Published in final edited form as:

Nat Rev Nephrol. 2021 November ; 17(11): 710–724. doi:10.1038/s41581-021-00463-x.

Multiomics integration in the age of million single cell data

Zhen Miao[1,2], **Benjamin D. Humphreys**[3], **Andrew P. McMahon**[4], **Junhyong Kim**[1,2,*]

[1]Department of Biology, University of Pennsylvania, Philadelphia, PA, USA

[2]Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[3]Division of Nephrology, Department of Medicine, Washington University in St. Louis, St. Louis, MO, USA

[4]Department of Stem Cell Biology and Regenerative Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

Abstract

An explosion in single cell technologies has revealed a previously underappreciated heterogeneity of cell types and novel cell state associations with sex, disease, development and other processes. Starting with transcriptome analyses, single cell techniques have been extended to multi-omics approaches, and now enable the simultaneous measurement of data modalities and cellular spatial context. Data are now available for millions of cells, for whole-genome measurements, and for multiple modalities. Although analyses of such multimodal datasets have potential to provide new insights into biological processes that cannot be inferred with a single mode of assay, the integration of very large, complex, multimodal data into biological models and mechanisms represents a considerable challenge. An understanding of the principles of data integration and visualization methods is required to determine what methods are best applied to a particular single cell data set. Each class of method has advantages and pitfalls in terms of its ability to achieve various biological goals, including cell type classification, regulatory network modeling, and biological process inference. In choosing a data integration strategy, consideration must be given to whether the multiome data are matched (that is, measured on the same cell) or unmatched (that is, measured on different cells) and, more importantly, the overall modelling and visualization goals of the integrated analysis.

[*] junhyong@sas.upenn.edu .

Competing interests

APM is a scientific advisor to Novartis, eGENESIS, TRESTLE Therapeutics and IVIVA Medical. The other authors declare no competing interests.

Related links:
Rebuilding a Kidney (https://www.rebuildingakidney.org/)
GenitoUrinary Development Molecular Anatomy Project (gudmap.org/)
KidneyCellExplorer (https://cello.shinyapps.io/kidneycellexplorer/)
Susztaklab Kidney Biobank (https://susztaklab.com/)
Kidney Interactive Transcriptomics (http://humphreyslab.com/SingleCell/)
HuBMAP portal (https://portal.hubmapconsortium.org/)

Introduction

Many different technologies are now available to measure various properties of any biological system. For instance, the same physical–chemical properties can be measured with different instruments (e.g., RNA sequencing and RNA microarray) and different physical–chemical properties can be measured for the same object (e.g., protein and RNA content of a cell). In the past few years, genome-scale technologies have led to the systematic generation of very large-scale quantitative datasets that comprise multiple measurement modalities. Although such multimodal datasets have potential to provide unprecedented insights into biological systems, their analysis and interpretation can be complicated due to modality-specific technical problems and modeling challenges to drawing common inference from different kinds of information.

The term "biological data integration" has been used to describe analytic methods that combine information from multiple sources into a single biological inference. At one level, biological data integration might represent an extremely broad concept, such as the integration of diverse information types including data from Electronic Medical Records, genomic analyses, phenotypic assays and literature reviews into a broad scientific model or hypothesis[1]. In this broad context, the term lacks technical meaning and is not pursued further here. Rather, we focus on biological data integration in the context of integrating large-scale omics data, especially at the single cell level[2,3]. These types of data have a high-degree of multiplexing, for example, with tens of thousands of gene measurements, leading to high-dimensional datasets. If we consider the expression level of a single gene to be one "dimension" of our data set; then a set of 10,000 genes would create a dataset of 10,000 dimensions. Each of these dimensions is commonly called a "feature" of the dataset. Single cell measurements also tend to have considerable noise and technical artifacts — a problem that is somewhat counterbalanced by the ability of new technologies to obtain measurements from thousands or even millions of objects, for example cells, in a given tissue[4,5]. This large number of cell measurements alleviates some of the problems associated with high dimensional data and noise, but creates additional challenges associated with high computational demand and biological complexity.

Despite the challenges associated with high-dimensional data, high-noise and large numbers of measurements, high throughput, single cell omics methodologies have already provided key insights into kidney biology. For example, single cell analyses have identified over 30 cell types along the continuous epithelial network; greater diversity likely exists in regional and sex-related cell states amongst these groupings[6,7]. As another example, time-series single nucleus RNA-sequencing (snRNA-seq) has been used to identify dynamic and spatially distinct proinflammatory and profibrotic subsets of proximal tubule cells that fail to repair after acute kidney injury (AKI)[7,8]. Single cell data can also be combined with clinical parameters such as those regulated by the kidney including blood pressure, blood pH, osmolarity and estimated glomerular filtration rate (eGFR). One single cell transcriptomic profiling study of human kidneys with eGFRs above and below 60 ml/min/1.73 m^2 identified *AP1* and *NKD1* as candidate drivers of kidney fibrosis in patients with chronic kidney disease (CKD)[9].

The abovementioned studies uncovered novel insights into kidney biology using single cell transcriptomics alone. However, in the last five years, many single cell measurement modalities beyond single cell transcriptomics have been developed, including approaches to measure multiple data types in the same cell (so-called multiomics single cell data). More than 30 single cell multiomics techniques[10,11] have been developed since 2015. Although these techniques offer invaluable opportunities to interrogate the properties of cells, the integration of information from these different modalities presents an acute challenge. The high-dimensionality, high noise, and large number of observations underlie this challenge, in which the goal is to reconcile and make comparable distinct modalities into a coherent biological inference.

Even without explicit computational integration, combining of information from different genome-scale data types can yield synergistic inferences. For example, cell-specific gene expression data can be coupled with chromatin status information in the region of a SNP variant, enabling the prioritization of causal variants for further experimental validation[12]. Multi-modal data often augment independent evidence from each mode. For example, one study[13] found that single nucleus assay for transposase-accessible chromatin sequencing (snATAC-seq) refined kidney cell type clusters obtained via snRNA-seq, revealing more clusters with potential clinical relevance. In another study, use of both single cell RNA sequencing (scRNA-seq) and snATAC-seq enabled the identification of a cell-specific regulatory network by inferring upstream regulators from analyses of cis-element motifs[14]. In that study, the identification of cis-regulatory elements with ATAC-seq helped overcome difficulties in detecting regulatory genes, such as transcription factors, in transcriptome data as a consequence of their low abundance. This study and others exemplify that the use of multiple modes of omics information can enable combined inferences that cannot otherwise be obtained from any single mode. Thus, the integration of multi-modal omics data has potential to synthesize more knowledge than would be gained as a sum of individual measurements.

Here, we review developments in computational methods for multi-omics data integration. We first provide a general overview of the principles of data integration. Then, we take a more practical data-centric view of what methods might be applied to a particular data set, starting with a discussion of methods for integrated analyses of multiomics data measured on the same cell, followed by discussion of methods for integrated analyses of multiomics data measured on different cells. We then consider data visualization methods that can integrate different measurement modalities and finally discuss current and future challenges for single cell data integration and prospects for application to kidney biology. Throughout our review, we focus on principles and general factors that determine strengths and challenges of different approaches.

Overview of single cell data integration

As described above, available studies demonstrate that even *ad hoc* integration of multi-modal data can yield inferences that cannot be made with single mode of assay. Many different more principled computational methods are now available to aid the integration of single cell multi-modal omics data, each with different advantages and drawbacks (Figure

1). Here, we first overview the general principles of these different approaches before describing details of the methods in the next section.

Quantitative causal modelling

The most principled form of multi-modal data integration is that which takes into account the actual biological processes that generate the measurements (Fig. 1a). For example, chromatin states, RNA levels, and protein levels represent different aspects of a single system-level molecular dynamics of a cell, where a causal relationship exists between the epigenome state, the number of RNA molecules, and the number of protein molecules. An accurate quantitative systems model of the cell (Box 1) would allow associating multi-modal measurements to parameters of the model, leading to an integrated inference of the dynamic state of the cell. Some computational approaches incorporate partial systems model of a cell's molecular dynamics. For example, the popular algorithm for RNA velocity[15] posits a differential equation model of the kinetics of transcription, splicing and degradation, and estimates the parameters of the model using exonic and intronic reads, in effect integrating the two types of read data into a single model inference. The computational tool protaccel[16] extends this kinetic model to include a differential equation term for proteins, allowing a model-based integration of RNA and protein data, such as can be obtained using methods that enable simultaneous measurement of proteins and mRNAs in single cells (e.g., CITE-seq[17] and REAP-seq[18]). A cell systems model-based data integration approach is ideal for integration of multi-modal data, but currently made impossible by the lack of dependable models for most dynamic molecular processes in a cell — especially models that can predict the dynamics of small finite numbers of molecules in a single cell or complex processes like chromosome remodeling.

Statistical modelling

In the absence of a causal kinetic model, another possible integration approach is to relate different measurement modalities to each other with a statistical model (Fig. 1b). For example, a statistical relationship could be modeled between RNA levels and protein levels[19]; or for example, between the location and amount of open chromatin around a gene and its RNA levels (so-called gene activity models[20]). Therefore, one possible class of methods for integrating different data modalities is to create a statistical model between two or more modalities such that the value from one data type can be mapped to another type. Such models could be calibrated (that is, the model parameters estimated) from reference datasets or fit to the dataset of interest. When such an approach is used, in effect, all of the data points of one modality are converted to (mapped to) the other modality, potentially augmenting the power of the dataset. The downside of this approach is that such translation does not provide additional insights into biological processes related to each data type since this process merely converts one data type into another.

