic changes in embryonic stem cells.

5.5. Regulatory Network Analysis

Understanding the structure of GRNs has been a central goal of systems biology research. While many methods infer GRNs from scRNA-seq data (reviewed in 139), identifying these networks from transcription data alone is challenging (140), and analysis of single-cell multimodal data can provide complementary for examining regulation. In particular, multimodal data containing epigenomic (e.g., snATAC-seq, methylation) and transcriptomic measurements can help identify these regulatory relationships (scRNA-seq and proteomic data can also be helpful for examining signaling networks). Genetic perturbation screens by methods such as CRISPR/Cas9 also help infer GRNs (141). These screens can be assayed by scRNA-seq [e.g., Perturb-seq (142)] or multimodal methods such as G&T-seq for examining both expression and genome editing.

Following integration of multimodal data, regulatory relationships can be explored by examining the correlations and colocalizations of open regions, methylation sites, or regulatory motifs with gene expression. These relationships are often explored by cluster, using pseudobulk summarization. For example, Welch et al. (95) calculated correlations between methylation signal and the expression of known TFs in order to examine cell-type-specific regulation patterns. Using accessibility data, Duren et al. (78), constructed GRNs using edges between proximal-cluster-specific accessible peaks and expressed genes. Duren et al. (78), Jin et al. (83), and Stuart et al. (58) also all identified cluster-specific TF motifs using RNA expression of TFs and enrichment of TF motifs in accessible regions.

In addition to postintegration comparisons at the cluster level, several methods have been developed specifically for identifying GRNs from multimodal data, including MIRA (143), scBPGRN (single-cell back-propagation GRN) (144), and Symphony (145).

6. CHALLENGES

Single-cell multiomic analysis presents several unique challenges. Improvements to single-omic technologies that help address noise and limitations of these methods will improve multiomic technologies as well. For example, for scRNA-seq and the multimodal technologies that include it, the incomplete capture of mRNAs presents a challenge, which newer microfluidic technologies attempt to tackle (146). Paired multiomic technologies can be costly and often involve an extra step in experimental processing. Protocols that use manual isolation to separate the nucleus and the cytoplasm (e.g., G&T-seq, scM&T-seq) can lose mRNA and DNA molecules, and this loss can be partially avoided by using other methods (e.g., DR-seq) that process the DNA and the RNA molecules together in a preamplification step (147). The options for paired multiomic technologies also do not cover all possible pairs and groupings of single-omic methods; future work is required to generate and optimize protocols for measuring novel combinations of omics technologies.

Computationally, integration across single-cell modalities can be challenging. All methods for single-cell multimodal integration must address a variety of challenges related to each modality’s noise, features measured, and information provided. Each data type presents its own bias: Spatial and protein data have both probe- and antibody-related background signals, scRNA-seq data have ambient RNA and problems of dropout, and epigenomic data involve sparse binary or almost binary measurements. Additionally, there are large differences in the numbers of features measured, ranging from tens or hundreds for spatial and protein data to tens of thousands for scRNA-seq data.

While many integration methods create a shared low-dimensional representation that can be visualized using UMAP or t-SNE, additional methods for visualizing the relationships between modalities are required (148). Models that can impute missing data or modalities are limited by the data on which they are trained, and since most analyses have focused on PBMCs and cell lines, further work is required to optimize protocols and analyses for tissue samples. Difficulties also exist that are related to the large size and sparsity of the single-cell data. Neural network and online learning methods help scale integration to increasingly large and atlas-sized datasets, but this issue, along with dataset sparsity, must be a continued area of consideration.

7. SUMMARY AND PERSPECTIVES

In the past decade, the ability to make high-dimensional measurements at the single-cell level has skyrocketed and continues to increase in throughput and affordability. These technologies have extended to multiomic measurements within and across cells, providing complementary and integrated information about cellular state and behavior. With these multimodal technologies have come a variety of methods for data integration and analysis, which we have reviewed here.