Latent space modelling

Converting one data type to another can be seen as constructing a (mathematical) function (often called a map) between one set of variables and another. The idea that measurements are related by functions motivates a more abstract framework for data integration. We might model data of type A, type B, and type C as a mathematical function from an abstract set of states, which we call “latent states” or “latent space” and corresponding variables “latent variables”. More concretely, the transcriptome, the proteome, and chromatin states might all be considered an aspect of an abstract “latent molecular state” of a cell (Fig. 1c). That is, if the cell is in a latent state X, then mathematical functions of X will predict the number of RNA molecules and protein molecules, and the parts of chromatin that are open. Many machine learning methods such as autoencoders (Box 1) and statistical methods such as factor models (see below) involve estimating a latent space, assumed to determine the observed multi-modal values. This latent space, in a sense, is a representation of the integrated data because it “explains” all of the observed data in different modalities. This concept of latent space from which the observations arise is one of the most common methods of data integration, as discussed below. Different approaches differ in the mathematical functions that map from the latent space to observations (e.g., linear functions versus non-linear functions), in how they model the observed data (e.g., as a probabilistic observation from the latent space), in whether they model only in the vicinity of the observed data or model the entire relationship between latent and measurement spaces, and in the notion of model fit that they use. In the absence of a more mechanical or causal model, the family of latent space models encapsulates the natural idea that different types of measurements must all represent some aspect of an unknown molecular state of the cell. The main downside of such models is that the latent space typically does not have a physical or chemical interpretation, making it difficult to know what the integrated space means in terms of the actual molecular state of a cell. In addition, the same set of cells may have different latent space representations that model different hidden biological states. For example, the same set of cells might have a latent space representation of their cell cycles, another latent space representation of their circadian rhythms, and yet another latent space representation of their cell type identities. Therefore, the utility and variety of the latent space as a model of data integration depends on the goals of the biological inference.

Late integration

The last class of methods for data integration might be called “late integration”[21] in the sense that this approach does not attempt to relate measurements to each other, but rather attempts to use each data modality to infer a model or result unique to that data type, and then attempts to integrate the output models or results (Fig. 1d). For example, we might infer gene regulatory networks from the transcriptome and from the proteome independently, and then apply an algorithm to create a consensus network. Another example might be estimating cell type clusters in each data modality independently before applying algorithms to reconcile the clusters. The above-described study that used snATAC-seq to uncover the dynamics of activity of transcription factors, which were then matched with single cell transcriptome data to identify gene regulatory circuits involved in kidney development[22] can also be thought of as a late integration approach, despite the fact that study did not use an explicit single computational algorithm.

In the best-case scenario, integrated multi-omics or multi-modal analysis can help derive a causal model of cellular processes[22], for example by using the different data modalities to fit a systems process model. Even without a causal model, analyses across modalities can generate a stronger biological inference than can be achieved with single modality analysis. As an example, one study[23] found that correlation between chromatin accessibility and gene expression better reflects chromatin conformation than chromatin accessibility information alone. Data from different modalities can also provide independent evidence for hypothesized processes. For example, motifs in the open cis-chromatin regions uncovered by ATAC-seq can be used to provide additional evidence for transcriptome-based gene regulatory relationships. Approaches that convert between different data modalities or construct a common latent space can augment mutual information derived from each modality and increase the power of subsequent inference. For example, clustering analysis on an integrated latent space might yield more stable estimates of cell types that more closely follow biological processes than that with single modality inference. For exploratory analyses of diseases, integrating multiple measurement modalities might also help narrow the molecular nature of the malfunctioning processes and help determine, for example, whether a disease-related change in gene expression is caused by changes in DNA methylation or chromatin accessibility. In sum, the different approaches of data integration can help the resulting inference become more than the sum of its parts. Below we take a more practical data-centric view of what methods one might apply given a particular set of data (Figure 2).

Integrating jointly profiled multiomics data

The greatest challenge for single cell measurements is recovering molecular fractions from limited amounts of material[24,25]. This problem of molecule recovery efficiency is exacerbated when attempts are made to recover different molecular compartments such as DNA and RNA. However, simultaneous measurements from the same cell alleviates one challenge of multi-modal data integration — mapping the measurement from one modality to another where each modality is measured on a different cell. Here, we refer to data with multimodal measurements on the same cell as matched data. The most popular matched multimodal technique is joint snRNA-seq and snATAC-seq, such as achieved using methods including sci-CAR[26], SNARE-seq[27], paired-seq[28], SHARE-seq[29] and the 10X Genomics multiome solution, which enable isolation and measurement of single cell nuclear transcriptomic and chromatin accessibility data. Techniques are also available for joint measurement of transcriptomic and surface protein data, such as achieved using CITE-seq[17] and REAP-seq[18]. Furthermore, technology has been built to measure single cell phenotypes along with transcriptomic data, providing an important additional dimension for single cell profiling[30]. The technologies used for matched multiomics have been reviewed elsewhere[10,31,32].

Naïve approaches

A number of methods have been developed for the integration of matched multimodal data (Table 1). A naive approach is to transform the data in such a way that all the features (i.e., the measured attributes) have homogeneous statistical characteristics. A classic approach in organismal systematic biology is to scale each feature by its variation across samples[33,34] (in our case, cells). However, this approach results in all features being considered equally important in determining cell variation, which is not biologically reasonable given their disparity in functional importance. A related approach is to give each value of a feature a probabilistic score, perhaps with different models for feature sets, such that the values can have consistent probabilistic interpretation. One example of a model that uses this approach, BREM-SC, assumes a multinomial distribution of each gene in each cell type, for both RNA and protein count matrices obtained using CITE-seq. This type of model enables a probabilistic clustering of cell types[35]. We note this approach is distinct from attempting to statistically translate measurements of one modality onto another. These naïve approaches are simple but ignore the biological context of the different modalities and instead attempt to harmonize the statistical characteristics of the different features, limiting their utility.

Latent space approaches

A more model-based theoretical approach is to consider each measurement, regardless of its modality, an "aspect" (or a "view") of an underlying relationship between the cells. That is, we would assume the existence of a common latent space. One tool that uses this approach to dissect heterogeneity in joint transcriptome and epigenome profiling data is called Single cell Aggregation and Integration (scAI)[36]. To solve the problem presented by the fact that typical epigenomic information such as that obtained through scATAC-seq is often sparse with a high false negative rate, scAI first replaces a cell's value with a similarly weighted average of a random selection of its neighbor's values to 'smooth over' sparse values. It then infers an underlying common latent space by assuming that the data matrix of the transcriptome and the epigenome can be approximated by a weighted linear function of the shared underlying space. An additional constraint (known as a sparseness constraint) is introduced to make the underlying space as simple as possible, along with another constraint that tries to optimize the preservation of original cell-to-cell relatedness in the underlying common space. Application of this method to joint transcriptome and epigenome data from kidney enabled the identification of two subpopulations with distinct open chromatin profiles but similar transcriptomes[36], indicating the need to consider both modalities in order to precisely characterize cell identities.

Latent space approaches can be thought of integrating at the level of features (that is, early integration). Multi-omics Factor Analysis (MOFA) and its updated version, MOFA+, implement group factor analysis to identify shared variation across multiple modalities[37,38]. The basic models of MOFA and MOFA+ are similar to that of scAI; that is, the observed data in each modality is considered a linear weighted function of an underlying common latent space. MOFA+ adds multiple underlying latent spaces to account for group effects such as different experimental batches. The main difference with scAI is that MOFA and MOFA+ explicitly attach a probability model such that each cell's feature value is a random variable that is a function of the common latent space. Thus, while the basic mathematical structure of the model is similar to that of scAI, the way MOFA associates the model to the data is different. Although not tailored for single cell data specifically, the utility of this tool to study a dataset with joint single-cell methylation and transcriptome profiles has been demonstrated[37].

Another tool, totalVI[39] also has similar structure to that of scAI and MOFA in that observed transcriptome and protein measurements (as achieved using CITE-seq[17]) are considered functions of a common latent space. TotalVI relates the observed data and modeled data with a machine learning model (deep neural network) that implements an encoder-decoder scheme (Box 1). The middle layer of this encoder-decoder neural network can be interpreted as a common latent space and used as the integrated variable set to carry out downstream analyses. A potential advantage of totalVI over scAI and MOFA methods is that the neural network architecture allows more complex (non-linear) relationships between the common latent space and measured features.

Late integration approaches

The above methods either explicitly or implicitly aim to infer a common representation space from multi-omics data. An alternative approach involves the integration of data at the level of inferred models (that is, late integration) such as affinity relationships in each modality. One such method[40] called Weighted Nearest Neighbor (WNN) analysis in Seurat V4 synthesizes a combined measure of cell-to-cell affinity from modality-specific affinity models; e.g., cell-cell relationships calculated using RNA data and protein data. We first note that data in each modality can be used to compute neighboring relationships of a cell; i.e., we can have a neighborhood by RNA data and neighborhood by protein data. WNN proposes to measure the informativeness of each kind of neighborhood by assessing how well the cells in each type of neighborhood predicts the RNA or protein value of a given cell. These computations are used to synthesize a weighted average of cell-to-cell affinities from that of each modality. Synthesizing affinity relationships based on a more principled computation idea called "message passing" was proposed in the method Similarity Network Fusion[41] (SNF). In this approach, first a neighborhood relationship is calculated for each object (i.e., cell) from the similarity (or affinity) matrix of each modality. Then the similarity matrices of each modality are "fused" together by "passing" the relationship information from the set of neighboring objects of one matrix to the other matrix, back-and-forth iteratively until they converge. This basic approach was implemented in CiteFuse[42] as a method to integrate affinity relationships from RNA and surface protein from CITE-seq.

Integrating independent multimodal data

With current technologies, a more common problem than the integration of matched datasets is the integration of two or more independently collected datasets (that is, unmatched data), with different modalities. The emergence of comprehensive, single modality, single cell datasets across whole organisms[43,44,45], has led to an abundance of highly accessible data of this type. In general, experimental approaches for joint measurements of certain modalities are still under development or maybe impossible. For example, approaches to simultaneously quantify single cell transcriptome and whole proteome are extremely challenging as single cell proteomics techniques are rapidly advancing but still lack resolution[46]. Single cell lipidomics has been more successful than proteomics at quantitatively identifying molecular species[47], but we are not aware of any attempts at multimodal measurement of lipidomics data. The key problem for unmatched data is that measurements from each modality are unlikely to have cell-to-cell correspondence. That is, in measurements from one set of cells using one modality, say proteins, and another set of cells for another modality, say the transcriptome, it is highly unlikely that there will be cells in each set that correspond exactly to the same cell state for both modalities. Thus, almost by definition, we cannot integrate information at the level of individual cells when measurements are not matched. Current integration approaches, therefore, attempt to match groups of cells, either at the level of distinct cell types, or at the level of local ensembles (neighboring cells). Alternatively, some methods try to statistically map one feature space to another feature space. Here, we classify these methods into three main categories: those that match by annotated cell groups; those that match by a shared feature set; and those that match without a common feature set (Table 2).