Multiomic data paired with computational integration approaches hold tremendous potential. Big cataloging efforts are underway that phenotype cell types and tissues in the human body at the single-cell level, including the Human Cell Atlas (149) and the Tabula Sapiens project (150); beyond this, multiomic approaches are being applied across patient populations as a means of understanding biological variation and disease states. As we move forward, it will be critical to share these data broadly and to reach a consensus around which methods are most effective for extracting biological information. In addition, we should seek to share these data and methodologies in a form that is readily accessible to the research community, such that limitations in data science expertise and compute infrastructure are not barriers to discovery. In doing so, we can open up new dimensions of single-cell patient profiling toward precision-level understanding of disease mechanisms, disease progression, and opportunities for treatment.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We would like to thank members of the Fragiadakis lab, the UCSF Bakar ImmunoX Initiative, and the UCSF CoLabs for discussion and guidance for this review. We would like to thank Dr. Walter Eckalbar, Dr. Kyle Marchuk, and Mohammad Nasser for scientific discussion and editing the manuscript. Further support came from the Bakar ImmunoX Initative and the ImmunoX Computational Biology Initiative at UCSF (funding for G.K.F.); NIH (National Institutes of Health) grants U01DE028891-01A1 and 5P30AR070155-05 and the Autoimmunoprofiler Initiative from UCSF and Eli Lilly (for E.F. and G.K.F; and NIH grants 3U19AI077439, R01AI093615-11, R01DK103735, U01AI168390, R01AI170239, P30AI027763-31, and R01DE032033 (for G.K.F.).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Gawad C, Koh W, Quake SR. 2016. Single-cell genome sequencing: current state of the science. Nat. Rev. Genet 17(3):175–88 [PubMed: 26806412]
2. Tang X, Huang Y, Lei J, Luo H, Zhu X. 2019. The single-cell sequencing: new developments and medical applications. Cell Biosci. 9:53 [PubMed: 31391919]
3. Haque A, Engel J, Teichmann SA, Lönnberg T. 2017. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 9:75 [PubMed: 28821273]
4. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161(5):1202–14 [PubMed: 26000488]
5. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, et al. 2017. Massively parallel digital transcriptional profiling of single cells. Nat. Commun 8:14049
6. Lebrigand K, Magnone V, Barbry P, Waldmann R. 2020. High throughput error corrected Nanopore single cell transcriptome sequencing. Nat. Commun 11:4025 [PubMed: 32788667]
7. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, et al. 2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science 357(6352):661–67 [PubMed: 28818938]
8. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, et al. 2012. Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods 9:72–74
9. Sena JA, Galotto G, Devitt NP, Connick MC, Jacobi JL, et al. 2018. Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. Sci. Rep 8:13121
10. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, et al. 2015. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat. Methods 12(6):519–22 [PubMed: 25915121]
11. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, et al. 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 344(6190):1396–1401 [PubMed: 24925914]
12. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. 2015. Integrated genome and transcriptome sequencing of the same cell. Nat. Biotechnol 33(3):285–89 [PubMed: 25599178]
13. Klemm SL, Shipony Z, Greenleaf WJ. 2019. Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet 20(4):207–20 [PubMed: 30675018]
14. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, et al. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523(7561):486–90 [PubMed: 26083756]
15. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, et al. 2020. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell 183(4):1103–16.e20 [PubMed: 33098772]
16. Bartosovic M, Kabbe M, Castelo-Branco G. 2021. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. Nat. Biotechnol 39(7):825–35 [PubMed: 33846645]
17. Wu SJ, Furlan SN, Mihalas AB, Kaya-Okur HS, Feroze AH, et al. 2021. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. Nat. Biotechnol 39(7):819–24 [PubMed: 33846646]
18. Zhu C, Zhang Y, Li YE, Lucero J, Behrens MM, Ren B. 2021. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. Nat. Methods 18(3):283–92 [PubMed: 33589836]
19. Pelizzola M, Ecker JR. 2011. The DNA methylome. FEBS Lett. 585(13):1994–2000 [PubMed: 21056564]
20. Li Y, Tollefsbol TO. 2011. DNA methylation detection: bisulfite genomic sequencing analysis. Methods Mol. Biol 791:11–21 [PubMed: 21913068]

21. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, et al. 2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat. Methods 11(8):817–20 [PubMed: 25042786]

22. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, et al. 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat. Methods 13(3):229–32 [PubMed: 26752769]

23. Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, et al. 2018. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat. Commun 9:781 [PubMed: 29472610]

24. Lay FD, Kelly TK, Jones PA. 2018. Nucleosome Occupancy and Methylome Sequencing (NOMe-seq). Methods Mol. Biol 1708:267–84 [PubMed: 29224149]

25. Bendall SC, Simonds EF, Qiu P, Amir ED, Krutzik PO, et al. 2011. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 332(6030):687–96 [PubMed: 21551058]

26. Nestorowa S, Hamey FK, Pijuan Sala B, Diamanti E, Shepherd M, et al. 2016. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood 128(8):e20–31 [PubMed: 27365425]

27. Darmanis S, Gallant CJ, Marinescu VD, Niklasson M, Segerman A, et al. 2016. Simultaneous multiplexed measurement of RNA and proteins in single cells. Cell Rep. 14(2):380–89 [PubMed: 26748716]

28. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, et al. 2017. Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods 14(9):865–68 [PubMed: 28759029]

29. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, et al. 2017. Multiplexed quantification of proteins and transcripts in single cells. Nat. Biotechnol 35(10):936–39 [PubMed: 28854175]

30. MacParland SA, Liu JC, Ma X-Z, Innes BT, Bartczak AM, et al. 2018. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nat. Commun 9:4383 [PubMed: 30348985]

31. Swanson E, Lord C, Reading J, Heubeck AT, Genge PC, et al. 2021. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. eLife 10:e63632

32. Fan Y, Braut SA, Lin Q, Singer RH, Skoultchi AI. 2001. Determination of transgenic loci by expression FISH. Genomics 71(1):66–69 [PubMed: 11161798]

33. Chen J, McSwiggen D, Ünal E. 2018. Single molecule fluorescence in situ hybridization (smFISH) analysis in budding yeast vegetative growth and meiosis. J. Vis. Exp 135:e57774

34. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. 2015. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 348(6233):aaa6090

35. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. 2022. An introduction to spatial transcriptomics for biomedical research. Genome Med. 14:68 [PubMed: 35761361]

36. 10× Genom. 2021. Inside Visium spatial capture technology. Tech. Rep, 10× Genom. Inc., Pleasanton, CA. https://pages.10xgenomics.com/rs/446-PBO-704/images/10x_BR060_Inside_Visium_Spatial_Technology.pdf