Matching by annotated cell groups

When different measurements are made on different sets of data, one coarse grained approach to integrate those measurements is to match groups of cells (e.g., clusters) between the modalities. The clusters in each modality can be associated manually if the clusters correspond to known cell types, which might have been inferred from expert knowledge (e.g., marker gene expression). If cluster label information is not available from established annotations, other features that are biologically informative can be used, such as the proximity of open chromatin to expressed genes, averaged over the ensemble of the cluster to match clusters from each modality. One study, for example, integrated scRNA and scATAC data[48] by linking open chromatin peaks of scATAC-seq cell clusters with the expression of scRNA-seq cell clusters through their proximity in the genome, from which they inferred enhancer–promoter pairs. These enhancer–promoter pairs were consistent with prior knowledge of regulatory networks, supporting the utility of this method. Another approach, MAESTRO[49], incorporates additional information from ChIP-seq databases to help define transcriptional regulators and match clusters based on scRNA and scATAC data.

Matching at the cell group level is also common practice in analyses of spatial transcriptome data. Most current spatial transcriptomics technologies either lack resolution or transcriptome complexity (reviewed elsewhere[50]); however, integrating scRNA-seq with spatial data can help overcome these two limitations. For example, training of a machine learning classifier, Support Vector Machine (SVM), on highly variable genes from annotated scRNA-seq clusters enabled the classifier to identify and map major cell types from sequential fluorescence in situ hybridization (seqFISH) data with which only 125 genes had been profiled[51]. For spatial transcriptomics data with low cellular resolution — such as that obtained using 10X Visium and slide-seq[52], scRNA-seq data can be used to deconvolute the spatially averaged low resolution readout and increase resolution by estimating frequencies of each cell type[53].

Matching with shared feature sets

In rare cases, measurement modalities might be different, but their common molecular basis can be used to match the features. For example, STvEA[54] matches CITE-seq data with multiplexed immunohistochemistry (mIHC) or flow cytometry data using measurements of protein abundance as the common factor. Matching is achieved through mutual nearest neighbor (MNN) correction[55] on the two data matrices, enabling automated annotation of mIHC (or flow cytometry) data with labels from CITE-seq data. Given two sets of objects and a notion of distance across the datasets, MNN identifies pairs of objects in the two sets that are considered to be each other's nearest neighbor. A classic application of MNN is in identifying homologs amongst gene paralogs; variations of the MNN principle have been used widely in data integration.

In the absence of a common molecular basis, measurements of one modality may be connected to features of another modality by some (biologically motivated) statistical model to enable joint analysis. For example, clonealign[56] assumes that increased DNA copy number (inferred from scDNA-seq data) in cancer cells will result in increased gene expression within the corresponding region. Many scRNA–scATAC integration methods synthetically construct a “gene activity matrix” from ATAC data, which is treated as a gene expression feature set. Multiple models have been proposed to infer gene activities from chromatin accessibility data. Seurat V3[57] aggregates all ATAC reads from −2kb of the transcription start site (TSS) throughout the whole gene body to predict expression levels. MAESTRO[49] assigns weights to each peak with an exponential decay based on the distance to the TSS. The Cicero model[20] is more complex and takes into consideration read depth and distal elements that are co-accessible with the TSS. Mapping features of one modality onto another often creates systematic differences that are similar to normalization problems and batch effects. Therefore, good calibration after feature conversion is essential for matching to be successful. Calibration can begin before integration: for example, Seurat V3 and STvEA carry out normalization of both datasets before integration, whereas this step is usually skipped by other models. Seurat V3 and MAESTRO pipelines implement canonical correlation analysis (CCA) to align the two datasets, which are then mapped to the same gene expression feature space. They then apply MNN correction[55] for additional alignment.

Integration of unmatched data by latent models

Similar to matched data cases, data from each modality can be modelled as maps from an abstract set of common factors (latent factors). LIGER[58] uses an integrative non-negative matrix factorization (iNMF[59]) approach to jointly factorize multiple cell-by-feature matrices into cell-by-factor matrices and factor-by-gene matrices using a set of common factors for all matrices and another set of factors specific to each matrix. Factors here refer to hypothetical underlying (latent) features that can be thought as abstract cell states that determine observed values. Multiple modalities can be integrated through statistical modelling of features; for example, by using a quantitative measure of gene accessibility from snATAC-seq to estimate gene activity for integration with scRNA-seq. The factor loadings of each gene are usually interpreted as “metagenes ” and the magnitude of modality-specific factors is constrained and regularized (Box 1). The factor loadings of each cell are used for clustering and cell matching. Matrix factorization methods assume that observed data are weighted linear functions of the latent decompositions but similar to above discussion of totalVI[39], more complex relationships can be modeled with neural networks. MAGAN[60] implements a type of neural network called dual Generative Adversarial Network (dual GAN) that uses a new architecture to map two datasets from different modalities reciprocally.

Obtaining a shared feature set by mapping between modalities can be challenging or even impractical when the measurements from each cell are vastly distinct. Rather than operating on a discrete set of observed data points, another approach is to consider modeling the entire “space” of data for each modality and map the spaces to each other. Manifold (Box 1) alignment and related methods assume that individual cells occupy some geometric subset of the feature space of a given modality. These geometric subsets have been called a “manifold” in the literature (with some abuse of the mathematical term). These manifolds can be thought as smooth curved surface that characterize a biologically feasible set of values for a given collection of cells. Manifold alignment methods assume that a shared latent structure (manifold) underlies each dataset and tries to learn a shared manifold among datasets to build correspondence between them. The approach is similar to linear latent variable models but with more generality.

Tools that implement manifold alignment include MATCHER[61], MMD-MA[62], and UnionCom[63]. These methods start with dimension reduction of the datasets. As an important first step, dimension reduction methods are chosen to be consistent with the model assumption and suitable to the data structure. MATCHER starts with the assumption that a one-dimensional structure exists along which all cells lie (this one-dimension can be interpreted as “pseudotime ”), MATCHER then fits a stochastic model] to infer a one-dimensional manifold structure (i.e., pseudotime) for each data modality. Subsequently, so-called monotonic warping function, is learned to match the two or more one-dimensional manifolds with pre-specified manifold orientation. Here, monotonic warping function means a function that associates two variables to each other that is strictly increasing or decreasing—i.e., order-preserving.). Schematically, MMD-MA maps geometric relationships within each modality feature space to a common space in a way that minimizes geometric distortions between each modality space while maintaining the intra-space configuration. UnionCom embeds each modality into a distance matrix that encapsulates a low-dimensional manifold for each modality. A well-defined pairwise distance matrix is sufficient to represent the complete geometric configuration of points. Thus, two matrices in UnionCom represent the estimated geometric relationships of the cells in each measurement modality. By optimizing a notion of difference between the two geometric configurations, the configurations of two modalities are matched and probabilistic cell correspondence between the two datasets are computed. Somewhat distinct from manifold alignment, SCOT[64] uses the notion of optimal transport, which tries to define a relationship between two sets of objects, each with a number of classes (e.g., cell types) and different frequency of objects in each class. The computed relationship takes into account both the frequency of objects in each class and a measure of distance between the objects.

Matching different modalities by aggregation is a natural idea but tends to lose individual cell resolution. Some approaches attempt to recover individual cell resolution through initial aggregated matching and then refinement, but the results from these approaches can be highly dependent on initial conditions. Matching by applying statistical models between the features of the different modalities can provide cell level resolution but this approach is highly dependent on the accuracy of the statistical models. Although a clear relationship exists between chromatin states and gene expression, the exact relationship, especially with respect to temporal dynamics is unclear. Matching by latent space or manifold alignment models are somewhat more principled approaches than those like aggregation and refinement, but the available models are complex and their interpretation in biological terms is often unclear. In sum, the available approaches have different strengths and weaknesses, and their utility is likely to be highly data and problem dependent.

Visualization of multiomics data

Computational visualization tools or interactive websites that allow user-friendly searches and display of features notably promote data sharing and reuse. Two large categories of data visualization exist in the context of single cell biology. One might be called “unbiased” visualization, and includes various dimension reduction approaches that attempt to display all data points. The other might be called “knowledge-driven” visualization, whereby certain curated aspects of the data (e.g., a focal subset of cells) are displayed. Although multiple tools have been developed for visualizing scRNA-seq data, tools for explicit visualization of single cell multiomics data are scarce. Below we provide a brief overview of current methods and discuss future directions for multi-modal single cell visualization.

Unbiased visualization

Dimension reduction and unbiased visualization has been critical for interpreting complex single cell data. The diverse cell types and states within a single cell dataset means that visualizing cells as a point in a two-dimensional or three-dimensional image is useful for evaluating data qualities, cell identities, developmental trajectories, and batch effects[65]. Various visualization methods have been implemented based on dimension reduction approaches, including tSNE, UMAP[66], PHATE[67], and force-directed graphs[68]. These methods extend from classic linear projection methods like Principal Component Analysis (PCA), which is based on projecting data points onto (orthogonal) directions of maximum variation, and embedding methods such as Multi-Dimensional Scaling (MDS). MDS embodies the general idea of computing one set of distance relationship in the original high-dimensions and then placing points in lower dimensions such that distance relationships in the lower dimensions are as similar to that of the original dimensions as possible. Variations of MDS involve different ways to define distances or measure the distortions between high and low dimensional distance relationships. The main problem faced by dimension reduction and visualization methods is that the configuration of points in a high dimension state cannot be represented in lower dimensions without error, and the methods therefore have to tradeoff the kinds of distortions that they allow. Typically, it is hard to uniformly spread out the distortions from smaller distances (e.g. within clusters) to those from larger distances (e.g., between clusters). These kinds of tradeoffs are determined by the approaches used to calculate distances and measure high-to-low distortions; typical options tradeoff accuracy at large distances for accuracy at smaller distances.

Methods such as tSNE, UMAP and PHATE add another twist to the dimension reduction approach by allowing inhomogeneous notions of distance **or** similarity. That is, a distance from point X to point Y might be different from that of point Y to point X. One interpretation of this approach is that the inhomogeneity in distances is related to curvature or (diffusion) velocity; thus, the distance of X to Y might be analogized to going uphill versus Y to X going downhill, or a particular region might have high curvature and is therefore hard to traverse. Modern methods of visualization also implement nonlinear notions of distance (**or** similarity) such that certain distances are emphasized whereas others are deemphasized, which often allows the resulting embeddings to highlight cluster relationships. These methods try to control the arbitrary freedom allowed by such flexibility by imposing user defined constraints (e.g., “perplexity” in tSNE (Box 1)). We caution that the high flexibility of these methods can complicate the interpretation of data. The visualizations can also be unstable, either because the algorithms start from random initial configurations or due to the sensitivity to the addition and subtraction of points. In-depth discussion of tools for the visualization of single cell data can be found elsewhere[69].

Unbiased visualization approaches naturally extend to multiomics single cell data as long as the above-described integration methods produce representation in a common space. Any of the available dimension reduction methods can be used to visualize integrated relationships within a common latent space, for instance, a shared gene expression space (by gene activity modeling) or a common layer in neural net. For example, the multiomics visualization arm of scAI[36], called VscAI, enables visualization of cells, genes, and (accessibility) loci by an embedding that reflects the low dimensional latent space. However, the nature of integrative analyses suggests the need for more complex multiple views of the data. For example, we might want to see single cells laid out in the common latent space and then also see their configuration in each of the measurement modalities, in particular with cell correspondences in each space. Although it is possible to switch views (as described elsewhere[44,6]), currently available methods do not easily show correspondence between layers. It would be desirable to have visualization systems similar to Geographic Information Systems (GIS), such as those used in landscape ecology[70], which have layers of multi-modal maps.

Knowledge-driven visualization

Single cell data is used by researchers to derive additional biological inferences —a process that is often called down-stream analysis. These downstream analyses result in the production of additional visual objects. Common examples of these visual objects include violin plots for visualizing cell type marker genes, di-graphs to visualize cellular interactions, or even simple annotation overlay to visualize a focal subset of cells. Other visual devices that focus on particular knowledge-driven assumptions include displays of motif enrichment along with the expression of corresponding transcription factors[14], visualization of sequence reads along genomic tracks[71], and other associated annotation data organized by genomic coordinates[71,72]. One important approach to incorporate existing knowledge for single cell data is to associate spatio-temporal information with single cell visualization. Temporal trajectories have been visualized using many different pseudotime methods; for example, the RNA velocity method[15] displays estimated displacement vectors to extrapolate the “flow” of cell differentiation states. Approaches for the visualization of single cell data in the context of anatomical ontology (e.g., KidneyCellExplorer[6]) or within detailed 3D models (e.g., NIH HuBMAP project[73]) is under development.

Future directions for data visualization

Additional visualization tools and frameworks are needed to fully appreciate the complexity of multimodal data (Figure 3). Visualization tools with greater flexibility to enable the display of multiple and coordinated views that link objects in various modalities will aid visual explorations of multi-modal relationships. However, even multiple layers of data visualization will be insufficient to fully explore the biological structure of multimodal data if the visualizations are static. Complex data are best explored with interactive systems that enable dynamic modifications of views, such as the ability to re-display subsets of data or dynamically switch between different modalities. One critical consideration is the computational speed required for such interactive visualizations and analyses, especially for very large datasets (e.g., those with 10^6 cell datasets[4,5]). As datasets scale to extremely large sizes, issues of where to store and compute the views — for example, in the cloud or on a client computer —become non-trivial.

Viscello[44], Cerebro[74], VscAI[36], and Giotto[75] are some of the tools that currently allow some degree of interactive multimodal single cell data visualization. Some consortia including ReBuilding a Kidney (RBK; https://www.rebuildingakidney.org/) and GenitoUrinary Development Molecular Anatomy Project (GUDMAP;gudmap.org/) integrate interactive single cell visualizers in their data archive. However, these tools are not fully interactive in the sense that they cannot recompute the visualization to an arbitrary choice of views or subsets of data.

Challenges for single cell multimodal data integration

We reviewed some of the existing approaches to data integration for single cell multi-modal data but our review only touches the surface of the very active on-going research in this area. For all of the approaches, there are some common challenges to be considered. These challenges can originate from the process of data collection, data conversion, and data interpretation. Here we discuss some of the most prominent challenges to single cell multi-modal data integration.

Accounting for data characteristics

It is well-acknowledged that single cell data are noisy. This noise arises from biological and technical variation. Common biological variation includes the stochastic bursting of genes[75], variation arising from circadian rhythm[77] and cell cycling[78], and variation arising from local cell environment. The contribution of technical variations is debated, but may include uneven dropouts and coverage[79,25], transcript contamination (from ambient RNA)[80], and multiplets [81]. In general, single cell assays, especially high-throughput assays, tend to be lossy because the technologies tradeoff sensitivity (e.g., efficient capture of the RNA molecules in a cell) for throughput, resulting in sparse datasets. This sparsity is a huge challenge and is typically approached by “borrowing” local information from nearby cells, which can introduce additional biases. Multi-omics approaches have the potential to resolve some confounding factors, sparsity, or noise in a single modality by ‘borrowing’ information in the other modality, but this integration does not always improve the prediction power achieved by a single modality[82]. Noise across multiple layers can be amplified, leading to a decrease in the signal strength[83]. An important problem in single cell analysis is that commonly used noise models typically use off-the-shelf parametric models, such as Poisson zero inflation models and negative binomial models[84,85], whereas in practice, single cell noise does not seem to be well modeled by these parametric models and systematic control experiments to measure the characteristics of the noise have been rare[79,86].

Although models have been built to distinguish biological and technical variation in scRNA-seq data[87], models to account for heterogenous noise across multiple modalities still need to be developed. In some cases, the problem of heterogeneous noise is handled best by “early integration”, whereby the input datasets themselves are operated on to make a single compatible matrix; for example, by applying weights and concatenating the datasets. In other cases, the problem is best approached by “intermediate integration”; for example, using the latent space methods approach to map the input data to theoretical common space features. In still other cases, “late integration” might be the best approach, whereby each modality is used to infer a model, such as a gene regulatory network, and then the inferred models are combined appropriately (e.g., using CiteFuse[41]). Each of these approaches have pros and cons depending on the modalities being integrated and other conditions of measurement (e.g., batch effects). Earlier integration might help increase the power of ultimate downstream analysis (e.g., the identification of cell clusters) by both increasing the size of the dataset and by bringing together (possibly) complementary information. Late integration can help the application of modality-specific models and methods to handle heterogeneous noise, and enable the individual inferences (e.g., clusters from each modality) to be combined to obtain a more robust inference.

Data types and cell composition compatibility

Although desirable to integrate information from all relevant sources, datasets that are to be integrated can be vastly distinct. At a simple level, gene expression profiles in scRNA-seq data are continuous variables whereas chromatin accessibility measurements are usually binarized to indicator variables[88] (also known as dummy variables). This integration of distinct datasets requires a consistent way to match metric variables with nominal variables, which can have both technical and conceptual challenges[89].

At a more complex level, traits such as cell morphology, while having a metric representation, are difficult to statistically characterize in a meaningful manner. The emergence of machine learning methods has led to the development of approaches to integrate morphology and expression data. Fascinating insights from these studies suggest that cell morphology might predict gene expression[90], but the functional connections of such relationships are still unclear. Another more common but important challenge is that subtypes of cells that are recovered and measured with high-throughput single cell methods can be very different for different measurement modalities. For example, immune cell populations are usually over-represented in scRNA-seq datasets, likely as a result of recovery bias, whereas snRNA-seq methodologies demonstrate bias in their recovery of different subpopulations[91]. Such differences in cell subtype distribution can complicate data matching, especially for nearest neighbor-based methods.

Computing millions of data points

With the development of combinatorial indexing technologies[92] and sample multiplexing strategies[93], datasets are now available at 10^6 scale[4,5]. Efficient computing over such big data matrices requires different strategies to those used for smaller datasets. We note that just to compute pairwise relationships for a dataset of size 10^6, ~10^{12} computations need to be considered. This scale of computing is prohibitive, resulting in the use of less intensive heuristic methods. In fact, even just laying out a million points for data visualization becomes a heavy computational burden and prevents researchers from exploring different views due to the wait time involved. Future integrative analyses of single cell data will require concerted efforts in algorithm development with incorporation of novel stochastic indexing strategies, streaming of algorithms, and careful heuristics, along with the development of carefully tuned high-performance codes. Some areas of computational biology such as phylogenetics and protein folding have long been acutely limited by computational speed, and advanced algorithmics have been an inherent part of those fields. We suspect single cell biology will soon demand similar levels of algorithm sophistication and high performance software engineering.

Modality Mapping

As discussed earlier, the integration of unmatched measurements is often achieved by mapping the values of one modality to the values of another — a key example is the conversion of chromatin states to gene expression values. However, such conversions assume an over-simplified model between different modalities, mostly due to a lack of knowledge of whole-genome gene regulatory logic. As previously reported[29], the temporal dynamics of the open chromatin states of a cell are not at the same phase as its corresponding RNA expression; rather, gene expression lags behind the opening of its proximal chromatin. Thus, accurate mapping between the modalities requires both a precise knowledge of the mechanisms connecting the measured molecules and the temporal dynamics of the mechanisms. Similar consideration would apply when mapping between the transcriptome and the proteome or, a more complicated scenario, the connection between molecular and morphological states.

Interpretability and Validation

Most data integration methods avoid detailed causal modeling. At the extreme end are purely data-driven machine learning methods, such as autoencoders (Box 1). For example, one autoencoder-based multiomics data integration method[94] has been trained to create a common latent space for many different modalities. Powerful computational tools such as this can indeed integrate multiple data types, automatically and regardless of the difficulties of comprehensive causal modeling. In a sense, machine learning methods completely avoid the careful modeling of mechanisms and instead apply a generically complex model to a very large reference dataset to produce a well-performing model with unknown parts. Thus, interpreting the details of a machine learning method in terms of biological correspondents is difficult. More importantly, training of complex machine learning models typically requires very large volumes of data. On the positive side, developments in high-throughput multiomics technologies promises the availability of such training data. On the negative side, for the models to be generalizable we need more than just replicate numbers but also large amounts of data across varying conditions, such as from different cell and tissue types. Until a mechanistic model of a cell with sufficient precision to enable integration of data under a causal model is available, both the utility and validation of any integration method must be evaluated in terms of their application; for example, by the recovery of the identities of biologically plausible cell types.

Conclusions and future directions

Integration of single cell multiomics data has been implemented in many real data analyses, revealing new biological insights. For example, multiomics integration has identified the presence of a pro-inflammatory, “failed repair proximal tubule cell” state in apparently healthy human kidneys[13,95]; it also facilitated the prioritization of GWAS loci through the identification of methylation and gene expression changes that are likely to mediate development of diabetic kidney disease[96], and has helped identify mechanisms of myofibroblast activation in CKD[8].