37. Hu KH, Eichorst JP, McGinnis CS, Patterson DM, Chow ED, et al. 2020. ZipSeq: barcoding for real-time mapping of single cell transcriptomes. Nat. Methods 17(8):833–43 [PubMed: 32632238]

38. Lee Y, Bogdanoff D, Wang Y, Hartoularos GC, Woo JM, et al. 2021. XYZeq: Spatially resolved single-cell RNA sequencing reveals expression heterogeneity in the tumor microenvironment. Sci. Adv 7(17):eabg4755

39. Delete in proof.

40. Moreno P, Huang N, Manning JR, Mohammed S, Solovyev A, et al. 2021. User-friendly, scalable tools and workflows for single-cell RNA-seq analysis. Nat. Methods 18(4):327–28 [PubMed: 33782609]

41. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21 [PubMed: 23104886]

42. Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol 34(5):525–27 [PubMed: 27043002]

43. Stegle O, Teichmann SA, Marioni JC. 2015. Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet 16(3):133–45 [PubMed: 25628217]

44. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp 2008(10):P10008

45. Traag VA, Waltman L, van Eck NJ. 2019. From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep 9:5233 [PubMed: 30914743]

46. van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. J. Mach. Learn. Res 9(86):2579–2605

47. McInnes L, Healy J, Saul N, Großberger L. 2018. UMAP: uniform manifold approximation and projection. J. Open Sour. Softw 3(29):861

48. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. 2017. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods 14(10):975–78 [PubMed: 28825706]

49. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. 2021. Single-cell chromatin state analysis with Signac. Nat. Methods 18(11):1333–41 [PubMed: 34725479]

50. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, et al. 2021. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat. Genet 53(3):403–11 [PubMed: 33633365]

51. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, et al. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43(7):e47 [PubMed: 25605792]

52. Zhang Y, Parmigiani G, Johnson WE. 2020. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom. Bioinform 2(3):lqaa078

53. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, et al. 2020. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 21:12 [PubMed: 31948481]

54. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, et al. 2022. Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods 19:41–50 [PubMed: 34949812]

55. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol 36(5):421–27 [PubMed: 29608177]

56. Hie B, Bryson B, Berger B. 2019. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol 37(6):685–91 [PubMed: 31061482]

57. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol 36(5):411–20 [PubMed: 29608179]

58. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, et al. 2019. Comprehensive integration of single-cell data. Cell 177(7):1888–1902.e21 [PubMed: 31178118]

59. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, et al. 2019. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods 16(12):1289–96 [PubMed: 31740819]

60. Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, Welch JD. 2020. Jointly defining cell types from multiple single-cell datasets using LIGER. Nat. Protoc 15(11):3632–62 [PubMed: 33046898]

61. Pedersen CB, Dam SH, Barnkob MB, Leipold MD, Purroy N, et al. 2022. cyCombine allows for robust integration of single-cell cytometry datasets within and across technologies. Nat. Commun 13:1698 [PubMed: 35361793]

62. Gong B, Zhou Y, Purdom E. 2021. Cobolt: integrative analysis of multimodal single-cell sequencing data. Genome Biol. 22:351 [PubMed: 34963480]

63. Rappoport N, Shamir R. 2018. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 46(20):10546–62 [PubMed: 30295871]

64. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, et al. 2021. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat. Commun 12:124 [PubMed: 33402734]

65. Meng C, Kuster B, Culhane AC, Gholami AM. 2014. A multivariate approach to the integration of multi-omics datasets. BMC Bioinform. 15:162

66. Lock EF, Hoadley KA, Marron JS, Nobel AB. 2013. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann. Appl. Stat 7(1):523–42 [PubMed: 23745156]

67. Tenenhaus A, Tenenhaus M. 2011. Regularized generalized canonical correlation analysis. Psychometrika 76(2):257–84

68. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, et al. 2018. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol 14(6):e8124

69. Chalise P, Fridley BL. 2017. Integrative clustering of multi-level ‘omic data based on non-negative matrix factorization algorithm. PLOS ONE 12(5):e0176278

70. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, et al. 2019. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics 35(17):3055–62 [PubMed: 30657866]

71. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, et al. 2014. Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods 11(3):333–37 [PubMed: 24464287]

72. Lock EF, Dunson DB. 2013. Bayesian consensus clustering. Bioinformatics 29(20):2610–16 [PubMed: 23990412]

73. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, et al. 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26(12):i237–245 [PubMed: 20529912]

74. Shen R, Olshen AB, Ladanyi M. 2009. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 25(22):2906–12 [PubMed: 19759197]

75. Cai Z, Poulos RC, Liu J, Zhong Q. 2022. Machine learning for multi-omics data integration in cancer. iScience 25(2):103798 [PubMed: 35169688]

76. Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. 2021. Computational principles and challenges in single-cell data integration. Nat. Biotechnol 39(10):1202–15 [PubMed: 33941931]

77. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, et al. 2018. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. Mol. Cell 71(5):858–71.e8 [PubMed: 30078726]

78. Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, et al. 2018. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. PNAS 115(30):7723–28 [PubMed: 29987051]