Ideally, the process of generating data integration models and evaluating the models should itself shed light on mechanisms of biological processes such as gene regulation. For example, cell identity is traditionally defined by the abundance of specific RNAs or proteins, but integration of these data with other omics datasets could effectively broaden the definition of cell types to other chemical–physical modalities of the cell. In addition, novel relationship across data modalities can be studied with multiomics data integration. For instance, correlating DNA methylation with gene expression in cis might reveal differential functional impact of methylation of different DNA elements (promoters or gene body). However, regardless of their utility in the modelling of biological processes, data integration often yields more or better resolved inferences than analyses of single datasets alone. For example, the addition of scATAC-seq to scRNA-seq data better distinguishes different segments of proximal tubules in the kidney[36] than does scRNA-seq data alone. Integrated data analyses can also identify underappreciated relationships that might lead to additional applications, such as drug target discovery or better causal SNP inference.

Currently available computational methods have generally followed the development of the measurements themselves. The number of available methods that attempt to integrate unmatched data far outweighs the number of methods that attempt to integrate matched data simply because multiomic measurements have only become widely available in the past 2 years. Methods for integrating cell morphologies[97], perturbations[98], spatial micro-environment[99,52], and subcellular measurements[100] (e.g., of organelles), are sparse, as are the corresponding data. However, we expect that methods to integrate these data will rapidly follow the availability of such data. In addition, most current computational methods are built to integrate two modalities; however, with the development of experimental methods that jointly profile three or more modalities, more flexible computational algorithms will be required.

Amongst the computational tools that lag behind the analytic methods are methods for visualization of complex multimodal data that interactively connect between different views and ancillary information. Some of the barriers to the development of these tools are the speed and capacity of the computers themselves. Approaches to enable the interactive visualization of extremely large volumes of data in the single cell field is non-trivial and may eventually require dedicated hardware.

As discussed above, one ideal way to integrate data is in terms of a causal model between the quantitative data and the underlying molecular processes, such as cell differentiation, physiology, and homeostasis. Conversely, we would hope that multimodal data, by providing measurements from multiple aspects of the biology of an organism could aid the development of such causal models. The era of multiomic single cell biology at the scale of millions of cells is just starting and we have no doubt that the data, analytical methods, and inferred models will advance our understanding of the kidney by leaps and bounds in years to come.

Acknowledgements

This work was supported in part by UC2DK126024 grant to JK, BDH, and APM and also by a Health Research Formula Fund of the Commonwealth of Pennsylvania, which did not play a direct role in the work.

Glossary

Assay for Transposase-Accessible Chromatin using sequencing

(ATAC-seq). A technique that profiles the accessibility of DNA elements based on the principle that the Tn5 transposase can insert a transposon only at accessible parts of the chromosome. The insertion location is identified through DNA sequencing.

Cis-regulatory elements

DNA elements proximal to a gene that are required for controlling gene expression. Such elements usually include promoters and enhancers, and often contain transcription factor binding sites.

Features and feature space

In machine learning, measured variables are often called features and the set of features comprise a feature space.

Molecule recovery efficiency

Single cell assays capture molecules, such as mRNAs or transposon-interrupted DNA fragments, and amplifies them for readout. Different protocols recover a given pool of molecules with different efficiencies e.g. a single podocyte might have 300,000 mRNA molecules and an RNA-seq protocol with a 10% recovery efficiency would recover ~30,000 of these.

Joint snRNA-seq and snATAC-seq

scRNA-seq attempts to recover RNA from the whole cell whereas snRNA-seq only isolates the nuclear fraction of the RNA; the two transcriptomes are related but different. Multiomics methods involving ATAC-seq and RNA-seq typically isolate the nucleus first resulting in snRNA-seq and snATAC-seq.

Sequential fluorescence in situ hybridization (seqFISH)

A technique that measures mRNA quantity through sequential fluorescent probes that have combinatorially encoded information for each targeted mRNA. For example, sequential signal from a spot of probe A then B might encode gene X while probe A then C might encode gene Y.

Read depth

Given a genomic region, say transcribed region, a quantity that measures, the number of times that sequencing reads cover that region. The region of interest may be a base pair or an entire transcribed region.

Canonical correlation analysis

A multi-variate statistical technique that computes correlation between two sets of variables, say X and Y. Canonical correlation analysis finds the linear combination of X and linear combination of Y that maximizes correlation.

Nonnegative Matrix factorization

A group of algorithms that decompose one matrix into a product of two (or more) matrices such that the elements in each matrix is nonnegative. Typically, each matrix has a model interpretation; e.g., a data matrix factorizes the matrix into one representing latent space features and another representing latent space features to cells.

Factorize (explained above with NNMF)

Metagene

A metagene is some (mathematical) function of a group of genes (e.g., linear combination), often relating some shared properties. For example, methods like NNMF compute matrices as the product between a gene-by-metagene matrix and a metagene-by-cell matrix.

Dimension reduction

A data transformation method that reduce the number of dimensions in the original feature space to a lower-dimensional (usually much lower than the original one) space while certain properties (e.g., the distance measures between observations) of the original data are preserved.

Pseudotime

In contrast to real time, pseudotime represents computationally inferred temporal stages of a collection of cells.

Principal Component Analysis

A common dimension reduction method that aims to project the original data to a fixed smaller dimension while minimizing the squared error during data reduction. Equivalently, this can be viewed as maximizing the variance in the projected data.

Embedding

In mathematics, embedding is a map from one set X to another set Y, where some characteristic of X is preserved. In single cell studies, the term embedding has been used for methods that “place” cells in a new feature space, possibly of lower dimension, such that notions of cell-to-cell distances are approximately preserved.

Ambient RNA

In droplet-based single cell RNA-seq approach, the measured mRNA molecules could be contaminated by mRNAs from other cells present in the suspension, say due to ruptured cells. These contaminating mRNAs are termed ambient RNA.

Multiplets

During high-throughput single cell (or single nuclei) isolation in droplets or similar vessels, two or more cells may be captured together creating a mixture of molecules. Computational methods have been developed to detect and remove such unwanted observations from the dataset.

High-performance codes

In programing there are many different ways to achieve the same computation. Some algorithms are inherently faster than others. For the same algorithm, programs can also be written differently to speed up the execution by careful use of hardware resources. High-performance codes try to use the fastest algorithms and fine tune the programs for optimal speed.

Dropouts

In single cell biology, dropout is usually referred to as the transcripts that are present in the cell but not captured during sequencing.

References:

1. Richardson S, Tseng GC & Sun W Statistical Methods in Integrative Genomics. Annu. Rev. Stat. Its Appl. 3, 181–209 (2016).
2. Yuan G-C et al. Challenges and emerging directions in single-cell analysis. Genome Biol. 18, 84 (2017). [PubMed: 28482897]
3. Eberwine J, Sul J-Y, Bartfai T & Kim J The promise of single-cell sequencing. Nat. Methods 11, 25–27 (2014). [PubMed: 24524134]
4. Yao Z et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. bioRxiv 2020.03.30.015214 (2020) doi:10.1101/2020.03.30.015214.
5. Cao J et al. A human cell atlas of fetal gene expression. Science 370, (2020).
6. Ransick A et al. Single-Cell Profiling Reveals Sex, Lineage, and Regional Diversity in the Mouse Kidney. Dev. Cell 51, 399–413.e7 (2019). [PubMed: 31689386] A comprehensive kidney single cell RNA-seq atlas with a visualization tool “Kidney Cell Explorer”
7. Kirita Y, Wu H, Uchimura K, Wilson PC & Humphreys BD Cell profiling of mouse acute kidney injury reveals conserved cellular responses to injury. Proc. Natl. Acad. Sci. 117, 15874–15883 (2020). [PubMed: 32571916]
8. Kuppe C et al. Decoding myofibroblast origins in human kidney fibrosis. Nature 1–9 (2020) doi:10.1038/s41586-020-2941-1.
9. Gerhardt MSL, et al. Single-nuclear transcriptomics reveals diversity of proximal tubule cell states in a dynamic response to acute kidney injury. Proc. Natl. Acad. Sci. In Press (2021)
10. Ma A, McDermaid A, Xu J, Chang Y & Ma Q Integrative Methods and Practical Challenges for Single-Cell Multi-omics. Trends Biotechnol. S0167779920300573 (2020) doi:10.1016/j.tibtech.2020.02.013. A comprehensive review of single cell multiomics technologies
11. Lee J, Hyeon DY & Hwang D Single-cell multiomics: technologies and data analysis methods. Exp. Mol. Med. 52, 1428–1442 (2020). [PubMed: 32929225]
12. Sullivan KM & Susztak K Unravelling the complex genetics of common kidney diseases: from variants to mechanisms. Nat. Rev. Nephrol. 16, 628–640 (2020). [PubMed: 32514149] A up-to-date review on efforts to gain further understanding of kidney disease associated GWAS variants
13. Muto Y et al. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. Nat. Commun. 12, 2190 (2021). [PubMed: 33850129]
14. Miao Zhen et al. Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. Nat. Commun. doi:10.1038/s41467-021-222266-1.

15. La Manno G et al. RNA velocity of single cells. Nature 560, 494–498 (2018). [PubMed: 30089906]

16. Gorin G, Svensson V & Pachter L Protein velocity and acceleration from single-cell multiomics experiments. Genome Biol. 21, 39 (2020). [PubMed: 32070398]

17. Stoeckius M et al. Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods 14, 865–868 (2017). [PubMed: 28759029]

18. Peterson VM et al. Multiplexed quantification of proteins and transcripts in single cells. Nat. Biotechnol. 35, 936–939 (2017). [PubMed: 28854175]

19. Zhou Z, Ye C, Wang J & Zhang NR Surface protein imputation from single cell transcriptomes by deep neural networks. Nat. Commun. 11, 651 (2020). [PubMed: 32005835]

20. Pliner HA et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol. Cell 71, 858–871.e8 (2018). [PubMed: 30078726]

21. Serra A, Fratello M, Greco D, & Tagliaferri R. Data integration in genomics and systems biology. in 2016 IEEE Congress on Evolutionary Computation (CEC) 1272–1279 (2016). doi:10.1109/CEC.2016.7743934.

22. Hasin Y, Seldin M & Lusis A Multi-omics approaches to disease. Genome Biol. 18, 83 (2017). [PubMed: 28476144]

23. Liu L et al. Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. Nat. Commun. 10, 470 (2019). [PubMed: 30692544]

24. Dueck H et al. Deep sequencing reveals cell-type-specific patterns of single-cell transcriptome variation. Genome Biol. 16, 122 (2015). [PubMed: 26056000]

25. Dueck HR et al. Assessing characteristics of RNA amplification methods for single cell RNA sequencing. BMC Genomics 17, 966 (2016). [PubMed: 27881084]

26. Cao J et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science 361, 1380–1385 (2018). [PubMed: 30166440]

27. Chen S, Lake BB & Zhang K High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat. Biotechnol. 37, 1452–1457 (2019). [PubMed: 31611697]

28. Zhu C et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. Nat. Struct. Mol. Biol. 26, 1063–1070 (2019). [PubMed: 31695190]

29. Ma S et al. Chromatin potential identified by shared single cell profiling of RNA and chromatin. 10.1101/2020.06.17.156943 (2020) doi:10.1101/2020.06.17.156943.

30. Han SH, Choi Y, Kim J & Lee D Photoactivated Selective Release of Droplets from Microwell Arrays. ACS Appl. Mater. Interfaces 12, 3936–3944 (2020). [PubMed: 31912738]

31. Stuart T & Satija R Integrative single-cell analysis. Nat. Rev. Genet. 20, 257–272 (2019). [PubMed: 30696980]

32. Li Y, Ma L, Wu D & Chen G Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine. Brief. Bioinform. (2021) doi:10.1093/bib/bbab024.

33. Sokal RR Distance as a Measure of Taxonomic Similarity. Syst. Biol. 10, 70–79 (1961).

34. Sneath PHA, Sneath PHA, Sokal RR & Sokal URR Numerical Taxonomy: The Principles and Practice of Numerical Classification. (W. H. Freeman, 1973).

35. Wang X et al. BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data. Nucleic Acids Res. 48, 5814–5824 (2020). [PubMed: 32379315]

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

37. Argelaguet R et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. 14, (2018).

38. Argelaguet R et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 21, 111 (2020). [PubMed: 32393329]

39. Gayoso A et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat. Methods 18, 272–282 (2021). [PubMed: 33589839]

40. Hao Y et al. Integrated analysis of multimodal single-cell data. Cell (2021) doi:10.1016/j.cell.2021.04.048.

41. Wang B et al. Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods 11, 333–337 (2014). [PubMed: 24464287] Introduced the SNF model which is widely applied in multiomics integration

42. Kim HJ, Lin Y, Geddes TA, Yang JYH & Yang P CiteFuse enables multi-modal analysis of CITE-seq data. Bioinformatics 36, 4137–4143 (2020). [PubMed: 32353146]

43. Han X et al. Construction of a human cell landscape at single-cell level. Nature 581, 303–309 (2020). [PubMed: 32214235]

44. Packer JS et al. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. Science 365, eaax1971 (2019). [PubMed: 31488706] A single cell atlas of C. elegans with the visualization tool visCello

45. Cao J et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature 566, 496–502 (2019). [PubMed: 30787437]

46. Slavov N Single-cell protein analysis by mass spectrometry. Curr. Opin. Chem. Biol. 60, 1–9 (2021). [PubMed: 32599342]

47. Neumann EK, Ellis JF, Triplett AE, Rubakhin SS & Sweedler JV Lipid Analysis of 30 000 Individual Rodent Cerebellar Cells Using High-Resolution Mass Spectrometry. Anal. Chem. 91, 7871–7878 (2019). [PubMed: 31122012]

48. Zhu Q et al. Developmental trajectory of prehematopoietic stem cell formation from endothelium. Blood 136, 845–856 (2020). [PubMed: 32392346]

49. Wang C et al. Integrative analyses of single-cell transcriptome and regulome using MAESTRO. Genome Biol. 21, 198 (2020). [PubMed: 32767996]

50. Asp M, Bergenstråhle J & Lundeberg J Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration. BioEssays 42, 1900221 (2020).

51. Zhu Q, Shah S, Dries R, Cai L & Yuan G-C Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nat. Biotechnol. 36, 1183–1190 (2018).

52. Rodriques SG et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science 363, 1463 (2019). [PubMed: 30923225]

53. Andersson A et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. Commun. Biol. 3, 1–8 (2020). [PubMed: 31925316]

54. Govek KW et al. Single-cell transcriptomic analysis of mIHC images via antigen mapping. Sci. Adv. 7, eabc5464 (2021). [PubMed: 33674303]

55. Haghverdi L, Lun ATL, Morgan MD & Marioni JC Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. 36, 421–427 (2018). [PubMed: 29608177] Introduced the mutual nearest neighbor methods that become a popular method in the single cell biology with multiple applications

56. Campbell KR et al. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. Genome Biol. 20, 54 (2019). [PubMed: 30866997]

57. Stuart T et al. Comprehensive Integration of Single-Cell Data. Cell 177, 1888–1902.e21 (2019). [PubMed: 31178118]

58. Welch JD et al. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell 177, 1873–1887.e17 (2019). [PubMed: 31178122]

59. Yang Z & Michailidis G A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinforma. Oxf. Engl. 32, 1–8 (2016).

60. Amodio M & Krishnaswamy S MAGAN: Aligning Biological Manifolds. 9.

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

62. Liu J, Huang Y, Singh R, Vert J-P & Noble WS Jointly Embedding Multiple Single-Cell Omics Measurements. in 19th International Workshop on Algorithms in Bioinformatics (WABI 2019) (eds. Huber KT & Gusfield D) vol. 143 10:1–10:13 (Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019).

63. Cao K, Bai X, Hong Y & Wan L Unsupervised topological alignment for single-cell multi-omics integration. Bioinformatics 36, i48–i56 (2020). [PubMed: 32657382]

64. Demetci P, Santorella R, Sandstede B, Noble WS & Singh R Gromov-Wasserstein optimal transport to align single-cell multi-omics data. 10.1101/2020.04.28.066787 (2020) doi:10.1101/2020.04.28.066787.

65. Li X et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat. Commun. 11, 2338 (2020). [PubMed: 32393754]

66. McInnes L, Healy J & Melville J UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat (2020).

67. Moon KR et al. Visualizing structure and transitions in high-dimensional biological data. Nat. Biotechnol. 37, 1482–1492 (2019). [PubMed: 31796933]

68. Costa F, Grün D & Backofen R GraphDDP: a graph-embedding approach to detect differentiation pathways in single-cell-data using prior class knowledge. Nat. Commun. 9, 3685 (2018). [PubMed: 30206223]

69. Wu Y & Zhang K Tools for the analysis of high-dimensional single-cell RNA sequencing data. Nat. Rev. Nephrol. 16, 408–421 (2020). [PubMed: 32221477] A comprehensive review of single cell RNA-seq data analysis pipelines and computational tools

70. Steiniger S & Hay GJ Free and open source geographic information tools for landscape ecology. Ecol. Inform. 4, 183–195 (2009).

71. Raney BJ et al. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics 30, 1003–1005 (2014). [PubMed: 24227676]

72. Ou J & Zhu LJ trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data. Nat. Methods 16, 453–454 (2019). [PubMed: 31133757]

73. Snyder MP et al. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature 574, 187–192 (2019). [PubMed: 31597973]

74. Hillje R, Pelicci PG & Luzi L Cerebro: interactive visualization of scRNA-seq data. Bioinformatics 36, 2311–2313 (2020). [PubMed: 31764967]

75. Dries R et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome Biol. 22, 78 (2021). [PubMed: 33685491]

76. Larsson AJM et al. Genomic encoding of transcriptional burst kinetics. Nature 565, 251–254 (2019). [PubMed: 30602787]

77. Chakrabarti S et al. Hidden heterogeneity and circadian-controlled cell fate inferred from single cell lineages. Nat. Commun. 9, 5372 (2018). [PubMed: 30560953]

78. Zhong L et al. Single cell transcriptomics identifies a unique adipose lineage cell population that regulates bone marrow environment. eLife 9, e54695 (2020). [PubMed: 32286228]

79. Lahens NF et al. IVT-seq reveals extreme bias in RNA sequencing. Genome Biol. 15, R86 (2014). [PubMed: 24981968]

80. Marquina-Sanchez B et al. Single-cell RNA-seq with spike-in cells enables accurate quantification of cell-specific drug effects in pancreatic islets. Genome Biol. 21, 106 (2020). [PubMed: 32375897]

81. Xi NM & Li JJ Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. Cell Syst. 12, 176–194.e6 (2021). [PubMed: 33338399]

82. Franzosa EA et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat. Microbiol. 4, 293–305 (2019). [PubMed: 30531976]

83. Tini G, Marchetti L, Priami C & Scott-Boyer M-P Multi-omics integration—a comparison of unsupervised clustering methodologies. Brief. Bioinform. 20, 1269–1279 (2019). [PubMed: 29272335]

84. Pierson E & Yau C ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 16, 241 (2015). [PubMed: 26527291]

85. Kharchenko PV, Silberstein L & Scadden DT Bayesian approach to single-cell differential expression analysis. Nat. Methods 11, 740–742 (2014). [PubMed: 24836921]

86. Marinov GK et al. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. Genome Res. 24, 496–510 (2014). [PubMed: 24299736]

87. Zhang L & Nie Q scMC learns biological variation through the alignment of multiple single-cell genomics datasets. Genome Biol. 22, 10 (2021). [PubMed: 33397454]

88. Fang R et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nat. Commun. 12, 1337 (2021). [PubMed: 33637727]

89. Velleman PF & Wilkinson L Nominal, Ordinal, Interval, and Ratio Typologies are Misleading. Am. Stat. 47, 65–72 (1993).

90. He B et al. Integrating spatial gene expression and breast tumour morphology via deep learning. Nat. Biomed. Eng. 4, 827–834 (2020). [PubMed: 32572199]

91. Wu H, Kirita Y, Donnelly EL & Humphreys BD Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. J. Am. Soc. Nephrol. 30, 23 (2019). [PubMed: 30510133]

92. Cao J et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science 357, 661–667 (2017). [PubMed: 28818938]

93. McGinnis CS et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat. Methods 16, 619–626 (2019). [PubMed: 31209384]

94. Yang KD et al. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. Nat. Commun. 12, 31 (2021). [PubMed: 33397893]

95. Dhillon P et al. The Nuclear Receptor ESRRA Protects from Kidney Disease by Coupling Metabolism and Differentiation. Cell Metab. 33, 379–394.e8 (2021). [PubMed: 33301705]

96. Sheng X et al. Systematic integrated analysis of genetic and epigenetic variation in diabetic kidney disease. Proc. Natl. Acad. Sci. 117, 29013 (2020). [PubMed: 33144501]

97. Wu P-H et al. Single-cell morphology encodes metastatic potential. Sci. Adv. 6, eaaw6938 (2020). [PubMed: 32010778]

98. Dixit A et al. Perturb-seq: Dissecting molecular circuits with scalable single cell RNA profiling of pooled genetic screens. Cell 167, 1853–1866.e17 (2016). [PubMed: 27984732]

99. Lindström NO et al. Spatial Transcriptional Mapping of the Human Nephrogenic Program. bioRxiv 2020.04.27.060749 (2020) doi:10.1101/2020.04.27.060749.