79. Minoura K, Abe K, Nam H, Nishikawa H, Shimamura T. 2021. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. Cell Rep. Methods 1(5):100071

80. Adossa N, Khan S, Rytkönen KT, Elo LL. 2021. Computational strategies for single-cell multi-omics integration. Comput. Struct. Biotechnol. J 19:2588–96 [PubMed: 34025945]

81. Stein-O’Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, et al. 2018. Enter the matrix: factorization uncovers knowledge from omics. Trends Genet. 34(10):790–805 [PubMed: 30143323]

82. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, et al. 2020. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 21:111 [PubMed: 32393329]

83. Jin S, Zhang L, Nie Q. 2020. scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. Genome Biol. 21:25 [PubMed: 32014031]

84. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, et al. 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 16:278 [PubMed: 26653891]

85. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, et al. 2021. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat. Methods 18(3):272–82 [PubMed: 33589839]

86. Gayoso A, Lopez R, Xing G, Boyeau P, Valiollah Pour Amiri V, et al. 2022. A Python library for probabilistic analysis of single-cell omics data. Nat. Biotechnol 40(2):163–66 [PubMed: 35132262]

87. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. 2018. Deep generative modeling for single-cell transcriptomics. Nat. Methods 15(12):1053–58 [PubMed: 30504886]

88. Ashuach T, Gabitto MI, Jordan MI, Yosef N. 2021. MultiVI: deep generative model for the integration of multi-modal data. bioRxiv 2021.08.20.457057. 10.1101/2021.08.20.457057

89. Ashuach T, Reidenbach DA, Gayoso A, Yosef N. 2022. PeakVI: A deep generative model for single-cell chromatin accessibility analysis. Cell Rep. Methods 2(3):100182

90. Zuo C, Chen L. 2021. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. Brief. Bioinform 22(4):bbaa287

91. Kim HJ, Lin Y, Geddes TA, Yang JYH, Yang P. 2020. CiteFuse enables multi-modal analysis of CITE-seq data. Bioinformatics 36(14):4137–43 [PubMed: 32353146]

92. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, et al. 2021. Integrated analysis of multimodal single-cell data. Cell 184(13):3573–87.e29 [PubMed: 34062119]

93. Lin Y, Wu T-Y, Wan S, Yang JYH, Wong WH, Wang YXR. 2022. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. Nat. Biotechnol 40(5):703–10 [PubMed: 35058621]

94. Yang Z, Michailidis G. 2016. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics 32(1):1–8 [PubMed: 26377073]

95. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. 2019. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell 177(7):1873–87.e17 [PubMed: 31178122]

96. Kriebel AR, Welch JD. 2022. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. Nat. Commun 13:780 [PubMed: 35140223]

97. Gao C, Liu J, Kriebel AR, Preissl S, Luo C, et al. 2021. Iterative single-cell multi-omic integration using online learning. Nat. Biotechnol 39(8):1000–7 [PubMed: 33875866]

98. Zhang Z, Yang C, Zhang X. 2022. scDART: integrating unmatched scRNA-seq and scATAC-seq data and learning cross-modality relationship simultaneously. Genome Biol. 23:139 [PubMed: 35761403]

99. Wen H, Ding J, Jin W, Wang Y, Xie Y, Tang J. 2022. Graph neural networks for multimodal single-cell data integration. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 4153–63. New York: Assoc. Comput. Mach.

100. Wang J, Ma A, Chang Y, Gong J, Jiang Y, et al. 2021. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. Nat. Commun 12:1882 [PubMed: 33767197]

101. Welch JD, Hartemink AJ, Prins JF. 2017. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biol. 18:138 [PubMed: 28738873]

102. Amodio M, Krishnaswamy S. 2018. MAGAN: aligning biological manifolds. Proc. Mach. Learn. Res 80:215–23

103. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, et al. 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst 63(11):139–44

104. Kohonen T. 1990. The self-organizing map. Proc. IEEE 78(9):1464–80

105. Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8(1):118–27 [PubMed: 16632515]

106. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol 33(5):495–502 [PubMed: 25867923]

107. Tibshirani R. 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B Methodol 58(1):267–88

108. Campbell KR, Steif A, Laks E, Zahn H, Lai D, et al. 2019. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. Genome Biol. 20:54 [PubMed: 30866997]

109. Hao Y, Stuart T, Kowalski M, Choudhary S, Hoffman P, et al. 2022. Dictionary learning for integrative, multimodal, and scalable single-cell analysis. bioRxiv 2022.02.24.481684. 10.1101/2022.02.24.481684

110. Wu KE, Yost KE, Chang HY, Zou J. 2021. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. Proc. Natl. Acad. Sci 118(15):e2023070118

111. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15(12):550 [PubMed: 25516281]

112. McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 40(10):4288–97 [PubMed: 22287627]

113. Das S, Rai A, Merchant ML, Cave MC, Rai SN. 2021. A comprehensive survey of statistical approaches for differential expression analysis in single-cell RNA sequencing studies. Genes 12(12):1947 [PubMed: 34946896]

114. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, et al. 2021. Confronting false discoveries in single-cell differential expression. Nat. Commun 12:5692 [PubMed: 34584091]

115. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, et al. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol 32(4):381–86 [PubMed: 24658644]

116. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, et al. 2017. Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods 14(10):979–82 [PubMed: 28825705]

117. Kharchenko PV, Silberstein L, Scadden DT. 2014. Bayesian approach to single-cell differential expression analysis. Nat. Methods 11(7):740–42 [PubMed: 24836921]

118. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102(43):15545–50 [PubMed: 16199517]

119. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 1(6):417–25 [PubMed: 26771021]

120. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, et al. 2003. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet 34(3):267–73 [PubMed: 12808457]

121. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, et al. 2009. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 462(7269):108–12 [PubMed: 19847166]

122. Hänzelmann S, Castelo R, Guinney J. 2013. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinform. 14:7

123. Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, et al. 2016. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat. Methods 13(3):241–44 [PubMed: 26780092]

124. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, et al. 2017. SCENIC: single-cell regulatory network inference and clustering. Nat. Methods 14(11):1083–86 [PubMed: 28991892]

125. Zhang Y, Ma Y, Huang Y, Zhang Y, Jiang Q, et al. 2020. Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. Comput. Struct. Biotechnol. J 18:2953–61 [PubMed: 33209207]

126. Holland CH, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, et al. 2020. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol. 21:36 [PubMed: 32051003]

127. Zhang B, Horvath S. 2005. A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol 4:17

128. Aguet F, Brown AA, Castel SE, Davis JR, He Y, et al. 2017. Genetic effects on gene expression across human tissues. Nature 550(7675):204–13 [PubMed: 29022597]

129. van der Wijst M, de Vries D, Groot H, Trynka G, Hon C, et al. 2020. The single-cell eQTLGen consortium. eLife 9:e52155

130. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, et al. 2018. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat. Biotechnol 36(1):89–94 [PubMed: 29227470]

131. Wills QF, Livak KJ, Tipping AJ, Enver T, Goldson AJ, et al. 2013. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. Nat. Biotechnol 31(8):748–52 [PubMed: 23873083]

132. van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, Franke L. 2018. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat. Genet 50(4):493–97 [PubMed: 29610479]

133. Yazar S, Alquicira-Hernandez J, Wing K, Senabouth A, Gordon MG, et al. 2022. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. Science 376(6589):eabf3041

134. Saelens W, Cannoodt R, Todorov H, Saeys Y. 2019. A comparison of single-cell trajectory inference methods. Nat. Biotechnol 37(5):547–54 [PubMed: 30936559]

135. Tritschler S, Büttner M, Fischer DS, Lange M, Bergen V, et al. 2019. Concepts and limitations for learning developmental trajectories from single cell genomics. Development 146(12):dev170506

136. Street K, Risso D, Fletcher RB, Das D, Ngai J, et al. 2018. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genom. 19:477

137. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, et al. 2019. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. 20:59 [PubMed: 30890159]

138. Cannoodt R, Saelens W, Saeys Y. 2016. Computational methods for trajectory inference from single-cell transcriptomics. Eur. J. Immunol 46(11):2496–2506 [PubMed: 27682842]

139. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. 2020. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat. Methods 17(2):147–54 [PubMed: 31907445]

140. Efremova M, Teichmann SA. 2020. Computational methods for single-cell omics across modalities. Nat. Methods 17:14–17 [PubMed: 31907463]

141. Henkel L, Rauscher B, Schmitt B, Winter J, Boutros M. 2020. Genome-scale CRISPR screening at high sensitivity with an empirically designed sgRNA library. BMC Biol. 18:174 [PubMed: 33228647]

142. Dixit A, Parnas O, Li B, Chen J, Fulco CP, et al. 2016. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell 167(7):1853–66.e17 [PubMed: 27984732]

143. Lynch AW, Theodoris CV, Long HW, Brown M, Liu XS, Meyer CA. 2022. MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. Nat. Methods 19(9):1097–108 [PubMed: 36068320]

144. Xuan C, Wang Y, Zhang B, Wu H, Ding T, Gao J. 2022. scBPGRN: integrating single-cell multi-omics data to construct gene regulatory networks based on BP neural network. Comput. Biol. Med 151:106249

145. Bachireddy P, Azizi E, Burdziak C, Nguyen VN, Ennis CS, et al. 2021. Mapping the evolution of T cell states during response and resistance to adoptive cellular therapy. Cell Rep. 37(6):109992

146. Dal Molin A, Di Camillo B. 2019. How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives. Brief. Bioinform 20(4):1384–94 [PubMed: 29394315]

147. Macaulay IC, Ponting CP, Voet T. 2017. Single-cell multiomics: multiple measurements from single cells. Trends Genet. 33(2):155–68 [PubMed: 28089370]

148. Miao Z, Humphreys BD, McMahon AP, Kim J. 2021. Multi-omics integration in the age of million single-cell data. Nat. Rev. Nephrol 17(11):710–24 [PubMed: 34417589]

149. Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA. 2017. The Human Cell Atlas: from vision to reality. Nature 550(7677):451–53 [PubMed: 29072289]

150. Tabula Sapiens Consort., Jones RC, Karkanias J, Krasnow MA, Pisco AO, et al. 2022. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. Science 376(6594):eabl4896