100. Khaladkar M et al. Subcellular RNA Sequencing Reveals Broad Presence of Cytoplasmic Intron-Sequence Retaining Transcripts in Mouse and Rat Neurons. PLOS ONE 8, e76194 (2013). [PubMed: 24098440] The first subcellular RNA sequencing methods

Box 1: Computational terminology

Model:

The term “model” is fairly generic. Here, we use the term ‘model’ in two different senses. In the first use, a model is a set of quantitative causal descriptions of biological processes, often abstracted to a simple form. An example would be a differential equation that describes RNA levels as a function of the rates of transcription, export and degradation. A second use of the term ‘model’ is to describe statistical models that relate measurements to each other; e.g., a “linear model” that relates latent space variables to observed variables as a linear mathematical function. This class of models might include more biology motivated models such as a gene activity model that posits a statistical relationship between the number of cis open chromatin regions and levels of gene expression.

Machine learning:

Machine Learning (ML) is a family of computational models that tries to associate a set of input features to a set of output features. Typically, output features are discrete labels such as “proximal tubule cells” or “podocytes”. ML methods separate into “supervised” methods and “unsupervised” methods. In supervised methods, some observations of “true” label assignment is known; e.g., input features might be gene expressions and true cell-type labels are available for some cells. Such ground-truth data are called “training data”. ML methods try to tune (learn) various mathematical functions to find the association between input features and the training data’s known output features. In unsupervised methods, training data is not available input features are available only for some observations. The typical goal of unsupervised methods is classify the input observations into groups (e.g., clusters) to reveal their grouping patterns.

Neural Networks and Deep Learning:

A Neural Network (NN), sometimes called Artificial Neural Network (ANN) to distinguish from biological brains, is a subset of machine learning methodologies t motivated by the modelling of a biological brain. The basic idea is to associate input features to output features using a set of mathematical functions called “nodes”. A node generates output values as a function of all the input values. Thus, a node emulates the metaphor of a neuron integrating all the synaptic input to an axonal output firing. Multiple nodes can be applied to the input features, each of which generates values, resulting in a set of values that can be treated as input features to another set of nodes. Each set of nodes used in this manner is called a “layer.” The complexity of the ANN can depend on the number of nodes in each layer, the number of layers, the input-output relationships between nodes, and the type of mathematical function in each node. Deep Learning (DL) is a non-technical term to refer to the development of methods that have very large number of nodes and layers.

Regularization:

Many statistical models can be complicated and overfit the data. For example, in the popular tSNE data visualization method, each data point has its own scale of distance, which can make pairwise relationships arbitrary. A common technique to prevent overfitting is to add some additional constraint, for example, a penalty for high model complexity, to prevent the model from being degenerate. For example, with tSNE a constraint called “perplexity” is introduced that constrains the observed data relationship to a certain pairwise distribution. The class of techniques to constrain the model complexity is called regularization. Regularization methods typically have a tunable parameter that controls how much regularization constraint is applied.

Manifold:

In mathematics, a manifold is a smooth topological space that locally resembles Euclidean space (i.e., space where distances between points can be defined as square root of sum of coordinate differences). In single cell studies, the term “manifold” refers to the idea that ensembles of the cells may lie in a lower dimension of the measurement space, which may have non-linear characteristics such as curvature and local folds.

Kernel functions:

A kernel function is a mathematical function that can be used to generalize the notion of distance between two objects (points). For example, the standard Euclidean distance (see Manifold) can be derived from a particular kernel function, the “dot product”. In machine learning, different kinds of kernel functions are used to change the pairwise relationship of objects, in a sense changing the geometric configuration of objects.

Loss function:

In machine learning, loss functions are functions that need to be optimized to obtain the desired performance given the data and model. For example, in the least squares’ regression model, the loss function is the sum of the squared error; and in Lasso regression, the loss function is the sum of squared error with regularization of regression coefficients. The design of loss function is key to a successful machine learning model.

Encoder-Decoder:

A commonly used architecture in machine learning where a neural net is constructed with a set of nodes that map the input to a middle layer (encoder) and another set of nodes, usually the inverse of the encoder architecture, that maps the middle layer to an output (decoder). The middle layer typically is simpler than the input, for example, with lower dimensions, and tends to encapsulate an abstract characteristic of the input dataset. The decoder then attempts to map this abstracted representation back to some observable data. In an autoencoder, the decoder tries recapitulate the input data. If successful, the middle layer is thought to represent the essential characteristics of the input data.

Key Points:

- With the development of single cell multiomics techniques, tools and models for data integration are critically important
- Integration problems in single cell biology can be divided into those associated with the integration of matched and unmatched data
- Strategies for integrating matched data include joint latent space inference, consensus of individual inferences, and biological causal modeling
- Strategies for integrating unmatched data include annotated group matching, matching with common features, and aligning spaces
- Visualization methods for integrated multi-modal single cell data are still underdeveloped
- Future challenges include accounting for specific noise related to each modality, overcoming the need for computing efficiency, and developing biologically interpretable integration strategies.

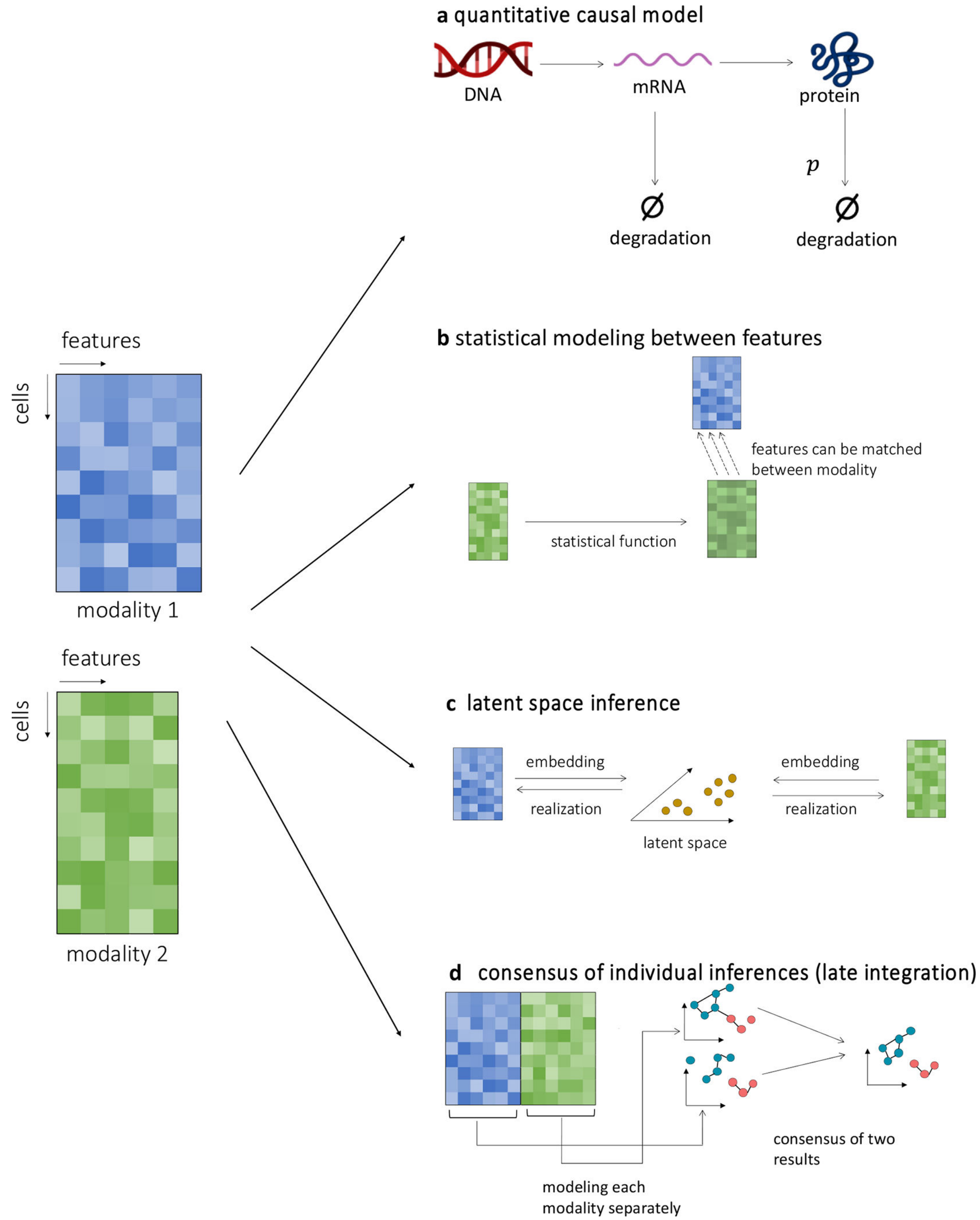

Figure 1. Frameworks for the integration of single cell multiomics data

Computational methods enable the integration of measured attributes (that is, features) obtained using multiomics approaches (for example, transcriptome and protein data) from single cells. These methods can be classified into four broad categories.

(a) Integration based on quantitative causal models. For example, the rates of RNA synthesis, splicing, translation and degradation might be modelled by differential equations and single cell multiomics data (for example, gene and protein expression data) can be used to estimate parameters (*p*) in the model. After obtaining the parameters, current and future cell states can be inferred.

(b) Statistical modeling between features. A statistical function is used to associate data in one modality to another modality, such that the two sets of features (again, for example, gene or protein expression data) can be harmonized into one modality for downstream analyses. Such models can be calibrated from reference datasets or potentially fit to the dataset of interest.

(c) Latent space modeling. Data from different modalities are assumed to be generated from a common latent space, and integrated based on the assumption that specific mapping functions are able to map the common latent space onto different modalities. The latent space can be viewed as an integrated low dimensional embedding of the multiomics or multi-modal data and the mapping functions can be regarded as a model of the abstract latent space to real observations.