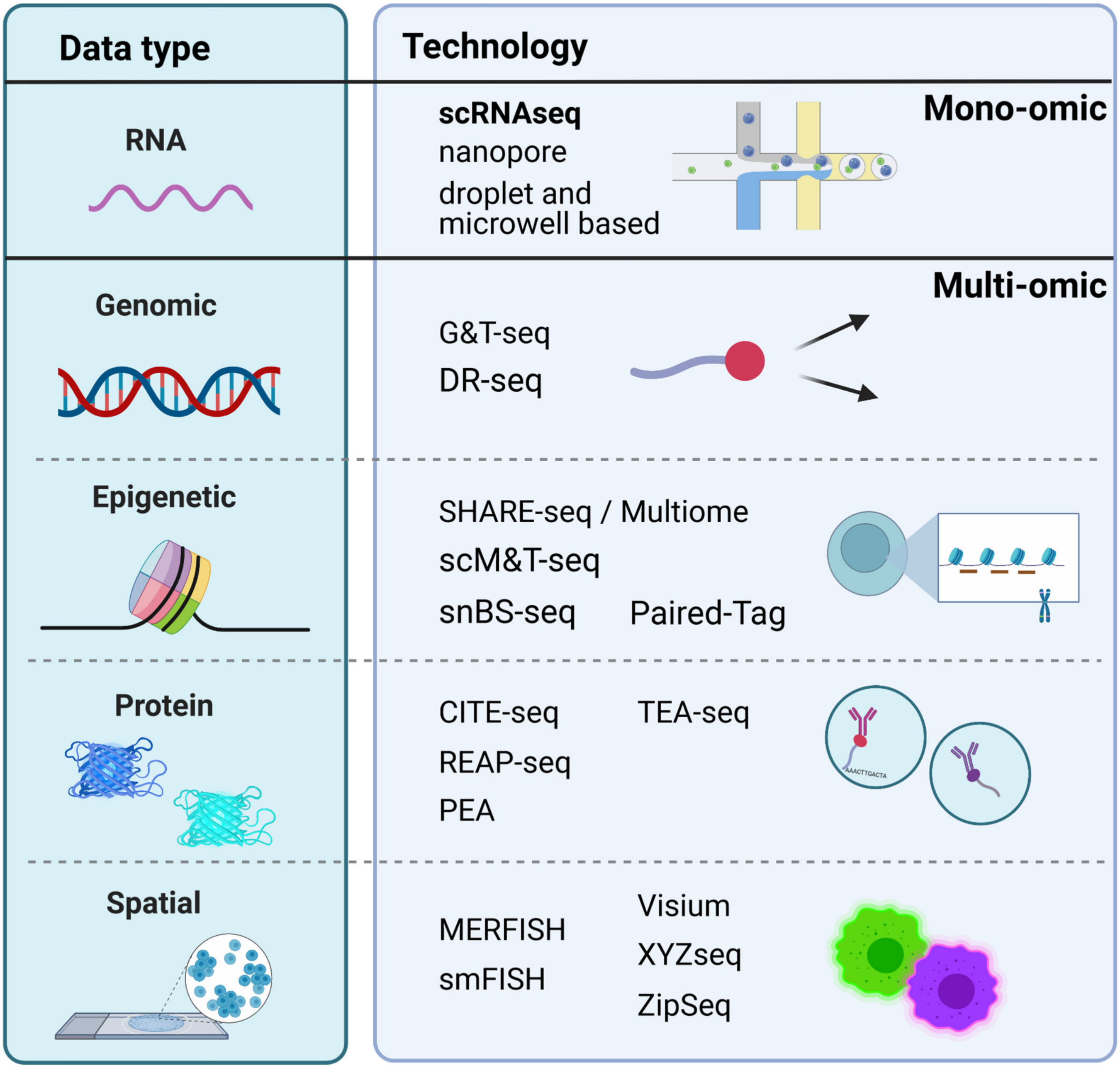

Figure 1.

An overview of multimodal technologies. Multimodal technologies listed involve the simultaneous measurement of transcriptomic, genomic, epigenetic, proteomic, or spatial information. Technologies listed are not comprehensive but represent many of the most prevalent technologies used, as described in more detail in the text. Abbreviations: CITE-seq, cellular indexing of transcriptomes and epitopes; DR-seq, genomic DNA–messenger RNA sequencing; FISH, fluorescence in situ hybridization; G&T-seq, genome and transcriptome sequencing; MERFISH, multiplexed error-robust FISH; Paired-Tag, parallel analysis of individual cells for RNA expression and DNA from targeted tagmentation by sequencing; PEA, proximity extension assay; REAP-seq, RNA expression and protein sequencing; scBS-seq, single-cell bisulfite sequencing; scM&T-seq, single-cell methylome and transcriptome sequencing; scRNA-seq, single-cell RNA sequencing; SHARE-seq, simultaneous high-throughput ATAC and RNA expression with sequencing; smFISH, single-molecule FISH; TEA-seq, simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility. Figure adapted from images created with BioRender.com.

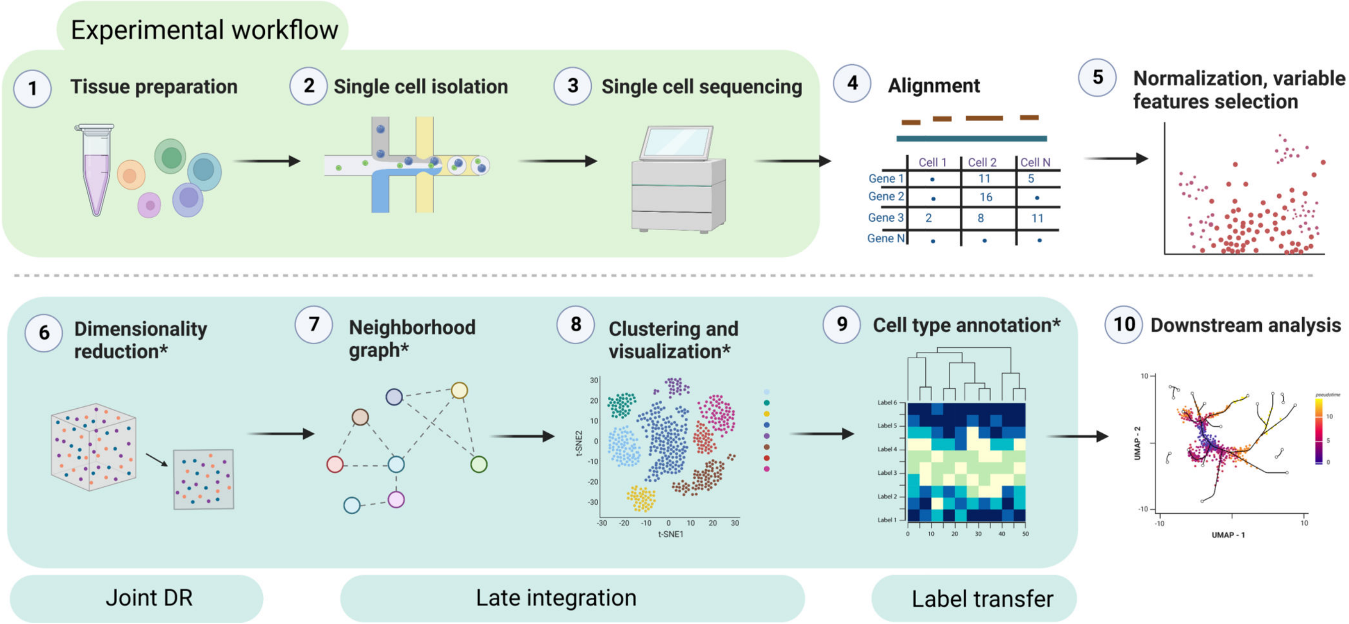

Figure 2.

Experimental and data processing workflows of single-cell sequencing data. Typical processing involves tissue preparation, single-cell isolation, and sequencing (experimental steps are highlighted in green), followed by alignment, normalization, dimensionality reduction (DR), neighborhood graph generation, and cell clustering. This is followed by cell type annotation and downstream analysis. Steps where multimodal integration is often performed are highlighted in light blue and marked with an asterisk, with the integration phase (joint DR, late integration, or label transfer) listed underneath. Figure adapted from images created with BioRender.com.