(d) Consensus of individual inferences (late integration). Analyses (such as clustering or dimension reduction) are performed for each individual data modality after which the results are combined to obtain common consensus outputs or complementary evidence.

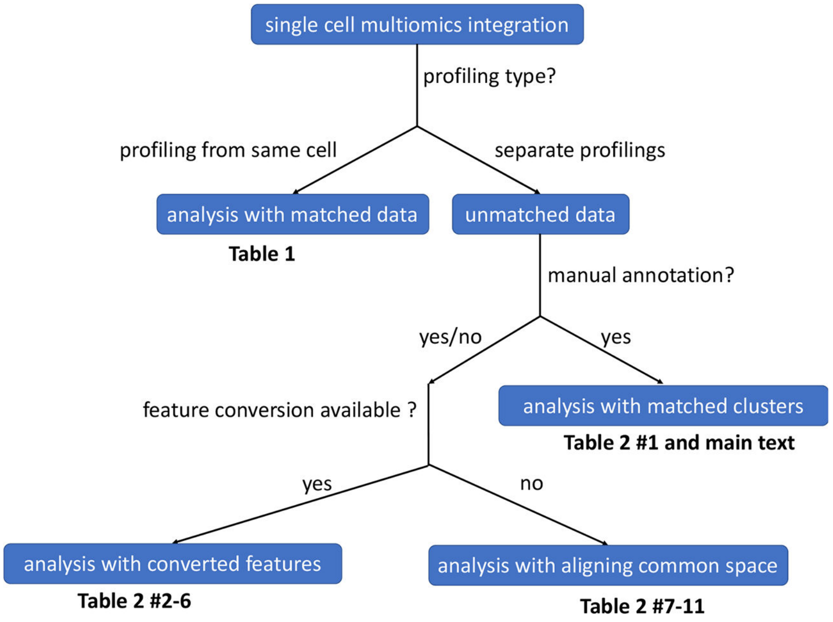
single cell multiomics integration
profiling type?
profiling from same cell
separate profilings
analysis with matched data
Table 1
unmatched data
manual annotation?
yes/no
yes
feature conversion available ?
analysis with matched clusters
Table 2 #1 and main text
yes
no
analysis with converted features
Table 2 #2-6
analysis with aligning common space
Table 2 #7-11

Figure 2. Considerations for choosing an integration method for single cell multiomics analysis.

Various data integration methods can be used depending on the nature of the data and whether they are matched (different modalities were profiled from the same cell) or unmatched (different modalities were profiled from different cells). For unmatched data, analyses can be performed with matched clusters if manual annotations of cell types are available, for example, if we are only interested in the cell-type level relationship between open chromatin and DNA metholation, we can perform clustering and cell type annotation for each modality, and integrate at the level of cell type. If manual annotations are not available or a higher resolution of integration is needed, two different strategies are available depending on whether feature conversion is possible. For data with a common feature set or converted features (e.g., open chromatin to gene activity), tools developed for matching with converted features can be used. For data without common features or feature conversion, integration by aligning common spaces can be applied.

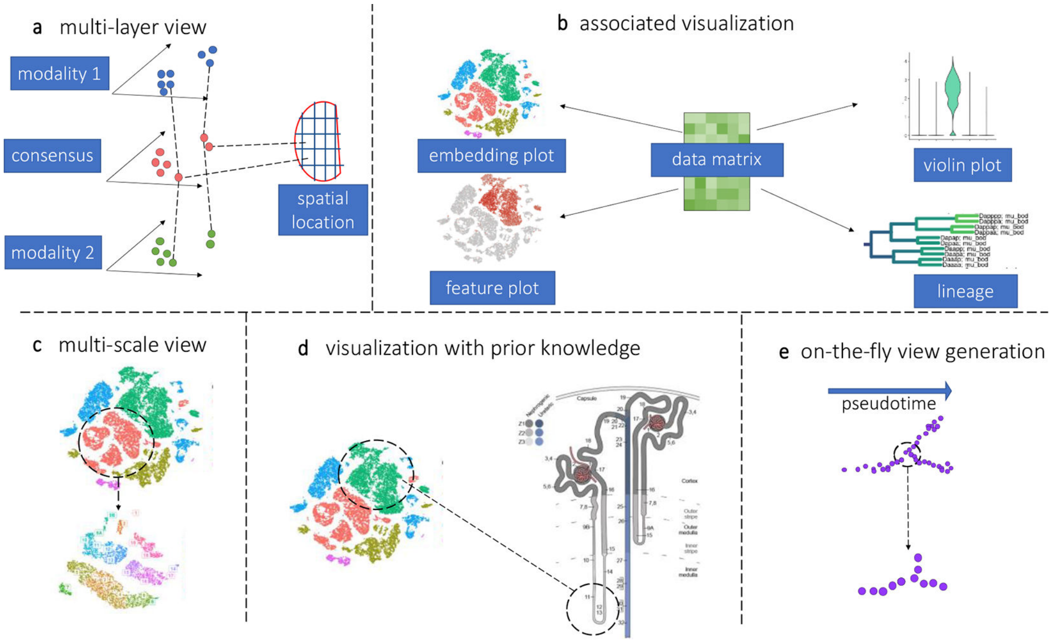

Figure 3. Desired properties and functionalities of visualization tools for single cell

Visualization of multiomics data requires additional functionalities given the complex data structure, for example, the ability to switch the view between different modalities. Some other desirable features include:

a) Multiple layers of data visualization based on data obtained for different modalities with mapping between each layer. Ideally, the mapping between each observation and their spatial location can also be displayed as another layer of information.

b) The addition of knowledge-based visualizations that incorporate down-stream analyses or prior knowledge.

c) Multi-scale views with multiple resolutions to assist the dissection of very large datasets.

d) Integration of prior knowledge such as ontology and anatomy with multiomics data to help anchor biological knowledge to the data

e) Tools that enable on-the-fly or dynamic visualization of data to enable more flexible data visualization

Table 1.

Methods for matched data analysis

Tool	Data type	Model	Additional notes	Documentation	Ref
BREM-SC	T+P	Early integration, probabilistic modeling	This method models the observed data by multinomial distributions and assumes data from both modalities to be generated in a cluster-specific manner	https://github.com/tarot0410/BREMSC	35
scAI	T+C	Early integration, latent space modeling	scAI iteratively updates a regularized matrix factorization model to obtain an optimal common cell loading matrix across two modalities	https://github.com/sqjin/scAI	36
MOFA+	T+C	Early integration, latent space modeling	MOFA and MOFA+ were built upon the framework of group Factor Analysis but extend the model to enable integration of different data types (count vs binary)	https://github.com/bioFAM/MOFA2	38
totalVI	T+P	Early integration, latent space modeling	This method uses a variational autoencoder framework built upon scVI. In this method, the protein measurements are modelled with a negative binomial mixture distribution to account for background reads	https://github.com/YosefLab/scvi-tools	39
CiteFuse	T+P	Late integration, latent space modeling	The similarity measurement for protein data is based on proportionality coefficient, and similarity measurement for RNA data is constructed with Pearson’s correlation	https://github.com/SydneyBioX/CiteFuse	42
Seurat 4.0	T+P	Late integration, latent space modeling	Compute a weighted average cell affinity matrix from modality-specific affinity matrices. The weights are computed to reflect the predictive information within a cell’s local neighborhood defined within each modality.	https://github.com/satijalab/seurat	40

T, transcriptome; C, chromatin accessibility; P, proteome.

Table 2.

Methods for unmatched data analysis

Strategy	Tool	Data type	Feature matching	Algorithm	Additional notes	Documentation	Ref
Group matching	Stereoscope	T+ST	R	Deconvolution	This method assumes negative binomial distributions of genes, and tolerates differential gene capture efficiencies between two technologies	https://github.com/almaan/stereoscope	53
	MAESTRO	T+C	R	CCA+MNN	This method implement ChIP-seq data-based TF enrichment score calculators to define core TFs in each cell type cluster.	https://github.com/liulab-dfci/MAESTRO	49
Comon features	STvEA	MI+ET	R	MNN	This method also provides a framework to transfer cell type annotations from one modality to the other modality	https://github.com/CamaraLab/STvEA	54
	Clonealign	T+D	R	Variational Bayes	This method assumes correlation between DNA copy number and gene expression within the same region	https://github.com/kieranrcampbell/clonealign	56
	Seurat 3.0	T+C	R	CCA +SNN	This method identifies anchor cells between datasets based on shared nearest neighbors across modality. These anchor cells serve as a bridge for matching	https://github.com/satijalab/seurat	57
	LIGER	T+M, T+C	R	iNMF	The relative contribution of dataset-specific factors and shared factors is determined by a hyperparameter λ, which can be used to fine-tune the integration results	https://github.com/welch-lab/liger	58
Aligning spaces	MAGAN	MI+T	R	GAN	This method identifies cell-to-cell correspondence by adding a loss function defined by similarity of cell matching. Such loss function requires at least some shared features between two datasets	https://github.com/KrishnaswamyLab/MAGAN	60
	MATCHER	T+C	NR	Manifold alignment	This method assumes 1D structure (pseudotime) with pre-specified direction	https://github.com/jw156605/MATCHER	61
	MMD-MA	T+M	NR	MMD	In addition to the MMD loss, the loss function also has a distortion loss and a penalty to ensure dimensionality and orthogonality of each projection	https://bitbucket.org/noblelab/2019_mmd_wabi/src/master/	62
	UnionCom	T+M	NR	GUMA	The algorithm generalizes the GUMA method to achieve soft matching between datasets, enabling matching with different number of cells	https://github.com/caokai1073/UnionCom	63
	SCOT	T+C	NR	GWOT	This is a late integration method where similarity matrix is constructed by each modality separately, then, probabilistic transportation between datasets is achieved GWOT.	https://github.com/rsinghlab/SCOT	64

T, transcriptome; ST, spatial transcriptome; MI, multiplexed immunohistochemistry; ET, simultaneous epitope and transcriptome; D, DNA; M, methylome; C, chromatin accessibility; P, proteome; TF, transcription factors

R, required; NR, not required

CCA, canonical component analysis; iNMF, integrative non-negative matrix factorization; GWOT, Gromov-Wasserstein optimal transport; MMD, maximum mean discrepancy; GUMA, generalized unsupervised manifold alignment; SOM, self organizing maps; MMN, mutual nearest neighbors; GAN, generative adversarial networks; SNN, shared nearest neighbors