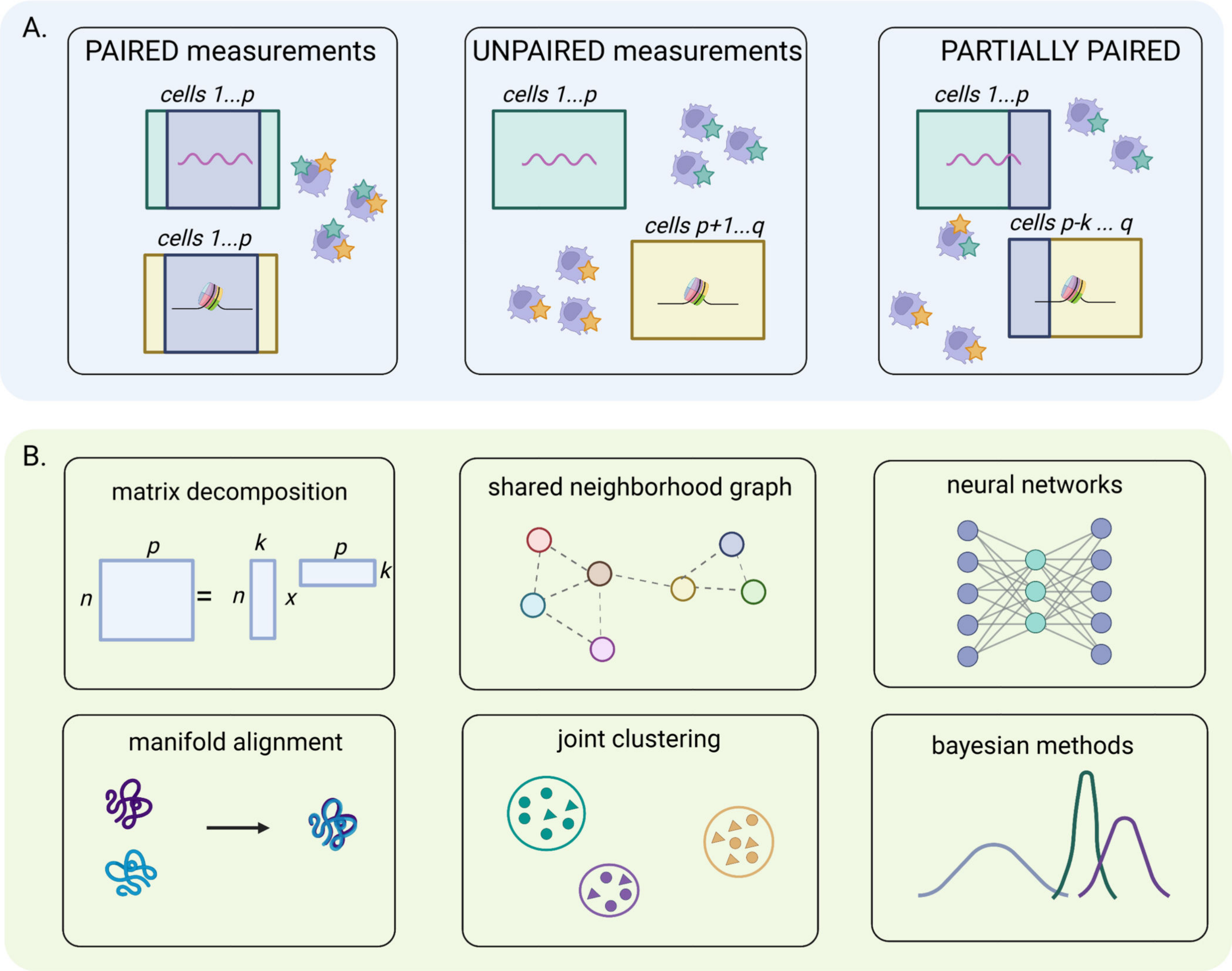

Figure 3.

Computational strategies for single-cell multiomic integration. (a) Types of input data: paired, unpaired, and partially paired. Different colors represent different modalities, and cell diagrams show whether these are measured on the same or separate cells. (b) Integration techniques: matrix decomposition, shared neighborhood graphs, joint clustering, manifold alignment, Bayesian methods, and neural networks. Figure adapted from images created with BioRender.com.

Table 1

List of computational methods for multiomic integration[a]

Method	Reference	Data type	Integration phase	Method type
MOFA+	Argelaguet et al. (82)	Paired	Joint reduction	Matrix decomposition
scAI	Jin et al. (83)	Paired	Joint reduction	Matrix decomposition
totalVI	Gayoso et al. (85)	Paired	Joint reduction	Neural network
multiVI	Ashuach et al. (88)	Paired	Joint reduction	Neural network
scMVAE	Zuo & Chen (90)	Paired	Joint reduction	Neural network
scMM	Minoura et al. (79)	Paired	Joint reduction	Neural network
Cobolt	Gong et al. (62)	Paired	Joint reduction	Neural network
CiteFuse	Kim et al. (91)	Paired	Late integration	Joint neighborhood graph
WNN (Seurat v4)	Hao et al. (92)	Paired	Joint reduction	Joint neighborhood graph
scJoint	Lin et al. (93)	Unpaired	Label transfer	Neural network
Coupled NMF	Duren et al. (78)	Unpaired	Joint reduction	Matrix decomposition
LIGER	Welch et al. (95)	Unpaired	Joint reduction	Matrix decomposition
UINMF	Kriebel & Welch (96)	Unpaired	Joint reduction	Matrix decomposition
CCA and anchor finding (Seurat v3)	Stuart et al. (58)	Unpaired	Joint reduction	Matrix decomposition
scDART	Zhang et al. (98)	Unpaired	Joint reduction	Neural network
scMoGNN	Wen et al. (99)	Unpaired	Joint reduction	Neural network
MATCHER	Welch et al. (101)	Unpaired	Joint reduction	Manifold alignment
MAGAN	Amodio et al. (102)	Unpaired	Joint reduction	Manifold alignment, neural network
Harmony	Korunsky et al. (59)	Unpaired	Joint reduction	Joint clustering
CyCombine	Pedersen et al. (61)	Unpaired	Late integration	SOM, joint clustering
Seurat spatial integration	Satija et al. (106)	Unpaired	Late integration	Bayesian
clonealign	Campbell et al. (108)	Unpaired	Late integration	Bayesian
Multimodal bridge integration (Seurat)	Hao et al. (109)	Partially paired	Imputation	Dictionary learning
BABEL	Wu et al. (110)	Unimodal only	Imputation	Neural network

[a] Methods are organized based on data type, integration phase, computational technique, and single-cell data types to which they can be applied. Researchers can make use of the full Supplemental Table 1 for analysis methods selection by taking into account the integration type, as well as parameters such as their input data modality, programming language, and resource availability.

Abbreviations: ATAC-seq, assay for transposase-accessible chromatin with sequencing; CCA, canonical correlation analysis; GNN, graph neural network; LIGER, linked inference of genomic experimental relationships; MAGAN, manifold aligning generative adversarial network; MATCHER, manifold alignment to characterize experimental relationships; MOFA+, multiomic factor analysis, version 2; NMF, nonnegative matrix factorization; RNA-seq, RNA sequencing; scAI, single-cell aggregation and integration; scDART, single-cell deep learning model for ATAC-seq and RNA-seq trajectory integration; scMVAE, single-cell multimodal variational autoencoder; SOM, self-organizing map; VI, variational inference; WNN, weighted nearest neighbors.